確率統計の話題から －条件つき確率の話題を中心に－

杉浦　誠

平成 24 年 8 月 23 日

1　確率を計算しよう

この節では具体的に確率を計算することで、直感と計算結果の違いを比較してみましょう。

例題 1.1 (誕生日) 40 人のクラスで、同じ誕生日の生徒はいるか。その確率を求めよ。ただし、簡単のため 1 年は 365 日とし、365 日のどの日に生まれる確率も等しく $\dfrac{1}{365}$ であると仮定する。[*1]

確率を $\dfrac{40}{365} \fallingdotseq 0.1096$ とするのはもちろん誤りです。正しい解答を見てみましょう。

解答: 余事象を考え、まず 40 人の誕生日がすべて異なる確率 q を求める。

40 人を順に比較していくという方針をとる。

まず、2 人の誕生日が異なる確率は、2 人目が 1 人目と誕生日が異なればよいので $\dfrac{364}{365}$.

次に、3 人目が 1 人目, 2 人目と誕生日が異なる確率は $\dfrac{363}{365}$ であるから、3 人の誕生日が異なる確率は $\dfrac{364}{365} \times \dfrac{363}{365}$ となる。同様に、4 人の誕生日が異なる確率を求めると $\dfrac{364}{365} \times \dfrac{363}{365} \times \dfrac{362}{365}$.

これを繰り返し、$q = \dfrac{364}{365} \times \dfrac{363}{365} \times \cdots \times \dfrac{326}{365} = 0.108768\cdots$ となる[*2] 。

以上より、求める確率は $1 - q \fallingdotseq 0.8912$ である。　□

問 1.1 次の場合に 5 人の誕生日がすべて異なる月となる確率を考える。

(1) どの月に生まれる確率も等しく $\dfrac{1}{12}$ である場合に、5 人の誕生日がすべて異なる月となる確率を求めよ。

(2) 2 月以外の 11 ヶ月に誕生日がある確率を p, 2 月にある確率を $1 - 11p$ とする。ただし、$0 < p < \dfrac{1}{11}$ とする。このとき、5 人の誕生日がすべて異なる月となる確率 $f(p)$ と、$f(p)$ が最大となる p の値を求めよ。

問 1.2 次のようなサイコロを 3 回投げたとき、3 回とも異なる目が出る確率を求めよ。

(1) どの目の出る確率も等しく $\dfrac{1}{6}$ の場合。

(2) 1 と 6 の目の出る確率が $\dfrac{1}{7}$, そのほかの目が出る確率が $\dfrac{5}{28}$ の場合。

1654 年のある日、フランスの数学者パスカルは、ド・メレという貴族から、ある質問を受けた。その質問とは次のような問題であった。パスカルは、この問題を同じ数学者のフェルマーと手紙をやり取りして研究し、その結果生まれたのが、「確率論」という分野である[*3]。

[*1] 実際には季節などによって生まれる確率が異なるので、同じ誕生日の人がいる確率はこれより高くなる (*cf.* 問 1.1, 問 1.2)。

[*2] これを求めるのは電卓では大変です。たとえば、$364 \times 363 \times \cdots \times 325 \fallingdotseq 9.2 \times 10^{98}$ です。[8] に Diaconis & Mosteller による簡便な近似計算法がある。これは平均値の定理より得られる近似式 $\log(1-x) \approx -x$ を利用する方法で、これを用いれば e^x の値が求まる電卓があれば計算できる。

[*3] 現在の確率論はルベーグ積分論を用いて定式化された。これはロシアの数学者コルモゴロフによってなされた (*cf.* [7])。[7] は統計学に関わる人物の業績をその人となりとあわせて (数学的な記述はなく) 書かれいる楽しい本です。

例題 1.2 (ド・メレからパスカルへの質問) 同額の賭け金を出し合い、先に 3 勝したほうが勝ちとするゲームで、時間の関係で途中でやめることになった。その時点で私が 2 勝 1 敗で勝っていたのだが、賭け金の分配方法がよくわからなかった。結局私が 3 分の 2、相手が 3 分の 1 ということにしたのだが、これでよかったのだろうか。

解答: ここでは両者の勝つ確率は等しいと仮定しよう。実際は「私」がリードしているので実力差があると仮定してもよいかもしれないがやめておく[*4]。

このゲームは 5 試合やれば必ず勝負がつくので、この勝負の残り 2 試合をしたとするとゲームの勝敗は以下の表のようになる。ただし、「私」の勝ちを W, 負けを L で表し、現在までの勝敗は 2 勝 1 敗なので順序を考えないとし「(WWL)」と表す。

現在までの勝敗		4 回戦	5 回戦	勝者
(WWL)	→	W	W	私
(WWL)	→	W	L	私
(WWL)	→	L	W	私
(WWL)	→	L	L	相手

2 人の実力は同じという前提なので、この 4 つの場合はどれも $\frac{1}{4}$ で起こる (このため途中で勝敗が決まる場合も最後まで書いた)。つまり、「私」は確率 $\frac{3}{4}$ で勝者のなったはずであるので、したがって賭け金もその割合で配分されなくてはならない。正しい配分は「私」が $\frac{3}{4}$, 相手が $\frac{1}{4}$ の賭け金を取るべきとなる。 □

問 1.3 A 氏と B 氏が同額の賭け金を出し合い、先に 5 勝したほうが勝ちとするゲームを行い、時間の関係で途中でやめることになった。賭け金を両者それぞれの勝つ確率にしたがって配分するとき、次の場合に A 氏が受け取るべき賭け金の割合を決定せよ。ただし、2 人の実力は同じとして考えよ。

(1) その時点で A 氏が 4 勝 2 敗で勝っていた場合

(2) その時点で A 氏が 3 勝 2 敗で勝っていた場合

問 1.4 (ド・メレの 2 つのサイコロ) ド・メレは次のような (1), (2) の賭けを行ったところ、(1) では勝てることが多かったが、(2) では損をよくした。

(1) 1 つのサイコロを 4 回投げて、1 回でも 6 の目が出れば自分の勝ち。

(2) 2 つのサイコロを同時に 24 回投げて、1 回でも 2 つとも 6 の目が出れば自分の勝ち。

それぞれの賭けに勝てる確率を求めることで、原因を調べよ。また、(2) の賭けでは何回以上投げることにすれば勝てる確率が 0.5 より大きくなるか求めよ。

確率論における重要な定理に「大数の法則」がある。これはヤコブ・ベルヌイ (1654–1705) によって紹介された。「大数の法則」とは、簡単に言えば、「個々の事象の予測は無理 (もしくは極めて困難) であっても、充分に多くの試行がなされると仮定するなら、全体像はかなり正確に予想しうる」とする法則である。式で書くと、繰り返し同じ試行を行うとき、その結果の列を $X_1, X_2, \ldots$ とすると、ある定数 c があって

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = c$$

となる[*5]。この c は X_1 の平均と一致する。この定理の厳密な証明はコルモゴロフによってなされた。

[*4] この「私」が勝つ確率を調べるというのが統計学の役割である。この場合**百分率に関する区間推定の精密法** ($cf.$ [6]) を使って区間推定を行うと「私」の勝つ確率 p は 90% の確率で $[0.135, 0.983]$ の範囲にあることがわかる。したがって、「両者の勝つ確率は等しい」という仮説は間違いとは言えないこととなる。

[*5] ここでは収束の意味を説明しない。詳しくは、確率統計学の教科書、例えば [2] を見よ。[2] はエッセイ的なところもあって、確率統計に関する読み物としても楽しめる本です。

ギャンブル産業が成立するのはこの定理に保証されているといえる ($cf.$ [9])。というのは、胴元が有利な賭けにおいては、数回の賭けでは損をすることがあっても、1 月, 1 年という長い期間を見ればかならず一定割合が収益として計算できからである。それを次の例題で見てみよう。

例題 1.3 ルーレットで赤か黒に賭けて勝つ確率は、どちらも $\dfrac{18}{38}$ であるとする。このルーレットにチップを 1 枚ずつ賭け、90 枚持っているチップを 100 枚に増やしたい。100 枚になるか、0 枚になるかまで続けるものとする。100 枚に到達する確率を求めよ。

解答: 1 回の賭けで勝つ確率を $p(=\dfrac{9}{19})$, 負ける確率を $q=1-p(=\dfrac{10}{19})$ とし、目標とする枚数を $N(=100)$ とかく。また、k 枚のチップを持っている人が目標枚数 N 枚に達する確率を a_k とする。

まず、一回の賭けで確率がどう変わるかを考える: $1 \leqq k \leqq N-1$ のとき

$$k\text{ 枚} \quad \begin{array}{c}\text{勝つ}\\ \nearrow \\ \\ \searrow \\ \text{負ける}\end{array} \quad \begin{array}{ll} k+1\text{ 枚} & \text{ここから勝つ確率は } a_{k+1} \\ k-1\text{ 枚} & \text{ここから勝つ確率は } a_{k-1}\end{array}$$

これより、1 回の賭けに勝つ確率は p であったから、

最初に勝って、それから最終的に勝つ確率 $= p \times a_{k+1}$

最初に負けて、それから最終的に勝つ確率 $= q \times a_{k-1}$

となる。よって、最初の賭けは勝ちか負けしかないから、漸化式

$$a_k = p\,a_{k+1} + q\,a_{k-1} \tag{1.1}$$

を得る。一方、所持金がなくなればもう賭けをすることができないから、N 枚に到達できないので $a_0=0$, N 枚に達したらこれ以上賭けをしなければよいので $a_N=1$ となる。

では、この漸化式を解こう。まず、$x=px^2+q$ について、$(px-q)(x-1)=0$ と変形できるので $x=\dfrac{q}{p}, 1.$ ($p \neq \dfrac{1}{2}$ より $\dfrac{q}{p}=\dfrac{1-p}{p} \neq 1$ となることに注意する。) これを踏まえ、漸化式を

$$a_{k+1}-a_k = \frac{q}{p}(a_k - a_{k-1})\text{と変形し、}a_{k+1}-a_k = \left(\frac{q}{p}\right)^k (a_1-a_0) = \left(\frac{q}{p}\right)^k a_1$$
$$a_{k+1}-\frac{q}{p}a_k = a_k - \frac{q}{p}a_{k-1}\text{と変形し、}a_{k+1}-\frac{q}{p}a_k = a_1 - \frac{q}{p}a_0 = a_1 \quad \text{を得る。}$$

上式の最後の等号は $a_0=0$ を用いた。これより、

$$\left(-1+\frac{q}{p}\right)a_k = \left\{\left(\frac{q}{p}\right)^k - 1\right\} a_1$$

を得る。ここで、$k=N$ とし、$a_N=1$ を用いると、$-1+\dfrac{q}{p} = \left\{\left(\dfrac{q}{p}\right)^N - 1\right\} a_1.$ これより a_1 を求め、代入することで、

$$a_k = \frac{\left(\frac{q}{p}\right)^k - 1}{\left(\frac{q}{p}\right)^N - 1}$$

を得る[*6]。以上より、$\dfrac{q}{p}=\dfrac{10}{9}$ で $k=90$ のときだから、求める確率は $\dfrac{\left(\frac{10}{9}\right)^{90}-1}{\left(\frac{10}{9}\right)^{100}-1} \fallingdotseq 0.34866$ となる。[*7] □

[*6] もし $p=\frac{1}{2}$ であれば、(1.1) を変形すると、$a_{k+1}-a_k = a_k - a_{k-1} = \cdots = a_1 - a_0$ となるので、$a_k = k(a_1-a_0)$. ここで $a_N=1$, $a_0=0$ を代入して、$a_k=\frac{k}{N}$ となる。

[*7] 90 枚持っているチップをもっている人が、100 枚になるか 0 枚になるかまで賭けを続けるとき、その賭けの平均回数は約 1047.5 回となる。これは、k 枚のチップを持っている人の N 枚か 0 枚に達するまでの賭けの平均回数を t_k とすると、その漸化式が

この場合は、10 枚のチップを一度に賭けるのが一番よい戦略である。賭けを行う回数を増やすほど、N 枚に到達できる確率は減ってしまう。ちなみに、この例題の方法で 900 枚のチップを 0 枚になる前に 1000 枚にまで増やせる確率を計算すると $\dfrac{\left(\frac{10}{9}\right)^{900}-1}{\left(\frac{10}{9}\right)^{1000}-1} \fallingdotseq \left(\dfrac{10}{9}\right)^{-100} = 0.0000265614\cdots$ となる。

2 条件つき確率とベイズの定理

この節では条件つき確率を導入して、いろいろな例を計算してみます。特に、最近様々に応用されているベイズの定理について考えましょう[*8]。

定義 2.1 事象 A, B について、事象 A が起こったときの事象 B の起こる条件つき確率 $P_A(B)$ を次で定義する。

$$P_A(B) = \frac{P(A \cap B)}{P(A)}$$

ただし、$P(A) > 0$ の場合のみに定義するものとする。[*9]

例 2.1 (シンプソンのパラドックス) A 高校と B 高校からそれぞれ 40 人を選び国語と数学のどちらが好きか調査したところ、左の表のような結果を得た。ここで、事象 A, B はそれぞれ生徒が A 高校, B 高校に属するという事象を、事象 R は国語より数学が好きという事象、事象 $\overline{R}$ は数学より国語が好きという事象を表す。このとき、A 高校で国語より数学が好きという生徒の割合は $20/40 = 0.5$ となる。一方、B 高校では $16/40 = 0.4$ となる。これより、A 高校のほうが B 高校より国語より数学が好きという生徒の割合が多いことがわかる。

	R	$\overline{R}$	計
A	20	20	40
B	16	24	40
計	36	44	80

ところが、ある先生が性別によって結果が異なるかも知れないと、性別を考慮してデータを見たところ、左の表のような結果を得た。このとき、男子 (M) について、国語より数学が好きという生徒の割合は A 高校では $18/30 = 0.6$, B 高校では $7/10 = 0.7$ であり、女子 (F) についての割合は A 高校では $2/10 = 0.2$, B 高校では $9/30 = 0.3$ となる。つまり、男子であれ女子であれ、B 高校のほうが A 高校より国語より数学が好きという生徒の割合が多いことがわかる。

	R_M	$\overline{R}_M$	M小計	R_F	$\overline{R}_F$	F小計	計
A	18	12	30	2	8	10	40
B	7	3	10	9	21	30	40
計	25	15	40	11	29	40	80

このように全体の傾向が、新しい要因を組み込んだとき全面的に否定されてしまうような結果を得ることをシンプソンのパラドックスという ($cf.$ [10]) [*10]。

これを条件つき確率の記号で表すと次のようになる。

$t_k = p(t_{k+1}+1) + q(t_{k-1}+1)$, $t_0 = t_N = 0$, となることから従う。
実際、$s_k = t_{k+1} - t_k$ とすると、漸化式は $s_{k+1} = \frac{q}{p}s_k - \frac{1}{p}$, $s_1 = t_1 - t_0$ となり、これは、$s_k = (t_1 - t_0 - \frac{1}{q-p})(\frac{q}{p})^k + \frac{1}{q-p}$ と解ける。よって、$t_n = \sum_{k=1}^{n-1}(t_{k+1}-t_k) + t_0 = (t_1 - t_0 - \frac{1}{q-p})\frac{(q/p)^n - 1}{(q/p)-1} + \frac{1}{q-p} + t_0$ となるが、$t_0 = t_N = 0$ より、t_1 を定めることで、$t_k = \frac{k}{q-p} - \frac{N}{q-p}\frac{(q/p)^k - 1}{(q/p)^N - 1}$ となることがわかる。これに数値を代入すればよい。

[*8] コンピューターの分野においては Mozilla Thunderbird は迷惑メールの判定にベイズの定理を使用している (wikipedia より)。CNET JAPAN の 2003/3/10 の記事に「グーグル、インテル、MS が注目するベイズ理論」がある。経済分野では [5] で、ゲームの理論と関連させた興味深い結果を見ることができる。ベイズ推定を実際に活用するためには複雑な計算を伴う。このため、計算機の発達もベイズ理論を利用のために必要であった ($cf.$ [3])。

[*9] 通常は $P(B|A)$ と表します。少なくとも私は高校教科書やその参考書以外で $P_A(B)$ の記号は見たことがありません。この講義は、中学高校の数学教員を対象として行うため $P_A(B)$ を用います。また、A の余事象に $\overline{A}$ は m 用いず、A^{c} を用いることが通例です。

[*10] 相関係数についても同様に、全体で見ると正の相関を示すが、部分で見るとどちらも負の相関を示すことがある ($cf.$ 注意 3.1)。

A, B をそれぞれ選んだ生徒が A 高校, B 高校の生徒であるという事象、R を国語より数学が好きであるという事象とすると、前半の表より

$$P_A(R) = \frac{20}{40} = 0.5, \quad P_B(R) = \frac{16}{40} = 0.4, \qquad \text{よって} \quad P_A(R) > P_B(R).$$

後半は、それにその生徒が男子であるという事象 M と女子であるという事象 B 組み込むと、

$$P_{A\cap M}(R) = \frac{18}{30} = 0.6, \quad P_{B\cap M}(R) = \frac{2}{10} = 0.2, \qquad \text{よって} \quad P_{A\cap M}(R) < P_{B\cap M}(R),$$
$$P_{A\cap F}(R) = \frac{7}{10} = 0.7, \quad P_{B\cap F}(R) = \frac{9}{30} = 0.3, \qquad \text{よって} \quad P_{A\cap F}(R) < P_{B\cap F}(R)$$

と表される。このように条件つき確率は直感が働かないことが多い。

条件つき確率の性質をいくつか調べよう。

$P_A(\cdot)$ は全事象を A に制限した確率とみなせる。また、$P_A(U) = P_A(A) = 1$ (U は全事象), $P_A(\emptyset) = 0$ であり、事象 B, C が排反 ($B \cap C = \emptyset$) なら

$$P_A(B \cup C) = P_A(B) + P_A(C)$$

となる。また、次の乗法定理が成立する。これは定義より明らかであろう。

定理 2.2 (乗法定理) 2 つの事象 A, B がともに起こる確率 $P(A \cap B)$ は

$$P(A \cap B) = P(A)P_A(B)$$

定理 2.3 (ベイズの定理) A および $C_1, C_2, \cdots, C_n$ は事象であり、全事象 U に対して

$$C_1 \cup C_2 \cup \cdots \cup C_n = U \qquad C_i \cap C_j = \emptyset \ (i \neq j)$$

を満たすとする。このとき、

$$P_A(C_i) = \frac{P(C_i)P_{C_i}(A)}{P(C_1)P_{C_1}(A) + P(C_2)P_{C_2}(A) + \cdots + P(C_n)P_{C_n}(A)} \quad (i = 1, 2, \cdots, n) \tag{2.1}$$

が成立する。特に B を事象とし、$n = 2$, $C_1 = B$, $C_2 = \overline{B}$ (B の余事象) とすると次のようになる。

$$P_A(B) = \frac{P(B)P_B(A)}{P(B)P_B(A) + P(\overline{B})P_{\overline{B}}(A)} \tag{2.2}$$

証明: 乗法公式により $P(C_i)P_{C_i}(A) = P(C_i \cap A)$. また、

$$\begin{aligned} P(C_1)P_{C_1}(A) + P(C_2)P_{C_2}(A) + \cdots + P(C_n)P_{C_n}(A) &= P(C_1 \cap A) + P(C_2 \cap A) + \cdots + P(C_n \cap A) \\ &= P(A) \end{aligned}$$

第 2 の等号は $(C_i \cap A) \cap (C_j \cap A) = \emptyset$ $(i \neq j)$ と $C_1 \cup C_2 \cup \cdots \cup C_n = U$ を用いた。よって、これを (2.1) の右辺に代入することで主張を得る。 □

まず、次のような例題を考えましょう。

例題 2.2 ある病原菌の検査試薬は、病原菌がいるのに誤って陰性と判断する確率が 1%, 病原菌がいないのに誤って陽性と判断する確率が 2% である。全体の 1% がこの病原菌に感染している集団から 1 つの個体を取り出す。この検査結果が陽性だったときに、実際には病原菌に感染していない確率を求めよ。[*11]

[*11] 実際の検診の例では、乳がん検診でのマンモグラフィーにおいて、乳がんなのに誤って陰性とすることは (ほとんど) ないが、乳がんでないのに誤って陽性とする確率が 9% で、40 歳代での罹病率は 0.3% だそうである (NHK ためしてガッテン、数字トリック見破り術、2011 年 7 月 6 日放送から)。

解答: 取り出した個体が感染しているという事象を A, 検査結果は陽性であるという事象を E とする。このとき、与えられた条件を式にすると次のようになる。

$$P_A(\overline{E}) = 0.01, \quad P_{\overline{A}}(E) = 0.02, \quad P(A) = 0.01$$

求めるべきは $P_E(\overline{A})$ である。$P(\overline{A}) = 1 - P(A) = 0.99$ より、

$$\begin{aligned} P(E) &= P(A \cap E) + P(\overline{A} \cap E) = P(A)P_A(E) + P(\overline{A})P_{\overline{A}}(E) \\ &= 0.01 \times (1 - 0.01) + 0.99 \times 0.02 = 0.99 \times 0.03 \end{aligned}$$

よって、$P_E(\overline{A}) = \dfrac{P(\overline{A} \cap E)}{P(E)} = \dfrac{0.99 \times 0.02}{0.99 \times 0.03} = \dfrac{2}{3}$ となる。[*12] $\square$

問 2.1 ある製品を製造する 2 つの工場 A,B があり、A 工場の製品には 3%, B 工場の製品には 4% の不良品が含まれているとする。A 工場の製品と B 工場の製品を、4 : 5 の割合で混ぜた大量の製品の中から 1 個を取り出す。それが不良品であったときに、A 工場の製品である確率を求めよ。

問 2.2 ある工場では、機械 M_1, M_2, M_3 で全製品のそれぞれ 60%, 30%, 10% を製造していて、これらの機械で生じる不良品の割合は 2%, 3%, 6% である。いま、1 個の不良品が見つかったとき、それが機械 M_3 で製造されたものである確率を求めよ。

例題 2.3 (3 囚人問題) 3 人の囚人 A, B, C がいる。1 人が恩赦になって釈放され、残り 2 人が処刑されることがわかっている。誰が恩赦になるか知っている看守に対し、A が「B と C のうち少なくとも 1 人処刑されるのは確実なのだから、2 人のうち処刑される 1 人の名前を教えてくれても私についての情報を与えていることにはならないだろう。1 人を教えてくれないか。」と頼んだ。看守は A の言い分に納得して「B は処刑される。」と答えた。それを聞いた A は「これで釈放されるのは自分と C だけになったので、自分の助かる確率は 1/3 から 1/2 に増えた。」といって喜んだ。実際には、この答えを聞いたあと、A の釈放される確率はいくらになるか。

解答: A, B, C をそれぞれ囚人 A, B, C が恩赦される事象とすると、A, B, C が恩赦される確率は等しいと考えられるので、$P(A) = P(B) = P(C) = \dfrac{1}{3}$ となる。
次に、F で看守が「B は処刑される」と告げる事象をあらわすと、
もし A が恩赦されるのであれば、看守は B, C のどちらと告げてもよいので $P_A(F) = \dfrac{1}{2}$.
もし B が恩赦されるのであれば、看守が「B は処刑される」と告げる可能性はないので、$P_B(F) = 0$.
もし C が恩赦されるのであれば、看守は必ず「B は処刑される」と告げるので、$P_C(F) = 1$.
よって、求める確率は $P_F(A)$ であるから、ベイズの定理を用いて

$$P_F(A) = \frac{P(A)P_A(F)}{P(A)P_A(F) + P(B)P_B(F) + P(C)P_C(F)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 0 + \frac{1}{3} \times 1} = \frac{1}{3}$$

となる。 $\square$

これは、冷静に考えれば明らかと思えるだろう。これと同型の次の問題を考えてみよう。

[*12] ここで、集団を「全体の 1/2000 がこの病原菌に感染している」と取りかえると $P_E(\overline{A}) = 0.9758\cdots$ となる。
これは、世間一般の水準からいえばめったにない強い証拠に見えても、極めて珍しいことに比べれば頻繁に起こるに過ぎない場合、頻繁に起こりうる結果をもってより珍しい原因の証拠とはできないことを意味している。例えば殺人事件において、血液型の一致が主な証拠での冤罪事件がこれにあたるであろう。証拠自体がどれほどしっかりしていても、偶然証拠に合致する無実の人にいきあたる確率のほうが犯罪者に出会う確率よりはるかに大きいからである。とくに珍しい事件に対してはそれを上回るまれな事実でないと証拠にならないことを肝に銘じて、危険な偏見を避けるべきである。また、「大地震の前兆として起こる現象」とされているものの多くはこれに相当するのではないだろうか ($cf.$ [2])。

例題 2.4 (3 ドア問題, モンティ・ホールのジレンマ) 3 つの扉のうち 1 つだけに賞品が入っていて、回答者はそれを当てたら賞品がもらえる。ただし扉は次のように 2 段階で選ぶことができる。

1. まず回答者は 3 つの扉からどれか 1 つを選ぶ、
2. 次に、答を知っている司会者が、選んでいない扉で賞品の入っていない扉 1 つを開けてみせる。ただし、回答者が当たりの扉を選んでいる場合は、残りの扉からランダムに 1 つを選んで開けるとする。このあと回答者は扉を 1 回選び直してもよい。

2 で扉を換えるのと換えないのと、どちらが当る確率が高いか？

解答: 扉を A, B, C とし、回答者が選んだ扉を A とし、司会者が選んで開けた扉が B だったとする。
A, B, C でそれぞれ A, B, C の扉に賞品があるという事象とし、司会者が B の扉を開けるという事象を S とすると、3 囚人問題の場合と全く同様に $P_S(A) = 1/3$, $P_S(C) = 2/3$ となる。よって、扉を換えるほうが当る確率が高い。[*13]　□

問 2.3 例題 2.4 で扉が A, B, C, D, E の 5 つの扉のうち 1 つだけに賞品が入っていている場合を考える。回答者が選んだ扉が A であり、次の (1), (2) のように司会者が扉を選んで開けたとする。このとき、賞品が C にある (事後) 確率を計算せよ。ただし、司会者は回答者が選んでいない扉で賞品が入っていないものからランダムに選んで開けるものとする。

(1) 司会者が B の扉を開けたとき。

(2) 司会者が B と E の扉を開けたとき。

例題 2.4 において、最初は C の扉に賞品がある確率が $P(C) = \dfrac{1}{3}$ ということから、司会者が B の扉を開けるという新たな情報が加わったことにより、C の扉に賞品がある確率は $P_S(C) = \dfrac{2}{3}$ となった。このように試行を行う前の判断の確率 $P(C)$ を**事前確率**, 試行を行った結果の条件の下での判断の確率 $P_S(C)$ を**事後確率**という。ベイズの定理は事前確率から事後確率を導く公式と考えられる。[*14]

次に変形 3 囚人問題を考える ([4] に詳しい)。これは更に直感と異なる結果となる。

例題 2.5 (変形 3 囚人問題) 3 人の囚人 A, B, C がいて、2 人が処刑され 1 人が釈放されることがわかっている。釈放される確率は、A, B, C それぞれが 1/4, 1/4, 1/2 であった。誰が釈放されるか知っている看守に対し、A が「B と C のうち少なくとも 1 人処刑されるのは確実なのだから、2 人のうち処刑される 1 人の名前を教えてくれても私の釈放についての情報を与えていることにはならないだろう。1 人を教えてくれないか。」と頼んだ。看守は A の言い分に納得して「B は処刑される。」と答えた。この答えを聞いたあと、A の釈放される確率はいくらになるか。

解答: 例題 2.3 と同じ記号を用いると、事前分布は $P(A) = P(B) = \dfrac{1}{4}$, $P(C) = \dfrac{1}{2}$ となる。
また、F で看守が「B は処刑される」と告げる事象をあらわすと、$P_A(F) = \dfrac{1}{2}$, $P_B(F) = 0$, $P_C(F) = 1$.
よって、求める確率は $P_F(A)$ であるから、ベイズの定理を用いて

$$P_F(A) = \frac{P(A)P_A(F)}{P(A)P_A(F) + P(B)P_B(F) + P(C)P_C(F)} = \frac{\frac{1}{4} \times \frac{1}{2}}{\frac{1}{4} \times \frac{1}{2} + \frac{1}{4} \times 0 + \frac{1}{2} \times 1} = \frac{1}{5}$$

となる。　□

[*13] [4] によると、2 つのドアの賞品のある確率は 1/2 ずつであると考えてしまう人が、ほとんで、更に、「確率が同じなら、最初に選んだほうを選び続けるほうがいい」と多くの人は考える。これはわざわざ変えてはずれるほうが、悔いが残るということのようである。実際に実験的検討がなされ「選ぶドアを変えない」という回答者が圧倒的に多くなるとあった。

[*14] 現実の問題において、事前確率をどのように設定するかはたいへん難しい問題である。また事前確率の概念そのものに設定者の主観が入り込む余地がある (主観主義) としての批判もある。

例題 2.3 では囚人 A が釈放される確率は 1/3 のままだから、「残った囚人は A と C だけで、もともとが釈放される確率の比は 1 : 2 だったから、1 を比例配分して 1/3 となる」と考えることも出来ると述べた。しかし、この場合では釈放される確率は 1/4 から 1/5 と減ってしまう。つまりこの推論は誤りだったことがわかる。

問 2.4 例題 2.5 で 3 人の囚人 A, B, C が釈放される事前確率がそれぞれが 1/4, 1/2, 1/4 であったとき、看守の答え「B は処刑される。」を聞いたあとの、A の釈放される確率はいくらになるか。また、事前確率が A, B, C それぞれが 1/2, 1/4, 1/4 であったときはどうか。

問 2.5 問 2.3 と同様に A, B, C, D, E の 5 つの扉のうち 1 つだけに賞品が入っていている場合を考える。ただし、扉 A, B, C, D, E に賞品が入っている事前確率は 1/6, 1/6, 1/6, 1/4, 1/4 であるとする。回答者が選んだ扉が A であり、次の (1), (2) のように司会者が扉を選んで開けたとする。このとき、賞品が A にある事後確率を計算せよ。ただし、司会者は回答者が選んでいない扉で賞品が入っていないものからランダムに選んで開けるものとする。

(1) 司会者が B の扉を開けたとき。

(2) 司会者が B と E の扉を開けたとき。

3 データの分析

2012 年度から数学 I でも記述統計を扱うようになりました。この機会に、記述統計の話題をいくつか扱ってみましょう[*15]。

3.1 1 次元データ

ここでは身長や数学の試験の得点などデータを構成する量が一つの数字で表されるものを考える。

変量 x の n 個のデータの値が $x_1, x_2, \cdots, x_n$ とする。

a. 中心的傾向をあらわすもの

- **平均値**　　$\overline{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$
- **中央値** (**メジアン**)　データを大きさの順に並び替えたものを $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n)}$ とする。このとき、

$$\text{中央値} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ が奇数のとき} \\ \dfrac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & n \text{ が偶数のとき} \end{cases}$$

とする。

例題 3.1 次のデータの平均値と中央値を求めよ。

(1) $42, 38, 40, 44, 52$　　(2) $42, 38, 40, 44, 52, 198$

解答: (1) 平均値: $\overline{x} = \dfrac{42 + 38 + 40 + 44 + 52}{5} = 43.2$

中央値: データを大きさの順に並べると $38 < 40 < 42 < 44 < 52$ となるので、42.

(2) 平均値: $\overline{x} = \dfrac{42 + 38 + 40 + 44 + 52 + 198}{6} = 69$

中央値: データを大きさの順に並べると $38 < 40 < 42 < 44 < 52 < 198$ となるので、$\dfrac{1}{2}(42 + 44) = 43.$

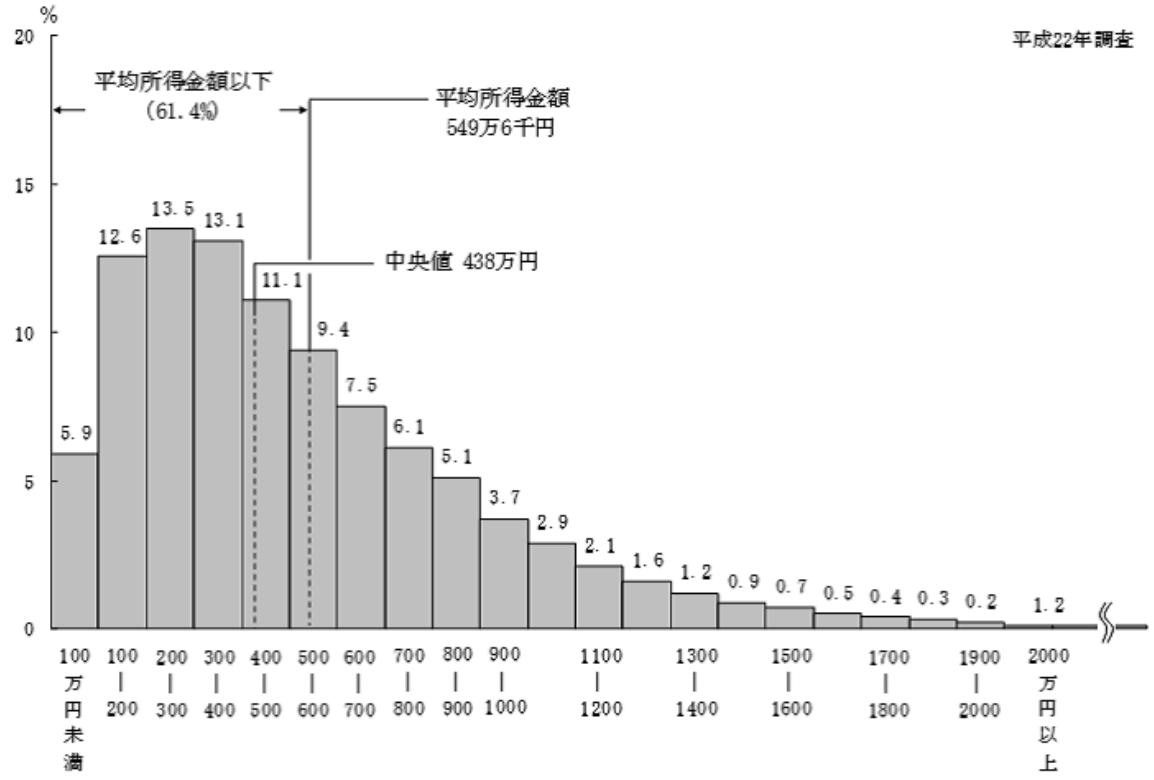
□

[*15] 数学 I で学ぶ記述統計学に対し、新課程の数学 B で学ぶ統計的な推測 (標本から母集団の特性値について推定や検定を行う) を推測統計学という。その歴史的な発展や数学と統計の違い、またどのような分野で応用されているかは [8] に簡潔にまとめられている。講師の専門は数学であって、統計ではないので受講者の皆さんのほうが詳しいかもしれません。間違いや勘違いなどがあればお教え願えればと思います。この部分は数学 I の教科書 (数研出版, 東京書籍のもの) のほか、文献表の [8] や [1] を参照しています。

注意 この例で、(1) から (2) へはデータを一つ増やしただけである。これによって (1) と (2) では平均値が大きく変わってしまった。一方、中央値はあまり影響を受けていない (安定している)。

このように、平均値は他のデータからかけ離れた値をもつ「はずれ値」の影響を受けやすいが、中央値はそうでないことがわかる。しかし中央値を求めるためにはデータを大きさの順に並べかえる必要があり、データが多い場合は、その計算は大変である[*16]。また平均値は数学的にいろいろよい性質をもっており、通常は平均値を用いることが多い。

平均値と中央値のどちらが日常用いる「平均」に近いか見るために、厚生労働省による平成 22 年国民生活基礎調査による所得金額階級別にみた世帯数のヒストグラムを見てみよう。[*17]



元データから平均値 549.6 万円であり、中央値が 438 万円であることがわかっている。また、このヒストグラムから最頻値 (度数が一番高い階級) は 200–300 万円であることがわかる。このように、平均値、中央値、最頻値は同じ階級にあるとは限らない。

もう少し極端な例として、平成 22 年度の二人以上世帯調査における金融資産保有額の分布を見てみよう[*18]。

保有額 (万円)	非保有	1-99	100-199	200-299	300-399	400-499	500-699
相対度数 (%)	22.3 (24.2)	5.5 (6.0)	5.7 (6.2)	4.4 (4.8)	4.3 (4.7)	3.6 (3.9)	7.3 (7.9)

700-999	1000-1499	1500-1999	2000-2999	3000–	無回答	計
7.1 (7.7)	9.7 (10.5)	5.1 (5.5)	7.0 (7.6)	10.1 (11.0)	7.8	99.9 (100.0)

この場合も平均値 1169 万円であり、中央値が 500 万円であることがわかっている。また、上の度数分布表から最頻値は非保有の階級となる。

[*16] 最近では、表計算ソフトを用いてデータを大きさの順に並べかえることで中央値は容易に求めることができる。

[*17] http://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa10/2-2.html
この分布の様子は異様に思えるかもしれないが、所得の分布はこのような形状を取ることがよく知られている。

[*18] 家計の金融行動に関する世論調査による。http://www.shiruporuto.jp/finance/chosa/yoron2010fut/index.html
相対度数の ( ) 内の数字は無回答を除いて計算した相対度数である。

これらの 3 種類の代表値を、どのように使い分けるかについては、明確な基準はない。多くの場合には、簡便さも含め平均値を用いればよいが、給与や貯蓄額のようにハッキリした上限がないようなデータの代表値として平均値を用いる場合には、注意が必要であろう。また、はずれ値が出易いデータの場合には、安定性の観点から、中央値を用いるのがよいであろう。最頻値を代表値として用いることは、現実にはめったにない ($cf.$ [8])。

b. 散らばりをあらわすもの

変量 x の n 個のデータの値は $x_1, x_2, \cdots, x_n$ であり、データを大きさの順に並び替えたものが $x_{(1)} \leqq x_{(2)} \leqq \cdots \leqq x_{(n)}$ であった。

- **範囲** $x_{(n)} - x_{(1)}$ (データの最大値と最小値の差)

- **四分位数**

$n = 2m$ が偶数のとき、

$x_{(1)}, x_{(2)}, \cdots, x_{(m)}$ を**下位のデータ**, $x_{(m+1)}, x_{(m+2)}, \cdots, x_{(2m)}$ を**上位のデータ**と、

$n = 2m+1$ が奇数のとき、

$x_{(1)}, x_{(2)}, \cdots, x_{(m)}$ を**下位のデータ**, $x_{(m+2)}, x_{(m+3)}, \cdots, x_{(2m+1)}$ を**上位のデータ** という。

$n = 2m+1$ のときは上位下位ともに m 個のデータがあることに注意する。このとき、

第 1 分位数 Q_1 は 下位のデータの中央値　　**第 3 分位数** Q_3 は 上位のデータの中央値

と定める。なお、第 2 分位数 Q_2 はデータ全体の中央値 (通常の中央値) とする。

これを用いて、　**四分位範囲** を $Q_3 - Q_1$,　**四分位偏差** を $\dfrac{1}{2}(Q_3 - Q_1)$　と定める。

また、データの 最小値・第 1 分位数・中央値・第 3 分位数・最大値 を図にしたのが箱ひげ図である。

例題 3.2 次のデータの第 1 分位数 Q_1 と第 3 分位数 Q_3 を求めよ。

(1) 65, 70, 47, 78, 89, 65, 89, 95, 59,　　(2) 65, 70, 47, 78, 89, 65, 89, 95, 59, 73

解答: (1) データを順に並べると $47 < 59 < 65 = 65 < 70 < 78 < 89 = 89 < 95$ であるから、

$Q_1 = \dfrac{59+65}{2} = 62$, $Q_3 = \dfrac{89+89}{2} = 89$ となる。

(2) 演習問題とする。(答 $Q_1 = 65$, $Q_3 = 89$)　　□

- **分散, 標準偏差**

$$\textbf{分散}\quad s^2 = \frac{1}{n}\left\{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\right\}$$

$$\textbf{標準偏差}\quad s = \sqrt{\frac{1}{n}\left\{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \cdots + (x_n-\overline{x})^2\right\}}$$

変量 x の測定単位が例えば「点」のとき、分散の単位は「点^2」となってしまう。一方、標準偏差は変量と同じ測定単位となる。また、分散が 0 となるのはすべてのデータの値が一致するときに限ることに注意する。

定理 3.1 $s^2 = \overline{x^2} - \overline{x}^2$. ただし、$\overline{x^2}$ は変量 x^2 のデータ ${x_1}^2, {x_2}^2, \cdots, {x_n}^2$ の平均値を表す。

証明:
$$s^2 = \frac{1}{n}\sum_{k=1}^{n}({x_k}^2 - 2\overline{x}x_k + \overline{x}^2) = \frac{1}{n}\sum_{k=1}^{n}{x_k}^2 - 2\overline{x}\frac{1}{n}\sum_{k=1}^{n}x_k + \frac{1}{n}\sum_{k=1}^{n}\overline{x}^2$$
$$= \overline{x^2} - 2\overline{x}\cdot\overline{x} + \frac{1}{n}\cdot n\overline{x} = \overline{x^2} - \overline{x}^2 \qquad \square$$

注意 分散や標準偏差は数学的にいろいろよい性質をもっている。特に、データ数が十分多いとき、そのヒストグラムの形状が適当なスケーリングのもとで標準正規分布の密度関数で近似できることが知られている (中心極限定理)。この性質は、偏差値など身近なところで用いられている。

偏差値の求め方: 平均値が $\overline{x}$, 標準偏差が s のとき、x_1 点だった人の偏差値は

$$50 + 10 \times \frac{x_1 - \overline{x}}{s}$$

となる。逆に、偏差値が a であれば、$z=(a-50)/10$ の値を正規分布表と比較することで、自分がおおよそ全体で上位何 % の位置にいるか判断できる。(正規分布表は統計的な推測を扱っている教科書、現行では高校数学 C の教科書にある。)

問 3.1 変量 x のデータ $x_1, x_2, \cdots, x_m$ と変量 y のデータ $y_1, y_2, \cdots, y_n$ をあわせた $m+n$ 個のデータを変量 z とする。変量 x, y, z の平均値を $\overline{x}$, $\overline{y}$, $\overline{z}$ と、分散を ${s_x}^2$, ${s_y}^2$, ${s_z}^2$ と表すとき、次を示せ。

(1) $\overline{z}=\dfrac{m}{m+n}\overline{x}+\dfrac{n}{m+n}\overline{y}$　　(2) ${s_z}^2=\dfrac{m}{m+n}{s_x}^2+\dfrac{n}{m+n}{s_y}^2+\dfrac{mn}{(m+n)^2}(\overline{x}-\overline{y})^2$

次に度数分布表に基づいた平均値と分散を定義しよう。

定義 3.2 変量 x のデータ n 個が次のような度数分布表にまとめられたとする。

階級値	x_1	x_2	$\cdots\cdots$	x_r	計
度数	f_1	f_2	$\cdots\cdots$	f_r	n

このとき、各 k に対して x_k の値のデータが f_k 個あるとみなして、平均値 $\overline{x}$ と分散 s^2 を

$$\overline{x}=\frac{1}{n}\sum_{k=1}^{r}x_kf_k,\quad s^2=\frac{1}{n}\sum_{k=1}^{r}(x_k-\overline{x})^2f_k$$

と定める。また、分散の非負の平方根を標準偏差という。

問 3.2 変量 x のデータ n 個が定義 3.2 の表の場合に、$s^2=\overline{x^2}-\overline{x}^2$ となることを示せ。

問 3.3 a と b を定数とする。変量 x のデータ n 個が定義 3.2 の度数分布表のように与えられ、変量 y の度数分布表をその階級値は $y_k=ax_k+b$, $k=1,2,\cdots,r$, とし度数は変量 x の度数と同じとして定めるとき、変量 x, y の平均値 $\overline{x}$, $\overline{y}$ と、分散 ${s_x}^2$, ${s_y}^2$ について次の関係式が成り立つことを示せ。

(1) $\overline{y}=a\overline{x}+b$　　(2) ${s_y}^2=a^2{s_x}^2$

問 3.4 次の数値は、あるクラスの 50 人の学生についての中間テストの点数である。

65	70	54	78	89	65	89	95	59	73
28	93	100	68	88	26	95	73	66	56
64	66	65	87	50	54	69	71	89	61
37	91	73	62	32	39	46	89	45	51
56	80	65	78	75	70	95	61	45	85

これを度数分布表にまとめると次のようになった。

階級値	25	35	45	55	65	75	85	95	計
度数	2	3	4	6	14	8	7	6	50

ただし、21 点以上 30 点以下の階級値を 25 とし、他も同様に 35, 45,$\cdots$, とした。例えば、階級値 55 点に入る点の範囲は 51 点以上 60 点以下である。このとき、次の問いに答えよ。

(1)　この度数分布表を用いて平均 $\overline{x}$ と分散 ${s_x}^2$ を計算せよ。

(2)　このデータの第 1 四分位数 Q_1 を求めよ。ヒント: まずどの階級にあるかを考えよ。

(3)　このデータの中央値 m を求めよ。

3.2　2 次元データ

クラス 40 人の数学と英語の点になんらかの関係があるかどうかなど、2 つの変量をもつ場合を考える。
ここでは、2 つ変量 x, y のデータが n 個の x, y の値の組として、次のように与えられているとする。

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_n, y_n)$$

- **散布図** 上記の x, y の値の組を座標とする点を平面上にとったもの。

- **共分散**, **相関係数**

$x_1, x_2, \cdots, x_n$ と $y_1, y_2, \cdots, y_n$ の平均値をそれぞれ $\overline{x}, \overline{y}$ で標準偏差を s_x, s_y で表す。
このとき、x と y の共分散 s_{xy} を

$$s_{xy} = \frac{1}{n}\left\{(x_1 - \overline{x})(y_1 - \overline{y}) + (x_2 - \overline{x})(y_2 - \overline{y}) + \cdots + (x_n - \overline{x})(y_n - \overline{y})\right\}$$

と定め、x と y の相関係数 r を

$$r = \frac{s_{xy}}{s_x s_y}$$

と定める。ただし、$s_x > 0$ かつ $s_y > 0$ のときのみ相関係数は考えるものとする。

定理 3.3 (1)　相関係数 r について、$-1 \leqq r \leqq 1$ となる。
(2)　$r = 1$ となるのは、n 個のデータが正の傾きをもつ直線上に集中しているとき、
(3)　$r = -1$ となるのは、n 個のデータが負の傾きをもつ直線上に集中しているときに限る。

証明: コーシー・シュワルツの不等式: $(a_1b_1 + a_2b_2 + \cdots + a_nb_n)^2 \leqq ({a_1}^2 + {a_2}^2 + \cdots + {a_n}^2)({b_1}^2 + {b_2}^2 + \cdots + {b_n}^2)$ で $a_k = x_k - \overline{x}$, $b_k = y_k - \overline{y}$ を代入することで (1) はすぐにわかる。また、この不等式で等号が成立するための条件は、ある定数 c があってすべての k に対して $b_k = ca_k$ となることであるから、[*19]
　$c > 0$ のとき $r = 1$ であり $y_k - \overline{y} = c(x_k - \overline{x})$ となること、
　$c < 0$ のとき $r = -1$ であり $y_k - \overline{y} = c(x_k - \overline{x})$ となること
から (2), (3) は従う。　　□

問 3.5 $s_{xy} = \overline{xy} - \overline{x}\,\overline{y}$ を示せ。 ただし、$\overline{xy}$ は変量 xy のデータ $x_1y_1, x_2y_2, \cdots, x_ny_n$ の平均値を表す。

問 3.6 3 つのデータ $(-1, 1), (1, -1), (a, a)$ について相関係数 r を求め、それが a が変化するとき $-1 \leqq r < 1$ のとなることを確かめよ。また、$r = -1$, $r = 0$ となる a の値をそれぞれ求め、そのとき 3 点 $(-1, 1)$, $(1, -1)$, (a, a) が散布図 (座標平面) においてどのような位置関係になるか調べよ。

- **正の相関**, **負の相関**　　変量 x と y の間に、

一方の値が増加すると他方も増加する傾向があるとき、2 つの変量 x, y の間に正の相関があるという。
一方の値が増加すると他方は減少する傾向があるとき、2 つの変量 x, y の間に負の相関があるという。
正の相関も負の相関もみられないとき、相関がないという。
　おおよその目安となる基準は以下のようである ($cf.$ [8], p.60)。
(i)　相関係数 $= 0.7 \sim 1.0$ (または $= -0.7 \sim -1.0$): かなり強い正の相関 (負の相関) がある。
(ii)　相関係数 $= 0.4 \sim 0.7$ (または $= -0.4 \sim -0.7$): 中程度の正の相関 (負の相関) がある。
(iii)　相関係数 $= 0.2 \sim 0.4$ (または $= -0.2 \sim -0.4$): 弱い正の相関 (負の相関) がある。
(iv)　相関係数 $= -0.2 \sim 0.2$: ほとんど相関がない。

[*19] コーシー・シュワルツの不等式とその等号成立のための条件は、$\sum_{k=1}^{n}(a_kt + b_k)^2$ を t について平方完成することで証明できる。

注意 3.1 (1) 二つの変量 x, y に強い正の相関があっても、実際にその二つの間に因果関係があるとは限らない。例えば、「サラリーマンの年収と血圧を調べると正の相関がある」について (実際に調べるとかなり強い正の相関があるらしい)、これは年収と血圧がともに年齢とともに上昇する傾向があることによっている。このように実際に因果関係があるかは相関係数だけではなく他の要因も調べなければならない。

社会科学の分野では、ポール・ラザースフェルドが 1959 年に、次の 3 つの基準を挙げた。

1. 原因は結果に先行する。
2. 2 つの変量は経験的に相関している。
3. その相関は、別の第三の変数によって説明されない。

自然科学の分野では、米国公衆衛生局長諮問委員会が 1964 年に喫煙と肺がんの因果関係を諮問されたときの判断基準がある。詳しくはいくつかの用語を導入しなければいないので省略する ($cf.$ [1], p.102)。

(2) 一般に、データをまとめ上げてしまうと、部分的に存在する関係等が良く見えなくなってしまう場合が多い。例えば、理系科目が得意の生徒だけが集まったクラスと文系科目が得意の生徒だけが集まったクラスがあったとしよう。それぞれのクラスでは、国語と数学の試験の点数には正の相関があったとしても、二つのクラス全体のデータから国語と数学の試験の点数の間の相関係数を計算すると負になることもあり得る。

このように、部分的な関係も把握できるように、属性やデータの値などによって、データをいくつかの部分集合に分けて (層別にして) 解析を行うことが重要となる。

一方、一部のデータのみに基づいて計算された相関係数は、実際の相関係数より小さくなりやすいことも注意する必要がある。例えば、大学入試の成績 x と入学後の成績 y の相関関係を考えてみよう。これがある正の相関をもつと想定することは自然である。しかし、このデータを調べることは不可能である。なぜなら、不合格者は大学に入学できないから、入学後の成績のデータが得られない。特に、競争倍率が高く合格者の割合が少ない場合など、合格者のみのデータによって計算される x と y の相関係数は低くなり、場合によっては負の相関となってしまう場合も珍しくない。

このようなある値より小さい (または大きい) 値を持つデータしか存在しない場合は、それは「切断データ」とよばれ、少なくとも一方が切断されている場合には、計算された相関係数の値は一般に低くなる ($cf.$ [8])。

参考文献

[1] 青木 繁伸: 統計数字を読み解くセンス 当確はなぜすぐにわかるのか?, 化学同人, 2009.

[2] 服部 哲弥: 統計と確率の基礎, 学術図書出版社, 2006.

[3] 伊庭 幸人: ベイズ統計と統計物理, 岩波講座 物理の世界, 2003.

[4] 市川 伸一: 確率の理解を探る 3 囚人問題とその周辺, 認知科学モノグラフ, 共立出版, 1998.

[5] 小島寛之: 確率的発想法 数学を日常に活かす, NHK ブックス, 2004.

[6] 国沢 清典 編: 確率統計演習 2 統計, 培風館, 1966.

[7] デイヴィッド サルツブルグ (竹内惠行, 熊谷悦生 訳): 統計学を拓いた異才たち, 日経ビジネス人文庫, 2010.

[8] 田栗 正章, 藤越 康祝, 柳井 晴夫, C.R. ラオ: やさしい統計入門, 講談社ブルーバックス, 2007.

[9] 谷岡 一郎: 確率・統計であばくギャンブルのからくり, 講談社ブルーバックス, 2001.

[10] 渡部 洋: ベイズ統計学入門, 福村出版, 1999.

問の解答

1.1 (1) $1\cdot\dfrac{11}{12}\cdot\dfrac{10}{12}\cdot\dfrac{9}{12}\cdot\dfrac{8}{12}=\dfrac{55}{144}(\fallingdotseq 0.3819)$.

(2) 2 月生まれの人を含まないとき $11p\cdot 10p\cdot 9p\cdot 8p\cdot 7p$, 含むときその順序を考慮して、$5(1-11p)\cdot 11p\cdot 10p\cdot 9p\cdot 8p$. これを加えて、$f(p)=7920p^4(5-48p)$. $f(p)$ の最大値については、$f(p)$ を微分して $f'(p)=0$ を解き増減表を書けば $p=\dfrac{1}{12}$ のとき最大となることがわかる。

1.2 (1) $\dfrac{6}{6}\cdot\dfrac{5}{6}\cdot\dfrac{4}{6}=\dfrac{5}{9}(=0.55555\cdots)$.

(2) 1, 6 の目を含まないとき、1 回だけ含むとき、2 回階含むときと分けて考え、1 か 6 の目の出る順番を考慮すると $\dfrac{4\cdot5}{28}\cdot\dfrac{3\cdot5}{28}\cdot\dfrac{2\cdot5}{28}+{}_3\mathrm{C}_1\cdot\dfrac{2}{7}\cdot\dfrac{4\cdot5}{28}\cdot\dfrac{3\cdot5}{28}+{}_3\mathrm{C}_2\cdot\dfrac{2}{7}\cdot\dfrac{1}{7}\cdot\dfrac{4\cdot5}{28}=\dfrac{5\cdot303}{2^3\cdot7^3}=\dfrac{1515}{2744}$ $(=0.5521137\cdots)$. ((1) より小さくなる。)

1.3 9 試合やれば必ず勝負がつくことに注意し、例題 1.2 と同様の表を作れば、(1) $\dfrac{7}{8}$ (2) $\dfrac{11}{16}$ となる。

(1)

現在までの勝敗		7	8	9	勝者
(WWWWLL)	→	W	W	W	A 氏
	→	W	W	L	A 氏
	→	W	L	W	A 氏
	→	W	L	L	A 氏

現在までの勝敗		7	8	9	勝者
(WWWWLL)	→	L	W	W	A 氏
	→	L	W	L	A 氏
	→	L	L	W	A 氏
	→	L	L	L	B 氏

(2)

現在まで		6	7	8	9	勝者
(WWWLL)	→	W	W	W	W	A 氏
	→	W	W	W	L	A 氏
	→	W	W	L	W	A 氏
	→	W	W	L	L	A 氏
	→	W	L	W	W	A 氏
	→	W	L	W	L	A 氏
	→	W	L	L	W	A 氏
	→	W	L	L	L	B 氏

現在まで		6	7	8	9	勝者
(WWWLL)	→	L	W	W	W	A 氏
	→	L	W	W	L	A 氏
	→	L	W	L	W	A 氏
	→	L	W	L	L	B 氏
	→	L	L	W	W	A 氏
	→	L	L	W	L	B 氏
	→	L	L	L	W	B 氏
	→	L	L	L	L	B 氏

1.4 例題 1.1 と同様に余事象を考えればよい。

(1) 4 回とも 6 の目が出ない確率は $\left(\dfrac{5}{6}\right)^4$. よって、勝つ確率は $1-\left(\dfrac{5}{6}\right)^4\fallingdotseq 0.5177$ となり、勝てることが多いと予想される。

(2) 二つとも 6 の目が出ないことが 24 回続く確率は $\left(\dfrac{35}{36}\right)^{24}$. よって、勝つ確率は $1-\left(\dfrac{35}{36}\right)^{24}\fallingdotseq 0.4914$ となり、負けることが多いと予想される。また、$\left(\dfrac{35}{36}\right)^{25}\fallingdotseq 0.4945$ なので、$1-\left(\dfrac{35}{36}\right)^{24}<0.5<1-\left(\dfrac{35}{36}\right)^{25}$ となり、25 回以上投げることにすれば勝てる確率が 0.5 より大きくなる。

2.1 A, B でそれぞれ A の工場, B の工場の製品である事象とし、F で不良品である事象とする。
仮定より $P_A(F)=0.03$, $P_B(F)=0.04$, $P(A)=\frac{4}{9}$, $P(B)=\frac{5}{9}$ であり、求める確率は $P_F(A)$ であるから、

$$P_F(A)=\frac{P(A\cap F)}{P(F)}=\frac{P(A)P_A(F)}{P(A)P_A(F)+P(B)P_B(F)}=\frac{4\cdot3}{4\cdot3+5\cdot4}=\frac{3}{8}$$

2.2 A_1, A_2, A_3 でそれぞれ機械 M_1, M_2, M_3 の製品である事象とし、F で不良品である事象とする。
仮定より $P(A_1) = 0.6$, $P(A_2) = 0.3$, $P(A_3) = 0.1$, $P_{A_1}(F) = 0.02$, $P_{A_2}(F) = 0.03$, $P_{A_3}(F) = 0.06$ であり、求める確率は $P_F(A_3)$ であるから、

$$P_F(A_3) = \frac{P(A_3)P_{A_3}(F)}{P(A_1)P_{A_1}(F) + P(A_2)P_{A_2}(F) + P(A_3)P_{A_3}(F)} = \frac{1 \cdot 6}{6 \cdot 2 + 3 \cdot 3 + 1 \cdot 6} = \frac{2}{9}$$

2.3 A, B, C, D, E でそれぞれ A, B, C, D, E の扉に賞品があるという事象とするとき、$P(A) = P(B) = P(C) = P(D) = P(E) = \frac{1}{5}$.

(1) 司会者が B の扉を開けるという事象を S_1 とすると、例題 2.4 と同様に、$P_A(S_1) = \frac{1}{4}$, $P_B(S_1) = 0$, $P_C(S_1) = P_D(S_1) = P_E(S_1) = \frac{1}{3}$. よって、

$$P_{S_1}(C) = \frac{P(C)P_C(S_1)}{P(A)P_A(S_1) + P(B)P_B(S_1) + P(C)P_C(S_1) + P(D)P_D(S_1) + P(E)P_E(S_1)} = \frac{4}{15}$$

(2) 司会者が B, E の扉を開けるという事象を S_2 とすると、(1) と同様に、$P_A(S_2) = \dfrac{1}{{}_4\mathrm{C}_2} = \dfrac{1}{6}$, $P_B(S_2) = P_E(S_2) = 0$, $P_C(S_2) = P_D(S_2) = \dfrac{1}{{}_3\mathrm{C}_2} = \dfrac{1}{3}$. よって、$P_{S_2}(C) = \dfrac{2}{5}$.

2.4 例題 2.5 と同じ記号を用いると、$P_A(F) = \frac{1}{2}$, $P_B(F) = 0$, $P_C(F) = 1$. よって、事前確率が A, B, C それぞれが 1/4, 1/2, 1/4 であったとき、$P(A) = P(C) = \frac{1}{4}$, $P(B) = \frac{1}{2}$ より、$P_F(A) = \frac{1}{3}$.
また、1/2, 1/4, 1/4 のとき、$P_F(A) = \dfrac{1}{2}$ となる。

2.5 問 2.3 の解答と同じ記号を用いると、$P(A) = P(B) = P(C) = \frac{1}{6}$, $P(D) = P(E) = \frac{1}{4}$. これより、問 2.3 と全く同様に (1) $P_{S_1}(A) = \dfrac{3}{19}$, (2) $P_{S_2}(A) = \dfrac{1}{6}$ となる。

3.1 (1) $(m+n)\overline{z} = m\overline{x} + n\overline{y}$ より明らか。

(2) $(m+n){s_z}^2 = (m+n)\overline{z^2} - (m+n)\overline{z}^2 = m\overline{x^2} + n\overline{y^2} - \dfrac{1}{m+n}(m\overline{x} + n\overline{y})^2$
$= m(\overline{x^2} - \overline{x}^2) + n(\overline{y^2} - \overline{y}^2) + \left(m - \dfrac{m^2}{m+n}\right)\overline{x}^2 + \left(n - \dfrac{n^2}{m+n}\right)\overline{y}^2 - \dfrac{2mn}{m+n}\overline{x} \cdot \overline{y}$
$= m{s_x}^2 + n{s_y}^2 + \dfrac{mn}{m+n}(\overline{x} - \overline{y})^2$ となり主張を得る。

3.2 $s^2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{r}({x_k}^2 - 2\overline{x}x_k + \overline{x}^2)f_k = \frac{1}{n}\sum_{k=1}^{r}{x_k}^2 f_k - 2\overline{x} \cdot \frac{1}{n}\sum_{k=1}^{r}x_k f_k + \overline{x}^2\frac{1}{n}\sum_{k=1}^{r}f_k$
$= \overline{x^2} - 2\overline{x} \cdot \overline{x} + \overline{x}^2 = \overline{x^2} - \overline{x}^2$.

3.3 (1) $\overline{y} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{r}y_k f_k = \frac{1}{n}\sum_{k=1}^{r}(ax_k + b)f_k = a\frac{1}{n}\sum_{k=1}^{r}x_k f_k + b\frac{1}{n}\sum_{k=1}^{r}f_k = a\overline{x} + b$.

(2) ${s_y}^2 = \dfrac{1}{n}\displaystyle\sum_{k=1}^{r}(y_k - \overline{y})^2 f_k = \frac{1}{n}\sum_{k=1}^{r}\{ax_k + b - (a\overline{x} + b)\}^2 f_k = \frac{1}{n}\sum_{k=1}^{r}a^2(x_k - \overline{x})^2 f_k = a^2{s_x}^2$.

3.4 (1) 階級値 x_k に対して $y_k = \dfrac{x_k - 5}{10}$ とすると、$\overline{x} = 10\overline{y} + 5$, ${s_x}^2 = 10^2{s_y}^2$ となることに注意する。
$\overline{y} = \dfrac{1}{50}(2 \cdot 2 + 3 \cdot 3 + 4 \cdot 4 + 5 \cdot 6 + 6 \cdot 14 + 7 \cdot 8 + 8 \cdot 7 + 9 \cdot 6) = 6.18$ より $\overline{x} = 66.8$.
$\overline{y^2} = \dfrac{1}{50}(2^2 \cdot 2 + 3^2 \cdot 3 + \cdots + 9^2 \cdot 6) = 41.58$ より ${s_y}^2 = \overline{y^2} - \overline{y}^2 = 3.3876$. よって、${s_x}^2 = 338.76$.

(2) データ数が 50 だから下位のデータは 25 個であるので、Q_1 は小さいほうから 13 番目のデータとなる。よって、階級値 55 の階級に属しており、その小さいほうから 4 番目のデータとなる。55 の階級値に属するデータを抜き出すと 54, 59, 56, 54, 51, 56 であるから、これを並べかえ $51 < 54 = 54 < 56 = 56 < 59$ となるので、$Q_1 = 56$.

(3) 小さいほうから 25 番目と 26 番目のデータの平均値なので、階級値 65 の階級に属しており、その大きいほうから 4 番目と 5 番目のデータとなる。55 の階級値に属するデータを抜き出すと 65, 70, 65, 68, 66, 64, 66, 65, 69, 61, 62, 65, 70, 61 であるから、これを並べかえて $m = \dfrac{66+68}{2} = 67$.

3.5 $s_{xy} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_k y_k - \overline{x}y_k - \overline{y}x_k + \overline{x}\,\overline{y}) = \dfrac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{x}\dfrac{1}{n}\sum_{k=1}^{n} y_k - \overline{y}\dfrac{1}{n}\sum_{k=1}^{n} x_k + \dfrac{1}{n}\sum_{k=1}^{n}\overline{x}\,\overline{y}$
$= \overline{xy} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}$.

3.6 $\overline{x} = \overline{y} = \dfrac{a}{3}$, $\overline{x^2} = \overline{y^2} = \dfrac{a^2+2}{3}$, $\overline{xy} = \dfrac{a^2-2}{3}$ より、$s_x = s_y = \dfrac{\sqrt{2a^2+6}}{3}$, $s_{xy} = \dfrac{2a^2-6}{9}$.
よって $r = \dfrac{a^2-3}{a^2+3}$. これより $r < 1$ となる。また $r+1 = \dfrac{6}{a^2+3}$ より $r > -1$ となる。
$r = -1$ とすると $a = 0$. このとき、この 3 点は直線 $y = -x$ 上にある。
$r = 0$ とすると $a = \pm\sqrt{3}$. $a = \sqrt{3}$ のとき $(\sqrt{3}, \sqrt{3})$ から $(-1, 1)$, $(1, 1)$ までの距離はともに $\sqrt{(\sqrt{3}+1)^2 + (\sqrt{3}-1)^2} = 2\sqrt{2}$ となり、$(-1, 1)$ から $(1, 1)$ までの距離と一致するので、3 点は正三角形をなす。$a = -\sqrt{3}$ のときも同様である。