確率統計の話題から －高校数学 I, A に関連した話題を中心に－

杉浦　誠

平成 25 年 8 月 24 日

1　確率を計算しよう

この節ではいくつかの確率論の起源となった問題について、その確率を具体的に計算してみましょう。

例題 1.1 (ド・メレの 3 個のサイコロ) イタリアの貴族ド・メレは 3 個のサイコロを投げて目の総和は 9 になる場合よりも 10 になる場合のほうが多いことを経験的に気づき不思議に思ってガリレイに問うたと言われている。これはどうしてか。

疑問の根拠は、3 個のサイコロの出る目の組み合わせがそれぞれ

9 のとき: (1,2,6), (1,3,5), (1,4,4), (2,2,5), (2,3,4), (3,3,3)

10 のとき: (1,3,6), (1,4,5), (2,2,6), (2,3,5), (2,4,4), (3,3,4)

の 6 通りであるためと考えられる。この推論は 3 個のサイコロが見た目で区別できないため、もし「すべての根元事象の起こる確率が等しい」なら、目の総和が 9 になる確率と 10 になる確率は等しいはずである。

これについて、ガリレイは「3 個のサイコロがたとえ見た目には区別ができなくても、別物である以上、思考上はこれを区別したうえで考えるべきである」と指摘し、

9 のとき: $6 \times 3 + 3 \times 2 + 1 = 25$ より 25 通り

10 のとき: $6 \times 3 + 3 \times 3 \quad = 27$ より 27 通り

となり、9 になる場合よりも 10 になる場合のほうが多いことを示した。(注意: 9 になる確率は 25/216, 10 になる確率は $27/216 = 1/8$ です。)　□

問 1.1 (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)　同時に 2 つのサイコロを 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。

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[*1]。

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。

[*1] 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。

これに対してパスカルは以下のような解答を与えた。

解答 1: ここでは両者の勝つ確率は等しいと仮定しよう[*2]。

このゲームの勝負の残りをしたとするとそ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	–	私
(WWL)	→	L	W	私
(WWL)	→	L	L	相手

両者の勝つ確率は等しいので、上記の起こる確率は順に $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ である。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。 □

これに対し、「勝負に勝つ確率は過去の実績を反映させるべきである」という判断から以下のような解答も考えられる ($cf.$ [4])。

解答 2: 次回に「私」が勝つ確率は過去の実績によって $\frac{2}{3}$ と推定し、それを仮定する[*3]。

解答 1 の勝敗表で、その起こる確率は順に $\frac{2}{3}, \frac{1}{3}\times\frac{2}{3}, \frac{1}{3}\times\frac{1}{3}$ である。つまり、「私」は確率 $\frac{2}{3}+\frac{1}{3}\times\frac{2}{3}=\frac{8}{9}$ で勝者のなったはずなので、正しい配分は「私」が $\frac{8}{9}$, 相手が $\frac{1}{9}$ の賭け金を取るべきである。 □

問 1.2 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 (a), (b) の場合に A 氏が受け取るべき賭け金の割合を、(1) 両者の勝つ確率が等しいとして、(2) 最尤法で過去の実績を反映させた確率として、決定せよ。

(a) その時点で A 氏が 4 勝 2 敗で勝っていた場合

(b) その時点で A 氏が 3 勝 2 敗で勝っていた場合

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[*4]。

定義 2.1 事象 A, B について、$P(A) > 0$ とする。このとき、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[*5]。

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

[*2] 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** ($cf.$ [5]) 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 (棄却されない)。

[*3] これには**最尤法**という統計学的裏付けがある。これは、もし「私」が勝つ確率を p とすると、n 回中 k 回勝つ確率は $f(p) = {}_nC_k p^k (1-p)^{n-k}$ であるが、この $f(p)$ を最大にする p の値 $p = k/n$ を、p の代表値 (**最尤推定値**という) とする方法である。

[*4] 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。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。

[*5] 通常は $P(B|A)$ と表します。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 $\overline{A}$ は用いず、A^c を用いることが通例です。(一般向けの書籍やインターネットなどを参照などする際はご注意ください。)

つまり、$P_A(B)$ とは「事象 A の中で、事象 $A \cap B$ の起こる確率」を表す。

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]) [*6]。

これを条件つき確率の記号で表すと次のようになる。

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

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

[*6] 実は、これはデータの個数がアンバランスであることに起因する。一般に、割合や平均を計算するもとになっているデータの個数がアンバランスな場合やグループ間で変数の関係が異なる場合には、様々なことが生じる可能性がある ($cf.$ [1])。

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、

$$P_A(C_i) = \frac{P(C_i)P_{C_i}(A)}{P(C_1)P_{C_1}(A) + P(C_2)P_{C_2}(A) + \cdots + P(C_n)P_{C_n}(A)} \quad \cdots\cdots\cdots\cdots\cdots\cdots \ ①$$

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。

$$P_A(B) = \frac{P(B)P_B(A)}{P(B)P_B(A) + P(\overline{B})P_{\overline{B}}(A)} \quad \cdots\cdots\cdots\cdots\cdots\cdots \ ②$$

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 (1) の右辺に代入することで主張を得る。 □

まず、ベイズの定理の応用例として、迷惑メールの防止フィルターを考える。

例題 2.2 迷惑メールの防止フィルターを、本文にある特定のワード (NG ワード) が含まれているか否かで判定する。私の主観では、私に届くメールのうち 60% は迷惑メール (Spam) で 40% は通常のメール (Ham) である。迷惑メールのうち 80% のメールは NG ワードを含んでおり、通常のメールのうちそれを含むものは 5% であった。このとき、NG ワードを含むメールが、迷惑メールである確率を求めよ。[*7]

解答: メールが NG ワードを含んでいるという事象を A, 迷惑メールであるという事象を S とする。

60% が迷惑メールなので、$P(S) = 0.6$, $P(\overline{S}) = 0.4$,

迷惑メールのうち 80% のメールは NG ワードを含んいるから、$P_S(A) = 0.8$,

通常のメールのうちそれを含むものは 5% であるから、$P_{\overline{S}}(A) = 0.05$.

したがって、求める確率 $P_A(S)$ はベイズの定理より、

$$P_A(S) = \frac{P(S)P_S(A)}{P(S)P_S(A) + P(\overline{S})P_{\overline{S}}(A)} = \frac{0.6 \times 0.8}{0.6 \times 0.8 + 0.4 \times 0.05} = \frac{48}{50} = 0.96. \qquad \square$$

試行を行う前の判断確率 $P(S)$ を**事前確率**, 試行を行った結果の条件の下での判断確率 $P_A(S)$ 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。

例題 2.3 自治体のがん検診で乳がんのマンモグラフィー検査を受けたところ「がんの疑い」と判定され、精密検査を受けることになった A さん。不安で家事も手につかない状態になりました。

では、A さんが「乳がんである可能性」はどのくらいでしょうか?

データによれば、乳がんでない女性が、間違って「がんの疑い」と判定されてしまう確率は 9% で、A さんの属する 40 歳台での罹病率は 0.3% です。[*8]

[*7] このとき、この確率が許容確率 (例えば $p^* = 0.8$) を超えれば迷惑メールと判断する。実際の迷惑メールフィルターでは、NG ワードを学習分類し、学習量が増えるとフィルタの分類精度が上昇するように設計されている。

[*8] 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。また [9] を参考にした。
番組では実数に置き換えて説明しています。具体的には、こうです。まず 1,000 人が検査を受けたものとします。この中に乳がんの人が 3 人おり、みな「乳がんの疑い」と判定されます。残りの 997 人は健康ですが、このうち $997 \times 0.09 \fallingdotseq 90$ 人が「乳がんの疑い」と判定されます。したがって、「乳がんの疑い」と判定された人計 93 人中で実際に乳がんであるのは 3 人だけなので、マンモグラフィーで陽性でも、乳がんである確率は $3 \div 93 \fallingdotseq 0.032$ となり約 3% であるとわかります。

A さんは「間違って『がんの疑い』と判定されてしまう確率は 9%」だから、「自分は 91% の確率でがん」だと思ったようです。冷静になって正しい確率を求めてみましょう。

解答: 実際にがんであるという事象を A, マンモグラフィー検査の結果が陽性であるという事象を E とする。

A さんの属する 40 歳台での罹病率は 0.3% より、$P(A) = 0.003$.

乳がんでない女性が、「がんの疑い」と判定されてしまう確率は 9% だから、$P_{\overline{A}}(E) = 0.09$.

問題文にはないが、ここでは乳がんの女性は必ず「がんの疑い」と判定されるとして、$P_A(E) = 1$.

したがって、求める確率 $P_E(A)$ はベイズの定理より、

$$P_E(A) = \frac{P(A)P_A(E)}{P(A)P_A(E) + P(\overline{A})P_{\overline{A}}(E)} = \frac{0.003 \times 1}{0.003 \times 1 + (1 - 0.003) \times 0.09} = \frac{3}{92.73} \fallingdotseq 0.032. \qquad \square$$

これより、マンモグラフィー検査で陽性でも、乳がんである確率はたった 3% ほどだとわかります。[*9]

問 2.1 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、実際に病原菌に感染している確率を求めよ。また、全体の 0.01% が感染している集団ではどうか調べよ。[*10]

問 2.2 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 2.3 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 2.4 (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、

[*9] マンモグラフィーをはじめとするがん検査が無意味というわけではない。実際、上記の例では検査前の事前確率 0.3% から、検査後には事後確率 3.2% と増加しており、精密検査はぜひ受けるべきであると私は思う。([9] によると、乳がん検診の効果は 40 歳台の女性についてははっきりしないが、50 歳以上については、死亡率を低下させていることがわかっているそうです。)

[*10] この問題から、事前確率の変化が事後確率に与える影響がわかる。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。

例えば、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。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、偶然に証拠と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。(この偏見は事前確率としてつい取り入れがちである。) 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])。

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。　□

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。

例題 2.5 (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。[*11]　□

問 2.4 例題 2.5 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 (等確率で) 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。

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

例題 2.6 (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 2.4 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、$P_A(F) = \dfrac{1}{2}$, $P_B(F) = 0$, $P_C(F) = 1$.

[*11] [3] 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。これはわざわざ変更してはずれるほうが、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。

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。　□

例題 2.4 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。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 2.5 例題 2.6 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 2.6 問 2.4 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

今回の新課程で重視されるようになった統計の分野から、特に記述統計の話題をいくつか扱ってみましょう[*12]。

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。

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

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) $42, 38, 40, 44, 52, 198$

[*12] 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。統計学の歴史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。[6] によると、ハーバード大学のメディカルスクールで使われている統計学の教科書の冒頭には「1903 年、H.G. ウェルズは将来、統計学的思考が読み書きと同じようによく社会人として必須の能力になる日がくると予言した」と書かれているそうです。また、同書には統計学の特徴を「どんな分野の議論においても、データを集めて分析することで最速で最善の答えを出すことができる」と述べていますし、教育や医学をはじめ様々な分野でどのように用いられているかがわかりやすく楽しく解説されています。実際、統計学は IT の発達により、データを用いるすべての分野に用いられるようになってきています。

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.

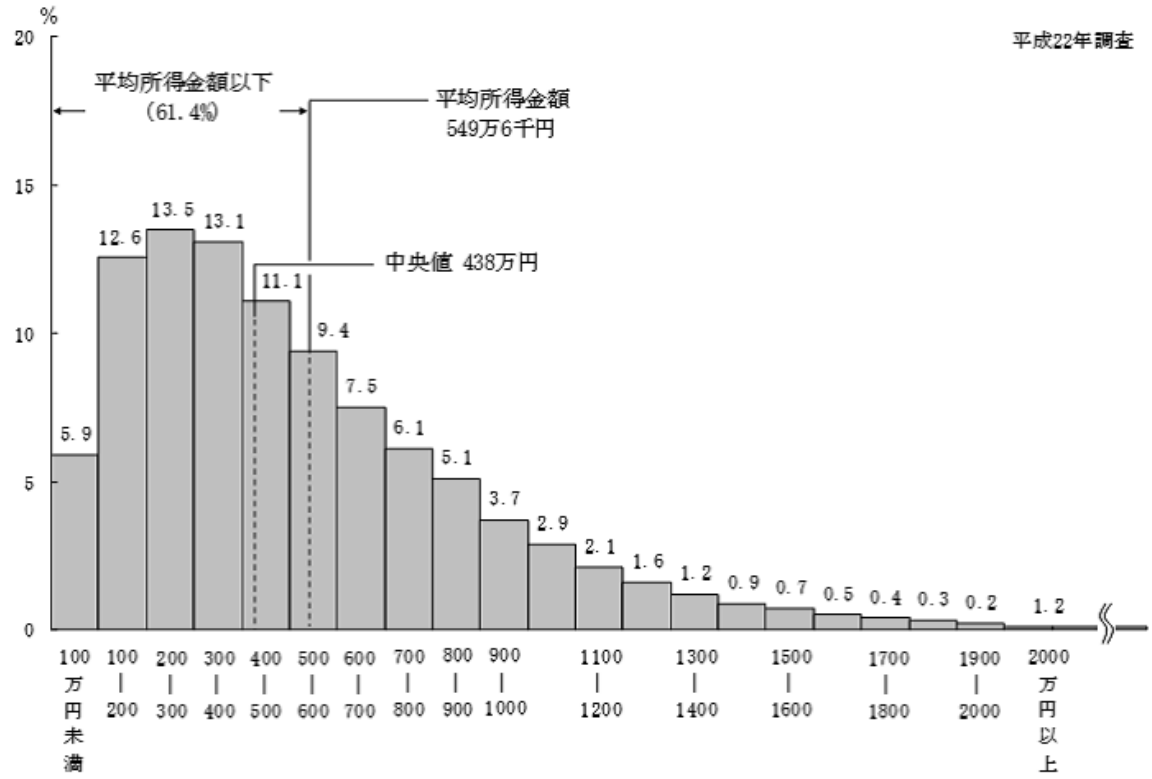
(2) 平均値: $\overline{x} = \dfrac{42+38+40+44+52+198}{6} = 69$

中央値: $38 < 40 < 42 < 44 < 52 < 198$ となるので、$\dfrac{1}{2}(42+44) = 43.$ □

注意 3.1 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[*13]、中央値はそうでない。しかし中央値を求めるためにはデータすべて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、それは大変な作業となる[*14]。一方、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。[*15]



元データから平均値は 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[*16]。

[*13] 通常、上側または下側四分位数から四分位範囲の 1.5 倍以上離れた値を「はずれ値」と定義する。

[*14] 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。

[*15] 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 (対数正規分布) を取ることがよく知られている。

[*16] 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。

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。

これらの 3 種類の代表値 (平均値、中央値、最頻値) をどのように使い分けるかについては、明確な規準はない。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、はずれ値が出やす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** (注意 3.2 も参照のこと)

$n = 2m$ 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 四分位数 Q_1 は 下位のデータの中央値　　**第 3 四分位数** Q_3 は 上位のデータの中央値

と定める。なお、第 2 四分位数 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。

例題 3.2 次のデータの第 1 四分位数 Q_1 と第 3 四分位数 Q_3 を求めよ。

(1) 65, 70, 47, 78, 92, 65, 89, 95, 59,　(2) 65, 70, 47, 78, 92, 67, 89, 95, 59, 73

解答: (1) データを小さいほうから並べると $47, 59, 65, 65, 70, 78, 89, 92, 95$ であるから、下位のデータは $47, 59, 65, 65$. よって、$Q_1 = \dfrac{59 + 65}{2} = 62$. 同様に上位のデータは $78, 89, 92, 95$ より $Q_3 = \dfrac{89 + 92}{2} = 90.5$.
(2) 順に並べると $47, 59, 65, 65, 70, 73, 78, 89, 92, 95$ であるから、$Q_1 = 65$, $Q_3 = 89$. 詳細は演習問題。 □

例題 3.3 次の数値は、ある授業の 30 人の学生についてのテストの点数である。

65	70	54	78	89	65
28	93	100	58	88	26
64	66	65	87	50	54
37	91	73	62	32	39
56	80	65	78	75	70

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	1	4	9	5	3	3	30

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。

このとき、このデータの第 3 四分位数 Q_3 を求めよ。ヒント: まずどの階級にあるかを考えよ。

解答: データ数が 30 だから上位のデータは 15 個であるので、Q_3 は大きいほうから 8 番目のデータとなる。よって、階級値 75 の階級に属しており、その大きいほうから 2 番目のデータとなる。この階級に属するデータを抜き出すと 78, 73, 80, 78, 75 であるから、これを順に並べると $73, 75, 78, 78, 80$ となるので、$Q_3 = 78$. □

問 3.1 例題 3.3 のデータの第 1 四分位数 Q_1 と中央値 m を求めよ。

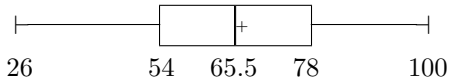
注意 3.2 四分位数の定義は複数ある。上記で定義したものは一般に Q_1 は下側ヒンジ、Q_3 は上側ヒンジと呼ばれている。例えば表計算ソフト Excel では、平面上の n 個の点 $(1, x_{(1)}), (2, x_{(2)}), \ldots, (n, x_{(n)})$ を順に折れ線で結んでできる関数 $y = f(t)$、即ち、$f(t) = \begin{cases} x_{(t)}, & t \text{ が自然数} \\ (\lceil t \rceil - t)x_{(\lfloor t \rfloor)} + (t - \lfloor t \rfloor)x_{(\lceil t \rceil)}, & \text{それ以外} \end{cases}$ を用い、$Q_q = f(1 + \frac{q}{4}(n-1)),\ q = 1, 3,$ と定めているようである。ここで、$\lceil t \rceil$ は t 以上の最小の整数、$\lfloor t \rfloor$ は t 以下の最大の整数を表す。この場合例題 3.2 の Q_3 は (1) $x_{(7)} = 89$, (2) $0.25x_{(7)} + 0.75x_{(8)} = 86.25$ となる。

- データの 最小値・第 1 四分位数・中央値・第 3 四分位数・最大値 を図にしたのが**箱ひげ図**である[*17]:

箱ひげ図は以下のように作成する。

1. データの第 1 四分位点 Q_1 と第 3 四分位点 Q_3 により、全データの半数が含まれる箱を描く。
2. 中央値 Q_2 を縦線で描く。
3. 平均値を「+」で描く（省略されることもある）。
4. 四分位範囲の 1.5 倍を箱の左右にとり、それを超えない内側のデータの最大値と最小値まで「ひげ」(左に「⊢——」, 右に「——⊣」) を引く。
5. 内境界点の外側の左右に四分位範囲の 1.5 倍の長さをとり（外境界）、その範囲にあるデータをはずれ値として「○」でプロットする（全データの最小値と最大値まで「ひげ」を引く方法ではこれは描かない）。
6. 外境界点の外側にあるデータを極値として「∗」でプロットする（同上）。

例題 3.3 のデータの場合、平均値が 65.6, 最小値 26, 最大値 100 であるから、右のようになる。ただし、平均値の数値は中央値に近いため記入しなかった。



- **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 3.3 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。

[*17] 「稲葉芳成: 箱ひげ図について」を参考にした。数学 I の教科書では、4 で内側すべてに「ひげ」を引き、5 の「はずれ値」と 6 の「極値」を省力している。(この方法も一般的ですが、はずれ値を加える図も見かけます。)

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。(正規分布表は数学 B の教科書などを参照。)

問 3.2 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。(最近の教科書では扱われていない。)

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

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

と定める。また、分散の非負の平方根を標準偏差という。

問 3.3 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 3.4 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) ${s_y}^2 = a^2{s_x}^2$

例題 3.4 例題 3.3 の度数分布表から、その平均と分散を求めよ。

解答: 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。
$\overline{y} = \dfrac{1}{30}(2\cdot 2 + 3\cdot 3 + 4\cdot 1 + 5\cdot 4 + 6\cdot 9 + 7\cdot 5 + 8\cdot 3 + 9\cdot 3) = 5.9$ より $\overline{x} = 59$.
$\overline{y^2} = \dfrac{1}{30}(2^2\cdot 2 + 3^2\cdot 3 + 4^2\cdot 1 + 5^2\cdot 4 + 6^2\cdot 9 + 7^2\cdot 5 + 8^2\cdot 3 + 9^2\cdot 3) = 38.5$ より ${s_y}^2 = \overline{y^2} - \overline{y}^2 = 3.69$.
よって、${s_x}^2 = 369$. □

問 3.5 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

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

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、[*18]

$c > 0$ 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。 $\square$

問 3.6 $s_{xy} = \overline{xy} - \overline{x}\,\overline{y}$ 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。

- **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。

[*18] 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。

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。

これは「$xy > 0 \Leftrightarrow x$ と y は同符号 (x, y の双方とも正、または双方とも負)」、「$xy < 0 \Leftrightarrow x$ と y は異符号」に注意する。平均値からのずれ (つまり偏差) を考慮し、n 個の平均値をとったものが共分散である。つまり、

・平均値からの偏差の符号が同じデータが多い $\longrightarrow$ 正の相関関係がある

・平均値からの偏差の符号が異なるデータが多い $\longrightarrow$ 負の相関関係がある　と考えられることによる。

($cf.$ 丸木和彦: 新学習指導要領における「数学Ⅰ データの分析」の指導方法の考察)

注意 3.4 (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。

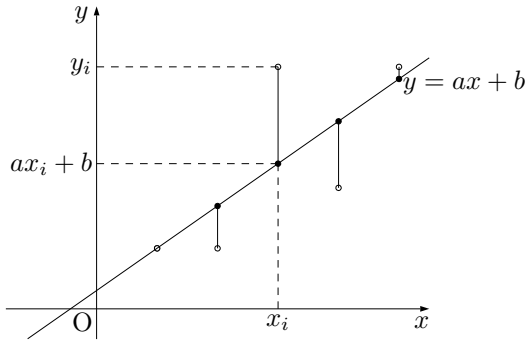
このように、部分的な関係も把握できるように、属性やデータの値などによって、データをいくつかの部分集合に分けて (層別にして) 解析を行うことが重要となる。

一方、一部のデータのみにもと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])。

● **回帰直線**　最後にこれも高校の教科書では扱われていませんが回帰直線を考えましょう。

2 次元データにある程度強い相関があるとき、変量 x と y の間に、$y = a + bx$ に近いの関係がある (a, b は定数) と考えられる。



● **最小二乗法**

x_i から予測される値 $ax_i + b$ と現実の値 y_i との差の二乗の和

$Q(a,b)=\sum_{i=1}^{n}\{y_i-(ax_i+b)\}^2$ が最小となるように係数 a,b の値を定める。

$$\begin{aligned}\frac{1}{n}Q(a,b)&=\frac{1}{n}\sum_{i=1}^{n}(y_i^2+a^2x_i^2+b^2-2ax_iy_i-2by_i+2abx_i)\\&=\overline{y^2}+a^2\overline{x^2}+b^2-2a\overline{xy}-2b\overline{y}+2ab\overline{x}=\{b-(\overline{y}-a\overline{x})\}^2+(\overline{x^2}-\overline{x}^2)a^2-2(\overline{xy}-\overline{x}\,\overline{y})a+\overline{y^2}-\overline{y}^2\\&=\{b-(\overline{y}-a\overline{x})\}^2+s_x^2a^2-2s_{xy}a+s_y^2=\{b-(\overline{y}-a\overline{x})\}^2+s_x^2\Big(a-\frac{s_{xy}}{s_x^2}\Big)^2-\frac{s_{xy}^2}{s_x^2}+s_y^2\end{aligned}$$

よって、$a=\dfrac{s_{xy}}{s_x^2}$, $b=\overline{y}-a\overline{x}=\overline{y}-\dfrac{s_{xy}}{s_x^2}\overline{x}$ のとき最小となるため、回帰直線の方程式は $y-\overline{y}=\dfrac{s_{xy}}{s_x^2}(x-\overline{x})$ と表される。(厳密には y の x への回帰直線という。)

例えば、経験的に親の身長と子供の身長は正の相関がある、すなわち、「背の高さは遺伝する」と思っている。英国人のゴルトンは 1885 年に約 1000 人を調べたデータを発表した。(実は彼の興味は「優秀な親からは優秀な子どもが生まれる」という現象の実証に興味があったとされている。) 彼のデータによると、

$$\text{子どもの身長} \;=\; 74.7 \;+\; 0.57 \;\times\; \text{両親の身長の平均値} \quad (\text{cm})$$

となる。ここで、0.57 という係数に着目されたい。これより相関係数は正であるから「背の高さは遺伝する」は事実ではありそうである。しかし、その係数が 1 より小さいということは、「身長が高い親の子どもほど実際にはそれほど高くない、とか、身長が低い親の子どもだって実際にはそれほど低くない」ということである。これを「平凡への回帰」あるいは「平均への回帰」とよぶ。

身長という測定誤差が小さく遺伝的要素が強いものでさえそうなのだから、知能についてはなおさらだろう。知能の高い両親から生まれた子どものほうが平均的には知能も高いのかもしれないが、それだけで十分予測ができるかというとそれほどでもない。だから人類が二極化するような進化をすることもないし、遺伝や人種にもとづいて人間を差別するメリットもないのである。([6] より。)

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] 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.
[4] 河野敬雄: 確率概論, 京都大学出版会, 1999.
[5] 国沢 清典 編: 確率統計演習 2 統計, 培風館, 1966.
[6] 西内 啓: 統計学が最強の学問である, ダイヤモンド社, 2013.
[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] 高橋 洋一: 統計・確率思考で世の中のカラクリがわかる, 光文社新書, 2011.
[10] 渡部 洋: ベイズ統計学入門, 福村出版, 1999.

問の解答

1.1 ともに余事象を考える。

(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 $\left(\frac{35}{36}\right)^{24}$. よって、勝つ確率は $1-\left(\frac{35}{36}\right)^{24} \fallingdotseq 0.4914$ となり、負けることが多いと予想される。また、$\left(\frac{35}{36}\right)^{25} \fallingdotseq 0.4945$ なので、$1-\left(\frac{35}{36}\right)^{24} < 0.5 < 1-\left(\frac{35}{36}\right)^{25}$ 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。

1.2 (a), (b) それぞれのゲームの勝負の残りをしたとすると、その勝敗は以下の表のようになる。

(a)

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

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

(b)

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

現在まで		6	7	8	9	勝者
(WWWLL)	→	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	–	B 氏

(1) (a) $\frac{1}{2}+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^3=\frac{7}{8}$. (b) $\left(\frac{1}{2}\right)^2+2\times\left(\frac{1}{2}\right)^3+3\times\left(\frac{1}{2}\right)^3=\frac{11}{16}$.

(2) (a) $\frac{4}{6}+\frac{2}{6}\cdot\frac{4}{6}+\left(\frac{2}{6}\right)^2\cdot\frac{4}{6}=\frac{26}{27}$. (b) $\left(\frac{3}{5}\right)^2+2\times\left(\frac{3}{5}\right)^2\cdot\frac{2}{5}+3\times\left(\frac{3}{5}\right)^2\cdot\left(\frac{2}{5}\right)^2=\frac{513}{625}$.

2.1 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。
仮定より $P_A(\overline{E})=0.01$, $P_{\overline{A}}(E)=0.02$, $P(A)=0.01$ であり、求める確率は $P_E(A)$ であるから、

$$P_E(A)=\frac{P(A)P_A(E)}{P(A)P_A(E)+P(\overline{A})P_{\overline{A}}(E)}=\frac{0.01\times(1-0.01)}{0.01\times(1-0.01)+0.99\times0.02}=\frac{1}{3}$$

$P(A)=0.0001$ の場合も同様に、$P_E(A)=\frac{1}{203}$.

2.2 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、

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

2.3 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、

$$P_F(A_3)=\frac{P(A_3)P_{A_3}(F)}{P(A_1)P_{A_1}(F)+P(A_2)P_{A_2}(F)+P(A_3)P_{A_3}(F)}=\frac{1\cdot6}{6\cdot2+3\cdot3+1\cdot6}=\frac{2}{9}$$

2.4 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 2.5 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、$P_{S_2}(C)=\dfrac{2}{5}$.

2.5 例題 2.6 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。

2.6 問 2.4 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 2.4 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 Q_1 は小さいほうから 8 番目のデータなので、階級値 55 の階級に属しており、その小さいほうから 2 番目のデータとなる。この階級に属するデータを抜き出し小さいほうから順に並べると $54, 54, 56, 58$ となるので、$Q_1=54$.
m は小さいほうから 15 番目と 16 番目のデータの平均なので、ともに階級値 65 の階級に属しており、その小さいほうから 5 番目と 6 番目のデータの平均となる。この階級に属するデータを抜き出し小さいほうから順に並べると $62, 64, 65, 65, 65, 66, 67, 70, 70$ となるので、$m=\dfrac{65+66}{2}=65.5$.

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

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

3.3 $s^2=\dfrac{1}{n}\displaystyle\sum_{k=1}^{r}({x_k}^2-2\overline{x}x_k+\overline{x}^2)f_k=\frac{1}{n}\sum_{k=1}^{r}{x_k}^2f_k-2\overline{x}\cdot\frac{1}{n}\sum_{k=1}^{r}x_kf_k+\overline{x}^2\frac{1}{n}\sum_{k=1}^{r}f_k$
$=\overline{x^2}-2\overline{x}\cdot\overline{x}+\overline{x}^2=\overline{x^2}-\overline{x}^2$.

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

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

3.5 (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。
$\overline{y}=\dfrac{1}{50}(2\cdot2+3\cdot3+4\cdot4+5\cdot6+6\cdot14+7\cdot8+8\cdot7+9\cdot6)=6.18$ より $\overline{x}=66.8$.
$\overline{y^2}=\dfrac{1}{50}(2^2\cdot2+3^2\cdot3+\cdots+9^2\cdot6)=41.58$ 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し並べかえると $51, 54, 54, 56, 56, 59$ 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 $m=\dfrac{66+68}{2}=67$.

3.6 $s_{xy}=\dfrac{1}{n}\displaystyle\sum_{k=1}^{n}(x_ky_k-\overline{x}y_k-\overline{y}x_k+\overline{x}\,\overline{y})=\frac{1}{n}\sum_{k=1}^{n}x_ky_k-\overline{x}\frac{1}{n}\sum_{k=1}^{n}y_k-\overline{y}\frac{1}{n}\sum_{k=1}^{n}x_k+\frac{1}{n}\sum_{k=1}^{n}\overline{x}\,\overline{y}$
$=\overline{xy}-\overline{x}\,\overline{y}-\overline{y}\,\overline{x}+\overline{x}\,\overline{y}=\overline{xy}-\overline{x}\,\overline{y}$.