法と統計学　第 9 回
ベイズ更新
記述統計学 1

森　大輔

熊本大学法学部

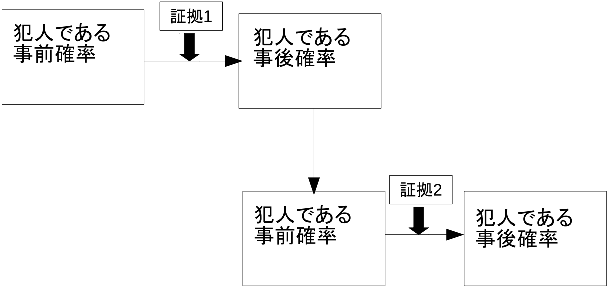
2017 年

本日の流れ

1. ベイズ更新と逐次性
2. 記述統計学
3. 中心の位置を測る尺度 (代表値)
   - 平均 (算術平均、相加平均)
   - 中央値
   - 幾何平均
4. ばらつきを測る指標
   - 分散と標準偏差
   - 標準偏差に関する事件

裁判とベイズ更新

- 裁判で行われていることも、逐次的なベイズ更新と見ることができる



裁判とベイズ更新

- 最初に被告人が法廷に来たときは、裁判官は例えば被告人が犯人である確率を 50%だと思っているとする
  - まだ犯人とも犯人でないとも言えないから理由不十分の原理（ただし、別の確率の値を考えることもできる）
  - これが一番最初の事前確率
- そこから、証拠が次々と提示されていく
- そのたびに裁判官は、被告人が犯人である確率を更新していく
- そして、被告人が犯人である確率がある値を超えたら、被告人が犯人だとみなして有罪判決
  - この境となる値のことを、法学では証明度と呼ぶ
  - 送られてきたメールが迷惑メールである確率がある値を超えたら、そのメールを迷惑メールとみなして迷惑メールフォルダに入れるのとちょうど同じ感じ

練習問題 1：裁判とベイズ更新

　ある被告人が法廷に連れてこられたとき、裁判官は、その人が犯人である確率を 50%と考えていた。
　まず証拠として、犯行現場の監視カメラに写っていた不鮮明な映像が、その人に似ているという事実が出された。経験則上、犯人であるときに、映像と似ている確率は 90%である。また、犯人でないのに、映像と似ている確率は 10%である。
　さらに次の証拠として、犯行時刻に被告人にアリバイがないという事実が出された。経験則上、犯人であるときに、アリバイがない確率は 80%である。また、犯人でないのに、アリバイがない確率は 30%である。
　これらの証拠を踏まえると、この人が犯人である確率はいくらと考えられるか。ただし、1 番目の証拠と 2 番目の証拠は、独立だと考えること。

統計学の 2 分類

- 記述統計学…集めたデータの特徴を明らかにするために、要約統計量などを求めて整理する
  - 要約統計量…データの持つ情報を要約して表したもの
    - 1 つの変数の情報の要約…データの個数、平均、分散、標準偏差など
    - 2 つ以上の変数間の関係に関する情報の要約…相関係数など
- 推測統計学…データを使って、その背景にある全体（母集団）の特徴を推測する
  - 仮説検定
  - 区間推定

例題 1

7 個の観測値 $\{1, 3, 5, 6, 6, 9, 12\}$ がある。

(1) 平均 (算術平均) を求めなさい。
(2) 中央値 (メディアン) を求めなさい。
(3) 幾何平均を求めなさい。
(4) 標準偏差を求めなさい。
(5) 7 個の観測値のうち 12 を 100 に変えた場合の平均、中央値、幾何平均の値を求めなさい。

算術平均

- 代表値のうち、最もよく知られているのは、平均 (算術平均、相加平均)
  - 他の種類の平均もある (後に見る)
- 計算は、観測値の総和 ÷ 観測値の数
  - n 個の観測値 $\{x_1, x_2, \cdots, x_n\}$ があるとき
  - 平均 $= (x_1 + x_2 + \cdots + x_n) \div n$
- 例えば、7 個の観測値 $\{1, 3, 5, 6, 6, 9, 12\}$ の平均
  - これは例えば、7 人の所持金がそれぞれ 1 万、3 万、5 万、6 万、6 万、9 万、12 万円のときの、7 人の所持金の平均を求めることなどを考えればよい
  - 平均$=(1 + 3 + 5 + 6 + 6 + 9 + 12) \div 7 = 6$

中央値（メディアン）

- 観測値を小さい順に並べたときに中央に位置する値
  - 7 個の観測値 $\{1, 3, 5, 6, 6, 9, 12\}$ の中央値は、4 番目の値である 6
- 観測値が奇数個のときはちょうど中央の値、偶数個のときは真ん中の 2 つの数字を足して 2 で割った値
  - 6 個の観測値 $\{1, 3, 5, 6, 6, 9\}$ の中央値は、3 番目の値 5 と 4 番目の値 6 を足して 2 で割って 5.5
    - $(5 + 6) \div 2 = 5.5$

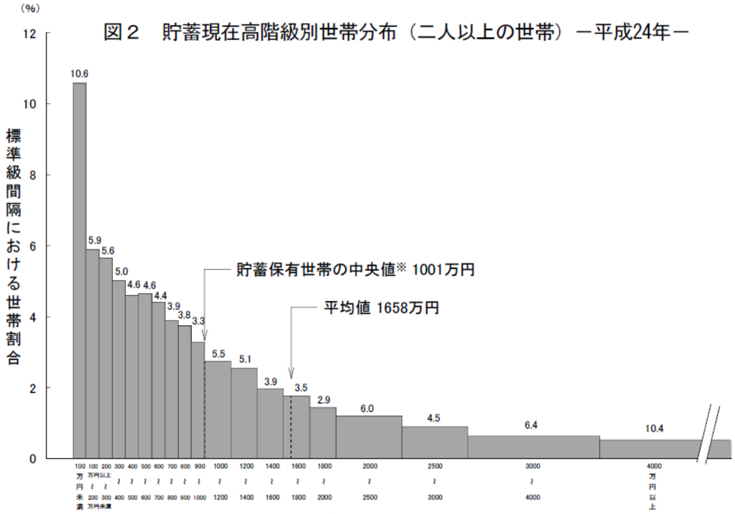
中央値と外れ値

- 7 個の観測値 $\{1, 3, 5, 6, 6, 9, 12\}$
  - 平均は $(1 + 3 + 5 + 6 + 6 + 9 + 12) \div 7 = 6$、中央値は 6
- 次に、上の 7 個のデータのうち、最後のデータを 12 から 100 に取り替えてみる
  - 100 が外れ値
  - 7 個の観測値 $\{1, 3, 5, 6, 6, 9, 100\}$
  - 平均は $(1 + 3 + 5 + 6 + 6 + 9 + 100) \div 7 \fallingdotseq 18.57$、中央値は 6
- 平均は外れ値 100 に引っ張られて大きくなってしまう
  - 平均 18.57 よりも大きな観測値は 100 のみ、後はすべて 18.57 より小さい
- それに対して、中央値は外れ値があっても変わらない
  - 中央値は外れ値の影響を受けにくい

中央値と外れ値

- 貯蓄、収入、負債などで、外れ値の影響が大きい
  - 一部のお金持ちが、ものすごい額のお金を銀行に預けていたりするから (外れ値)
- このような場合、平均よりも中央値の方が、中心の位置を測る尺度として、人々の感覚に合う
  - 例えば、平成 24 年度の日本の 2 人以上世帯の貯蓄の平均は、1658 万円
    - みんな本当にこんなに貯蓄があるのか？
  - 中央値は 1001 万円
    - 平均は、一部のお金持ちの貯蓄額にかなり引きずられて大きくなっていることがわかる
- 貯蓄、収入、負債などは、柱上のグラフ (ヒストグラム) を作ると、一番大きい柱が左の方に寄る
  - 逆にこういう形になっているときは、平均だけでなく中央値も見てみるべき

中央値とヒストグラム (『家計調査報告（貯蓄・負債編）平成 24 年平均結果速報』p.2 より)



例題 2：中央値に関する事件 1

1. 裁判所が平均と中央値を問題にした例その 1
   - 鉄道会社の線路への財産税の問題

- 財産税は、資産に対して課される税
- 財産税は、資産の査定価格を基準にして課される
  - 税を課す前に、妥当と思われる価格を算出する
  - 査定価格が高いほど、たくさんの税金を払う必要
- 査定価格は、市場価格とは異なる
  - 市場価格は、市場に売りに出した時に付く価格
- 税金をたくさん払わせたいために、州が査定価格を高く査定することが行われる可能性
  - そうすると、市場価格に比べて査定価格が高くなる
  - 査定価格/市場価格、という比の値が大きくなる

中央値に関する事件 1

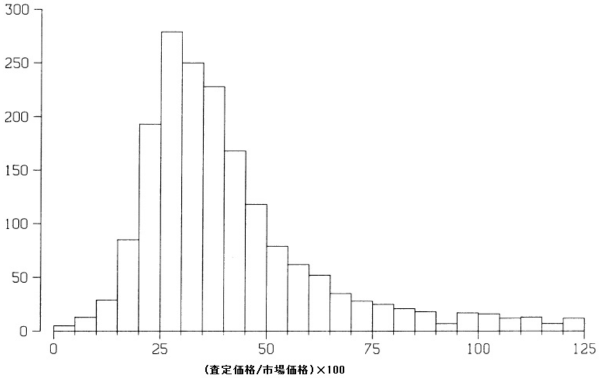
- 連邦法の規定
  - 州などが、他の会社や組織等の資産の場合の比の値 (査定価格/市場価格) よりも大きな比の値で、鉄道会社の線路に財産税を課すことを禁止
  - 鉄道会社を他と平等に扱うための規定
- ここで問題
  - 「他の会社や組織等の資産の場合の比の値」というのは、平均で見るのか中央値で見るのか？
    - 「他の会社や組織等」には、いろいろな会社や組織が含まれる
    - それらを平均したもので見るのか、中央値で見るのか

中央値に関する事件 1

- ある裁判所は、平均で見る、とした
  - Southern Pacific v. California State Board of Equalization (1983)
    - カリフォルニア州の連邦地裁
    - 連邦法を作るとき、連邦議会で average という言葉が議論で使われた
    - average は平均の意味でしかとれない、という理由
- しかし、別の裁判所は、中央値で見る、とした
  - Clinchfield R.R. Company v. Lynch (1983)
    - 第 4 巡回区連邦控訴裁判所
    - より統計学の中身に踏み込んだ議論
    - 大きな比の値 (査定価格/市場価格) になっている少数の会社や組織に引きずられて、平均は大きくなる（たくさん税金を取るために州によって意図的に高く査定されていた会社や組織があった）
    - よって、中央値で見る方が適切

中央値に関する事件 1

- 会社・組織ごとの査定価格/市場価格の分布の例



(Finkelstein & Levin (2001:11) より)

中央値に関する事件 2

2 裁判所が平均と中央値を問題にした例その 2
- 悪性腫瘍に関する医療過誤

- Heinrich v. Sweet (2002)
  - 第 1 巡回区連邦控訴裁判所
  - 悪性の脳腫瘍で死亡した患者の相続人が、医療過誤による死亡だとして患者を担当した医師を訴えた
  - 医師は新しい治療法を試みて患者の死期を早めた、と相続人が主張
  - それを示すために、同じ病気に罹患した 70 人の患者を対象にした研究を引用
    - この 70 人では、最初の診断からの余命の平均は 17 ヵ月
    - それに対して、本事件の患者は 7 ヵ月しか生きられなかった

中央値に関する事件 2

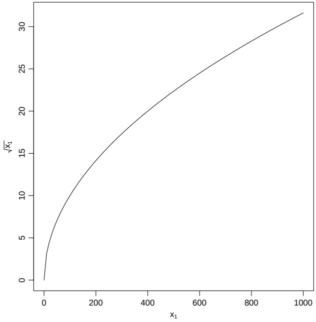
- しかし、裁判所は、この 70 人の研究では、余命の中央値は 8.3 ヵ月であることを指摘
- この研究を行った研究者自身も、中央値で見る方が適切だと述べている
  - 余命の平均は、例外的に長く生存した患者の余命に引きずられるから
- 中央値ならば、本事件の患者の余命 7 ヵ月とかなり近い

幾何平均（相乗平均）

- 幾何平均の計算の仕方
  - n 個の観測値 $\{x_1, x_2, \cdots, x_n\}$ があるとき
  - 幾何平均$=\sqrt[n]{x_1 x_2 \cdots x_n}$
  - 例 1：2 個の観測値 $\{10, 20\}$ があるとき、幾何平均は
    - $\sqrt{10 \times 20} = \sqrt{200} \fallingdotseq 14.14$
  - 例 2：7 個の観測値 $\{1, 3, 5, 6, 6, 9, 12\}$ があるとき、幾何平均は
    - $\sqrt[7]{1 \times 3 \times 5 \times 6 \times 6 \times 9 \times 12} = \sqrt[7]{58320} \fallingdotseq 4.80$
  - 例 3：7 個の観測値 $\{1, 3, 5, 6, 6, 9, 100\}$ があるとき、幾何平均は
    - $\sqrt[7]{1 \times 3 \times 5 \times 6 \times 6 \times 9 \times 100} = \sqrt[7]{486000} \fallingdotseq 6.49$
- 幾何平均は、観測値が 0 以上の場合しか計算できない
  - 負の値の平方根などは実数の範囲では存在しない

幾何平均の性質

- 幾何平均は外れ値の影響を受けにくい
- 幾何平均では大きな数が割り引いて考えられるから
  - 例えば 2 つの観測値 $\{x_1, x_2\}$ の幾何平均 $\sqrt{x_1 x_2}$
  - x_1 の値を 0 から大きくしていった場合の $\sqrt{x_1}$ の値は下図
    - x_1 が相当大きくなっても $\sqrt{x_1}$ は小さいまま



幾何平均の適用例

- このため、幾何平均は、少数の異常に大きな観測値があっても影響されにくい
  - 算術平均はこうした外れ値の影響を受けやすかった
- よって、観測値が 0 以上で、かつ少数の異常に大きな観測値の影響を防ぎたいと考える場合に、幾何平均を採用することがある
- 例：水質基準における大腸菌の数
- 例：メディケアに関する入院日数
  - メディケアとは米国の高齢者などのための公的な医療保険制度
    - 国民がメディケアの機関に保険料を支払い、メディケアの機関が病院に医療費を払う

幾何平均の適用例

- 米国では、高齢者の医療費の増大に頭を悩ませていた
- そのため、定額払い制度を導入
  - 一定の病気については、実際の治療にいくらかかったかに関係なく、メディケアの機関は定額で医療費を払う
- こうすると、病院側としてはなるべく患者の入院日数を短くしようとする
  - 定額しか払われないので、治療にかかる費用を少なくするほど病院の得になる
- 病院が経営状態を把握するために、多数の患者の入院日数の平均を調べる
  - このときに、算術平均だけでなく幾何平均も使われる
  - 幾何平均であれば、少数の異常に長く入院している患者に平均が引きずられることがないから

ばらつき

- データの平均を知るだけで十分か？
- → データのばらつき具合も重要
- 例：バス停へのバスの到着時刻
  - 時刻表では 7 時 30 分到着となっている
  - 4 日間、本当にその時刻に着くか確かめてみた

- パターン A

7 時 31 分	7 時 29 分	7 時 32 分	7 時 28 分

- パターン B

7 時 55 分	7 時 45 分	7 時 15 分	7 時 5 分

ばらつき

- パターン A
- $(31 + 29 + 32 + 28) \div 4 = 30$ なので、到着時刻の平均は 7 時 30 分

7 時 31 分	7 時 29 分	7 時 32 分	7 時 28 分

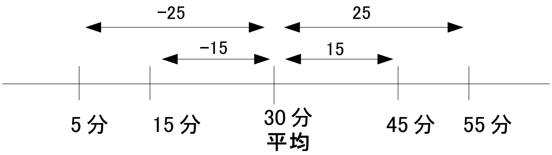
- パターン B
- $(55 + 45 + 15 + 5) \div 4 = 30$ なので、到着時刻の平均は 7 時 30 分

7 時 55 分	7 時 45 分	7 時 15 分	7 時 5 分

- $\rightarrow$ どちらも平均は同じ
- しかし、パターン B のバスには乗りたくないのでは？
- 到着時刻のばらつきが大きすぎるから

ばらつきの測り方

- ばらつきを測るには？



- 案 1 :「平均と各データの差」の平均を求める
  - 各データから平均を引いた値（偏差）を足す
  - 0 になってしまうので、だめ
  - パターン B の場合、$\{25 + 15 + (-15) + (-25)\} \div 4 = 0$
  - -15 や -25 のマイナスを消す必要

分散と標準偏差

- 案 2:「平均と各データの差」の絶対値の平均を求める
  - こうするとマイナスは消える
  - ただし、これよりも次の 2 乗を使う案の方が、数学的に扱いやすいことが多い
- 案 3:「平均と各データの差」の 2 乗の平均を求める
  - これを分散という
  - パターン B の場合、$\{25^2 + 15^2 + (-15)^2 + (-25)^2\} \div 4 = 425$
  - しかし、2 乗したせいで数が大きくなりすぎる
- 案 4:「平均と各データの距離」の 2 乗の平均を求めて、その平方根をとる
  - これを標準偏差という
  - パターン B の場合、$\sqrt{425} \fallingdotseq 20.6$

標準偏差に関する事件

- データのばらつきが重要な場合、平均や中央値に標準偏差が併記される
- 例 1：Asher v. Baxter Int'l, Inc. (2004)
  - 第 7 巡回区連邦控訴裁の事件
  - 製薬会社のバクスター・インターナショナルが将来の収益の見積もりを発表
    - しかし見積もりが楽観的すぎた
  - 実際の収益が発表されると、バクスター社の株価が大きく下がった
  - バクスター社は訴えられた
    - 誤解を与える見積もりを発表したという理由
  - 裁判所は、バクスター社は正確な株価を形成するのに十分な情報を市場に与えなければならないとする

標準偏差に関する事件

- 具体的には裁判所は、バクスター社は自分の見積もりの平均、中央値、標準偏差を開示すべきだったと判断
  - バクスター社は内部では複数の見積もりをしているはず
  - それらの平均、中央値、標準偏差
- 「平均値は中央値よりも大きいという知識、または標準偏差が重要であるという知識は、市場価格を決定するような取引を行うプロの投資家にとってとりわけ役に立つであろう」
- 例 2：Exxon Shipping Co. v. Baker (2008)
  - 連邦最高裁の事件
  - 石油タンカーのエクソン・ヴァルディーズ号が座礁
  - 原油がアラスカのプリンス・ウィリアム湾に流出して大規模な環境破壊

標準偏差に関する事件

- 連邦控訴裁でエクソン社に填補的損害賠償 2 億 8700 万ドル、懲罰的損害賠償 25 億ドルの支払いが命じられる
  - 填補的損害賠償とは、生じた損害を埋めあわせるため、損害額と同額分が課される賠償
  - 懲罰的損害賠償とは、悪性の強い加害者を罰するためなどに、損害額分とは別に課される賠償
    - 日本には填補的損害賠償しかない
    - 懲罰的損害賠償は、米国の特徴
    - 米国では企業などに巨額の懲罰的損害賠償が課されることがある
- エクソン社は連邦最高裁に上告
  - 懲罰的損害賠償が巨額すぎると主張
  - それまでにも、あまりに巨額な懲罰的損害賠償は合衆国憲法違反と連邦最高裁は判断していた

標準偏差に関する事件

- 連邦最高裁は、現在、懲罰的損害賠償の額が著しく予測困難になっていることを問題視
  - 州の民事訴訟の賠償額に関する研究を引用
    - いろいろな州の様々な事件で、填補的損害賠償と懲罰的損害賠償が課されている
  - (懲罰的損害賠償/填補的損害賠償) の平均は 2.90
    - つまり、填補的損害賠償の 2.9 倍の額の懲罰的損害賠償が平均で課されている
  - しかし、(懲罰的損害賠償/填補的損害賠償) の中央値は 0.62
    - 中央値が平均値に比べてずいぶん小さくなっている
    - 外れ値の存在
    - つまり、一部の事件で填補的損害賠償の何十、何百倍にもなる懲罰的損害賠償が課され、それに平均が引っ張られている可能性

標準偏差に関する事件

- (懲罰的損害賠償/填補的損害賠償) の標準偏差は 13.81
  - 懲罰的損害賠償が填補的損害賠償の何倍になるかは、事件によってばらつきが非常に大きい
- 連邦最高裁は、この研究から次のように結論
  - 分布の範囲は非常に広く、填補的損害賠償がかすんでしまうほどの懲罰的損害賠償を被告に課しているのは外れ値に当たる事例である
- よって連邦最高裁は懲罰的損害賠償を減額
- しかし、連邦最高裁のこの研究の引用の仕方には批判
  - この研究の研究者 Theodore Eisenberg 本人から批判
  - この分析は填補的損害賠償の額を考慮に入れていない
  - Exxon Shipping Co. v. Baker 事件のように填補的損害賠償も巨額な場合についてのみ見ると、(懲罰的損害賠償/填補的損害賠償) の標準偏差はずっと小さい

標準偏差に関する事件

- 填補的損害賠償の額ごとに (懲罰的損害賠償/填補的損害賠償) について調べると下の表

填補的損害賠償の額	(懲罰的損害賠償/填補的損害賠償) の中央値	(懲罰的損害賠償/填補的損害賠償) の平均	(懲罰的損害賠償/填補的損害賠償) の標準偏差	事件数
$0〜999	24.69	101.47	175.44	11
$1 千〜9,999	1.00	9.64	39.37	43
$1 万〜99,999	0.56	1.68	3.58	162
$10 万〜999,999	0.55	1.62	3.32	151
$100 万〜9,999,999	0.42	1.46	3.71	57
$1,000 万〜1 億未満	0.57	1.12	1.31	13
$1 億以上	2.41	2.41	-	1

Eisenberg, Heise & Wells (2010:18) より

参考文献

- マイクル O. フィンケルスタイン (2014)『法統計学入門：法律家のための確率統計の初歩』木鐸社の第 1 章.
- ゲルト・ギーゲレンツァー (2010)『リスク・リテラシーが身につく統計的思考法：初歩からベイズ推定まで』早川書房.
- 小島寛之 (2015)『完全独習 ベイズ統計学入門』ダイヤモンド社.
- Eisenberg, T., M. Heise & M. T. Wells (2010) "Variability in Punitive Damages: Empirically Assessing Exxon Shipping Co. v. Baker," *Journal of Institutional and Theoretical Economics* 166, 5–26.
- Finkelstein, M. O. & B. Levin (2001) *Statisticf for Lawyers* (2nd ed.), Springer の第 1 章.