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 30 年 8 月 18 日 (2018 年 8 月 18 日修正)

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[*1]。

例題 1.1 トスカナ大公は「3 個のサイコロ投げで、目の和は 9 より 10 の方が出やすいのはなぜか?」とガリレイに問うたと言われている。出やすいのはどうしてか。

トスカナ大公の疑問は、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 通りであり、また、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と推論したためと考えられる。

これに対し、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 になる場合の方が出やすいことを示した。(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 (2 つのサイコロ, ド・メレからパスカルへの質問 1) 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。

賭けに勝つ確率をそれぞれについて求めることで、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[*2]。

例題 1.2 (分配問題, ド・メレからパスカルへの質問 2) 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[*3]。

[*1] これらの歴史的な事項については安藤著 [1] を参考にした。哲学的側面から確率の歴史が述べられているものに [3] がある。[3] ではその先史についても触れられている。(参考文献リストは最後のページ p.16 にあります。)

[*2] 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. [15])。

[*3] この問いはルカ・パチョーリによる『スムマ (Sũma)』(1494 年刊) にすでに書かれている。1637 年頃メルセンヌのアカデミーで話題になっており、当時 14 歳のパスカルは父親に連れられてこのアカデミーに出入りしていたようである。この問題は 16 世紀にもカルダノやタルタリアをはじめ多くの数学者によって考察され、パスカルとフェルマーが最初に正解にたどり着いた。

解答 1: 両者の勝つ確率は等しいと仮定する。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。 □

次に数学 B で学ぶ二項分布を用いる解き方も見てみよう。

解答 2: 両者の勝つ確率は等しいと仮定する。5 回戦するものとし、X で残り 2 戦で「私」が勝つ回数を表すと、X は二項分布 $B\left(2, \frac{1}{2}\right)$ に従う[*4]。「私」はあと 1 勝すればよいので、求める確率は

$$P(X \geqq 1) = {}_2C_1 \frac{1}{2} \cdot \frac{1}{2} + {}_2C_2 \left(\frac{1}{2}\right)^2 = \frac{2+1}{4} = \frac{3}{4}.$$

よって、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。

(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

問 1.3 A 氏, B 氏, C 氏の 3 人が先に 4 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。賭け金を三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, B 氏, C 氏が受け取るべき賭け金の割合を決定せよ。ただし、3 人の実力は同じとして考えよ[*5]。

(a) その時点で A 氏 3 勝, B 氏 2 勝, C 氏 2 勝だった場合

(b) その時点で A 氏 3 勝, B 氏 2 勝, C 氏 1 勝だった場合

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[*6]。

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[*7]。

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

[*4] 確率変数 X が二項分布 $B(n,p)$ に従うとは $P(X=k) = {}_nC_k p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$, となるときにいう。ここで ${}_nC_k$ はパスカルの三角形で計算できる数であり、興味深いと思い取り上げた。ただし、この三角形はパスカルより前から知られていた。

[*5] パスカルとフェルマーの間で 3 人の場合を考察した手紙も残っている。この場合、例題 1.2 の解答 2 のように解くと複雑になる。

[*6] 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。マグレイン著 [10] ではベイズ理論の歴史、その多彩な応用例など詳しい記述がある。(数式はほとんど出てこない。)

[*7] 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利用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 ($cf.$ [18]) [*8]。

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 F を組み込むと、

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

と表される。

条件つき確率の性質をいくつか述べる。

$P(A) > 0$ とする。$P_A(\cdot)$ 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 に対して $P(A) > 0$ であれば

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

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

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

[*8] 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 ($cf.$ [2])。
この例はデータの分析における質的データ間の関係として考察する方が適切であるが、第 3 節で質的データを取り扱わなかったためここに記述した ($cf.$ 学習指導要領解説高等学校 数学編 理数編 平成 29 年 3 月公示 [11] p.47)。

を満たすとする。このとき、$P(F)>0$ かつ $P(C_i)>0,\ i=1,2,\dots,n,$ であれば

$$P_F(C_i)=\frac{P(C_i)P_{C_i}(F)}{P(C_1)P_{C_1}(F)+P(C_2)P_{C_2}(F)+\cdots+P(C_n)P_{C_n}(F)} \tag{1}$$

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表せる。

$$P_F(B)=\frac{P(B)P_B(F)}{P(B)P_B(F)+P(\overline{B})P_{\overline{B}}(F)} \tag{2}$$

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

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

第 2 の等号は $(C_i\cap F)\cap(C_j\cap F)=\emptyset\ (i\neq j)$ 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。[*9]

解答: メールが NG ワードを含んでいるという事象を F, 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、$P_S(F)=0.8,$

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

したがって、求める確率 $P_F(S)$ はベイズの定理より、

$$P_F(S)=\frac{P(S)P_S(F)}{P(S)P_S(F)+P(\overline{S})P_{\overline{S}}(F)}=\frac{0.6\times0.8}{0.6\times0.8+0.4\times0.05}=\frac{48}{50}=0.96. \qquad \square$$

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。[*10]

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 F とする。

---

*9 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。

*10 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 [17] 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 $997\times0.09\fallingdotseq 90$ 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、$P_{\overline{A}}(F) = 0.09$.

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、$P_A(F) = 1$.
したがって、求める確率 $P_F(A)$ はベイズの定理より、

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

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。*11

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。*12

問 2.2 ([9], [10] より) A 市で強盗殺人事件が起こり、X 氏が容疑者として逮捕された。現場の血痕から、犯人の血液型は 1000 人に一人という珍しいものであることがわかり、血液型の一致する X 氏が逮捕されたのだが、X 氏は果たして犯人なのだろうか。次の場合に X 氏が犯人である確率を求めよ。ただし A 市近郊の総人口は 100 万人とする。
(a) X 氏は犯人か犯人でないかの二つに一つだから、犯人であるという事前確率は 1/2 とした場合。
(b) 犯人が A 市の人間だとしても、A 市近郊には 100 万人の人間がいるのだから、X 氏が犯人であるという事前確率はどう大きく見積もっても 10 万分の 1 とした場合。

次に、モンティ・ホールの 3 ドア問題を考える。*13

例題 2.4 (モンティ・ホールの 3 ドア問題) 3 つの扉のうち 1 つだけに車が、残りの扉には山羊が入っていて、回答者は車の入っている扉を当てたら車がもらえる。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、車の入っている扉を知っている司会者 (モンティ・ホール) が、選んでいない扉で車の入っていない扉 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 で扉を変えると、当る確率はどのように変化するか、または、変化しないか?

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 だった場合を考える。

---

*11 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。[10] や [17] 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。また、[10] にはマンモグラフィー検査では乳がんの人を「がんの疑い」と判定する確率は 80% ($P_A(F) = 0.8$) とありました。

*12 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、血液型や初期の DNA 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 (*cf*. 問 2.2 とその解答)。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。とくに珍しい事象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。)「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*. [4])。

*13 モンティ・ホールの 3 ドア問題とまったく同値な問題に 3 囚人の問題がある。ローゼンハウス著 [14] によると、マーティン・ガードナーによる 1959 年の『サイエンティフィック・アメリカン』誌の連載記事が、3 囚人問題が紹介された最も古い文献のようである。[14] はモンティ・ホール問題についての書で、以下しばしば引用する。

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すると、その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、回答者が選んだ扉は A で司会者が開けた扉が B であるという事象を F とすると、

もし A に車があれば、司会者は B, C の扉をランダムに開けるので $P_A(F) = \dfrac{1}{2}$.

もし B に車があれば、司会者は B の扉を開けることはないので $P_B(F) = 0$.

もし C に車があれば、司会者は B の扉を必ず開けるので $P_C(F) = 1$.

このとき、A の扉に車のある確率は $P_F(A)$ であるから、ベイズの定理を用いて

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

となり、したがって $P_F(C) = 2/3$ となる。よって、扉を変えれば当る確率は 2 倍となる。[*14] □

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 A,C にある (事後) 確率をそれぞれ計算せよ。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 ドア問題 ([5] による) を考える。これは更に直感と異なる結果となる。[*15]

例題 2.5 (変形 3 ドア問題) 例題 2.4 でこの番組の熱心な視聴者である回答者は、それまでの番組の観察を通して、車のある位置が A, B, C の扉にそれぞれ 1/4, 1/4, 1/2 の確率で車が配置されること、一方、司会者は回答者が当たりの扉を選んでいる場合は、残りの扉から等確率で 1 つを選んで開ける傾向があるとの情報を得た。この場合、回答者が A の扉を選択し、その後、司会者が B を開けたとすると、A の扉に車のあ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。
司会者は回答者が選んだ A と車のある扉以外を等確率で開けるので、$P_A(F) = \dfrac{1}{2}$, $P_B(F) = 0$, $P_C(F) = 1$.
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。 □

市川と下條 ([5]) は、統計学をある程度知っている大学院生に予備的に面接した結果に基づいて、人がこのような問題を解くときに用いる推論について、出発点となる仮説を立てた。その仮説には、次の三つの**主観的定理** (数学的な定理ではない) の利用が含まれている:

[*14] 1990 年 9 月 9 日発行、ニュース雑誌 Parade にて、マリリン・ボス・サヴァントが連載するコラム欄「マリリンにおまかせ」において読者投稿による質問に「正解は『ドアを変更する』である。なぜなら、ドアを変更した場合には景品を当てる確率が 2 倍になるからだ」と回答したところ、読者から「彼女の解答は間違っている」との約 1 万通の投書が殺到したことにより、この問題が知られるようになった。投書には 1000 人近い博士号保持者からのものも含まれており「ドアを変えても確率は五分五分（2 分の 1）であり、3 分の 2 にはならない」と主張した (wikipedia「モンティ・ホール問題」の事項より)。この顛末は [14] に詳しい。同書によるとポール・エルデッシュでさえ、問題を取り違えただけでなく、しばらくは正しい答えを認めようとしなかった。また、パーシー・ディアコニスは「私たちの脳は、確率の問題をうまく処理するようにできていないので、間違いがあっても私は驚かない。」と述べている。ちなみに、当のモンティ・ホール氏は扉を変えることで確率が増加することを知っていたようであるとあった。
　認知科学の書籍 [5] によると、2 つの扉で車のある確率は 1/2 ずつであると考えてしまう人がほとんどで、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。[14] には [5] 以降考察された認知科学の結果も記載されている。

[*15] [14] には司会者がランダムに (車のある扉を知らない) 場合や司会者が扉を開けて回答者が選び直す行為を複数回繰り返す漸進モンティ・ホール問題など、様々なモンティ・ホール問題の変形が紹介されている。

「場合の数」定理　あらゆる選択肢の数が N のとき、それぞれの選択肢の確率は $1/N$ である。

「等比率」定理　一つの選択肢が除外されても、残った選択肢どうしの比は事前確率と同じである。

「不変」定理　一部の選択肢 $(A_1, A_2, \ldots, A_k)$ にうち少なくとも一つが除外されることが確実な場合、その選択肢が除外されるかを特定する情報が与えられても、その一部以外の選択肢 $(A_{k+1}, \ldots, A_N)$ の確率は変わらない。

例題 2.5 では、4 通りの解き方 (ベイズの定理と三つの主観的定理) が異なる答えを導くこととなる。詳細は、以下のようになる。分数は 4 つの方法それぞれを介して二つの問題について得られた $P_F(A)$ の値を表す。(文章は [14] より引用しています。詳細は [5] もしくは [14] を参照ください。)

定理	例題 2.4	例題 2.5
ベイズの定理	1/3	1/5
「場合の数」定理	1/2	1/2
「等比率」定理	1/2	1/3
「不変」定理	1/3	1/4

問 2.4 例題 2.5 で A, B, C の各扉に車がある事前確率がそれぞれが 1/4, 1/2, 1/4 であったとき、A に車がある事後確率はいくらになるか。また、事前確率が A, B, C それぞれ 1/2, 1/4, 1/4 であったときはどうか。

問 2.5 例題 2.5 と同様に A, B, C の各扉に車がある事前確率がそれぞれが 1/4, 1/4, 1/2 であったとき、もし、回答者が A の扉を選択し、その後、司会者が C を開けたなら、A の扉に車のある確率はいくらになるか。

問 2.6 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, C, D にある事後確率をそれぞれ計算せよ。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。

3　データの分析

ここでは、記述統計の話題をいくつか扱ってみましょう[*16]。

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。[*17]

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。

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

[*16] 数学 I で学ぶ記述統計 (与えられたデータの構造を明らかにすることを目的にする) に対し、数学 B では推測統計 (標本から母集団の構造に関する推論を目的とし、推定や仮説検定を扱う) を学ぶ。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 [16], p.265 に簡潔にまとめられている。

[12] 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き社会人として必須の能力になる日が来ると予言した」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。

[*17] 平成 29 年 3 月公示の学習指導要領解説によると、四分位範囲や箱ひげ図は中学校数学科第 2 学年で学ぶとある。

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.

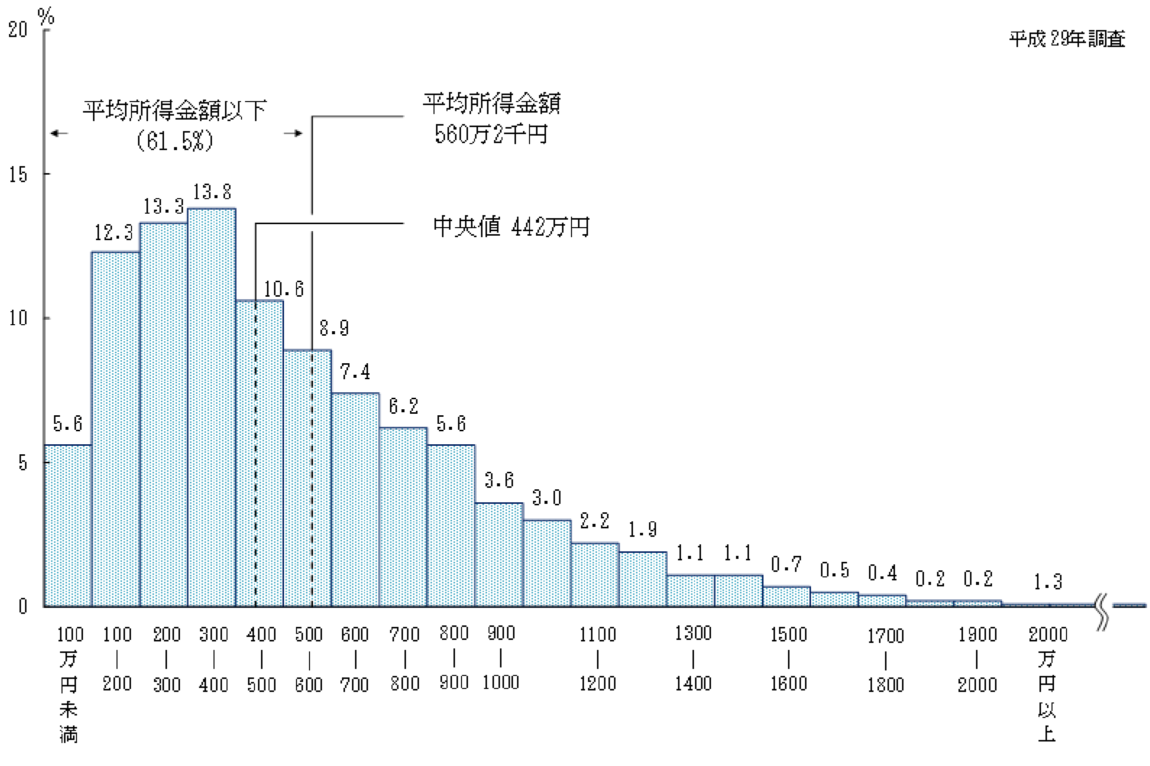
(2) 平均値: $\overline{x} = \dfrac{42+38+40+44+52+198}{6} = 69$

中央値: $38 < 40 < 42 < 44 < 52 < 198$ 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「外れ値」の影響を受けやすいが (「外れ値」については p.10 の箱ひげ図の書き方の脚注を参照のこと)、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[18]。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」に近いか考えるために、厚生労働省による平成 29 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。[19]



元データから平均値は 560.2 万円であり、中央値が 442 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 300–400 万円であることがわかる。

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、所得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、外れ値が出やすいデータの場合

[18] もちろん表計算ソフトを用いれば平均値も中央値も容易に求めることができます。

[19] `https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa17/index.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。X が対数正規分布に従うとは、その対数 $\log X$ が正規分布に従うと定義される。対数正規分布は正規分布に等に比べ外れ値が出やすい。平成 29 年度調査では 1500 万円以上が外れ値に当たり、全体の 3.3% となっている。

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 ($cf.$ [16])。

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** [20]

$n = 2m$ 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。 □

注意 3.2 四分位数の定義は複数ある。表計算ソフト Excel の QUARTILE 関数は、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

$$y = f(t) = \begin{cases} x_{(t)}, & t \text{ が自然数} \\ (\lceil t \rceil - t) x_{(\lfloor t \rfloor)} + (t - \lfloor t \rfloor) x_{(\lceil t \rceil)}, & \text{それ以外} \end{cases}, \quad 1 \leqq t \leqq n,$$

とし、$Q_q = f(1 + \frac{q}{4}(n-1))$, $q = 1, 3$, と定めているようである[21]。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 は次のようになる。[22]

(1) $1 + \dfrac{3}{4}(9-1) = 7$ より $Q_3 = x_{(7)} = 89$.

(2) $1 + \dfrac{3}{4}(10-1) = 7.75$ より $Q_3 = 0.25 x_{(7)} + 0.75 x_{(8)} = 86.25$ となる。

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	69
28	93	100	58	88	26
64	66	65	87	50	54
37	98	73	62	33	39
56	79	65	77	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	0	5	7	7	3	3	30

階級値 25 は 20 点以上 30 点未満のとし、35, 45,$\cdots$, も同様とした。ただし、100 点は 95 の階級に含めた。

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。

---

[20] ここでは高校数学 I で学ぶ定義を紹介する。注意 3.2 も参照のこと。

[21] 中央値は n が奇数、偶数にかかわらず $m = Q_2 = f(1 + \frac{1}{2}(n-1))$ と表せる。

[22] これに類するものにテューキーの「ヒンジ」があるが、これは n が奇数の場合は注意 3.2 と同じであるが、n が偶数の場合は高校数学 I で習う定義と同様に上位のデータ、下位のデータの中央値と定義する。

解答: データ数が 30 だから上位のデータは 15 個なので、Q_3 は大きいほうから 8 番目のデータとなる。これは階級値 75 の階級に属しており、その大きいほうから 2 番目となる。この階級に属するデータを抜き出すと 70, 78, 73, 79, 77, 75 であるから、これを並べ直すと $70, 70, 73, 75, 77, 78, 79$ 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。(まずどの階級にあるかを考えよ。)

問 3.2 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	3	7	12	9	8	6	50

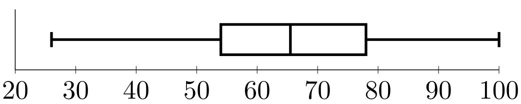
ただし、階級値 25 は 20 点以上 30 点未満のとし、35, 45,$\cdots$, も同様とした。また、100 点は 95 の階級に含めた。このデータの第 1 四分位数 Q_1 と中央値 m を求めよ。

- 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[*23]:

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. 四分位範囲のの左右に最大値と最小値まで「ひげ」(左に「├───」, 右に「───┤」) を引く。

例題 3.3 のデータの場合、最小値 26, 第 1 四分位数 54, 中央値 65.5, 第 3 四分位数 78, 最大値 100 であるから、箱ひげ図は右のようになる。



問 3.3 問 3.2 のデータについて、その解答 (第 1 四分位数, 中央値) と、最小値, 第 3 四分位数, 最大値が順に 26, 78, 100 であることを用いて、その箱ひげ図を完成させよ。

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

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

[*23] 数学 I の教科書にあるように、平均値を「+」で描くこともある。(ここでは省略する。) 箱ひげ図の発案者テューキーの流儀では、3 で四分位範囲の 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」を引く。さらに、その外側の左右にあるデータを「外れ値」として「○」でプロットする。また、測定ミス・記入ミスなど原因がわかっているものは「異常値」として区別することもある。平成 29 年 3 月公示の学習指導要領解説では「外れ値を見出す意義を理解できるようにする」とあり、この方法で描かれた箱ひげ図が紹介されている ([11], 高等学校 数学編 p.48)。そこには、標準偏差 σ を用いて、平均値より $\pm 2\sigma$ (事象によっては $\pm 3\sigma$) 以上離れた値とする「外れ値」の定義も紹介されている。
最小値・第 1 四分位数・中央値・第 3 四分位数・最大値を並べたものを五数要約ということがある (テューキーが命名, *cf*. [13])。

定理 3.2 n 個のデータ $x_1, x_2, \cdots, x_n$ と数値 p に対して、各データ x_k と p の距離の平方の和を $f(p)$ とする。このとき、$f(p)$ は $p = \overline{x}$ (平均値) で最小値 $f(\overline{x}) = ns^2$ (データ数 × 分散) をとる。

証明:
$$\begin{aligned} f(p) &= (x_1 - p)^2 + (x_2 - p)^2 + \cdots + (x_n - p)^2 \\ &= {x_1}^2 + {x_2}^2 + \cdots + {x_n}^2 - 2(x_1 + x_2 + \cdots + x_n)p + np^2 \\ &= n\overline{x^2} - 2n\overline{x}^2 + np^2 = n(p - \overline{x})^2 + n\overline{x^2} - n\overline{x}^2 = n(p - \overline{x})^2 + ns^2 \qquad \square \end{aligned}$$

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)*24。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 $50 + 10 \times \dfrac{x_1 - \overline{x}}{s}$ 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下から何 % 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 3.4 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) ${s_z}^2 = \dfrac{m}{m+n}{s_x}^2 + \dfrac{n}{m+n}{s_y}^2 + \dfrac{mn}{(m+n)^2}(\overline{x} - \overline{y})^2$

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 (正確にはピアソンの積率相関係数という)

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

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、*25

*24 所得や株価は何 % 増・何 % 減と積の形で増減するので、その対数が正規分布に従うことになる。

*25 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、$\sum_{k=1}^{n}(a_k t + b_k)^2$ を t について平方完成することで証明できる。

$c > 0$ 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 次の数値は、ある授業の 10 人の学生についてのテストの点数である。このとき、科目 x, y の平均値、分散および科目 x, y の共分散と相関係数を (電卓を用いて) 求めよ。

科目 x	80	56	65	71	56	47	42	53	45	60
科目 y	90	36	49	71	57	53	37	76	30	40

- **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 ($cf.$ [16], 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、因果関係を推定する五箇条がある。これ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である (詳細は [2], p.102 を参照ください)。

関連の一致性　他の集団でも同じ現象が観察される。
関連の強固性　相関係数などいくつかの指標で評価する。
関連の特異性　原因と結果とが必要十分である。
関連の時間性　原因として疑われるものは、必ず結果に先行する。
関連の整合性　既知の事実との合致・無矛盾性。

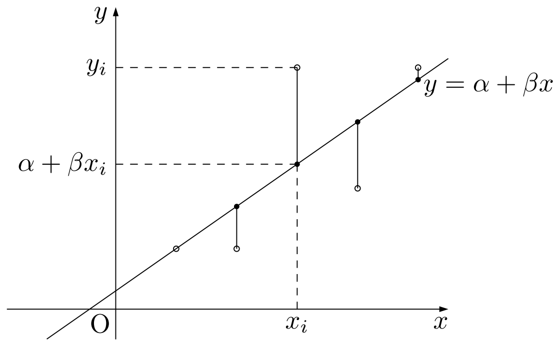
因果関係を実証するため、ランダム化比較実験 (集団をランダムにグループ分けし、一方にのみ介入を行い、介入を行わなかったグループと比較をする手法) が用いられ、例えば「少人数教育は子どもの教育のために良いのか」といったことも調べられている ($cf.$ [6])。

(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、上記の「サラリーマンの年収と血圧」がそれにあたる。これには、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 ($cf.$ [16])。



- **回帰直線**　最後に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、$y=\alpha+\beta x$ に近いの関係がある (α, β は定数) と考えられる。x を独立変数、y を従属変数という。

- **最小二乗法**

x_i から予測される値 $\alpha+\beta x_i$ と現実の値 y_i との差の二乗の和 $Q(\alpha,\beta)=\sum_{i=1}^{n}\{y_i-(\alpha+\beta x_i)\}^2$ が最小となるように係数 α,β の値を定める。

$$
\begin{aligned}
\frac{1}{n}Q(\alpha,\beta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i^2+\alpha^2+\beta^2x_i^2-2\alpha y_i-2\beta x_iy_i+2\alpha\beta x_i) \\
&= \overline{y^2}+\alpha^2+\beta^2\overline{x^2}-2\alpha\overline{y}-2\beta\overline{xy}+2\alpha\beta\overline{x}=\{\alpha-(\overline{y}-\beta\overline{x})\}^2+(\overline{x^2}-\overline{x}^2)\beta^2-2(\overline{xy}-\overline{x}\,\overline{y})\beta+\overline{y^2}-\overline{y}^2 \\
&= \{\alpha-(\overline{y}-\beta\overline{x})\}^2+s_x^2\beta^2-2s_{xy}\beta+s_y^2=\{\alpha-(\overline{y}-\beta\overline{x})\}^2+s_x^2\Big(\beta-\frac{s_{xy}}{s_x^2}\Big)^2-\frac{s_{xy}^2}{s_x^2}+s_y^2
\end{aligned}
$$

よって、$\beta=\dfrac{s_{xy}}{s_x^2}$, $\alpha=\overline{y}-\beta\overline{x}=\overline{y}-\dfrac{s_{xy}}{s_x^2}\overline{x}$ のとき最小となるため、回帰直線の方程式は

$$
y-\overline{y}=\frac{s_{xy}}{s_x^2}(x-\overline{x}) \qquad \text{あるいは変形して} \qquad \frac{y-\overline{y}}{s_y}=r\frac{x-\overline{x}}{s_x}
$$

と表される。(厳密には y の x への回帰直線という。) ここで、r は相関係数である。

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。(実は彼の関心は「優秀な親からは優秀な子どもが生まれる」ことの実証にあったとされている。) 彼のデータによると、

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

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し、経験的にも「背の高さは遺伝する」は事実でありそうである。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 ([12] より)。

問 3.7 問 3.6 の場合に y の x への回帰直線の方程式を求めよ。(各係数は有効数字 4 桁で求めよ。)

4 仮説検定について

高等学校学習指導要領解説数学編 [11] p.48 に「具体的な事象において仮説検定の考え方を理解するとともに，不確実な事象の起こりやすさに着目し，主張の妥当性について，実験などを通して判断したり，批判的に考察したりすること」とあります。この機会に統計的仮説検定の考え方を見ておきましょう。*26

例題 4.1 1920 年代末のイギリスにて、陽射しの強いある夏の午後。何人かの英国紳士と婦人たちが屋外のテーブルで紅茶を楽しんでいたときのことだった。その場にいたある婦人はミルクティについて「紅茶を先に入れたミルクティ」か「ミルクを先に入れたミルクティ」か、味が全然違うからすぐにわかると言ったらしい。

その場にいた紳士たちのほとんどは、婦人の主張を笑い飛ばした。彼らが学んだ科学的知識に基づけば、紅茶とミルクが一度混ざってしまえば何ら化学的性質の違いなどない。

だが、その場にいた 1 人の小柄で、分厚い眼鏡をかけ髭を生やした男だけが、婦人の説明を面白がって「その命題をテストしてみようじゃないか」と提案したそうだ。その男こそが、現代統計学の父、ロナルド A. フィッシャーである。

彼はさっそくティカップをずらりと並べ、婦人に見えない場所で 2 種類の違った淹れ方のミルクティを用意した。そしてランダムな順番で婦人にミルクティを飲ませ、婦人の答えを書き留めた後でちょっとした確率の計算を行った。婦人は出された 5 杯のミルクティをすべて正確に言い当てることができた。*27

仮説検定では、

1. まず主張したい内容と逆の仮説である**帰無仮説** H_0 と
   帰無仮説が棄却されたとき、代わりに採択される仮説である**対立仮説** H_1 を設定する
2. 帰無仮説が正しい場合に、誤って棄却される確率の限界である**有意水準** α を決める。
3. 帰無仮説のもとその現象が起こりうる確率 (p 値という) を計算する。あるいは有意水準から帰無仮説を棄却する範囲である**棄却域**を決める。
4. 3 で求めた確率が有意水準より小さければ帰無仮説を**受容**し、大きければ帰無仮説を**棄却**する。

として、帰無仮説が棄却されれば対立仮説を採択し、受容されれば帰無仮説は否定されないとします。

例題 4.1 を婦人がミルクティを正しく選ぶ確率を p として，帰無仮説 $\mathrm{H}_0 : p = \dfrac{1}{2}$, 対立仮説 $\mathrm{H}_1 : p > \dfrac{1}{2}$ (片側検定という) とし、有意水準はもっとも一般的な 5% で検定しましょう。

H_0 のもと、婦人が 5 杯とも正解する確率は $\left(\dfrac{1}{2}\right)^5 = \dfrac{1}{32}$ で 5% より小さいので、H_0 は棄却される。すなわち「婦人はミルクティを正しく選ぶことができる」となる。

もう少しまじめに、教科書にある例題を考えましょう。

例題 4.2 あるコインを 8 回投げたところ 7 回表が出た。このコインは表が出る確率は $\dfrac{1}{2}$ であるか。有意水準 5% で検定せよ。

解答: 表が出る確率を p、帰無仮説 $\mathrm{H}_0 : p = \dfrac{1}{2}$, 対立仮説 $\mathrm{H}_1 : p \neq \dfrac{1}{2}$ (両側検定という) とし検定する。

X で 8 回投げたときの表の出る回数とすると、「帰無仮説のもとその現象が起こりうる」場合は、より起こりにくいと考えられる場合であるすべてが表となる場合と、両側検定なので裏が 7 回または 8 回出る確率を合

*26 数学 B では正規分布に基づく仮説検定を学びますが、ここではより簡単な二項分布に基づく場合を考えます。仮説検定は農学、生物学、医学薬学、心理学、経済学、金融保険をはじめデータを扱う様々な分野で用いられ発展しています。興味のある方は、インターネットで調べたり、例えば統計学図鑑 [7] など書物を調べたりしてください。

*27 [12] p.102 より。これはランダム化比較実験の例ですが、ここでは仮説検定の問いとして考察します。
同書 p.106 には「王立化学協会が 2003 年に発表した「一杯の完璧な紅茶の淹れ方」というウィットにとんだプレリリースがある。」とあります。興味のある方はインターネットで「一杯の完璧な紅茶の淹れ方」と検索してみてください。

計したものが p 値となる。したがって、

$$p\text{ 値} = {}_8C_0\left(\frac{1}{2}\right)^8 + {}_8C_1\left(\frac{1}{2}\right)^8 + {}_8C_7\left(\frac{1}{2}\right)^8 + {}_8C_8\left(\frac{1}{2}\right)^8 = \frac{9}{128} = 0.070\cdots.$$

これは 5% より大きいので H_0 は受容される。すなわち表が出る確率が $\frac{1}{2}$ でないとはいえない。 □

例題 4.2 では帰無仮説が $p = 1/2$ であるため、p 値を求めることができた。一般には p 値を求めず棄却する範囲 (棄却域) を定めて検定する方法を用いる。両側検定の場合、棄却域は $X \leqq k$, $X \geqq l$ となる確率がともに (有意水準)/2 を超えない最大の範囲として定義する。そこで X の実際の値 (実現値) が棄却域に入っていれば H_0 を棄却し、そうでなければ H_0 を受容する。片側検定の場合は対立仮説 $\mathrm{H}_1 : p > p_0$ であれば $X \geqq l$ となる確率が有意水準を超えない最大の範囲を棄却域とする。$\mathrm{H}_1 : p < p_0$ の場合は $X \leqq k$ とし同様に定める。

例題 4.3 和子さんは、商店街で買いものをして、5 本のうち 3 本が当たりというスピードくじをひいた。ところが、和子さんは 12 本もひいたのに、4 本しか当たらなかった。このスピードくじは、正しくつくられていないのだろうか。それとも、ただ和子さんが運が悪かっただけだろうか。([8] p.183 より。)

解答: このくじが当たる確率を p、帰無仮説 $\mathrm{H}_0 : p = \frac{3}{5}$, 対立仮説 $\mathrm{H}_1 : p \neq \frac{3}{5}$, 有意水準 5% で検定する。$X$ で 12 回ひいたときの当たり本数とすると、X の確率分布は次の表で与えられる。

X	0	1	2	3	4	$\cdots$	10	11	12
確率	0.00002	0.00030	0.00249	0.01236	0.04204	$\cdots$	0.06385	0.01741	0.00218

両側検定なので $X \leqq k$, $X \geqq l$ となる確率がそれぞれ 0.025 以下となるように棄却域を定めると、棄却域は $X \leqq 3$ また $X \geqq 11$ となる。

和子さんは 4 本当たったが、実現値 $X = 4$ は棄却域に入らない。したがって、H_0 は受容される。すなわち当たる確率が $\frac{3}{5}$ でないとはいえない。 □

問 4.1 5 本のうち 3 本が当たりというスピードくじを、8 本ひいたところ 2 本当たった。次の (1), (2) について有意水準 5% で検定せよ。ただし、X で当たる確率が 0.6 のくじを 8 回ひいたときの当たり本数とすると、X の確率分布は次の表で与えられる。

X	0	1	2	3	4	5	6	7	8
確率	0.00066	0.00786	0.04129	0.12386	0.23224	0.27869	0.20902	0.08958	0.01680

(1) このスピードくじは正しくつくられているか (両側検定)。

(2) このスピードくじの当たりの本数は 5 本中 3 本より低いのではないか (片側検定)。

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 $\left(\frac{5}{6}\right)^4$. よって、勝つ確率は $1 - \left(\frac{5}{6}\right)^4 \fallingdotseq 0.5177$ となり、勝てることが多くなる。

(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、負けることが多くなる。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つ確率が 0.5 より大きくなる。

1.2 A, B でそれぞれ A 氏, B 氏の勝を表すと、その勝敗は以下の表のようになる。

	現在まで		7	8	9	勝者	現在まで		7	8	9	勝者
(a)	(AAAABB)	→	A	–	–	A 氏	(AAAABB)	→	B	B	A	A 氏
		→	B	A	–	A 氏		→	B	B	B	B 氏

(b)

現在まで		6	7	8	9	勝者
(AAABB)	→	A	A	–	–	A 氏
	→	A	B	A	–	A 氏
	→	A	B	B	A	A 氏
	→	A	B	B	B	B 氏
	→	B	A	A	–	A 氏

現在まで		6	7	8	9	勝者
(AAABB)	→	B	A	B	A	A 氏
	→	B	A	B	B	B 氏
	→	B	B	A	A	A 氏
	→	B	B	A	B	B 氏
	→	B	B	B	–	B 氏

よって (a) $\frac{1}{2}+\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^3=\frac{7}{8}$ より A 氏が $\frac{7}{8}$, B 氏が $\frac{1}{8}$.

(b) $\left(\frac{1}{2}\right)^2+2\times\left(\frac{1}{2}\right)^3+3\times\left(\frac{1}{2}\right)^3=\frac{11}{16}$ より A 氏が $\frac{11}{16}$, B 氏が $\frac{5}{16}$.

1.3 A, B, C でそれぞれ A 氏, B 氏, C 氏の勝を表すと、その勝敗は以下の表のようになる。

(a)

現在まで		8	9	10	勝者
(AAA	→	A	–	–	A 氏
BBCC)	→	B	A	–	A 氏
	→	B	B	–	B 氏
	→	B	C	A	A 氏
	→	B	C	B	B 氏
	→	B	C	C	C 氏

現在まで		8	9	10	勝者
(AAA	→	C	A	–	A 氏
BBCC)	→	C	B	A	A 氏
	→	C	B	B	B 氏
	→	C	B	C	C 氏
	→	C	C	–	C 氏

(b)

現在まで		7	8	9	10	勝者
(AAA	→	A	–	–	–	A 氏
BBC)	→	B	A	–	–	A 氏
	→	B	B	–	–	B 氏
	→	B	C	A	–	A 氏
	→	B	C	B	–	B 氏
	→	B	C	C	A	A 氏
	→	B	C	C	B	B 氏
	→	B	C	C	C	C 氏
	→	C	A	–	–	A 氏

現在まで		7	8	9	10	勝者
(AAA	→	C	B	A	–	A 氏
BBC)	→	C	B	B	–	B 氏
	→	C	B	C	A	A 氏
	→	C	B	C	B	B 氏
	→	C	B	C	C	C 氏
	→	C	C	A	–	A 氏
	→	C	C	B	A	A 氏
	→	C	C	B	B	B 氏
	→	C	C	B	C	C 氏
	→	C	C	C	–	C 氏

(a) A 氏: $\frac{1}{3}+2\times\left(\frac{1}{3}\right)^2+3\times\left(\frac{1}{3}\right)^3=\frac{17}{27}$. B 氏, C 氏: $\left(\frac{1}{3}\right)^2+2\times\left(\frac{1}{3}\right)^3=\frac{5}{27}$ より A 氏が $\frac{17}{27}$, B 氏, C 氏がそれぞれ $\frac{5}{27}$.

(b) A 氏: $\frac{1}{3}+2\times\left(\frac{1}{3}\right)^2+3\times\left(\frac{1}{3}\right)^3+3\times\left(\frac{1}{3}\right)^4=\frac{19}{27}$, B 氏: $\left(\frac{1}{3}\right)^2+2\times\left(\frac{1}{3}\right)^3+3\times\left(\frac{1}{3}\right)^4=\frac{2}{9}$, C 氏: $\left(\frac{1}{3}\right)^3+3\times\left(\frac{1}{3}\right)^4=\frac{2}{27}$ より A 氏が $\frac{19}{27}$, B 氏が $\frac{2}{9}$, C 氏が $\frac{2}{27}$.

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

$$P_F(A)=\frac{P(A)P_A(F)}{P(A)P_A(F)+P(\overline{A})P_{\overline{A}}(F)}=\frac{0.01\times(1-0.01)}{0.01\times(1-0.01)+0.99\times 0.02}=\frac{1}{3}.$$

$P(A)=0.0001$ の場合も同様に計算でき、$P_F(A)=\frac{1}{203}$.

2.2 X 氏が犯人であるという事象を A, X 氏の血液型が犯人の血液型と一致するという事象を F とする。このとき、X 氏が犯人であれば血液型は犯人のものと一致するから $P_A(E)=1$. X 氏が犯人でなければ血液型が一致するのは 1000 分の 1 と考えられるから $P_{\overline{A}}(E)=0.001$. これとベイズの定理より

$$P_E(A)=\frac{P(A)P_A(E)}{P(A)P_A(E)+P(\overline{A})P_{\overline{A}}(E)}=\frac{1000P(A)}{1000P(A)+P(\overline{A})}.$$

(a) $P(A) = P(\overline{A}) = \dfrac{1}{2}$ であるから、$P_E(A) = \dfrac{1000}{1000+1} \fallingdotseq 0.999$ となる。
すなわち、99.9% の確率で X 氏が犯人である。

(b) $P(A) = \dfrac{1}{100,000}$ より、$P_E(A) = \dfrac{0.01}{0.01+0.99999} \fallingdotseq 0.00990$ となる。
すなわち、X 氏が犯人である確率は 1% 未満。

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、$P(A) = P(B) = P(C) = P(D) = P(E) = \dfrac{1}{5}$.

(1) 回答者が A の扉を選び司会者が B の扉を開けるという事象を F_1 とすると、
例題 2.4 と同様に、$P_A(F_1) = \dfrac{1}{4}$, $P_B(F_1) = 0$, $P_C(F_1) = P_D(F_1) = P_E(F_1) = \dfrac{1}{3}$. よって、

$$P_{F_1}(A) = \frac{P(A)P_A(F_1)}{P(A)P_A(F_1) + P(B)P_B(F_1) + P(C)P_C(F_1) + P(D)P_D(F_1) + P(E)P_E(F_1)} = \frac{3}{15},$$

同様に $P_{F_1}(C) = \dfrac{4}{15}$.

(2) 回答者が A の扉を選び司会者が B, E の扉を開けるという事象を F_2 とすると、(1) と同様に、
$P_A(F_2) = \dfrac{1}{{}_4\mathrm{C}_2} = \dfrac{1}{6}$, $P_B(F_2) = P_E(F_2) = 0$, $P_C(F_2) = P_D(F_2) = \dfrac{1}{{}_3\mathrm{C}_2} = \dfrac{1}{3}$.
よって、$P_{F_2}(A) = \dfrac{1}{5}$, $P_{F_2}(C) = \dfrac{2}{5}$.

2.4 例題 2.5 と同じ記号を用いると、$P_A(F) = \dfrac{1}{2}$, $P_B(F) = 0$, $P_C(F) = 1$. 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、$P(A) = P(C) = \dfrac{1}{4}$, $P(B) = \dfrac{1}{2}$ より、$P_F(A) = \dfrac{1}{3}$.
また、1/2, 1/4, 1/4 のとき、$P_F(A) = \dfrac{1}{2}$ となる。

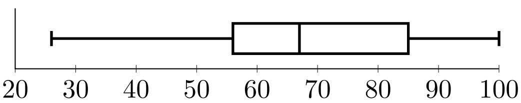
2.5 例題 2.5 と同じ記号を用いると、G で回答者が A の扉を選び司会者が C の扉を開けるという事象を表すとき、$P(A) = P(B) = \dfrac{1}{4}$, $P(C) = \dfrac{1}{2}$ で、$P_A(G) = \dfrac{1}{2}$, $P_B(G) = 1$, $P_C(G) = 0$ より $P_G(A) = \dfrac{1}{3}$.

2.6 問 2.3 の解答と同じ記号を用いると、$P(A) = P(B) = P(C) = \dfrac{1}{6}$, $P(D) = P(E) = \dfrac{1}{4}$. これより、問 2.3 と全く同様に (1) $P_{F_1}(A) = \dfrac{3}{19}$, $P_{F_1}(C) = \dfrac{4}{19}$, $P_{F_1}(D) = \dfrac{6}{19}$,
(2) $P_{F_2}(A) = \dfrac{1}{6}$, $P_{F_2}(C) = \dfrac{1}{3}$, $P_{F_2}(D) = \dfrac{1}{2}$ 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 3 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昇順に並べると $50, 54, 54, 56, 58$ 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昇順に並べると $62, 64, 65, 65, 65, 66, 67$ となるので、$m = \dfrac{65+66}{2} = 65.5$.

3.2 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 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し昇順に並べかえると $50, 51, 54, 54, 56, 56, 59$ となるので、$Q_1 = 56$.
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 2 番目と 3 番目のデータとなる。65 の階級値に属するデータを抜き出すと 65, 65, 68, 66, 64, 66, 65, 69, 61, 62, 65, 61 であるから、これを降順に並べかえて 69, 68, 66, 66, . . ., となるので、$m = \dfrac{66+68}{2} = 67$.

3.3



3.4 (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。

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

$= \overline{xy} - \overline{x}\,\overline{y} - \overline{y}\,\overline{x} + \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y}.$

3.6 平均値 $\overline{x} = 57.5$, $\overline{y} = 53.9$. 分散は $\overline{x^2} = 3434.5$, $\overline{y^2} = 3254.1$ より ${s_x}^2 = 128.25$, ${s_y}^2 = 348.89$. 共分散は $\overline{xy} = 3245.7$ より $s_{xy} = 146.45$, 相関係数は $r = 0.69233\cdots$.

3.7 $\beta = \dfrac{s_{xy}}{{s_x}^2} = 1.1419\cdots, \alpha = \overline{y} - \beta\overline{x} = -11.759\cdots$ より、$y = -11.76 + 1.142x$.

4.1 このくじが当たる確率を p とする。

(1) $\mathrm{H}_0 : p = 0.6$, $\mathrm{H}_1 : p \neq 0.6$ で有意水準 5% なので表より棄却域は $X \leqq 1$, $X \geqq 8$ となる。実現値 $X = 2$ は棄却域に入らないので、H_0 は受容される。すなわち当たる確率が 0.6 でないとはいえない。

(2) $\mathrm{H}_0 : p = 0.6$, $\mathrm{H}_1 : p < 0.6$ で有意水準 5% なので表より棄却域は $X \leqq 2$ となる。実現値 $X = 2$ は棄却域に入る (p 値 $= 0.04981$) ので、H_0 は棄却される。すなわち当たる確率は 0.6 より低いと考えられる。

注意 このように両側検定より片側検定のほうが H_0 を棄却しやすい。[7] p.80 には「片側検定を使いたくなるけど普段は両側検定を使っておいた方がよさそうだ」とある。

参考文献

[1] 安藤 洋美: 確率論の生い立ち, 現代数学社, 1992.

[2] 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.

[3] イアン ハッキング (広田 すみれ, 森元 良太 訳): 確率の出現, 慶應義塾大学出版会, 2013.

[4] 服部 哲弥: 統計と確率の基礎, 学術図書出版社, 2014.

[5] 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.

[6] 伊藤 公一朗: データ分析の力 因果関係に迫る思考法, 光文社新書, 2017.

[7] 栗原伸一, 丸山敦史, ジーグレイプ: 統計学図鑑, オーム社, 2017.

[8] 黒田 孝郎, 小島 順, 野崎 昭弘, 森 毅: 高等学校の確率・統計, ちくま学芸文庫, 2011.

[9] 楠岡 成雄: 確率・統計, 森北出版, 1995.

[10] シャロン バーチュ マグレイン (冨永 星 訳): 異端の統計学 ベイズ, 草思社, 2013.

[11] 学習指導要領解説 (平成 29 年 3 月公示)
中学校 数学編 `http://www.mext.go.jp/a_menu/shotou/new-cs/1387016.htm`
高等学校 数学編 理数編 `http://www.mext.go.jp/a_menu/shotou/new-cs/1407074.htm`

[12] 西内 啓: 統計学が最強の学問である, ダイヤモンド社, 2013.

[13] 奥村 晴彦: R で楽しむ統計学, 共立出版, 2016.

[14] ジェイソン ローゼンハウス (松浦俊輔 訳): モンティ・ホール問題, 青土社, 2013.

[15] 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.

[16] 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.

[17] 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.

[18] 渡部 洋: ベイズ統計学入門, 福村出版, 1999.