数理統計学 講義ノート（2011 年，(電情＋医)/2 の 2 年用，担当：原隆）

（このノートは 2011 年 4 月現在の暫定版で，最初の部分しかありません．講義が進むに連れて，すこしずつ加筆訂正されるでしょう．講義ノートの章立ては教科書とは異なります —— 教科書に比べて，かなり細切れ．）

1 確率論の基礎

（教科書の第 2 章から入ります．）まずは確率論の基礎（枠組み）から考えて行こう．

1.1 確率論の舞台 — 事象と標本空間[1]

現実の問題の「確からしさ」を議論するのはなかなか大変である．そこで，数学ではまず，現実から少し切り離した形で，考えやすい舞台を設定する．(確率そのものはもう少し後で導入)．以下のような「実験」[2]を行うことを考える．

例 1： コインを一回だけ投げる．

例 2： コインを 2 回投げる．(この場合，2 回続けて投げたものを一回の「実験」と考える．)

例 3： さいころを一回だけ投げる．

例 4： さいころを 2 回投げる．

例 5： ５２枚あるトランプから一枚取り出す．

このような例では，まず，上の「実験」の結果は何通りかある．一回「実験」をやった場合にその結果が何になるかは分からないが —— だからこそ「確率論」がでてくる ——，少なくとも**可能な結果の全体**はわかっている．そこで，以下の定義を行おう．

定義 1.1.1 「実験」をやる場合，**可能な結果の全体**からなる集合を**標本空間**（sample space）$\mathcal{S}$ と言う．標本空間の元（つまり，一回の「実験」の結果になりうるもの）を**標本点**または**根元事象**と言う．

- 例 1 では $\mathcal{S} = \{H, T\}$．ここで H は表が出ること，T は裏が出ることで，根元事象は T と H．
- 例 2 では $\mathcal{S} = \{(H,H), (H,T), (T,H), (T,T)\}$．ここで例えば (T,H) は一回目に表，2 回目に裏がでること．
- 例 3 では $\mathcal{S} = \{1,2,3,4,5,6\}$．ここで i はさいころの i の面が出ること（$i = 1,2,\ldots,6$）
- 例 4 では $\mathcal{S} = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),\ldots,(6,5),(6,6)\} = \{(i,j) \big| i,j = 1,2,\ldots,6\}$．ここで (i,j) は一回目に i の面，2 回目に j の面が出ること．
- 例 5 では $\mathcal{S} = \{$ ハートのエース，ハートの 2，ハートの 3，$\ldots\}$ と全部で５２個の要素からなる集合．

以下では有限な標本空間，および有限からのアナロジーで考えられる場合のみを考察する[3]．

さて，我々は根元事象のみに興味があるわけではない．たとえば例 2 で,「一回目に表が出ること」を知りたかったり，例 3 で「さいころで偶数の目が出ること」を知りたかったり，例 5 で「ハートが出ること（数字は問わない)」を知りたかったりする．このような問いに答えるため，**事象**と言う概念を導入する．

定義 1.1.2 事象とは実験の結果が持っている性質のこと．数学的に厳密に言うと，**事象**とは単に**標本空間の部分集合**，つまり「根元事象の集まり」のことである．なお，事象には空集合（起こり得ないこと)，および標本空間全体も含めて考える．

「部分集合」と言うと大げさだが，普通に我々の言っている「出来事」に相当していることを，下の例で納得されたい．

[1] 教科書の 2.1 節，a) の 1)
[2] 「実験」と言っているが,「観測」などと思った方が良い場合も含める
[3] 有限でない場合はいろいろとややこしい（＝数学的に面白い）ことが起こるが，すべて略

- 例１では可能な事象は $\varnothing$ （起こり得ない）, $\{H\}$ （「表が出た」）$\{T\}$ （「裏が出た」）, $\mathcal{S}=\{H,T\}$ （「表または裏が出た」）.
- 例２での事象の例は（根元事象で無いものを書くと） $\{(H,H),(H,T)\}$ （「一回目に表が出た（２回目は何でも良い)」）, $\{(H,T),(T,T)\}$ （「２回目に裏が出た（１回目は何でも良い)」）, $\{(H,H),(T,T)\}$ （「２回とも同じ目が出た」）など.
- 例３では $\{1,3,5\}$ （「奇数の目が出た」）, $\{1,2,3,4\}$ （「４以下の目が出た」）など.
- 例４では $\{(1,j)\,\big|\,j=1,2,\ldots,6\}$ （「１回目に１が出た」）, $\{(i,j)\,\big|\,i+j=$ 偶数 $\}$ （「１回目と２回目の数字を足すと偶数」）など.
- 例５では { ハートのエース，ハートの２，ハートの３, ..., ハートの１３ } （「ハートが出た」）, とか { ハートの３，スペードの３，ダイヤの３，クローバーの３ } （「３が出た」）など.

事象を標本空間の部分集合として定義するのは，以下の事象の演算ともあっている．まず，２つの事象 E,F に対して，その**和事象**を集合としての和集合 $E\cup F$ として，またその**積事象**を集合としての交わり $E\cap F$ として定義する（事象の場合，$E\cap F$ を EF と略記することが多い）．日常言語に直せば，$E\cup F$ とは E または F の**どちらかが起こること**，$E\cap F=EF$ とは E と F の**両方が起こること**を意味する．更に，E^c を $\mathcal{S}\backslash E$ （E の補集合）をして定義し，E の **余事象**と言う．これは日常言語では「事象 E が起こらないこと」に相当する．

- 例１で，$E=\{H\},F=\{F\}$ とすると，$E\cap F=\varnothing$．これは「表と裏が同時に起こることは無理」という直感にあっている．$E^c=\{T\}$ であるが，裏が出るというのは「表が出ない」ことでもあるから，これも余事象の定義にあっている．また，$E\cup F=\mathcal{S}$ であるが，これは「表または裏が出る」と言うのは要するに可能性全部だから．
- 例２で，$E=\{(H,H),(H,T)\},F=\{(H,T)\},G=\{(T,H)\},D=\{(T,T)\}$ とすると，$E\cap F=\{(H,T)\}$, $E\cap G=\varnothing$，$E\cup G=\{(H,H),(H,T),(T,H)\}$ などとなる．また，$D^c=E\cup G$ であるが，確かに「『２回とも裏』と言うことはない」という事象になっている．

なお，$A\cap B=\varnothing$ の時,「A と B は互いに背反」という．

1.2 数学における確率[4]

今までは単に確率をやる舞台を設定したにすぎない．これからいよいよ,「確率」を割り振っていこう．

数学ではある意味で「天下りに」確率を定める．本当のところを言うと，確率の定め方そのものは数学の仕事ではなく，実験の行い方に即して物理学・化学・心理学... などに基づいて決めるべきものだ．しかし，通常は確率を定めるところから始めることになる．

ただし，ここでどのような p_j を選ぶか，は個々の問題に応じてうまく決めてやる必要がある．

- 例１で，コインが裏表同じように出やすいのなら，$P(H)=P(T)=1/2$ とするのが良いだろう．
- 例３で，さいころのどの目も同じように出やすいのなら， $P(j)=1/6$ とすべし．しかし，イカサマさいころで 6 が出やすく，1 が出にくい，のなら，例えば $P(1)=\frac{1}{12},P(6)=\frac{3}{12},P(2)=P(3)=P(4)=P(5)=\frac{1}{6}$ ととるのが良いかも知れない．

今までの話を，標本空間が $\mathcal{S}=\{e_1,e_2,\ldots,e_N\}$ になる実験について一般化しておく（e_j が根元事象）．上で見たように，数学的に確率を決めるというのは，それぞれの根元事象の確率（起こり易さ） p_j （$j=1,2,\ldots,N$） を与えることである．それでこの根元事象の起こり易さ（確率）は現実をできるだけ反映するように決めるのだった．

しかし，この根元事象の確率 p_j はいくつかの性質を満たすべきである．まず，これは確率だから 0 と 1 の間にないといけない．更に，$\mathcal{S}$ そのものというのは全事象だから（いつでも起こる）この確率は 1 であるべし．要するに

$$0\le p_j\le 1,\qquad \sum_{j=1}^{N}p_j=1 \tag{1.2.1}$$

[4]教科書の 2.1 節，a の 2) と c の一部

であればよい，ということになる．そして，根元でない事象 $E = \{e_1, e_2, e_3, \ldots, e_m\}$ については，

$$(E \text{ の確率}) = \sum_{j=1}^{m} p_j \tag{1.2.2}$$

となるはずである．と言うのも，E とは「e_1 か，e_2 か，…，e_m の**どれか**が起こる」事象だから，それぞれの事象の確率の和になるのが自然．

これが数学での確率論の出発点である．要するに

- 標本空間 $\mathcal{S}$ 上に根元事象の確率 p_j を (1.2.1) を満たす形で与え，
- 根元事象でない一般の事象 E の確率を (1.2.2) で計算する．

それで，**このルールを満たすものを全て確率と認める**のである．(しつこいが，どのように p_j を選ぶか，は個々の問題に応じてうまく決める．)

さて，上のように決めた「それぞれの事象の確率」はどんな性質を満たしているだろうか？上では根元事象から確率を決めたが，そうでない場合 —— つまり，根元事象の和事象である色々な事象の確率から決めた方が楽な場合 —— も（後でたくさん）出てくる．そのために，(根元事象から出発しない場合にもなりたつ）抽象的な確率の性質を公理としてまとめておく．

定義 1.2.1 (確率の公理) 標本空間 $\mathcal{S}$ が与えられたとき，$\mathcal{S}$ 上の**確率**（または確率測度）とは，以下を満たす関数（数の組） P のこと：$\mathcal{S}$ の部分集合（事象） E のそれぞれについて値 $P[E]$ が定まり，かつ

1. 全ての $E \subset \mathcal{S}$ に対して $0 \le P[E] \le 1$ （確率は E を超えない）
2. $P(\mathcal{S}) = 1$ （全確率は E）
3. E_1, E_2 が**排反**，つまり 「$E_1 \cap E_2 = \varnothing$」，のとき，$P\big[\,E_1 \cup E_2\big] = P[E_1] + P[E_2]$

なお，標本空間 $\mathcal{S}$ とその上の確率測度 P をあわせて**確率空間**と言う．

上の性質を満たしている P なら何でも確率と認めてしまおう，と言うわけ．しつこいけども，実際にどのような P を採用するかは考えている具体的問題によって，適当に決める．

命題 1.2.2 確率について，以下が成り立つ（ベン図を書いて意味を確認しよう）．

$$P[E^c] = 1 - P[E] \qquad (E^c \text{ は } E \text{ が起こらない事象のこと}) \tag{1.2.3}$$

$$E \subset F \quad \Longrightarrow \quad P[E] \le P[F] \tag{1.2.4}$$

$$P[E \cup F] = P[E] + P[F] - P[EF] \tag{1.2.5}$$

根元事象から考えるよりも，他の事象から考えた方が確率を割り振りやすい例として，２枚のイカサマコインを投げる場合を考えよう．2 枚のコインがあり，１枚目は表が p，裏が $1-p$ の確率で出る．２枚目は表が q，裏が $1-q$ の確率で出る，としよう．

このとき標本空間は $\{(H,H),(H,T),(T,H),(T,T)\}$ である．さて，この４つの根元事象にどのように確率を割るふるべきか，だが：１枚目と２枚目の出方は無関係と思うのが良いだろう（数学的には「独立」という；後述）．すると，

$$P[\,1 \text{ 枚目が表}] = p, \qquad P[\,2 \text{ 枚目が表}] = q \tag{1.2.6}$$

ととるのが良いのでは？これは根元事象の言葉では

$$P[\{(H,H),(H,T)\}] = p, \qquad P[\{(H,H),(T,H)\}] = q \tag{1.2.7}$$

と言うことになるね．後，基本的性質から

$$P[\{(T,H),(T,T)\}] = 1 - p, \qquad P[\{(H,T),(T,T)\}] = 1 - q \tag{1.2.8}$$

も言えているわけだ．でもこれだけでは４つの根元事象の確率は決まらない．実際，

$$P[\{(H,H)\}]=a,\quad P[\{(H,T)\}]=b,\quad P[\{(T,H)\}]=c,\quad P[\{(T,T)\}]=d \tag{1.2.9}$$

と書くと，上のは

$$a+b=p,\quad a+c=q,\quad c+d=1-p,\quad b+d=1-q \tag{1.2.10}$$

となって，不定方程式になる．でも，この場合はやはり余分な仮定をおくのが良いだろう．１枚目と２枚目が「独立」なのなら，

$$P[\{(H,H)\}]=P[\,1\text{枚目が表，}\,2\text{枚目も表}]=P[\,1\text{枚目が表}]\times P[\,2\text{枚目が表}]=pq \tag{1.2.11}$$

と考えるのがよいだろう．その他も同様に考えると，

$$P[\{(H,T)\}]=P[\,1\text{枚目が表，}\,2\text{枚目は裏}]=P[\,1\text{枚目が表}]\times P[\,2\text{枚目が裏}]=p(1-q) \tag{1.2.12}$$

$$P[\{(T,H)\}]=P[\,1\text{枚目が裏}]\times P[\,2\text{枚目が表}]=(1-p)q \tag{1.2.13}$$

$$P[\{(T,T)\}]=P[\,1\text{枚目が裏}]\times P[\,2\text{枚目が裏}]=(1-p)(1-q) \tag{1.2.14}$$

となる．

1.3 数の数え方の復習（高校の復習）

（始めに）以下のようなことは頭から覚え込むのではなく，自分で納得して理解するようにすべし．まず記号を導入する．

定義 1.3.1
- $n>0$ に対して，$n!:=n\cdot(n-1)\cdot(n-2)\cdots3\cdot2\cdot1$，また $0!=1$ と定義する．
- $0\le k\le n$ に対して，$\binom{n}{k}:=\dfrac{n!}{k!(n-k)!}$ と定義し，「二項係数」と呼ぶ．
- $0\le n_i\ (i=1,2,\dots,r),\ \sum_{i=1}^{r}n_i=n$ のとき，$\binom{n}{n_1\,n_2\,n_3\,\cdots\,n_r}:=\dfrac{n!}{n_1!\,n_2!\,n_3!\,\cdots\,n_r!}$ を**多項係数**と言う．

さて，上の記号は何に使うかというと：1 から n までの数字を書いた n 枚のカードがあって，これから k 枚を取り出す場合を考える．取り出し方（戻し方）に応じて，大体３とおりある．

Case 1: n 枚のカードから繰り返しを許して k 枚とり，その結果を並べる場合．この場合の結果は $(a_1,a_2,\dots,a_k)$ と言う列になる（a_j は j 番目に出たカードの目）．ここでそれぞれの a_j は勝手に 1 から n の値をとれるので，結果の総数（場合の数）は

$$n\cdot n\cdot n\cdots n=n^k \tag{1.3.1}$$

となる．

Case 2: n 枚のカードから繰り返しを許さないで k 枚とり，その結果を並べる場合．やはり結果は $(a_1,a_2,\dots,a_k)$ の形になるが，今回は a_j は全て別のものにならざるを得ない．a_1 は n 通り，a_2 は a_1 をよけるから $(n-1)$ 通り，と考えて行くと，結果は

$$n\cdot(n-1)\cdot(n-2)\cdots(n-k+1)=\frac{n!}{(n-k)!} \tag{1.3.2}$$

となる．高校ではこの数を ${}_nP_k$ と書いた．

Case 3: n 枚のカードから繰り返しを許さないで k 枚とるが，その順序は気にしない場合．やはり結果は case 2 のように $(a_1,a_2,\dots,a_k)$ の形になるが，今は a_j の順序を気にしない（順序が異なっても同じものと見なす）．従って場合の数は $Case\ 2$ のものを 「k 個の数字を並べる並べ方」$k!$ で割ったものになる：

$$\frac{n!}{(n-k)!}\times\frac{1}{k!}=\binom{n}{k}={}_nC_k \tag{1.3.3}$$

１つだけ，これらの応用例を挙げておく．この証明は帰納法でもできるし， Case 3 の数え方を使う方法もある．

命題 1.3.2 (二項定理，高校でやったかな) $1 \leq n$ では，$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k\, y^{n-k}$.

Case 4. なお，補足的に Case 3 の一般化を考えておく．n 枚のカードを，それぞれ $n_1, n_2, \ldots, n_r$ 枚のカードからなる r 個のグループに分ける場合（$\sum_{i=1}^{r} n_i = n$）．この場合はまず n 枚から n_1 枚を取り出し，次に $n-n_1$ 枚から n_2 枚を取り出し，次に $n-n_1-n_2$ 枚から n_3 枚を取り出し... と考えて

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-n_1-n_2}{n_3} \times \cdots \times 1 = \frac{n!}{n_1!\, n_2!\, n_3! \cdots n_r!} = \binom{n}{n_1\, n_2\, n_3 \cdots n_r} \tag{1.3.4}$$

となることがわかる．

1.4 条件付き確率[5]

前回は確率を考える舞台（標本空間）とその上の確率の満たすべき性質，を導入した．これだけでは簡単すぎて何をやりたいのか混乱した人もいるだろうから，もう少し自明でないものに進むことにする．ここでは「条件付き確率」の概念を導入する．

定義 1.4.1 (独立な事象) 確率空間 $(\mathcal{S}, P)$ 中の事象 E, F が，

$$P[E \cap F] = P[E]\, P[F] \qquad \text{（E と F が起こる確率は E, F それぞれが起こる確率の積）} \tag{1.4.1}$$

を満たすとき，F と E は独立な事象 であると言う．

日常言語で言えば，E と F が独立とは，E と F の**起こり方が無関係**（F が起こっても起こらなくても，E の起こり方には影響がない）と言う場合にあたる（この事情は以下の「条件付き確率」を考えた方がわかりやすいかも）．

E, F が独立でない場合は F の起こり方が E の起こり方に影響しているわけだ．影響の度合いを測るため,「条件付き確率」を導入する．

定義 1.4.2 (条件付き確率) 確率空間 $(\mathcal{S}, P)$ 中の事象 E, F を考える．$P[F] \neq 0$ の場合に，

$$P[\, E \,|F\,] := \frac{P[E \cap F]}{P[F]} \tag{1.4.2}$$

を F の下で E が起こる条件付き確率 と言う.（ベン図で感じをつかもう！）

註 1.4.3 E と F が独立の場合はもちろん，$P[E|F] = P[E]$ となる．これがまさに，E と F が独立なら,「F が起こっても起こらなくても E の起こる確率は変わらない」という意味である．

さて，$P[E]$ そのものよりも $P[E|F]$ と $P[F]$ の方が良くわかる場合が往々にしてある．この場合（条件付き確率の定義からすぐに出てくる式）

$$P[E] = P[E|F]\, P[F] + P[E|F^c]\, P[F^c] \tag{1.4.3}$$

を用いて $P[E]$ を計算することができる．条件付き確率そのものに興味がある場合もあるが，このような計算や後述のベイズ推定において，**条件付き確率を計算の中間段階として利用する**場合も非常に多い．

例 2.A： 袋の中に赤玉が１０個，白玉が３個，黒玉が４個入っている．目をつぶって１つ取り出すとき：

1. 白が出る確率は？
2. 「出た玉は赤ではない」ことがわかった場合，取り出した玉が白である確率は？

[5]教科書 2.1 節の b

例 2.B： 男と女の生まれる確率は $\frac{1}{2}$ ずつとする．A さんちには子供が二人いる．(まあ，探偵がこの家のことをいろいろと調べていると思って下さい．)

1. 二人とも男の子である確率は？
2. 「少なくとも一人が男の子だとわかっている」場合，二人とも男の子である確率は？

例 2.C： 袋の中に赤サイコロが 1 個，白のサイコロが 2 個入っている．白の方は普通の 1 〜 6 が書かれたサイコロだが，赤の方は 1，2，3 が 2 つずつ書かれている変態サイコロである．この袋から目をつぶってサイコロを一つ取り出して転がした．1 の目が出る確率を求めよ．

例 2.D： （これはあくまで例．深読みはしないように）．僕はある大学で 200 人の学生に物理を教えているが，そのうちの 4 割は高校で物理を履修しており，残りの 6 割は未履修である．過去の経験から，僕の物理の講義に受かる確率は,「高校での物理既習者では 0.9，物理未修者では 0.3」と予測される．以上から，僕の物理の講義に受かる学生は 200 人中何人くらいと考えられるか？

例 2.E： 2 個のサイコロ（6 つの面が 1/6 の確率ででるものとする）を一回ずつ転がすことを考える．2 つのサイコロの目が異なる場合，少なくとも一方が 6 をだした確率はいくらか？

1.5 ベイズの公式と推定[6]

ここでは条件付き期待値の，今までとは少し違った解釈を考えよう．これまでの解釈では $P[F\,|E]$ は 「E が起こったという条件の下で F が起こる確率」だったが，新しい解釈として 「E が起こったという情報を知った後で F の確率をどのように設定する（見積もる）のがよいか」を示す式とも考えられる．この節では，このような解釈に基づく推論を考える．

まずは，この節の議論の元になる公式を述べよう．

命題 1.5.1 (Bayes の公式) 確率空間 $(\mathcal{S}, P)$ を考える．すると，$E, F \subset \mathcal{S}$ に対して

$$P[F\,|\,E] = \frac{P[F \cap E]}{P[E]} = \frac{P[E\,|F]\,P[F]}{P[E\,|F]\,P[F] \;+\; P[E\,|F^c]\,P[F^c]} \tag{1.5.1}$$

が成立する．事象が 3 つ以上の場合に一般化すると，事象 F_i （$i = 1, 2, \ldots, k$）が互いに排反（$F_i \cap F_j = \varnothing$ for $i \neq j$），かつ $\bigcup_{i=1}^{k} F_i = \mathcal{S}$ を満たすときは,

$$P[F_j\,|E] = \frac{P[F_j \cap E]}{P[E]} = \frac{P[E\,|F_j]\,P[F_j]}{\sum_{i=1}^{k} P[E\,|F_i]\,P[F_i]} \tag{1.5.2}$$

が成立する．

上の式は単に条件付き確率の定義

$$P[F\,|E] = \frac{P[F \cap E]}{P[E]} \tag{1.5.3}$$

と (1.4.3) の一般化

$$P[E] = \sum_{i=1}^{k} P[E\,|F_i]\,P[F_i] \tag{1.5.4}$$

を組み合わせた**だけ**のものであるから**無理に暗記しない**方がよい．$P[E]$ の計算に (1.5.4) が不可欠な事例が多々あるから，応用上は非常に役立つ．また，解釈としても，左辺は E で条件づけているのに，右辺は F_i で条件付けていて，**条件付けの立場が逆転**しているように見えるのも面白い．

[6]教科書の 2.1 節，d

残念ながら，時間の関係から，ベイズの公式を用いた面白い問題については詳しく述べることはできない．以下に過去の講義で用いた例題をいくつか挙げるにとどめる．

まずは条件付き確率を使った全確率の計算

問 1.5.2 僕はある大学で 200 人の学生に物理を教えている．学生の

- ４割（$= r_1$）は高校で物理 I, II を履修
- ２割（$= r_2$）は高校で物理 I のみを履修
- 残りの４割（$= r_0$）は物理を未履修

である．過去の経験から，僕の物理の講義に受かる確率は，

- 物理 I, II の既習者では 0.9（$= p_1$），
- 物理 I のみの既習者では 0.6（$= p_2$），
- 未修者では 0.3（$= p_0$）

と予測される．以上から，僕の物理の講義に受かる学生は 200 人中何人くらいと考えられるか？

つづいてベイズ型の推定について

問 1.5.3 上の例 2.D や上の問 1.5.2 と同じ状況を考える．僕のクラスの A 君は健闘むなしく，僕の物理の単位が取れなかった．A 君は高校で物理（I まで，II まで？）を履修してきたのだろうか？(物理 II まで履修して来た確率はどのくらいと考えるのが妥当か？)

言うまでもないことであるが，上のような問いかけは余りにも安易である．単位が取れる — より正確には講義内容が身につく — かどうかは多分に本人のやる気や努力によるわけで，高校時代にどれくらいやったかで単純に推し量ることはできない．この問では現実的でないくらいの非常な単純化を行っていることには注意されたい．(将来，実際にこのような手法を用いる際にはくれぐれも単純化のしすぎに注意！)

上の２問が典型的な問題である．以下では数学的には同じ構造であるが応用としては異なった場面を述べる．

問 1.5.4 （再録）かなり稀な病気の血液テストを考える．このテストの誤差の入り方は，

- この病気にかかっている人をテストすると $(1-p)$ の確率で「病気だ」と正しく判定するが，残りの p の確率で見逃してしまう
- 健康な人をテストすると $(1-q)$ の確率で「健康だ」と正しく判定するが，残りの q では（健康なのに）「病気だ」と言ってしまう

となっている．さて，独立な疫学的調査から病気の人の割合は r であるだろうとわかっている（p, q, r はすべてゼロに近いがゼロではない）．

僕の検査結果は陽性（病気だ）だった．僕が本当に病気である確率，健康なのに間違って病気と診断された確率，をそれぞれ求めよ．

問 1.5.5 ○○科目の期末試験は（数学ではあり得ないことに）○×式の問題で，各問は m 個の選択肢から一つ正解を選ぶ形になっています．A 君はかなり怠けていたので，実力で（つまり，まぐれ無しで）正しく答えられる確率は各問毎に p であると思われます（$P < 1/2$）．答を正しく知っているときは勿論，A 君はその正解を答えますが，答がわからないときはヤケクソで m 個の答から等確率で 1 個を選びます．さて，

1. ある一問に対して（まぐれであれ何であれ）A 君が正解を答える確率はいくらでしょう？
2. ある一問をテストしてみたところ，A 君は正解を答えました．このとき，A 君が実際に答を知っていた（まぐれ当たりではない）確率はいくらでしょう？
3. 以上の結果を解釈せよ． どのような p, m の値の場合に「マグレ当たり」が多くなるか，考えてみよう．

問 1.5.6 行方不明の飛行機を捜索中である．現在，墜落した可能性のあるのは 1, 2, 3 の 3 地区に限ること，およびこれらの 3 地区に墜ちている確率は等しい（つまり 1/3）こと，までは絞り込んだ．これから捜索に入るが，厳しい気象条件のため，確実に見つけられる保証はない — 実際に i-地区に墜ちていたとしても，確率 p_i で見逃すだろうと思われる（$p_i \ll 1$）．

まず 1-地区を捜索したところ，飛行機は見つからなかった．この事実から，i-地区に墜ちている確率を推定せよ（$i = 1, 2, 3$）．

問 1.5.7 (Laplace) $i = 0, 1, 2, \ldots, k$ と（非常に小さな）印が付けられた $(k + 1)$ 個のコインが壺に入っている．これらは非常にいびつなコインで，i 番目のコインを投げたときに表が出る確率は i/k となるように調節されている．目隠しをしたままこの壺から一枚のコインを選んで実験をする．以下の問いに答えよ．

1. 取り出したコインを一回投げたところ，表が出た．このコインが i 番目のコインである確率はいくらか？（$i = 0, 1, 2, \ldots, k$）
2. 取り出したコインを更に投げ続け，合計 n 回投げた．結果は全て表だった．このコインが i 番目のコインである確率はいくらか？（$i = 0, 1, 2, \ldots, k$）
3. 取り出したコインを更にもう一回（つまり通算で $(n + 1)$ 回目）投げる事にした．このとき，やはり表が出る確率はいくらか？
4. 上の小問 2, 3 の答はそれほど簡単にならなかったかも知れない．そこでこれらの確率が $k \to \infty$ の極限でどうなるか，求めてみよう．結果は直感と合うだろうか？

（注）この問では，コインは最初に一枚取り出したら，同じ物を使い続ける．コインを何回か投げるとき，一回ごとの結果は独立だとする．また，コインについている印は大変小さいので，取り出したコインがどれかは見ただけではわからないものとする．(そうでないと，小問 2, 3 が面白くない.)

問 1.5.8 3 人の射撃手（$1, 2, 3$）が 200m 離れた，同じ的を狙う．今までの練習成績から，射撃手 i が一発で的に当てる確率はそれぞれ p_i と考えられる（$i = 1, 2, 3$）．さて，3 人が一発ずつ撃ったところ，的には**丁度一発だけ**当たっていた．この当たった一発が射撃手 i のものである（つまり，他の二人ははずした）確率について，以下の問いに答えよ．

1. まず，計算を始める前に，直感的に答を推定してみよう．
2. では，講義での説明に基づき,「正しく」計算してみよう．
3. 2 の結果は直感とあっているか？例えば，$p_1 = 0.2, p_2 = 0.4, p_3 = 0.6$ として，射撃手 1 が当てた確率はいくらになっているか？(勿論，1, 2 の答が一緒になった人は立派なものである．僕にはこの結果は意外だったけどね.)

2 確率変数と期待値

中心極限定理に入る準備として,「確率変数」についての基本事項をまとめておこう.

2.1 確率変数（離散版）[7]

今まではランダムな事象を考えてきた（例：このクラスの学生から一人選んだら男であった，とか）．事象はそれが起こるか起こらないかの２通りしかない．しかし，実際には選ばれた標本の数値的な性質を問題にすることも多い（例：選んだ学生の身長はいくらか）.

このような問題では（我々の注目する）実験の結果が数値で表されている．つまり，実験の結果として**ランダムな数値**が出てくるわけだ．そこで，このようにランダムに値がきまる数値のことを**確率変数**と呼ぶ（ちょっとえーかげん）.

確率変数には「離散的な確率変数」と「連続な確率変数」がある．まずは簡単な「離散的」なものから考える.

「離散的な確率変数」とはとびとびの（有限個の）値しかとらないもので[8]，例は以下の通り.

例 2.1.A： サイコロを一回振る実験を考える．X を出た目の数とすると，X のとりうる値は $1,2,3,4,5,6$ の６通り．また，それぞれの値をとる確率は（マトモなサイコロなら）

$$P[X=1]=P[X=2]=\ldots=P[X=6]=\frac{1}{6} \tag{2.1.1}$$

と考えるのが自然だろう．また，Y を「出た目が４以下なら 0，出た目が５以上なら 10」である確率変数とすると，Y のとりうる値は $0,10$ で，その確率は

$$P[Y=0]=\frac{4}{6}=\frac{2}{3},\qquad P[Y=10]=\frac{2}{6}=\frac{1}{3} \tag{2.1.2}$$

例 2.1.B： サイコロを２個振る実験を考える．Z を出た目の和とすると，Z のとりうる値は $2,3,4,\ldots,12$ の１１通り．また，それぞれの値をとる確率は（マトモなサイコロなら）

$$P[Z=2]=\frac{1}{36},\qquad P[Z=3]=\frac{2}{36}=\frac{1}{18},\qquad \text{（場合が多すぎて書ききれない）} \tag{2.1.3}$$

などとなる.

上の例でもわかるように，離散的な確率変数を記述するには「確率変数のとりうる値」と「それぞれの値をとる確率」を全て与えれば良い．つまり，確率変数 X が $x_1,x_2,\ldots,x_n$ の値をとりうる場合，X がそれぞれの x_i をとる確率，つまり $P[X=x_i]$ （$i=1,2,\ldots,n$）を与えればよいわけだ.

2.2 期待値と分散（離散版）[9]

では，確率変数が与えられたとき，この確率変数の分布をどのように特徴づけたらよいか，考えていこう．もちろん，完全に特徴づけるには，$P[X=x_i]$ を（すべての x_i について）与えないといけない．これは大変すぎるし，そもそも，このようにすべてを知ったとして，分布の特徴がつかめるとは限らない．そうではなくて，**もっと少ない情報量で分布の特徴を捉える**ことを考えたいのだ.

定義 2.2.1 離散的な確率変数 X が $x_1,x_2,\ldots,x_n$ の値をとり，その確率が

$$P[X=x_i]=p_i \qquad \left(\text{もちろん,}\sum_{i=1}^{n}p_i=1\right) \tag{2.2.1}$$

[7]教科書の 2.2 節，a と b 前半
[8]とびとびの値しかとらないけど，全体としては無限個の値をとりうる例もある．が，話を簡単にするため，ここはごまかした
[9]教科書の 2.2 節, b 後半

と与えられているとする．このとき，X **の期待値**を

$$E[X] := \langle X \rangle := \sum_{i=1}^{n} x_i \, p_i \tag{2.2.2}$$

により定義する．(数学では $E[X]$ の記号を，物理などでは $\langle X \rangle$ の記号を用いることが多い．）また，X **の分散**を

$$\mathrm{Var}[X] := E\Big[\big(X - E[X]\big)^2\Big] = E\big[X^2\big] - E[X]^2 = \big\langle X^2 \big\rangle - \langle X \rangle^2 = \Big\langle \big(X - \langle X \rangle\big)^2 \Big\rangle \tag{2.2.3}$$

により定義する．その平方根

$$\sigma := \sqrt{\mathrm{Var}[X]} \qquad （これによると \quad \mathrm{Var}[X] = \sigma^2 \quad となる）$$

を X の**標準偏差**と呼ぶ．

期待値とは，要するに平均値（ただし，p_i の重みを用いた加重平均）のことであり，確率変数の分布の「中心」を表す（どのような意味で中心かは要注意）．

分散とは平均からのズレ（の２乗）の平均だから，分散の平方根（標準偏差）が分布の「拡がり」を表す．

（少し脱線）事象 F の確率を期待値の形で書くことができる．すなわち，関数 $I[F]$ を

$$I[F] := \begin{cases} 1 & (F \text{が起こるとき}) \\ 0 & (F \text{が起こらないとき}) \end{cases} \tag{2.2.4}$$

として定義すると，

$$P[F] = E[\, I[F]\,] = \langle I[F] \rangle \tag{2.2.5}$$

となる．つまり，F の起こる確率は関数 $I[F]$ の期待値 なのである．

教科書の 2.2 節の c には,「代表的な離散確率分布」が載っている．講義でも説明したが，各自で学修しておいてもらいたい．

2.3 確率変数（連続版）[10]

「連続的な確率変数」とは文字通り，連続な値をとりうる確率変数だ．例を見るのが良いだろう．

例 2.3.A： X は区間 $[0,1]$ 内の全ての値を，同じ確率でとりうる確率変数である．

例 2.3.B： Y はこのクラスの学生を一人選んだ場合の学生の身長である（ただし，身長はいくらでも細かく測るものとする）．

例 2.3.C： Z は学研都市の駅で，福岡方面の地下鉄に乗る場合の待ち時間（ただし，時間を計る場合にいくらでも細かく測定するものとする）である．

例 2.3.A では，X のとりうる値は連続無限個あり，これらの確率は同じと仮定しているから，X が特定の値（例：$X = \frac{1}{2}$）をとる確率はゼロだ．(ゼロでなかったら，全確率が無限大になってしまう！)

このように，連続な確率変数を記述するには，離散的な確率変数のような $P[X = x_i]$ を与えるやり方は使えない．仕方がないので， $P[X = x_i]$ に相当するものとして，

$$P[a \le X \le b] = \int_a^b f(x) dx \tag{2.3.1}$$

のように，**確率密度関数** $f(x)$ を用いて積分の形で表すことにする．

[10] 教科書の 2.2 節，d

例 2.3.A の場合は $f(x) = 1$ である．例 2.3.B や例 2.3.C の分布関数は厳密にはわかりそうにないが，大体の感じは書けそうだ．

離散的な場合と同じく，連続な確率変数に対しても期待値や分散を定義する．

定義 2.3.1 連続な確率変数 X（その確率密度関数は $f(x)$）に対しては，(2.2.2) の代わりに X **の期待値**を

$$E[X] := \langle X \rangle := \int_{-\infty}^{\infty} x\, f(x)\, dx \tag{2.3.2}$$

とするにより定義する．また，X **の分散**を

$$\mathrm{Var}[X] := E\Big[\big(X - E[X]\big)^2\Big] = E\big[X^2\big] - E[X]^2 = \big\langle X^2 \big\rangle - \langle X \rangle^2 = \Big\langle \big(X - \langle X \rangle\big)^2 \Big\rangle \tag{2.3.3}$$

により定義する．その平方根

$$\sigma := \sqrt{\mathrm{Var}[X]} \qquad （これによると \quad \mathrm{Var}[X] = \sigma^2 \quad となる）$$

を X の**標準偏差**と呼ぶ．

教科書の 2.2 節の e には,「代表的な連続確率分布」が載っている．講義でも説明したが，各自で学修しておいてもらいたい．

2.4 多変数の確率変数[11]

さて，確率変数が 2 つ以上ある場合を考えよう．まずは離散的な場合から始める．今，確率変数 X が値 $x_1, x_2, \ldots, x_n$ をとり，確率変数 Y が値 $y_1, y_2, \ldots, y_m$ をとるとする．これらがそれぞれの値をとる確率は

$$P[X = x_i \text{ かつ } Y = y_j] = p_{ij} \tag{2.4.1}$$

であるとしよう．

このとき，Y の値は気にしないで，X のみの分布に着目すると，

$$P[X = x_i] = \sum_{j=1}^{m} P[X = x_i \text{ かつ } Y = y_j] = \sum_{j=1}^{m} p_{ij} \tag{2.4.2}$$

となる．これを X の**周辺分布**という．同様に，Y のみの分布は

$$P[Y = y_j] = \sum_{i=1}^{n} P[X = x_i \text{ かつ } Y = y_j] = \sum_{i=1}^{n} p_{ij} \tag{2.4.3}$$

で与えられる．

期待値の重要な性質はその**線形性**である．大事なので，命題の形にまとめておく．(線形性というと大げさだが，要するに以下の命題にある関係式がなりたつということだ.)

命題 2.4.1 確率空間 $(\mathcal{S}, P)$ における確率変数 X, Y と実定数 $a > 0$ に対しては以下が成り立つ：

$$E[X + Y] = E[X] + E[Y] \tag{2.4.4}$$

$$E[aX] = aE[X] \tag{2.4.5}$$

[11] 教科書 2.3 節

$$\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X] \tag{2.4.6}$$

また，X と Y の**共分散**を

$$\mathrm{Cov}(X,Y) := \langle (X - \langle X\rangle)(Y - \langle Y\rangle)\rangle \tag{2.4.7}$$

と定義すると，

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}(X,Y), \tag{2.4.8}$$

もなりたつ．

註： これらの結果は X, Y の分布が独立で**なくても**，いつでも成り立つ．

Proof. 簡単のため，離散の場合のみを考える．X のとりうる値を x_i （$i = 1, 2, \ldots, N$），Y のとりうる値を y_j （$j = 1, 2, \ldots, M$），それぞれの値をとる確率を $P[X = x_i \text{ かつ } Y = y_j] = p_{ij}$ とおく．すると，

$$E[X+Y] = \sum_{ij} p_{ij}(x_i + y_j) = \sum_{ij} p_{ij}x_i + \sum_{ij} p_{ij}y_j \tag{2.4.9}$$

であるが，$\displaystyle\sum_{j=1}^{M} p_{ij} = P[X = x_i \text{ かつ } Y \text{ は何でも良い}] = P[X = x_i]$ であるので，

$$\sum_{ij} p_{ij}x_i = \sum_{i=1}^{N} x_i \Big(\sum_{j=1}^{M} p_{ij}\Big) = \sum_{i=1}^{N} x_i\, P[X = x_i] = E[X] \tag{2.4.10}$$

が成り立つ．同様に

$$\sum_{ij} p_{ij}y_j = E[Y] \tag{2.4.11}$$

なので，$E[X+Y] = E[X] + E[Y]$ が証明された．

次に，$E[aX]$ については，

$$E[aX] = \sum_{i=1}^{N} P[X = x_i](ax_i) = a\sum_{i=1}^{N} P[X = x_i]\, x_i = a\,E[X]. \tag{2.4.12}$$

また，$\mathrm{Var}[aX]$ については

$$E[(aX)^2] = E[a^2X^2] = a^2E[X^2] \tag{2.4.13}$$

であることと線形性から

$$\mathrm{Var}[aX] = E[(aX)^2] - \big(E[aX]\big)^2 = a^2\,E[X^2] - \big(aE[X]\big)^2 = a^2\,E[X^2] - a^2\,\big(E[X]\big)^2 = a^2\,\mathrm{Var}[X]. \tag{2.4.14}$$

(2.4.8) も同様に証明できる． □

確率変数 X と Y が任意の $A, B \subset \mathbb{R}$ に対して

$$P[X \in A \text{ かつ } Y \in B] = P[X \in A]\,P[Y \in B] \tag{2.4.15}$$

を満たすとき， X と Y は**独立**な確率変数と言う．X と Y が**独立**な場合には，

$$E[XY] = E[X]\,E[Y], \qquad \mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] \tag{2.4.16}$$

が成り立つ．

問 2.4.2 さいころを続けて n 回投げることを考える．この n 回のうちに出る**異なった目**の数を N_n としよう．N_n の期待値はいくらか？(注：例えば 5 回投げたとき，$(1,3,2,1,1)$ とでたら，異なった目は $1,2,3$ なので，$N_5 = 3$ と言うこと．)

問 2.4.3 駅の切符売り場や銀行での行列の作り方を考える．窓口は M 個あり，全体で N 人のお客が並んでいる．このとき，

1. 一列待ち：お客を一列に並べておいて，開いた窓口へ誘導していく
2. M 列待ち：お客を勝手に，それぞれの窓口に並ばせる

のどちらが良い（苦情が少ない）だろうか．待ち時間の期待値や分散を考えてみよう．

連続的な確率変数 X, Y がある場合には，その分布は**同時密度関数**$f(x, y)$ を用いて表される．すなわち，

$$P[a < X \leq b \quad \text{かつ } c < Y \leq d] = \int_a^b dx \int_c^d dy\, f(x, y)$$

と書けるような関数 f を X, Y の同時密度関数という．また，Y の値を気にしないで X の分布のみを見る場合には，つまり X の周辺分布は

$$P[a < X \leq b] = P[a < X \leq b \quad \text{かつ} -\infty < Y \leq \infty] = \int_a^b dx \left[\int_{-\infty}^{\infty} dy\, f(x, y) \right]$$

で与えられる．つまり，X の分布密度は

$$f_1(x) = \int_{-\infty}^{\infty} dy\, f(x, y)$$

である．

連続版の確率変数に対しても，期待値の線形性などの命題 2.4.1 はなりたつが，くりかえさない．

3 つ以上の確率変数がある場合も，同様に議論できるが，一言だけ注意を．確率変数 $X, Y, \ldots, Z$ が**独立**であるとは，これらの確率変数の分布が，それぞれの確率変数の周辺分布の積に分解することをいう．つまり，離散の場合に書けば，

$$P[X = x_i, Y = y_j, \ldots, Z = z_k] = P[X = X_i]\, P[Y = y_j] \ldots P[Z = z_k] \tag{2.4.17}$$

となることをいう．

最後に，n 個の確率変数の和の期待値などについてまとめておく．まず，期待値の線形性から

$$\langle X_1 + X_2 + \cdots + X_n \rangle = \langle X_1 \rangle + \langle X_2 \rangle + \cdots \langle X_n \rangle \tag{2.4.18}$$

である．これは X_j が独立でなくても，いつでも成り立つ事はすでに強調した．特に，$X_1, X_2, \ldots$ が全く同じ期待値をもつならば，

$$\langle X_1 + X_2 + \cdots + X_n \rangle = n \langle X_1 \rangle \tag{2.4.19}$$

となる．つまり，n この和の期待値は期待値の n 倍になる．これは自然．

次に分散に移る．残念ながら，一般の n 個の確率変数の分散は簡単には書けない．Cov が一杯出て来るからだ．しかし，**確率変数がすべて独立ならば**事情は簡単になる．この場合，Cov がすべて 0 になるので，

$$\mathrm{Var}[X_1 + X_2 + \cdots + X_n] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \cdots + \mathrm{Var}[X_n] \tag{2.4.20}$$

と，分散も和に分解できる．特に，n 個の確率変数の分散がすべて等しいなら，

$$\mathrm{Var}[X_1 + X_2 + \cdots + X_n] = n\mathrm{Var}[X_1] \tag{2.4.21}$$

となる．するとこの場合，標準偏差については，

$$\sqrt{\mathrm{Var}[X_1 + X_2 + \cdots + X_n]} = \sqrt{n} \times \sqrt{\mathrm{Var}[X_1]} \tag{2.4.22}$$

となる．n 個の和であるのに，標準偏差は $\sqrt{n}$ 倍であることに注意しよう．

以前に,「標準偏差は分布のバラツキの度合いを表す」事を注意した．上の結果によると，n この和の分布のバラツキは n 倍ではなく，$\sqrt{n}$ 倍になる訳だ．この事実はこれから非常に重要になって来る．

2.5 チェビシェフの不等式とその仲間[12]

今までにも,「標準偏差は確率変数のばらつきの目安を与える」と言ったが，ここではもう少し定量的な議論を行う．ここでも確率空間 $(\mathcal{S}, P)$ 上の確率変数 X を考える．

まず，$A \in \mathbb{R}$ について

$$P[a \le X \le b] = \langle I[a \le X \le b] \rangle \tag{2.5.1}$$

であることに注意しておこう．ここで $I[\cdots]$ とは，カッコの中の $\cdots$ が満たされているときに 1，満たされていないときに 0 である関数である．

命題 2.5.1 (マルコフの不等式) 正の値のみをとる確率変数 X と任意の正の数 a に対して，

$$P[X \ge a] \le \frac{\langle X \rangle}{a} \tag{2.5.2}$$

が成立.（勿論，右辺の期待値が存在しないときは右辺には意味がないけど.）

命題 2.5.2 (チェビシェフ の不等式) 確率変数 X の期待値を μ，分散を $\mathrm{Var}[X]$ と書くと，任意の正の数 a に対して，

$$P[|X - \mu| \ge a] \le \frac{\mathrm{Var}[X]}{a^2} \tag{2.5.3}$$

が成立.（勿論，右辺の分散が存在しないときは右辺には意味がないけど.）

これらの不等式は勿論，右辺の期待値が存在しなければ意味がないが，存在する場合には（特に $a \to \infty$ について）強力なものになる．実際の応用については後述.

（証明の概略）これらの不等式は (2.5.1) を用いると簡単に証明される．マルコフの不等式なら

$$\langle X \rangle \ge \langle X\, I[X \ge a] \rangle \ge \langle a\, I[X \ge a] \rangle = a\, \langle I[X \ge a] \rangle = a\, P[X \ge a]. \tag{2.5.4}$$

チェビシェフの不等式なら

$$\mathrm{Var}[X] = \big\langle |X - \mu|^2 \big\rangle \ge \big\langle |X - \mu|^2, I[X \ge a] \big\rangle \ge \big\langle a^2\, I[X \ge a] \big\rangle = a^2\, \langle I[X \ge a] \rangle = a^2\, P[X \ge a]. \tag{2.5.5}$$

□

（以下はおまけ）調子に乗って似たような不等式を作ることもできる．例えば,

$$P[|X - \mu| \ge a] \le \frac{\langle |X - \mu|^n \rangle}{a^n} \qquad (a > 0,\ n \text{ は任意の正の整数}) \tag{2.5.6}$$

同様に，任意の $a, b > 0$ に対して

$$P[|X - \mu| \ge a] \le \frac{\big\langle e^{b|X-\mu|} \big\rangle}{e^{ab}}. \tag{2.5.7}$$

また，マルコフの不等式の仲間として,（X が非負の値しかとらないとき）

$$P[X \ge a] \le \frac{\big\langle e^{bX} \big\rangle}{e^{ab}} \tag{2.5.8}$$

など.

2.6 正規分布について[13]

正規分布とは一般に（μ を実数，σ は正の数として）

$$P[a \le X \le b] = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Big[-\frac{1}{2}\big(\frac{x-\mu}{\sigma}\big)^2\Big]\, dx \tag{2.6.1}$$

[12]教科書には該当部分はない
[13]教科書の 2.4 節

を満たすような分布のことを言う.（これは $N(\mu,\sigma^2)$ と書かれる.）また，上のような分布をもった確率変数 X は正規分布に従う確率変数という.

実際に計算してみるとすぐにわかることだが，上の正規分布の期待値は μ，分散は σ^2，標準偏差は σ である.

特に，$\mu=0,\sigma=1$ の正規分布を「標準正規分布」とよぶ．通常

$$\Phi(x) := \int_{-\infty}^{x} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \tag{2.6.2}$$

と書く．以下に $1-\Phi(x)=\displaystyle\int_x^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$ のいくつかの値を載せておく：

x	0	1	1.645	1.960	2	2.326	2.576	3	4
$1-\Phi(x)$	$\frac{1}{2}$	0.1587	$\frac{1}{20}$	$\frac{1}{40}$	0.02275	$\frac{1}{100}$	$\frac{1}{200}$	1.350×10^{-3}	3.167×10^{-5}

さて，積分の変数変換を用いると，一般の正規分布の分布確率を標準正規分布の分布確率から求めることができる．つまり，X が $N(\mu,\sigma^2)$ に従うときに，新しい確率変数

$$Z := \frac{X-\mu}{\sigma} \tag{2.6.3}$$

を定義すると Z が標準正規分布になることが容易にわかる．もちろん，この場合 X と Z のズレを考慮して

$$P[a\le X\le b] = P\Big[\frac{a-\mu}{\sigma}\le Z\le \frac{b-\mu}{\sigma}\Big] \tag{2.6.4}$$

とやる必要はあるが.

ともかく，このようなわけで，いろいろある正規分布は，標準正規分布になおして計算できる.

3 大数の法則と中心極限定理

さていよいよ，推定や検定の基本となる「大数の法則」「中心極限定理」について学ぶ．ここでは以下のような典型的な問題を考える．

問 3.A：マトモ（表と裏が $\frac{1}{2}$ の確率で出る）な硬貨を 10000 回投げたとしよう．表は何回くらい出るだろうか？もちろん，答は 0 回から 10000 回まで，どれでもアリだけど，このうちのどの答が一番ありそうだろうか？また，そのありそうな答えになる確率はどうだろう？

この節では上のような問題を主に考える．上では硬貨の例を取り上げたが，もっと一般に「独立な」実験の結果を扱う．次の第 5 章以降では，このような問題の逆に相当する，以下の問題を考える．

問 3.B：ある硬貨を 10000 回投げたら，表が 4500 回出た．この硬貨が表を出す確率 p はどのくらいと考えられるか？

問 3.C：ある硬貨を 10000 回投げたら，表が 1000 回だけ出た．この硬貨はマトモ（表・裏とも確率 $\frac{1}{2}$ で出る）であると思って良いか？

これらの問題に共通するのは 独立な確率変数の和 の振る舞いを見ようとしていることである．以下に用語の意味も含めて説明していこう．

3.1 大数の法則[14]

問 3.A を考える．我々は直感的に「表は 5000 回」と言いたくなるが，既に断ったように，5000 回きちんと出るとは言えない．言えるのはあくまで「○○回以上が表になる確率はこのくらい小さい」「出る回数は 5000 回を中心にこのくらいでばらつく」などという確率評価である．

少しだけ抽象的になるが，定理の形で書いておく．まず，考える対象（独立な確率変数の和）を導入する．考えるのは $X_1, X_2, X_3, \ldots$ という確率変数の列で，特にその和 $S_n := \sum_{i=1}^{n} X_i$ を考える．硬貨を投げる例では，X_i は i 回目に投げた硬貨の結果（表なら $X_i = 1$，裏なら $X_i = 0$ と決める）で，この場合 S_n は「硬貨を n 回投げたときに表の出た回数」を表す．

更にここで，確率変数 $X_1, X_2, \ldots$ は「独立」かつ「同分布」だと仮定する．

> 確率変数 $X_1, X_2, \ldots$ が **独立**であるとは，X_1 の結果と X_2 の結果と，X_3 の結果と… が全く無関係であることをいう．硬貨の例では，一回目の結果によって，２回目以降の結果が左右されない，などのことを言う．一応正確な定義を書くと，確率変数 X が値 x_i を確率 p_i で，確率変数 Y が値 y_j を確率 q_j でとる時（$i = 1, 2, \ldots, n$，$j = 1, 2, \ldots, m$），X, Y が独立であるとは，
>
> $$P[\,X = x_i \text{ かつ } Y = y_j\,] = P[X = x_i]\, P[Y = y_j] \qquad \text{（確率が積になる）}$$
>
> ことを言う．
>
> 確率変数 $X_1, X_2, \ldots$ が **同分布**であるとは，X_i がとりうる値とその確率が i によらず同じであることを言う．

上のように書くとヤヤコシイが，要するに硬貨やサイコロを何回も投げる場合の i 回目の結果が X_i だと思えばよい．

このとき，大数の法則は以下のようになる．

[14] 教科書 2.5 節の a

Theorem 3.1.1 (大数の弱法則) 独立・同分布な確率変数の列 $X_1, X_2, \ldots$ と $S_n := \sum_{i=1}^{n} X_i$ を考える．X_i の期待値を μ，X_i の分散を $Var[X_1]$ と書くと，

$$Var[X_1] < \infty \text{ ならば} \qquad \lim_{n\to\infty} P\Big[\frac{S_n}{n} \neq \mu\Big] = 0 \tag{3.1.1}$$

が成り立つ（上のはちょっとえーかげんな書き方）．より正確にはどんな正の数 $\epsilon > 0$ に対しても

$$P\Big[\,|\frac{S_n}{n} - \mu| > \epsilon\,\Big] \leq \frac{Var[X_1]}{n\,\epsilon^2} \tag{3.1.2}$$

が成り立つ．

定理の形にするとややこしいが，要するに $\boxed{S_n/n \text{ は } n\to\infty \text{ で } \mu \text{ に収束する}}$ と主張しているわけだ．これは我々の直感を支持するものである．例えばマトモな硬貨を何回も投げると，大体半分くらいが表になるだろう．上の定理は「硬貨を無限回（！）投げると，その半分くらいは表だよ」と主張していることになる．

（対数の弱法則の証明の “説明”）

先週やったチェビシェフの不等式を確率変数 $\frac{S_n}{n}$ に応用するだけなのだが，それには $\frac{S_n}{n}$ の期待値と分散を計算しないといけない．そこで，確率変数 $X_1, X_2, \ldots$ の和である S_n について，その期待値や分散がどうなるか，考えてみよう．重要なので命題の形にまとめると：

命題 3.1.2 確率空間 $(\mathcal{S}, P)$ における確率変数 X, Y と実定数 $a > 0$ に対しては以下が成り立つ：

$$E[X+Y] = E[X] + E[Y], \qquad E[aX] = aE[X] \tag{3.1.3}$$

$$\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X] \tag{3.1.4}$$

また，X, Y が**独立**である場合には以下が成り立つ：

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]. \tag{3.1.5}$$

これを認めて対数の弱法則を証明しよう．上の線形性から

$$E[S_n] = \sum_{i=1}^{n} E[X_i] = n\mu, \qquad E\Big[\frac{S_n}{n}\Big] = \mu \tag{3.1.6}$$

および

$$\mathrm{Var}[S_n] = \sum_{i=1}^{n} \mathrm{Var}[X_i] = n\mathrm{Var}[X_1], \qquad \mathrm{Var}\Big[\frac{S_n}{n}\Big] = \frac{1}{n^2}\mathrm{Var}[S_n] = \frac{1}{n}\mathrm{Var}[X_1] \tag{3.1.7}$$

を得る．よってチェビシェフの不等式に代入して

$$P\Big[\,|\frac{S_n}{n} - \mu| > \epsilon\,\Big] \leq \frac{1}{\epsilon^2}\mathrm{VarVar}\Big[\frac{S_n}{n}\Big] = \frac{\mathrm{Var}[X_1]}{n\,\epsilon^2} \tag{3.1.8}$$

（大数の弱法則の証明の説明終わり）

上の要点は，S_n の分散が n に比例してしか増えない（よって S_n/n **の分散は** $1/n$ **に比例して** $n\to\infty$ **でゼロに行く**）ことである．分散（の平方根）というのは確率変数のばらつきの程度を表すから，分散がゼロになると言うことは S_n/n がその平均値からばらつかない，ことを意味する．これが上の証明とチェビシェフの不等式の意味だった．

（命題 3.1.2 の証明；興味のある人だけ見ればよい）X のとりうる値を x_i $(i = 1, 2, \ldots, N)$，Y のとりうる値を y_j $(j = 1, 2, \ldots, M)$，それぞれの値をとる確率を $P[X = x_i \text{ かつ } Y = y_j] = p_{ij}$ とおく．すると，

$$E[X+Y] = \sum_{ij} p_{ij}(x_i + y_j) = \sum_{ij} p_{ij}x_i + \sum_{ij} p_{ij}y_j \tag{3.1.9}$$

であるが，$\sum_{j=1}^{M} p_{ij} = P[X = x_i \text{ かつ } Y \text{ は何でも良い}] = P[X = x_i]$ であるので，

$$\sum_{ij} p_{ij} x_i = \sum_{i=1}^{N} x_i \Big(\sum_{j=1}^{M} p_{ij} \Big) = \sum_{i=1}^{N} x_i \, P[X = x_i] = E[X] \tag{3.1.10}$$

が成り立つ．同様に

$$\sum_{ij} p_{ij} y_j = E[Y] \tag{3.1.11}$$

なので，$E[X+Y] = E[X] + E[Y]$ が証明された．

次に，$E[aX]$ については，

$$E[aX] = \sum_{i=1}^{N} P[X = x_i](a x_i) = a \sum_{i=1}^{N} P[X = x_i] \, x_i = a \, E[X]. \tag{3.1.12}$$

また，$\mathrm{Var}[aX]$ については $E[(aX)^2] = E[a^2 X^2] = a^2 E[X^2]$ であることと線形性から

$$\mathrm{Var}[aX] = E[(aX)^2] - \big(E[aX]\big)^2 = a^2 \, E[X^2] - \big(aE[X]\big)^2 = a^2 \, E[X^2] - a^2 \left(E[X]\right)^2 = a^2 \, \mathrm{Var}[X]. \tag{3.1.13}$$

(3.1.5) の証明はスペースの都合で略． □

硬貨投げの例に戻って考えよう．この場合，$E[X_i] = \frac{1}{2}, \mathrm{Var}[X_i] = \frac{1}{4}$ であるので，大数の弱法則から

$$P\Big[\big| \frac{S_n}{n} - \frac{1}{2} \big| > \epsilon \Big] \leq \frac{1}{4n\,\epsilon^2} \tag{3.1.14}$$

が得られる．

(練習問題)

問題 3.1.3 マトモな（どの面も同じ確率で出る）サイコロを何回も投げることを考え，i 回目に出た目を X_i で表す．

- X_i の期待値と分散，標準偏差を求めよ．
- $S_n := \sum_{i=1}^{n} X_i$ とするとき，$\frac{S_n}{n}$ の期待値と標準偏差を求めよ．
- 大数の弱法則を用いて，$n \to \infty$ の時に $\frac{S_n}{n}$ がどのような値になりそうか，議論せよ．

問題 3.1.4 （少しムズイかも：次節へのつなぎ）３つの小問からなるテストがある．それぞれの小問は４つの選択肢から１つの正解を選ぶ選択式である．全く勉強していない学生達が当てずっぽうでテスト問題に答えることを考える．

- 一人の学生が当てずっぽうでこれらの問題に答えた場合，正解した小問の数を X で表そう．X の期待値と分散，標準偏差を求めよ．
- N 人の学生がこのテストを受けた場合の正解された小問の総数を S_N と書く．S_N の期待値と分散，標準偏差を求めよ．(ヒント：i 番目の学生が正解した小問の数を X_i で表すと，$S_N = \sum_{i=1}^{N} X_i$ とかける．)
- 大数の弱法則を用いて，$n \to \infty$ の時に $\frac{S_n}{n}$ がどのような値になりそうか（要するにこれらの学生達の平均点はどのくらいか)，議論せよ．

（上の問題では学生は互いに答案を見せあったりしないものとする — これは数学の言葉で何の条件を満たさせるためかわかるかな？）

3.2 正規分布と中心極限定理[15]

前節では大数の法則をやった．これは要約すると，

分散が有界な独立・同分布な確率変数 $X_1, X_2, \ldots$ の和を考え（X_i の期待値を μ），
$S_N := \sum_{i=1}^{N} X_i$ とすると，$\lim_{N\to\infty} P\Big[\frac{1}{N}S_N \neq \mu\Big] = 0$ が成り立つ

と言うものだった．更にその証明（チェビシェフの不等式を使った）によると，S_N がその平均値の周り $\sqrt{N}$ くらいのところに集中していった．そこで，集中していく様子をもっと細かく見たい，と思うのが人情であり．これに答えてくれるのが中心極限定理である．この定理はこれからの検定・推定の議論の基礎になる，非常に重要なものである．

定理 3.2.1 X_i （$i = 1, 2, \ldots$）を独立，かつ同分布な確率変数とし，その平均と，標準偏差をそれぞれ

$$\mu := E[X_i], \qquad \sigma := \sqrt{\mathrm{Var}[X_i]} \tag{3.2.1}$$

とする．このとき，

$$S_N := \sum_{i=1}^{N} X_i, \qquad Z_N := \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{N} \big(X_i - \mu\big) = \frac{S_N - \langle S_N\rangle}{\sigma\sqrt{N}} \tag{3.2.2}$$

を定義すると，任意の $a < b$ に対して

$$\lim_{N\to\infty} P\Big[\, a \le Z_N \le b \,\Big] = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \tag{3.2.3}$$

が成り立つ．

上の定理の主張をもう少し述べておく．S_N や $S_N - N\mu$ 自身は N 個のものの和だから，N が大きくなると（普通は）大きくなる．けれども，$S_N - N\mu$ の大きくなり方は N に比例するのではなく，$\sqrt{N}$ に比例する，と言うのが前節までの話だった．そこで上の定理では $S_N - N\mu$ を $\sqrt{N}$ で割ることによって Z_N を定義した．こうすることで，$N \to \infty$ でも（大抵は）有限にとどまるような量を定義したわけである．それで，定理は，この Z_N が $N \to \infty$ で「標準正規分布」に近づいていくことを主張している．

本来ならばここで中心極限定理の証明をすべきだが，これはこの講義のレベルを遙かに超えている．代わりに実例を挙げ，中心極限定理は証明無しに認めてもらうことにする．

二項分布

中心極限定理の一番簡単な例として，前回と同じく，コインを何回も投げることを考えよう．(ただし，一回投げたときに表の出る確率は p とする．) i 回目に表が出れば 1，裏が出れば 0 となる確率変数を X_i と書くと，$S_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ は N 回のうちで表が出た回数である．N 回のうち，丁度 m 回だけ表になる確率は

$$P\big[S_N = m\big] = \binom{N}{m} p^m (1-p)^{N-m}, \qquad \binom{N}{m} := {}_N C_m := \frac{N!}{m!\,(N-m)!} \tag{3.2.4}$$

と計算できる．上の分布を（パラメーターが p の）「二項分布」と言う．(ここで上の導出を説明)．

さて，上の二項分布について平均と分散を計算してみよう．定義通りに行うと（$q := 1 - p$），

$$\langle X_1\rangle = 1\cdot p + 0\cdot(1-p) = p, \qquad \mathrm{Var}X_1 = (1-p)^2\cdot p + (0-p)^2\cdot(1-p) = p(1-p) = pq \tag{3.2.5}$$

[15]教科書 2.5 節の b

であるので，中心極限定理に出てくる Z_N は

$$Z_N := \frac{S_N - Np}{\sqrt{pqN}} \tag{3.2.6}$$

となるはずである．実際に $N \to \infty$ に従って Z_N が正規分布に近づいていく様子は次ページに載せてある．(標語的には「二項分布は N が大きいときに正規分布に近づく」と言える．)

問題 3.2.2 問 3.1.3 と同じく，マトモな硬貨を N 回投げる．表の出る回数が投げた回数の 49% から 51% に入る確率を，中心極限定理を用いて考えたい．$N = 100, 1000, 10000$ に対して，この確率がどのような積分で表されるか，求めよ．(注：積分そのものの値は計算できないと思うので，やらなくて良い．)

問題 3.2.3 問 3.2.2 の続き．今度は「表の出る回数が投げた回数の 49% から 51% にほとんど確実に入る」ような N を求めたい．「ほとんど確実」と言うのはいい加減な書き方だから，具体的に「表の出る回数が投げた回数の 49% から 51% に入る確率が 0.95 以上になる」ような，そんな N を求めよ．

中心極限定理の使い方について．

問 3.1.3. これはやるだけ，ね．

X_i は 1 から 6 までの値を確率 $\frac{1}{6}$ ずつでとるから，

$$\langle X_i \rangle = \frac{1}{6} \times (1+2+3+4+5+6) = \frac{7}{2}, \qquad \left\langle (X_i)^2 \right\rangle = \frac{1}{6} \times (1^2+2^2+3^2+4^2+5^2+6^2) = \frac{91}{6}, \tag{3.2.7}$$

$$\mathrm{Var}[X_i] = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}, \qquad \sigma = \sqrt{\frac{35}{12}} \tag{3.2.8}$$

で，X_i が独立であるから

$$\left\langle \frac{S_n}{n} \right\rangle = \langle X_1 \rangle = \frac{7}{2}, \qquad \mathrm{Var}\Big[\frac{S_n}{n}\Big] = \frac{1}{n}\mathrm{Var}[X_1] = \frac{1}{n}\frac{35}{12} \tag{3.2.9}$$

大数の弱法則から

$$P\Big[\Big|\frac{S_n}{n} - \frac{7}{2}\Big| > \epsilon\Big] \le \frac{35}{12} \times \frac{1}{\epsilon^2\, n} \tag{3.2.10}$$

である．つまり，$\frac{S_n}{n}$ は $\frac{7}{2}$ に近づく．

問 3.1.4. これは二項分布になる．4 項目から 1 つを当てずっぽうで選択する，のだから，小問の一つ一つに正解する確率は $\frac{1}{4}$ と考えられる．各小問の結果が独立であると仮定すると，正解の数が i である確率は（$i = 0, 1, 2, 3$）

$$P[X = i] = \binom{3}{i}\left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{3-i} \tag{3.2.11}$$

である．これから定義通りに計算して，

$$\langle X \rangle = \frac{3}{4}, \qquad \mathrm{Var}[X] = \frac{9}{16}, \qquad \sigma = \frac{3}{4} \tag{3.2.12}$$

X の独立性から，

$$\langle S_N \rangle = N\,\langle X \rangle = \frac{3}{4}N, \qquad \mathrm{Var}[S_N] = N\,\mathrm{Var}[X] = \frac{9}{16}N. \tag{3.2.13}$$

大数の弱法則から平均の正解数は $\frac{3}{4}$.

問 3.2.2. 中心極限定理を使うには，まず Z_N を作らないといけない．i 回目に表が出れば $X_i = 1$，裏が出れば $X_i = 0$ とすると，$S_N = \sum_{i=1}^{N} X_i$ と書けるから，今までに考えてきた形である．さて，

$$\langle X_i \rangle = \frac{1}{2}, \qquad \mathrm{Var}X_i = \frac{1}{4}, \qquad \sigma = \frac{1}{2} \tag{3.2.14}$$

であるから，中心極限定理にでてくる Z_N は

$$Z_N = \frac{S_N - \frac{N}{2}}{\sqrt{\frac{1}{4}N}} = \frac{2S_N - N}{\sqrt{N}} \tag{3.2.15}$$

となっている．さて，表が 49% から 51% 出る，と言うことは

$$0.49 \le \frac{S_N}{N} \le 0.51 \qquad \Longleftrightarrow \Big|\frac{S_N}{N} - \frac{1}{2}\Big| \le \frac{1}{100} \qquad \Longleftrightarrow |Z_N| \le \frac{\sqrt{N}}{50} \tag{3.2.16}$$

と言うことだ．だから，中心極限定理を少しええ加減に使うと，この確率は

$$P\Big[\,0.49 \le \frac{S_N}{N} \le 0.51\,\Big] = P\Big[\,|Z_N| \le \frac{\sqrt{N}}{50}\,\Big] \approx \int_{-\frac{\sqrt{N}}{50}}^{\frac{\sqrt{N}}{50}} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \tag{3.2.17}$$

となるわけだ．(ヤヤコシイが，積分の上下は $\frac{\sqrt{N}}{50}$.) N に具体的な数を入れると，

$$N = 100 \text{ なら} \quad \int_{-1/5}^{1/5} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \approx 0.1585, \tag{3.2.18}$$

$$N = 1000 \text{ なら} \quad \int_{-\sqrt{10}/5}^{\sqrt{10}/5} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \approx 0.4729, \tag{3.2.19}$$

$$N = 10000 \text{ なら} \quad \int_{-2}^{2} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \approx 0.9545, \tag{3.2.20}$$

(最後の積分の値は数値的に出したもので，皆さんに対しては要求しない.)

問 3.2.3. 今度は

$$\int_{-\frac{\sqrt{N}}{50}}^{\frac{\sqrt{N}}{50}} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \ge 0.95 \tag{3.2.21}$$

となるような N を求めればよい．この積分は手計算ではできないから，この前のプリントにあった Φ で書き直し，表を使うしかない．定義から

$$\Phi(x) = \int_{-\infty}^{x} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \tag{3.2.22}$$

であった．(3.2.21) の積分を上の Φ で表すには，一般に（講義で説明）

$$\int_a^b e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = \int_{-\infty}^{b} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} - \int_{-\infty}^{a} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = \Phi(b) - \Phi(a) \tag{3.2.23}$$

とするのが良い．特に $a < 0$ の場合は，対称性から

$$\int_{-\infty} ae^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = \int_{-a}^{\infty} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = 1 - \Phi(-a) \tag{3.2.24}$$

を使う．結局,

$$\int_{-\frac{\sqrt{N}}{50}}^{\frac{\sqrt{N}}{50}} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = \Phi\Big(\frac{\sqrt{N}}{50}\Big) - \{1 - \Phi(\frac{\sqrt{N}}{50})\} = 2\Phi(\frac{\sqrt{N}}{50}\Big) - 1 \tag{3.2.25}$$

となる．よって，(3.2.21) の条件は

$$2\Phi\Big(\frac{\sqrt{N}}{50}\Big) - 1 \ge 0.95 \qquad \Longleftrightarrow \qquad 1 - \Phi\Big(\frac{\sqrt{N}}{50}\Big) \le 0.025 = \frac{1}{40} \tag{3.2.26}$$

となる．前回のプリントの表を見ると，こうなるには

$$\frac{\sqrt{N}}{50} \ge 1.960 \qquad \Longrightarrow N \ge (50 \times 1.960)^2 = 9604 \tag{3.2.27}$$

となる．まあ，余り細かいことを言っても仕方ないので，$N \ge 9600$ ぐらい，と言うのが答え．

4 推定と検定（おおざっぱに）

漸くこれで統計に入る準備が整った．この講義のもう一つのテーマである,「推定と検定」に入ろう．

4.1 考える問題

以下の問題群を考えてみる．

> 問 4.1．コインを 10 回投げたところ，10 回も表が出た．どう見たってこのコインはイカサマ（表が出やすい）と思うが，これは本当か？

正直この問には**確実には**答えることが出来ない．マトモなコインを投げても，**たまたま全部おもてになることもある**から，**確実に**イカサマだとは結論できない．でも，常識的に考えて，**ほとんど間違いなく**イカサマだと思うよね．どのくらいの確率でイカサマだと言えるか（言えないか），を考えるのが「仮説検定」の問題である．

> 問 4.2．上と同じ問題（10 回中 10 回とも表が出た）で，コインを一回投げて表が出る確率 p はどのくらいと考えるのが自然か？

イカサマかどうか，の定性的な問だけでなく，どのくらいイカサマか，を問うのが上の問題である．この場合，p の値をぴったりこれくらい，と言うことは出来ないだろう．出来るのはせいぜい,「p は大体○○以上，××以下と考えられる」と言う感じの，p の存在範囲（存在区間）を与える事である．これが「区間推定」の問題である．

以下ではこのような２つの問題を主に考える．もちろん，もっと複雑な状況を考えもするが，基本的なところはこれで尽きている．

4.2 仮説検定

まずは上の問 4.1 を考える．再掲するとこんな問題だった：

> 問 4.1．コインを 10 回投げたところ，10 回も表が出た．どう見たってこのコインはイカサマ（表が出やすい）と思うが，これは本当か？

この問題を解くため，以下のように考えてみる．このコインがマトモかどうかを問題にしているのだから，**マトモだと仮定して 10 回表になる確率を計算し**，その結果が大きいか小さいかでマトモかどうかを推定するのである．

具体的に計算すると，まともなコインを 10 回投げて 10 回とも表になる確率は，言うまでもなく

$$\left(\frac{1}{2}\right)^{10} \approx 9.77 \times 10^{-4} \approx 10^{-3} \tag{4.2.1}$$

である．これは非常に小さい！でも問 5.1 では，こんなに小さい確率で起こるはずの事象が起こったと主張している．この場合，以下の２通りの可能性があり，どちらかを排除することはできない：

a. コインは**マトモ**なのだが，10^{-3} と言うような，小さな確率でしか起こらない事象が，**たまたま**起こってしまったのだ．
b. いやいや，コインはそもそもかなり**いびつで，表が出やすく**できていた．そのために 10 回とも表になったのだ．

しつこいが，確実にどちらかとは言い切れない．どちらも起こり得て排除できないことは，いくら強調してもしすぎることはない．ただし，前者の確率は非常に小さいので，a の立場をとるのは心情的にもかなり困難である．

仮説検定ではこの心情的な見方の通りに判断し，上の立場 a は認めず，立場 b を採用する．つまり，

1. 9.77×10^{-4} なんて確率で起こるはずの事象は非常に起こりにくい．
2. でもそいつが実際に起こったんだ．
3. 実際に起こった事象に対して 9.77×10^{-4} などという確率を与えた確率計算は間違っている！
4. より正確には，その確率計算の元になった仮説「コインがマトモ」が，そもそもおかしい．だから，この仮説「コインはマトモ」は葬り去られるべきだ，

と考えるのだ．

ここで少し言葉を導入すると共に，上でやったことを整理しよう．

- 上での本音は「このコインはイカサマだ」を主張することだった．この本音の主張を**対立仮説**と言う．
- しかし，上ではこの対立仮説そのものは扱わず，その否定命題[16]の「このコインはマトモだ」に基づいていろいろと計算した．その結果,「コインはマトモ」の仮説がおかしいと結論した．このように対立仮説の否定命題（結果的に「その仮説はおかしい」と結論したいもの）を**帰無仮説**と言う．こう呼ぶのは上で見たとおり，結果的に「この仮説はおかしい」と否定される（無に帰する）はずの命題だから．
- 帰無仮説を疑う際に使ったのは「(この帰無仮説で計算すると）起こった事象の確率がこんなに小さくなる．でも，実際にこの事象が起こっている．だから，元の帰無仮説は許せない」という考えだった．この場合,「どのくらい小さい確率の事象は起こるはずがないと判断するか」の境目を決めておく必要があり，この境目の値を**危険率**（または有意水準）と言う．危険率は通常，α で表す．
  通常，危険率は 0.05 または 0.01 くらいにとる．つまり，確率 5% または 1% くらいの事象は「あり得ない，起こるはずがない」と判断し，その確率計算の元になった仮説を疑いにかかるのである．
- 言うまでもなく，危険率をどうとるかは，**仮説検定を行う前**に決めておくべきである．検定の結果を見て,「いやいや，危険率が高すぎたからもうちょっと下げよう」などとやるのは，自分の導きたい結論に（無意識のうちにでも）誘導している可能性が大だから，やってはいけない．

仮説検定についての，非常に重要な注意：

上の例では「コインはマトモ」という仮説が最終的に否定された[17]ので，めでたしめでたしだった．つまり，この場合，帰無仮説が否定（無に帰する）されたので，対立仮説が復活し,「このコインはイカサマ」と結論できたわけである[18]．では，**帰無仮説が否定できない場合**はどうなるのだろうか？

例として,「コインを 10 回投げると，6 回が表，4 回が裏であった」場合に，やはりイカサマかどうか考えてみよう．今までのように,「コインはマトモ」を帰無仮説として計算してみると,「6 回が表，4 回が裏」の確率は

$$\binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = \frac{210}{2^{10}} \approx 0.2051 \tag{4.2.2}$$

となる．確率 0.2 と言うのはまあ，5 回に 1 回と言うことだから，そんなに珍しいことでもないわけで，これではもう,「コインはマトモ」を否定する訳にもいかないだろう．では何が結論できるのか？

この場合，何も結論できない．もっと正確には,「『コインはマトモ』の仮説（帰無仮説）を**否定することはできない**」が結論である．しかししかし,（国会答弁ではないが），**否定することはできない**と言うのはほとんど何も言っていないに等しい.「コインはマトモ」を否定してやろうと思って計算したが，結果としては**証拠不十分で結論が出なかった**，ということだ．特に,「コインはマトモ」であると積極的に主張しているのでは**決してない**．

このように「帰無仮説が否定できない」場合は「何も結論できない」と考えるのが正しい．決して「帰無仮説が採用できる」の**ではない**，ことをしつこく強調しておく．

---

[16] ここは，論理学の意味での厳密な否定命題になっていない場合もある．でもまあ大体否定命題と思って良い．時間があれば後述

[17] 非常にしつこいけども,「否定された」とは言っても，これは「否定するのが確率的に自然である」という意味である．実際には「コインはマトモ」なのに，間違ってこれを否定してしまっている可能性は，わずかではあるが，存在する

[18] しつこいけども，もう一回．結論できたとは言っても,「イカサマの可能性が高い」という結論である．確実にイカサマとは言い切れないのは何度も強調した通りだ

しつこいけど，もう一つ注意.

上でもくり返し強調したように，帰無仮説を否定した場合（上の「10 回とも表ならマトモではない」）でも，この結論が間違っている可能性はある．まともな硬貨だって，9.77×10^{-4} の確率では「10 回とも表」が出るわけだから，これだけの確率で間違って「イカサマ」の烙印を押してしまうわけだ．上での「危険率」と言う呼び名は，正にこれを指している．つまり,「帰無仮説が正しいのにも関わらず否定してしまう」確率の目安が危険率なのだ.

少し用語の説明

以上の仮説検定では，以下の 2 種類の過った判断をする可能性がある：

- まず,「本当は帰無仮説が正しいのに，帰無仮説が間違って棄却されてしまう誤り」を**第一種の過誤**という．危険率はこの「第一種の過誤」が起こる確率である．コインの例で言えば，コインが本当は fair であるのに，たまたま 10 回続けて表になってしまい,「fair ではない」と濡れ衣を着せてしまう場合が第一種の過誤にあたる.
- 次に,「本当は帰無仮説は間違っているのに，実験結果がそれを否定するほど強くないために，帰無仮説を棄却できなかった」場合もある．これを**第二種の過誤**という．コインの例では,「このコインは表が出やすいように作ってあるのだが，10 回投げたら 5 回表になった」（従って，コインをイカサマと見抜けなかった）ような場合が第二種の過誤にあたる.

いくつかの例題

問 4.3．コインを N 回投げたとき，N 回とも表が出た．このコインがイカサマである，と結論できるような N の範囲を，危険率 5% および 1% のそれぞれの場合について求めよ（対立仮説，帰無仮説を明示して議論すること）.

問 4.4．コインを 10 回投げたら n 回表が出たという．このコインがイカサマであると結論できるような n の範囲を，危険率 5% および 1% のそれぞれの場合について求めよ（対立仮説，帰無仮説を明示して議論すること）.

4.2.1 片側検定，両側検定

上の例題の解答から始めよう.

問 4.3 の場合，対立仮説（本音）は「コインはイカサマ」，帰無仮説は「コインはマトモ」であるが，この「イカサマ」の意味が問題だ．想定しているのは「表が沢山出たから，このコインは表が出やすく作ってあるのでは？」と言うような場合である．つまり，この場合の対立仮説は単に「イカサマ」と言うよりは,「表が出やすい」または「裏が出やすい」のどちらかと思った方が良い — この意味で対立仮説は厳密には帰無仮説の否定にはなっていない．これは問 4.4 で問題になる.

帰無仮説の「コインはマトモ」を仮定して計算すると，N 回のうちで N 回とも表になる確率は 2^{-N} である．これが 0.05 より小さくなるのは $N \geq 5$ の時で，この場合は危険率 5% で「イカサマ」と結論できる．同様に，$N \geq 7$ なら危険率 1% で「イカサマ」と結論できる． □

問 4.4 には非常に注意すべき所がある．その前に，X を表が何回出たか，の変数として，確率を計算しておくと

$$P[X=0] = P[X=10] = \frac{1}{1024}, \quad P[X=1] = P[X=9] = \frac{10}{1024}, \quad P[X=2] = P[X=8] = \frac{45}{1024}, \tag{4.2.3}$$

$$P[X=3] = P[X=7] = \frac{120}{1024}, \quad P[X=4] = P[X=6] = \frac{210}{1024}, \quad P[X=5] = \frac{252}{1024} \tag{4.2.4}$$

である.

さて，対立仮説（本音）「コインはイカサマ」，帰無仮説「コインはマトモ」の持っている意味は問 4.3 と同じだが，これが重要である．

例として，$n = 9$ だったとしてみよう.「9 回表」となる確率だけを考えると，これは 1% より小さいから,「コインはマトモ」の帰無仮説を棄却できるように見える．しかし，$n = 9$ の場合に棄却するのなら，$n = 10$ が出ても，当然棄却すべきだろう．と言うわけで，この場合に問題になるのは

$$P[X \geq 9] = \frac{11}{1024} \approx 0.0107 \tag{4.2.5}$$

なのだ．これは 0.01 より大きいから，この場合,「コインがマトモ」は棄却できないのだ！

このように考えると，危険率 1% で「イカサマ」と判断できるのは $n = 0, 10$ の時のみである．また，危険率 5% の場合は，$n = 0, 1, 9, 10$ の時のみ「イカサマ」と判断できる．($n = 2, 8$ の場合は上と同じ理由で棄却できない.)

上で「9 回表」の確率だけを考えず,「9 回以上表」とした理由については，コインを 10000 回（または一億回）投げることを考えると納得しやすい．この場合，投げた回数が非常に多いため，それぞれ特定の回数だけ表が出る確率は非常に小さくなる．例えば，10000 回投げた場合に 5000 回表，の確率は 0.00798 くらいであって，1% より小さい．ちょっと考えると「コインはマトモ」を棄却してしまいたくなるが，投げた回数の丁度半分しか表が出ていないのだから，このコインをイカサマと判断するのはおかしい．この場合，問題になるのは個々の確率の値ではなく，結果（表の回数）がこの区間に入っていればおかしい（イカサマ）と言う区間（棄却域）の設定である．

問 4.3 も同じ問題であるが，この場合，全部表だったので，問題が表面化することはなかった．

言葉：

上の問のように，対立仮説がある区間の片側に出てくるような場合を**片側検定**と言う．両側に出る場合を**両側検定**と言う．後者の例としては対立仮説が左右対称になっている場合，たとえば工場の製品が規格にあっているかどうかの検定などが挙げられる．

4.2.2 中心極限定理との連携

以下の例題を考える（問 4.4 の 1000 回バージョン）．

> 問 4.6．コインを 10000 回投げたら n 回表が出たという．このコインがイカサマであると結論できるような n の範囲を，危険率 5% および 1% のそれぞれの場合について求めよ（対立仮説，帰無仮説を明示して議論すること）．

この問題，10000 回も投げて一つ一つの n について確率を計算していると大変だから，中心極限定理を使う．今回も $n = 5000$ を中心に片側の確率を考えて（以下，$n < 5000$ とする），

$$P[X \leq n] < 0.01 \quad \text{（危険率）} \tag{4.2.6}$$

などとなるような n を求めたいわけだ．

さて，中心極限定理を使おう．コインがマトモだとすると（$\sigma = 1/2$），

$$Z_N = \frac{2}{\sqrt{N}}\Big(n - \frac{N}{2}\Big) = \frac{n - 5000}{50} \tag{4.2.7}$$

が中心極限定理に出てくる Z_N だ（$N = 10000$ だよ）．従って，中心極限定理のズルをすると

$$\begin{aligned} P[X \leq n] &\approx \int_{-\infty}^{(n-5000)/50} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} \\ &= \int_{(5000-n)/50}^{\infty} e^{-z^2/2}\frac{dz}{\sqrt{2\pi}} = 1 - \Phi\big(\frac{5000-n}{50}\big). \end{aligned} \tag{4.2.8}$$

$1-\Phi(x)$ の表から，危険率 1% の時には

$$\frac{5000-n}{50} \geq 2.326 \qquad \Longrightarrow \qquad n \leq 4880 \tag{4.2.9}$$

なら「イカサマ」と判断できることがわかる．(実は $n=5000$ の両側であるから $n \geq 5120$ くらいでもイカサマと判断できる．)

同様に危険率 5% ならば

$$\frac{5000-n}{50} \geq 1.645 \qquad \Longrightarrow \qquad n \leq 4920 \tag{4.2.10}$$

または $n \geq 5080$ くらいでイカサマ，と判断される．

問 4.7. 6 面あるサイコロを 10000 回投げたら 1 の目が 2000 回も出た．このサイコロはイカサマと結論できるだろうか？

4.3 区間推定

では問 4.2 に進もう．ここでも基本的な考え方は同じである．$p=\frac{1}{2}$ だとしたらあり得ないほど確率が小さくなったのは上で見た．ので，別の p の値をいろいろ設定し,「10 回とも表」の確率が危険率を超えるような p の範囲を求めればよい．答は当然，危険率の設定によるわけだが，ともかくまず，いろいろな p の値に対して,「10 回とも表」の確率を計算してみよう．

p の値	0.5	0.6	0.6309574	0.7	0.7411345	0.8	0.9	1.0
「10 回とも表」の確率	0.000977	0.00605	0.010000025	0.0282	0.050000034	0.107	0.349	1.0

これを見ると，もちろん，p が大きい方が「10 回とも表」が起こりやすくなっている．(極めつけは $p=1$ で，この場合は絶対に「10 回とも表」だ.）であるけども，$p=\frac{1}{2}$ を含め，p が余り小さいと,「10 回とも表」の確率はかなり小さく，ほとんど起こり得ない．そこで危険率と相談しながら,「起こりうる確率」になるような p の値を求めていく．

確率 0.01 とはかなり小さいが，ひとまずこれを危険率に設定しよう．要するに，0.01 以上の確率を持つ事象は起こっても良い，と判断するわけだ．表を見ると，$p \geq 0.6309574$ ぐらいでは「10 回とも表」の確率が 0.01 を超えており，ここでは確率 0.01 以上で「10 回とも表」になりうる．つまり,「10 回とも表」になったのはあり得る事象で，従って元にした仮説 $p \geq 0.6309574$ は許される．結果としてあり得る p の範囲は

危険率 1% では $p \geq 0.6309574$

になることがわかる．同様に,

危険率 5% では $p \geq 0.7411345$

となることもわかる．

これが区間推定の基本的な考え方である．要するに，起こった事象（問の場合は「10 回とも表」）が「あり得る事象」になるような（つまり，その事象の確率が危険率を超えるような）パラメータの範囲を求めればよい．(コインの例をもっと簡単に解く方法はあとで詳しく説明する．ここでは基本的な考えをわかればよい.)

註： しつこいけども，もう一回強調しておく．上で推定された区間はあくまで「p がこの範囲なら『10 回とも表』の結果と矛盾はない」と言うことで，それ以上ではない．

4.4 より実用的な区間推定

問 4.4′. あるコインを 400 回投げたら，220 回表，180 回裏が出た．このコインが表を出す確率 p はどのくらいと考えるのが妥当か？

400 回中，220 回表なのだから，$p \approx \frac{220}{400}$ くらい，と推定するのは当然だ．しかし，$\frac{220}{400}$ からどれくらい外れる可能性があるのだろうか？ここを区間推定で考えたい．

少し状況を一般化しておく．これまでと同じように，独立，同分布の確率変数 $X_1, X_2, X_3, \dots$ がある．コイン投げの場合などと対応させるには，X_j が j 回目の実験の結果だと思えば良い．さて今，N 回実験をやって，その結果が $X_1, X_2, \dots, X_N$ となったとしよう（確率変数と同じ記号を使うが混乱はないだろう）．この結果から，元の確率変数の分布について何かを言いたい．特に，その期待値 $\mu = E[X_1]$ や標準偏差 $\sigma = \sqrt{\mathrm{Var}[X_1]}$ について何が言えるか？

ここで N 回の結果から作られる量を 2 つ，定義しておく．まず，

$$\bar{X}_N = \frac{1}{N}\sum_{j=1}^{N} X_j \tag{4.4.1}$$

を**標本平均**という．また，

$$\bar{V}_N = \frac{1}{N-1}\sum_{j=1}^{N}\big(X_j - \bar{X}_N\big)^2 \tag{4.4.2}$$

を**不偏分散**という．これらは N が正の整数なら

$$\big\langle \bar{X}_N \big\rangle = \langle X_1 \rangle = \mu, \qquad \big\langle \bar{V}_N \big\rangle = \mathrm{Var}[X_1] = \sigma^2 \tag{4.4.3}$$

を満たす．更に，これらに対しては大数の強法則がなりたつ．つまり，$\bar{X}_N$ と $\bar{V}_N$ は，$N \to \infty$ で μ と σ^2 に収束するのである．

この事実から，N がかなり大きければ，μ の推定値として $\bar{X}_N$ を用いようとするのは自然なことであろう．問題は，この推定値がどのくらい正しそうか（言葉を変えれば，この推定値のまわりどのくらいの区間に真の μ がありそうか）ということである．

そのために，中心極限定理を（少し誤摩化して）利用する．中心極限定理によれば，N 個の確率変数 $X_1, X_2, \dots, X_N$ から新しい確率変数

$$Z_N = \frac{1}{\sigma\sqrt{N}}\sum_{j=1}^{N}\big(X_j - \mu\big) \tag{4.4.4}$$

を作ると，$N \to \infty$ **の極限で**，この Z_N は正規分布に従う，つまり，

$$\lim_{N\to\infty} \mathbb{P}\,[\,a < Z_N < b\,] = \int_a^b \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \tag{4.4.5}$$

ということであった（ただし，μ と σ は今のところ未知）．上の等号はあくまで**極限でしか**成立が保証されていないが，ここで大胆に，極限をとる前でも等号がなりたつとむりやり思ってみよう：

$$\mathbb{P}\,[\,a < Z_N < b\,] \approx \int_a^b \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz \tag{4.4.6}$$

さらに，ここに出ている Z_N は N 回の実験結果 $X_1, X_2, \dots, X_N$ から作られたものだと考えてみる．すると，Z_N の定義から

$$\sigma\sqrt{N}Z_N = \sum_{j=1}^{N}\big(X_j - \mu\big) = N\bar{X}_N - N\mu \qquad \Longrightarrow \qquad Z_N = \frac{\sqrt{N}}{\sigma}\left(\bar{X}_N - \mu\right) \tag{4.4.7}$$

の関係があることがわかる．

さて，我々は，得られた実験結果から作った上の Z_N が，$\mathbb{P}[a < Z_N < b] \geq 1-\alpha$ を満たしてほしい —— 満たさないような場合は，非常に起こりにくいものとして棄却する．ここで中心極限定理を無理矢理適用すると，このような a, b は

$$1-\alpha = \int_a^b \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \tag{4.4.8}$$

となるような a, b で与えられる．このような a, b は $a = -b$ にとるのが自然であるから，実際には，

$$1-\alpha = \int_{-b}^b \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz \tag{4.4.9}$$

となる正の数 b を一つ決めると，$|Z_N| < b$ でなければならない，ということになる．これを $\bar{X}_N$ と μ の言葉に直すと

$$\frac{\sqrt{N}}{\sigma}\left|\bar{X}_N - \mu\right| < b \qquad \text{つまり} \qquad |\bar{X}_N - \mu| < \frac{b\sigma}{\sqrt{N}} \tag{4.4.10}$$

となる．**もし，我々が何らかの理由で σ を知っているなら**，上式から直ちに μ の推定範囲が定まる．

実際には，σ はわからないことが多い．この場合には，次節の t-分布を用いることが多いが，N が大きければ，以下のようにも議論できる．すなわち，対数の強法則によれば，$\bar{V}_N$ は $N \to \infty$ で σ^2 に収束するのである．よって，上の式の σ を $\sqrt{\bar{V}_N}$ でおきかえてしまっても良いかもしれない．こうすると，μ の推定範囲として以下を得る．

$$|\bar{X}_N - \mu| < b\sqrt{\frac{\bar{V}_N}{N}} \tag{4.4.11}$$

4.5 更に進んだ話題

> 問 5.5．福岡の高校一年生（男子）の平均身長を求めたい．全員を計るのは大変なので，天神で 100 人を捕まえて計ったところ，平均が 170 cm だった．福岡の高校一年生（男子）の平均身長はどのくらいと考えられるか？

前に宣言したように，この問はそのままでは解けない．というのは元になる分布（高校生の身長の分布）の分散がわからないからである．ここを無理矢理，不変分散で置き換えてしまう方法を前節では紹介した．

この節ではもう少し良い方法 — t-分布 — を紹介する．

まず，中心極限定理の主張は，X_i が独立・同分布で各 X_i の期待値が μ，分散が σ^2 なら，

$$Z_N := \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^N (X_i - \mu) \tag{4.5.1}$$

は（$N \to \infty$ で）正規分布に従う，ことだった．さてここで上の Z_N に似ているけども少しだけ違う，

$$Y_N := \frac{1}{\sqrt{N\bar{V}_N}} \sum_{i=1}^N (X_i - \mu) = \sqrt{\frac{N}{\bar{V}_N}}(\bar{X}_N - \mu), \qquad \bar{X}_N := \sum_{i=1}^N X_i, \qquad \bar{V}_N := \frac{1}{N-1}\sum_{i=1}^N \left(X_i - \bar{X}_N\right)^2 \tag{4.5.2}$$

を考える．すると非常に都合の良いことに，(元々の X_i が独立・同分布で，何かの正規分布に従うならば)，Y_N は自由度 $(N-1)$ の t-分布 と呼ばれる分布に従うことがわかる．t-分布が具体的にどんな分布かは教科書や参考書を見てもらうことにして，ここでは t-分布の利点だけを強調しておく．

(4.5.1) を見ると，右辺には未知の量が２つ（μ と σ）入っている．高校生の問題では μ を求めたかったのだが，σ （高校生全体の標準偏差）がわからない限り手が出せない．ところがところが（！）(4.5.2) の Y_N の表式の右辺には，未知の量は μ 一つしか出ていない！σ の代わりに $\bar{V}_N$ という量が出ているが，これは $\bar{X}_N$ 同様，標本（100 人の高校生）から計算できる量である．

つまり，t-分布のウリは，σ がわからなくても何とかなる，ところにあるわけだ．実際の応用では個々の X_i は正規分布に従わない場合がほとんどだが，そこは中心極限定理と同じノリで，N が大きければ Y_N が t-分布に従うとみなしても良いわけだ．

5 推定と検定（各論）

前節で推定，検定のおおざっぱなところを述べたので，この節では，個々の状況に応じて，実際にどのように使って行くかをまとめて行く．

5.1 1 標本問題（自由度大）

（この節は教科書の 3.2 節に相当する.）

まず，一番簡単な実例として，一標本問題を考える．ただし，中心極限定理を使いたいので，大標本の（＝標本の大きさが大きい）場合を考える．

5.1.1 平均の推測（大標本 — 母分散は標本分散で置き換え)

（概要）標本の大きさが大きい場合を考える．この場合，母分散を標本分散で近似して良いので，話がかなり簡単になる．母分散も推定する方法は後で述べる．

（状況の例）非常に大きな母集団（例：日本人の 20 歳男子の全体）がある．そこから n 人を取り出して身長を測った．この測定結果から，日本人 20 歳男子の全体の身長の平均を知りたい．

（用語）

- 取り出して測った n 人を**標本**，n を**標本の大きさ**，n 人の平均を**母平均**という．
- 一方，元になった集団（日本人の 20 歳男子の全体）を**母集団**，母集団の平均を**母平均**という．
- 分散，標準偏差についても，標本のものか，母集団のものかを,「標本」や「母」を頭に付けて区別する．
- この節で考える問題は，標本の大きさ n の標本でのデータから母集団の母平均を求めよ，ということである．

> 上の状況で，n をどう呼ぶかには用語の混乱がある．この n は正しくは「標本の大きさ」「標本サイズ」というらしい．しかし，人によっては，これを「標本数」ということもあり，実際，教科書の一部にはそう書いてある部分がある．一方で，標本数というのは，n 人とりだすことを何セット行ったか，の「何セット」を意味する場合もあり，混乱の元である．この点を考え，6/30 の講義以降では，n は「標本の大きさ」ということにした．
> なお，上の状況では,「n 人からなる 1 セット」を取り出している訳だが，これを無理矢理，1 人からなる標本を n セット取り出した，と思えない事もない —— ただし，その際には，同じ人が 2 回以上入ってしまう事もあるから，厳密には「n 人 1 セット」とは少しズレるのだが．この意味で,「標本数 n」と言っても，そんなにおかしな事にはなっていない．用語が混乱しがちな理由はこの辺りにも原因があるように思う．
> また，慣習として，大文字 N は母集団のメンバーの数を表す事がおおく，標本の大きさは小文字の n にする事が多い．この点を考慮し，教科書にも併せて，小文字の n を使う．

（更に用語）標本平均や標本分散などは，以下のように定義する．標本の大きさ n の標本を考える．測定している量（今の例なら身長）のデータが n 個あるから，それを x_j $(j = 1, 2, \ldots, n)$ とする．このとき

$$\bar{\mu} = \text{標本平均} := \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad \bar{V} = \text{標本分散} := \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{\mu})^2, \qquad \bar{\sigma} = \text{標本標準偏差} := \sqrt{\bar{V}} \tag{5.1.1}$$

と定めておく（教科書では，標本平均は $\bar{X}$，標本標準偏差は S と書いている）．我々の問題は，上のようにデータから求まる標本平均などを用いて，母集団の平均を推定せよ，ということだ．

（解法）中心極限定理と大数の法則を近似的に使いまくる．

$\bar{\mu}$ の分布がどうなるかを考えてみると，大数の法則と中心極限定理より，平均値 μ，分散は σ^2/n の正規分布になる（少なくとも n が大きい場合）はずである —— 分散を n で割っている理由は，$\sum_{j=1}^{n} X_j$ の分散は $n\sigma^2$ であるので，それを n で割った $\bar{\mu}$ の分散は $n\sigma^2/n^2$ になるからである．

このままでは，σ が未知なので困るのだが，n が大きければ，母集団の標準偏差 σ と標本の標準偏差 $\bar{\sigma}$ はかなり近いと考えて良い．従って，

$$Z' := \frac{\bar{\mu} - \mu}{\bar{\sigma}/\sqrt{n}} = \sqrt{n} \times \frac{\bar{\mu} - \mu}{\bar{\sigma}} \tag{5.1.2}$$

は，大体，標準正規分布と思っても良いだろう．従って，この Z' は確率 0.95 で，

$$|Z'| < 1.96 \tag{5.1.3}$$

の範囲にあると思って良いだろう．これを元の変数に直すと，

$$|Z'| < 1.96 \qquad \Longrightarrow \qquad \left|\sqrt{n} \times \frac{\bar{\mu} - \mu}{\bar{\sigma}}\right| < 1.96 \qquad \Longrightarrow \qquad |\mu - \bar{\mu}| < \frac{1.96 \times \bar{\sigma}}{\sqrt{n}} \tag{5.1.4}$$

となる．最後の式が，確率 0.95 で，母平均 μ の推定区間を与える式である（我々の持っているデータは $\bar{\mu}, \bar{\sigma}$ であったことを思い出そう）．

（用語）上の推定区間を μ の 95% 信頼区間 という．つまり，(有限の n なのに中心極限定理を用いてごまかしたところは目をつぶったとして）確率 0.95 で上の範囲の中に μ が入ってるよ，というわけだ．残りの 0.05 の確率で上の区間からはみ出しているわけだが，これはマア，推定や検定にはつきものなので，仕方ないと考える．

なお，0.95 という確率では足りない，もっと信頼度を上げたい，というのであれば,「99% 信頼区間」を考えることもある．この場合は，上の 1.96 の代わりに，2.576 をとればよい．

5.1.2 平均についての仮説の検定（大標本 — 母分散は標本分散で置き換え)

やっていることは先の小節と全く同じである．

先の小節と同じ状況（n 人の身長を測った）を考え,「母平均 μ は 175cm である」などという仮説が正しいかどうかの検定を考えてみよう．175cm を一般に μ_0 と書くことにすると，検定したいのは

$$H_0 : \mu = \mu_0 \text{ である} \tag{5.1.5}$$

という仮説（帰無仮説）である.(有意水準が 5%の場合を考える.）この仮説が正しいとするならば，

$$Z' := \frac{\bar{\mu} - \mu_0}{\bar{\sigma}/\sqrt{n}} = \sqrt{n} \times \frac{\bar{\mu} - \mu_0}{\bar{\sigma}} \tag{5.1.6}$$

は標準正規分布と思って良いはずだ．そこで，対立仮説を適切にとって，検定を行う．

(a) 両側検定，つまり対立仮説が $H_1 : \mu \neq \mu_0$ の場合は Z' が 0 から両側に遠くはなれていて，その離れ具合が確率 0.05 以下なら，H_0 を棄却する．具体的には

$$|Z'| > 1.96 \qquad \text{つまり} \qquad |\mu_0 - \bar{\mu}| > \frac{1.96 \times \bar{\sigma}}{\sqrt{n}} \tag{5.1.7}$$

ならば，H_0 を棄却し，$\mu \neq \mu_0$ と結論する.(ただし，しつこく強調したように，実は $\mu = \mu_0$ なのに，濡れ衣をきせて H_0 を棄却してしまった可能性は 5% くらいある）．

(b) 片側検定，つまり対立仮説が $H_1 : \mu > \mu_0$ の場合は Z' が正の方に大きくて，その実現確率が 0.05 以下なら，H_0 を棄却する．具体的には

$$Z' > 1.65 \qquad \text{つまり} \qquad \bar{\mu} - \mu_0 > \frac{1.65 \times \bar{\sigma}}{\sqrt{n}} \tag{5.1.8}$$

なら H_0 を棄却して，$\mu > \mu_0$ と結論する．

(c) 片側検定の $H_1 : \mu < \mu_0$ の場合には，

$$Z' < -1.65 \qquad \text{つまり} \qquad \bar{\mu} - \mu_0 < -\frac{1.65 \times \bar{\sigma}}{\sqrt{n}} \tag{5.1.9}$$

なら H_0 を棄却して，$\mu < \mu_0$ と結論する．

5.1.3 比率の推定（二項分布を仮定）

（状況）表が出る確率が p であるコインを n 回投げたら，m 回，表がでた．表の出る確率 p を推定したい．
（状況′）ある工場での製品 n 個中，m 個が不良であった．この工場での不良率 p を推定せよ．

（解法）コイン投げの例で考える（工場の例でも全く同じ．適宜，読み替えて下さい）．最も単純には，n 回中 m 回の表なので，p は以下の

$$\bar{p} := \frac{m}{n} \tag{5.1.10}$$

くらいだろう，と考えたい．問題は，真の p の値が，上の $\bar{p}$ の周りにどのくらいふらついて分布しているのか，ということだ．以下のように考える．

n 回中の表の回数 X は二項分布に従う確率変数で，その確率分布は

$$P[X = m] = \binom{n}{m} p^m (1-p)^{n-m} \tag{5.1.11}$$

で与えられるはずだ．ところが，この「j 回目に表なら 1, 裏なら 0」という確率変数を X_j とすると，$X = \sum_{j=1}^{n} X_j$ と書ける．更に X_j は独立同分布なので，n が十分に大きいなら中心極限定理により

$$Z = \frac{X - pn}{\sqrt{np(1-p)}} \tag{5.1.12}$$

は標準正規分布と思って良い．(元々が二項分布なので，パラメーターは一個 p だけ．従って，分散も p の関数として決まっているのがミソ.)

従って，信頼区間 95% で議論するなら，

$$|Z| < 1.96 \qquad \text{つまり} \qquad \frac{|m - pn|}{\sqrt{np(1-p)}} < 1.96 \tag{5.1.13}$$

となるような p の範囲が求めるもの，ということになる．(今は m は n 回投げたあとの結果として得られているので，上の不等式の未知数は p だけである．よって，原理的に上のは p について解ける.）これをもう少しわかりやすく書いておくと，

$$\frac{|\bar{p} - p|}{\sqrt{p(1-p)}} < \frac{1.96}{\sqrt{n}} \tag{5.1.14}$$

となる．n が大きくなるに連れて，右辺がゼロに行く事，従って左辺もゼロに行く必要があるから，p の存在範囲が $\bar{p}$ の周りに絞られて行く事がわかる（存在範囲の幅は右辺から $\sqrt{n}$ のオーダーである）．

ただ，上の不等式を解くのは大変なので，分散については，標本のデータで置き換える事をよくやる．すなわち，上の式を

$$|\bar{p} - p| < \frac{1.96}{\sqrt{n}} \times \sqrt{p(1-p)} \tag{5.1.15}$$

と書いてみると，$\sqrt{p(1-p)} \leq 1/2$ であって，これは大して大きくない．しかも，$1/\sqrt{n}$ のお陰で右辺はもともと小さい．だから，右辺の p はその第零近似の $\bar{p}$ で置き換えても，そんなに違いはないだろう（n が十分に大きければこれは正当化できる）．ということで，

$$|p - \bar{p}| < 1.96 \times \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{5.1.16}$$

としてしまうのが一般的である．

5.1.4 比率の検定（二項分布を仮定）

同じようなノリであるが，もう一個．先の小節と同じ状況で，今度は $p = 0.5$ であるか否か（コインがイカサマか？），などを考えたい．つまり，帰無仮説　$H_0 : p = p_0$（p_0 は特定の値）を検定したい．

考え方は平均に関する場合と同じである．H_0 が正しいのであれば，

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} \tag{5.1.17}$$

は（n が大きい時に）標準正規分布に従うはずである．従って，

(a) 両側検定．$H_1 : p \neq p_0$ を対立仮説にした場合は，

$$|Z| > 1.96 \qquad \text{つまり} \qquad \frac{|\bar{p} - p_0|}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} > 1.96 \tag{5.1.18}$$

ならば H_0 を棄却して，$p \neq p_0$ と結論する．また

(b) 片側検定．$H_1 : p > p_0$ を対立仮説にした場合は，

$$Z > 1.65 \qquad \text{つまり} \qquad \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} > 1.65 \tag{5.1.19}$$

ならば H_0 を棄却して，$p > p_0$ と結論する．さらに，

(c) 片側検定．$H_1 : p < p_0$ を対立仮説にした場合は，

$$Z < -1.65 \qquad \text{つまり} \qquad \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)}} \times \sqrt{n} < -1.65 \tag{5.1.20}$$

ならば H_0 を棄却して，$p < p_0$ と結論する．

5.1.5 適合度の検定（大標本）

（状況）サイコロを n 回投げたら，j の目のでた回数が n_j になった（$j = 1, 2, \ldots, 6$）．これから（たとえば）このサイコロがイカサマであるかどうか判断したい．より一般には，j の目が出る確率 p_j が特定の値 p_j^0 であると思って良いか検定したい．(先の小節では結果が 2 通りしかない場合（コインなげ）を考えたが，この小節では結果がサイコロのように 3 通り以上ある場合を考える.)

（解法）仮説「$H_0 : j$ の目の出る確率 p_j は p_j^0 である（$j = 1, 2, \ldots, 6$）」を検定する．

この仮定の下で，j の目の出る回数を X_j とする．X_j は確率変数であるが，仮定により，その期待値は np_j である．しかし，実際には n_j 回が観測された．この差 $np_j - n_j$ から，p_j についての情報を得たい訳だ．

さて，もし仮説 H_0 が正しいとき，大標本では，

$$X^2 := \sum_{j=1}^{6} \frac{(n_j - np_j^0)^2}{np_j^0} \tag{5.1.21}$$

の分布は，**自由度** $(6-1) = 5$ **の** χ^2**-分布**というものに近づく（教科書 2.4b；ただし，そこでは正規分布の話をしているが，ここでは大標本であることから，n_j の分布を正規分布と近似して良い，と考えて議論する）．

もちろん，X^2 の値が大きければ大きいほど，(n_j と np_j^0 の差が大きいので）元の仮説がウソの可能性が高い．従って，いつもの検定のノリで，

$$X^2 > \chi^2_{6-1}(0.05) \tag{5.1.22}$$

ならば，確率 5% 程度で起こらない事が起こってしまっていると考えて，もとの仮説 H_0 を棄却する．

χ^2-分布については，今学期の最後の方で回帰分析というのをやる場合にもっと詳しくやるので，ここではこのくらいにしておきましょう．

5.2 2 標本問題（自由度大）

（この節は教科書の 3.3 節（の一部）に相当する.）

この節で考えるのは,「2 標本問題」と呼ばれるものである．でもこれは本当は「二母集団問題」と言った方が良いように思う（理由はすぐ後に）．

「比率の差の推測」が一番わかりやすいと思うので，それからやります．

5.2.1 比率の差の推定（二項分布を仮定）

（この節は教科書 3.3b の前半）

（状況）コインが 2 枚ある（10 円玉と 100 円玉など）．片方のコインを n_1 回なげたら，表が n_{1g} 回でた．もう一つのコインを n_2 回投げたら，n_{2g} 回表がでた．これらのコインが**表を出す確率に差はあるだろうか？**

（状況 ′）二つの母集団 Π_1, Π_2 があり，それぞれのメンバーは 2 種類に別れている（「良い」メンバーと「悪い」メンバーなど）．$j = 1, 2$ に対して Π_j から n_j 個のメンバーを取り出してみたところ，それぞれ n_{jg} 個が「良い」メンバーであった．元々の母集団での「良い」メンバーの比率 p_j に差があるかどうか，考えよ．

（新薬の治験[19]）ある病気に対する新薬の候補がある．これが従来の薬よりも効果があるのか，確かめるために，患者を 2 グループ（n_1 人と n_2 人）に分けて，グループ 1 には旧来の薬，グループ 2 には新薬候補を投与した．その結果，グループ 1 では m_1 人，グループ 2 では m_2 人の病気が治った．二つのグループの病気の深刻さは同じくらいだったとして，新薬候補が本当に旧薬よりも効果があるのか否かを判断せよ．

（解法 0）別に差などといわなくても，それぞれの母集団について，前節のやり方で推定を行い，p_1 と p_2 の 95% 信頼区間を求める．両者が重なってなければ差があると思うし，重なってたら差は（それほど）ないと思う．

このやり方は決して悪いものではないが，折角，二つのデータがあるのにそれを 100% 活かし切っていない感じが残る．そこで，もう少し良いやり方として，以下の解法を考える．

このような結果を整理するには，以下のような 2×2 分割表を用いるのが便利である．データの整理のためにも，まずは分割表を書く事をお勧めする．

	良い	悪い	合計
母集団 1	n_{1g}	n_{1b}	$n_1 = n_{1g} + n_{1b}$
母集団 2	n_{2g}	n_{2b}	$n_2 = n_{2g} + n_{2b}$
合計	$m_g = n_{1g} + n_{2g}$	$m_b = n_{1b} + n_{2b}$	$n = n_1 + n_2$

（解法）データから得られる p_j の第零近似としては

$$\bar{p}_j := \frac{n_{jg}}{n_j} \qquad (j = 1, 2) \tag{5.2.1}$$

があるが，大標本では中心極限定理から，これは（$j = 1, 2$ のそれぞれに対して）正規分布

$$N\Big(p_j, \frac{p_j(1-p_j)}{n_j}\Big) \tag{5.2.2}$$

で近似できると考えられる．正規分布の差もまた正規分布になることを用いると，$\bar{p}_1 - \bar{p}_2$ の分布は

$$N\Big(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\Big) \tag{5.2.3}$$

で近似できるはずである（分散は和になった事に注意）．

あとはこれを標準正規分布になおして，今までのように議論する．つまり

$$Z = \frac{(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \tag{5.2.4}$$

の分布は，標準正規分布のはずである．よって，$p_1 - p_2$ の 95% 信頼区間は，上の Z が $Z = 0$ の周りに確率 0.95 で存在する部分，つまり

$$|z| < 1.96 \qquad \text{つまり} \qquad \frac{\big|\,(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)\,\big|}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} < 1.96 \tag{5.2.5}$$

[19]実際の治験には条件を同じくするために非常に厳しい制約がある．特に二重盲検法と呼ばれる手法で，検査側の主観をできるだけ排除することが行われる．(このような厳しい試験に通っていない自称「治療法」は信用しない方が無難であろう.) また，ある種の人体実験であるから，非常に慎重に進める必要がある．この講義ではそのようなややこしい事はすべて無視して，統計をどう使うかだけをお話しする

という事になる．

実は 1 標本問題と違って，上の式は $p_1 - p_2$ だけの式にはなってくれないので，これだけでは情報不足である．仕方ないので，普通は分母の p_j を $\bar{p}_j$ で置き換える（ここは前節でやった近似と全く同じノリである — 平均はちゃんと推測したいが，分散は近似値で誤摩化す）．結果として

$$\frac{\big|\,(\bar{p}_1 - \bar{p}_2) - (p_1 - p_2)\,\big|}{\sqrt{\dfrac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \dfrac{\bar{p}_2(1-\bar{p}_2)}{n_2}}} < 1.96 \tag{5.2.6}$$

つまり

$$(p_1 - p_2) = (\bar{p}_1 - \bar{p}_2) \pm 1.96 \times \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}} \tag{5.2.7}$$

というのが，$p_1 - p_2$ の 95% 信頼区間になる[20]．特にこの信頼区間にゼロが入っておれば，両者に差があるとは言い切れない，と判断するのが無難であろう．(これまで何回もくり返して来たように，だからといって「差がない」という結論にもならない．)

5.2.2 比率の差の検定（二項分布を仮定）

（この節は教科書 3.3b の後半）

（状況）前小節と同じ状況を考えるが，今度は $H_0 : p_1 = p_2$ を帰無仮説として検定する事を考える．薬の治験の場合なら,「新薬には特に改善点はない」というのが H_0 で，これを（本当は）否定したい訳だ．

（解法）H_0 を仮定して，共通の p_j を p と書くと，

$$Z = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{p(1-p)\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} \tag{5.2.8}$$

が標準正規分布に従うはずなのである．またもや，分母の p がわからないのだが，この p の推定値としては 2 グループを併せた場合の「良い」メンバーの数，つまり

$$\bar{p} = \frac{n_{1g} + n_{2g}}{n_1 + n_2} = \frac{m_g}{n} \tag{5.2.9}$$

を使ってやろう．すると

$$Z' = \frac{(\bar{p}_1 - \bar{p}_2)}{\sqrt{\bar{p}(1-\bar{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} \tag{5.2.10}$$

が標準正規分布に近似的に従う，ということだ．後はこれまで通りに検定を行う．

両側検定（対立仮説が $H_1 : p_1 \neq p_2$ なら，$|Z| > 1.96$ で H_0 を棄却する．つまり，

$$\frac{|\bar{p}_1 - \bar{p}_2|}{\sqrt{\bar{p}(1-\bar{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} > 1.96 \qquad \text{つまり} \qquad \frac{(\bar{p}_1 - \bar{p}_2)^2}{\bar{p}(1-\bar{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)} > (1.96)^2 \approx 3.84 \tag{5.2.11}$$

ならば H_0 を棄却する．なお，教科書には，この式を n_j, n_{jg} などで書き直した式が (3.23) として載っているが，同じ事である．

同様に，片側検定で対立仮説が $H_1 : p_1 > p_2$ ならば,

$$\frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} > 1.65 \tag{5.2.12}$$

の時に H_0 を棄却する．対立仮説が $H_1 : p_1 < p_2$ なら

$$\frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\Big(\dfrac{1}{n_1} + \dfrac{1}{n_2}\Big)}} < -1.65 \tag{5.2.13}$$

の時に H_0 を棄却する．

[20] いうまでもなく，$X = a \pm b$ というのは，$a - b \leq X \leq a + b$ だと読んで下さい

5.2.3 平均の差

（この節は教科書 3.3a）

この場合もやる事はほとんど同じである．時間の関係で触れられないかもしれないので，講義ノートには時間ができたら書きます．

5.3 母集団が正規分布のとき（小標本でも）

さて，今までの節で見て来たのは，大標本（標本の大きさが大きい）場合であった．この場合，母集団の分布が何であれ，大標本であることが原因して，標本平均や標本分散がまあまあ，良い分布で近似できることが使えた．その結果,(時には分散の方は近似を粗く誤摩化してでも）平均についてはそこそこ良い結果を得る事ができた．

ところが，世の中には大標本の問題だけがあるわけではない．いやむしろ，標本の大きさが小さい事の方が多い（薬の治験だって，200 人，500 人と集められないこともある）．そんな場合にも何か言える事はないのだろうか？それがこの節のテーマである．

ただし，小標本でものを言うには，**母集団の分布について，かなりの仮定が必要**である．なぜなら，小標本の場合に成り立つ普遍的な極限定理などがないので，小標本のデータからもとを推測するのは（何らかの付加的仮定抜きでは）不可能だからである．

そこで母集団について何らかの仮定をする事になるが，一番考えやすく，たくさんの情報が得られるのは**母集団が正規分布に従うとき**である．ので，この節では**母集団が正規分布に従うとき**に限って，小標本のデータから何がいえるかを考えて行く．

5.3.1 正規分布のいくつかの性質

一部は既に正規分布を学んだところで述べたが，大事な性質をまとめておこう．

- 独立な確率変数 X_j（$j = 1, 2, \ldots, n$）が正規分布 $N(\mu_j, \sigma_j^2)$ に従うとき，その和 $Y = X_1 + X_2 + \cdots + X_n$ は $N(\mu, \sigma^2)$ に従う．ただし，$\mu = \mu_1 + \mu_2 + \ldots + \mu_n$，$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$ である．

- 独立な確率変数 X_j（$j = 1, 2, \ldots, n$）が標準正規分布に従うとき，$Y = X_1^2 + X_2^2 + \ldots + X_n^2$ は**自由度 n の χ^2-分布**に従う．Y の分布密度関数は

$$f(y) = \frac{1}{2\Gamma(n/2)} \left(\frac{y}{2}\right)^{n/2-1} e^{-y/2} \tag{5.3.1}$$

  である（$y \geq 0$）．この性質は，実際に分布関数を積分して求めれば納得できる．なお，Y が自由度 n の χ^2 分布に従うとき，$P[Y > C] = \alpha$ となる C の値の事を $\chi^2_{n-1}(\alpha)$ で表す（**上側 $100\alpha\%$ 点**，教科書 p.49）．

- X を標準正規分布，Y を自由度 k の χ^2-分布に従う確率変数，かつ，X, Y は独立とする．このとき，

$$T := \frac{X}{\sqrt{Y/k}} \tag{5.3.2}$$

  の分布を，**自由度 k の t-分布**という．実際に計算するとその分布密度は

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-(k+1)/2} \tag{5.3.3}$$

  で与えられることがわかる．なお，$P[|T| > C] = \alpha$ となる C を $t_k(\alpha)$ で表す．実際の値は分布表から求められる．

	0.99	0.975	0.95	0.05	0.025	0.01
1	0.00	0.00	0.00	3.84	5.02	6.63
2	0.02	0.05	0.10	5.99	7.38	9.21
3	0.12	0.22	0.35	7.81	9.35	11.34
4	0.30	0.48	0.71	9.49	11.14	13.28
5	0.55	0.83	1.15	11.07	12.83	15.09
6	0.87	1.24	1.64	12.59	14.45	16.81
7	1.24	1.69	2.17	14.07	16.01	18.48
8	1.65	2.18	2.73	15.51	17.53	20.09
9	2.09	2.70	3.33	16.92	19.02	21.67
10	2.56	3.25	3.94	18.31	20 48	23.21

表 1: $\chi^2_n(\alpha)$ の値．縦の列が自由度 n，横の列が α である．学期始めに紹介した，服部氏の著作から引用．

- $X_1, X_2, \ldots, X_n$ を，同じ正規分布 $N(\mu, \sigma^2)$ に従う互いに独立な確率変数とする．$\bar{\mu}$ をこの n 個のデータの標本平均，$\bar{\sigma}$ を標本標準偏差とする．このとき，

$$T := \sqrt{n-1}\,\frac{(\bar{\mu}-\mu)}{\bar{\sigma}} \tag{5.3.4}$$

  は自由度 $(n-1)$ の t-分布に従う．

- $X_1, X_2, \ldots, X_n$ を，同じ正規分布 $N(\mu, \sigma^2)$ に従う互いに独立な確率変数とする．$\bar{\mu}$ をこの n 個のデータの標本平均，$\bar{V}$ を標本分散，$\bar{\sigma}$ を標本標準偏差とする．このとき，

$$\chi^2 := \frac{\bar{\sigma}^2}{\sigma^2} = \frac{\bar{V}}{V} \qquad (\text{ここで } V = \sigma^2 \text{ は母集団の分散}) \tag{5.3.5}$$

  は自由度 $(n-1)$ の χ^2-分布に従う．

これらの性質は，具体的に計算する事で確かめられる．

なお，教科書では敢えて使っていないが，データを扱う際には，標本分散だけでなく**不偏分散**と呼ばれる以下の量を用いることがある：

$$\bar{V}_{\text{不偏}} := \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{\mu})^2 \tag{5.3.6}$$

(通常の標本分散と異なり，$(n-1)$ で割っている)．これを用いると，上の T は

$$T := \sqrt{n-1}\,\frac{(\bar{\mu}-\mu)}{\bar{\sigma}} = \sqrt{\frac{n}{\bar{V}_{\text{不偏}}}}(\bar{\mu}-\mu) \tag{5.3.7}$$

とも書ける（両方の表式が等しい事は定義からすぐに確かめられる)．同様に，上の χ^2 は

$$\chi^2 := \frac{\bar{\sigma}^2}{\sigma^2} = \frac{\bar{V}}{V} = (n-1)\frac{V_{\text{不偏}}}{V} \tag{5.3.8}$$

とも書ける．

教科書には $\chi^2_{n-1}(0.05)$ と $\chi^2_{n-1}(0.01)$ しか載っていないので，参考までに他の値も表 5.3.1 に挙げておく．もちろん，このような値を覚える必要は全くない．

5.3.2 1 標本問題の平均の推定・検定

では，上の知識をつかって，一標本問題を考えて行こう．

（問題 1）母集団は正規分布に従う事はわかっているが，その平均 μ と分散 σ が未知である．この時に，n 個からなる標本をとると，その標本平均は $\bar{\mu}$，標本標準偏差は $\bar{\sigma}$ であった．これから母集団の平均を推測せよ．

（解法）前小節でまとめた T 分布の性質をモロに用いる．それによれば，

$$T := \sqrt{n-1}\,\frac{(\bar{\mu}-\mu)}{\bar{\sigma}} = \frac{\bar{\mu}-\mu}{\bar{\sigma}/\sqrt{n-1}} \tag{5.3.9}$$

は自由度 $n-1$ の t-分布に従うのだった．従って，これまでと全く同じノリで，$|T| < t_{n-1}(0.05)$ となる T を満たすような μ の範囲が 95% 信頼区間ということになる．つまり，

$$|\mu - \bar{\mu}| < t_{n-1}(0.05) \times \frac{\bar{\sigma}}{\sqrt{n-1}} \tag{5.3.10}$$

が μ の 95% 信頼区間である．$t_{n-1}(0.05)$ の値は数表になっているから，実際の問題を解く時には，その数表を用いれば良い（教科書の最後にもある）．

（問題 2）問題 1 と同様の状況を考えるが，今度は $\mu = \mu_0$（μ_0 は適当に推測した値）であるか否かを検定せよ．

（解法）帰無仮説は $H_0 : \mu = \mu_0$ とする．検定に際して用いるのは，上と同じく，

$$T = \frac{\bar{\mu}-\mu_0}{\bar{\sigma}/\sqrt{n-1}} \tag{5.3.11}$$

である（ただし，H_0 を仮定しているので，μ は μ_0 になってる）．この T は仮説 H_0 が正しいならば自由度 $(n-1)$ の t-分布に従うはずなので，あとは対立仮説によって以下のように議論する．

（対立仮説が $H_1 : \mu \neq \mu_0$ の時）この場合は普通に

$$|T| > t_{n-1}(0.05), \qquad \text{つまり} \qquad \frac{|\bar{\mu}-\mu_0|}{\bar{\sigma}/\sqrt{n-1}} > t_{n-1}(0.05) \quad \text{ならば} \quad H_0 \text{ を棄却} \tag{5.3.12}$$

する．

（対立仮説が $H_1 : \mu > \mu_0$ の時）この場合は片側検定である．(片側 5% ということは両側に直せば 10% なので)

$$T > t_{n-1}(0.10), \qquad \text{つまり} \qquad \frac{\bar{\mu}-\mu_0}{\bar{\sigma}/\sqrt{n-1}} > t_{n-1}(0.10) \quad \text{ならば} \quad H_0 \text{ を棄却} \tag{5.3.13}$$

する．

（対立仮説が $H_1 : \mu < \mu_0$ の時）この場合は片側検定で，上の正負逆バージョンであるから，

$$T < -t_{n-1}(0.10), \qquad \text{つまり} \qquad \frac{\bar{\mu}-\mu_0}{\bar{\sigma}/\sqrt{n-1}} < -t_{n-1}(0.10) \quad \text{ならば} \quad H_0 \text{ を棄却} \tag{5.3.14}$$

する．

上の何れの場合も，棄却できない場合は「何も言えない」という結論になるのはいままでと同じ．なお，このような検定を t-検定という．

5.3.3 1 標本問題の分散の推定・検定

この内容は教科書には無いようだが，話を完結させるために述べておく．

では，上の知識をつかって，一標本問題を考えて行こう．

（問題 1′）母集団は正規分布に従う事はわかっているが，その平均 μ と分散 σ が未知である．この時に，n 個からなる標本をとると，その標本平均は $\bar{\mu}$，標本標準偏差は $\bar{\sigma}$ であった．これから母集団の分散を推測せよ．

（解法）この問題は先の小節とほとんど同じだが，平均でなく分散を調べてほしい，というところが異なる．これは当然，χ^2-分布を使って解くべきだ．5.3.1 節のまとめによると

$$\chi^2 := \frac{\bar{\sigma}^2}{\sigma^2} = \frac{\bar{V}}{V} \tag{5.3.15}$$

は自由度 $(n-1)$ の χ^2-分布に従う．上の χ^2 には未知数は V （または σ）だけしか入っていない．だから，自由度 $(n-1)$ の χ^2 が確率 0.95 以上で存在する範囲を求めれば，$V=\sigma^2$ の存在範囲がわかるはずである．つまり，V の 95% 信頼区間は

$$\chi^2_{n-1}(0.975) < \frac{\bar{V}}{V} < \chi^2_{n-1}(0.025) \tag{5.3.16}$$

を満たすような V の区間である（95% 信頼区間という事は，χ^2 が余りにも小さすぎるのと大きすぎるのを排除すべし，ということなので，両側から 0.025 ずつを避けた）．

分散の検定も同様に行う．基本的なアイディアはこれまでと同じ，また χ^2-分布を使う所は上の推定と同じなので，詳細は省略する．

5.3.4 2 標本問題

（問題）母集団が二つある．ともに正規母集団に従うが，その平均や分散はわかっていない．ただし，**分散は二つの母集団で等しい**と仮定する．つまり，二つの母集団は $N(\mu_1,\sigma^2)$ と $N(\mu_2,\sigma^2)$ に従い，μ_1,μ_2,σ^2 が未知数である場合を考える．

さて，このときに母集団 1 から n_1 個の標本をとったら，その標本平均が $\bar{\mu}_1$，標本標準偏差が $\bar{\sigma}_1$ であったとしよう．また，母集団 2 から n_2 個の標本をとると，その標本平均が $\bar{\mu}_2$，標本標準偏差が $\bar{\sigma}_2$ であったとしよう．このときに，母集団平均の差 $\mu_1-\mu_2$ を推測したい．

（解法）ノリは 1 標本問題とほとんど同じであるが，うまく t-分布に従う統計量を作る（見つける）のがキーである．

仮定から，$\bar{\mu}_1$ の分布は $N(\mu_1,\sigma^2/n_1)$ に従うはずである．同様に，$\bar{\mu}_2$ の分布は $N(\mu_2,\sigma^2/n_2)$ に従うはずである．従って，正規分布の足し算，引き算の性質を思い出すと，$\bar{\mu}_1-\bar{m}u_2$ は $N(\mu_1-\mu_2,\sigma^2/n_1+\sigma^2/n_2)$ に従うはずだ．

これから，標準正規分布に従う量として

$$Z=\frac{(\bar{\mu}_1-\bar{\mu}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\,\sigma} \tag{5.3.17}$$

が考えられる．ところが，この量には未知数として $\mu_1-\mu_2$ のみならず σ が入っていて，これだけでは扱えない．

仕方ないので，σ^2 をその推定量

$$V=\frac{n_1\bar{\sigma}_1^2+n_2\bar{\sigma}_2^2}{n_1+n_2-2} \tag{5.3.18}$$

で置き換えて

$$T=\frac{(\bar{\mu}_1-\bar{\mu}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\,\sqrt{V}} \tag{5.3.19}$$

を考えよう．この T 中には未知数は $\mu_1-\mu_2$ しかないから，この T の分布がわかれば，これまでと同じノリで推定や検定を行える．

さて，この T の分布は正規分布ではなく，自由度 (n_1+n_2-2) の t-分布 になることがわかる（ノートを書く時間がなかった．教科書 p.88 の下半分を参照のこと）．あとはこれまでと同じく，t-推定を行えばよい．また，元の問題が $\mu_1=\mu_2$ を検定する問題なら，t-検定を行えば良い．

6 回帰分析の初歩

最後に,（工学部なら実験解析等で散々使ってるはずの）回帰分析に簡単に触れる.

6.1 単回帰

（問題 1）ある物質の電気抵抗を測りたい．オームの法則から，電圧と電流には $V = IR$ の比例関係があるはずである．電圧をいろいろ変えて測定したところ，全部で n 個の電圧電流のデータ $(V_1, I_1), (V_2, I_2), \ldots, (V_n, I_n)$ が取れた．このデータから抵抗 R を決めよ（できるだけ信頼できる値を出せ）.

（問題 1′）電流と電圧に限らず，二つの物理量 (x, y) の間に線型な関係（一次関数の関係）$y = ax + b$（a, b は定数）があると期待されている．実際に n 個のデータが $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ と得られた場合，これらから定数 a, b をできるだけうまく定めよ.（問題 1 は $b = 0$ の特殊な場合であった.）

このような状況は皆さん，中学（小学校？）から何回も遭遇した事と思う.（電流と電圧に限らず，塩の溶解度と温度の関係だか，いろいろ.）このような実験の場合，大抵グラフ用紙に測定値を書き込んで，それらの点を「うまく結ぶように」直線を引け，と習ったと思うが，それをもうちょっときちんとやろうということである.（もちろん，大学での実験では既に「最小二乗法」などを用いる事を習っているだろうが，その理論的根拠を確認しておきたいということ.）

さて，実は上の問い，重要な仮定を書いていなかった．その仮定とは

> n 個のデータのそれぞれの誤差は，同じ正規分布にしたがって分布するだろう

という事である．より正確に言うと.

> n 個の $x_1, x_2, \ldots, x_n$ は，それぞれ真の値からバラツくはず（測定誤差）だが，そのバラツキは，n 個のデータが同じ正規分布（平均はゼロ）$N(0, \sigma_x^2)$ にしたがって分布するだろう．また，n 個の $y_1, y_2, \ldots, y_n$ も，それぞれ真の値からバラツくはず（測定誤差）だが，そのバラツキは，n 個のデータが同じ正規分布 $N(0, \sigma_y^2)$ にしたがって分布するだろう．ただし，σ_x, σ_y の値はよくわからない.

ということである.

以下，このような状況で，どのように a, b を決めるべきかを考える.

まず，本来，測定誤差がなければ，n 個のデータは線型の関係を厳密に満たすはずであった：

$$y_j = ax_j + b_j \qquad \text{または} \qquad y_j - (ax_j + b) = 0 \qquad (j = 1, 2, \ldots, n) \tag{6.1.1}$$

しかし，実際には測定誤差があるため，上の関係式は近似的にしかなりたたない．そのずれを ϵ_j と書いてみよう：

$$\epsilon_j = y_j - (ax_j + b) \tag{6.1.2}$$

上の仮定によると y_j, x_j はそれぞれが平均ゼロの正規分布に従うので，上の特別な組み合わせ $y_j - (ax_j + b)$ も平均ゼロの正規分布に従うはずだ．つまり，上の ϵ_j は平均ゼロの正規分布（分散の値はわからない）に従っているはずである．話を具体的にするため，この正規分布を $N(0, \sigma^2)$ と書くことにする.

さて，我々の目的は「正しい」係数 a, b を決める事だった．その方法を考えるため，ここで敢えて正しくない（つまり，本来成り立っているはずの線型関係からズレた係数の）a, b を考えてみる．つまり，本来の関係は $y = ax + b$ なのだが，これを敢えて間違った係数を用いて $y = a'x + b'$ だと思ってみるのだ.

我々はこの間違った係数がどのくらい間違っているかわからないから，

$$\epsilon'_j := y_j - (a'x_j + b') \tag{6.1.3}$$

を考えたくなる．これは係数 a, b を用いて書き直すと

$$\epsilon'_j = y_j - (a'x_j + b') = y_j - (ax_j + b) - (a' - a)x_j - (b' - b) = \epsilon_j - (a' - a)x_j - (b' - b) \tag{6.1.4}$$

となっている．つまり，間違った係数を用いて計算した誤差 ϵ'_j は ϵ_j から $-(a'-a)x_j-(b'-b)$ だけズレている訳だ．

さて，ここで間違った係数と正しい係数でそれぞれ，誤差の二乗の和（の期待値）を考えてみよう．正しい係数の場合の誤差は ϵ_j そのもので，これは $N(0,\sigma^2)$ に従う．従って，その二乗の期待値は

$$\left\langle \sum_{j=1}^{n} \big(\epsilon_j\big)^2 \right\rangle = \sum_{j=1}^{n} \left\langle \big(\epsilon_j\big)^2 \right\rangle = n\sigma^2 \tag{6.1.5}$$

である．

一方，間違った係数を使った場合の誤差は ϵ'_j だから，その二乗の期待値は

$$\left\langle \sum_{j=1}^{n} \big(\epsilon'_j\big)^2 \right\rangle = \sum_{j=1}^{n} \left\langle \big(\epsilon'_j\big)^2 \right\rangle = \sum_{j=1}^{n} \left\langle \big\{\epsilon_j-(a'-a)x_j-(b'-b)\big\}^2 \right\rangle \tag{6.1.6}$$

となる．ここで x_j が実は誤差を含んでいて確定しない量なのだが，近似的にその値を真の値で置き換えて考えると (真の値を x_j^0 と書く)，

$$= \sum_{j=1}^{n} \left\langle \big(\epsilon_j\big)^2 \right\rangle + \sum_{j=1}^{n} \left\langle \big\{(a'-a)x_j+(b'-b)\big\}^2 \right\rangle = n\sigma^2 + \sum_{j=1}^{n} \big\{(a'-a)x_j+(b'-b)\big\}^2 \tag{6.1.7}$$

となる．右辺の $n\sigma^2$ は正しい係数と同じだが，右辺第 2 項がついている．しかもこの第 2 項は二乗の和だから非負であって，かつ，$a=a', b=b'$ の時に最低値 0 をとる．

すなわち，間違った値で計算した結果は，かならず正しい値での計算結果以上である．逆にいうと，正しい値を知りたければ，$\left\langle \sum_{j=1}^{n} \big(\epsilon'_j\big)^2 \right\rangle$ を**最小にするように**a', b' を決めれば良い．これが**最小二乗法**の考え方である．

具体的な計算は以下のように行う．a, b は最終的には「正しい」値にしたいが，今暫くは未知の数だとしておこう．我々の目的は

$$f(a,b) := \sum_{j=1}^{n} \big\{y_j-(ax_j+b)\big\}^2 \tag{6.1.8}$$

を最小にするような a, b を求める事である（最小にするような a, b が「正しい」値のはず）．これは単に上の 2 変数函数の極値問題であって，簡単に解ける．

答えを書くと

$$a = \frac{\bar\sigma_{xy}}{\bar\sigma_x^2}, \qquad b = \bar y - a\bar x \tag{6.1.9}$$

となる．ここで

$$\bar\sigma_x^2 := \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar x)^2, \qquad \bar\sigma_{xy} := \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar x)(y_j-\bar y), \tag{6.1.10}$$

は x_j の標本分散，x, y の標本共分散である．