統計数学入門

山上　滋

平成31 年2 月5 日

はじめに

これは確率の数学が現実的な問題に如何に役立つかを、統計的な基本問題を題材にして解説した半年用の講義ノートです。

統計学的な「お作法」にはあまりこだわらずに、確率の考え方を現実の問題にどのように結びつけ得るか、に的をしぼって見ました。その意味で、かなり中途半端な内容にはなりましたが、この方面の知識の第一歩としては、ある程度の論理性を確保したつもりです。

これを手がかりにさらに進んだ内容に向かうもよし、あるいはまた、現実の様々な統計的な分析結果に対して理をもって対処するだけの心得とすることも可能でしょう。

とかく「教条的」になりがちな統計学のエッセンスが伝えられれば幸い。サイエンスの基本は「信じることではなく疑うこと」であるべきなので。意味を忘れて形式だけに囚われてはいけません。

試験の成績で、平均点以上の点数を取る人の割合は、

(i) ちょうど５０%、

(ii) ５０%より大きい、

(iii) ５０%よりも小さい

のどれが正しいか？という「インチキ」な問題を考えてみよう。

大きい数と小さい数、対数の導入と計算。近似の考え方。この近似に対する圧倒的経験不足の現状が様々な劣化を引き起こしているような。大丈夫か、日本人。

目 次

1 二項分布

確率という言葉は実社会でも頻繁に使われているが（たとえば天気予報での降水確率、模擬試験における大学の合格率など)、それを数学的に厳密な仕方で定式化するのは意外に難しいものである。

常識的には、大量のデータに基づいた規則を比率の形で表現したものを確率と呼んでいるようである。ただ単にデータと言った場合には様々なものが考えられるが、統計学では何らかの方法で数字によって記述されるものを問題にすることが多い。

例題 1.1. 試験の成績、バクテリアの数、原子から放射される γ 線のエネルギー、1 日の最高気温・最低気温。円の為替相場。

このように測定あるいは実験結果を数字で表した時、観測を数多く繰り返すにつれて（観測回数を N とするとき、$\lim_{N\to\infty}$ ということ)、特定の観測数値が得られる割合が一定の値（0 と 1 の間の）に近づくならば、その極限値をもってその特定の現象が起こる**確率** (probability) とよび、そのような観測数値の対応のさせ方を**確率変数** (random variable) と呼ぶ。

例題 1.2.

(i) 入学試験の受験者の中から、でたらめに一人を選んでその入学試験の成績を調べる。点数が x である受験者の人数を N_x で表すと、仮に 100 点満点とし、調べた結果が x 点である確率は

$$\frac{N_x}{N_0 + \cdots + N_{100}}$$

となるであろう。

(ii) 一定の条件（温度・湿度などが）のもとで一定時間経過した後のバクテリアの増加数 x。N 回実験を行った中で結果が x となった回数を $\{N_x\}$ で表せば、x 個という結果が得られる確率は

$$\lim_{N\to\infty} \frac{N_x}{N}$$

で与えられるであろう。もちろん、このような極限が存在する保証は一般的には何もない訳であるが、存在するものと仮定して、それから導かれる結論が実際の結果をうまく説明できるならば、「確率が存在する」と考えることにする。（すべての科学的な仮説とは，このようなものである。）

Remark . 単に確率変数 X といった場合にはその背景の条件（どういう観察をするか）も一緒に考えているものとする。確率変数について議論をするときに、その背景についての条件が曖昧になりがちなので、注意を要する。

以下、確率変数は X などのアルファベット大文字で表す。個々の観測結果は a, b, x, y 等で示す。観測結果が x である確率を $P(X = x)$（あるいは単に P_x）で表す。個々の観測結果にどういう確率が対応するかの規則を**確率分布** (distribution) という。

$\cdots$	x	$\cdots$
$\cdots$	P_x	$\cdots$

, $\qquad \sum_x P_x = 1.$

確率変数そのものの意味が違っていても同じ確率分布をもつことは多々ある。

範囲の確率：$a < b$ に対して、$P(a < X < b)$ で観測結果が (a, b) の範囲に見出される確率を表す。すなわち

$$P(a < X < b) = \sum_{a<x<b} P_x.$$

例題 1.3. ある商店で宣伝のためにくじ引きを行うことにした。大量のくじの中に当たりのくじを一定の割合（$0 < p < 1$）で混ぜ、先着 100 人に引かせることにした。ところが、店主はけちなので、当たりが一人も出ない確率を $1/2$ 以上にしようと思った。 p はどの程度小さくすべきか？

Proof. 条件は $(1-p)^{100} \geq 1/2$ である。常用対数をとって

$$100 \log(1-p) \geq -\log 2 = -0.301 \iff \log(1-p)^{-1} \leq 0.00301.$$

これを 10 の肩にのせて（指数をとって）

$$\frac{1}{1-p} \leq 1.007 \iff 1-p \geq 0.993 \iff p \leq 0.007.$$

もう少し手を抜いた方法として、二項展開の近似式 $(1-p)^{100} = 1 - 100p$ を使うと、$1 - 100p \geq 1/2$ から $p \leq 0.005$ が得られる。 □

問 1. 200 人、300 人として同様の計算をしてみる。手抜きの近似計算の精度はよくなるか悪くなるか？

Remark . Windows で使える「電卓」として、
スタート＞すべてのプログラム＞アクセサリ＞電卓
がある。対数計算、冪計算を行うためには、電卓の「表示」で関数電卓を選択すると良い。

上の問題で、n 人くじを引くとき当たりの出る回数を表す確率変数を X とする。n 人がくじを引いた結果は、当たり（A）とはずれ（H）の列で表される。例えば、

$$\mathrm{AAHAHAAHHHA}\cdots$$

A が k 回 H が l 回現れる確率は $p^k q^l = p^k q^{n-k}$ である（$q = 1 - p$）。k を指定したとき、このような列の可能な組み合わせの数は ${}_nC_k = \binom{n}{k}$ だけあるので、

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

この式で表される確率分布を**二項分布** (binomial distribution) とよぶ。名前の由来は二項展開

$$1 = (p + q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

例題 1.4. インフルエンザが大流行し４人に１人は風邪をひいていたが、A 社のかぜ薬をのんだ１０人は１人だけ風邪にかかり、B 社の薬をのんだ２０人は２人だけ風邪にかかった。どちらが良く効くか？

Proof. どちらもまったく効き目がなかったとして、上記のような結果が得られる確率を計算してみよう。確率の小さい方が、実際にかぜ薬として効いている可能性が高いと考えることにする（他にも色々な判断の仕方があるだろうが、ここでは二項分布を使う前提でこう考える）。

$$A = \binom{10}{0} \left(\frac{3}{4}\right)^{10} + \binom{10}{1} \frac{1}{4} \left(\frac{3}{4}\right)^{9},$$

$$B = \binom{20}{0} \left(\frac{3}{4}\right)^{20} + \binom{20}{1} \frac{1}{4} \left(\frac{3}{4}\right)^{19} + \binom{20}{2} \left(\frac{1}{4}\right)^{2} \left(\frac{3}{4}\right)^{18}.$$

関数電卓を使って、

$$A = \left(\frac{3}{4}\right)^{10} \frac{13}{3} = 0.244,$$

$$B = \left(\frac{3}{4}\right)^{20} \frac{259}{9} = 0.0912.$$

□

問 2.

(i) ２０人中３人と１０人中１人を比較せよ。

(ii) 一見上の問題で両者の割合は同じに見える。１００人中１０人、１０００人中１００人といった場合の確率の傾向について考察せよ。また、プログラム電卓等を使って具体的な数値を求める工夫をしてみよ。

(iii) 上の解答で

$$A = \binom{10}{1}\frac{1}{4}\left(\frac{3}{4}\right)^9 = 0.1877, \qquad B = \binom{20}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{18} = 0.0669$$

としないのは何故か？

例題 1.5. 超能力の実験。３桁以下の数字を片方が思いうかべもう片方が言い当てる実験を始めたところ、いきなり１０８という数字を言い当てた。３桁以下の数字は $0 \sim 999$ の 1000 個あるので、108 という数字を思いうかべる確率は千分の一。もう片方がやはり 108 を答える確率もやはり千分の一。従って偶然このような一致がおこる確率は

$$\frac{1}{1000} \times \frac{1}{1000} = \frac{1}{1000000}$$

すなわち百万分の一という、けっして偶然とは思えない程小さい確率である。よって、二人の間にはテレパシーが存在したに違いない。

問 3. 上の推論の問題点を検討せよ。刑事事件の裁判で実際に上のような誤った推論に基づき 99.9 % 犯人であると鑑定され、冤罪を着せられた例が実際にあるという。

2　平均値と分散

確率変数 X の**平均値** (mean) または**期待値** (expectation) を

$$\langle X \rangle = \sum_{x \in \mathbb{R}} x P_x$$

で定める。もっと一般に、関数 $f(x)$ に対して、

$$\langle f(X)\rangle = \sum_{x\in\mathbb{R}} f(x)P_x$$

とおく。期待値という言葉はもともと賭けにおいて配当がどれだけ期待できるかということに由来する。

例題 2.1. １等１０００万円一本、２等５００万円３本、３等１００万円１０本の宝くじを１枚１００円で売り出すとして、主宰者（胴元）が損しないためには何枚以上の宝くじを売らないといけないか。またそのときの１枚あたりの期待値は？

例題 2.2 (St. Petersburg Paradox)**.** コインを投げて n 回目に初めて表が出たとき 2^n 円受け取るという賭の期待値は？

無制限に儲かるということなので、どんなに高い参加料を支払ってでも賭けるべきだ。たとえば、１００万円払う？

何かがおかしい。問題点を考察する。

Proof. 百万円に近い金額がもらえる n を計算すると、

$$2^n = 10^6 \quad \Rightarrow \quad n = \frac{6}{\log 2} = 19.9$$

であるから、大体 20 回目が目安である。ところが、２０回目以降にお金を受け取れる確率は

$$\left(\frac{1}{2}\right)^{20} + \left(\frac{1}{2}\right)^{21} + \cdots = \left(\frac{1}{2}\right)^{19}$$

という小さい数で、そのようなことが起こるのは

$$2^{19} = 524288$$

回に１回程度の割合である。したがって、仮にこれだけの回数の賭けを行うことができたとしても（毎日１０００回賭けを行っても、約１年半かかる！）、百万円当てるまでには、１００万円x５２万回＝５千２百億円程度支払わなければならず（１９回以前での配当金の期待値は１９円で５２万倍しても１０００万円程度にしかならない）、儲けるまでには、莫大な時間と資金（不可能ほどの）が必要であることがわかる。 □

平均値は確率分布に関する最も基本的な情報であるが、同じ平均値 0 の確率分布でも次の 2 つは全く性質の異なる分布である。

山一つと山二つの図

（山が二つの場合は二つの要因の合成を疑って見るべきである。）

普通は平均値の近くで確率の高い観測値が密集している。しからば、密集の度合は如何？これの目安を与える量が標準偏差と呼ばれるものである。

確率変数 X の平均値を μ で表すとき、$(X-\mu)^2$ の平均値を X の**分散**（variance）と呼びその平方根を**標準偏差** (standard deviation) といい記号 σ で表す。公式

$$\langle (X-\mu)^2 \rangle = \langle X^2 \rangle - \langle X \rangle^2$$

が成り立つ。いずれも確率分布の広がり程度を表す量であるが、分散の単位は元の確率変数のと異なるのに対して、標準偏差の方は同一の単位であることに注意。（例えば、X が長さを表しているとき、分散は面積の単位、標準偏差は長さの単位をもつ。）

例題 2.3. 二項分布の平均値と分散は、

$$\mu = np, \qquad \sigma = \sqrt{np(1-p)}$$

で与えられる。

Proof. **母関数** (generating function)

$$F(t) = \sum_k \binom{n}{k} p^k q^{n-k} t^k = (pt+q)^n$$

を使う。両辺を t について微分して

$$np(pt+q)^{n-1} = \sum_k \binom{n}{k} p^k q^{n-k} k t^{k-1}.$$

ここで $t=1$ とおくと、

$$np = \langle X \rangle.$$

もう一度微分して

$$n(n-1)p^2(pt+q)^{n-2} = \sum_k \binom{n}{k} p^k q^{n-k} k(k-1) t^{k-2}$$

それから $t=1$ とおくと、

$$n(n-1)p^2 = \langle X(X-1)\rangle = \langle X^2\rangle - \langle X\rangle.$$

すなわち、

$$\langle X^2\rangle = n(n-1)p^2 + np.$$

したがって、分散は

$$\langle X^2\rangle - \langle X\rangle^2 = np(1-p) = npq.$$

□

例題 2.4. 天気予報で、気温が平年よりも５度も低かったあるいは高かったといった表現をよく耳にするが、これの意味するところは、気温の分布のばらつき（分散）が結構大きいということで、さほど驚くべきことではないのであろう。

問 4. このことを現実のデータに当って確かめてみよ。

命題 2.5 (Chebyshev's inequality)**.**

$$P(|X-\mu| \geq r\sigma) \leq r^{-2}.$$

Proof.

$$\begin{aligned}
\sigma^2 &= \sum_x (x-\mu)^2 P_x && (1)\\
&\geq \sum_{|x-\mu|\geq r\sigma} (x-\mu)^2 P_x && (2)\\
&\geq \sum_{|x-\mu|\geq r\sigma} r^2\sigma^2 P_x && (3)\\
&= r^2\sigma^2 P(|X-\mu| \geq r\sigma). && (4)
\end{aligned}$$

□

例題 2.6. 区間 $[a,b]$ を n 等分し、区切り上に $n+1$ の点をならべ、各点に同じ重み（$1/(n+1)$）の確率を分布させる。

(i) 平均値と標準偏差を求めよ。（$\mu=(a+b)/2$, $\sigma^2=(b-a)^2\dfrac{2n+1}{6n}$）

(ii) $P(|X-\mu|\geq r\sigma)$ を計算しこれと r^{-2} とを比較する。（r の関数とみてグラフを書く。）

例題 2.7. ある超能力者が、コイン投げ５００回のうち２７０回を当てたとする。これが偶然の出来事だとすると、二項分布の平均値からのずれは２０回。一方標準偏差は $\sqrt{npq}=\sqrt{125}=11.2$ 回、比は $20/11.2=1.8$。結構まれな出来事ではある。

(i) １０００回の内５４０回当てたとき。

(ii) ２０００回のうち１０８０回当てたとき。

チェビシェフの不等式の理論的応用として、

命題 2.8 (S.N. Bernstein)**.** 閉区間 $[0,1]$ 上ので定義された連続関数 $f(x)$ に対して x の多項式 $Q_n(x)$ を

$$Q_n(x)=\sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k}f(k/n)$$

で定める。このとき、

$$\max_{0\leq x\leq 1}|f(x)-Q_n(x)|\to 0 \qquad (n\to\infty)$$

が成り立つ。

3 Poisson 分布

宝くじはなかなか当たらないことで有名であるが、仮に 100 枚のくじに平均１枚の当たりくじ（何等でもよいから）が含まれているとしよう。すなわち、１枚だけ買ったとき、それが当たりである確率は $p = 1/100$ であるとする。このような状況で、２００枚の宝くじを買ったとしよう。当たりくじの枚数の確率分布は、二項分布

$$P_k = \binom{200}{k} \left(\frac{1}{100}\right)^k \left(\frac{99}{100}\right)^{200-k}$$

で与えられる。この分布の期待値は $np = 200 \times 1/200 = 2$ であるから、P_k の値は $k = 2$ のとき最も大きくなりそうであるが、実際そうであろうか。

$k = 0, 1, 2, 3$ の場合の P_k の式を具体的に書いてみると

$$\begin{aligned}
P_0 &= \left(\frac{99}{100}\right)^{200} \\
P_1 &= 200\left(\frac{1}{100}\right)\left(\frac{99}{100}\right)^{199} = 2\left(\frac{99}{100}\right)^{199} \\
P_2 &= \frac{200\cdot 199}{2}\left(\frac{1}{100}\right)^2\left(\frac{99}{100}\right)^{198} = \frac{199}{100}\left(\frac{99}{100}\right)^{198} \\
P_3 &= \frac{200\cdot 199\cdot 198}{6}\left(\frac{1}{100}\right)^3\left(\frac{99}{100}\right)^{197} = \frac{1}{3}\frac{199}{100}\frac{198}{100}\left(\frac{99}{100}\right)^{197}.
\end{aligned}$$

ここで、$\left(\frac{99}{100}\right)^{197} \sim \left(\frac{99}{100}\right)^{200}$ はほとんど等しいから、P_0, P_1, P_2, P_3 の比は、ほぼ

$$P_0 : P_1 : P_2 : P_3 = 1 : 2 : 2 : \frac{4}{3}$$

に一致し、$P_0 < P_3 < P_1 = P_2$ となっている。期待に相違して、確率のピークの位置は、$k = 1$ と $k = 2$ の二つにまたがっている。

以上の考察は、次の形に一般化される。いま、二項分布

$$P_k = \binom{n}{k} p^k (1-p)^{n-k}$$

で、p の値は小さく n は大きく $\mu = np$ は 1 と比較できる程度の大きさであるとする。さて、

$$\frac{P_k}{P_{k-1}} = \frac{n-k+1}{k}\frac{p}{1-p}$$

で、考える k の範囲を μ の大きさ程度に限定すると、μ は n と比較して小さいから、

$$\frac{P_k}{P_{k-1}} = \frac{\mu}{k}$$

がよい近似で成り立つ。従って、

$$P_k = P_0 \frac{\mu^k}{k!}$$

もよい近似式である。$P_0 = (1-p)^n$ の値は、

$$\log P_0 = n \log(1-p) = n\left(-p - \frac{p^2}{2} - \frac{p^3}{3} - \cdots\right) = -\mu\left(1 + \frac{p}{2} + \frac{p^2}{3} + \cdots\right)$$

より、ほぼ $e^{-\mu}$ に等しい。以上により次の近似式（**Poisson 近似**）が得られた。

定理 3.1. $0 < p < 1$ は小さく n は大きく $\mu = np$ は 1 と比較できる大きさであるとき、

$$\binom{n}{k} p^k (1-p)^{n-k} \fallingdotseq e^{-\mu} \frac{\mu^k}{k!}$$

という近似式が成り立つ。

例題 3.2. 先ほどの宝くじの例では、$\mu = 2$ であるから、Poisson 分布表から、

k	0	1	2	3	4	5
P_k	.135	.271	.271	.180	.090	.036

となって、当たりくじが 2 枚以上含まれる確率は、

$$1 - P_0 - P_1 = 1 - 0.406 = 0.594$$

とそれ程確実なわけではないことがわかる。

問 5. ９９%の確かさで、２枚以上当てるためには、何枚以上宝くじを買わないといけないか。

Poisson 分布表から、$P_0 + P_1$ の値は、$\mu = 6$ のとき 0.017, $\mu = 7$ のとき 0.007 であるから、７００枚程度購入する必要がある。

問 6. $\mu = 2$ の場合には、Poisson 近似は、$P_1 = e^{-\mu}\mu = e^{-\mu}\mu^2/2 = P_2$ となるが、二項分布では

$$\frac{P_2}{P_1} = \frac{200-1}{2}\frac{p}{1-p} = \frac{199}{198}$$

となって、ほんの少し P_2 が P_1 より大きい。

単位時間あたり λ 回おこる現象を T 時間観測するとき、現象の観測される回数の分布を求めてみよう。

観測時間 T を n 当分する。n が大きければ、T/n 時間内に現象のおこる確率は小さいので、この小時間内に 2 回以上続けて現象が観測される確率は、さらに小さく無視して良いであろう。すなわち小時間あたり、現象は 1 回起こるか起きないかのいずれかであり、起こる確率 p は小さいと考えてよい（近似）。このとき、小時間の観測を n 回繰り返した T 時間での観測においては、平均 np 回の現象が起こるはずである。この値は、λT に一致しないといけないので、$p = \lambda T/n$。今、n は大きく取っていて $\mu = np = \lambda T$ は一定の値であるから、Poisson 近似が成立し、n を大きくする程、T/n 時間内に現象が 2 回以上起こらないという仮定が正しいものとなるから、$n \to \infty$ の極限では、近似が厳密な分布に近づき、次の結果が得られた。

定理 3.3. 単位時間あたり λ 回観測される現象を T 時間観測した際の現象の出現回数 X は、**Poisson 分布**

$$P(X = k) = e^{-\lambda T}\frac{(\lambda T)^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

で与えられる。

例題 3.4. Poisson 分布の平均と標準偏差。

Proof. 母関数の方法による。関数

$$f(t) = \sum_{k\geq 0}\frac{t^k}{k!} = e^t$$

を考える。$f(t)$ を t で微分して $t = \mu$ とおくと

$$e^{\mu} = \sum_{k\geq 0}\frac{\mu^{k-1}}{k!}k = \frac{1}{\mu}\sum_{k\geq 0}\frac{\mu^k}{k!}k.$$

これから、

$$\langle X \rangle = \mu.$$

$tf'(t)$ を微分して $t = \mu$ とおくと、

$$e^{\mu}(\mu + \mu^2) = \sum_{k \geq 0} \frac{\mu^k}{k!} k^2$$

となって、

$$\sigma^2 = \langle X^2 \rangle - \langle X \rangle^2 = \mu + \mu^2 - \mu^2 = \mu.$$

□

例題 3.5. バスの待ち時間。１０分間隔で運行されているとして、１０分待っても１台も来ない確率は、$1/e = 0.368$ 程度。

問 7. ７００頁の本に３５０箇所のミスプリント。ミスプリントが２ヶ所以上ある頁はどれくらいあるか？

１ページ当り平均 0.5 ヶ所の間違いがあるので、２ヶ所以上の間違いが見つかる確率は、$\mu = 0.5$ に対する Poisson 分布 P_k を使って、

$$1 - P_0 - P_1 = 0.09$$

となるので、そのページ数は $700 \times 0.09 = 63$ 程度。

例題 3.6. M市では１日平均 0.7 人が交通事故により死亡するという。たまたま、ある１週間で、平均を大幅に上回る９人の死者が出た。この事実を、異常な事態と表現することの妥当性について検討する。

4 正規分布

二項分布の平均値・標準偏差のところで考えた超能力者の問題を再び取り上げる。（$p = 1/2$, $n = 500$）Poisson 近似のところでやったように、

$$P_{k+1}/P_k = \frac{n-k}{k+1}\frac{p}{q}$$

である。

$$\frac{n-k}{k+1} = \frac{n+1}{k+1} - 1$$

は n から始まって単調に減少し最後は $1/n$ で終わる。従って、確率の比は $\dfrac{np}{q}$ から始まって p/nq に向かって単調に減少する。$\mu = np$ が十分大きいときには、$k = (n+1)p - 1$ の付近で比が 1 を通過する。

以上を総合すると、平均値 np が 1 に比べて十分大きいとき、二項分布のグラフは、次のようになる。

さて np が大きくなるとグラフの山がどんどん右に行ってしまいおまけに分布の幅（$\sigma = \sqrt{npq}$）もどんどん広がって収拾がつかなくなるので、平均値が中央にくるようにずらしてやりさらに単位を無次元化して次のような確率変数を導入する（**確率変数の基準化**）。

$$Z = \frac{X - \mu}{\sigma}.$$

Z の平均値は 0 で標準偏差は 1 である。

問 8. これを確かめよ。

確率変数 Z は、$-\mu/\sigma = -\sqrt{np/q}$ と $nq/\sigma = \sqrt{nq/p}$ の間を $\sigma^{-1} = 1/\sqrt{npq}$ 刻みで値を取る。従って、n が大きくなると、確率変数 Z の分布の裾野は $-\infty$ から $+\infty$ に広がり、変数の取り得る値の刻みはどんどん小さくなる。結局、$n = \infty$ という極限では、Z は、総ての実数を取り得るようになり、それと同時に特定の観測値が得られる確率は 0 に近づく。こういった場合でも、個々の確率の大きさを一刻みの幅の帯の面積で置き換えて、その高さを表す関数 $P_z/\Delta z$ を考えてやると、これがある関数 $\rho(z)$ に近づく。そして、範囲の確率は

$$P(a < Z < b) = \int_a^b \rho(z)\, dz$$

と面積で表せるようになる。

このような場合にも Z を確率変数と呼び、$\rho(z)$ は Z の（確率）**密度関数** (density function) と呼ぶ。

例題 4.1 (一様分布 (uniform distribution))**.** 密度関数

$$\rho(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{otherwise} \end{cases}$$

によって定められる確率分布の平均値と標準偏差を求めよ。($\mu = (a+b)/2$, $\sigma = (b-a)/2\sqrt{3}$)

n が大きいときに、二項分布の近似を与える密度関数 $\rho(z)$ を求めてみよう。上で与えた確率変数の基準化に対応して $k \in \{0, 1, \cdots, n\}$ の代わりに

$$z = \frac{k - np}{\sigma}$$

なる変数を導入する。n が大きくなるとき、密度関数の近似式は

$$\rho_n(z) = \frac{P_k}{1/\sigma} = \sigma P_k.$$

極限 $\rho(z) = \lim_{n\to\infty} \rho_n(z)$ を求めるために、Poisson 近似のところでやったように比の関係式を考えよう。

$$\begin{aligned} \frac{\rho_n(z + \Delta z)}{\rho_n(z)} &= \frac{\sigma P_{k+1}}{\sigma P_k} && (5) \\ &= \frac{n-k}{k+1}\frac{p}{q} && (6) \\ &= \frac{nq - \sigma z}{\sigma z + np + 1}\frac{p}{q} && (7) \\ &= \frac{n - \sigma z/q}{n + (\sigma z + 1)/p} && (8) \\ &= \frac{1 - (z/q)(\sigma/n)}{1 + p^{-1}(\sigma z + 1)/n}. && (9) \end{aligned}$$

ここで、p, q は定数であるし、z は特定の点での確率（密度）を考察中ということで固定して考えると、

$$\frac{\sigma}{n} = \sqrt{\frac{pq}{n}}$$

は $\sqrt{n}$ の逆数のスピードで 0 に近づく。という訳で、$n \to \infty$ の時、比はどんどん 1 に近くなる。そこで、この 1 に近い比をもっと詳しく調べ

るために $f_n(z) = \log \rho_n(z)$ という関数を用意して、いま導いた比の関係式の対数をとれば、

$$f_n(z+\Delta z) - f_n(z) = \log(1 - zq^{-1}\sigma n^{-1}) - \log(1 + p^{-1}(\sigma z + 1)n^{-1}) \tag{10}$$

$$= -\frac{z}{q}\frac{\sigma}{n} - \frac{z}{p}\frac{\sigma}{n} + O(n^{-1}) \tag{11}$$

$$= -\frac{\sigma z}{npq} + O(n^{-1}) = -\frac{z}{\sigma} + O(1/n) \tag{12}$$

となって、両辺を $\Delta z = 1/\sigma$ で割って、$\sigma/n \to 0$ に注意して極限 $\lim_{n\to\infty}$ を取れば、微分方程式、

$$\frac{df}{dz}(z) = -z$$

が得られる。これを解けば、$f(z) = -\frac{1}{2}z^2 + \text{const}$, すなわち

$$\rho(z) = Ce^{-z^2/2}$$

が得られる。定数 C は、ρ が確率密度であることから

$$1 = C\int_{-\infty}^{+\infty} dz e^{-z^2/2} = \sqrt{2\pi}C$$

より求まる。

定理 4.2 (Laplace の近似公式)**.** 試行回数 n、平均値 $\mu = np$、標準偏差 $\sigma = \sqrt{np(1-p)}$ の二項分布に従う確率変数 X に対して、X の観測結果 x が、$a \le x \le b$ という範囲に含まれる確率 $P(a \le X \le b)$ は、n が大きいとき近似的に、

$$P(a \le X \le b) = \frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta} e^{-z^2/2}dz,$$
$$\alpha = \frac{a-\mu}{\sigma}, \quad \beta = \frac{b-\mu}{\sigma}$$

によって計算できる。

近似の精度は、p の大きさによって変化するが、p の大きさが $0.1 \le p \le 0.9$ 程度の場合には、$n \ge 100$ を目安とするとよいかな。あるいは、p の値が 0.5 に近いときには、$n \ge 50$ 程度でも使えるらしい。

例題 4.3. 超能力者の問題、再考。正規分布表の使い方。

$\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ の確率を表から読み取る。the point of 5 percent ($z = 1.64$).

Proof. $p = 1/2$, $n = 500$ の二項分布を考えると、Chebyshev の不等式から

$$P(X \geq 270) = \frac{1}{2}P(|X - \mu| \geq 1.8\sigma) \leq \frac{1}{2}(1.8)^{-2} = 0.154321$$

であるが、正規分布による近似を使えば、

$$\frac{1}{2}P(|X - \mu| \geq 1.8\sigma) = \frac{1}{2}P(|Z| \geq 1.8) = \frac{1}{2} - 0.464 = 0.036$$

と計算できて、結構小さい値であることがわかる。 □

5 正規分布と変数変換

$$P(a \leq X \leq b) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b dx\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

の形の密度関数をもつ確率分布を**正規分布** (normal distribution) という。ここで、μ と σ はそれぞれ平均値、標準偏差になっている。確率密度関数 $\rho(x)$ をもつ確率変数 X を考える。

$$P(a \leq X \leq b) = \int_a^b dx\rho(x).$$

このとき、X の関数式 $Y = f(X)$ は新たな確率変数を定める。$f(X)$ の密度関数 $p(x)$ を求めてみよう。具体例で示す。

正規分布の基準化。これは X のかわりに

$$Z = \frac{X - \mu}{\sigma}$$

という新しい確率変数を考えることで、Z の確率分布は

$$P(a \leq Z \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b dz\, e^{-\frac{z^2}{2}}$$

となる。

例題 5.1. Z^2 の分布。

$$P(a < Z^2 < b) = \frac{1}{\sqrt{2\pi}} \int_a^b dv v^{-1/2} e^{-v/2}.$$

仮説検定の原理 (hypothesis testing)

(i) 考察する確率現象に対して確率分布を仮定する。

(ii) 観測結果を反映する確率変数 X でその確率分布が計算可能であるもの（容易に計算できるもの）を選ぶ。

(iii) (ii) で選んだ確率変数 X を利用して観測されたデータが得られる確率を求める。

(iv) (iii) の計算結果に基づいて最初の仮定が正しかったかどうかについて判断する。

注意すべき点：

(iii) で求めた確率は小さい程 (i) の仮定を疑うことになる。では、(iii) の確率が小さくなかったときはどうかというと、(i) の仮定は合っているかもしれないしあるいは違っているかも知れない。つまり、特定の結論は何も得られないということである。従って、(i) での仮定は否定されて初めて価値があるという意味で、**帰無仮説** (null hypothesis) と呼ばれる。一方でまた、帰無仮説が否定された場合に、では、どういう結論とするかを予め状況から判断して設定しておくことが多い。これを**対立仮説** (alternative hypothesis) と呼ぶ。背理法との類似点。p 値との関係。

さらにまた、(iii) の確率がどの程度の大きさであれば小さいと言えるのか（これを**有意水準または危険率**, the level of significance という）というのは考察する問題にもよるし、判断する人の考え方にもよる。その意味で本来、非常に主観的なものであるが、統計の教科書ではしばしば 5%, 1% との大小で判断されることが多い。（人の命に関わる問題であれば 1% でも大きすぎるだろうし、予想が外れても深刻でない問題ならば 60% の成算で十分な場合もあるだろう。）この 1%, 5% の数値は、p 値を簡単には計算できなかった時代の名残りか。

仮説検定ではよく、

「有意水準 5% で帰無仮説は否定され、よって対立仮説が採択される」

といった「厳かな」表現が使われるが、その実態をよく把握して、見かけの「統計学的権威」に惑わされないようにすべきである。

例題 5.2. サイコロを１００回投げたところ、偶数の目が５８回、奇数の目が４２回出た。サイコロは正常といえるだろうか。$\mu = 50$, $\sigma = 5$, $(58 - 50)/5 = 1.6$.

ここで、**危険率**（有意水準）の意味について。サイコロは正しいという仮説をもとにしても、対立仮説として、（１）サイコロは正しくない、をとるか、（２）サイコロは偶数が出やすいをとるかで危険率の範囲を片側によせるか、両側にとるかで変ってくる。すなわち、何を問題にするかで同じ確率分布を仮定して同じ危険率を採用しても判断が変ってくるわけである。

これは一見パラドックスであるが、両側検定の方が帰無仮説を棄却するのにより慎重な態度であるといえる。したがって、片側検定を行う場合には、それだけの根拠を持った上で行うのがよい。例えば、薬の有効成分の含有率を検定する際には、(i) もしその薬がある一定の量以上含まれていれば効果があり、多すぎても副作用等の影響が心配ないときには、片側検定で十分であるが、(ii) 多量摂取した場合の副作用が深刻である場合とかそもそも効き目があるのかどうかがはっきりしない場合には、慎重な両側検定を行うのがよいだろう。

以上の (i), (ii) の区別についても、どの程度を多量とみるのか、また有効成分の含有量のばらつき（分散）との関係等、諸々の事情関わってくるので、状況をよく理解した上で適用しないといけない。単に統計の本に載っている例をそのまま真似して機械的に判断するのはとても危険である。

パラメータの推定

確率現象によっては、確率分布の形が予め（理論的あるいは経験的に）予測されており、観測されたデータから確率現象を特徴付けるパラメータの推測が問題になる。この場合、上の (ii)、(iii) の計算を逆用して推測に役立てることができる（いわゆる**区間推定** (interval estimation)）。

２項分布の確率変数を X として、パラメータ p の値を推定するために、

$$P\left(\left|\frac{x - np}{\sqrt{np(1-p)}}\right| \leq K_\alpha\right) = 1 - \alpha$$

を逆用して X の実現値 x が得られるならば、p は**信頼度** (the level of confidence) $1-\alpha$ で不等式

$$\left|\frac{x-np}{\sqrt{np(1-p)}}\right| \leq K_\alpha$$

を満たすと考える。

例題 5.3. 内閣支持率 p を誤差が１%以内になるように信頼度９５%で推定するためには、何人以上を調査しないといけないか。

Proof. n 人に調査から得られた平均値を p_0 とすると ($x = np_0$)、信頼度９５%で、

$$p_0 - K_\alpha\sqrt{\frac{p(1-p)}{n}} < p < p_0 + K_\alpha\sqrt{\frac{p(1-p)}{n}}$$

が成り立つ。これは、p についての２次不等式になるが、n が大きいときは、ルートの中の p を p_0 で置き換えた不等式

$$p_0 - K_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} < p < p_0 + K_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$$

が近似的に成り立つ。したがって、

$$1.96\sqrt{\frac{p_0(1-p_0)}{n}} \leq 0.01$$

を解いて

$$n \geq (1.96/0.01)^2 p_0(1-p_0).$$

ここで、$p_0(1-p_0)$ の最大値が $1/4$ であることに注意すれば、悪くても

$$n \geq (1.96/0.01)^2/4 = \frac{196^2}{4} = 9604$$

とすればよい。 □

問 9. 支持率の誤差を、信頼度９５%で、２%、３%以内にするためには、調査の人数をそれぞれ何人程度にするべきか。

$$\frac{1}{4}\left(\frac{1.96}{0.02}\right)^2 = 2401, \quad \frac{1}{4}\left(\frac{1.96}{0.03}\right)^2 = 1067.$$

例題 5.4. 内閣支持率を１００人に聞いたところ、４０人が指示すると答えた。信頼度９５%で、支持率の範囲を求めよ。

Proof. 上の式で、$n = 100$, $p_0 = 0.4$, $K_\alpha = 1.96$ を代入して、

$$0.35 < p < 0.45.$$

□

6 χ^2 検定（適合度）

二項分布の確率変数 X に対して $Y = n - X$ とおいて、

$$V = \frac{(X - np)^2}{np} + \frac{(Y - nq)^2}{nq}$$

という確率変数を考えると、

$$V = \frac{(X - np)^2}{Npq} = Z^2$$

となるので、V は密度関数

$$\rho(v) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} v^{-1/2} e^{-v/2} & \text{if } v > 0 \\ 0 & \text{otherwise} \end{cases}$$

を持つ確率変数になる。

これを一般化すると次の定理になる。

定理 6.1. 結果が r 通りの確率分布 $p_1, \cdots, p_r$ をもつ確率変数 X で記述される観察を n 回繰り返した結果、j 番目の結果が得られる回数を表す確率変数を X_j とする。このとき、

$$V = \frac{(X_1 - np_1)^2}{np_1} + \cdots + \frac{(X_r - np_r)^2}{np_r}$$

は n, np_i が大きいとき近似的に

$$P(a \le V \le b) = C \int_a^b dv\, v^{(r-3)/2} e^{-v/2}$$

という確率分布に従う。この右辺で与えられる確率分布を**自由度** (the degree of freedom) $r - 1$ の**カイ二乗分布** (χ^2 distribution) と呼ぶ。

問 10. (i) 関数 $f_d(v) = v^{(d-2)/2}e^{-v/2}$ $(v > 0)$ のグラフの様子を調べる。

(ii)

$$\int_0^\infty dv f_d(v) = 2^{d/2}\Gamma(d/2)$$

を示せ。

例題 6.2. さいころの目の出た度数が

1	2	3	4	5	6
95	98	107	93	103	104

であるとすると、

$$\begin{aligned}\chi^2 &= (95-100)^2/100 + (98-100)^2/100 + (107-100)^2/100 + (93-100)^2/100 \\ &\quad + (103-100)^2/100 + (104-100)^2/100 \\ &= 1.52.\end{aligned}$$

自由度 5 の χ^2 分布の表から $\chi^2 \geq 1.52$ という結果が得られる確率は 0.9 より大きく 0.95 より小さい（どちらかというと 0.9 に近い）。したがって歪んでいる兆候はみられない。

例題 6.3. 相対度数は不変でも実験を行う回数が N 倍にふえると、χ^2 の値も N 倍になり、起こる確率が小さくなる。

上の V の実現値を

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

と表示することが多い。O は observed value で E は expected value の意味。

例題 6.4. 次の表はある選挙での政党別の得票数と当選者数を表す。
A:50%, B:30%, C:20%
A:100, B:60, C:20
この選挙は、投票者の意志が十分反映された選挙といえるだろうか。

Proof.

$$\chi^2 = \frac{(100-90)^2}{90} + \frac{(60-180\times 0.3)}{180\times 0.3} + \frac{(20-180\times 0.2)^2}{180\times 0.2} = 8.8888.$$

自由度は 2 で、

$$0.01 < P(\chi^2 > 8.89) < 0.025$$

であるから、選挙制度を見直す必要があるだろう。 □

例題 6.5. サイコロを６００回投げたところ、１の目が１０１回、２の目が１００回、３の目が９９回、４の目が１００回、５の目が１０２回、６の目が９８回出た。

このサイコロは正常といってよいだろうか。

Proof. 正常と仮定すると、

$$\chi^2 = \frac{1+0+1+0+2^2+2^2}{100} = 0.1$$

となって、自由度５の χ^2 を適用すれば、$P(\chi^2 \geq 0.1) = 0.9998$ であることから、ばらつきが異常に少なく正常なサイコロではない。 □

メンデルのデータが整いすぎていること。雑種第一世代の交配。エンドウ豆の色。黄色:緑色＝６０２２:２００１というデータから、$\chi^2 = 0.0166$ となるので、自由度１の χ^2 分布表から、

$$P(\chi^2 \leq 0.016) = 0.1$$

となる。

分布のパラメータが未定である場合。

例題 6.6. くじを１０回ひいて当たりのでた回数の分布。

0	1	2	3	4
3	5	22	11	9

これは、二項分布といっていいか。$p = 0.226$.

7 χ^2 検定（独立性）

２つの確率変数 X, Y が**独立** (independent) であるとは、

$$P(a \leq X \leq b, c \leq Y \leq d) = P(a \leq X \leq b)P(c \leq Y \leq d)$$

が成り立つこと。

いま X, Y の定める確率分布をそれぞれ $p_1, \cdots, p_m$, $q_1, \cdots, q_n$ であるとすると、

$$P(X = x_i, Y = y_j) = p_i q_j$$

であるし、逆にこの式で与えられる２次元の分布をもつ X, Y は独立になる。

同様にして、X, Y が密度関数 $f(x)$, $g(y)$ で与えられるとき、X, Y が独立になるための必要十分条件は (X, Y) の密度関数が $f(x)g(y)$ で与えられることである。

独立な確率変数 X, Y については、

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$$

という形の加法法則が成り立つ。

確率変数 X, Y が独立であるなしにかかわらず、

$$P(X = x_i, Y = y_j) = p_{i,j}$$

あるいは、

$$P((X, Y) \in D) = \int_D dxdy\, \rho(x, y)$$

と表すとき、

$$\mu_X = \langle X \rangle = \sum_{i,j} p_{i,j} x_i \quad \mu_Y = \langle Y \rangle = \sum_{i,j} p_{i,j} y_j$$

あるいは、

$$\mu_X = \int_D dxdy\, \rho(x, y)x, \quad \mu_Y = \int_D dxdy\, \rho(x, y)y$$

となる。さらに、

$$\sigma^2_X = \int_D dxdy\, \rho(x, y)(x - \mu_X)^2$$

等である。

ここで、**共分散** (covariance) を

$$\langle (X-\mu_X)(Y-\mu_Y)\rangle = \int_D dxdy\,\rho(x,y)(x-mu_X)(y-\mu_Y)$$

で定める。この量は、$x-\mu_X$ と $y-\mu_Y$ の符号が一致する傾向にあるときに大きい値をとり、2つの確率変数 X, Y の相関傾向を表していると考えられる。より客観的には、標準偏差の積との比を取って

$$\rho_{X,Y} = \frac{\langle (X-\mu_X)(Y-\mu_Y)\rangle}{\sigma_X\sigma_Y}$$

を**相関係数** (correlation coefficient) と呼ぶ。これは二種類の観測値の増減が揃う傾向にあるかどうかの目安を与えるもので、$\rho > 0$ のとき（正の相関）は揃う傾向を、$\rho_{X,Y} < 0$ のとき（負の相関）は増減が逆に連動する傾向を示す。

命題 7.1. 相関係数は次の性質をみたす。

(i) $-1 \le \rho_{X,Y} \le 1$.

(ii) X と Y が独立ならば、$\rho_{X,Y} = 0$.

(iii) $\rho_{X,Y} = \pm 1$ となるのは、X, Y が一次関係式 $Y = aX + b$ を満たすときで、このとき、$\rho_{X,Y} = a/|a|$.

例題 7.2. さいころを n 回投げるとき1の目のでる回数 X と2の目のでる回数 Y。$P(X=0,Y=0) = \left(\frac{4}{6}\right)^n$, $P(X=0) = \left(\frac{5}{6}\right)^n = P(Y=0)$ であるから X と Y とは独立ではない。

例題 7.3. 上の問題で $P(k,l) = P(X=k, Y=l)$ を求めよう。母関数

$$F(t_1,\cdots,t_6) = \sum_{n_1,\cdots n_6} P(n_1,\cdots,n_6)t_1\cdots t_6 = 6^{-n}(t_1+\cdots t_6)^n$$

で $t_3 = \cdots = t_6 = 1$ とおくと、

$$F(s,t) = \sum_{k,l} P(k,l)s^k t^l = 6^{-n}(s+t+4)^n = \left(\frac{2}{3}\right)^n (1+s/4+t/4)^n.$$

これから

$$P(k,l) = \left(\frac{2}{3}\right)^n \frac{n!}{k!l!(n-k-l)!} \left(\frac{1}{4}\right)^{k+l}$$

となる。(相関係数は、$-1/5$)

例題 7.4. コイン投げを１０回繰り返すとき、１回目と２回目の結果を表す確率変数は独立。

二次元のデータ
身長と体重、英語と数学、温度と圧力など。
散布図 (scattered diagram) による表現。

$$N = \sum_{i,j} O_{ij}.$$

$$\mu_X = \sum_i \frac{a_i}{N} x_i, \qquad \mu_Y = \sum_j \frac{b_j}{N} y_j,$$

$$\sigma_X^2 = \sum_i \frac{a_i}{N} (x_i - \mu_X)^2, \sigma_Y^2 =$$

$$\rho_{X,Y} = \frac{1}{\sigma_X \sigma_Y} \sum_{i,j} \frac{O_{ij}}{N} (x_i - \mu_X)(y_j - \mu_Y).$$

独立性の検定とその自由度

例：購読新聞の種類と支持政党。新型の抗癌剤の効果。A紙、M紙、Y紙、J党、S党、K党、M党、C党。神社と宝くじ。インフルエンザの新薬、A、B。ひいたかひかないか。飲んだか飲まないか。２つの 2×2 表。効き目の判定。

２種類の項目ごとに分類した観測度数の表を用意する。

O_{11}	O_{12}	$\cdots$	O_{1n}	a_1
O_{21}	O_{22}	$\cdots$	O_{2n}	a_2
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
O_{m1}	O_{m2}	$\cdots$	O_{mn}	a_m
b_1	b_2	$\cdots$	b_n	N

次に、

$$p_i = \frac{a_i}{N}, \qquad q_j = \frac{b_j}{N}$$

とおいて、2つの要因が独立に作用していると仮定したときの期待度数を

$$E_{ij} = Np_iq_j = \frac{a_ib_j}{N}$$

と考える。このとき、

$$\sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

は、近似的に自由度 $(m-1)(n-1)$ の χ^2 分布になる。

もし、p_i, q_j が予めわかっているときの自由度は $mn-1$ であるが、上の計算では、p_i, q_j は観測データを利用して計算してあるので、自由度は $(m-1)+(n-1)$ だけ減少する。

$$\sum_i a_i = N = \sum_j b_j$$

という関係があるので、$a_1, \cdots, a_m$ のうち自由なのは $m-1$ 個である。

例題 7.5. 2種類の風邪の予防薬の効果の違いを検証するために、２００人の被験者に薬を飲んでもらったところ、次のような結果が得られた。

A	10	70	80
B	20	140	160
	30	210	240

「風邪の罹患率が半減する新薬！」と宣伝してよいかどうか。

Proof. 計算すると $\chi^2 = 3.92$ なるので、自由度１の χ^2 分布表から、薬の効果が一緒でもこのようなことが起こる確率は、0.025 より大きく 0.05 よりは小さいことがわかる。

営業部長であれば、これで十分と判断するかもしれないが、科学者としては、効果に違いがあることすら判断に迷う数字である。まして、罹患率が半減すると言い切ることはできない。 □

問 11. 上の問題で信頼度９９%以上で、効果の違いを確認するためには、観測回数を５倍程度にしないといけない。

8　t 検定

独立な確率変数 X, Y の和 $X+Y$ の平均値と分散は、

$$\mu_{aX+bY} = a\mu_X + b\mu_Y, \quad \sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$$

となる。

科学的なデータは、同一現象の繰り返し観測によって得られる。1回の観測で得られるデータが確率変数 X によって記述されるならば、そのような観測を n 回繰り返した場合に、k 回めの観測データを記述する確率変数を X_k で表すことにすれば、各 X_k は X と同じ確率分布を持ち、X_j, X_k は互いに独立であると考えられる。

このような、同一の確率分布に従う互いに独立な確率変数の組 $X_1, \cdots, X_n$ を大きさ n の**標本**（sample）と呼ぶ。そうすると、観測結果の平均値・分散は、

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n},$$

$$V = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n}$$

なる確率変数によって記述されるだろう。

先ほどの公式から

$$\langle \overline{X} \rangle = \mu_X, \quad \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n}$$

であるが、さらに

$$\langle V \rangle = \frac{n-1}{n}\sigma^2_X$$

が成り立つ。実際、

$$V = \frac{1}{n}\sum_i (X_i - \mu)^2 - (\overline{X} - \mu)^2$$

と書き直して期待値をとると分かる。

$$S^2 = \frac{n}{n-1}V = \frac{(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1}$$

を標本の**不偏分散**（unbiased variance）という。S^2 のかわりに U^2 の記号で表すことも多い。

ちなみに、数学では、最初に断りさえすればどのような（標準的であろうとなかろうと）記号を使おうと自由であるが、「統計学」の分野では、

アルファベットの大文字小文字の区別含めて、「正しい記号」以外は間違いである、とする考え方があるらしい。この辺が「実用科学」と「趣味の科学」の違いなのかも知れない。

誤差の分布

科学実験等の観測値は、様々な要因で真の値 μ のまわりに揺らぎを生ずる。簡単のため個々の要因の効果は、観測値を $\epsilon > 0$ だけ増やすか減らすかのどちらかであり、それぞれ確率 $1/2$ でおこるとする。そして様々な要因はお互いに独立に発生するものとする。

そうすると、観測の結果を与える確率変数は

$$X = \mu + \epsilon(X_1 + \cdots + X_n) = \mu + \epsilon(2Y - n)$$

で与えられることになる。ここで、

$$Y = \sum_{j=1}^{n} \frac{X_j + 1}{2}$$

は、各要因ごとにくじを引いて、$+$ の結果が得られたら当たり、$-$ の結果が得られたらはずれ、と見たときの当たりの総数を表す。

さて、X の分散を求めてみると、

$$4\epsilon^2\sigma_Y^2 = \epsilon^2 n$$

となるので、X の標準偏差は $\sigma = \epsilon\sqrt{n}$ で与えられる。

そうすると、

$$P(a \leq X - \mu \leq b) = P(a \leq 2\epsilon(Y - n/2) \leq b)$$

は近似的に

$$(2\pi)^{-1/2} \int_{a/\epsilon\sqrt{n}}^{b/\epsilon\sqrt{n}} dz e^{-z^2/2}$$

で与えられる（Y の標準偏差が $\sqrt{n}/2$ であることに注意）。

確率変数 X の標準偏差 $\sigma = \epsilon\sqrt{n}$ を使って書き表すと、

$$P(a \leq X - \mu \leq b) = (2\pi\sigma^2)^{-1/2} \int_a^b dx e^{-(x-\mu)^2/2\sigma^2}$$

となる。

このことから、科学実験の観測値を表す確率変数は、密度関数が

$$(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}$$

で与えられる確率分布に従うと考えられる。これも正規分布と呼ぶ。いいかえると**偶然誤差** (accidental error) は正規分布で与えられる。

さて観測データから、観測対象の μ（**母平均**、population mean と呼ぶ）あるいは σ^2（**母分散**、population variance）（これらは、パラメータ、parameter と総称される）をどうやって推定するか。観測データのばらつき具合（すなわち S の大きさ）にもよるが、大量の観測を行えばデータから得られる平均値と標準偏差は真のそれとよい精度で一致するであろう。しかしながら実験によっては大量観察が極めて困難な場合もある。そういった場合でも数少ないデータからどのようにして客観的な情報を引き出したらよいか。

定理 8.1. 個々の観測データがパラメータ (μ, σ) の正規分布に従う大きさ n の標本 $\{X_1, \cdots X_n\}$ に対して、確率変数

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$$

は次の形の分布関数で記述される。

$$\rho(t) = C_{n-1}\left(1 + \frac{t^2}{n-1}\right)^{-n/2}.$$

これを自由度 $n-1$ の **t 分布**（*W.S. Gosset* の *t-distribution*）と呼ぶ。

注意：n が大きいとき、この分布は正規分布（$\mu = 0$, $\sigma = 1$）に近づく。従って、実際問題として上の t 分布が使われるのは、n が大きくないとき（**小標本**、small sample という）だけ。

例題 8.2. 喘息の新薬の効果を判定するために、８人の患者に薬を服用してもらったところ、喘息の発作の回数（１週間の）が次のように変化した。

服用前	3	7	2	6	9	11	5	7
服用後	4	5	2	5	6	7	6	3
改善度	-1	2	0	1	3	4	-1	4

効果があると考えて良いか。

Proof. 患者の服用前の発作の回数の平均値は 6.25, 服用後の平均値は 4.75 で、効果がありそうな変化ではある。

そこで、薬の効果はまったくなく偶然このような結果が起こる確率を考える。喘息を起こす頻度は、諸々の要因が関係していると思えば、正規分布に従うと考えて良いだろう。そこで、回数の差を表す確率変数の標本平均は 0 であったと仮定して、上記のようなことが実現される確率を求めてみる。不偏分散の実現値は、

$$s^2 = \frac{(-1-1.5)^2 + (2-1.5)^2 + \cdots + (4-1.5)^2}{8-1} = \frac{30}{7} = 4.3$$

となり、T の実現値は

$$t = \frac{\overline{x} - 0}{\sqrt{s^2/8}} = 2.05$$

となる。自由度は、$8-1=7$ なので、t 分布表から、

$$0.025 < P(T \geq t) < 0.05$$

となる。

危険率５％なら、効果があると判定できるが、危険率１％では、そのような判断をすることは危険である。

もし新薬にめだった副作用がないのなら、効果を認めてより多くの臨床データを求めることになろうが、副作用の程度如何では慎重な投与が必要であろう。 □

例題 8.3. 次は、ある新素材の比熱の測定値である。

2.31	2.25	2.28	2.30

信頼度９０％で、比熱の範囲を求めよ。

Proof. 測定データの平均値、不偏分散の値は、

$$\overline{x} = 2.285, \qquad s^2 = 0.0007$$

となる。

自由度 $4-1$ の信頼度９０％の T の実現限界は、

$$t = \pm 2.3534$$

なので、母平均 μ は、信頼度９０％で、不等式

$$\left| \frac{\overline{x} - \mu}{\sqrt{s^2/4}} \right| \leq 2.3534$$

をみたす。したがって、求める比熱の範囲は、

$$[2.254, 2.316]$$

となる。 □