

『基礎統計』を振り返る

れいな(仮)

$$f(x) = \frac{1}{\sqrt{2\pi}\beta} \exp\left(-\frac{(x - \alpha)^2}{2\beta^2}\right)$$

まえがき

この TeX ノートを編集し始めたとき、国会では「統計不正問題」について、毎日野党が与党を追求していた。いろんな統計の不正疑惑が出てきている。統計の不正疑惑は大問題だと野党議員は毎日言っているが、野党議員は統計学についてしっかりと理解できているのか。まあ、与党議員もしっかりと統計学を理解できている人は多くはないだろう。神戸市議会議員の橋本健との不倫疑惑を報じられた今井絵里子議員は絶対理解していないだろう。

統計学については、現在の高校の学習指導要領では、数学 1A の「データの分析」ぐらいしか習わない。けれど、センター試験では 6 ページも問題がある。高校数学の新学習指導要領では、数学 B で統計学が必修となり、「仮説検定」や「区間推定」といった現在の数学 B の選択分野の統計学よりも高度な内容を習う。その代わりに、現在数学 B で習う「ベクトル」が数学 C にはね飛ばされるらしい。機械学習やビッグデータの活用など統計学がより重要になってきたから統計学を数学 B で必修化されたと見ることもできるらしい。でも、「ベクトル」が数学 C にいっちゃった。

大学で習う基礎的な統計学では、時々微分や積分を使用する。そこからさらに発展させた内容では、高度な解析学の他に、線形代数の知識も必要になる。真に統計学を理解するには、高校までの内容では無理があり、中途半端になってしまう気がする。そうなら、わざわざ高校数学で必修単元とする必要はあるのか。

そんな統計学の授業『基礎統計』についてまとめたのがこの TeX ノートである。この TeX ノートは、2018 年度 S セメスター中に、私が作成した『基礎統計』(廣瀬教員) の授業の内容をまとめた TeX ノートを編集し直したものです。2018 年の S セメスターの時は、廣瀬教員作成のレジュメと、『入門統計解析』を中心を作成した。

2019 年春、大学 2 年の春休み。短期バイトで出版社の編集補助のアルバイトをした。その際、東京大学の計数工学科に進学予定の学生なら、統計学には精通しているだろうと言われたが、私はそうではなかった。その言葉を機に、大学 2 年の夏学期に作成したノートを整理しながら、確率論や統計学の基礎を復習しようと思い、この TeX ノートを作成することとした。

2019 年夏休みは、参加予定のインターンシップの関係上、確率論や統計学を知っておくと良いと思われる所以、この TeX ノートをまとめながら、ボチボチ復習することとした。その際、『基礎統計』の内容だけでなく、3 年生の S セメスターで受講した『数理手法 4』(測度論的確率論) の講義で学んだこともいかしてまとめることとした。

れいな (仮)
2019 年 8 月 5 日

目 次

第 1 章 統計解析とは？	7
1.1 Abema TV で紹介された統計トリック	7
1.2 母集団と標本	9
1.3 データの分類	10
1.3.1 データの次元	10
1.3.2 データの型	10
1.3.3 質的データと量的データ	11
第 2 章 1 次元データの分析	13
2.1 1 次元データの整理	13
2.1.1 度数分布表の作成	13
2.1.2 ヒストグラムの形	15
2.1.3 階級幅の調整	16
2.1.4 層別	16
2.2 データの特徴を表す代表的な値	18
2.2.1 平均値	18
2.2.2 中央値・最頻値・ヒストグラムの形	19
2.3 データ分布の散らばり	19
2.3.1 分散	19
2.3.2 標準偏差	20
2.3.3 範囲と四分位偏差	21
2.3.4 度数分布表と原データから求める種々の量	21
2.3.5 変動係数	22
2.3.6 データの変換	23
2.3.7 基準化変量	23
2.4 データ分布の形状の指標を示す	24
2.4.1 歪度	24
2.4.2 尖度	25
第 3 章 2 次元データの分析	29
3.1 2 次元データの整理と散らばり	29
3.1.1 散布図と相関関係	29
3.1.2 共分散	30
3.1.3 相関係数	32
3.2 相関に関する注意点 (みかけ上の相関)	35
3.3 回帰モデル	35
3.4 最小 2 乗法による回帰直線の推定	36
3.4.1 最小 2 乗法とは?	36
3.4.2 最小 2 乗推定値の導出	36

3.4.3 「大学進学率」と「平均給与」の関係に関する回帰直線	37
3.5 決定係数	38
3.5.1 予測値と残差	38
3.5.2 決定係数	39
3.5.3 決定係数と相関係数	41
第 4 章 確率モデル	43
4.1 身近な確率	43
4.1.1 降水確率	44
4.1.2 テレビの視聴率	45
4.2 確率モデルの基礎	45
4.2.1 標本空間と事象	46
4.2.2 事象の演算	47
4.2.3 確率の定義	47
4.2.4 確率の基本公式	49
4.3 条件つき確率	50
4.3.1 条件つき確率のイントロ	50
4.3.2 条件つき確率	51
4.3.3 全確率公式	53
4.3.4 ベイズの定理	53
4.3.5 事前確率と事後確率	55
4.4 事象の独立性	56
第 5 章 離散型確率分布の性質	59
5.1 離散型確率変数	59
5.1.1 確率変数	59
5.1.2 離散型確率変数	60
5.2 期待値	61
5.3 確率変数の散らばり	63
5.3.1 分散と標準偏差	63
5.3.2 基準化変量	64
5.4 ベルヌーイ試行と 2 項分布	65
5.4.1 ベルヌーイ試行	65
5.4.2 2 項分布	65
5.5 ポアソン分布	67
5.6 幾何分布	70
5.6.1 無限級数に関する公式	70
5.6.2 幾何分布の平均と分散	72
5.6.3 幾何分布の性質	73
第 6 章 連続型確率分布の性質	75
6.1 連続型確率分布の導入	75
6.1.1 連続型確率変数とは?	75
6.1.2 密度関数と分布関数	77
6.1.3 連続型確率変数の例題 (1)	78
6.1.4 連続型確率変数の平均と分散	79

目次	5
6.2 一様分布	80
6.3 正規分布	82
6.3.1 Gauss 積分	82
6.3.2 正規分布の定義	83
6.3.3 標準正規分布	85
6.3.4 正規分布の歪度と尖度	88
6.4 指数分布	88
6.5 確率変数の変換	91
6.5.1 離散型確率変数の変換	91
6.5.2 連続型確率変数の変換	91
参考文献およびデータの引用元	93

第 1 章 統計解析とは？

1.1 Abema TV で紹介された統計トリック

Google で「統計不正問題 わかりやすく」と検索した。すると、AbemaTV の動画 (1/27 の Abema 的ニュースショー) が出てきた。この section では、AbemaTV の動画の要旨を追いかながら、TV で紹介されるレベルの「統計」についてまず理解しよう。「統計不正問題」で野党が激怒しているということもあり、「統計」が(少しだけ)注目されている今日、TV や新聞に出てくる統計データをまとめたグラフの解釈には注意が必要だというのが、メインテーマだった。その前に、AbemaTV で軽く紹介された「統計不正問題」の abstract をまず記す。

- 毎月勤労統計の不正問題が発覚
- その他、22 の基幹統計でもミスが発覚
- 信頼性に欠ける統計データを根拠にして、法整備を進めようとしている。
- 厚労省でヒアリング調査が行われたが、半数近くが身内による調査だったことが判明

「毎月勤労統計」は企業の雇用状態や給与、労働時間を調べたデータである。この TeX ノートでは、データの分析の方法の例を述べる際に、最低賃金を例に考える。この統計データをもとに、労災や失業手当の額が決まるので、統計データは正確で信頼できるものでなければならない。病気や出産などで仕事ができなくなった時、あるいは、何らかの理由で解雇されて仕事を探している時、いくら支給すれば良いのかを決める指標になる。

最初にイントロとして、このようなことが取り上げられた後、専門家が出てきて「統計トリック」の解説を始めた。1つ目は「基準のトリック」だった。専門家は「A 大学では就職率 9 割越えを実現しています」という事を例にトリックを話した。「9 割」というからには、10 割(全体)があるのだが、その全体集合の構成要素は何か。この 9 割に、大学院に進学する人が含まれないのは納得できるだろう。ただ、場合によっては、公務員志望の人を除いた人数かもしれないと専門家は注意を呼びかけている。何を分母にとると割合が最大になるのか、色々な場合を考えて、割合の最大値だけをアピールしているかもしれない。

2つ目に「グラフのトリック」を挙げた。AbemaTV では少年犯罪の発生件数を例に説明されたが、ここでは、「昭和 45 年の交通事故月別死者件数」を例に説明する⁽¹⁾。

⁽¹⁾データは、e-stat という統計データが見られるサイトから引用した。URL は、<https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00130002&tstat=000001032727&cycle=1&year=20180&month=24101210> である。

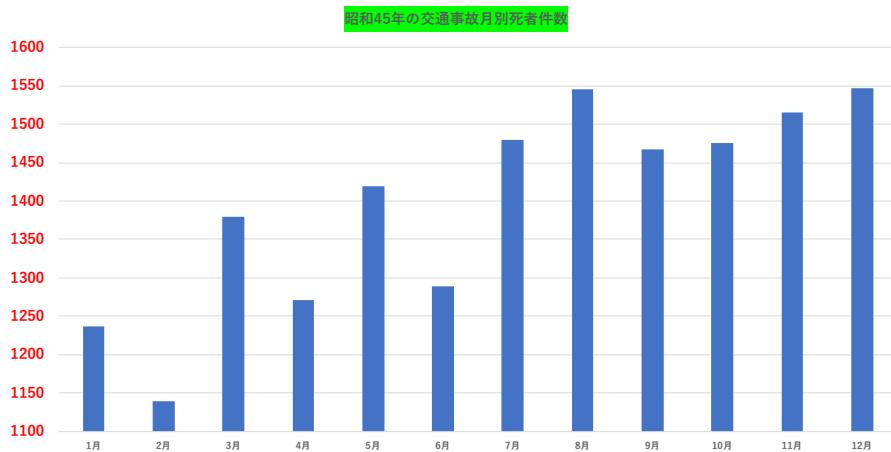


図 1.1: 昭和 45 年の交通事故月別死者件数 (1)

さあ、この図を見て、何を考えるか。8月と12月の死者件数が非常に多いように見える。だからこそ、8月と12月の死者件数が多い理由を考える。8月はなぜ多いのか？グループディスカッションの課題だったら、このように答える人は必ずいるだろう。

免許取り立て、あるいは、免許を取得して数年の大学生が、夏休みに友達とドライブに行った時に事故を起こしてしまうから。

では、12月に事故が多いのはなぜか？

12月に事故が多いという結果だが、実は、その多くは年末に起きている。年末に実家に帰省するときに事故を起こしてしまうのだ。ベテランドライバーでも、普段運転しない道を走行していると、その道の特性を知らないために、事故を起こす可能性が高くなる。…

ここで考えてみよう。8月と12月はそんなに死者件数が多いのか。確かに図 1.1 を見ると、8月と12月がとても多いように見える。逆に、2月は死者件数がとても少ないように見える。でも、縦軸をよく見てほしい。最低値は 0 ではなく、1100 である。縦軸を 0 にとってグラフを作り直すと、以下の図 1.2 のようになる。

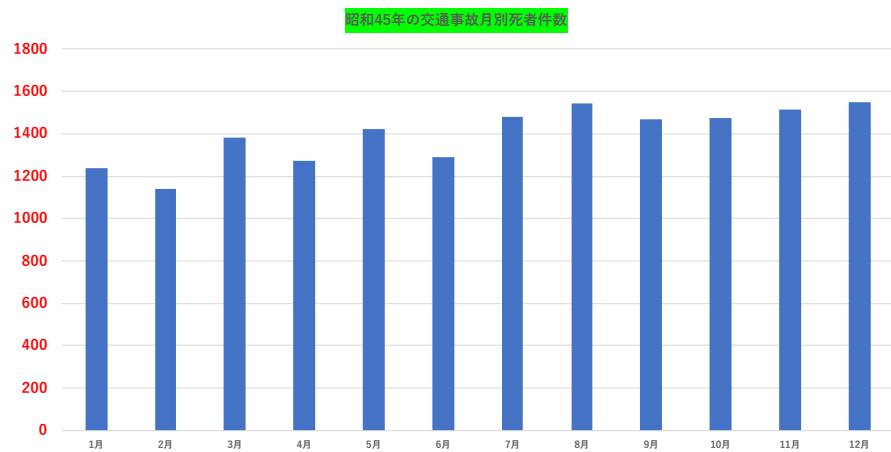


図 1.2: 昭和 45 年の交通事故月別死者件数 (2)

縦軸の最低値を 0 にとらなくても、図 1.1 からわかるように、死者件数が最も少ない 2 月でも 1100 件以上の死者が出る交通事故が起きている。実は 2 月が(交通事故による)死者が特別少ないわけではない。そうではあるが、図 1.1 のようなグラフを見ると、2 月は(交通事故による)死者が極端に少ない月だと感じてしまう。これが「グラフのトリック」である。

専門家は、3 つ目に「人の心が生み出すサバ読みのトリック」を挙げた。AbemaTV で挙げられた例は、高校 3 年生の男子の身長データである。169cm の男子の人数が少ないらしい。ここでいう「少ない」とは、高校 3 年生の男子の身長の平均は 170.7cm であることを考慮すると、もう少し 169cm の男子がいてもおかしくないという意味である。この理由は何だろうか。専門家は、「生徒の中には 169cm ではなく 170cm にしてほしいと言った人がいたのかもしれない」という事を理由の 1 つとして挙げた。まあ、男子にとって、169cm と 170cm は大きな差がある。小さい子が「あと 1cm で遊園地のジェットコースターに乗れたのに」というのと同じである。データの中には恣意的に操作されているものもあるかもしれないということを指摘しているのだろうか。3 つ目のトリックを取り上げた専門家は、複数のデータ、文献などを比較することで、客観的なデータ解釈をする必要があるということを言いたかったのか。

1.2 母集団と標本

- 母集団：実験や調査などにおいて計測や観測の対象となる人やものの集まり
- 個体：母集団に含まれる要素

集団全体の特徴を調べるには、すべての個体について調べるのが理想である。これを全数調査という。しかし、実際に調べるとなると、膨大な時間と費用がかかる。そこで、多くの場合、母集団から一部の個体を取り出して調査を行う。この調査を標本調査という。

- 標本：標本調査の際に選ばれた個体の集まり
- 標本の大きさ：標本として選ばれた個体の数

統計学では、標本調査の標本の大きさとしてどれぐらいにすれば、データとして信頼できるかを数学的に判定する手法などを扱う。この手法に基づき、標本の大きさを適切に設定してから、標本調査は行われる。標本の大きさが全数に近ければ近いほど、得られたデータ分布はより正確な分布に近くなる。国勢調査では、標本の大きさだけではなく、男女、出身都道府県の比率、回答書の年齢分布なども、実際の数に近い方が正確なデータが得られることが予想される。

少し、脇道にそれる。以下は、総合科目 C 系列「法と社会」の講義のレジュメの文章である。今後は、国会議員の構成メンバーが現実の社会を反映するように目指す動きにならい、統計分析のためのデータが多様性を持つことで、現実のモデルを正確に反映できるようにすることが求められるだろう。

国民代表制に関する課題

国民代表制に関して新たに最近指摘されている問題として、ジェンダーやエスニシティの点から見て、国会議員の構成が有権者一般との乖離が激しいことが挙げられる（2018年現在で、日本の女性衆議院議員割合（10.1%），参議院議員割合（20%）である。これらの割合は先進諸国において著しく低い。）。そこでフランスでは2000年以降、憲法改正を行うことによって国政および地方選挙において女性と男性の国会議員の比率を近づける⁽²⁾パリテ政策が積極的に導入されてきた。

パリテ政策の導入の目的として⁽³⁾、女性がもっぱら女性に投票することや、女性が議会でもっぱら女性の利益を主張することは、必ずしも求められていない。そこでは、議会に女性が男性と同数存在すること自体が、女性が日々直面する男性にはない社会的経験を議会の場に持ち込むものとして大きな価値があることだと考えられている。

1.3 データの分類

1.3.1 データの次元

- 1次元データ：各個体につき1つの変数を計測して得られるデータ
ex. 各都道府県における最低賃金（2017年）

2017年度都道府県別最低賃金									
北海道	810	埼玉	871	岐阜	800	鳥取	738	佐賀	737
青森	738	千葉	868	静岡	832	島根	740	長崎	737
岩手	738	東京	958	愛知	871	岡山	781	熊本	737
宮城	772	神奈川	956	三重	820	広島	818	大分	737
秋田	738	新潟	778	滋賀	813	山口	777	宮崎	737
山形	739	富山	795	京都	856	徳島	740	鹿児島	737
福島	748	石川	781	大阪	909	香川	766	沖縄	737
茨城	796	福井	778	兵庫	844	愛媛	739		
栃木	800	山梨	784	奈良	786	高知	737		
群馬	783	長野	795	和歌山	777	福岡	789		

- 2次元データ：各個体につき2つの変数を計測して得られるデータ
ex. 小学生の身長と体重の測定結果を分析

1.3.2 データの型

- 離散型：とびとびの値しかとりえない変数やデータ
ex. 交通事故死者数のデータ（平成18年から平成27年）

(2)方法の例として、比例代表の名簿の順番を工夫して、男女同数に近づけるように努めることが挙げられる。

(3)社会の意見が忠実に反映されるようにする「社会学的代表説」をうまく実現するのは難しい。フランス社会における白人と黒人の比率と、フランス国会の白人と黒人の比率を一緒にすることは困難である。他にも、フランスの全人口のアジア系、アフリカ系、ヨーロッパ系の比率と国会議員のアジア系、アフリカ系、ヨーロッパ系の比率を一緒にすることは厳しい。しかし、男女の比率を一緒にすることはこれらに比べて容易であるといえる。白人や黒人、アジア系かヨーロッパ系などに関係なく、男女比は1:1なので、国会議員の比率を1:1にすることを目指せば良い。人種で人間を分けると人種差別の問題が関係してくるが、性別で人間を分けた時にそのような問題は発生しない。男女平等は1つの達成すべき目標であるために、国会内の男女平等を目指すことは悪くはない。ただ、最近ではLGBTなど「性」の多様化が加速していて、このような新しい問題に対して、どう対処していくかが今後の課題である。

6415, 5796, 5209, 4979, 4948, 4691, 4438, 4388, 4113, 4117

(1人単位でのデータで、それ以上細かい単位でのデータはとれない)

- **連續型** : 連續値を取り得るデータ
ex. 健康診断で得られた身長や体重のデータ

1.3.3 質的データと量的データ

- **質的データ** : 数値として観測できず、あるカテゴリーに属していることや、ある状態にあることがわかるだけのデータ。(ex. 満足度を5段階評価した時の1,2,3,4,5)
 - **名義尺度** : 分類や区分を表す変数。データを数値で表すことができず、ダミー変数を導入して数値化する。性別や職種などが該当する。数量の間隔が等間隔でなく、ダミー変数どうしの四則計算に意味はない。
 - **順序尺度** : 順序関係や大小関係のある分類や区分を表す変数。順位づけなどが該当する。順序関係のみ意味があり、四則演算を行うことができない。
- **量的データ** : 何かの量を表すデータ。四則演算が意味をもつ場合がある⁽⁴⁾。
 - **間隔尺度** : 間隔に意味のある変数。「0」は無を表しているのではない。西暦などが該当する。順序情報をもっていて、測定値間の加減演算が可能である。
 - **比尺度** : 感覚だけでなく比率に意味のある変数。物の長さや速さなどが該当する。原点(絶対零)を有する。測定値間の四則演算が可能である。

(4) 「8cmは4cmの2倍である」と「8°Cは4°Cの2倍である」の2つの文について、長さの方は正しく、温度の方は間違っている。よって、長さは比尺度で、温度は間隔尺度である。しかし、「8cmと4cmの差は4cmである」と「8°Cと4°Cの差は4°Cである」の2つの文は両方とも正しい。これは、長さが間隔尺度だからである。このように、(定義から明らかであるが)、比尺度は間隔尺度の条件を満たしている。この時、比尺度は間隔尺度よりも水準が高いという。同様に、順序尺度は名義尺度の条件を満たしている。そのため、順序尺度は名義尺度よりも水準が高い。

第 2 章 1 次元データの分析

2.1 1 次元データの整理

この section の前半では、「各都道府県における最低賃金 (2017 年)」をもとに、1 次元データの整理方法についてまとめる。section の最後には、データを整理する際の注意点として「層別」にふれる。

2017年度都道府県別最低賃金								
北海道	810	埼玉	871	岐阜	800	鳥取	738	佐賀
青森	738	千葉	868	静岡	832	島根	740	長崎
岩手	738	東京	958	愛知	871	岡山	781	熊本
宮城	772	神奈川	956	三重	820	広島	818	大分
秋田	738	新潟	778	滋賀	813	山口	777	宮崎
山形	739	富山	795	京都	856	徳島	740	鹿児島
福島	748	石川	781	大阪	909	香川	766	沖縄
茨城	796	福井	778	兵庫	844	愛媛	739	
栃木	800	山梨	784	奈良	786	高知	737	
群馬	783	長野	795	和歌山	777	福岡	789	

図 2.1: 各都道府県における最低賃金 (2017 年)

2.1.1 度数分布表の作成

n 個のデータから度数分布表を作ることを考える。

n 個のデータを小さい順に並べかえて、順番に x_1, x_2, \dots, x_n とする。 $(x_1 \leq x_2 \leq \dots \leq x_n)$

- データの最小値 x_1 と最大値 x_n を導出する。(小さい順に並べかえる)
- $R = x_n - x_1$ (範囲, range) を参考にして、いくつの階級に分けるか決める。
(階級の数を m とし、第 k 階級 ($k = 1, 2, \dots, m$) の境界値を決める。下限を a_k 、上限を a_{k+1} とする。)
- 各階級の階級値 c_k を求める。階級値は各階級を代表する値のことである。一般的には、各階級の中では観測地は一様に分布していると仮定して、階級の上限と下限の中間の値を階級値とする。

$$c_k = \frac{a_k + a_{k+1}}{2} \quad (2.1)$$

- データを各階級に分類後、集計し、各階級に属する度数 (frequency) $f_k^{(1)}$ を求める。

⁽¹⁾ f_k ($1 \leq k \leq m$) は $\sum_{k=1}^m f_k = n$ を満たす。つまり、度数の総和は標本の大きさ n に等しい。

5. 第*i*階級までの度数の和(累積度数) F_i を求める。

$$F_i = \sum_{k=1}^i f_k \quad (2.2)$$

6. 第*i*階級における度数の全データに占める割合(相対度数⁽²⁾) p_i を求める。

$$p_i = \frac{f_i}{n} \quad (2.3)$$

7. 累積度数の全データ数に占める割合(累積相対度数⁽³⁾) P_i を求める。

$$P_i = \frac{F_i}{n} \quad (2.4)$$

図2.1の「各都道府県における最低賃金(2017年)」のデータから、階級の幅を20円として、度数分布表とヒストグラムを作成すると以下のようになる。(相対度数は四捨五入した値を用いたため、総和をとると1.000にならない。実際、1.001になる。)

階級	階級値	度数	相対度数
720以上 740未満	730	14	0.298
740-760	750	3	0.064
760-780	770	6	0.128
780-800	790	9	0.191
800-820	810	5	0.106
820-840	830	2	0.043
840-860	850	2	0.043
860-880	870	3	0.064
880-900	890	0	0.000
900-920	910	1	0.021
920-940	930	0	0.000
940-960	950	2	0.043
合計		47	1.000

図2.2: 「各都道府県における最低賃金(2017年)」の度数分布表

データを度数分布表にまとめて、データ分布の特徴を把握しづらい。データを適切に把握するためにも、データ分布をグラフとして見やすいようにするのが良い。データ分布を棒グラフで表したものヒストグラムという。図2.3のヒストグラムは、図2.2をもとに作成したヒストグラムである。ヒストグラムを作ると、以下のことが視覚的にわかる。

- ピークの位置と個数(→最頻値の把握)
- データ分布の中心の位置(→平均値の把握)

(2) p_i は f_i の性質から、 $\sum_{i=1}^m p_i = \sum_{i=1}^m \frac{f_i}{n} = \frac{1}{n} \sum_{i=1}^m f_i = \frac{1}{n} \times n = 1$ となる。

(3) データによっては、度数や相対度数より、累積度数や累積相対度数の方を利用した方が良い場合がある。各都道府県における最低賃金については、800円から820円の間の都道府県がいくつあるよりも、800円以下の都道府県はいくつあるか。また、900円以上の都道府県はいくつあるかといったことが重要かもしれない。これは階級幅を大きくした場合を考えていることと等価であるといえよう。この後の2.1.3「階級幅の調整」の内容とも関係しているが、初めから、階級幅を大きくすると(最低賃金の例の場合は「100円刻みでとる」とことなど)、元データの本質的な特徴を見失うかもしれない。注意が必要である。

- データの散らばり具合 (\rightarrow 標準偏差の把握)
- データ分布に歪みや尖りはあるか
- 極端に外れた値があるか

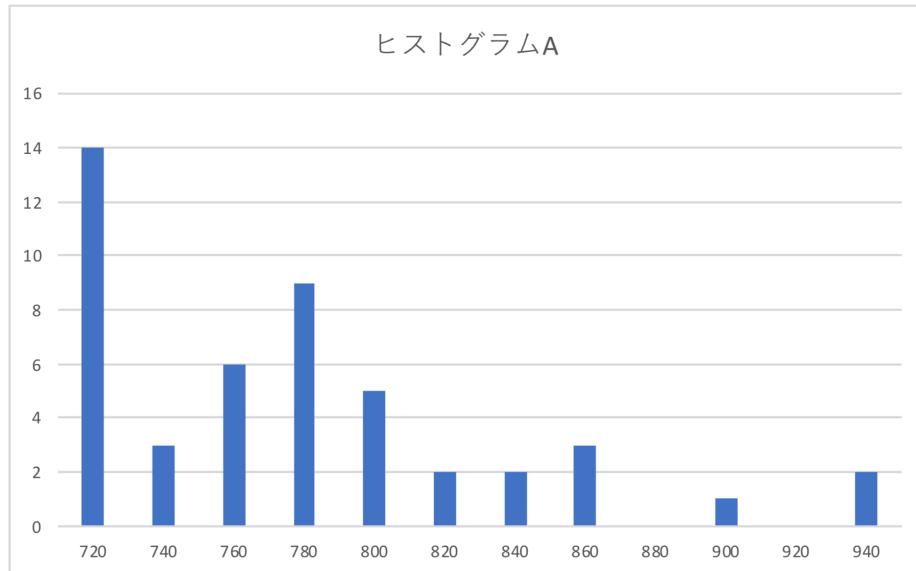


図 2.3: 「各都道府県における最低賃金(2017年)」のヒストグラム(1)

- モード: 度数が最大となる階級の階級値。モードは、度数分布表の作り方に依存する。
 「各都道府県における最低賃金(2017年)」のデータのモードについて、「720円以上740円未満」の階級の度数が14で最大なので、この階級の階級値である730円が、モードである。

2.1.2 ヒストグラムの形

ヒストグラムを作成すると、データ分布の特徴が視覚的にわかることは既に言及した。ヒストグラムの形には主に以下の3パターンがある⁽⁴⁾。

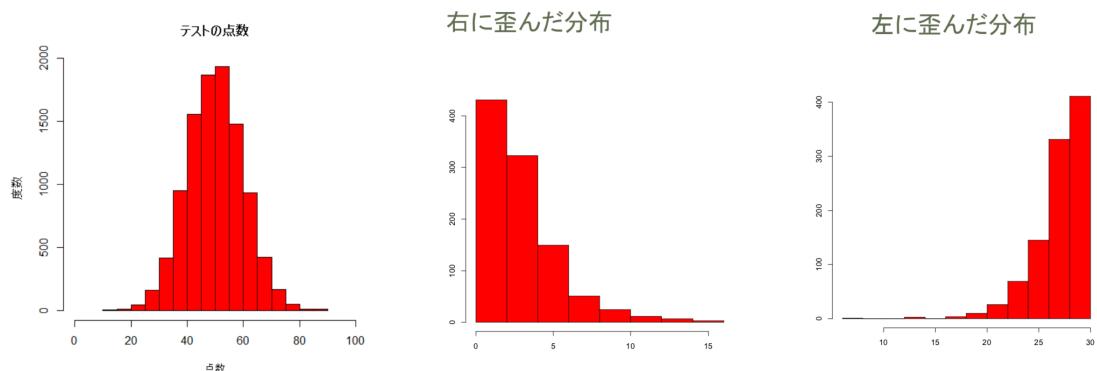


図 2.4: 様々なヒストグラムの形

(4) 「右に歪んだ分布」と「左に歪んだ分布」は逆で覚えててしまうかもしれないが、(覚えるなら、)十分注意する必要がある。

2.1.3 階級幅の調整

階級幅の決め方にはルールがない。データを分析する上で、適切と考えられる階級幅をとればよい。13ページの図2.1から度数分布表を作る事を考えよう。2.1.1では、例として、階級幅を20円として度数分布表を作った。階級幅を5円で作ったらどうなるだろうか。また、階級幅を50円で作ったらどうなるか。階級幅が大きすぎたり、逆に小さすぎたりすると、その後の分析を適切に行えない可能性がある。

一般的に全ての階級の階級幅が等しいのが望ましい。ただ、分布の両端に近い所では、度数が中心付近と比較して極端に小さい場合がある。そのような時は階級の幅を広げたり、最初の度数分布表作成時の階級幅が適切かどうか検討することが必要であるかもしれない。図2.3のヒストグラムで、860円以上のデータを全て一括りにして、ヒストグラムを作成すると以下の図2.5のようになる。（「860円以上960円未満」という階級を新たに作ったことになる。）

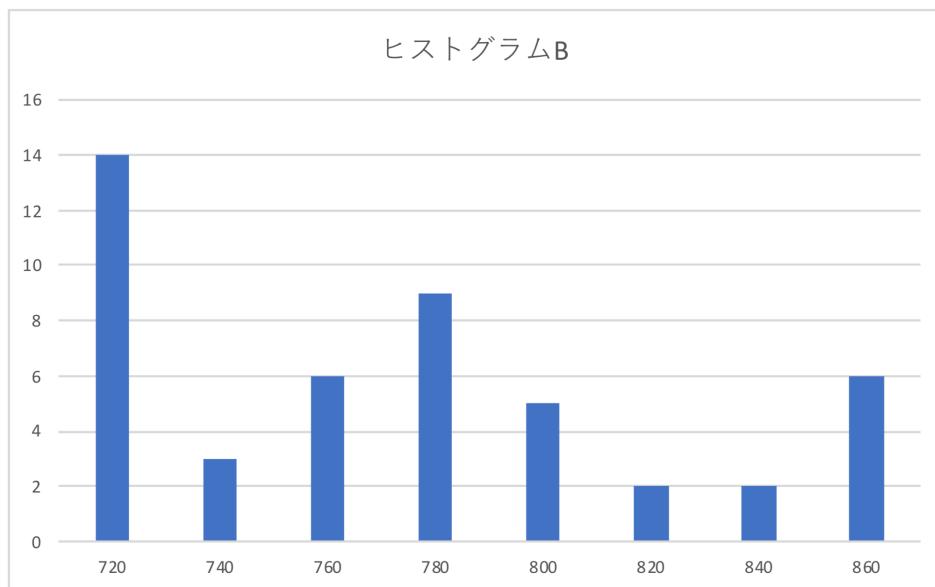


図2.5: 「各都道府県における最低賃金(2017年)」のヒストグラム(2)

2.1.4 層別

次の図2.6は、あるデータをもとに作成したグラフである。グラフの横軸には140～194の数値が書かれている。さらに、特徴として、160と170でピーク値をとることがわかる。このグラフは何を表すデータか？

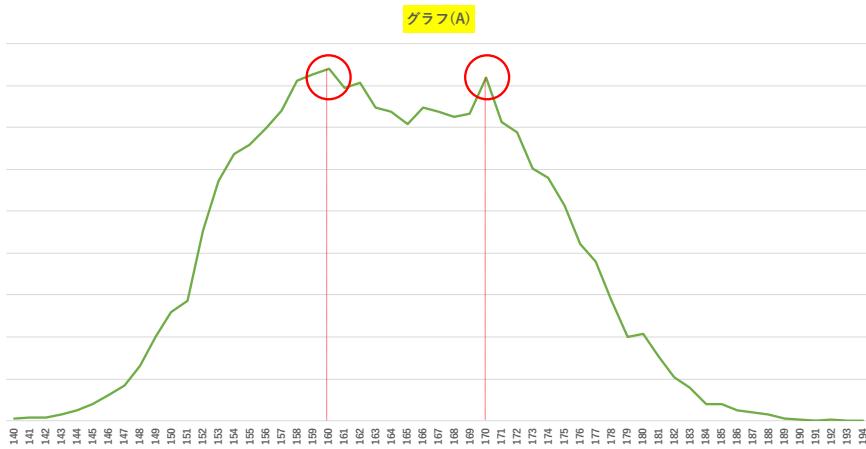


図 2.6: 謎のグラフ : グラフ (A)

実は、図 2.6 は、以下の 2 つの曲線をもとに生み出したグラフである。オレンジのグラフの方程式を $y = w(x)$ 、青のグラフの方程式を $y = m(x)$ とおこう。実は、図 2.6 のグラフ（緑のグラフ）の方程式は $y = w(x) + m(x)$ なのである⁽⁵⁾。

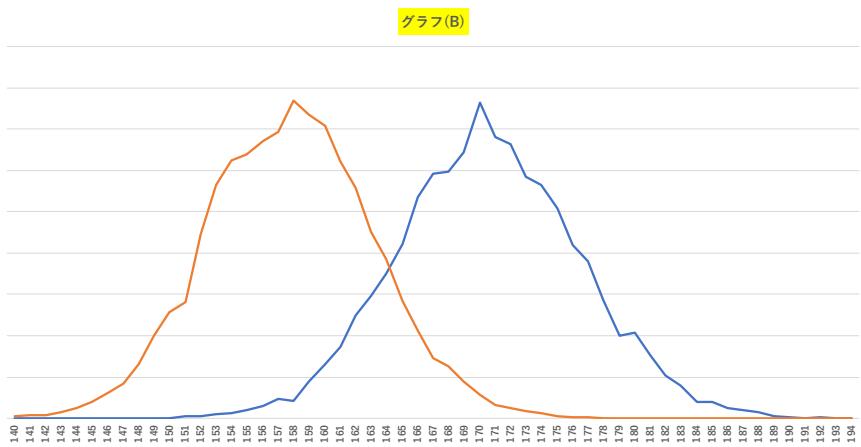


図 2.7: 謎のグラフ : グラフ (B)

今、 $w(x)$ と $m(x)$ と書いたのは、 $w(x)$ の方のグラフは「高校 3 年生の女子の身長分布」、 $m(x)$ のグラフは「高校 3 年生の男子の身長分布」をもとに作成したグラフだからである。ということは、緑のグラフには（実は）意味がない。

さて、この subsection で何を言いたかったのかまとめよう。データ分布をグラフにすると、図 2.6 のようになったとする。ピークが 2 つ以上あるとき、「男性のデータと女性のデータが混在している」などの理由が考えられる。身長分布を調べるとき、男女を区別せず総合的にデータを分析することも必要である。男子だけ、女子だけのデータを知りたい時は、データを区別してピークが 1 つの分布に分離することが求められる。ピークが複数個ある時、データを分離すると、よりデータの特性を理解しやすいことがある。そのためデータを分離することを層別という。

(5) 「 $x = x_i$ におけるオレンジのグラフの値を w_i 、青のグラフの値を m_i とすると、同じ x_i に対して、緑のグラフの値は $w_i + m_i$ なのである」と書いた方が正確かもしれない。

2.2 データの特徴を表す代表的な値

前の section 「1次元データの整理」の前半では、「各都道府県における最低賃金(2017年)」をもとに、1次元データの整理方法をまとめた。特に、度数分布表やヒストグラムを用いた整理方法について書いた。この section では、数値によりデータの特徴を表すことを考える。このデータの特徴を表す代表的な値を代表値という。

この section では、 n 個のデータを x_1, x_2, \dots, x_n と書いた時、既に n 個のデータを小さい順に並べかえてあるものとする。すなわち、 $x_1 \leq x_2 \leq \dots \leq x_n$ であるとする。

2.2.1 平均値

データ x_1, x_2, \dots, x_n に対して、平均値 \bar{x} を以下のように定める。

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k \quad (2.5)$$

この平均値についていくつかの特徴がある。

- 原データと同一の単位を持つ。
- 全てのデータに等しい重みを与えるため、外れ値の影響を受けやすい。
(→ 外れ値の影響を受けないようにする方法はいくつかある。一番単純な方法は、明らかに異常値と思われるデータを無視して考えることである⁽⁶⁾。異常値を無視した時、データの総数を適切に補正すれば問題ない。)
- 平均値は最小二乗値である。

最小二乗値と平均値

与えられたデータ x_1, x_2, \dots, x_n に対して、各データとある定数 c との差の2乗の値の総和 $f(c)$ を考える。

$$f(c) = \sum_{k=1}^n (x_k - c)^2 \quad (2.6)$$

この時、 $f(c)$ を最小にする c の値は、 n 個のデータの平均値 \bar{x} である。

証明

$A = \sum_{k=1}^n (x_k)^2, B = \sum_{k=1}^n x_k$ とおく。すると、 $\bar{x} = \frac{B}{n}$ となる。

$$\begin{aligned} f(c) &= \sum_{k=1}^n (x_k)^2 - 2c \left(\sum_{k=1}^n x_k \right) + nc^2 \\ &= nc^2 - 2Bc + A \\ &= n \left(c - \frac{B}{n} \right)^2 + A - \frac{B^2}{n^2} \\ &= n(c - \bar{x})^2 + A - \frac{B^2}{n^2} \end{aligned}$$

⁽⁶⁾ 異常値を無視できるのは、データの総数が十分大きく、1つぐらいデータを無視しても全体の様子が把握できなくなるということが起こらないと予想される場合に限る。

となるから、 $c = \bar{x}$ の時、 $f(c)$ は最小。 □

2.2.2 中央値・最頻値・ヒストグラムの形

データ分布の特徴を表す平均値は外れ値の影響を受けやすいので、代表値として不適切な場合がある。このような時は代表値として中央値（メディアン、median）や、最頻値（モード、mode）が用いられる⁽⁷⁾。データを小さい順に並べたときに中央に位置する値を中央値という。小さい順に並べた時の両端にあることが多い外れ値を使用しないので、外れ値の影響を受けにくい。中央値 Md は以下のように定義される。

- n が奇数のとき : $Md = x_{(\frac{n+1}{2})}$
- n が偶数のとき : $Md = \frac{1}{2} \left\{ x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} \right\}$

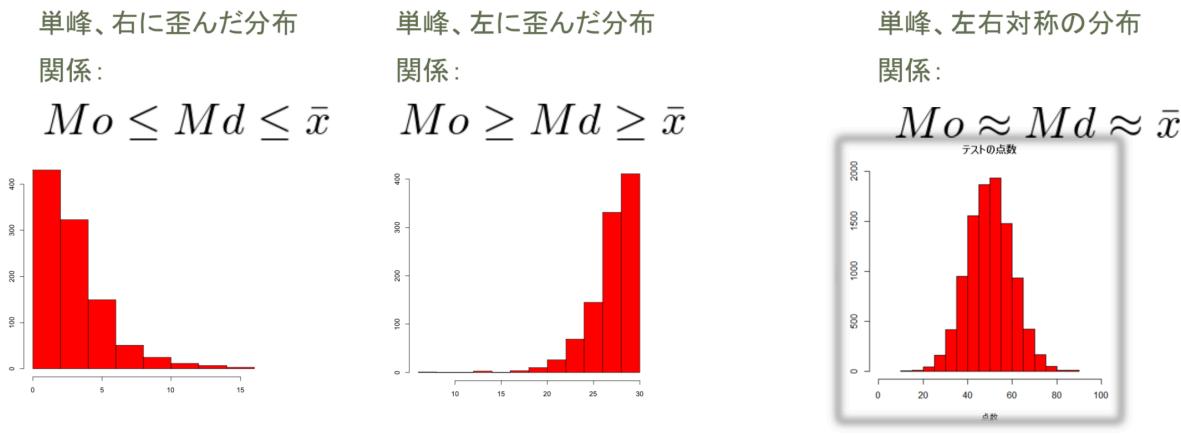


図 2.8: 単峰の時の平均値・中央値・最頻値の関係

2.3 データ分布の散らばり

データの特徴を正確に把握するには、データの特徴を示す代表値の他にも、データがどのように分散しているかを示す値も重要である。散らばりを表す指標として、分散（variance）、標準偏差（standard deviation）、変動係数などが使われる。

2.3.1 分散

データ x_k と平均 \bar{x} の差 $x_k - \bar{x}$ を偏差という。この偏差を利用して、分散 S^2 が定義される。

$$S^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 \quad (2.7)$$

分散は、平均値と各データの乖離度を利用した数値であり、散らばりを表す 1 つの指標である。分散の単位は原データの 2 乗になるので、解釈には注意する必要がある。

⁽⁷⁾ 最頻値（モード）については 15 ページを参照のこと

分散の2つ目の定義

分散の定義式を変形することで、もう一つの定義式が導ける。

$$\begin{aligned}
 S^2 &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n (x_k)^2 - 2\bar{x} \left(\frac{1}{n} \sum_{k=1}^n x_k \right) + \frac{1}{n} \cdot \sum_{k=1}^n (\bar{x})^2 \\
 &= \frac{1}{n} \sum_{k=1}^n (x_k)^2 - 2(\bar{x})^2 + (\bar{x})^2 \\
 &= \frac{1}{n} \sum_{k=1}^n (x_k)^2 - (\bar{x})^2 \\
 &= \bar{x}^2 - (\bar{x})^2
 \end{aligned} \tag{2.8}$$

分散を求める2つの公式

分散を求める公式としては、定義式(2.7)と、そこから派生してできる式(2.8)がある。状況に応じて使い分けると議論が楽になる。

- $S^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2$
- $S^2 = \bar{x}^2 - (\bar{x})^2$

2.3.2 標準偏差

分散は、単位が原データの2乗であるため、解釈がしづらい場合がある。そこで、分散の正の平方根である標準偏差 S が、原データと同一の単位をもつので、解釈の際にはよく使われる。

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2} \tag{2.9}$$

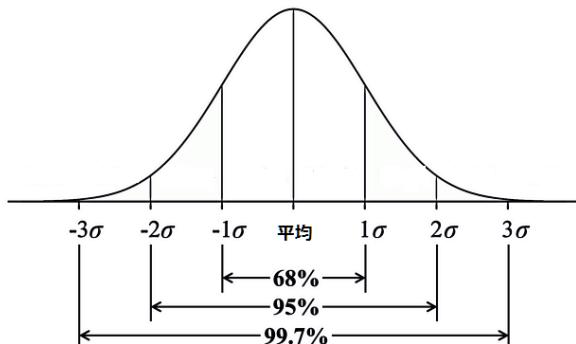
$$= \sqrt{\frac{1}{n} \{ \bar{x}^2 - (\bar{x})^2 \}} \tag{2.10}$$

標準偏差とシグマ区間

標準偏差 S に対して、 $\bar{x} - kS$ から $\bar{x} + kS$ までの区間を k シグマ区間という。データ分布が右の図のような左右対称な釣鐘型（正規分布）ならば、

- 1シグマ区間には全データの 68.3%
- 2シグマ区間には全データの 95.4%
- 3シグマ区間には全データの 99.7%

が含まれることが知られている。



2.3.3 範囲と四分位偏差

分散や標準偏差は、各データ x_i と平均 \bar{x} の差 $x_i - \bar{x}$ に基づいて散らばりを考えるものであった。これは、分散の定義式からもわかるように、計算がとても面倒である。そこで、散らばりの別の評価方法を考える。その評価方法が範囲 (range) と四分位偏差を用いる評価方法である。どちらも小さい順に並べればすぐにわかる値で、散らばりを表す指標としては単純なものである。

- データの最大値 x_n と最小値 x_1 の差 $x_n - x_1$ を範囲 (レンジ) という。
- レンジは散らばりを表す指標としては最も素朴なもので、最大値と最小値しか用いない。
(\Rightarrow そのため、外れ値の影響を受けやすい。)

最大値と最小値しか使わないと、外れ値の影響を受けやすく、散らばりの表現としては不適切な場合が多い。そこで、より中央値に近い値を使って(両端を無視して)、散らばりを評価することを考える。データを順番に並べた時の下位 25%、75% に位置する値を使う。

- 下位 25% に位置する値を第 1 四分位点 Q_1 、下位 50% に位置する値を第 2 四分位点 Q_2 、下位 75% に位置する値を第 3 四分位点 Q_3 という。第 2 四分位点 Q_2 は、中央値のことである。
- Q_1 と Q_3 に対し、 $\frac{Q_3 - Q_1}{2}$ を四分位偏差という。

2.3.4 度数分布表と原データから求める種々の量

資料によっては、度数分布表のみ利用できて、原データが利用できない場合がある。そのような時は、度数分布表から、平均値や分散、標準偏差の値を見積もることが可能である。以下の表(図 2.9)は、47 都道府県の最低賃金⁽⁸⁾について、「原データから求めた値」と「度数分布表から見積もった値」を比較したものである。

	原データから求めた値	度数分布表から見積もった値
平均値	790 円	788 円
中央値	781 円	780 円
最頻値	737 円	730 円
分散	3152	3289
標準偏差	56.14 円	57.35 円

図 2.9: 度数分布表と原データから求める種々の量の比較

度数分布表から見積もる時は、各階級の値は全て階級値に等しいとみなして計算する。47 都道府県の最低賃金に関する度数分布表は、14 ページの図 2.2 の通りである。その最初の部分を取り出すと以下のようである。

階級	階級値	度数	相対度数
720 以上 740 未満	730	14	0.298
740-760	750	3	0.064

(8) 最低賃金に関するデータは、13 ページの図 2.1 を参照。

今回の場合、「720円以上740円未満」の14個のデータを全て階級値730円に等しいとみなすことにする。また、「740円以上760円未満」の3個のデータを全て階級値750円に等しいとみなすことにする。このように階級値を利用することで、平均値や分散、標準偏差を見積もることができる。実際、近似値であるが、真値に近いことが図2.2からわかる。具体的なデータを使うより計算が楽なので、厳密な値を必要としないなら、階級値を利用して大体の平均値や標準偏差を求めればよいだろう。

2.3.5 変動係数

異なる2つのデータの散らばりを比較する際は、標準偏差を用いることが必ずしも適当でないことがある。

	数学の平均点	標準偏差
2017年駿台東大実戦(夏)	26.5	16.1
2017年駿台東大実戦(秋)	24.6	15.6
2017年河合塾東大オープン(夏)	30.3	20.7
2017年河合塾東大オープン(秋)	40.7	23.8

図2.10: 2016年度に実施された東大模試の数学の結果

上の表(図2.10)を見ると、2017年の東大オープン(秋)の標準偏差は一番大きく、2017年の駿台東大実戦(秋)の標準偏差が一番小さい。だからといって、2017年の東大オープン(秋)の散らばりは一番大きいといって良いだろうか。

異なるデータの散らばりを比較する時は、標準偏差 S を平均値 \bar{x} で割った変動係数 (coefficient of variation) が使われる。

$$CV = \frac{S}{\bar{x}} \quad (2.11)$$

変動係数は、単位を持たないため、異なる単位を持つデータの散らばり具合を比較することも可能である。また、平均の中心が著しく異なる場合(「日本と世界」など)の分布の散らばり具合を相対的に比較するのにも使える。上のデータについて変動係数を計算すると以下のようになる。

	数学の平均点	標準偏差	変動係数
2017年駿台東大実戦(夏)	26.5	16.1	0.608
2017年駿台東大実戦(秋)	24.6	15.6	0.634
2017年河合塾東大オープン(夏)	30.3	20.7	0.683
2017年河合塾東大オープン(秋)	40.7	23.8	0.585

変動係数を使うと、2017年の東大オープン(秋)の散らばりは、実は一番小さいということがわかる。

2.3.6 データの変換

データの変換

a, b を定数とする。与えられたデータ x_1, x_2, \dots, x_n を、 $y_k = ax_k + b$ と変換する時、データ y_1, y_2, \dots, y_n の平均や分散は以下のようになる。

$$\bar{y} = a\bar{x} + b \quad (2.12)$$

$$S_y^2 = a^2 S_x^2 \quad (2.13)$$

証明

実際に丁寧に計算することで導ける。

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{k=1}^n y_k = \frac{1}{n} \sum_{k=1}^n (ax_k + b) = a \cdot \frac{1}{n} \sum_{k=1}^n x_k + \frac{1}{n} \cdot nb \\ &= a\bar{x} + b \\ (S_y)^2 &= \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 = \frac{1}{n} \sum_{k=1}^n \{(ax_k + b) - (a\bar{x} + b)\}^2 \\ &= \frac{1}{n} \sum_{k=1}^n a^2(x_k - \bar{x})^2 \\ &= a^2 \cdot \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = a^2 \cdot (S_x)^2\end{aligned}$$

□

2.3.7 基準化変量

データ分布全体の中で特定のデータがどのような位置にあるか知るときに、模試などでは偏差値が使用される。「偏差値」という指標は、基準化変量の考えが利用されている。

基準化変量

データ x_1, x_2, \dots, x_n の平均を \bar{x} 、標準偏差を S とする。この時、

$$z_k = \frac{x_k - \bar{x}}{S} \quad (2.14)$$

で定まる量 z_k をデータ x_k の基準化変量 (standardized data) という。

式 (2.14) を変形すると、 $x_k = \bar{x} + z_k S$ とかけるので、基準化変量 z_k はデータ x_k が平均 \bar{x} から標準偏差いくつ分ずれているかを表す数であることがわかる。また、データ x_k に対して、 z_k を求めることを x_k を基準化するという。

基準化変量の平均と分散

データ x_1, x_2, \dots, x_n の平均を \bar{x} 、標準偏差を S とする。

この時、以下の式で定まる基準化変量の平均は 0、分散は 1 である。

$$z_k = \frac{x_k - \bar{x}}{S}$$

証明

元のデータ x_k を $\frac{1}{S}$ 倍して、 $\frac{\bar{x}}{S}$ 引けば z_k になるので、

$$\begin{aligned}\bar{z} &= \frac{1}{S} \cdot \bar{x} - \frac{\bar{x}}{S} = 0 \\ (S_z)^2 &= \left(\frac{1}{S}\right)^2 \cdot S^2 = 1\end{aligned}$$

□

受験生がとても気にする偏差値とは、テストの点数などを平均点を 50 点、標準偏差を 10 点になるように変換したものである。データの変換公式を考えると、基準化変量 z_k を 10 倍して、50 足せば偏差値が求められる。つまり、データ x_k に対して、その偏差値 t_k は

$$t_k = 10z_k + 50 = 10 \times \frac{x_k - \bar{x}}{S} + 50 \quad (2.15)$$

とかける。例えば、2017 年駿台東大実戦(夏)の数学(平均点:26.5 点、標準偏差 16.1 点)で 60 点を取った人の数学の偏差値は、

$$10 \times \frac{60 - 26.5}{16.1} + 50 = 70.81$$

である。

2.4 データ分布の形状の指標を示す

2.1 「1次元データの整理」では、データの整理方法として、度数分布表やヒストグラムを取り上げた。その後に 2.2 と 2.3 で、データの特徴を表す 2 つの指標として、まず代表値を取り上げ、次に標準偏差や分散といった散らばりを表す指標を取り上げた。この section では、データの特徴を表す新たな指標として、歪みや尖り具合を表す指標である歪度と尖度を取り上げる。

この Section では、データを x_1, x_2, \dots, x_n 、平均を \bar{x} とし、分散を S^2 とする。また、各データを基準化した値を z_1, z_2, \dots, z_n とする。

2.4.1 歪度

歪度

与えられたデータ x_1, x_2, \dots, x_n に対して、それらを基準化したものを z_1, z_2, \dots, z_n とする。この時、基準化変量の 3 乗の平均を歪度 (skewness) という。

$$b_1 = \frac{1}{n} \sum_{k=1}^n (z_k)^3 = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{S} \right)^3 \quad (2.16)$$

歪度は、歪みが強くなるほど、その絶対値が大きくなる。歪みが小さいほど、歪度の絶対値は 0 に近くなる。

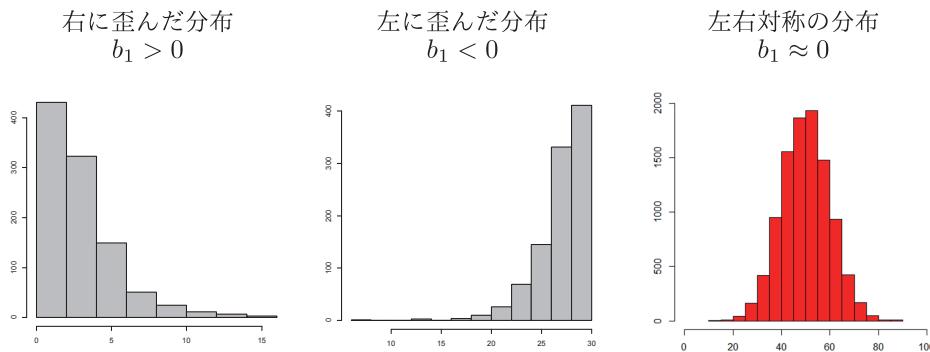


図 2.11: データ分布の歪みと歪度の関係

2.4.2 尖度

尖度

与えられたデータ x_1, x_2, \dots, x_n に対して、それらを基準化したものを z_1, z_2, \dots, z_n とする。この時、基準化変量の 4 乗の平均を尖度という。

$$b_2 = \frac{1}{n} \sum_{k=1}^n (z_k)^4 = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{S} \right)^4 \quad (2.17)$$

尖度は、データ分布が釣鐘型（正規分布）に近いかどうかを計る指標である。正規分布では、尖度は 3 になることが知られている。

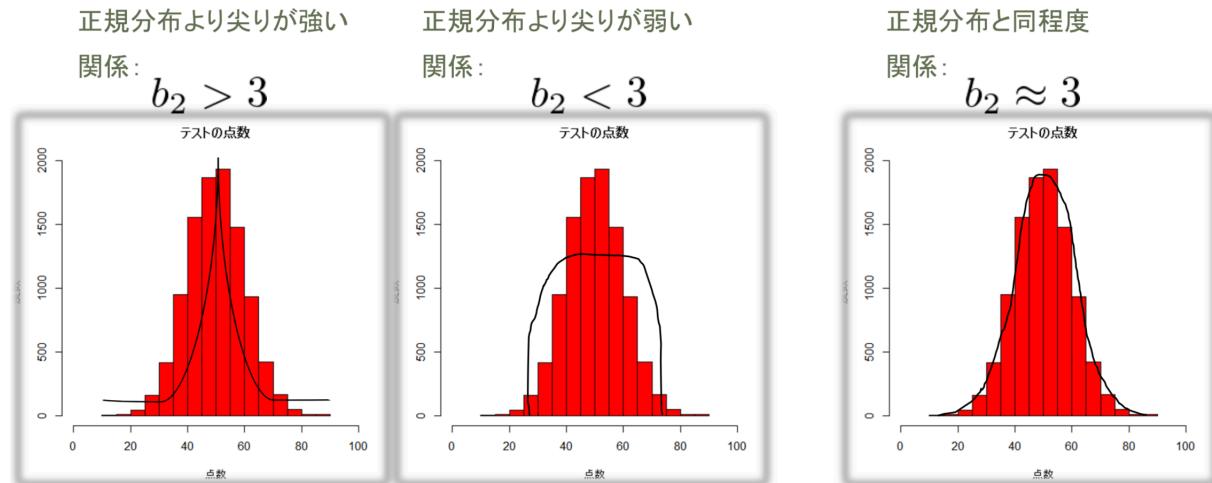


図 2.12: データ分布の尖り具合と尖度の関係

さて、第1回、第2回で扱った各都道府県ごとの最低賃金に関するデータについて歪度と尖度を計算すると、歪度は1.31、尖度は4.38と求められる。歪度が1より大きいので、ヒストグラムは右に歪んだ形になると予想できる。実際にヒストグラムを作ると、予想通りの結果が得られる。

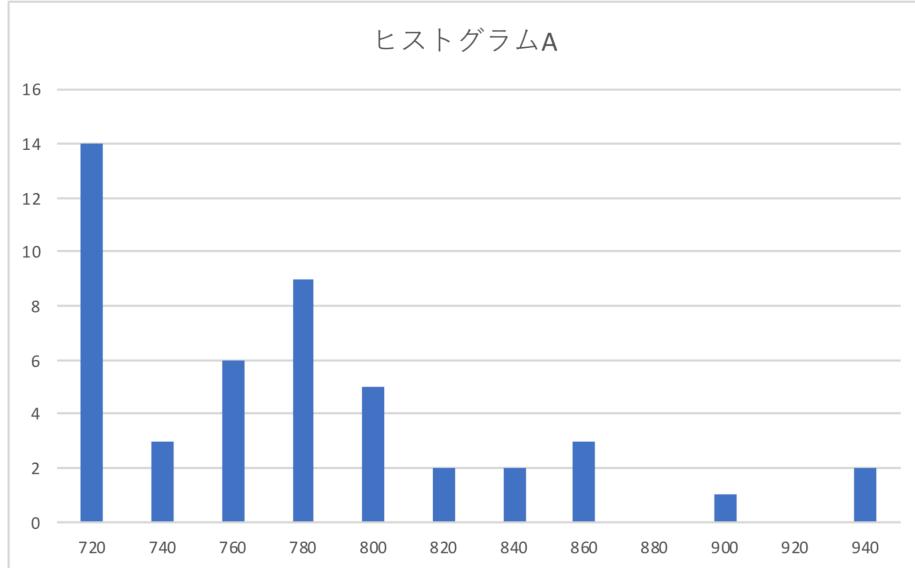


図 2.13: 「各都道府県における最低賃金(2017年)」のヒストグラム(1) / (図 2.3と同じ)

この subsection の最後に、式 (2.17) で定義した尖度 b_2 の大小で、データ分布の尖り具合を表せる理由を考える。結論を書くと、 $y_k = (z_k)^2$ とおくと

$$b_2 = \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 + 1 \quad (2.18)$$

が成立するからである。

証明

式 (2.18) の右辺を変形して左辺と一致することを確認する。

$$\begin{aligned} b_2 &= \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2 + 1 = \frac{1}{n} \sum_{k=1}^n (y_k)^2 - 2 \cdot \frac{1}{n} \sum_{k=1}^n y_k \cdot \bar{y} + \frac{1}{n} \sum_{k=1}^n (\bar{y})^2 + 1 \\ &= \frac{1}{n} \sum_{k=1}^n (y_k)^2 - 2 \cdot \bar{y} \cdot \underbrace{\frac{1}{n} \sum_{k=1}^n y_k}_{\sim\!\sim\!\sim\!\sim} + (\bar{y})^2 + 1 \\ &= \frac{1}{n} \sum_{k=1}^n (y_k)^2 - 2(\bar{y})^2 + (\bar{y})^2 + 1 \\ &= \frac{1}{n} \sum_{k=1}^n (y_k)^2 - (\bar{y})^2 + 1 \end{aligned} \quad (2.19)$$

ここで、基準化変量 z_k の平均が0、分散が1であることから、

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n (z_k)^2 = \frac{1}{n} \sum_{k=1}^n \underbrace{(z_k - \bar{z})^2}_{\sim\!\sim\!\sim\!\sim} = 1 \quad (2.20)$$

となる。ゆえに、式(2.19)と式(2.20)より、

$$b_2 = \frac{1}{n} \sum_{k=1}^n (y_k)^2 = \frac{1}{n} \sum_{k=1}^n (z_k)^4$$

となることがわかる。 \square

以上より、尖度は、 y_1, y_2, \dots, y_n の分散に 1 を足した値とも言えるので、尖度 b_2 は、基準化変量の 2 乗 $y_k = (z_k)^2$ の散らばりを表す指標ともいえる。証明でも使った「基準化変量 z_k の分散が 1」であることから、

$$y_1 + y_2 + \dots + y_n = n \quad (2.21)$$

となる。 y_k は 0 以上の値をとるから、もし、ある y_i が n に十分近いなら、残りの y_k は限りなく 0 に近い。これは、 z_k の定義も考えると、 x_k が平均 \bar{x} に限りなく近いことと同値である。つまり、多くのデータが平均に近くなるので、尖り具合が強くなるといえる。したがって、 y_i が n に近いとき、尖度 b_2 も大きくなるので、尖度の大小で尖り具合を表すことができることがわかる。

第3章 2次元データの分析

3.1 2次元データの整理と散らばり

このSectionでは、2つの変数(x_k, y_k)の関係を表し、分析する方法を考える。第2章では最低賃金を例に分析方法をまとめたが、この第3章では「各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)」の関係を題材に、2次元データの分析手法をまとめる。

都道府県	大学進学率	平均給与						
北海道	42.8	266.4	石川	49.2	276.7	岡山	46.2	269.6
青森	37.2	234.8	福井	47.8	272.3	広島	53.9	297.6
岩手	37.3	236.8	山梨	56.4	279.9	山口	37.7	273.5
宮城	46.6	284.5	長野	43.5	275	徳島	46.2	274.6
秋田	37.6	240.8	岐阜	45.2	277.9	香川	47.3	277.7
山形	38.6	246.7	静岡	47.7	290.8	愛媛	45.5	262.7
福島	39.5	261.4	愛知	52.4	318.3	高知	40.8	258.3
茨城	51.3	299.8	三重	44	300	福岡	47.4	282.7
栃木	48.4	294.9	滋賀	48.2	295.8	佐賀	38.5	246.6
群馬	47	282.4	京都	65.2	311.6	長崎	39	253.4
埼玉	51.4	296.5	大阪	56.2	326.1	熊本	41.5	253.8
千葉	53.2	309.4	兵庫	54	294.8	大分	36.6	257.8
東京	72.7	377.5	奈良	56.1	300.7	宮崎	38.2	235.5
神奈川	54.4	329.8	和歌山	44	282.2	鹿児島	35.8	249.2
新潟	42.3	260.1	鳥取	39.3	254.2	沖縄	36.7	244.4
富山	44.8	267.6	島根	39.9	253.4			

図 3.1: 各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)

3.1.1 散布図と相関関係

2つの変数の関係を表す方法の1つに、データを平面上にプロットする方法がある。以下のような図を散布図という。

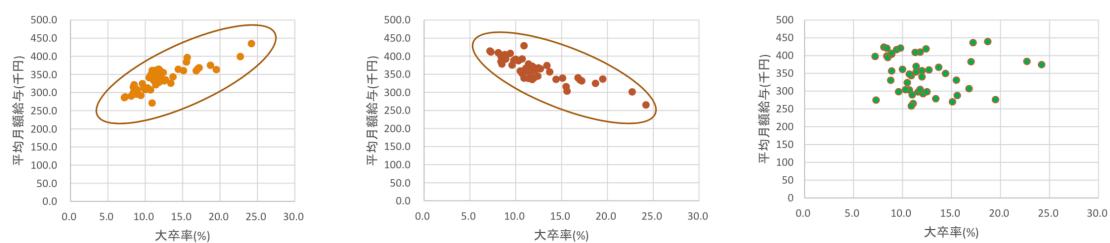


図 3.2: 散布図の例

図3.2を見てほしい。2次元データの分布については3種類あることがわかる。

- 正の相関：正の傾きを持つ直線のまわりにデータが集まっている。
- 負の相関：負の傾きを持つ直線のまわりにデータが集まっている。
- 無相関：直線的関係を持たない

では、各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)」の関係は、3種類のうち、どれに該当するだろうか。高学歴の人ほどたくさんの給料をもらえる可能性が高いのが、現在の日本社会である。このことをふまえると、大学進学率が高い都道府県の平均給与は高くなることが予想される。つまり、右上がりの散布図になることが予想される。

実際に、図3.1の表のデータを散布図にすると、以下のようにになる。縦軸に平均給与(千円)、横軸に大学進学率(%)をとった。図3.2の一番左の右上がりの散布図と同じようであることがわかる。

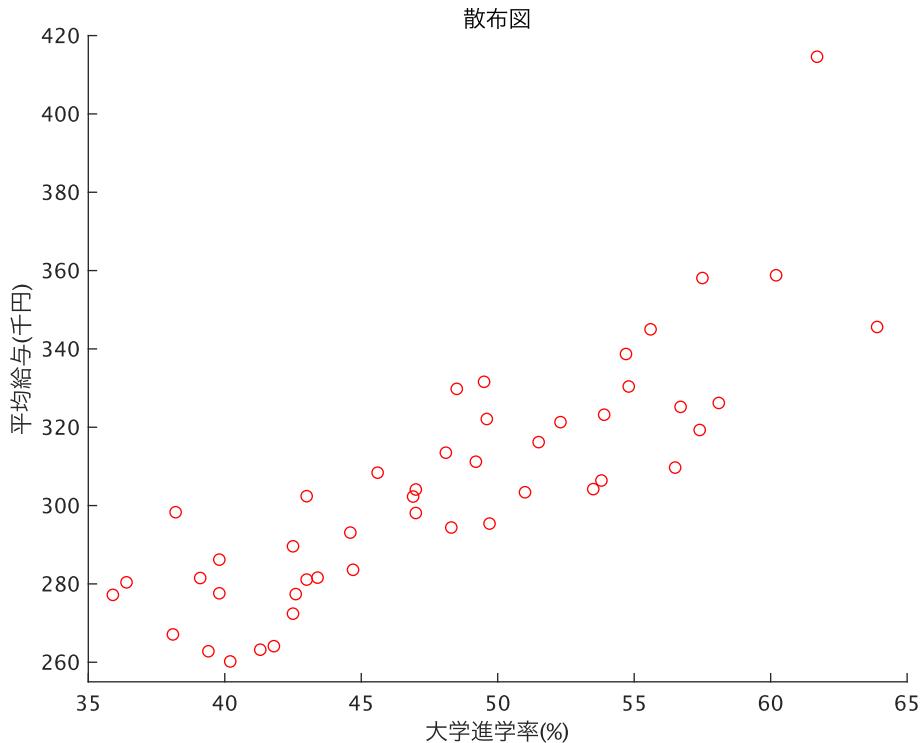


図3.3: 各都道府県の大学進学率(男性・平成29年)と各都道府県の平均給与(男性・平成29年)の関係の図示

散布図は、2つの変数の関係を示すために、データをプロットした図である。散布図を用いることで、2つの変数の関係を視覚的に把握できる。また、データの散らばりの程度や異常な値を視覚的に把握できる。「各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)」の関係の散布図を見ると、

- 大学進学率が高い都道府県は平均給与が高い
- 両者は直線的関係に近い関係を持つ

ということがわかる。

3.1.2 共分散

相関の有無や正負、強弱を量的に表す指標として共分散 (covariance)、相関係数 (correlation coefficient) がある。とりあえず、まず共分散を定義しよう。

共分散

与えられたデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して、共分散 (covariance) は次の式で定義される。

$$S_{xy} = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) \quad (3.1)$$

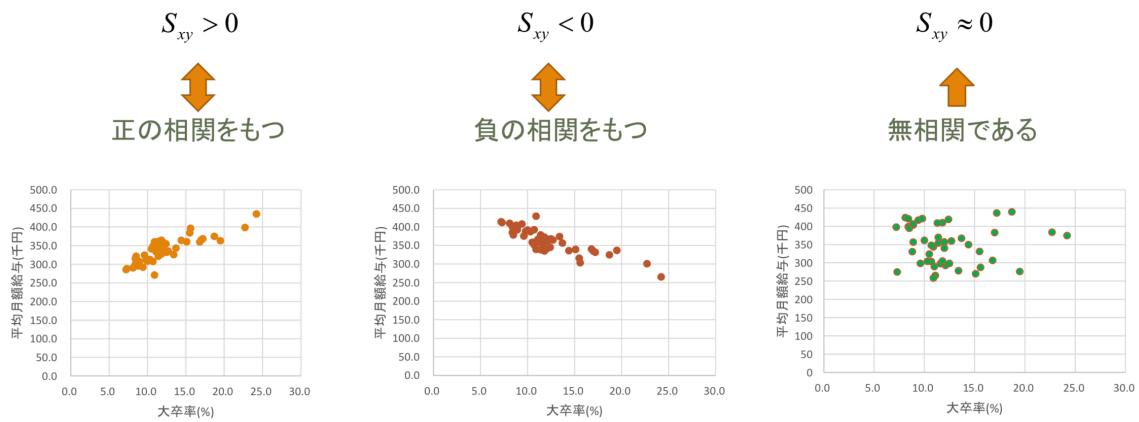


図 3.4: 共分散と相関関係

では、「各都道府県の大学進学率 (2016 年) と各都道府県の平均給与 (2017 年)」の関係について、共分散を求めてみよう。でも、この計算はとても面倒である。Excel でも使わないと大変なことになる。そこで、共分散に関する次の公式を取り上げる。

共分散の求め方 Part2

与えられたデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して、共分散は次の式でも求められる。

$$S_{xy} = \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{x} \cdot \bar{y} \quad (3.2)$$

証明

共分散の定義式、式(3.1)の右辺を変形していく。

$$\begin{aligned}
 S_{xy} &= \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{n} \sum_{k=1}^n (x_k y_k - \bar{y} \cdot x_k - \bar{x} \cdot y_k + \bar{x} \cdot \bar{y}) \\
 &= \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{y} \cdot \frac{1}{n} \sum_{k=1}^n x_k - \bar{x} \cdot \frac{1}{n} \sum_{k=1}^n y_k + \frac{1}{n} \sum_{k=1}^n \bar{x} \cdot \bar{y} \\
 &= \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{y} \cdot \bar{x} - \bar{x} \cdot \bar{y} + \bar{x} \cdot \bar{y} \\
 &= \frac{1}{n} \sum_{k=1}^n x_k y_k - \bar{x} \cdot \bar{y}
 \end{aligned}$$

□

式(3.1)を使って、「各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)」の関係の共分散を求めてみよう。

- 大学進学率の平均は46.07%
- 給与の平均は277.37(千円)
- 各都道府県の大学進学率と給与の積の平均は、12974.03(%×円)

と求められるので、大学進学率と平均給与の共分散は、

$$12974.03 - 46.07 \times 277.37 = 195.5941$$

である。共分散は正なので、大学進学率と平均給与の間には正の相関があることがわかる。

このsubsectionの最後に、1次元データの時にも出てきた「データの変換」の公式の2次元versionを取り上げる。

データの変換と共分散

与えられたデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して、 $z_k = ax_k + b, w_k = cy_k + d$ と変換するとき、 z と w の共分散 S_{zw} は

$$S_{zw} = acS_{xy} \quad (3.3)$$

と表せる。

証明

データの変換と平均の関係から、 $\bar{z} = a\bar{x} + b, \bar{w} = c\bar{y} + d$ が成り立つことを使う。

$$\begin{aligned}
 S_{zw} &= \frac{1}{n} \sum_{k=1}^n (z_k - \bar{z})(w_k - \bar{w}) = \frac{1}{n} \sum_{k=1}^n \{(ax_k + b) - (a\bar{x} + b)\} \{(cy_k + d) - (c\bar{y} + d)\} \\
 &= \frac{1}{n} \sum_{k=1}^n ac(x_k - \bar{x})(y_k - \bar{y}) \\
 &= acS_{xy}
 \end{aligned}$$

□

3.1.3 相関係数

共分散のデメリットは、共分散は単位を持つということである。そのため、他のデータと相関の程度を比較するときには不便である。そこで、測定単位に依存しない相関の指標として、相関係数 (correlation coefficient) を考える。

相関係数

与えられたデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して、相関係数 (correlation coefficient) は次の式で定義される。

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{S_x} \right) \left(\frac{y_k - \bar{y}}{S_y} \right) \quad (3.4)$$

(この定義式から、相関係数は x と y の基準化変量の共分散であることがわかる。)

相関係数 r は -1 と 1 の間の値をとることが知られている。相関係数の正負とデータの相関には以下の関係があり、相関係数の絶対値が 1 に近いほど、正または負の相関が強い。(絶対値が 1 のときは、全てのデータがある直線上にのっていることを表す。)

- 正の相関 : $0 < r_{xy} \leq 1$
- 負の相関 : $-1 \leq r_{xy} < 0$
- 無相関 : r_{xy} が 0 に近い。

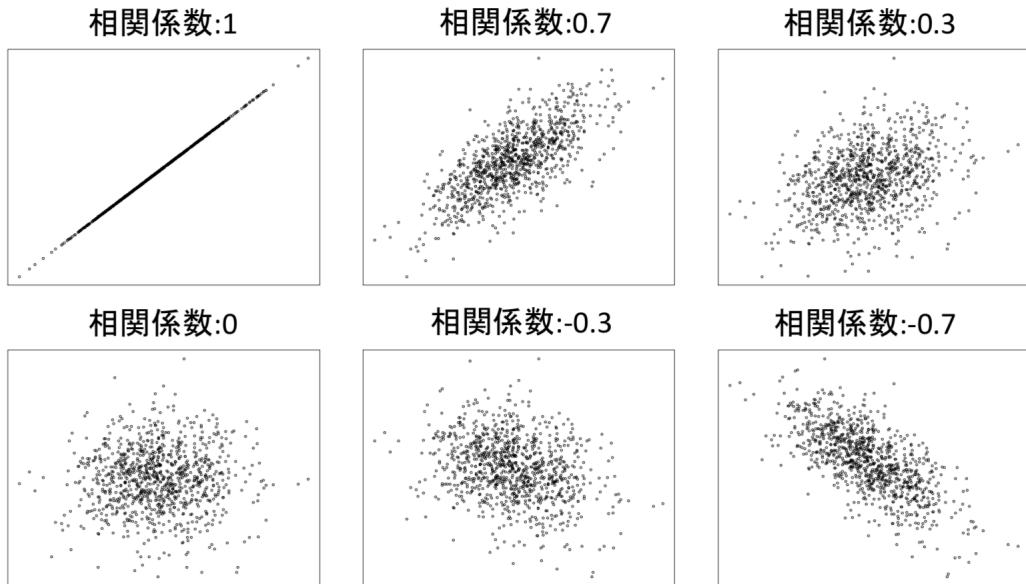


図 3.5: 相関係数と散布図の関係

式 (3.4) を使って、「各都道府県の大学進学率 (2016 年) と各都道府県の平均給与 (2017 年)」の関係の相関係数を求めてみよう。

- 共分散 $S_{xy} = 195.5941$

- 大学進学率 (x) の分散 $(S_x)^2 = 59.1283$
- 平均給与 (y) の分散 $(S_y)^2 = 787.8041$

と求めることができるので、相関係数 r_{xy} は、

$$r_{xy} = \frac{195.5941}{\sqrt{59.1283} \cdot \sqrt{787.8041}} = 0.9063$$

と求められる。相関係数が 0.9 なので、散布図は右上がりで、直線からの散らばりはあまり多くないことが予想される。30 ページの散布図 (図 3.3) を見てみると、この計算結果には納得できる。

2 次元データについて、データの変換と共に分散の関係は前の subsection の最後にふれたので、今度は、データの変換と相関係数の関係にふれる。

データの変換と相関係数

与えられたデータ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ に対して、 $z_k = ax_k + b, w_k = cy_k + d$ と変換するとき、 z と w の相関係数 S_{zw} は

$$S_{zw} = r_{xy} \quad (3.5)$$

と表せる。

証明

データの変換と共に分散の関係、および、データの変換と分散の関係より、

$$\begin{aligned} S_{zw} &= acS_{xy} \\ S_z &= aS_x \\ S_w &= cS_y \end{aligned}$$

となることより示される。 □

この subsection の最後に、相関係数 r は -1 と 1 の間の値をとることを示す。

証明

データ変換と相関係数の関係から、 x_k と y_k を基準化した場合を考えても、相関係数は変わらない。そこで、

$$z_k = \frac{x_k - \bar{x}}{S_x} \quad w_k = \frac{y_k - \bar{y}}{S_y}$$

と基準化する。

Remark

「2.3.7 基準化変量」の部分で紹介した通り、基準化変量の平均は 0 、分散は 1 である。
そのため、次の式が成立する。

$$\bar{z} = 0 \quad \bar{w} = 0$$

$$S_z = 1 \quad S_w = 1$$

これを使うと、 $r_{zw} = \frac{S_{zw}}{S_z S_w} = S_{zw}$ となることも導ける。

この Remark の内容を使って、 $(z_k + w_k)^2$ と $(z_k - w_k)^2$ の平均について考える。

$$\begin{aligned} \frac{1}{n} \sum_{k=1}^n (z_k + w_k)^2 &= \frac{1}{n} \sum_{k=1}^n (z_k)^2 + \frac{2}{n} \sum_{k=1}^n z_k w_k + \frac{1}{n} \sum_{k=1}^n (w_k)^2 \\ &= \frac{1}{n} \sum_{k=1}^n (z_k - \bar{z})^2 + \frac{2}{n} \sum_{k=1}^n (z_k - \bar{z})(w_k - \bar{w}) + \frac{1}{n} \sum_{k=1}^n (w_k - \bar{z})^2 \\ &= 1 + 2S_{zw} + 1 \\ &= 2 + 2r_{zw} \end{aligned}$$

(データ変換によって、相関係数は変化しないことより、 $r_{zw} = r_{xy}$ なので、)

$$= 2 + 2r_{xy} \geq 0$$

これより、 $r_{xy} \geq -1$ である。同様にして、 $2 - 2r_{xy} \geq 0$ となるので、 $r_{xy} \leq 1$ となる。 \square

3.2 相関に関する注意点(みかけ上の相関)

相関係数は、測定単位に依存しないから、他のデータと相関の度合いを比べるにはとても便利である。しかし、取扱に注意しなければならない。例えば、「小学生男子の足の大きさ」と「小学生男子が知っている漢字の数」について、相関を調べてみよう。すると、「足の大きい小学生ほど、多くの漢字を知っている」という結果が得られるだろう。だからといって、「小学生男子の足の大きさ」と「小学生男子が知っている漢字の数」に相関があるとは言えない。なぜなら、高学年の小学生ほど足が大きく、また、高学年の小学生ほど多くの漢字を知っているといえるからである。つまり、「学年」という第三者のせいで「足の大きさ」と「知っている漢字の数」の間に相関があるように見えるのである。

3.3 回帰モデル

2次元データによっては、変数 x が変数 y を決定するという関数関係が見られる時がある。例えば、「年齢と血圧の関係」の場合、一般的には、年齢が原因で、血圧が結果とみるのが自然である場合が多い。このような関係を因果関係という。相関関係ではない。どちらがどちらを決定すると考えられる関係を因果関係という。因果関係については、適当な関数 $f(x)$ を用いて、 $y = f(x)$ と表し、関数 $f(x)$ がどのような性質を持つか調べれば良い。この Section では、特に $f(x)$ が一次関数で $f(x) = \beta_0 + \beta_1 x$ となる場合について見ていく。

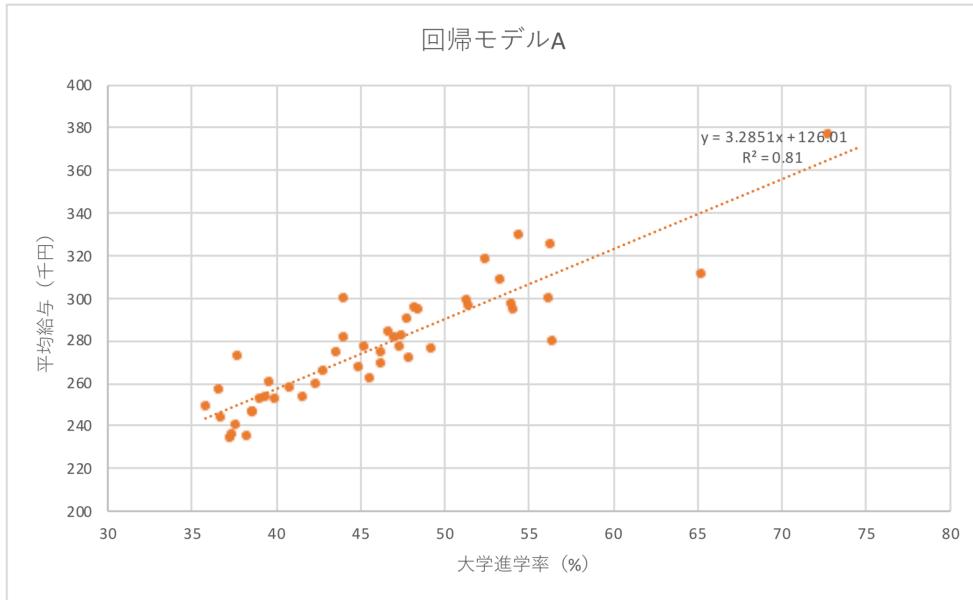


図 3.6: 各都道府県の大学進学率(2016年)と各都道府県の平均給与(2017年)の関係のグラフ

多くの場合、適当な一次関数をとっても、全てのデータが完全に、一次関数が表す直線上にのることはない。そこで、直線からの乖離を表す項 ε_k を導入して、

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k \quad (3.6)$$

という形で式に表すこととする。この式 (3.6) を回帰モデルといい、このモデルに基づいて 2 変数の関係を調べることを回帰分析という。ここで、いくつかの用語を紹介しよう。

- 変数 x : 独立変数
- 変数 y : 従属変数
- 直線の切片 β_0 と傾き β_1 : 回帰係数
- 誤差 ε_k : 誤差項

ただ、 (x_k, y_k) のデータしか持っていないときは、回帰係数と誤差項は未知数である。

3.4 最小2乗法による回帰直線の推定

3.4.1 最小2乗法とは？

与えられたデータに対して、どのような直線 $y = \beta_0 + \beta_1 x$ を当てはめるのが最適かを判断する方法を考える。つまり、 β_0, β_1 をどう取れば、与えられたデータに最もよく合うかを判断する。この方法として最も popular なのが最小2乗法である。与えられたデータ (x_k, y_k) に対して、

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$$

で定まる ε_k (誤差項)に対して、 ε_k の 2 乗 $(\varepsilon_k)^2$ の、 $k = 1, \dots, n$ の和を最小にする β_0, β_1 を求めればよい⁽¹⁾。

⁽¹⁾ 上のように定めた誤差項 ε_k の絶対値 $|\varepsilon_k|$ の、 $k = 1, \dots, n$ の和をとってもよいが、絶対値を含む計算は面倒なので、誤差項の 2 乗の値を使う。

3.4.2 最小 2 乗推定値の導出

β_0 と β_1 をどのように求めれば良いかは、数学的に導くことができる。まず、先に結論を書いてしまおう。

最小 2 乗推定値

データ $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ の回帰直線の方程式を

$$y = \beta_0 + \beta_1 x$$

とすると、この直線の傾きと切片は

$$\beta_1 = \frac{S_{xy}}{(S_x)^2} \quad (3.7)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (3.8)$$

で与えられる。

証明

$$\varepsilon_k = y_k - (\beta_0 + \beta_1 x_k)$$

とかけるから、 β_0 と β_1 の関数 $f(\beta_0, \beta_1)$ を

$$f(\beta_0, \beta_1) = \sum_{k=1}^n (\varepsilon_k)^2 = \sum_{k=1}^n (y_k - \beta_0 - \beta_1 x_k)^2 \quad (3.9)$$

と定義する。この $f(\beta_0, \beta_1)$ を最小にする β_0, β_1 を求める。

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{k=1}^n \{(y_k)^2 + (\beta_0)^2 + (\beta_1 x_k)^2 - 2\beta_0 y_k + 2\beta_0 \beta_1 x_k - 2\beta_1 x_k y_k\} \\ &= \left(\sum_{k=1}^n (y_k)^2 \right) + n(\beta_0)^2 + \beta_1^2 \left(\sum_{k=1}^n (x_k)^2 \right) - 2\beta_0 n \bar{y} + 2\beta_0 \beta_1 n \bar{x} - 2\beta_1 \left(\sum_{k=1}^n x_k y_k \right) \end{aligned}$$

なので、 β_0, β_1 に関して偏微分すると、

$$\frac{\partial f}{\partial \beta_0} = 2n(\beta_0 - \bar{y} + \beta_1 \bar{x}) \quad (3.10)$$

$$\frac{\partial f}{\partial \beta_1} = 2 \left\{ \beta_1 \left(\sum_{k=1}^n (x_k)^2 \right) + n\beta_0 \bar{x} - \left(\sum_{k=1}^n x_k y_k \right) \right\} \quad (3.11)$$

となる。これらがともに 0 になるには、式 (3.10) より、

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (3.12)$$

が成立することが必要である。この下で、 β_1 について、

$$\beta_1 \left(\sum_{k=1}^n (x_k)^2 \right) + n \bar{x} (\bar{y} - \beta_1 \bar{x}) - \left(\sum_{k=1}^n x_k y_k \right) = 0$$

が成立するので、

$$\begin{aligned} \beta_1 &= \frac{(\sum_{k=1}^n x_k y_k) - n \bar{x} \cdot \bar{y}}{(\sum_{k=1}^n (x_k)^2) - n (\bar{x})^2} \\ &= \frac{n S_{xy}}{n (S_x)^2} = \frac{S_{xy}}{(S_x)^2} \end{aligned} \quad (3.13)$$

と求められる。 □

さて、式(3.12)より、 $\bar{y} = \beta_0 + \beta_1 \bar{x}$ とかけるので、回帰直線 $y = \beta_0 + \beta_1 x$ は、

$$y - \bar{y} = \beta_1(x - \bar{x}) \quad (3.14)$$

とかける。これより、最小2乗法により求められる回帰直線は、点 (\bar{x}, \bar{y}) を通る傾き β_1 の直線であることがわかる。

3.4.3 「大学進学率」と「平均給与」の関係に関する回帰直線

このchapterの最初から扱っている「各都道府県の大学進学率」と「各都道府県の平均給与」のデータについて、回帰直線を求めよう。29ページより、 $S_{xy} = 195.5941$ 、 $(S_x)^2 = 59.1283$ なので、

$$\beta_1 = \frac{195.5941}{59.1283} = 3.3080$$

と、回帰直線の傾きが求められる。また、28ページより、大学進学率の平均は46.07、平均給与の平均は277.37なので、回帰直線の方程式は、

$$y = 3.3080(x - 46.07) + 277.37 = 3.3080x + 124.97$$

と求められる。35ページのグラフは、「各都道府県の大学進学率」と「各都道府県の平均給与」のデータに回帰直線を当てはめたものである。直線の横に、直線の方程式が表示されている。ただ、Excelにより導出された数値は、今の計算結果と異なる。これは以下の2つの理由による。

- 今回の計算では平均や分散、共分散は四捨五入された値を使っているから(近似値を使っているから)。
- 分散として、標本分散を使うか、不偏標本分散を使うかの違いである。

3.5 決定係数

3.5.1 予測値と残差

回帰直線 $y = \beta_0 + \beta_1 x$ が得られると、ある特定の x の値に対する y の値を予測することができる。

- 予測値：データ x_k に対して、 $y_k' = \beta_0 + \beta_1 x_k$ を y_k の予測値という。
- 残差： $\varepsilon_k' = y_k - y_k'$ で定まる ε_k' を残差といい、誤差 ε_k の推定値と解釈できる。

残差の性質

残差 ε_k' は次の式を満たす。

$$\sum_{k=1}^n \varepsilon_k' = 0 \quad (3.15)$$

$$\sum_{k=1}^n x_k \varepsilon_k' = 0 \quad (3.16)$$

証明**Remark**

$$\text{回帰係数 } \beta_0 \text{ と } \beta_1 \text{ の関係 : } \beta_1 = \frac{S_{xy}}{(S_x)^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\text{共分散 : } S_{xy} = \frac{1}{n} \left(\sum_{k=1}^n x_k y_k \right) - \bar{x} \cdot \bar{y}$$

$$\text{分散 (1 次元データ) : } (S_x)^2 = \frac{1}{n} \left(\sum_{k=1}^n (x_k)^2 \right) - n(\bar{x})^2$$

$$\begin{aligned} \varepsilon'_k &= y_k - y'_k \\ &= y_k - (\beta_0 + \beta_1 x_k) \\ &= y_k - (\bar{y} - \beta_1 \bar{x} + \beta_1 x_k) = (y_k - \bar{y}) - \beta_1(x_k - \bar{x}) \end{aligned}$$

となり、

$$\sum_{k=1}^n \varepsilon'_k = \sum_{k=1}^n \{(y_k - \bar{y}) - \beta_1(x_k - \bar{x})\} = 0 \quad (3.17)$$

である。次に、2つ目の式について

$$\begin{aligned} \sum_{k=1}^n x_k \varepsilon'_k &= \sum_{k=1}^n \{x_k(y_k - \bar{y}) - x_k \beta_1(x_k - \bar{x})\} \\ &= \sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \bar{y} - \sum_{k=1}^n \beta_1(x_k)^2 + \sum_{k=1}^n x_k \bar{x} \beta_1 \\ &= \left\{ \left(\sum_{k=1}^n x_k y_k \right) - n \bar{x} \cdot \bar{y} \right\} - \beta_1 \left\{ \left(\sum_{k=1}^n (x_k)^2 \right) - n(\bar{x})^2 \right\} \\ &= n S_{xy} - \frac{S_{xy}}{(S_x)^2} \cdot n(S_x)^2 = 0 \end{aligned}$$

と示される。 □

3.5.2 決定係数

回帰直線 $y = \beta_0 + \beta_1 x$ が得られた時、その直線が実際のデータに当てはまっているかどうかを評価する指標として決定係数を導入する。当てはまりが良いとは、回帰直線とデータの差、残差が 0 に近いかどうかで判断できる。そこで、残差の 2 乗の和⁽³⁾が 0 に近いかどうかを check すればよい。つまり、

$$C = \sum_{k=1}^n (\varepsilon'_k)^2 \quad (3.18)$$

と定義した C が 0 に近いかを check すればよい。 C が 0 に近いなら、よく当てはまっていると解釈できる。しかし、上記のように定めた C は単位に依存するので、 C の値だけでよく当てはまっているかどうかを判断するのは難しい。ちなみに、Excel の「分析ツール」を使うと、図 3.7 の C13 セルに表示される値が、残

⁽³⁾最小 2 乗法の時と同じで、 $\sum_{k=1}^n |\varepsilon'_k|^2$ を評価するのは大変なので、2 乗の和を評価する。

差の2乗の和の値である。「各都道府県の大学進学率」と「各都道府県の平均給与」の関係については、 C の値は 528.06791 であることがわかる。

	A	B	C	D	E	F	G	H	I	J
1	概要									
2										
3	回帰統計									
4	重相関 R	0.89998954								
5	重決定 R2	0.80998117								
6	補正 R2	0.80575853								
7	標準誤差	3.42561563								
8	観測数	47								
9										
10	分散分析表									
11		自由度	変動	分散	観測された分散比	有意 F				
12	回帰	1	2250.96145	2250.96145	191.8186345	7.7529E-18				
13	残差	45	528.06791	11.7348424						
14	合計	46	2779.02936							
15										
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
17	切片	-22.314984	4.96313361	-4.4961482	4.81979E-05	-32.311248	-12.31872	-32.311248	-12.31872	
18	X 値 1	0.2465619	0.01780248	13.8498605	7.75293E-18	0.21070586	0.28241794	0.21070586	0.28241794	
19										
20										

図 3.7: Excel によるデータ分析の結果

C の値は 528.06791 だからといって、当てはまりが良いのかは容易に想像できない。そこで、次にとりあげる「変動の分解」が使われる。

変動の分解

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (y'_k - \bar{y})^2 + \sum_{k=1}^n (\varepsilon'_k)^2 \quad (3.19)$$

証明

$y_k = y'_k + \varepsilon'_k$ と書けるので、 $y_k - \bar{y} = (y'_k - \bar{y}) + \varepsilon_k$ となる。

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (y'_k - \bar{y})^2 + \sum_{k=1}^n (\varepsilon'_k)^2 + 2 \sum_{k=1}^n (y'_k - \bar{y}) \varepsilon'_k$$

と書けるので、第3項が 0 になることを示せばよい。

Remark

- 最小2乗法により求められる回帰直線は、点 (\bar{x}, \bar{y}) を通る傾き β_1 の直線である。
 - 点 (x_k, y'_k) はこの直線上にある。(予測値の定義)
- ⇒ 以上の2点から、 $y'_k - \bar{y} = \beta_1(x_k - \bar{x})$ が成立する。

上の **Remark** の内容から、

$$\begin{aligned}\sum_{k=1}^n (y'_k - \bar{y})x_k &= \beta_1 \sum_{k=1}^n (x_k - \bar{x})\varepsilon'_k \\ &= \beta_1 \sum_{k=1}^n x_k \varepsilon'_k - \beta_1 \bar{x} \sum_{k=1}^n \varepsilon'_k = 0\end{aligned}$$

□

式 (3.19) の意味をもう少し具体的に考えよう。

- 全変動 $A = \sum_{k=1}^n (y_k - \bar{y})^2$ ： 観測値 y の変動の大きさを表す。
- 回帰変動 $B = \sum_{k=1}^n (y'_k - \bar{y})^2$ ： 予測値 $y'_k = \beta_0 + \beta_1 x_k$ の変動の大きさを表す。
- 残差変動 $C = \sum_{k=1}^n (\varepsilon'_k)^2$ ： 全変動と回帰変動の誤差

A , B , C を上のように定義すると、全変動 (A) は回帰変動 (B) と残差変動 (C) に分解できるのが、式 (3.19) の「変動の分解」の意味である。図 3.7 の C12 セル、C13 セル、C14 セルは表全体を見ればわかるが、C12 が回帰変動、C13 が残差変動、C14 が合計の変動 (全変動) を表している。確かに、 $A = B + C$ の関係が成立していることがわかる。

この A , B , C を使って、単位に依存しない当てはまりの指標を定めることができる。

決定係数

全変動に占める回帰変動の割合で決定係数 R^2 を定める。

$$R^2 = \frac{B}{A} = 1 - \frac{C}{A} \quad (3.20)$$

この決定係数 R^2 は $0 \leq R^2 \leq 1$ を満たしていて、 R^2 が 1 に近いほど、(C が 0 に近くなるので、) 回帰直線がデータによく当てはまっていることを示す。

3.5.3 決定係数と相関係数

決定係数と相関係数

決定係数 R^2 は、相関係数 r_{xy} の 2 乗に等しい。

$$R^2 = (r_{xy})^2 \quad (3.21)$$

証明

$$y'_k - \bar{y} = \beta_1(x_1 - \bar{x}) \text{ なので、} \sum_{k=1}^n y'_k - \bar{y} = \sum_{k=1}^n \beta_1(x_1 - \bar{x}) \text{ となる。}$$

$$\begin{aligned} R^2 &= \frac{\sum_{k=1}^n (y'_k - \bar{y})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} \\ &= \frac{(\beta_1)^2 (S_x)^2}{(S_y)^2} \\ &= \frac{\left(\frac{S_{xy}}{(S_x)^2}\right)^2 \cdot (S_x)^2}{(S_y)^2} \\ &= \frac{(S_{xy})^2}{(S_x S_y)^2} = r^2 \end{aligned}$$

□

決定係数は、相関係数の絶対値が大きくなるほど大きくなる。そのため、決定係数が大きいほど相関関係が強いということがわかる。しかし、決定係数だけでは、正の相関か負の相関かは判定できない。

このSectionの最後に、「大学進学率」と「平均給与」の関係について、決定係数を求めよう。決定係数の定義式のように A, B, C を求めるのは面倒なので、決定係数が相関係数の2乗であることを使う。共分散 $S_{xy} = 195.5941$ 、大学進学率 (x) の分散 $(S_x)^2 = 59.1283$ 、平均給与 (y) の分散 $(S_y)^2 = 787.8041$ なので、相関係数 r_{xy} は、

$$r_{xy} = \frac{195.5941}{\sqrt{59.1283} \cdot \sqrt{787.8041}} = 0.9063$$

と求められ、決定係数 R^2 は

$$R^2 = (0.9063)^2 = 0.8214$$

と求められる。

第 4 章 確率モデル

4.1 身近な確率

この section では、『理工基礎 確率とその応用』(サイエンス社) の第 1 章の『確率ア・ラ・カルト』の 1.1 と 1.2 の内容をもとに作成した。第 1 章では「内閣の支持率調査」、「エイズの計算結果」などの気になるテーマについても記されているが、それについてはここでは紹介しない。これから確率の話をした時に、少しずつ取り上げることにする。



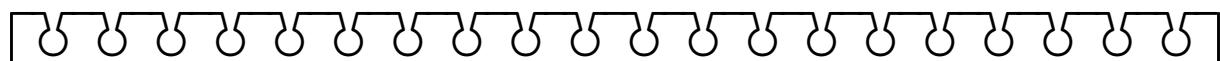
図 4.1: Google で「確率 面白い」と調べた結果

この chapter では「確率モデル」を扱う。「確率」という言葉はよく身近で使われるけれど、その本質をしっかり理解しているだろうか。図 4.1 に記されている「モンティ・ホール問題」は、「条件つき確率」の考え方を理解すると受け入れやすいが、そうでないと不思議なことを言っているだけとしか思えない。

条件つき確率など確率に関する議論を進める前に、私たちの身近で使われている「確率」と名のつくものについて見ていくことにしよう。

4.1.1 降水確率

梅雨の時期は毎日必ず天気予報⁽¹⁾を見るだろう。梅雨の時期は何を気にするか?「いつ、どれくらいの雨が降るか」を気にするだろう。天気が怪しい時は「降水確率」をもとに傘を持っていくかどうかを決める人は多いだろう。では、そこで使われている「降水確率」の定義は何か。気象庁のホームページを見てみよう。



降水確率

(a) 予報区内で一定の時間内に降水量にして 1 mm 以上の雨または雪の降る確率(%)の平均値で、0, 10, 20, …, 100 % で表現する(この間は四捨五入する)。

(b) 降水確率 30 % とは、30 % という予報が 100 回発表されたとき、その内のおよそ 30 回は 1 mm 以上の降水があるという意味であり、降水量を予報するものではない。

(https://www.jma.go.jp/jma/kishou/know/yougo_hp/yoho.html より)



降水確率 0 %

- 降水確率が 5 % 未満のこと。降水確率は 1 mm 以上の降水を対象にしているので、1 mm 未満の降水予想である場合は「降水確率 0 %」でもよい。ただし、実用上の見地からは雨または雪の降りにくい状態に用いることが好ましい。

(https://www.jma.go.jp/jma/kishou/know/yougo_hp/yoho.html より)

この定義によると、降水確率が低い日に限って土砂降りの雨が降っても、降水確率が高い日に折り畳み傘を持っていてたのに使わなくても、気象庁の言っていることに間違いはないのである。だって、降水確率 100 % でない限り、雨が降らなくても問題ないからだ。「降水確率 30 %」という予報を出したなら、たまたま今回は 70 回のハズレのうちの 1 回だったと言えば良いのである。

1 mm 以上の雨が降るかどうかで決まるので、パラパラ雨でもザアーザー雨でも良い。そうは言うけれど、降水確率が高い時はたいていザアーザー雨が降っているという感じがする。降水確率が高いほど、雨が降る可能性が高いというのは、「確率」という言葉からもなんとなくわかる。ただ、同時に降水確率が高いほど強い雨が降る場合が多い気がするので、降水確率と雨の強さを結びついている気がしてもおかしく

(1) 朝のニュース番組は何を見るかは、各家庭によって異なる。各局の朝のニュース番組はこんなものだろうか。私は千葉県に住んでいるのでこんな感じである。

- NHK ニュースおはよう日本 (NHK 総合テレビ)
- ZIP! (日本テレビ)
- あさチャン! (TBS)
- めざましテレビ (フジテレビ)
- グッド!モーニング (テレビ朝日)
- おはスタ (テレビ東京)

ちなみに、私の家はなんか知らないけど『めざましテレビ』である。朝はかやちゃんの天気予報を見ている。私は朝 6 時 27 分頃の天気予報を見る。CanCam が選んだコーデの部分より、天気予報の方を見ている。

はないだろう。こういう点が「確率」という数学的概念の捉え方の勘違いにつながるのかもしれない。

4.1.2 テレビの視聴率

<https://www.videor.co.jp/tvrating/> によると、2019年7月15日～7月21日のドラマの視聴率ランキングは以下の通りである⁽²⁾。

表 4.1: 2019年7月15日～7月21日のドラマの視聴率ランキング

番組名	放送局	番組平均世帯視聴率 (%)
連続テレビ小説・なつぞら	NHK 総合	21.2
刑事7人	テレビ朝日	13.1
監察医朝顔	フジテレビ	12.3
偽装不倫	日本テレビ	11.2
金曜ドラマ・凪のお暇	TBS	10.3
木曜ミステリー・科搜研の女	テレビ朝日	10.2
サイン・法医学者柚木貴志の事件	テレビ朝日	9.5
Heaven?・ご苦楽レストラン	TBS	9.2
土曜時代ドラマ・雲霧仁左衛門2	NHK 総合	8.9
TWOWEEKS	フジテレビ	8.4
ボイス 110緊急指令室	日本テレビ	8.4

テレビ局スタッフ、出演している俳優・女優が気にする「視聴率」はどのようにして決められるのか？ Wikipediaによると、視聴率の測定は基本的に、モニター世帯に設置されるテレビに接続した専用の機器から得られるデータを基にしているらしい。さらに、<https://www.videor.co.jp/service/media-data/tvrating.html> によると、モニターに指定されている世帯は関東地区ではたったの 900 世帯である。いうまでもなく、関東地区にある世帯は 900 よりはるかに多い。900 のデータだけで全体の特性を把握できるのはある程度予測できるのは何故なのか？その理由は確率や統計が教えてくれる。

4.2 確率モデルの基礎

確率論を議論するためには、その枠組みである標本空間 (sample space) を的確に理解しないといけないのだが、これが結構難しい。まずは、確率モデルの基礎となる事柄である標本空間と事象 (event) について確認する。

この section 以降は、**定義** と **例** に番号をつけることにする。数学の書物のような「定義」と「命題」「定理」「系」の羅列が続いて読む気が失せるあのタイプのノートに、この TeX ノートも近づいていくだろう。

(2) ちなみに、私が 2019 年の夏ドラマで見ているのは、「監察医朝顔」、「凪のお暇」、「Heaven?」、「ボイス」、「ルパンの娘」、「あなたの番です 反撃編」、「ノーサイドゲーム」の 7 つである。「あなたの番です」、「ノーサイドゲーム」はこの週は、参議院議員選挙特番のため放送されなかったため、表には載っていない。「ルパンの娘」はあの深田恭子が主演なのに視聴率が高くないのか、ランキング漏れとなってしまったらしい。

4.2.1 標本空間と事象

まず、試行 (trial) と標本空間 (sample space) を定義する。

定義 4.1

- 試行 : 実験や観測
- 標本空間 Ω : 試行の結果として起こりうるもの全体、すなわち試行の結果を要素とする集合

例 4.1

コインを1回投げ、表(1)か裏(0)を観測する試行に対する標本空間 Ω は以下のようにかける。

$$\Omega = \{0, 1\}$$

続けて、事象 (event) を定義する。「事象」とは、「直感的で非数学的な書き方をするなら、」「試行の結果として起こりうる事柄」、あるいは「将来起きるかもしれないランダムな出来事」のことである。だが、統計学や確率論では、次のように「事象」を定義する。

定義 4.2

- 事象 : 標本空間 Ω の部分集合
- 「事象 A が起こった」とは、試行の結果として A に含まれる要素の1つが起こったこと。
例えば、コインを2回投げる(表:1, 裏:0)時、Aを「1回目に表が起こる」とすると、(1,0), (1,1) が起こることを指す。
- 標本空間 Ω を全事象 (whole event) と呼ぶ。(標本空間 Ω 自身も Ω の部分集合)
- 要素のない事象(決して起こらない事象)を空事象 (empty event) ϕ と呼ぶ。
(要素が何もない事象も Ω の部分集合)
- 根元事象 : ただ1つの要素のみからなる事象
先ほどのコインを2回投げる(表:1, 裏:0)場合における事象 A は、(1,0) と (1,1) の2つの要素からなるので、根元事象ではない。

4.2.2 事象の演算

事象に対して、いくつかの演算を定義する。

定義 4.3

- 余事象 (complementary event) A^c : 事象 A に含まれない要素からなる事象

$$A^c = \{ \text{A に含まれない要素} \} = \{ \text{A が起こらない} \} \quad (4.1)$$

- 和事象 (union of events) $A \cup B$: A または B の少なくとも一方に含まれる事象

$$\begin{aligned} A \cup B &= \{ \text{A または B の少なくとも一方に含まれる要素} \} \\ &= \{ \text{A または B が起こる} \} \end{aligned} \quad (4.2)$$

- 積事象 (intersection of events) $A \cap B$: A と B の両方に含まれる事象

$$A \cap B = \{ \text{A と B の両方に含まれる要素} \} = \{ \text{A と B の両方が起こる} \} \quad (4.3)$$

- 上記の事象の演算と「事象」の定義より次の式が成立することがわかる。

$$A \cup A^c = \Omega \quad (4.4)$$

定義 4.4

2つの事象 A と B に対して、 $A \cap B = \phi$ が成り立つとき、A と B は互いに排反 (disjoint) であるという。すなわち、A と B は同時に起こらず、一方が起きている時、他方は起きないならば、この 2つの事象を互いに排反という。

事象の演算については以下のようないくつかの関係も成り立つ⁽³⁾。

事象の演算の性質

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \quad (4.5)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C) \quad (4.6)$$

$$(A \cup B)^c = A^c \cap B^c \quad (4.7)$$

$$(A \cap B)^c = A^c \cup B^c \quad (4.8)$$

4.2.3 確率の定義

ラプラスの定義

ラプラスによる確率の定義は次のようにある。試行の根元事象が全部で N 個あって、それらは同様に確

⁽³⁾ 式 (4.7) と式 (4.8) は、特にド・モルガンの法則と言われる。

からしい (equally likely) とする。この時、それが出ればその事象 A が起こるような根元事象が R 個あれば、事象 A の起こりやすさ $P(A)$ は

$$P(A) = \frac{R}{N}$$

と定義される。

例えば、「サイコロを 1 回ふった時に奇数の目が出る」(これを事象 A とする) 確率 $P(A)$ を考える、根元事象は 6 つあり ($N = 6$)、このうち、1,3,5 の目が出れば良いので、A が起こるような根元事象は 3 つある ($R = 3$)。よって、 $P(A) = \frac{3}{6} = \frac{1}{2}$ である。

この定義の最大の利点は、確率を場合の数の数え上げに帰着できることであり、順列や組み合わせの諸定理が使えることである。しかし、その一方で、「同様に確からしい」という仮定のもとでしか正しくない。そのため、歪みのないサイコロに関する確率など、特定の場合のみしか(数学的に) 正しくない。そうはいいうけど、確率論の例題の多くは「同様に確からしい」という仮定のもと、このラプラスの定義を使って確率を計算する。

なぜ、「同様に確からしい」という仮定が OK かということを説明するのが、『統計学入門』(東京大学出版会) でいう理由不充分の原則や、『理工基礎 確率とその応用』(サイエンス社) でいう等可能性の原理である。(理想的な⁽⁴⁾) サイコロの目やコインの裏表のように、標本空間が有限で、各々の「目」が出る事象が他の事象よりも起きやすいという明確な根拠がない限り、同程度の確かさで出現すると考えるのが妥当であるという原理(原則)のことである。この原理(原則)は否定できない限り、この原理(原則)を正しいと考えるしかないということである。

頻度による確率の定義

各根元事象の起こる確率が同様に確からしくないときは、ラプラスの確率の定義が使えない。そこで、次に頻度による確率の定義を考える。

事象 A を生みうる実験を n 回繰り返して、A が n_A 回起きたとする。このとき、 $\frac{n_A}{n}$ により、A が起きた割合(相対頻度)が定義できる。頻度により確率を定義するときは、 $n \rightarrow \infty$ の極限がいくつに収束するかを考える。つまり、

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

により確率 $P(A)$ を定義する。

しかし、この定義にも問題点がある。

- n が大きくなるほど、相対頻度は真の確率に近づくが、非常に多く実験を繰り返しても、($n \rightarrow \infty$ でない限り、) 真の確率とは一致しない。
- 各根元事象の起こる確率が同様に確からしくないので、極限の収束値はいつも同じとは限らない。

確率の公理主義的定義

数学者コルモゴロフは、確率論の公理を決めることで、ラプラスの定義や頻度による定義の問題点を解決を試みた。数学的には、以下の定義を満たすなら、どのような数も確率であるとした。この定義が一番わかりづらいが、測度論的確率論といった高級な概念を理解するためには、確実に押さえなければならない。

(4) サイコロの 1 の目～6 の目を均一な密度を持つ立方体に丸を適切な数だけ貼り付けることで作るとする。この時、1 の目は丸を 1 枚貼るだけでいいが、6 の目は 6 枚も貼る必要がある。すると、わずかであるが、6 の目の近くは 1 の目の近くよりも重くなっている。適切な方法でサイコロを作らない限り、サイコロの目の出方は(厳密には) 同様に確からしくないといえる。

確率論の公理

1. 確率は非負である。全ての事象 A に対して、 $0 \leq P(A) \leq 1$ である。

2. 全事象の確率は 1 である。一方、空事象の確率は 0 である。

$$P(\Omega) = 1, \quad P(\emptyset) = 0 \quad (4.9)$$

3. 事象 A_1, A_2, \dots, A_n が互いに排反ならば、これらのうち少なくとも 1 つが起こるという事象 $A_1 \cup A_2 \cup \dots \cup A_n$ の起こる確率は、各事象の和に等しい⁽⁵⁾。(シグマ加法性)

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad (4.10)$$

4.2.4 確率の基本公式

確率の基本公式

1. 事象 A に対して、 A の余事象 A^c の起こる確率 $P(A^c)$ は次のように表される。

$$P(A^c) = 1 - P(A) \quad (4.11)$$

2. 2 つの事象 A と B が $A \subset B$ ならば、不等式 $P(A) \leq P(B)$ が成立する。

3. 必ずしも排反ではない 2 つの事象 A と B に対しては、次の式が成立する。

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.12)$$

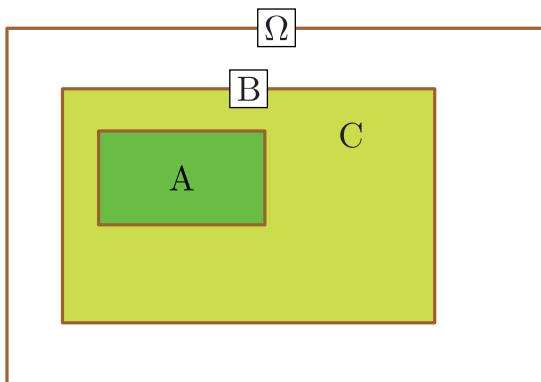
証明

図 4.2: 「確率の基本公式」の 2 の証明

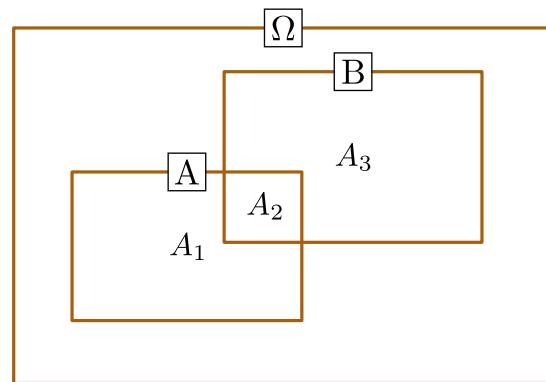


図 4.3: 「確率の基本公式」の 3 の証明

1について

$\Omega = A \cup A^c$ であり、 A と A^c は互いに排反なので、

$$P(\Omega) = P(A) + P(A^c) \quad (4.13)$$

(5) この n は有限である必要性はない。

が成立する。確率論の公理より $P(\Omega) = 1$ なので、 $P(A^c) = 1 - P(A)$

2について

包含関係は図4.2のようにかける。そのため、不等式 $P(A) \leq P(B)$ が成立するのは明らかである。数式を使うなら、Bを排反な2つの事象AとCの和事象 $A \cup C$ で表して考えればよい。すると、 $P(B) = P(A) + P(C)$ であるが、確率論の公理より、 $P(C) \geq 0$ なので、 $P(B) \geq P(A)$ となる。

3について

$A \cup B$ を図4.3のように排反な3つの集合 A_1, A_2, A_3 に分ける。すると、

$$P(A) = P(A_1) + P(A_2), \quad P(B) = P(A_2) + P(A_3)$$

である。集合 A_2 が、 $A \cap B$ であることに注意すると、

$$\begin{aligned} P(A \cup B) &= \underbrace{P(A_1)}_{=P(A)} + \underbrace{P(A_2)}_{=P(A \cap B)} + P(A_3) \\ &= P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A \cap B) \end{aligned}$$

となる。 □

4.3 条件つき確率

4.3.1 条件つき確率のイントロ

このsectionでは、section4.1の最初に言葉だけ紹介した「モンティ・ホール問題」を考える枠組みである「条件つき確率」について考える。その前にイントロダクションとして、こんな問題の答えを考えてみよう⁽⁶⁾？

Aさんには子供が2人いて、一人は男の子である。もう一人の子供は、男の子と女の子のどちらでしょうか？

男か女かの $1/2$ に決まっていると考えるのは間違いである。そもそも、男性と女性の出生率は完全に $50 : 50$ ではなく、 $51 : 49$ ぐらいであるらしいから、 $1/2$ ではないというわけではない。確率の考え方が間違っているのである。

子供が2人いるときの組み合わせは、上の子と下の子を区別すると、(男・男)、(男・女)、(女・男)、(女・女)の4パターンに限られる。標本空間 Ω は以下のようになる。

$$\Omega = \{ (\text{男} \cdot \text{男}), (\text{男} \cdot \text{女}), (\text{女} \cdot \text{男}), (\text{女} \cdot \text{女}) \}$$

この中で興味があるのは、以下の3つに限られる。

$$B = \{ (\text{男} \cdot \text{男}), (\text{男} \cdot \text{女}), (\text{女} \cdot \text{男}) \}$$

この問題では、議論の土台を全体集合 Ω から B に制限する必要があるのである。この制限をどのように表現するかが条件つき確率なのである。議論の土台を B に制限すると、2人の子どものうち、1人が女の子なら、 $2/3$ の確率でもう1人は女の子であるということがすぐにわかる。

⁽⁶⁾ <https://tidbits.jp/probability-problem/> を参考にした。

4.3.2 条件つき確率

定義 4.5

事象 B が起きたことがわかっているという条件の下で、事象 A の起こる確率を、事象 B が与えられた時の事象 A の条件つき確率 (**conditional probability**) という。これを $P(A|B)$ と書くと、

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (4.14)$$

と定義される。また、この式を変形することで、次の公式 (確率の乗法公式) が得られる。

$$P(A \cap B) = P(A|B) \cdot P(B) \quad (4.15)$$

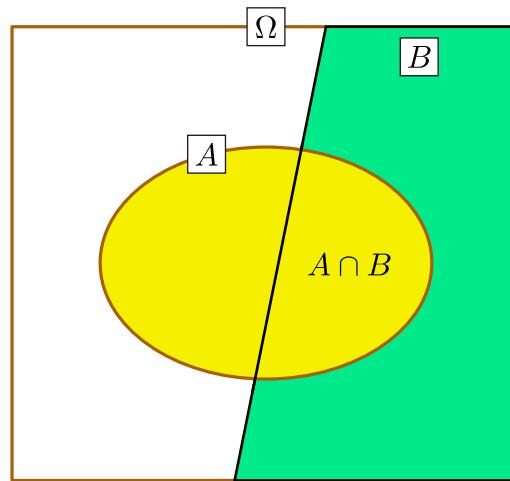


図 4.4: 条件つき確率のイメージ

「事象 B が起きたことがわかっている」という条件のもとで考えるなら、考える対象とする標本空間 Ω を B に制限すればよい。「事象 B が起きたことがわかっている」時に、事象 A が起こることと、事象 $A \cap B$ が起こることが等価であることから、式 (4.14) のような比を考えれば良いことがわかる。

そうはいっても、この説明だけではわからないので、以下の問題を考えてみることにしよう。

例 4.2

(<https://kamelink.com/public/2018/16.6-18 新潟大・理 2 文 3.pdf> より)

袋 A には赤玉 2 個と白玉 5 個、袋 B には赤玉 2 個が入っている。まず、袋 A から 3 個の玉を同時に取り出し、玉の色は確認せず、そのまま袋 B に入れ、よくかき混ぜて、袋 B から 2 個の玉を同時に取り出す。

- (1) 袋 A から取り出された 3 個の玉が、赤玉 1 個と白玉 2 個である確率、白玉 3 個である確率をそれぞれ求めよ。
- (2) 袋 B から取り出された玉が 2 個とも白玉である確率を求めよ。

- (3) 袋Bから取り出された玉が2個とも白玉であったとき、袋Bに白玉が残っている条件付き確率を求めよ。

解答

- (1) 赤玉1個と白玉2個である確率

$$\frac{2 \times {}_5C_2}{{}_7C_3} = \frac{20}{35} = \frac{4}{7}$$

白玉3個である確率

$$\frac{{}_5C_3}{{}_7C_3} = \frac{10}{35} = \frac{2}{7}$$

- (2) 袋Bから取り出された玉が2個とも白玉であるのは以下の2パターンに限る。この2パターンは排反である。

- (a) 袋Aから赤玉1個と白玉2個を取り出して、袋Bに入れたのち、赤玉3個、白玉2個が入っている袋Bから白玉を2個取り出す。
- (b) 袋Aから白玉3個を取り出して、袋Bに入れたのち、赤玉2個、白玉3個が入っている袋Bから白玉を2個取り出す。

(a) が起こる確率は、以下のように計算できる。

$$\frac{4}{7} \times \frac{1}{{}_5C_2} = \frac{4}{7} \times \frac{1}{10} = \frac{2}{35}$$

また、(b) が起こる確率は、以下のように計算できる。

$$\frac{2}{7} \times \frac{3}{{}_5C_2} = \frac{2}{7} \times \frac{3}{10} = \frac{3}{35}$$

よって、求める確率は、 $\frac{2}{35} + \frac{3}{35} = \frac{5}{35} = \frac{1}{7}$

- (3) 袋Bから2個の白玉を取り出しても、まだ白玉が残っているのは(b)の場合。

$$\left(\frac{3}{35}\right) / \left(\frac{5}{35}\right) = \frac{3}{5}$$

上の問題について、(1) は簡単な「組合せ」の計算の問題。(2) と (3) が定義4.5の内容を使った問題である。というけど、(2) の問題を「条件つき確率」と関連づけて考える人はあまり多くはないかもしれない。ただ、袋Aから赤玉1個と白玉2個を取り出す確率と、その後に、袋Bから白玉2個を取り出す確率をかけば良いと、受験生時代に手法だけ身につけてしまったのではないか。

後半の「袋Bから白玉2個を取り出す確率」は厳密には「袋Aから赤玉1個と白玉2個を取り出した後に袋Bから白玉2個を取り出す確率」であり、「袋Aから赤玉1個と白玉2個を取り出す」という結果がわかった状況と制限されているのである。公式の $A \cap B$ も、「袋Aから赤玉1個と白玉2個を取り出す」事象と「袋Bから白玉2個を取り出す」事象が両方とも起こるという意味では積事象になっている。

私は「同時に2つの条件を満たす」というイメージと積事象を結びつけて覚えている人なので、どうしても(2) の問題を「条件つき確率」と関連づけて考えることができない。とにかく、実は条件つき確率は大活躍しているのである。

4.3.3 全確率公式

全確率公式

標本空間 Ω が、互いに排反な n 個の事象 H_1, H_2, \dots, H_n の和でかけるとき、つまり、 $\Omega = H_1 \cup H_2 \cup \dots \cup H_n$ のとき、任意の事象 A の確率は以下のようになる。

$$P(A) = \sum_{k=1}^n P(A \cap H_k) = \sum_{k=1}^n P(A|H_k) \cdot P(H_k) \quad (4.16)$$

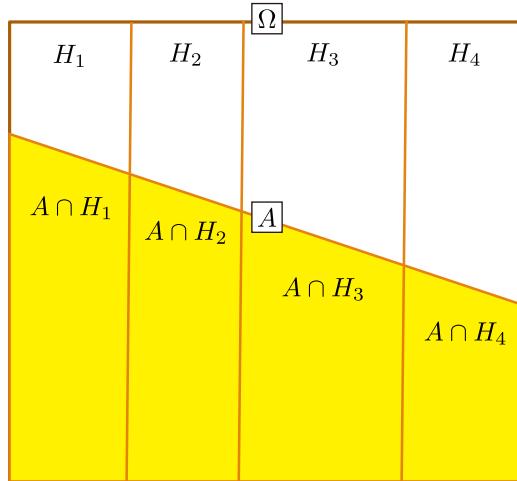


図 4.5: 全確率公式のイメージ

証明

$A = A \cap \Omega$ なので、

$$\begin{aligned} A &= A \cap \Omega = A \cap (H_1 \cup H_2 \cup \dots \cup H_n) \\ &= (A \cap H_1) \cup (A \cap H_2) \cup \dots \cup (A \cap H_n) \end{aligned}$$

となる。(上図からも明らかな通り、) $k = 1, 2, \dots, n$ に対して、各 $A \cap H_k$ は互いに排反なので、

$$P(A) = \sum_{k=1}^n P(A \cap H_k)$$

□

4.3.4 ベイズの定理

多くの受験生は、大学受験の前に必ず模試を受ける。ここでは、2016年11月に実施された駿台東大実戦模試⁽⁷⁾を例に考える⁽⁸⁾。さらに、ここでは理科一類についてのみ考える。

(7) 2019年度から河合塾の東大模試「東大オープン」はZ会と共に催をやめて、代わりに駿台の東大模試「東大実戦」がZ会と共に催にならざるを得ない。これにより、河合塾、駿台の東大模試がどう変わるのが気になる。

(8) ここでは簡単にするために、この模試を受けた人全員が、東大を受験したものとし、第一段階選抜（いわゆる、足切り）はないものとする。また、この模試を受けていない人は東大を受験しないものとする。まあ、東大を受験する人の多くは東大模試を受けているはずなので、2つ目の仮定はある程度正しいといえよう。



図 4.6: 2019 年夏の東大実戦模試のパンフレット

例 4.3

- $T = \{ \text{東大合格} \}$ とし、 $T^c = \{ \text{東大不合格} \}$ とする。
なお、標本空間 Ω は $\Omega = \{ \text{東大実戦受験者} \}$ とする。すると、 $\Omega = T \cup T^c$ である。
- 模試の判定は A から E の 5 段階評価である。以下では、

$$\begin{aligned} A &= \{ \text{A 判定} \}, B = \{ \text{B 判定} \}, C = \{ \text{C 判定} \} \\ D &= \{ \text{D 判定} \}, E = \{ \text{E 判定} \} \end{aligned}$$

とする。このとき、 $\Omega = A \cup B \cup C \cup D \cup E$ である。

- この模試および過去のデータから次のことがわかっているとする⁽⁹⁾。

$$\begin{aligned} P(A) &= 0.148, P(B) = 0.078, P(C) = 0.099, P(D) = 0.167, P(E) = 0.508 \\ P(T|A) &= 0.9, P(T|B) = 0.7, P(T|C) = 0.5, P(T|D) = 0.3, P(T|E) = 0.1 \end{aligned}$$

受験生にとっては、合格するまでは、自分が何判定であったか、自分が何点とれたか、 $P(A)$ から $P(E)$ の値、 $P(T|A)$ から $P(T|E)$ の値が重要である。でも、3月10日に、東大本郷キャンパスの正門から入った後に掲示される掲示板で自分の受験番号があるのを確認した後はどうでもいい。諸手続きを済ませ、オリ合宿が始まると、とりあえず周りにいるクラスメートとの話のネタとして、センター試験の点数や $P(A|T)$ から $P(E|T)$ の値が話題になるだろう。そこで、上記のデータから $P(A|T)$ から $P(E|T)$ の値を求めてみることにする。

⁽⁹⁾ ここに書いた数値は、模試結果の返却の際に駿台から配布された資料をもとに私が(勝手に)計算した結果である。実際のデータではない。

例 4.3 (続き)

例 4.3 の続きとして、 $P(A|T)$ 、東大に合格した人で、東大模試で A 判定だった人の割合を計算する。

まず、条件つき確率の定義より、

$$P(A|T) = \frac{P(A \cap T)}{P(T)} \quad (4.17)$$

である。 $P(A \cap T) = P(T \cap A)$ なのは明らかである。確率の乗法公式より、

$$P(T \cap A) = P(T|A)P(A) = 0.9 \times 0.148 = 0.1332 \quad (4.18)$$

となる。一方、 $P(T)$ は全確率公式より、

$$\begin{aligned} P(T) &= P(T|A)P(A) + P(T|B)P(B) + P(T|C)P(C) + P(T|D)P(D) + P(T|E)P(E) \\ &= 0.9 \times 0.148 + 0.7 \times 0.078 + 0.5 \times 0.099 + 0.3 \times 0.167 + 0.1 \times 0.508 = 0.3382 \end{aligned} \quad (4.19)$$

と求められる。よって、 $P(A|T)$ 、すなわち、東大に合格した人で、東大模試で A 判定だった人の割合は、

$$P(A|T) = \frac{P(A \cap T)}{P(T)} = \frac{0.1332}{0.3382} = 0.394$$

となる。これより、東大に合格した人の約 4 割が東大模試で A 判定をとっていたことがわかる。

同様に計算すると、 $P(B|T) = 0.161$, $P(C|T) = 0.146$, $P(D|T) = 0.148$, $P(E|T) = 0.150$ と求めることができる。

以上の操作を一般化したものがベイズの定理 (Bayes' theorem) である。

ベイズの定理

標本空間 Ω が、互いに排反な n 個の事象 H_1, H_2, \dots, H_n の和でかけるとき、つまり、 $\Omega = H_1 \cup H_2 \cup \dots \cup H_n$ のとき、任意の事象 E に対して、E が起きたという条件下で H_k が起こる条件つき確率は以下のようになる。

$$\begin{aligned} P(H_k|E) &= \frac{P(E \cap H_k)}{P(E)} \\ &= \frac{P(E|H_k) \cdot P(H_k)}{P(E|H_1) \cdot P(H_1) + P(E|H_2) \cdot P(H_2) + \dots + P(E|H_n) \cdot P(H_n)} \end{aligned} \quad (4.20)$$

上記の定理は、E という事象の起きる原因として n 個の原因 H_1, H_2, \dots, H_n が考えられるという状況において、各原因ごとに事象 E が起きる確率 $P(E|H_k)$ と、各原因の発生確率 $P(H_k)$ が分かっているということを仮定している。この仮定のもとで、結果から原因 $P(H_k|E)$ を探るという、逆問題の形をしているのがベイズの定理である。例 4.3 では東大合格を「結果」、模試の判定を「原因」としているのである。

4.3.5 事前確率と事後確率

例 4.3 より、 $P(A|T) = 0.394$, $P(B|T) = 0.161$, $P(C|T) = 0.146$, $P(D|T) = 0.148$, $P(E|T) = 0.150$ である。

ある日本人と外国人のハーフ、東京アレス太郎君が東大模試を受けたことがわかったとする。では、東京アレス太郎君が東大模試で A 判定をとった確率はいくらか。彼が東大に合格したという情報がなければ、A 判定を出す確率が 0.148 なので、当然 0.148 となるだろう。しかし、彼が東大に合格していたことがわかると、前のページの計算より、東大合格者の約 4 割（正確には 39.4%）が A 判定をとっているので、Harverd 君が A 判定をとった確率は 0.394 となる。

このように、情報を得る前後で、事象の確実性は変化する。情報を得る以前の確率 0.148 のことを事前確率（prior probability）、情報を得た後の確率 0.394 のことを事後確率（posterior probability）という。

4.4 事象の独立性

定義 4.6

複数の事象があるとき、それらの起こり方が互いに無関係の時、これらは独立であるという。事象 A と B が独立であることを数学的には次のように定義する。

$$P(A \cap B) = P(A) \cdot P(B) \quad (4.21)$$

事象の独立性は、「事象 B が起きた後に事象 A が起こる確率は、事象 A が単独で起こる確率に等しい」ということをいう。そのため、感覚的には次の式で定義した方がわかりやすいかもしれない。

$$P(A|B) = P(A) \quad (4.22)$$

で独立を定義することもできる。2つの定義は（当然）同値である。式 (4.21) と式 (4.22) は同値であることは以下のようにして示すことができる。

証明

- 式 (4.21) から式 (4.22) を示す。

条件つき確率の定義式（式 (4.14)）より、

$$P(A|B) \cdot P(B) = P(A \cap B) = P(A) \cdot P(B)$$

となるから、「 $P(A \cap B) = P(A) \cdot P(B) \implies P(A|B) = P(A)$ 」を示すことができた。

- 式 (4.22) から式 (4.21) を示す。

条件つき確率の定義式（式 (4.14)）より、

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

となるから、「 $P(A|B) = P(A) \implies P(A \cap B) = P(A) \cdot P(B)$ 」を示すことができた。

□

さて、この証明では「条件つき確率」を利用したが、条件つき確率の式の定義式は分数で与えられるので、分母に相当する $P(B)$ は 0 であってはならない。つまり、厳密には、 $P(B) > 0$ の時に限り、式 (4.21) と式 (4.22) は同値であるといえる。条件つき確率を利用して式 (4.22) で定義するときは、暗黙の前提として $P(B) > 0$ を仮定しているのである。とはいって、「事象 B が起きたことがわかっている」という条

件の下で、事象 A の起こる確率」を考える条件つき確率で、 $P(B) > 0$ は当然のことではないかと私は思う。B が起きていない時に条件つき確率を考えないのでと言いたくなってしまう。そのため、上の証明を厳密に正しいと考えても良いかも知れない。

A と B が独立ならば、事象 A が起きた後に事象 B が起こる確率は、事象 B が単独で起こる確率に等しい。つまり、 $P(B|A) = P(B)$ である。したがって、条件つき確率の定義は、A と B をひっくり返しても良い対称性を持っていて欲しい。そのため、式 (4.21) を条件つき確率の定義式とするのが良いといえる。

条件つき確率は独立性の判定で役に立つことがある。そこで、以下の性質を紹介する。

独立な事象の性質

事象 A と B が独立なとき、A と B が独立であることと

$$P(A|B) = P(A|B^c) \quad (4.23)$$

が成り立つことは同値である。

証明

- (A と B が独立であることを仮定した場合)
A と B は独立なので、式 (4.21) より、

$$P(A \cap B) = P(A)P(B), \quad P(A|B) = P(A)$$

が成立する。この時、

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \\ &= P(A)\{1 - P(B)\} \\ &= P(A)P(B^c) \end{aligned}$$

となるから、A と B^c は独立である。したがって、 $P(A|B^c) = P(A) = P(A|B)$ となる。

- 式 (4.23) が成り立つと仮定した場合
式 (4.23) は次のように変形できる。

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

余事象の性質より、 $P(B^c) = 1 - P(B)$ なので、これを上の式に代入して整理する。

$$\begin{aligned} \frac{P(A \cap B)}{P(B)} &= \frac{P(A \cap B^c)}{1 - P(B)} \\ (1 - P(B)) \cdot P(A \cap B) &= P(A \cap B^c) \cdot P(B) \\ P(A \cap B) &= P(B)\{P(A \cap B) + P(A \cap B^c)\} \end{aligned}$$

ここで、 $A = (A \cap B) \cup (A \cap B^c)$ であり、 $(A \cap B)$ と $(A \cap B^c)$ は互いに排反なので、

$$P(A \cap B) + P(A \cap B^c) = P((A \cap B) \cup (A \cap B^c)) = P(A)$$

となり、 $P(A \cap B) = P(A)P(B)$ が成立。よって、A と B は独立である。

□

また、2個の事象から n 個の事象に拡張した場合の独立性は次のように定義される。

定義 4.7

事象 E_1, E_2, \dots, E_n に対して、これらの事象から m 個($2 \leq m \leq n$)を選んで、 F_1, F_2, \dots, F_m とする。この時、任意の m および、 n 個の事象から m 個の事象を選ぶ任意の組み合わせに対して、以下の式が成立する時、事象 E_1, E_2, \dots, E_n は独立であるという。

$$P(F_1 \cap F_2 \cap \dots \cap F_m) = P(F_1) \cdot P(F_2) \cdots P(F_m) \quad (4.24)$$

第 5 章 離散型確率分布の性質

5.1 離散型確率変数

このTEXノートは教養学部時代に受講した「基礎統計」のレジュメや教科書をもとに作成している。ただ、確率変数の説明は、3年生のSセメスターで受講した「数理手法4」での説明が個人的には気に入っている。そこで、chapter5以降、確率変数の概念が登場する箇所は適宜「数理手法4」の説明を導入し、それに合わせて2年生のSセメスターで作成した内容を改良することにした。

5.1.1 確率変数

chapter5では、離散型確率変数 (discrete random variable) の種々の性質を調べていく。そこで、確率変数 (random variable) という概念を導入したい。このTEXノートでは、確率変数を難しそうな「確率論」の教科書にのっているような書き方で定義する。すなわち、確率や確率変数を集合から実数 \mathbb{R} への写像 (集合関数) の形で考えるとする。まずは、確率 (probability) というものを再定義する。

定義 5.1

標本空間 Ω の部分集合の集合 (べき集合 (Power set)) を \mathcal{F} と書くことにする。

写像 $P : \mathcal{F} \rightarrow [0, 1]$ が以下の性質を満たす時、 P は確率であるという。

- 全事象の確率は1である。一方、空事象の確率は0である。

$$P(\Omega) = 1, \quad P(\emptyset) = 0 \quad (5.1)$$

- $A, B \in \mathcal{F}$ が $A \cap B = \emptyset$ を満たす時、次の式が成立する。

$$P(A \cup B) = P(A) + P(B) \quad (5.2)$$

定義 5.2

- 確率変数 (random variable) とは、写像 $X : \Omega \rightarrow \mathbb{R}$ のことをいう。
- データは確率変数の実現値と定義される。

このように、写像として確率や確率変数を定義しても、その意味が全くわからないので、次の例で確認する。

例 5.1

理想的なサイコロを1回投げる。出る目は

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

である。この時、1の目がでる確率は

$$P(\{1\}) = \frac{1}{6}, \quad P(\{1, 2, 3\}) = \frac{1}{2}$$

とかける。また、 X をサイコロを1回振った時に出た目を表す確率変数とする。サイコロの目が1であるということは、 $1 \in \Omega$ に対して、 $X(1) = 1$ となることを表す。これは、1の目が出るという事象 $\{1\}$ を関数 X に入力したら、実数 1 が outputされたということを表している。

Y をサイコロを1回振った時に出た目の10倍を表す確率変数とする。この時は、 $1 \in \Omega$ に対して、 $Y(1) = 10$ となる。これは、1の目が出るという事象 $\{1\}$ を関数 X に入力したら、実数 10 が outputされたということを表している。

5.1.2 離散型確率変数

定義 5.3

標本空間 Ω が有限集合である時、確率変数 X を離散型確率変数 (**discrete random variable**) という。離散型確率変数 X の性質は、実現値 x_k を観測する確率を p_k がどのように分布しているかによって定めることができる。

$$P(\{\omega \mid X(\omega) = x_k\}) = p_k \quad (5.3)$$

式 (5.3) を確率変数 X の従う確率分布 (**probability distribution**) という。以下では、式 (5.3) を

$$P(X = x_k) = p_k \quad (5.4)$$

と略記することがある。

離散型確率変数の性質

X を離散型確率変数とする時、次の2つが成り立つ。

1. $P(X = x_k) \geq 0$
2. $\sum_k P(X = x_k) = 1$

証明

- 1について：確率は非負なので OK。

- 2について

X は x_1, x_2, \dots, x_n のいずれかの値をとるとすると、

$$P(\{X = x_1\} \cup \{X = x_2\} \cup \dots \cup \{X = x_n\}) = 1 \quad (5.5)$$

となる。ここで、各 k に対して、事象 $\{X = x_k\}$ は互いに排反なので、左辺は、 $\sum_{k=1}^n P(\{X = x_k\})$ と書ける。

□

例 5.2

コインを 10 回投げ、何回表が出るかを観測する。

表の出る回数を X とすると、 X は離散型確率変数となる。標本空間 Ω は、

$$\Omega = \{(x_1, x_2, \dots, x_{10}) \mid \{0, 1\} \in x_i \ (\forall i)\}$$

とかけて、 X のとりうる値は、0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 のどれかであり、

$$P(X = n) = \frac{\binom{10}{n}}{2^{10}} = \frac{\binom{10}{n}}{1024}$$

と確率分布を表現することができる。

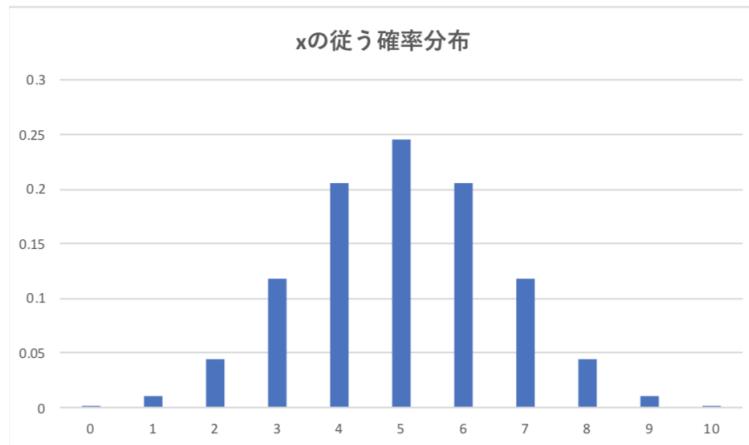


図 5.1: 表の出る回数 X の従う確率分布 (例 5.2)

5.2 期待値

確率変数 X の性質もデータの記述や要約で用いた平均や分散、標準偏差、基準化変量を導入することを考える。まず、期待値を導入する。

定義 5.4

離散型確率変数 X に対して、

$$\mu = E(X) = \sum_k x_k \cdot P(x = x_k) \quad (5.6)$$

で定義される量を平均 (mean)、または期待値 (expectation) という。

例 5.3

ロト 6 の当選金とその確率は次の表 5.1 になっている。宝くじを 1 枚購入したときの当選金を X 円とす

ると、 X は表 5.1 のような離散型確率分布で、 X の期待値 $E(X)$ は、

$$\begin{aligned} E(X) &= 200000000 \times \frac{1}{6096454} + 10000000 \times \frac{1}{1016076} \\ &\quad + 300000 \times \frac{1}{28224} + 6800 \times \frac{1}{610} + 1000 \times \frac{1}{39} = 81.21 \quad (5.7) \end{aligned}$$

となる⁽¹⁾。つまり、宝くじが 80 円以下で売られていない限り、宝くじを買うと損する可能性が高いといえる。

表 5.1: ロト 6 の当選金と当選確率

	1 等	2 等	3 等	4 等	5 等
当選金 (円)	2 億円	1000 万円	30 万円	6800 円	1000 円
確率	$\frac{1}{6096454}$	$\frac{1}{1016076}$	$\frac{1}{28224}$	$\frac{1}{610}$	$\frac{1}{39}$

定義 5.5

確率変数 X の取りうる値が x_1, x_2, \dots, x_n であるとする。確率変数 X と関数 $g(x) : \mathbb{R} \rightarrow \mathbb{R}$ を考える。このとき、確率変数 $Y : \Omega \rightarrow \mathbb{R}$ を次のように定める。

$X(\omega) = x_k$ を満たす $\omega \in \Omega$ に対して、 $Y(\omega) = g(x_k)$ とする。

このとき、確率変数 Y は、 $Y = g(X)$ と表記する。

さらに、離散型確率変数 $g(X)$ の期待値 $E[g(X)]$ は、次の式で定義される。

$$E[g(X)] = \sum_{k=1}^n g(x_k) \cdot P(X = x_k) \quad (5.8)$$

期待値の線型性

a, b を定数とする。このとき、次の 2 つの関係が成り立つ。

1. $E(aX + b) = aE(X) + b$
2. 任意の関数 $f(x)$ と $g(x)$ に対して、 $E(f(X) + g(X)) = E(f(X)) + E(g(X))$

証明

1. 期待値の定義 $E(X) = \sum_{k=1}^n x_k p_k$ と、確率変数の性質 $\sum_{k=1}^n p_k = 1$ ($p_k = P(X = x_k)$) に注意する。

$g(x) = ax + b$ とおくと、

$$\begin{aligned} E(aX + b) &= E(g(X)) = \sum_{k=1}^n g(x_k) p_k = \sum_{k=1}^n (ax_k + b)p_k \\ &= a \sum_{k=1}^n x_k p_k + b \sum_{k=1}^n p_k \\ &= aE(X) + b \end{aligned}$$

⁽¹⁾ $X = 0$ となる確率は計算するのは面倒だが、期待値の計算で、その確率に 0 をかけて足すので、計算の上では不必要である。

となる。

2. $h(x) = f(x) + g(x)$ とすると、

$$\begin{aligned} E(f(X) + g(X)) &= E(h(X)) = \sum_{k=1}^n h(x_k)p_k \\ &= \sum_{k=1}^n \{f(x_k) + g(x_k)\}p_k \\ &= \sum_{k=1}^n f(x_k)p_k + \sum_{k=1}^n g(x_k)p_k \\ &= E(f(X)) + E(g(X)) \end{aligned}$$

となる。

□

5.3 確率変数の散らばり

5.3.1 分散と標準偏差

確率変数でも、散らばりを表す指標として、分散と標準偏差を定義する。

定義 5.6

確率変数 X の取りうる値が x_1, x_2, \dots, x_n であるとする。また、 $p_k = P(X = x_k)$ とする。離散型確率変数 X の分散 (variance) $V(X)$ は、次の式で定義される。

$$\sigma^2 = V(X) = \sum_{k=1}^n (x_k - \mu)^2 p_k \quad (5.9)$$

つまり、期待値の記号を使うと、 $\sigma^2 = V(X) = E[(X - \mu)^2]$ と書ける。

また、 $V(X)$ の正の平方根を $D(X)$ とし、これを X の標準偏差 (standard deviation) と定義する。

$$\sigma = D(X) = \sqrt{V(X)} \quad (5.10)$$

確率変数の分散の性質

分散は次のように求めることができる。

$$\sigma^2 = V(X) = E(X^2) - \mu^2 \quad (5.11)$$

証明

期待値の線型性を利用する。

$$\begin{aligned} E((X - \mu)^2) &= E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + E(\mu^2) \\ &= E(X^2) - 2\mu \cdot \mu + \mu^2 \\ &= E(X^2) - \mu^2 \end{aligned}$$

□

次に、期待値の場合と同様に、確率変数 X を線型変換した $aX + b$ の分散と標準偏差を考える。

分散・標準偏差と線型性

a, b を定数とする。ただし、 $a > 0$ とする。この時、次の 2 つの関係が成立する。

$$V(aX + b) = a^2 \cdot V(X) \quad (5.12)$$

$$D(aX + b) = a \cdot D(X) \quad (5.13)$$

証明

$Y = aX + b$ とおく。すると、 $\mu_Y = E(Y)$ とすると、 $V(Y) = E[(Y - \mu_Y)^2]$ である。

期待値の線型性より、 $E(Y) = aE(X) + b = a\mu + b$ であるから、

$$\begin{aligned} V(Y) &= E[\{(aX + b) - (a\mu + b)\}^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2 \cdot E[(X - \mu)^2] \\ &= a^2 \cdot V(X) \end{aligned}$$

となる。この式の両辺の正の平方根をとると、 $D(Y) = a \cdot D(X)$ となる。 □

5.3.2 基準化変量

確率変数 X に対しても、基準化変量を定義する。

定義 5.7

確率変数 X の基準化変量 Z は、 X の期待値 μ と標準偏差 σ を用いて、

$$Z = \frac{X - \mu}{\sigma} \quad (5.14)$$

と定義する。また、期待値・分散と線型性の関係より、

$$E(Z) = 0 \quad (5.15)$$

$$V(Z) = 1 \quad (5.16)$$

となる。

証明

期待値と分散を線型変換させる公式の a, b が $a = \frac{1}{\sigma}$, $b = -\frac{\mu}{\sigma}$ となっている場合を考えると、

$$E(Z) = \frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$$

$$V(Z) = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$$

□

5.4 ベルヌーイ試行と2項分布

5.4.1 ベルヌーイ試行

定義 5.8

2種類の可能な結果（「成功」or「失敗」）を生じる実験あるいは観測があり、それらが起こる確率をそれぞれ $p, 1-p$ とする。この実験（観測）を、同じ条件でかつ独立（各回の事象は互いに無関係）に n 回繰り返す試行を「成功確率 p 、長さ n のベルヌーイ試行（bernoulli trial）」という。

写像を使うと、 $\Omega = \{0, 1\}$ 、 $0 \leq p \leq 1$ に対して、写像 $P : \mathcal{F} \rightarrow \mathbb{R}$ (\mathcal{F} は Ω に対応するべき集合) が、

$$P(\{0\}) = 1 - p, \quad P(\{1\}) = p \quad (5.17)$$

となることが「成功確率 p 」に対応していて、この操作を各回が独立になるように n 回繰り返すことがベルヌーイ試行である。この「0」は「失敗」を観測するという事象、「1」は「成功」を観測するという事象に対応する。

例 5.4

（コイン投げ）

表が出る確率が p ($0 < p < 1$) であるようなコインを独立に n 回投げるという試行は、コインの表裏を「成功」や「失敗」と対応させると、上記のベルヌーイ試行の条件を満たしている。

5.4.2 2項分布

上記の成功確率 p 、長さ n のベルヌーイ試行を実行したとき、何回成功するかを表す確率変数を導入することを考えよう。この確率変数が従う分布は**2項分布**（binomial distribution）と呼ばれる。写像を使って書くと、2項分布の定義は以下のようになる。

定義 5.9

標本空間 Ω を次のように定める。

$$\Omega = \{(x_1, x_2, \dots, x_n) \mid \forall i; x_i = 0 \text{ or } 1\}$$

このとき、確率変数 $X : \Omega \rightarrow \mathbb{R}$ を次のように定義する。

$\omega = (x_1, x_2, \dots, x_n) \in \Omega$ に対して、 $x_i = 1$ となっている成分の個数を $X(\omega)$ とする。

成功確率 p 、長さ n のベルヌーイ試行に対して、第 i 回目の成功、失敗を x_i に対応させたとき、 n 回のうち何回成功したかを表す確率変数は、まさに上記の X であり、以下の式で表される分布に従う。

$$P(X = x) = {}_n C_x p^x (1 - p)^{n-x} \quad (5.18)$$

以下では、確率変数 X が式 (5.18) の分布に従うことを、

$$X \sim B(n.p) \quad (5.19)$$

と記すことにする。

ここで、今後使用する組み合わせに関する公式を 1 つ証明しておくことにする。

Remark

$$x \cdot {}_n C_x = n \cdot {}_{n-1} C_{x-1} \quad (5.20)$$

証明

$$x \cdot {}_n C_x = x \cdot \frac{n!}{x!(n-x)!} = \frac{n \cdot (n-1)!}{(x-1)! \cdot (n-x)!} = n \cdot {}_{n-1} C_{x-1}$$

□

2 項分布の平均と分散

確率変数 X が**2 項分布** $B(n,p)$ に従う ($X \sim B(n,p)$) とき、2 項分布の平均 $E(X)$ と分散 $V(X)$ は次のようにになる。

$$E(X) = np \quad (5.21)$$

$$V(X) = np(1-p) \quad (5.22)$$

証明

以下では、 $q = 1 - p$ とする。

$$\begin{aligned} E(X) &= \sum_{x=0}^n x \cdot {}_n C_x p^x q^{n-x} = \sum_{x=1}^n x \cdot {}_n C_x p^x q^{n-x} \\ &= \sum_{x=1}^n n \cdot {}_{n-1} C_{x-1} p^x q^{n-x} \end{aligned}$$

ここで、 $m = n - 1$, $y = x - 1$ とおくと

$$\begin{aligned} &= n \sum_{y=0}^m {}_m C_y p^{y+1} q^{m-y} \\ &= np \sum_{y=0}^m {}_m C_y p^y q^{m-y} \\ &= np(p+q)^m = np \end{aligned}$$

次に、式 (5.11) より、 $V(X) = E(X^2) - \{E(X)\}^2$ である。期待値の線型性より、

$$V(X) = E\{X(X-1)\} + E(X) - \{E(X)\}^2 \quad (5.23)$$

と書ける。期待値を求めた場合と同様に式処理をする。

$$\begin{aligned} E\{X(X-1)\} &= \sum_{x=0}^n x(x-1) \cdot {}_n C_x p^x q^{n-x} \\ &= \sum_{x=2}^n (x-1) \cdot n \cdot {}_{n-1} C_{x-1} p^x q^{n-x} \\ &= \sum_{x=2}^n n(n-1) \cdot {}_{n-2} C_{x-2} p^x q^{n-x} \end{aligned}$$

ここで、 $m = n - 2$, $y = x - 2$ とおくと

$$\begin{aligned} &= n(n-1) \sum_{y=0}^m {}_m C_y p^{y+2} q^{m-y} \\ &= n(n-1)p^2 \sum_{y=0}^m {}_m C_y p^y q^{m-y} \\ &= n(n-1)p^2(p+q)^m = n(n-1)p^2 \end{aligned}$$

よって、 $V(X) = n(n-1)p^2 + np - (np)^2 = np - np^2 = np(1-p)$ となる。 \square

5.5 ポアソン分布

例 5.5

東京大学の 2018 年度入試の合格者数は、次のようになっている。

文 1	文 2	文 3	理 1	理 2	理 3	合計
404	361	472	1130	549	98	3014

全科類合わせて 3014 人が入学した東京大学の新 1 年生の中に、1 月 1 日生まれの人が 1 人もいない確率を求めよ。ただし、1 年は 365 日とする。

(解答)

ある 1 人の誕生日が 1 月 1 日である確率は $\frac{1}{365}$ である。すると、新 1 年生全員を 1 月 1 日生まれかそうでないかの 2 グループに分ける問題を考えることになる。新 1 年生の中で、1 月 1 日生まれの人数を表す確率変数を X とすると、 X は $n = 3014$, $p = 1/365$ の 2 項分布 $B\left(3014, \frac{1}{365}\right)$ に従う。よって、求める確率は

$$P(X = 0) = {}_{3014}C_0 \left(\frac{1}{365}\right)^0 \times \left(\frac{364}{365}\right)^{3014} = 0.00026 = 2.6 \times 10^{-4}$$

となる。つまり、求める確率は 0.026% であり、3014 人も新入生がいれば、1 月 1 日生まれの人が少なくとも 1 人はいる可能性が高いと予想できる。

では、この問題が、「1 月 1 日生まれの人が 100 人いる確率を求めよ。」であったらどうするか。「東大生の 30 人に 1 人は 1 月 1 日生まれなんてありえないから確率は 0 に決まってるよ。100 以上あるクラス全てにおいて誰か 1 人は 1 月 1 日生まれって状況を考えるということでしょ。どう考えてもそんな状況はありえないでしょ。」と答えるても何の問題もない気がする。ただ、今回はそういうのは NG として真面目に何 % かを考える。その場合は、

$$P(X = 100) = {}_{3014}C_{100} \left(\frac{1}{365}\right)^{100} \times \left(\frac{364}{365}\right)^{2914}$$

で求められるが、これを計算するの面倒である。 n と p が次の 2 つの特徴をもつからである。

- n が非常に大きい。(上の場合、 $n = 3014$)
- p が非常に小さい。(上の場合、 $n = 0.0027 = 2.7 \times 10^{-3}$)

そのせいで、 ${}_{3014}C_{100}$, $(1/365)^{100}$, $(364/365)^{2914}$ を計算するのは面倒である。要するに、 n と p が極端すぎるるのである。ただ、 $np = \frac{3014}{365} = 8.26$ と np は中程度である。このような時は、2項分布の極限として定義されるポアソン分布 (**Poisson distribution**) が役に立つ。

定義 5.10

確率変数 X が次の確率分布を持つ時、 X はポアソン分布 $Po(\lambda)$ に従うといい、 $X \sim Po(\lambda)$ と書く。

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad (5.24)$$

(ただし、 $x = 0, 1, \dots$, $\lambda > 0$ とする。)

2項分布の計算がしやすいように導入したのがポアソン分布である。そのため、2項分布からポアソン分布の式を導くことを考える。

2項分布の極限によるポアソン分布の定義

2項分布 $B(n, p)$ において、 $np = \lambda$ を一定に保ちながら、 $n \rightarrow \infty$ の極限をとると、

$${}_nC_x p^x (1-p)^{n-x} \rightarrow \frac{e^{-\lambda} \cdot \lambda^x}{x!} \quad (5.25)$$

となる。

証明

$p = \lambda/n$ ので、

$$\begin{aligned} {}_nC_x p^x (1-p)^{n-x} &= \frac{n!}{(n-x)! \cdot x!} \cdot \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{\lambda^x}{x!} \times \frac{n!}{(n-x)! \cdot n^x} \times \left(1 - \frac{\lambda}{n}\right)^n \times \left(1 - \frac{\lambda}{n}\right)^{-x} \end{aligned}$$

となる。ここで、 $n \rightarrow \infty$ の極限をとると、

$$\begin{aligned} \frac{n!}{(n-x)! \cdot n^x} &= \frac{n \cdot (n-1) \times \cdots \times (n-x+1)}{n^x} \\ &= \frac{n}{n} \times \left(1 - \frac{1}{n}\right) \times \cdots \times \left(1 - \frac{x-1}{n}\right) \rightarrow 1 \\ \left(1 - \frac{\lambda}{n}\right)^n &= \left(1 + \frac{1}{-\frac{\lambda}{n}}\right)^{(-\frac{n}{\lambda}) \times (-\lambda)} \rightarrow e^{-\lambda} \\ \left(1 - \frac{\lambda}{n}\right)^{-x} &\rightarrow 1 \end{aligned}$$

となるから、 $n \rightarrow \infty$ の極限はポアソン分布になる。 \square

次に、ポアソン分布の平均と分散を考える。 $np = \lambda$ の条件の下、極限を考えたのだから、2項分布の時の平均 np がポアソン分布では λ となるだろう。分散については、2項分布では $np(1-p)$ だった。 $np = \lambda$ の条件下で、 $n \rightarrow \infty$ の極限をとるので、 $p \rightarrow 0$ となる。すると、 $1-p \rightarrow 1$ なので、分散も λ となると予想できる。このことを計算で確かめる。

ポアソン分布の平均と分散

確率変数 X がポアソン分布に従う時、 X の平均と分散は次のようになる。

$$E(X) = \lambda \quad (5.26)$$

$$V(X) = \lambda \quad (5.27)$$

証明

まず、平均(期待値)が λ であることを示す。

$$E(X) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda} \cdot \lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}$$

$y = x - 1$ と変数変換をすると、

$$\begin{aligned} &= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda \end{aligned}$$

続けて、ポアソン分布の分散が λ となることを示す。2項分布の場合と同様に次の式(5.23)を使う。

$$V(X) = E[X(X-1)] + E(X) - \{E(X)\}^2 \quad (5.23)$$

ということで、 $E[X(X-1)]$ を求める。

$$\begin{aligned} E[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1) \cdot \frac{e^{-\lambda} \cdot \lambda^x}{x!} \\ &= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} \end{aligned}$$

$y = x - 2$ と変数変換をすると、

$$\begin{aligned} &= \lambda^2 e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \\ &= \lambda^2 e^{-\lambda} \cdot e^{\lambda} = \lambda^2 \end{aligned}$$

よって、 $V(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$ となる。 \square

5.6 幾何分布

5.6.1 無限級数に関する公式

Remark

x が $0 < x < 1$ を満たすとき、以下の無限級数は収束する。

$$\sum_{n=1}^{\infty} nx^n = \frac{x}{(1-x)^2} \quad (5.28)$$

証明

(その1) 高校数学で習う方法で示す。

$$\begin{aligned} S_n &= 1 \cdot x + 2 \cdot x^2 + 3 \cdot x^3 + \cdots + n \cdot x^n \\ xS_n &= \qquad 1 \cdot x^2 + 2 \cdot x^3 + \cdots + (n-1)x^n + nx^{n+1} \end{aligned}$$

よって、辺々引くと、

$$\begin{aligned} (1-x)S_n &= x + (x^2 + x^3 + \cdots + x^n) - nx^{n+1} \\ S_n &= \frac{1}{1-x} \sum_{k=1}^n x^k - \frac{nx^{n+1}}{1-x} \\ &= \frac{1}{1-x} \cdot \frac{x(1-x^n)}{1-x} - \frac{nx^{n+1}}{1-x} \\ &= \frac{x}{(1-x)^2} - \frac{x^{n+1}}{(1-x)^2} - \frac{x}{1-x} \cdot (nx^n) \end{aligned}$$

となり、 $0 < x < 1$ ならば、 $n \rightarrow \infty$ のとき、

$$\lim_{n \rightarrow \infty} S_n = \frac{x}{(1-x)^2}$$

(その2) 項別微分の考え方を利用する。

項別微分が可能かどうかを厳密に議論するところから本当はスタートすべきなのだろうが、今回はその議論は飛ばす。 $0 < x < 1$ のとき、

$$1 + x + x^2 + \cdots = \sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad (5.29)$$

が成立する。この両辺を x で微分すると、

$$\sum_{n=0}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2} \quad (5.30)$$

となる⁽²⁾。この両辺に x をかける。 $n = 0$ の時、 $nx^n = 0$ なので、和を $n = 1$ からにしてもよいので、

$$\sum_{n=1}^{\infty} nx^n = \frac{x}{(1-x)^2}$$

(2) $1 + x + x^2 + \cdots$ を微分した時に x^{-1} の項は登場しないので、和を取るときは、 $n = 1$ からでないとおかしいと考える人もいるだろう。ただ、私の微分積分学の教官は、 nx^{n-1} は $n = 0$ の時に 0 になって、和に影響を与えないから、和をとるのは $n = 0$ からでも良いのではないかといっていた。このTeXノートでは後者の考え方を採用する。

としてよい。 □

Remark

x が $0 < x < 1$ を満たすとき、以下の無限級数は収束する。

$$\sum_{n=1}^{\infty} n^2 x^n = \frac{x(x+1)}{(1-x)^3} \quad (5.31)$$

証明

(その1) 高校数学で習う方法で示す。

$$\begin{aligned} T_n &= 1 \cdot x + 4 \cdot x^2 + 9 \cdot x^3 + \cdots + n^2 \cdot x^n \\ xT_n &= \qquad \qquad 1 \cdot x^2 + 4 \cdot x^3 + \cdots + (n-1)^2 x^n + n^2 x^{n+1} \end{aligned}$$

よって、辺々引くと、

$$\begin{aligned} (1-x)T_n &= x + 3x^2 + 5x^3 + \cdots + (2n-1)x^n - n^2 x^{n+1} \\ &= \sum_{k=1}^n (2k-1)x^k - n^2 x^{n+1} \\ &= 2S_n - \sum_{k=1}^n x^k - n^2 x^{n+1} \end{aligned}$$

となる。(ただし、 S_n は前のページで定義したものと同一である。)

すると、 T_n は以下のようになる。

$$T_n = \frac{2x}{(1-x)^3} - \frac{2x^{n+1}}{(1-x)^3} - \frac{2x}{(1-x)^2} \cdot (nx^n) - \frac{x(1-x^n)}{(1-x)^2} - \frac{x}{1-x} \cdot (n^2 x^n)$$

よって、 $0 < x < 1$ ならば、 $n \rightarrow \infty$ のとき、

$$\lim_{n \rightarrow \infty} T_n = \frac{2x}{(1-x)^3} - \frac{x}{(1-x)^2} = \frac{x(x+1)}{(1-x)^3}$$

(その2) 項別微分の考え方を利用する。

ここでも、項別微分が可能かどうかの議論は飛ばす。 $0 < x < 1$ のとき、式 (5.30) のように nx^{n-1} の無限和が表される。

$$\sum_{n=0}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2} \quad (5.30)$$

が成立する。この両辺を x で微分すると、

$$\sum_{n=0}^{\infty} n(n-1)x^{n-2} = \frac{2}{(1-x)^3} \quad (5.32)$$

となる。一方、式 (5.30) より、

$$\sum_{n=0}^{\infty} nx^{n-2} = \frac{1}{x(1-x)^2} \quad (5.33)$$

が成立する。よって、式(5.32)と式(5.33)より、

$$\sum_{n=0}^{\infty} n^2 x^{n-2} = \frac{2}{(1-x)^3} + \frac{1}{x(1-x)^2} = \frac{x+1}{x(1-x)^3} \quad (5.34)$$

となるから、両辺に x^2 をかけて、さらに $n=0$ のとき $n^2 x^n = 0$ を考慮すると、

$$\sum_{n=1}^{\infty} n^2 x^n = \frac{x(x+1)}{(1-x)^3} \quad (5.35)$$

と式(5.31)を導くことができる。 \square

5.6.2 幾何分布の平均と分散

成功確率 p 、長さ n のベルヌーイ試行を実行した時に、何回成功したかを表す確率変数が従う分布が2項分布であった。今度は、成功確率 p を試行回数を決めずに何度も繰り返すことを考える。このようなベルヌーイ試行において、初めて成功するのが何回を表す確率変数が従う分布を今度は考える。その分布が幾何分布 (geometric distribution) である。幾何分布はある事象が起こるまでの時間 (待ち時間) を表すのに適した確率分布である。

定義 5.11

確率変数 X が次の確率分布を持つ時、 X は幾何分布 $Ge(p)$ に従うといい、 $X \sim Ge(p)$ と書く。

$$P(X=x) = p(1-p)^{x-1} \quad (5.36)$$

(ただし、 $x=1, 2, \dots$ とする。)

ポアソン分布の平均と分散

確率変数 X が幾何分布に従う時、 X の平均と分散は次のようになる。

$$E(X) = \frac{1}{p} \quad (5.37)$$

$$V(X) = \frac{1-p}{p^2} \quad (5.38)$$

証明

$q = 1 - p$ とする。すると、 $P(X=x) = pq^{x-1}$ となり、期待値 $E(X)$ は次のようにになる。

$$E(X) = \sum_{x=1}^{\infty} x \cdot pq^{x-1}$$

ここで、次のように変形して、無限級数に関する公式を使う。

$$\begin{aligned} E(X) &= \frac{p}{q} \sum_{x=1}^{\infty} x \cdot q^x = \frac{p}{q} \cdot \frac{q}{(1-q)^2} \\ &= \frac{p}{(1-q)^2} \\ &= \frac{p}{p^2} = \frac{1}{p} \end{aligned}$$

次に、分散 $V(X)$ を考える。式 (5.11) を使って、 $V(X)$ を変形する。

$$V(X) = E(X^2) - \{E(X)\}^2$$

$E(X^2)$ について、無限級数に関する公式を使う。

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{\infty} x^2 \cdot pq^{x-1} = \frac{p}{q} \sum_{x=1}^{\infty} x^2 \cdot q^x = \frac{p}{q} \cdot \frac{q(q+1)}{(1-q)^3} \\ &= \frac{p}{q} \cdot \frac{q(2-p)}{p^3} \\ &= \frac{2-p}{p^2} \end{aligned}$$

ゆえに、求める分散 $V(X)$ は、

$$V(X) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}$$

となる。 \square

5.6.3 幾何分布の性質

幾何分布の確率計算

確率変数 X が幾何分布 $Ge(p)$ に従うとき、

$$P(X > a) = q^a \quad (5.39)$$

(ただし、 $q = 1 - p$ とする。)

証明

余事象 $\{\omega \mid X(\omega) \leq a\}$ の起こる確率を求める。

$$\begin{aligned} P(X \leq a) &= \sum_{x=1}^a P(X = x) = \sum_{x=1}^a pq^{x-1} \\ &= p \cdot \frac{1 - q^a}{1 - q} \\ &= p \cdot \frac{1 - q^a}{p} \\ &= 1 - q^a \end{aligned}$$

余事象の確率が $1 - q^a$ なので、求める確率は q^a である。 \square

この chapter の最後に、幾何分布が待ち時間を表す分布であることについて考える。離散型確率分布なので、時間を離散時間 (Discrete time)で考える。つまり、時間を $0, 1, 2, 3, \dots$ と考える。 $X \sim Ge(p)$ ならば、 $E(X) = 1/p$ となることから、発生確率が p ならば、平均的に $1/p$ 待つということを幾何分布は表している。

例 5.6

事象 A が起こるまでの待ち時間を表す確率変数を X とする。例えば、事象 A が「あるコンビニ M において、次の客が来る」という事象とする。つまり、 X を客の来店間隔とする。この時、事象 A の起こり方がデタラメならば、それが起こるまでの待ち時間 X も当然デタラメとなる。

もう少し具体的に考える。コンビニ店員の太郎は、すでに 10 分新しい客が来ていないことを知っている。そこに、店員の花子がきて、太郎と花子は一緒にレジに立った。では、花子がきてから 5 分以内に次の客がくる確率はいくらか。

- (太郎の立場) : $P(X = 15 | X > 10)$

「10 分がお客様が来店していない」という条件を $X > 10$ と記した。その下で、5 分待つと、太郎は 15 分待っていることになるので、条件付き確率を用いて表す必要がある。

- (花子の立場) : $P(X = 5)$

お客様が来店する間隔がデタラメなら、10 分待っていたからといって、5 分以内に客がくる確率が上がるということはない。よって、次の式が成立する。

$$P(X = 15 | X > 10) = P(X = 5)$$

幾何分布と無記憶性

幾何分布は無記憶性を持つ。すなわち、 $X \sim Ge(p)$ のとき、次の式が成立する。

$$P(X = a + b | X > b) = P(X = a) \quad (5.40)$$

(ただし、 $a, b = 1, 2, \dots$ とする。)

証明

(左辺) は、条件つき確率の定義より、

$$P(X = a + b | X > b) = \frac{P\{(X = a + b) \cap (X > b)\}}{P(X > b)}$$

となる。 $a > 0$ なので、 $\{(X = a + b) \cap (X > b)\} = (X = a + b)$ である。よって、

$$\begin{aligned} P(X = a + b | X > b) &= \frac{P(X = a + b)}{P(X > b)} = \frac{pq^{a+b-1}}{q^b} \\ &= pq^{a-1} \\ &= P(X = a) \end{aligned}$$

□

第 6 章 連続型確率分布の性質

確率分布で最も重要な正規分布の標本空間 Ω は、離散的な値でかつ有限であるのではなく、実数 \mathbb{R} 全体であり、連続的である。このような枠組みで確率変数を捉える時は、離散的な場合とは異なる手法を採用する。連続的な確率変数を用いた方が良い場合は身長や体重、気温など多岐にわたる。この chapter では連続型確率分布の基本的な取り扱い方を紹介する。

6.1 連続型確率分布の導入

第 4 章の脚注 (1)(44 ページ) で、私の家では、平日の朝はめざましテレビのかやちゃんの天気予報を見ていることを記した。毎朝、今日の全国各地の予想最高気温は何 °C と画面に出てくる。そして、夕方のニュース番組になると、「今日は暑かったですね～～」的なことを「明日の天気予報」の間に気象予報士の方がいう。そこで、「今日の都心の最高気温は何 °C でした」的なこともいうだろう。

気象予報士の方がいう温度は離散的な値である。2019 年 7 月 1 日から 7 月 10 日の東京の最高気温は、表 6.1 のようである。0.1 °C 刻みで温度が記されているが、これは実際の値であるか。もちろん No である。温度は、実際は連続的に変化する量であるが、温度計は真の値に非常に近い値しか表示できない。温度のように真の値は連続的に変化する値(量)については、離散的な確率変数を用いるのは不適切である。

表 6.1: 2019 年 7 月 1 日から 7 月 10 日の東京の最高気温

7/1	7/2	7/3	7/4	7/5	7/6	7/7	7/8	7/9	7/10
24.3	28.1	29.1	25.2	24.5	23.7	20.8	24.8	21.8	24.8

6.1.1 連続型確率変数とは？

温度を無限に精度の高い温度計で測定できるとしたら、私たちは温度の真の値を知ることができる。しかし、そのような温度計は理想的なものであり存在しない。でも、本当にそのような温度計が存在して、真の値を入手できるなら、私たちは度数分布表を作ることができるだろう。その度数分布表はどうなっているだろうか。無限桁の測定値が完全に一致することはありえないと考えられるので、どの値の度数も 1 である。

どれも度数が 1 というのは嬉しいことではない。そこで、私たちは階級幅を設定して、「 a と b の間に真の値が存在する」という風にデータを取り扱う。温度のような連続的に値が変化する量に対しては、このように区間を用いて議論する必要がある。

確率変数 X が連続の値をとる時、 X を連続型確率変数といわれる。身長や体重、気温や為替レートなどは連続型確率変数である。連続型確率変数の場合は、 $X = a$ といった具体的な値より、 X が a と b の間にある といった区間の方が重要となる。確率変数が連続的という概念を導入するために、分布関数という概念をまず導入する。

定義 6.1

確率変数 X について、 $X \leq a$ となる確率 $P(X \leq a)$ を x の関数とみて $F(a)$ とかき、 X の分布関数 (**distribution function**) という。

例 6.1

離散型確率変数の場合、分布関数は

$$F(a) = P(X \leq a) = \sum_{x_i \leq a} P(X = X_i)$$

と書ける。

定義 6.2

確率変数 X について、 X の分布関数 $F(x) = P(X \leq x)$ が任意の $x \in (-\infty, +\infty)$ で連続であるとき、確率変数 X は連続型確率変数 (**continuous random variable**) であるといい、 X の確率分布を連続型確率分布 (**continuous distribution**) という。

分布関数を利用することで、連続型確率変数 X が a と b の間にある確率 $P(a < X \leq b)$ を定義することができる。 $a \leq b$ ならば、

$$\{\omega \mid X(\omega) \leq b\} = \{\omega \mid X(\omega) \leq a\} \cup \{\omega \mid a < X(\omega) \leq b\}$$

という性質と、 $\{\omega \mid X(\omega) \leq a\}$ と $\{\omega \mid a < X(\omega) \leq b\}$ が互いに排反であることから、

$$P(a < X \leq b) = F(b) - F(a) \quad (6.1)$$

と書ける。こうして、分布関数により、連続型確率変数 X がある区間にある確率を定義することができる。

式 (6.1)において、 $b \rightarrow a$ の極限を考えよう。左辺は $P(X = a)$ に近づこうとして、右辺は 0 に近づく。そのため、連続型の確率変数がある 1 つの特定の値をとる確率は 0 であるといえる。この他にも、分布関数の定義と確率の性質から次の 3 つの性質がわかる。この 3 つの性質は確率変数が離散型か連続型に依らず成立する。

分布関数の性質

- 分布関数は非減少関数である。すなわち、 $a < b$ なら $F(a) \leq F(b)$ である。
(確率が非負であることから上と同様に考えることで示せる。)
- $\lim_{x \rightarrow -\infty} F(x) = 0$ ($x \rightarrow -\infty$ とすると、 $X \leq x$ という事象は空事象。)
- $\lim_{x \rightarrow \infty} F(x) = 1$ ($x \rightarrow \infty$ とすると、 $X \leq x$ という事象は全事象。)

6.1.2 密度関数と分布関数

微分積分学の基本的知識として、「微分可能な関数は連続である」という性質がある。一方、「連続な関数は必ずしも微分可能ではない」という性質もある。分布関数が連続である場合について、その分布関数 $F(x)$ が微分可能であると仮定する。そうすると、分布関数は

$$F(x) = \int_{-\infty}^x f(u) du$$

と書けるはずである。

定義 6.3

連続型確率変数 X の分布関数 $F(x)$ と \mathbb{R} 全体で非負の関数 $f(x)$ の間に、次の関係式が成立するとき、 $f(x)$ を(連続型)確率変数 X の確率密度関数 (probability density function) という。

$$F(a) = P(X \leq a) = \int_{-\infty}^a f(x) dx \quad (6.2)$$

微分積分学の基本定理から、 $F'(x) = f(x)$ が成立することがわかる。

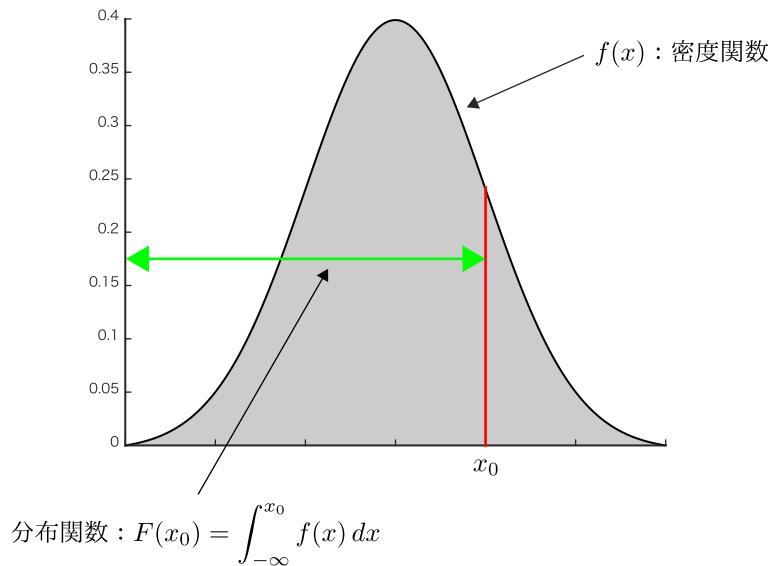


図 6.1: 分布関数と確率密度関数

式 (6.2) のように、確率密度関数 $f(x)$ を定義すると、連続型確率変数 X が a と b の間にある確率 $P(a \leq X \leq b)$ は、

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad (6.3)$$

と書ける。 $F(b) - F(a) = P(a < X \leq b)$ だが、連続型確率変数の場合、確率変数がある特定の 1 つの値になるという確率 $P(X = a)$ は 0 と考えてよいので、 $P(a \leq X \leq b) = P(a < X \leq b)$ とした。このように連続型確率変数の場合、区間の端点の厳密性はあまり重要ではなく、

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

と考えてよい。

連続型確率分布について、確率としての性質が成り立つように、密度関数に次のような制約を設ける。

1. 密度関数 $f(x)$ の定義域は、 $-\infty \leq x \leq +\infty$ とし、この定義域全体で非負の値をとる。
2. 確率変数 X が任意の $\omega \in \Omega$ に対して値をとらない場合は、 $f(x) = 0$ とする。
3. $f(x)$ が 0 以外の値をとる区間では、密度関数は必ず連続である。
(→ そうでないと、分布関数 $F(x)$ が連続にならない)
4. $P(-\infty \leq X \leq +\infty) = 1$ が成り立つ。すなわち、 $\int_{-\infty}^{\infty} f(x) dx = 1$ である。

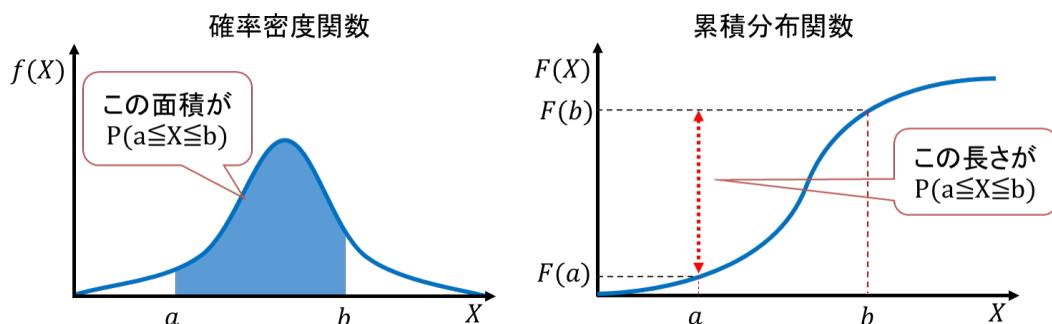
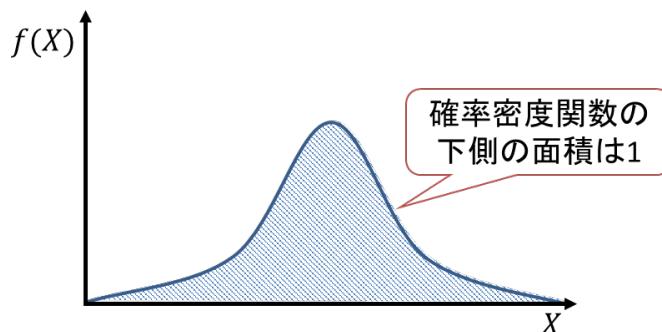


図 6.2: 確率密度関数と分布関数、確率の関係
(<https://bellcurve.jp/statistics/course/6710.html> より)

6.1.3 連続型確率変数の例題 (1)

例 6.2

連続型確率変数 X が、次のような密度関数 $f(x)$ を持つ時、 X の分布関数 $F(x)$ を求めよ。

$$f(x) = \begin{cases} 2x & (0 < x < 1) \\ 0 & (\text{それ以外}) \end{cases}$$

(解答)

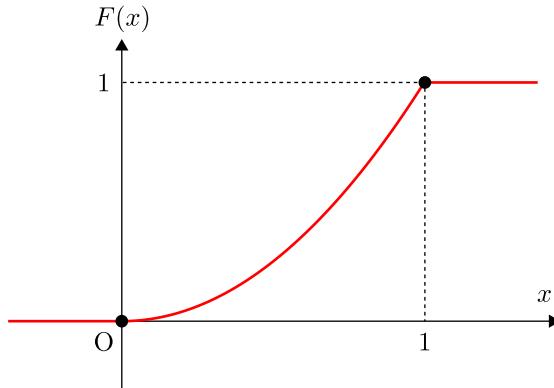
- $x \leq 0$ の時、 $F(x) = 0$

- $0 \leq x \leq 1$ の時、

$$F(x) = \int_{-\infty}^x f(x) dx = \int_0^x 2x dx = x^2$$

- $x \geq 1$ の時、

$$F(x) = \int_{-\infty}^x f(x) dx = \int_0^1 2x dx = 1$$



6.1.4 連続型確率変数の平均と分散

定義 6.4

連続型確率変数 X が密度関数 $f(x)$ を持つ時、 X の平均 $E(X)$ と分散 $V(X)$ は次の式で定義される。

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx \quad (6.4)$$

$$V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \quad (6.5)$$

(ただし、 $\mu = E(X)$ である。)

定義 6.5

連続型の確率変数 X の関数 $g(X)$ の期待値 $E[g(X)]$ は

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x) dx \quad (6.6)$$

で定義される。

さらに、離散型の時に示した「期待値の線型性」は、連続型でも成立する。これは、積分の線型性による。そのため、期待値と分散は離散型と同様に、

$$V(X) = E(X^2) - \{E(X)\}^2 \quad (6.7)$$

の関係で結ばれている。

例 6.3

連続型確率変数 X が、次のような密度関数 $f(x)$ を持ち、その平均(期待値)が $\frac{2}{3}$ となる。この時、定数 a, b の値と分散 $V(X)$ を求めよ。

$$f(x) = \begin{cases} a + bx^2 & (0 < x < 1) \\ 0 & (\text{それ以外}) \end{cases}$$

(解答)

密度関数の性質 $\int_{-\infty}^{\infty} f(x) dx = 1$ と $E(X) = \int_{-\infty}^{\infty} xf(x) dx = \frac{2}{3}$ が成り立つように a, b を決めれば良い。

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx = 1 &\iff \int_0^1 (a + bx^2) dx = 1 \iff a + \frac{1}{3}b = 1 \\ \int_{-\infty}^{\infty} xf(x) dx = \frac{2}{3} &\iff \int_0^1 (ax + bx^3) dx = \frac{2}{3} \iff \frac{1}{2}a + \frac{1}{4}b = \frac{2}{3} \end{aligned}$$

よって、 $\boxed{a = \frac{1}{3}, b = 2}$ である。つまり、 $0 < x < 1$ では、 $f(x) = \frac{1}{3} + 2x^2$ である。

次に分散を求める。

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 \left(\frac{1}{3}x^2 + 2x^4 \right) dx = \left[\frac{1}{9}x^3 + \frac{2}{5}x^5 \right]_0^1 = \frac{23}{45}$$

ゆえに、求める分散は、

$$V(X) = E(X^2) - \{E(X)\}^2 = \frac{23}{45} - \left(\frac{2}{3} \right)^2 = \boxed{\frac{1}{15}}$$

6.2 一様分布

ここから、連続型確率変数の分布をいくつか紹介する。一様分布は、ある区間のどの値も等しい確からしさでとるような連続型確率変数の分布である。

定義 6.6

確率密度関数 $f(x)$ が定数 α, β ($\alpha < \beta$) に対して、

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & (\alpha \leq x \leq \beta) \\ 0 & (\text{それ以外}) \end{cases} \quad (6.8)$$

と与えられる分布を一様分布 (uniform distribution) といい、 $U(\alpha, \beta)$ とかく。

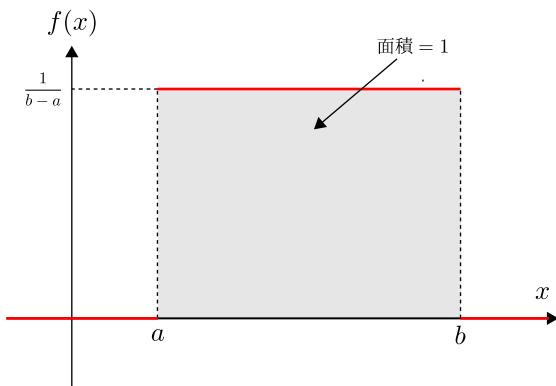


図 6.3: 一様分布の確率密度関数

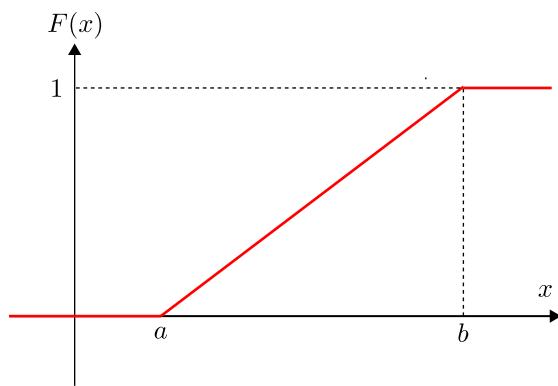


図 6.4: 一様分布の分布関数

一様分布の密度関数は図 6.2 のようなので、分布関数は次のようになる。

$$F(x) = \begin{cases} 0 & (x < \alpha) \\ \frac{x - \alpha}{\beta - \alpha} & (\alpha \leq x \leq \beta) \\ 1 & (x > \beta) \end{cases} \quad (6.9)$$

続けて、平均と分散を求めよう。式 (6.4) より平均は、

$$E(X) = \int_a^b x \cdot \frac{1}{\beta - \alpha} dx = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2} \quad (6.10)$$

となる。また、 $E(X^2)$ は次のようにになる。

$$E(X^2) = \int_a^b x^2 \cdot \frac{1}{\beta - \alpha} dx = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} = \frac{\alpha^2 + \alpha\beta + \beta^2}{3}$$

よって、分散 $V(X)$ は次のようにになる。

$$V(X) = E(X^2) - \{E(X)\}^2 = \frac{\alpha^2 + \alpha\beta + \beta^2}{3} - \left(\frac{\alpha + \beta}{2}\right)^2 = \frac{(\beta - \alpha)^2}{12} \quad (6.11)$$

6.3 正規分布

統計分析において最も重要な分布は正規分布 (normal distribution) である。正規分布は左右対称の釣鐘型の密度関数をもつ分布で、多くの現象は正規分布で近似できる場合が多い。そして、正規分布は、この後の章で述べる中心極限定理とも関係が深い。

6.3.1 Gauss 積分

正規分布の各種性質を調べるために、Gauss 積分について理解する必要がある。この subsection では、正規分布を紹介する前に Gauss 積分を紹介する。この積分の結果と公式は必ず暗記しておきたいものである。

Remark

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi} \quad (6.12)$$

証明

この積分を実行するために、 $f(x, y) = e^{-x^2-y^2}$ という関数の \mathbb{R}^2 上での積分を考える。

$$I = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy e^{-x^2-y^2} \quad (6.13)$$

この I は、以下のように書ける。

$$I = \left(\int_{-\infty}^{\infty} dx e^{-x^2} \right) \cdot \left(\int_{-\infty}^{\infty} dy e^{-y^2} \right) = \left(\int_{-\infty}^{\infty} dx e^{-x^2} \right)^2$$

e^{-x^2} は \mathbb{R} 全体で非負の値をとり、 $\lim_{x \rightarrow \pm\infty} e^{-x^2} = 0$ なので、積分値は正になる。すると、求める積分は \sqrt{I} である。

ということで I を計算しよう。元の積分を計算するのは大変だが、 I は簡単に計算できる。 $(x, y) \rightarrow (r, \theta)$ と変換して、極座標で考える。 $x = r \cos \theta$, $y = r \sin \theta$ とすると、 \mathbb{R}^2 全体は、 $r \geq 0$, $0 \leq \theta < 2\pi$ で表せる。Jacobian は、

$$\det \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \theta} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \det \begin{pmatrix} \cos \theta & \sin \theta \\ -r \sin \theta & r \cos \theta \end{pmatrix} = r$$

となるから、

$$\begin{aligned} I &= \int_0^{\infty} dr \int_0^{2\pi} d\theta r e^{-r^2} = 2\pi \int_0^{\infty} r e^{-r^2} dr \\ &= 2\pi \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty} \\ &= \pi \end{aligned} \quad (6.14)$$

となる。

$$\therefore \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

□

Gauss 積分の基本は、式 (6.12) であるが、一般に e の肩にのっている x^2 の係数は -1 ではない。そのような場合は変数変換をすればよい。

Remark

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}} \quad (a > 0) \quad (6.15)$$

証明

$ax^2 = t^2$ つまり、 $t = \sqrt{ax}$ とおくと、 x から t へと変数変換することができる。

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \int_{-\infty}^{\infty} e^{-t^2} \frac{1}{\sqrt{a}} dt = \sqrt{\frac{\pi}{a}}$$

□

さらに、この Gauss 積分の式、式 (6.15) の両辺を a で微分することを考える。微分と積分の交換ができるることを認めてしまえば、次の式を導くことができる。

Remark

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2} dx = \frac{1}{2a} \sqrt{\frac{\pi}{a}} \quad (a > 0) \quad (6.16)$$

証明

式 (6.15) を a で微分すると、以下のようにになるから OK。

$$\begin{aligned} (\text{左辺の微分}) &\rightarrow \frac{d}{da} \int_{-\infty}^{\infty} e^{-ax^2} dx = - \int_{-\infty}^{\infty} x^2 e^{-ax^2} dx \\ (\text{右辺の微分}) &\rightarrow \frac{d}{da} \sqrt{\frac{\pi}{a}} = -\frac{1}{2a} \sqrt{\frac{\pi}{a}} \end{aligned}$$

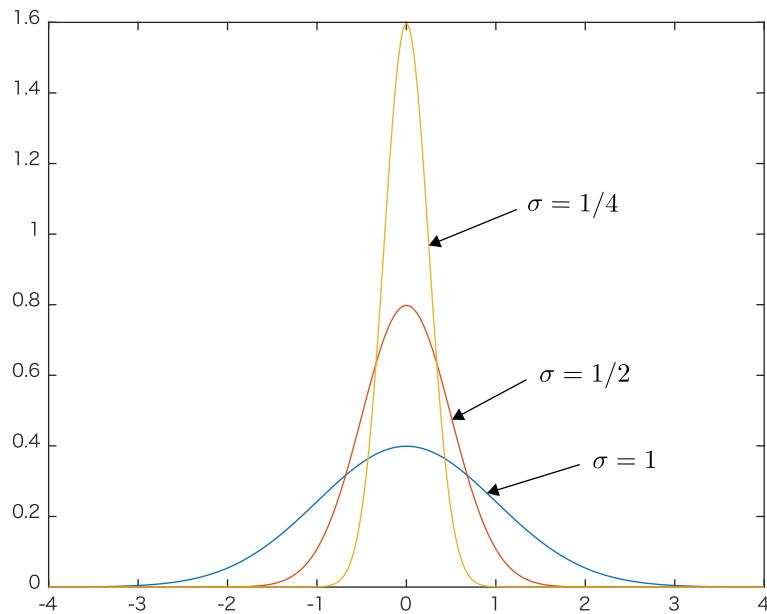
□

6.3.2 正規分布の定義

定義 6.7

$\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_{>0}$ とする。確率変数 X の密度関数 $f(x)$ が以下のように与えられる時、 X は正規分布 (normal distribution) $N(\alpha, \beta^2)$ に従うといい、 $X \sim N(\alpha, \beta^2)$ とかく。

$$f(x) = \frac{1}{\sqrt{2\pi}\beta} \exp\left(-\frac{(x-\alpha)^2}{2\beta^2}\right) \quad (-\infty < x < \infty) \quad (6.17)$$

図 6.5: $x = 0$ にピークを持つ正規分布のグラフ

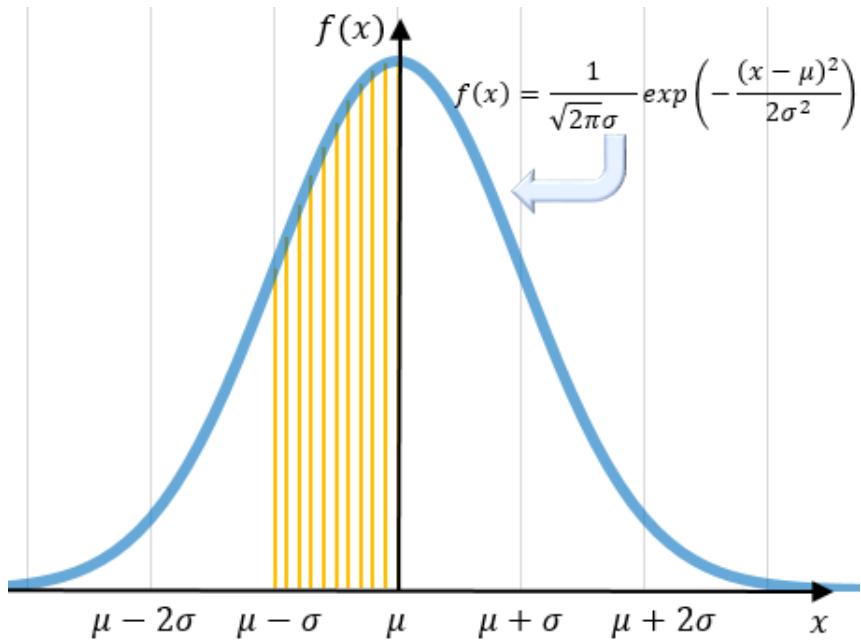
正規分布の平均と分散

$X \sim N(\alpha, \beta^2)$ の時、確率変数 X の平均と分散は以下のようになる。

$$E(X) = \alpha \quad (6.18)$$

$$V(X) = \beta^2 \quad (6.19)$$

そのため、 α は分布の中心、 β は分布のばらつきの大きさに関する情報を持つことがわかる。



正規分布の性質

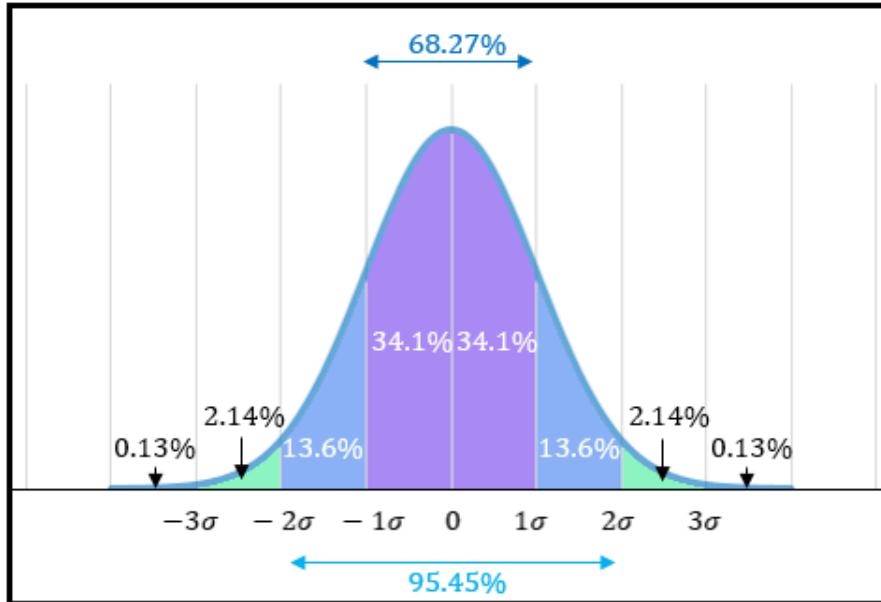
確率変数 X が正規分布 $N(\mu, \sigma^2)$ に従うとき、次の式が成り立つ。

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \simeq 0.683 \quad (6.20)$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \simeq 0.954 \quad (6.21)$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \simeq 0.997 \quad (6.22)$$

そのため、 $\mu - 3\sigma \leq X \leq \mu + 3\sigma$ の範囲に、事実上全てのデータが入っているとみなして良い。



6.3.3 標準正規分布

正規分布は、2項分布やポアソン分布、幾何分布とは異なり、確率変数 X を一次変換した $aX + b$ も正規分布になるという特徴を持つ。

一次変換と正規分布

確率変数 X が正規分布 $N(\mu, \sigma^2)$ に従うとき、任意の実数 a, b に対して、

$$aX + b \sim N(a\mu + b, a^2\sigma^2) \quad (6.23)$$

が成り立つ。

(証明)

連続型確率変数にも「期待値の線型性」が成り立つことより、一次変換後の期待値が $a\mu + b$ になることは明らか。同様に、分散が $a^2\sigma^2$ になることも明らか。

そのため、一次変換後も正規分布を保ち、平均や分散が上記のようになることを確かめれば良い。

X が正規分布 $N(\mu, \sigma^2)$ に従う時、

$$P(s \leq X \leq t) = \int_s^t \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

である。

ここで、 $Z = aX + b$ とおくと、 $X = \frac{Z - b}{a}$ であるから、

$$P(s \leq X \leq t) = P\left(s \leq \frac{Z - b}{a} \leq t\right)$$

積分の変数変換を行うと、

$$\begin{aligned} &= \int_{as+b}^{at+b} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-b-\mu)^2}{2\sigma^2}\right) dz \\ &= \int_{as+b}^{at+b} \frac{1}{\sqrt{2\pi}(a\sigma)} \exp\left(-\frac{(z-(a\mu+b))^2}{2(a^2\sigma^2)}\right) dz \end{aligned}$$

よって、 Z は正規分布 $N(a\mu + b, a^2\sigma^2)$ に従う。

以上より、 $Z = \frac{X - \mu}{\sigma}$ により基準化しても、正規分布が保たれて、平均は0、分散は1になる。正規分布 $N(0, 1)$ に従う分布を標準正規分布という。そして、標準正規分布の分布関数を $\Phi(a)$ と表す。

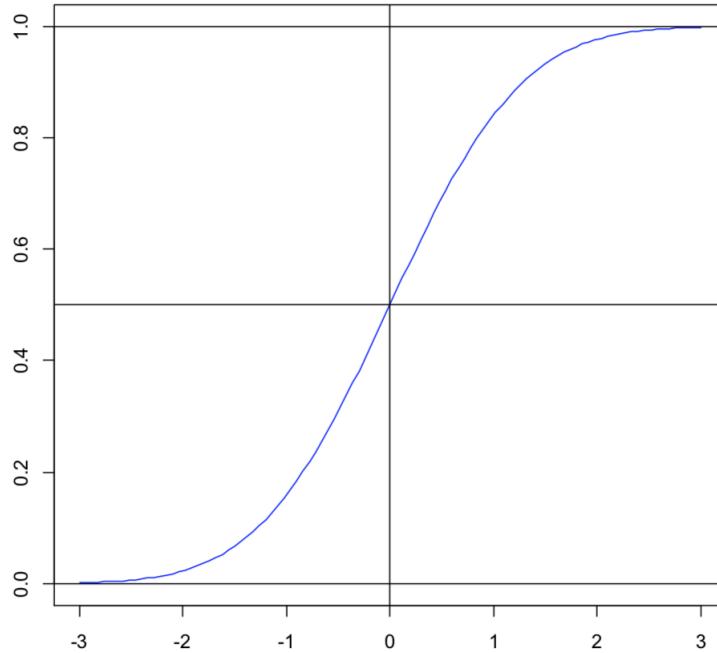
標準正規分布

確率変数 X が正規分布 $N(0, 1)$ に従うとき、その分布関数 $\Phi(a) = P(Z \leq a)$ は、次の性質を満たす。

$$\Phi(0) = \frac{1}{2} \quad (6.24)$$

$$\lim_{a \rightarrow -\infty} \Phi(a) = 0 \quad (6.25)$$

$$\lim_{a \rightarrow +\infty} \Phi(a) = 1 \quad (6.26)$$



確率変数 X が正規分布 $N(\mu, \sigma)$ に従う時、 $a \leq X \leq b$ となる確率をどう求めるか。一次変換が1対1に対応することを利用する。すなわち、確率変数 X を基準化した $Z = \frac{X - \mu}{\sigma}$ を使って確率を求める。なぜな

ら、標準正規分布 $N(0, 1)$ については、その分布関数の値が知られているからである。

標準正規分布を使って確率を求める

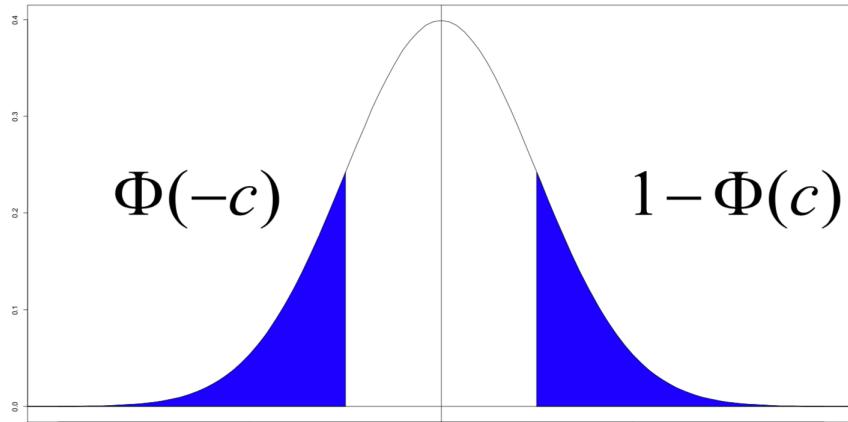
確率変数 X が正規分布 $N(\mu, \sigma)$ に従うとき、 X を $Z = \frac{X - \mu}{\sigma}$ により基準化すると、次の式が成り立つ。

$$P(a \leq x \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right) \quad (6.27)$$

$$= \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \quad (6.28)$$

この他にも、標準正規分布の対称性を利用すると、次の式が成り立ち、これもよく使う⁽²⁾。

$$\Phi(-c) = 1 - \Phi(c) \quad (6.29)$$



⁽²⁾標準正規分布の分布関数の値については、次のように具体的な値をまとめた表が存在する。しかし、これは、 $a \geq 0$ の時の場合の $P(Z \leq a)$ の値しかのっていない場合が多い。これは、正規分布の密度関数の対称性を利用して、 $a \leq 0$ の場合も求められるからである。

正規分布表

a に対して、 $\Phi(a)$ を与える。但し、 $\Phi(a)$ は標準正規分布の分布関数である。

すなわち、 $Z \sim N(0, 1)$ のとき、 $\Phi(a) = P(Z \leq a)$ 。
153頁(4.5.11)式を参照のこと。

a	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.504	0.508	0.512	0.516	0.520	0.524	0.528	0.532	0.536
0.1	0.540	0.544	0.548	0.552	0.556	0.560	0.564	0.567	0.571	0.575
0.2	0.579	0.583	0.587	0.591	0.595	0.599	0.603	0.606	0.610	0.614
0.3	0.618	0.622	0.626	0.629	0.633	0.637	0.641	0.644	0.648	0.652
0.4	0.655	0.659	0.663	0.666	0.670	0.674	0.677	0.681	0.684	0.688
0.5	0.691	0.695	0.698	0.702	0.705	0.709	0.712	0.716	0.719	0.722
0.6	0.726	0.729	0.732	0.736	0.739	0.742	0.745	0.749	0.752	0.755
0.7	0.758	0.761	0.764	0.767	0.770	0.773	0.776	0.779	0.782	0.785
0.8	0.788	0.791	0.794	0.797	0.800	0.802	0.805	0.808	0.811	0.813
0.9	0.816	0.819	0.821	0.824	0.826	0.829	0.831	0.834	0.836	0.839
1.0	0.841	0.844	0.846	0.848	0.851	0.853	0.855	0.858	0.860	0.862

6.3.4 正規分布の歪度と尖度

歪度と尖度(復習)

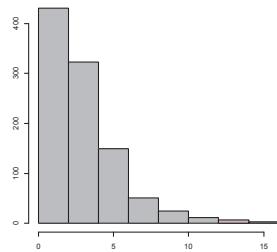
与えられたデータ x_1, x_2, \dots, x_n に対して、それらを基準化したものを z_1, z_2, \dots, z_n とする。この時、基準化変量の3乗の平均を歪度(b_1)といい、基準化変量の4乗の平均を尖度(b_2)という。

$$b_1 = \frac{1}{n} \sum_{k=1}^n (z_k)^3 = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{S} \right)^3 \quad (6.30)$$

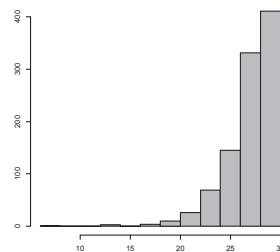
$$b_2 = \frac{1}{n} \sum_{k=1}^n (z_k)^4 = \frac{1}{n} \sum_{k=1}^n \left(\frac{x_k - \bar{x}}{S} \right)^4 \quad (6.31)$$

歪度と尖度を連続型の場合に適用できるようにすると、正規分布の場合、歪度は0、尖度は3となる。データ分布の歪度や尖度の指標である0や3は、正規分布の値である。

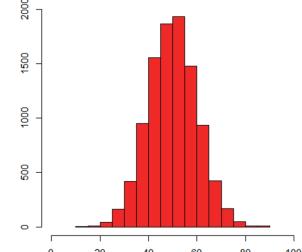
右に歪んだ分布
 $b_1 > 0$



左に歪んだ分布
 $b_1 < 0$

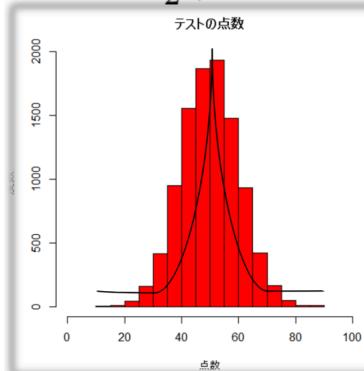


左右対称の分布
 $b_1 \approx 0$



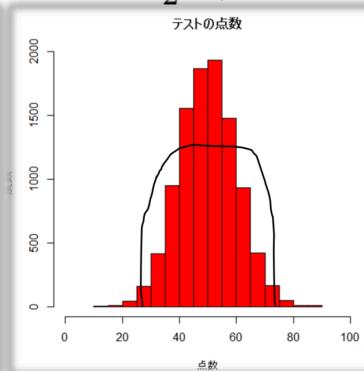
正規分布より尖りが強い

関係:
 $b_2 > 3$



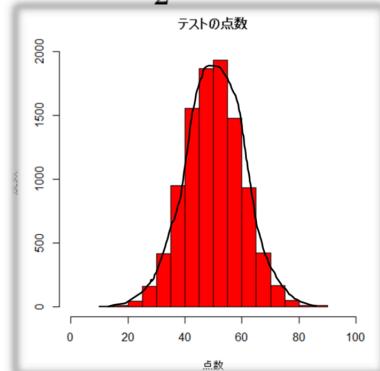
正規分布より尖りが弱い

関係:
 $b_2 < 3$



正規分布と同程度

関係:
 $b_2 \approx 3$



6.4 指数分布

離散型確率分布の「幾何分布」の連続版を考える。それが「指数分布」である。指数分布は、ある事象の起り方が偶発的要因に支配されている場合の、次の事象発生までの待ち時間を表現するのに適した分布である。

指数分布

確率変数 X の確率密度関数 $f(x)$ が次のように与えられる時、 X は指数分布 $Ex(\lambda)$ に従うといい、 $X \sim Ex(\lambda)$ とかく。ただし、 λ は正の定数である。

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & (x > 0 \text{ のとき}) \\ 0 & (x \leq 0 \text{ のとき}) \end{cases} \quad (6.32)$$

指数分布の性質

確率変数 X が指数分布 $Ex(\lambda)$ に従うとき、分布関数 $F(x)$ 、平均 $E(X)$ 、分散 $V(X)$ は次のようになる。

$$F(x) = \begin{cases} 1 - \lambda e^{-\lambda x} & (x > 0 \text{ のとき}) \\ 0 & (x \leq 0 \text{ のとき}) \end{cases} \quad (6.33)$$

$$E(X) = \frac{1}{\lambda} \quad (6.34)$$

$$V(X) = \frac{1}{\lambda^2} \quad (6.35)$$

(証明)

分布関数

(1) $x \leq 0$ のとき、 $F(x) = 0$ である。

(2) $x > 0$ のとき、

$$\begin{aligned} F(x) &= \int_{-\infty}^x f(x) dx \\ &= \int_{-\infty}^0 0 dx + \int_0^x \lambda e^{-\lambda x} dx \\ &= [-e^{-\lambda x}]_0^x \\ &= 1 - e^{-\lambda x} \end{aligned}$$

平均

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

$x \leq 0$ では $f(x) = 0$ なので

$$\begin{aligned} &= \int_0^{\infty} \lambda x e^{-\lambda x} dx \\ &= \lambda \left\{ \left[-\frac{1}{\lambda} x e^{-\lambda x} \right]_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx \right\} \\ &= \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

分散

いつも通り、 $V(X) = E(X^2) - \{E(X)\}^2$ を利用する。

$$\begin{aligned} E(X^2) &= \int_0^\infty \lambda x^2 e^{-\lambda x} dx \\ &= \lambda \left\{ \left[-\frac{1}{\lambda} x^2 e^{-\lambda x} \right]_0^\infty + \frac{1}{\lambda} \int_0^\infty 2x e^{-\lambda x} dx \right\} \\ &= \frac{2}{\lambda} \cdot \frac{1}{\lambda} \\ &= \frac{2}{\lambda^2} \end{aligned}$$

なので、

$$V(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}$$

指数分布は、幾何分布と同様に無記憶性をもつ。

指数分布の無記憶性

確率変数 X が指数分布 $Ex(\lambda)$ に従うとき、任意の正の実数 a, b に対して次の式が成り立つ。

$$P(X > a + b | X > b) = P(X > a) \quad (6.36)$$

(証明)

条件付き確率の定義より

$$P(X > a + b | X > b) = \frac{P(X > a + b)}{P(X > b)}$$

である。ここで、 $P(X > n)$ を求めると、

$$\begin{aligned} P(X > n) &= \int_n^\infty \lambda e^{-\lambda x} dx \\ &= [-e^{-\lambda x}]_n^\infty \\ &= e^{-n} \end{aligned}$$

となるから、

$$P(X > a + b | X > b) = \frac{P(X > a + b)}{P(X > b)} = \frac{e^{-(a+b)}}{e^{-b}} = e^{-a}$$

では、具体例を使って考える。

首都高の交通量の予測

2017年度の休日に、首都高の「箱崎JCT」から「都心環状線」に向かう方向へと通った車は、1日平均 66010 台であるという。つまり、1分あたりに平均 46 台であるという。この区間における車の流れがランダムであると仮定したとき、1台通過した後、次の車が通過するまでの時間間隔 X (分)について考える。

1分あたり平均 46 台の車が通過が通過するということは、 X の平均は $E(x) = \frac{1}{46}$ である。よって、 X は指数分布 $Ex(46)$ に従う。

そのため、例えば、10秒($= 1/6$ 分)以内に次の車が通過する確率は

$$P\left(X \leq \frac{1}{6}\right) = 1 - e^{-46 \times (1/6)} = 0.9995$$

となり、首都高の「箱崎 JCT」と「都心環状線」の間の区間は、絶えず車が通過することと予想される。

さて、同じ方法で、渋滞の名所ともいえる「小仏トンネル」について同じことを考える。

小仏トンネルの渋滞の分析

2016 年度の観光シーズンに、小仏トンネルを東京方面に通過した車は、1 日平均 36332 台であるといふ。つまり、1 分あたりに平均 25 台であるという。この区間における車の流れがランダムであると仮定したとき、10 秒 (= 1/6 分) 以内に次の車が通過する確率を指数分布を使って予測せよ。

1 台通過した後、次の車が通過するまでの時間間隔を X (分) とすると、 X は指数分布 $Ex(25)$ に従う。首都高の場合と同様に考えると、10 秒 (= 1/6 分) 以内に次の車が通過する確率は

$$P\left(X \leq \frac{1}{6}\right) = 1 - e^{-25 \times (1/6)} = 0.9845$$

となる。

98% の確率で、10 秒以内に次の車が通過する。首都高の「箱崎 JCT」と「都心環状線」の間は、ほぼ 100% の確率で 10 秒以内に次の車が通過する。数字の上では、首都高のこの区間の方が混んでいるように思えるが、実際は違う。この数学的予測との違いは十分検討の余地がある。

6.5 確率変数の変換

6.5.1 離散型確率変数の変換

6.5.2 連続型確率変数の変換

参考文献およびデータの引用元

参考文献

- (1) 「基礎統計」講義レジュメ
- (2) 倉田博史・星野崇宏『入門統計解析』(新世社)
- (3) 東京大学教養学部統計学教室編『基礎統計学 統計学入門』(東京大学出版会)
- (4) 繩田和満『東京大学工学教程 基礎系 数学 確率・統計 I』(丸善出版)
- (5) 前園宜彦『詳解演習 確率統計』(サイエンス社)
- (6) 逆瀬川浩孝『理工基礎 確率とその応用』(サイエンス社)
- (7) 杉山将『機械学習のための確率と統計』(講談社)
- (8) 「数理手法 4」講義ノート
- (9) 「確率数理工学」講義ノート
- (10) 「応用統計学」講義ノート

データの引用元

- (1) https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/koyou_roudou/roudoukijun/minimumichiran/
- (2) <https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00130002&tstat=000001032727&cycle=1&year=20180&month=24101210>
- (3) https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00400002&tstat=000001011648&cycle=0&tclass1=000001113655&tclass2=000001113656&stat_infid=000031685239&second2=1