



青の統計学

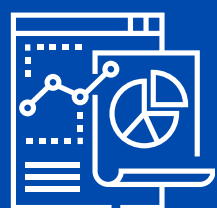
ビジネスマンが
データサイエンスを活かすための

統計学 基礎講座

はじめに

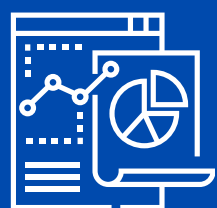
この資料の対象者

- **ビジネスにデータサイエンスを活かしたい方**
 - 現場で使えるような具体例が多め
- **統計学をこれから勉強したい方**
 - 高校数学の知識があれば、理解可能
- **統計検定などの資格取得に挑戦したい方**
 - 3級～2級で頻出のトピックを厳選



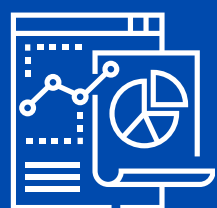


1. 記述統計
2. データの散らばりの指標
3. 確率と確率分布
4. 相関と回帰
5. 統計的推測





1. 記述統計
2. データの散らばりの指標
3. 確率と確率分布
4. 相関と回帰
5. 統計的推測



記述統計の目的

推測統計との比較

記述統計

手元のデータを可視化したり、まとめる方法。

例：商品の売上データから月ごとの売上平均や最大値を算出する。

▶ 主な目的は、データを最大値や最小値、中央値などの指標で**整理し、可視化することで、全体像を把握し、意思決定につなげる**こと

推測統計

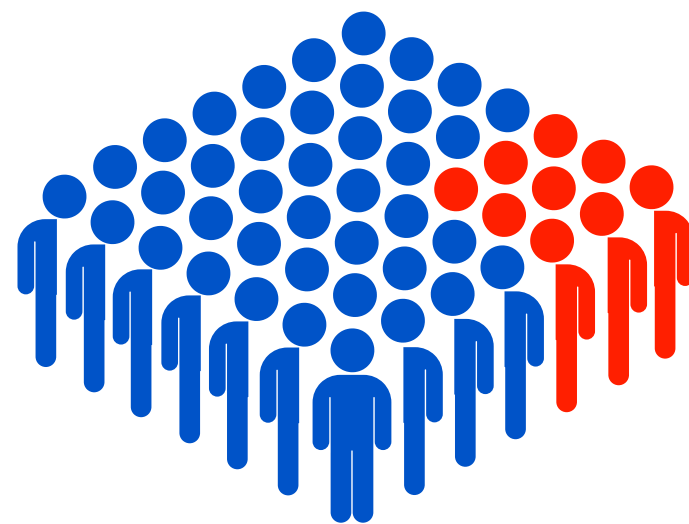
標本をもとに、母集団の特性を把握したり、予測を行う方法。

例：一部の顧客にアンケートを実施し、全顧客の満足度を推定する。

▶ 主な目的は、**不確実性を伴う意思決定を支援**すること

特徴	記述統計	推測統計
目的	データを整理・要約し、全体像を把握する	標本(サンプル)から、母集団の特徴を推測する
対象	手元のデータ（全数でも標本でもよい）	標本
例	テストの平均点の計算	新商品の市場シェアの予測

全数調査は対象全体を調べる方法。標本調査は一部を調べて全体を推測する方法。**費用や時間に加え、未回答・測定誤差・対象の偏りも考慮して選ぶ。**



ポイント

記述統計は**データそのものを描写するための方法**。

推測統計は限られた標本から母集団の特徴や変数の重みなどを推定するための方法。

代表値

平均値と中央値

平均値(mean)

データの値の合計を、データの個数で割った値。
外れ値の影響を強く受ける特徴がある。

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

中央値(median)

データを小さい順に並べたときの真ん中の値。
奇数個であれば、真ん中の値、偶数個であれば中央の二つの値の平均になる。
*下の添え字は並べ替え後の順序統計量を表す。

$$\text{median} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd} \\ \frac{1}{2} (x_{n/2} + x_{(n/2)+1}), & \text{if } n \text{ is even} \end{cases}$$

中央値は '非感度統計量 (Resistant Statistic) ' として知られ、外れ値に対する影響を抑える性質がある



ポイント

中央値は平均値に対して、外れ値の影響を受けにくい頑健（ロバスト）な性質

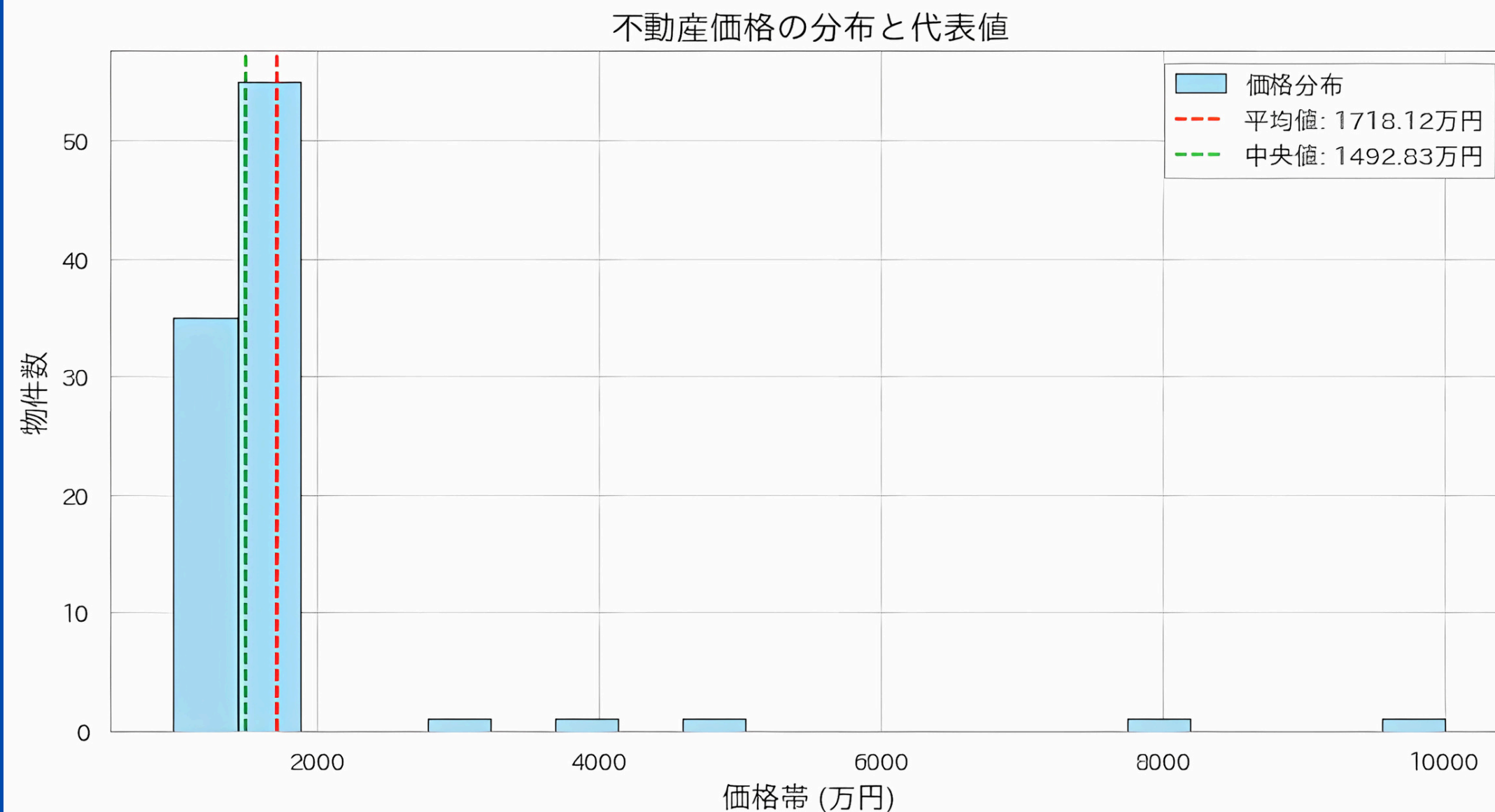
具体例

不動産の価格分析では、外れ値による影響を受けにくい中央値を使うことが多い。

- 平均値: 少数の高額物件に引き上げられるため、典型的な物件価格より高くなることがある。
- 中央値: 価格順の中央に位置する物件の水準を示し、高額物件の影響を受けにくい。

不動産価格は通常、右に長い尾を持つスキュー分布を形成する。

これは**非常に高額な物件（例: 高級マンションや一等地の物件）**が少数存在するため



代表値 最頻値

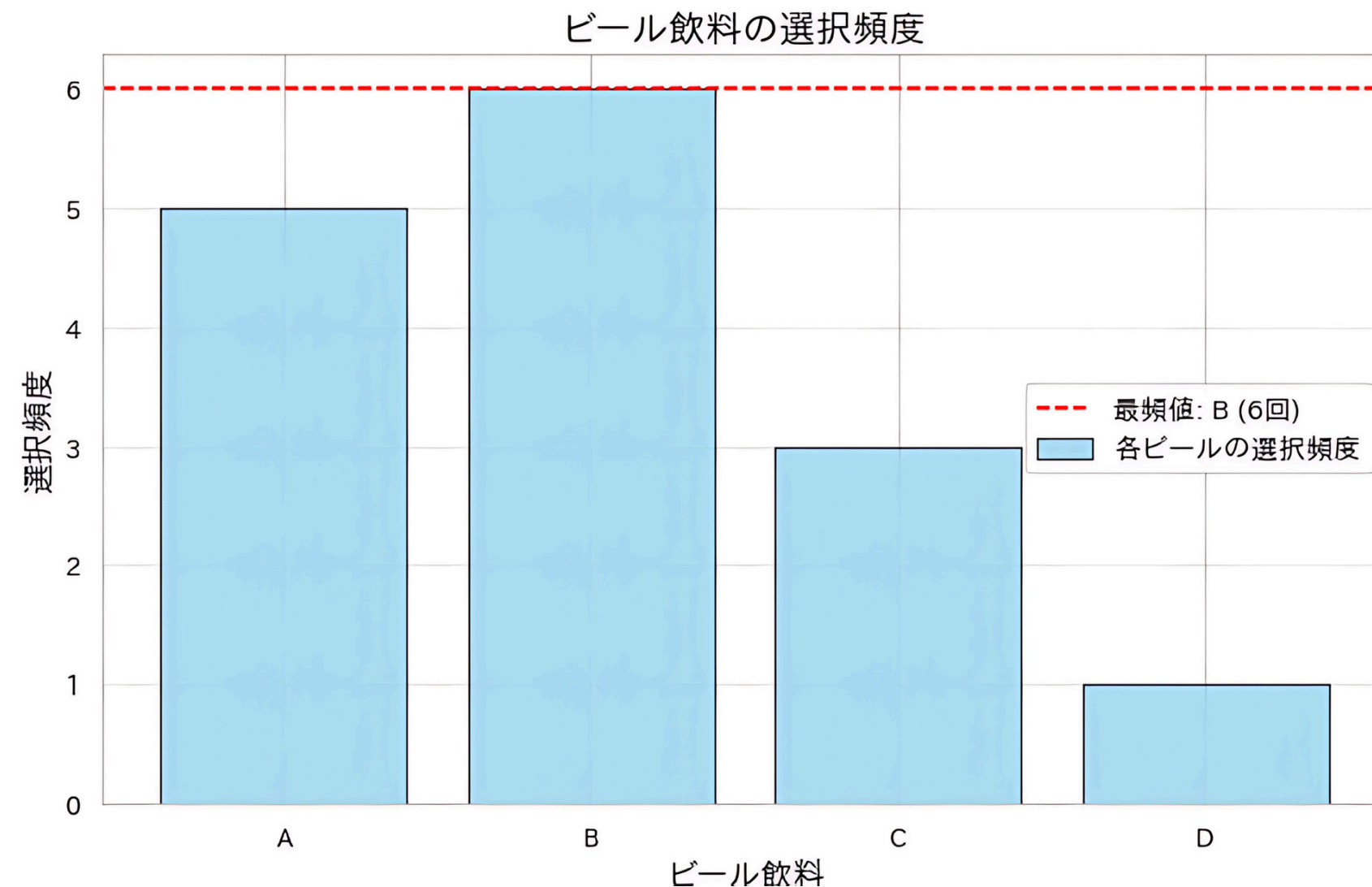
最頻値(mode)

データ内で最も頻繁に現れる値。
複数の最頻値が存在する場合もある。

下の例では最も多く選ばれたカテゴリBが最頻値（度数6）。
連続分布では密度のピークの位置を最頻値と呼ぶ。

具体例

ビール飲料A、B、C、Dに関するアンケートデータを基に、各飲料の選択頻度を棒グラフで視覚化し、最頻値（最も選ばれた飲料）を確認した。



実際、どんな意思決定ができるのか

- 製品ごとの広告予算の最適化
 - 最も選ばれた飲料（最頻値）を主力商品としてマーケティング予算を集中させる。
 - 在庫管理:
 - 最頻値に基づき、最も需要のある飲料を追加生産し、需要過多による欠品を防ぐ。
 - 製品改善
 - 最頻値以外の商品が低い頻度で選ばれている場合、その商品の改善やキャンペーンを検討する。（効果の差配的な意味合い）
- * 回答者の構成や購入量も確認し、選択頻度だけで需要全体や最適な広告予算を決めない。



分位数 定義と具体例

分位数

データセットをいくつかの等しい部分に分けるための基準となる値

パーセンタイル

データを100等分したときの位置に対応する値

あるp(パーセンタイル)に対応する分位数の位置Lは以下の通り

$$L = \frac{p}{100} (n + 1)$$

- pパーセンタイルは、小さい方からおよそp%の位置にある値。
- 標本からの計算には複数の方法があり、補間方法などによってソフト間で値が異なることがある。

代表的な分位数

- 25パーセンタイル（第1四分位数）
 - データの下位25%に位置する値
- 50パーセンタイル
 - データの中央値（50%の位置）
- 75パーセンタイル（第3四分位数）
 - 小さい方から75%の位置にある値

具体例

顧客の購入額による分類



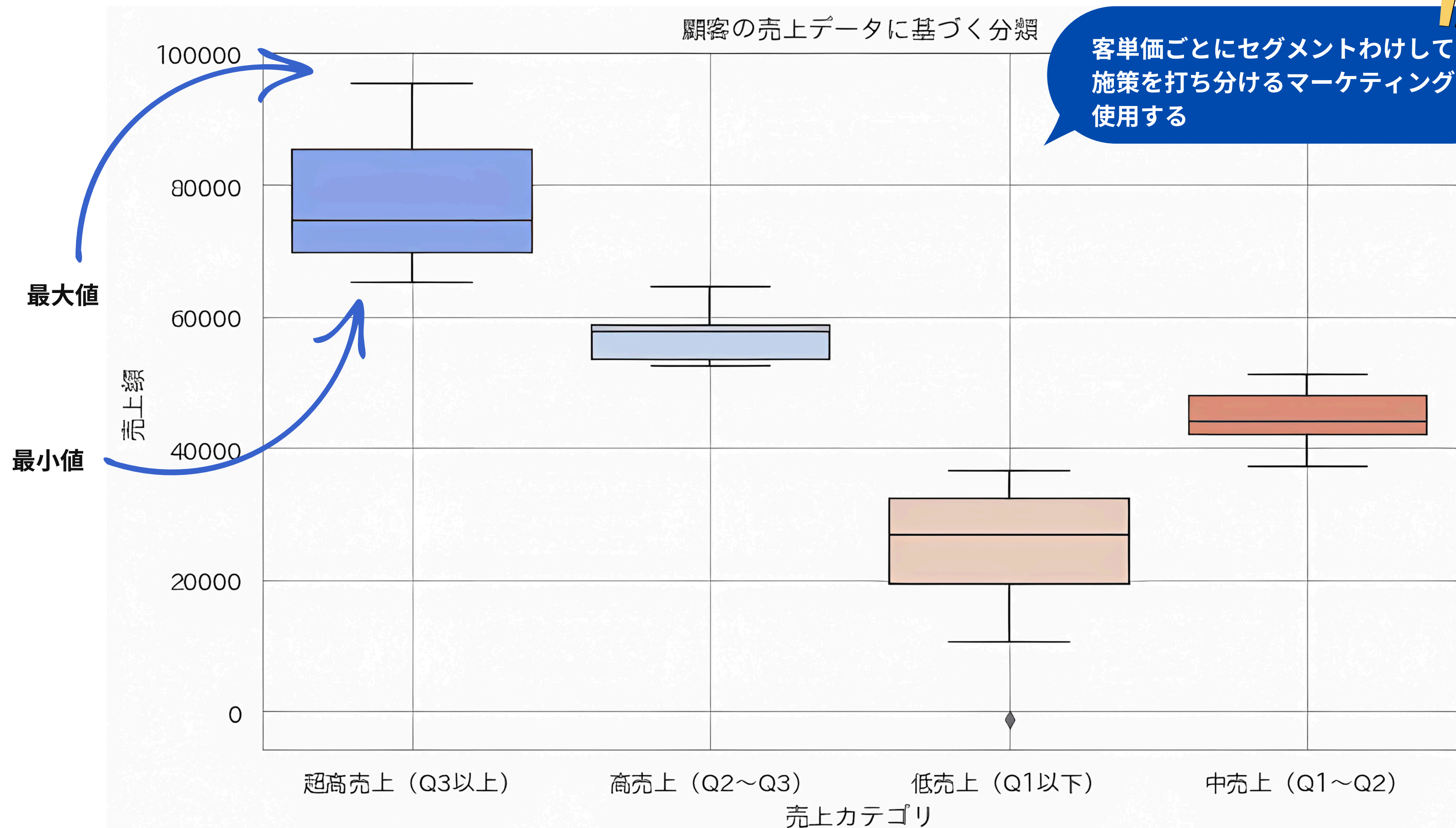
顧客ごとの年間購入額を小さい順に並べ、第1四分位数・中央値・第3四分位数を求める。
これらを区切りとして顧客をおおむね4つの層に分け、各層の購入頻度や継続率を比較する。
同額の顧客が多い場合、各層の人数は厳密に25%ずつにはならない。

▶ 購入額の分布を把握し、顧客層ごとの施策を検討する材料にできる。



ポイント

四分位数が一般的。
データを効率的に切り分けて、ビジネスの意思決定に利用することが多い。



度数

ある範囲にある、データの個数。

例えば店舗ごとの売上をある階級ごとに分けて、**各階級に属するデータをカウントしたものが度数**となる。

$$f_k = \sum_{j=1}^n \mathbf{I}(x_j \in C_k)$$

クラスkの度数は、指示関数 $\mathbf{I}(\cdot)$ を使って上のように表せる。括弧内の条件が成り立てば1、成り立たなければ0となり、階級 C_k に入るデータだけを数える。

累積度数

ある範囲以下に属するデータの総数。
それまでの度数を合計して算出する。

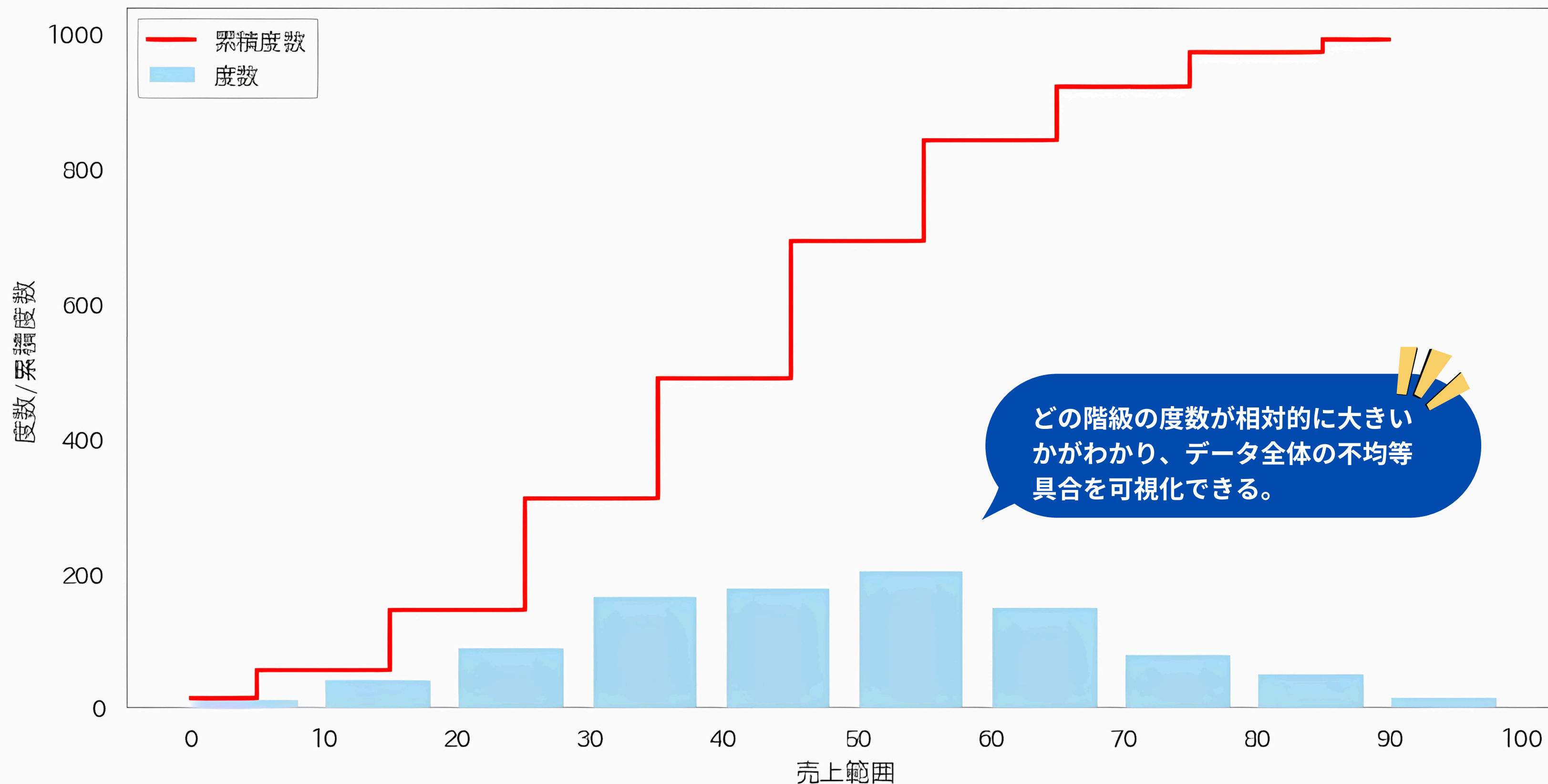
$$F_k = \sum_{j=1}^k f_j$$

階級値は、各階級の中央の値。
階級を代表する値として使われる。

階級 (万円)	階級値	度数	相対度数	累積相対度数
0以上10未満	5	13	0.013	0.013
10以上20未満	15	40	0.04	0.053
...	...	...	...	...
合計	-	1000	1.0	-

店舗売上の度数分布表

度数と累積度数の分布（1000店舗の売上データ）



学んだ内容を問題で確認

この章に対応する演習問題を3問選びました。リンクまたはQRコードから挑戦できます。

問題 1

レベル1

平均値と中央値

5つのデータから、平均値と中央値を計算する。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 2

レベル2

四分位数と箱ひげ図

データの第1四分位数と第3四分位数を求める。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 3

レベル1

度数分布表と相対度数

度数分布表から、相対度数が最も大きい階級を答える。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます



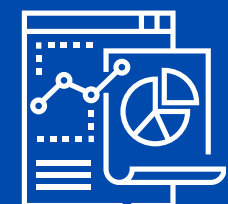
ポイント

この章のほかの問題にも、Web版「青の統計学」で挑戦できる。

統計検定2級・3級対策の演習問題と模擬試験を用意しています。 app.statisticsschool.com



1. 記述統計
2. データの散らばりの指標
3. 確率と確率分布
4. 相関と回帰
5. 統計的推測



期待値

期待値の定義と計算方法

期待値

確率変数がとりうる平均的な値を表す指標

定義

離散的な確率変数の期待値は、以下の式で定義される

 x_i

実現値

 p_i

その値が起こる確率

$$E[X] = \sum_{i=1}^n x_i p_i$$

具体例

保険会社が新しい医療保険商品を販売する。この商品では、年間保険料が10万円で、顧客が入院した場合に平均50万円の補償を支払う。1年間で入院する確率を2%とすると、一人の顧客当たりの期待収益はいくらか

入院しない場合

- 保険料収入: 10万円
- 支払い: 0円
- 純利益: 10万円

入院する場合

- 保険料収入: 10万円
- 補償金支払い: -50万円
- 純利益: -40万円



$$E[X] = 0.98 \times 10 + 0.02 \times (-40) = 9$$

「純利益」は管理費等を考慮していない簡略例



ポイント

期待値は**実現値**と**確率の積**の合計

期待値

平均値との違い

期待値

定義

確率分布に基づく理論上の平均値。
各値を、その値が起こる確率で重み付けする。

計算方法

$$E[X] = \sum_{i=1}^n x_i p_i$$

特徴

確率モデルから計算する。
将来の予測に限らず、分布の中心を表す。

計算例

サイコロの目（1～6）の期待値

$$1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

3.5という目は実際に出ないが、
「理論的な平均」として解釈される

標本平均

定義

観測したデータから計算する代表値。

計算方法

$$\text{平均値} = \frac{\text{データの合計}}{\text{データの個数}}$$

特徴

既存のデータから計算する実績値

計算例

テストの得点が60点、70点、
80点だった場合

$$\frac{(60+70+80)}{3} = 70$$



ポイント

期待値は確率分布の平均。
標本平均は観測データの平均であり、母平均を推定するためにも使う。

分散と標準偏差

定義と計算方法

分散

データの広がり具合やばらつきを定量的に示した指標
単位は元のデータの単位の2乗になる

$$V[X] = E[(X - E[X])^2] = \sum_{i=1}^n (x_i - E[X])^2 p_i$$

▼ より計算しやすい形式にすると・・・

$$V[X] = E[X^2] - (E[X])^2$$

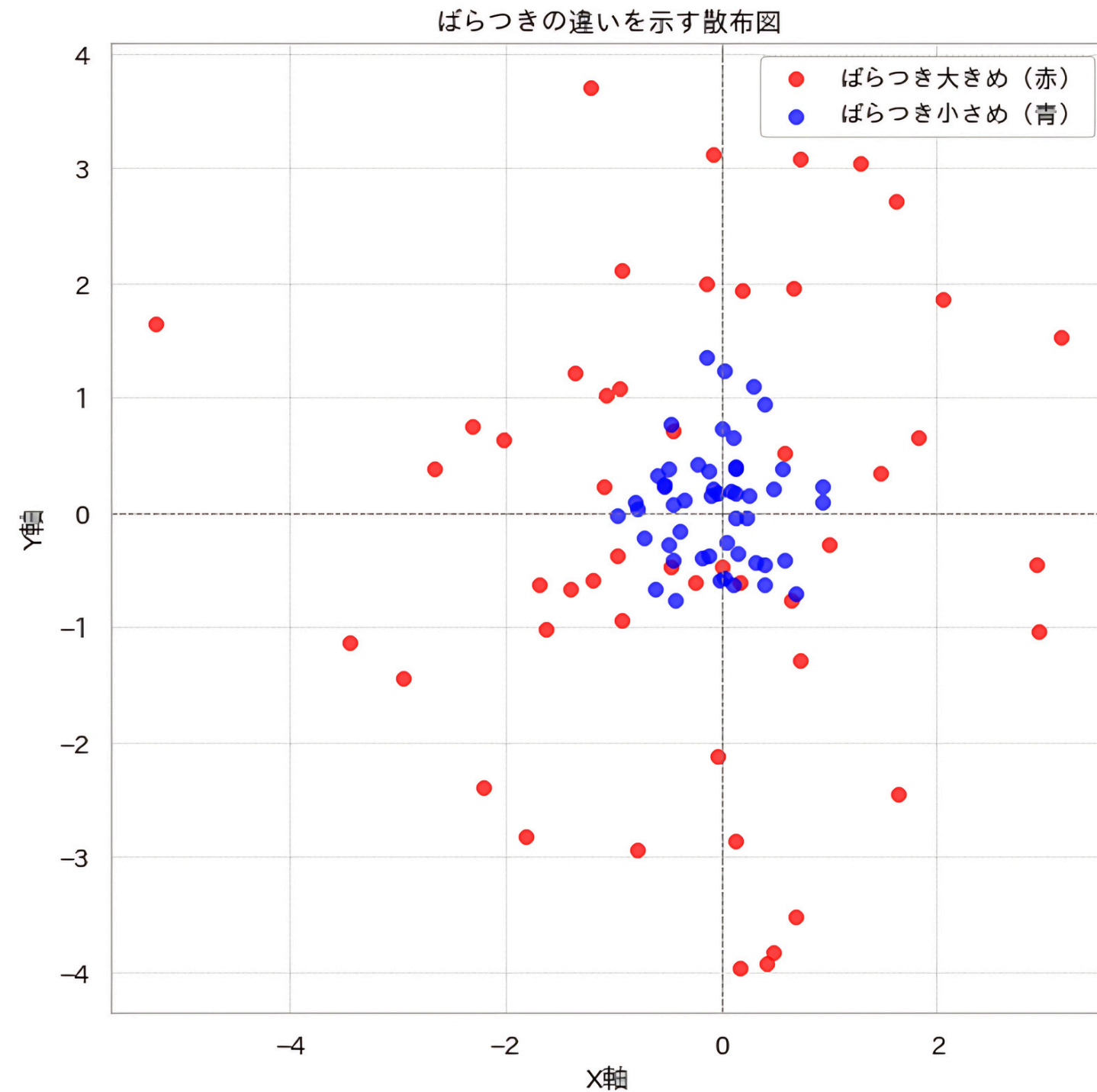
分散はデータの散らばり具合を表す統計量。期待値との差の二乗平均で計算される

標準偏差

データのばらつきの指標の一つ。
実際のデータのスケールに合わせるのが特徴で、分散の平方根として表される

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\sum_{i=1}^n (x_i - \mu)^2 p_i}, \quad \mu = E[X]$$

ここでは離散分布の標準偏差を示す。すべての値が等確率1/nで現れる場合は、偏差の二乗を単純平均してから平方根を取る。

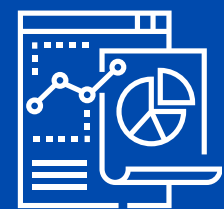


青い点の集合（ばらつき小さめ）

データが平均値付近に集中しており、ばらつきが小さい。分散が小さいと、値が平均に近い範囲に収まりやすくなる。

赤い点の集合（ばらつき大きめ）

データが広い範囲に分布しており、ばらつきが大きい。分散が大きい場合、値が平均から遠く離れる可能性が高くなる。



分散

定義

データのばらつきを表す指標で
データが平均値から
どれだけ離れているかの「二乗の平均」

使い分け

理論的な解析や確率論で使われることが多い。単位が「元のデータの単位の二乗」になるため、**直接の解釈が難しい場合がある。**

計算方法

$$V[X] = E[X^2] - (E[X])^2$$

標準偏差

定義

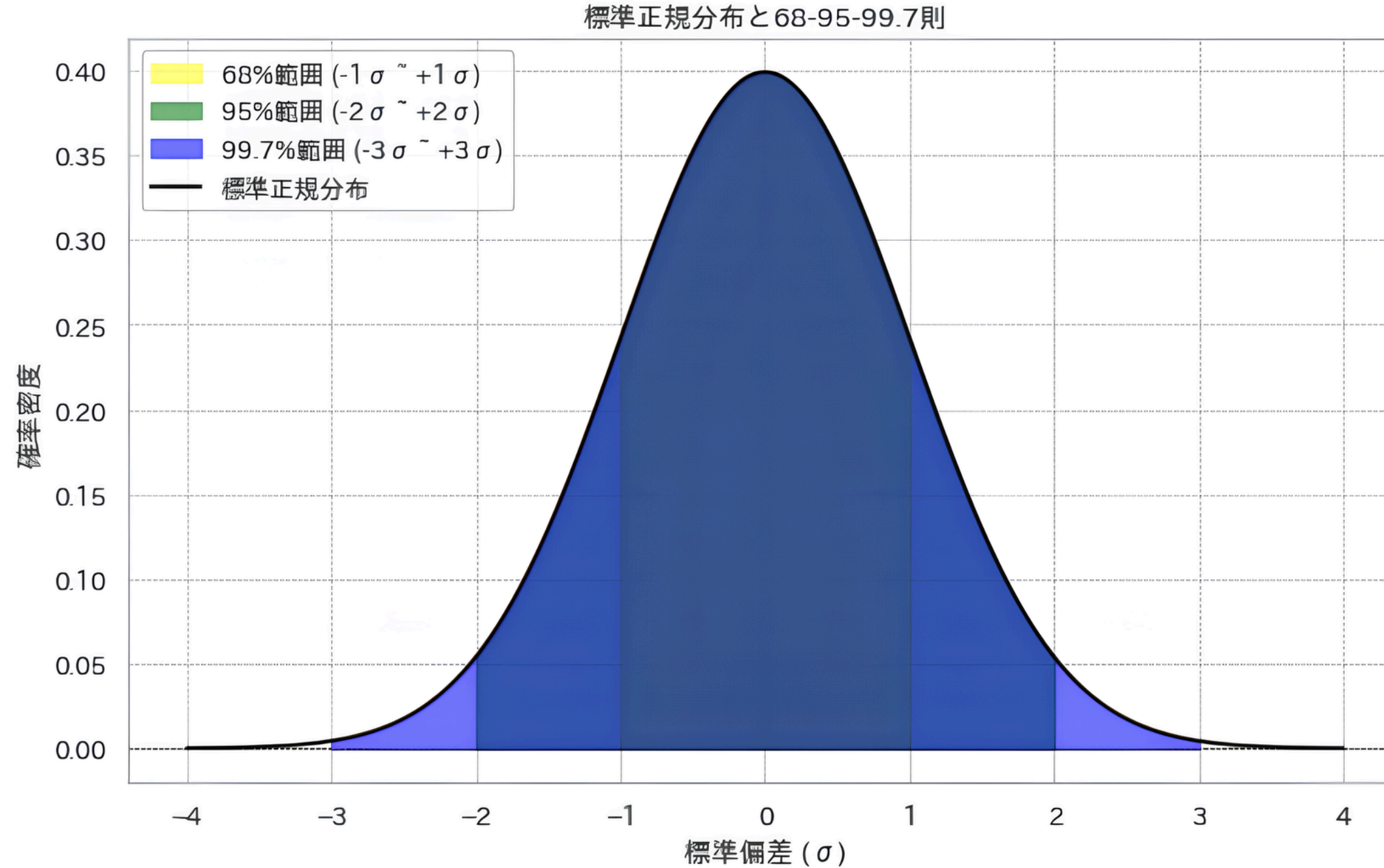
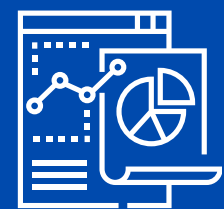
分散の平方根を取った値で
データのばらつきを「元の単位」で表す

使い分け

データのばらつきを直感的に理解したり、比較したりするために使用される。
元のデータと同じ単位なので解釈しやすい。

計算方法

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{\sum_{i=1}^n (x_i - \mu)^2 p_i}$$



$\mu \pm 1\sigma$ の範囲：約68.27%のデータが含まれる

$\mu \pm 2\sigma$ の範囲：約95.45%のデータが含まれる

$\mu \pm 3\sigma$ の範囲：約99.73%のデータが含まれる



「68-95-99.7則」と呼ばれる

分散と標準偏差

具体例

具体例

投資家が特定の株式 A に投資している。株式の将来のリターンの確率分布が右のように与えられるとする。ポートフォリオ内の株式リターンの分散を考えたい。

期待リターンの計算

$$E(R) = 0.2 \times (-0.05) + (0.5 \times 0.05) + (0.3 \times 0.15) = 0.06$$

期待リターンは6%となる。

分散の計算

①偏差の二乗を計算する

$$(R_1 - E(R))^2 = (-0.11)^2 = 0.0121$$

$$(R_2 - E(R))^2 = (-0.01)^2 = 0.0001$$

$$(R_3 - E(R))^2 = (0.09)^2 = 0.0081$$

②各項を確率で重み付けする

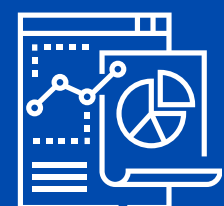
$$V(R) = 0.2 \cdot 0.0121 + 0.5 \cdot 0.0001 + 0.3 \cdot 0.0081 = 0.0049$$

この例では期待リターンは6%、リターンの標準偏差は7パーセントポイント。
リスクが大きいかどうかは、対象期間や比較する投資先を揃えて判断する。

$$SD(R) = 0.07$$

リターン (R)	確率
-5%	0.2
5%	0.5
15%	0.3





分散と標準偏差

標本分散と不偏分散

青の統計学

不偏性

推定量の期待値が真の母数に一致する性質

$$E[\hat{\theta}] = \theta$$

θ は、真の母数。モデルのパラメータなども θ で表すことが多い。

標本分散

標本内の散らばりを表す、分母 n の分散。本資料では v_n と書く。

$$v_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

一貫性

標本が無限に大きくなったときに推定量が母数に収束する性質を一貫性と呼ぶ。不偏性と合わせ、推定量の望ましい性質。



独立・同一分布の標本では、 $E[v_n] = (n-1)\sigma^2/n$ となり、母分散を平均的に小さく見積もる。



ポイント

標本分散は**実際の標本の散らばり**を示すが、統計的推測に重要な性質「**不偏性**」はもたない。不偏性は、推定量が**バイアスを持たない**ことを保証している。

分散と標準偏差

標本分散と不偏分散

なぜ標本分散だと不偏性がないのか

標本平均は、同じ標本のデータを使って決まる。

偏差の和が0になるという制約があるため、偏差平方和をnで割ると、母分散を平均的に小さく見積もる

$$E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = (n-1)\sigma^2, \quad E[v_n] = \frac{n-1}{n}\sigma^2$$

不偏分散

独立・同一分布の標本について、分母をn-1にすると期待値が母分散に一致する。
本資料では、この不偏分散を s^2 と書く。

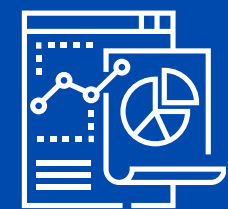
$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

これは「どの標本でも必ず小さくなる」という意味ではなく、
標本抽出を繰り返したときの平均についての性質。



ポイント

標本分散は**実際の標本の散らばり**を示す
不偏分散は**母集団の分散**を推定する際に用いる補正済みの値



Excelでは、どちらの分散が使われているのか

VAR.Pを使う場合

入力されたデータを対象全体として扱い、偏差平方和をnで割る

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

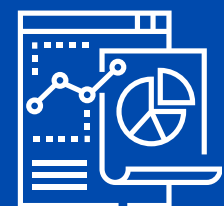
VAR.Sを使う場合

入力されたデータを標本として扱い、偏差平方和をn-1で割る。

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$



* 標準偏差を求めるExcel関数はSTDEV.P、STDEV.S。



変動係数

定義と計算方法

変動係数

データのばらつきの程度を比較するための指標で、標準偏差を平均で割った値

$$CV = \frac{\sigma}{\mu} \times 100$$

▶ 0に「量がない」という意味がある尺度で、平均が正のとき、平均に対する相対的なばらつきを比較する。
平均が0に近い場合は値が不安定になる。

計算例

店舗の日次売上

店舗A：平均80万円、標準偏差10万円。店舗B：平均100万円、標準偏差15万円。



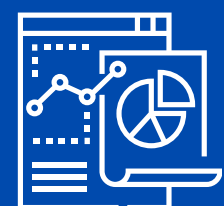
店舗Aの変動係数は $\frac{10}{80} \times 100 = 12.5\%$ 店舗Bの変動係数は $\frac{15}{100} \times 100 = 15\%$

この例では、店舗Bの方が平均売上に対する相対的なばらつきが大きい。
変動係数は、摂氏温度のように0が任意の尺度にはそのまま使えない。



ポイント

変動係数は元の大きさが全然違う2つのデータであってもそのばらつきを比較できる



練習問題

Web版「青の統計学」で腕試し

青の統計学

学んだ内容を問題で確認

この章に対応する演習問題を3問選びました。リンクまたはQRコードから挑戦できます。

問題 1

レベル1

分散と標準偏差

データ 2, 4, 6, 8, 10 の標準偏差を求める。



▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 2

レベル2

変動係数

平均50、標準偏差10のデータの変動係数を求める。



▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 3

レベル2

不偏分散の理解

標本分散の期待値と母分散の関係を考える。



▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます



ポイント

この章のほかの問題にも、Web版「青の統計学」で挑戦できる。

統計検定2級・3級対策の演習問題と模擬試験を用意しています。 app.statisticsschool.com

目次



1. 記述統計
2. データの散らばりの指標
- 3. 確率と確率分布**
4. 相関と回帰
5. 統計的推測

独立な試行

2つの事象A、Bについて、一方が起きたと知っても他方の確率が変わらないとき、独立という。
独立なら、同時に起きる確率は個別の確率の積になる。

事象AとBが独立だと、同時確率は個別の確率の積で表せる

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

iid: 独立かつ同一な分布に従う

多くの問題設定では前提とされるが、実データに当てはまらないこともある。

具体例



オンライン広告を表示したとき、ユーザーが広告をクリックするかどうかを分析する。

- 試行: 広告を1回表示すること。
- 結果: ユーザーがクリックする（成功）か、クリックしない（失敗）

分析では、各表示のクリック有無が独立で、クリック確率が共通であると仮定する。
同じユーザーへの繰り返し表示や配信条件の違いがある場合は、この仮定を確認する。



ポイント

独立性は、直接の因果関係がないことと同じではない。

共通の要因によって関連が生じる場合がある。時系列データでも独立性の確認が必要。

条件付き確率 ある事象が既に起こったという条件のもとで、別の事象が起こる確率

定義 事象 A が起こったという条件のもとで、事象 B が起こる確率

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

ちなみに、事象 A と B が独立な時は以下のように表せる

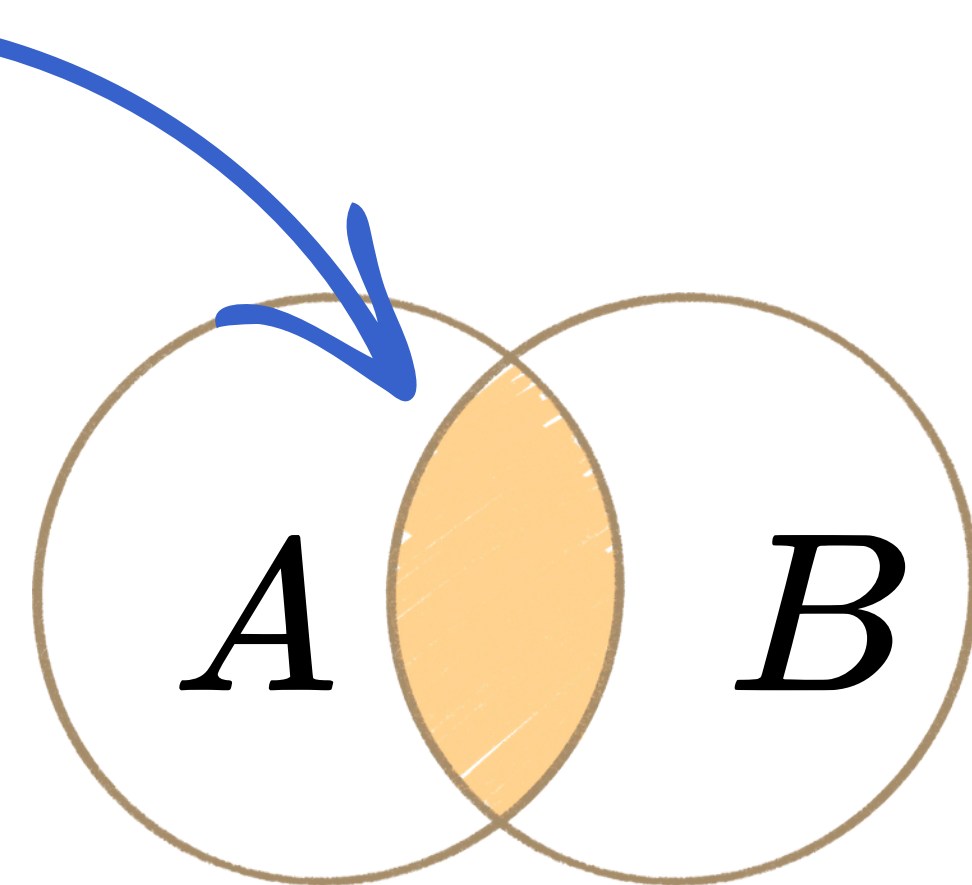
$$P(B|A) = P(B)$$

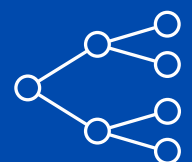
事象 A が起こることが分かっているにもかかわらず、事象 B が起こる確率は変わらないため、条件付き確率は単に事象 B の確率と等しくなる。



ポイント

条件付き確率では、ある事象が起こった条件のもとで別の事象が起こる確率を求める





具体例

一定期間のWebサービスのログから、「Aページを訪れたユーザーのうち、同じ期間にBページも訪れたユーザーの割合」を求める。訪問の前後関係は問わない。

データ

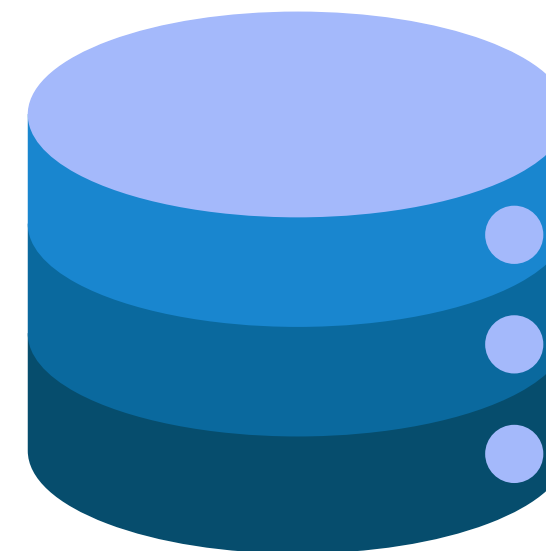
- 対象期間の全ユーザー数 $N=100,000$ 人
- Aページを訪れたユーザー数 $N(A)=20,000$ 人
- AとBの両方を訪れたユーザー数 $N(A \cap B)=5,000$ 人

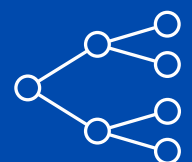
計算

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

$$P(B|A) = \frac{5000}{20000} = 0.25$$

▶ Aページを訪れたユーザーのうち、Bページも訪れたユーザーの割合は25%。





確率変数

確率変数は、試行の各結果に数値を対応させるもの。
試行前には値が不確定であり、どの値をどのくらいの確率で取るかを確率分布で表す

離散型

$$P(X = x_i) = p_i$$

有限個または数え上げられる個数の値を取る
確率変数。

例：サイコロを振る実験の結果・コインを投げる結果

連続型

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

連続的な範囲の数値を取る確率変数

例：身長、体重、温度、時間などの連続的な数値

ランダム性

通常は複数の値を取り得る。なお、確率1で同じ値を取る定数も、確率変数の一種として扱える。

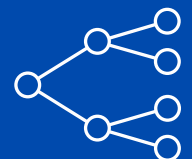
例えば、サイコロの目を表す確率変数 X は、1から6までの値を取るが、各値の確率は等しく

$$\frac{1}{6}$$



ポイント

確率変数は単なる変数ではなく、**不確実性を持つ量を数学的に表現する道具**



確率変数の期待値

実現値に対して、各事象が起きる確率を掛けたものの和であることには変わらない。

離散型

各値とその確率の積の和

$$E(X) = \sum x_i p_i$$

連続型

確率密度関数 $f(x)$ に基づき積分として求められる

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

期待値の性質

線形性

$$E(aX + b) = aE(X) + b$$

和の期待値 期待値が存在すれば、独立性を仮定しなくても成り立つ。

$$E(X + Y) = E(X) + E(Y)$$

分散の性質

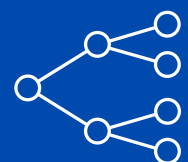
スカラー倍の変化

$$Var(aX + b) = a^2 \cdot Var(X)$$

分散の定義と計算式

$$\begin{aligned} Var(X) &= E[(X - E(X))^2] \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

変数に関係のない値は、外に出すと二乗されるというのがポイント



分散の加法性

和の分散には共分散が関わる。共分散が0なら、それぞれの分散を足せる。
以下では分散が有限であるとする

独立

$$Var(X + Y) = Var(X) + Var(Y)$$

独立なら共分散は0。ただし、共分散が0だからといって独立とは限らない。

独立
ではない

$$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$$

共分散

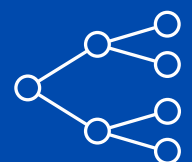
2つの確率変数がどのように一緒に変動するかを表す指標

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

変形



$$Cov(X, Y) = E[XY] - E[X]E[Y]$$



確率変数

確率密度関数

確率密度関数

連続型確率変数に対する確率の分布を表す関数

定義 任意の x に対して、値が常に非負

$$f(x) \geq 0 \quad \text{for all } x$$

全範囲で積分すると1になる

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

計算 確率変数 X が区間 $[a, b]$ に属する確率は、確率密度関数を積分することで求められる

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



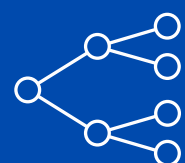
ポイント

- 確率密度関数 $f(x)$ 自体は確率を表さないが、**その面積（積分値）が確率になる。**
- 値が大きい部分は、その付近で確率変数 X が現れる可能性が高いことを意味する。

確率質量関数

$$P(X = x)$$

離散型確率変数に対する確率を表す関数。密度関数とは異なり、特定の値を取る確率そのものを直接表す。



ベルヌーイ分布

成功（1）または失敗（0）の二値を取る確率分布。

確率変数 X がベルヌーイ分布に従う場合、成功確率を p とした場合の確率質量関数は以下

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

$x = 0, 1$ 、 $0 \leq p \leq 1$ 。これ以外の x での確率は0。

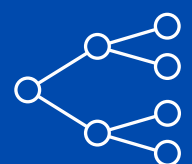
期待値

$$E[X] = \sum_{x=0}^1 x P(X = x) = 0 * (1 - p) + 1 * p = p$$

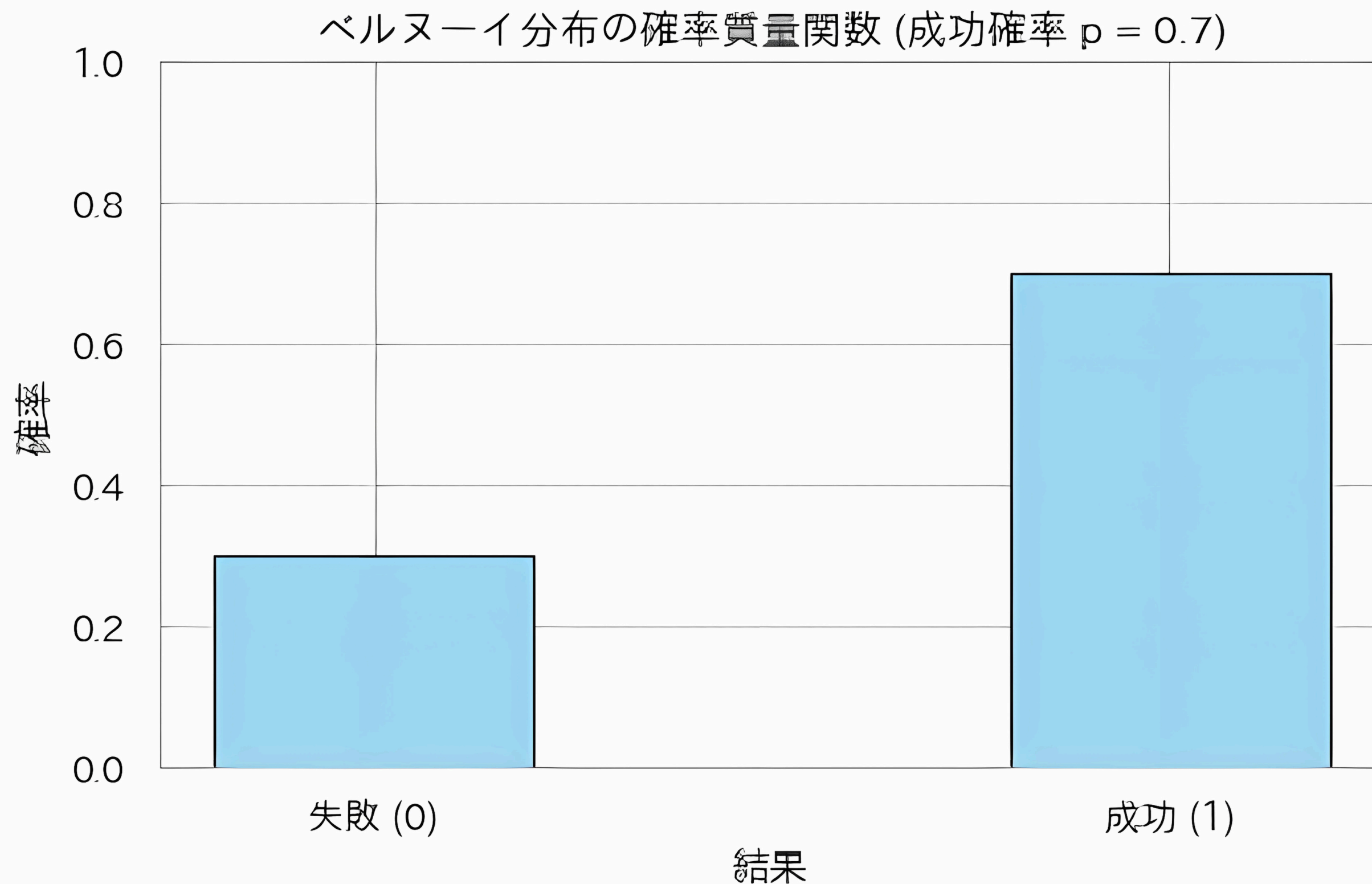
分散

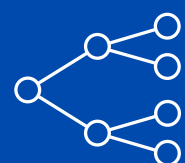
$$Var(X) = p(1 - p)$$





確率質量関数





二項分布

成功確率 p が共通の、 n 回の独立なベルヌーイ試行における成功回数の分布。

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

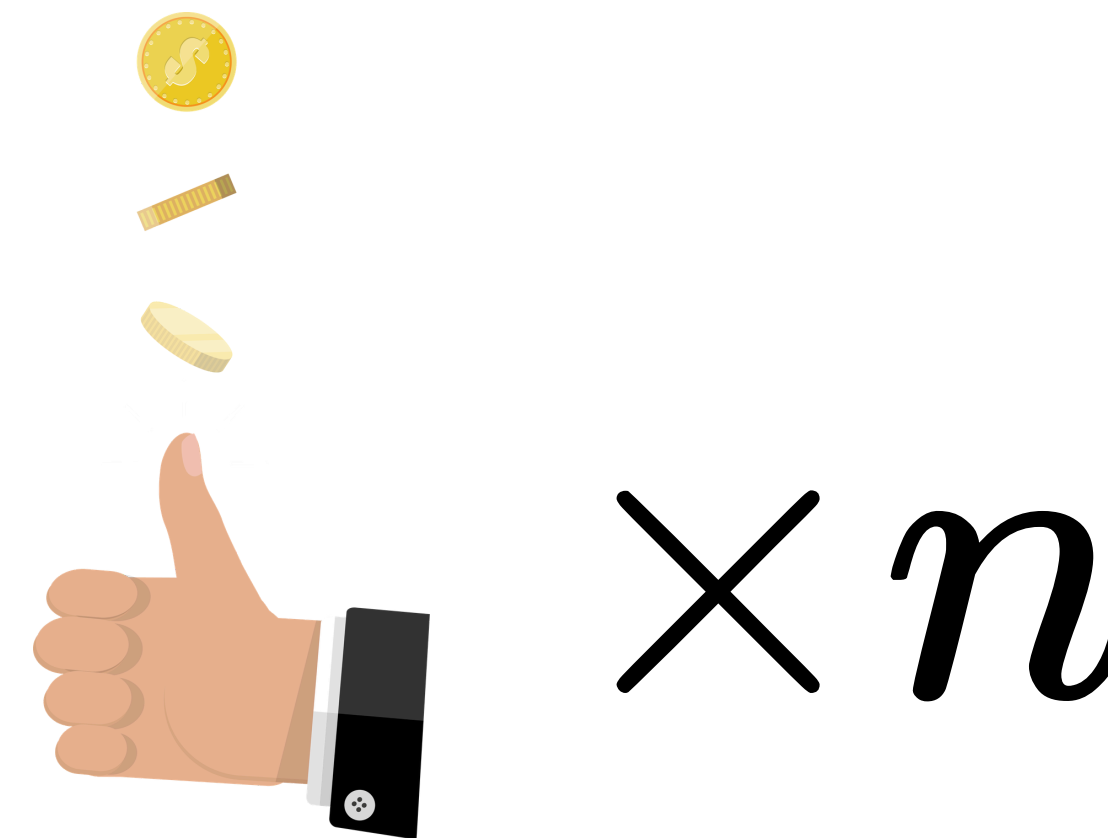
$$x = 0, 1, \dots, n, 0 \leq p \leq 1.$$

期待値

$$E[X] = np$$

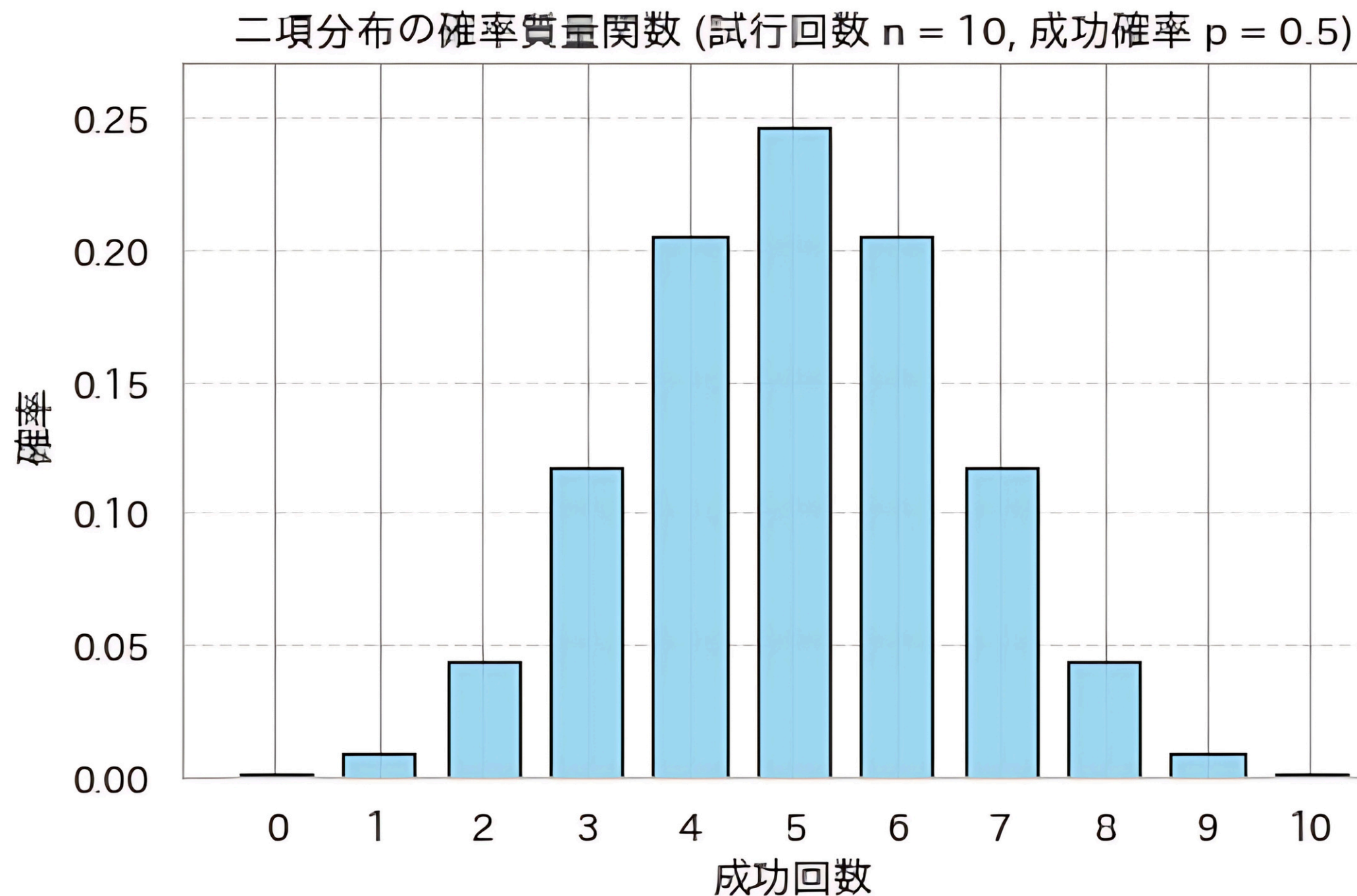
分散

$$\text{Var}(X) = np(1 - p)$$



ベルヌーイ試行を n 回繰り返したものの。コイン投げなど多くの実データに当てはめられる。

確率質量関数

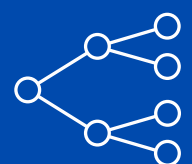


nを固定すると、平均 np は p とともに増える。分散 $np(1-p)$ は $p=0.5$ で最大となり、 p が0や1に近づくと小さくなる。



ポイント

平均 np は p とともに増加する。分散 $np(1-p)$ は $p=0.5$ で最大になる。
 $p=0.5$ の理論上の二項分布は厳密に左右対称であり、上の図もその理論分布。



正規分布

多くの現象に適合する連続確率分布。
平均値の周りにデータが対称的に分布する特徴を持つ。

確率密度関数

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

期待値

$$E[X] = \mu$$

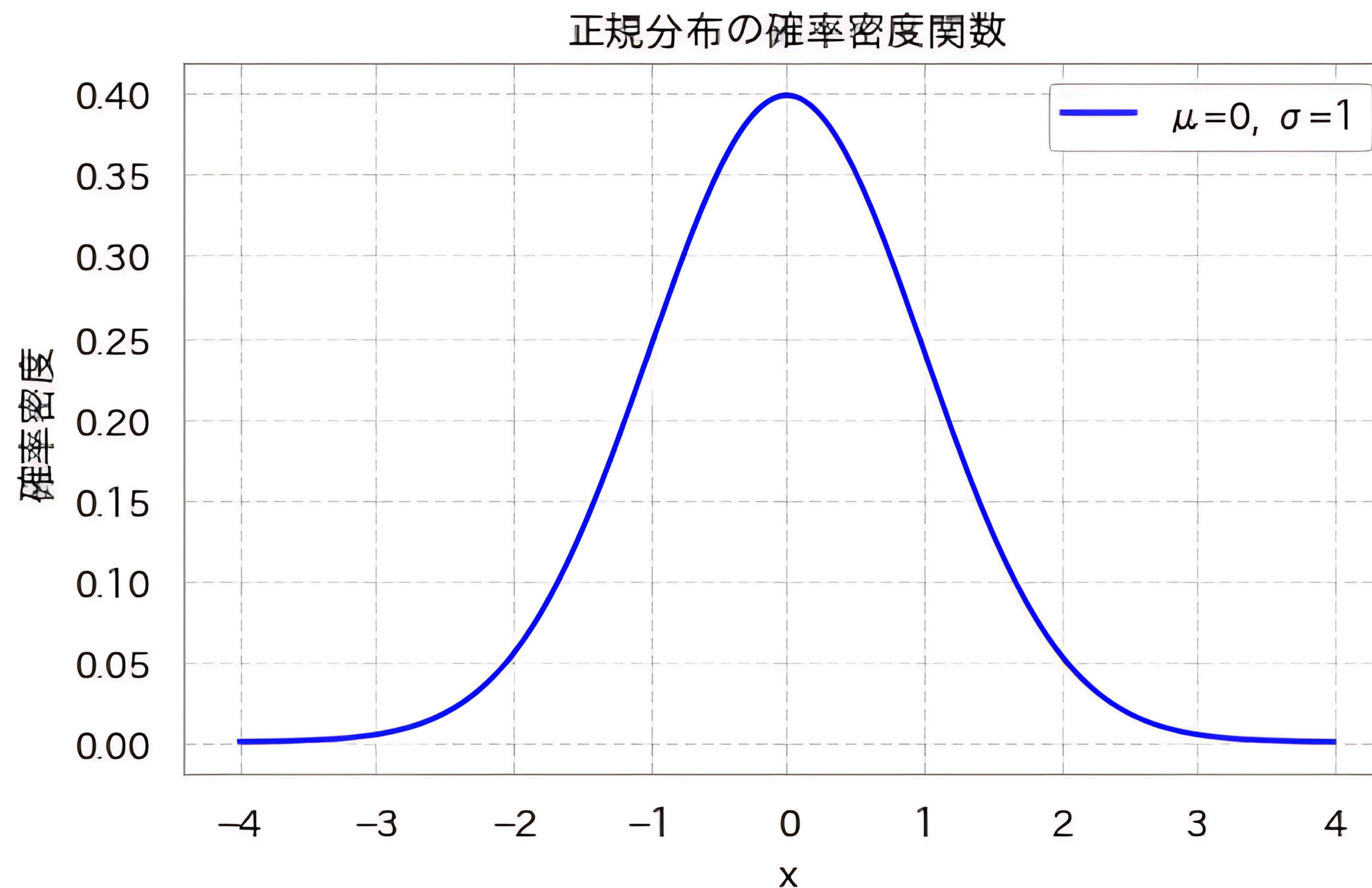
分散

$$Var[X] = \sigma^2$$

確率変数が、ある分布（正規分布）に従うという意味

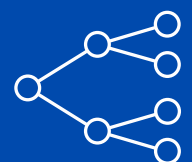
$$X \sim N(\mu, \sigma^2)$$

確率密度関数



ポイント

正規分布は**左右対称でベル型の形状**をしており、平均値で最大値を取る。分布の広がりは分散に依存し、分散が大きいほど分布は広がり、小さいほど尖る。



色々な確率分布

正規分布の特性

加法性

独立な正規分布に従う変数の和も正規分布に従う。

$$X \sim N(\mu_1, \sigma_1^2) \quad Y \sim N(\mu_2, \sigma_2^2) \quad \text{の場合}$$

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$



独立でない場合、和の分散には共分散が加わる。

ただし、各変数がそれぞれ正規分布に従うだけでは、和の正規性は保証されない。

$$V(X + Y) = \sigma_1^2 + \sigma_2^2 + 2Cov(X, Y)$$



ポイント

独立な正規確率変数の和は正規分布に従い、平均と分散はそれぞれ足し合わせられる。

確率変数同士の和も同じ種類の分布に従う性質を「再生性」と呼び、ポアソン分布などの他の分布も持っている

標準正規分布

正規分布の中でも、平均が0で分散が1のもの。
他の正規分布のデータを標準化（zスコア変換）することで変換できる。

確率密度関数

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

期待値

$$E[X] = 0$$

分散

$$\text{Var}[X] = 1$$

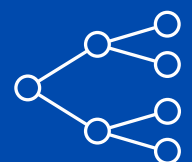
標準化：正規分布に従う確率変数X
を標準正規分布に従う変数zにできる

$$z = \frac{x - \mu}{\sigma}$$



ポイント

標準化することで、異なる単位のデータを同じ基準で評価できる。
機械学習等の前処理でも使われる



二項分布の正規近似

二項分布は、 np と $n(1-p)$ がともに十分大きいとき、正規分布で近似できる。
 n が大きくても、 p が0や1に非常に近い場合は注意が必要

期待値と分散

二項分布の期待値と分散は次の通り

期待値

$$E[X] = np$$

分散

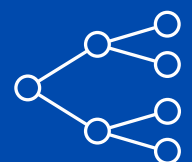
$$Var(X) = np(1 - p)$$



$$X \sim \text{Bin}(n, p), \quad X \overset{\text{approx}}{\sim} N(np, np(1 - p))$$

二項分布の平均・分散と同じ平均・分散を持つ正規分布で近似する。有限の n では、二項分布自体が正規分布に変わるわけではない。

この近似は**中心極限定理**に基づく。独立・同一分布で分散が有限などの条件の下で、標準化した和の分布が標準正規分布に近づく。



色々な確率分布

二項分布の正規近似

標準化

平均を引いて標準偏差で割り、平均0・分散1に標準化する。
正規近似の条件を満たすとき、標準正規分布で近似できる。

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \stackrel{\text{approx}}{\sim} N(0, 1)$$

標準化だけで正規分布になるわけではない。
二項分布の正規近似を使っている。

何が嬉しいのか？

具体例1

二項分布に基づく信頼区間を計算する際に、正規近似を利用する。
例えば、ある二項分布の確率 p を推定する時、以下のような信頼区間を考えられる

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

標準誤差

標本比率の分散 $p(1-p)/n$ に $\hat{p}$ を
代入し、その平方根を取る。



ポイント

$z_{\alpha/2}$ は標準正規分布の上側 $\alpha/2$ 点で、95%信頼区間では約1.96。
上の信頼区間はWald近似であり、小標本や $\hat{p}$ が0・1に近い場合には不正確になりやすい。

具体例2

母比率のz検定では、帰無仮説の下で標本比率を正規分布で近似し、検定統計量を計算する。

二項分布の成功確率 p についての帰無仮説に対しての検定統計量は以下になる

帰無仮説

$$H_0 : p = p_0$$

z統計量

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \stackrel{H_0, \text{approx}}{\sim} N(0, 1)$$

$$\hat{p}, p_0$$

左：サンプルの成功確率
右：帰無仮説での成功確率

独立な試行を仮定し、 np_0 と $n(1-p_0)$ がともに十分大きい場合の近似。小さい場合は二項分布による検定を検討する。



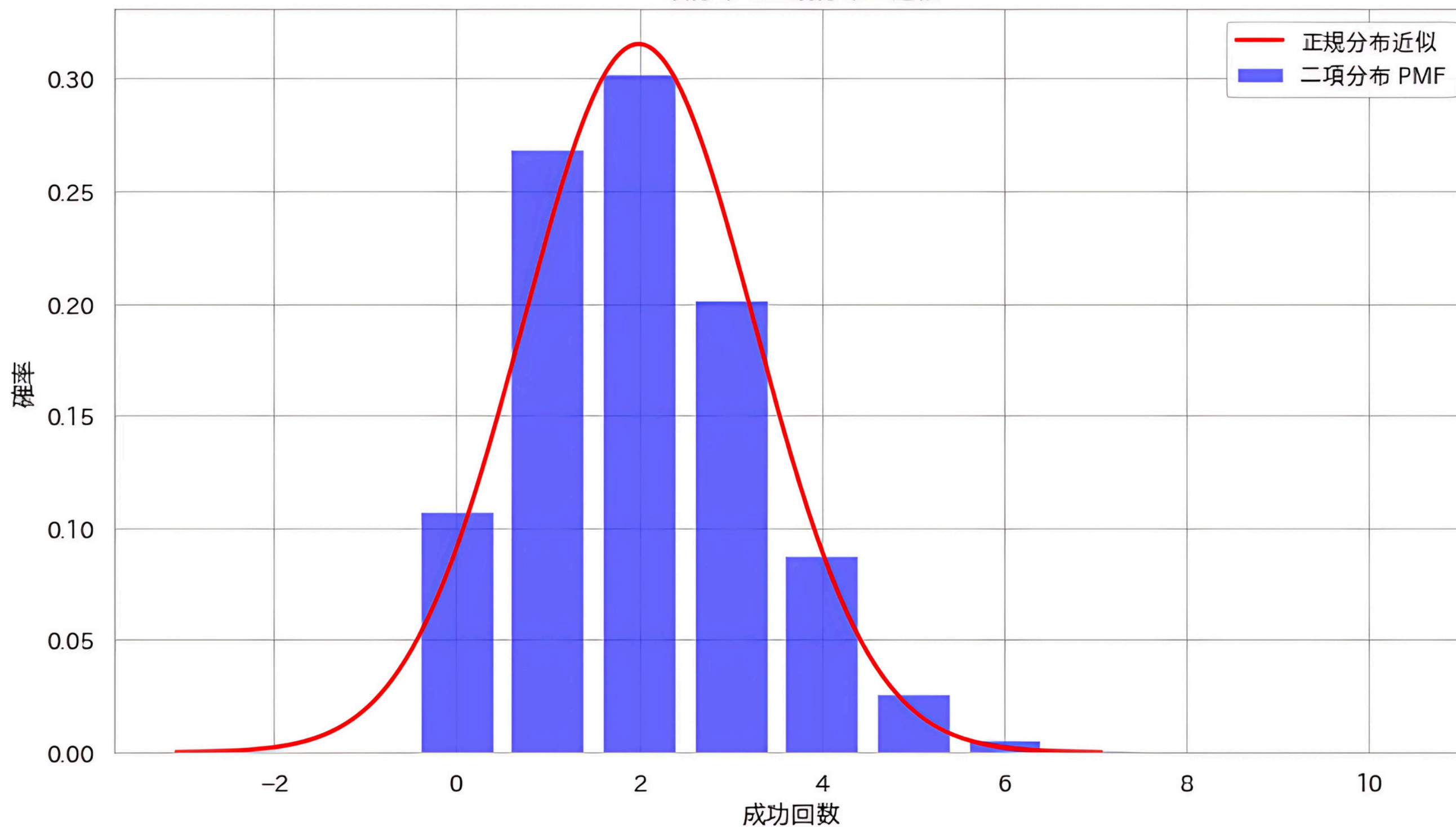
ポイント

zを棄却限界値と比較する、またはp値を有意水準 α と比較する。
右片側5%なら $z > 1.645$ 、両側5%なら $|z| > 1.96$ が通常の正規近似による基準。

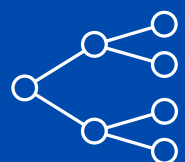
二項分布： $n=10, p=0.2$ の場合

$n=10$ 、 $p=0.2$ では $np=2$ と小さく、正規近似は粗い。この図は近似が十分でない例。

二項分布と正規分布の近似



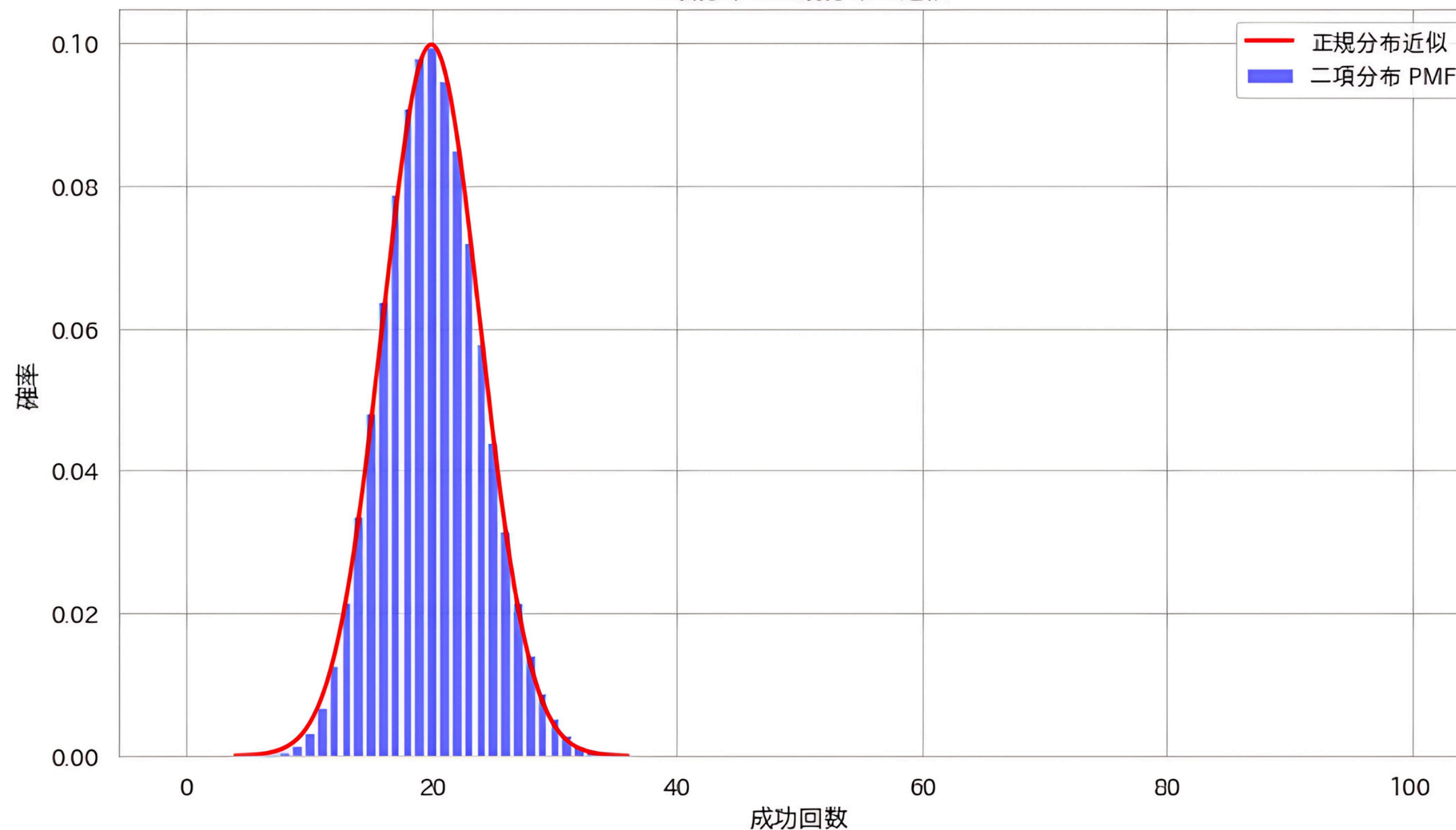
棒は各整数値の確率、曲線は正規分布の密度。確率を比較するときは曲線の面積を用いる。整数値 k の確率は、 $k \pm 0.5$ の区間の面積で近似できる



二項分布： $n=100, p=0.2$ の場合

$n=100$ 、 $p=0.2$ では $np=20$ 、 $n(1-p)=80$ となり、前ページより正規近似が良くなる。

二項分布と正規分布の近似



棒は各整数値の確率、曲線は正規分布の密度。確率を比較するときは曲線の面積を用いる。整数値 k の確率は、 $k \pm 0.5$ の区間の面積で近似できる

中心極限定理

独立で同一の分布に従う確率変数 $X_1, \dots, X_n$ について、平均 μ と分散 σ^2 ($0 < \sigma^2 < \infty$) が存在するとする。
 n が大きくなると、標準化した標本平均の分布は標準正規分布に収束する。

$$\bar{X}_n \stackrel{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

標準化



$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$$

サンプル数が多いとき、標準化された標本平均は標準正規分布 $N(0,1)$ に従うことがわかる。
左は大きな n での近似、右は $n \rightarrow \infty$ での分布収束を表す。



中心極限定理の成立条件

独立

各変数が互いに独立である。依存がある場合には、ここで紹介する定理をそのまま適用できない。

同一の分布に従う

各変数が共通の分布に従う。異なる分布の場合に使える拡張もあるが、追加条件が必要。

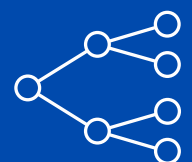
有限で性の分散

平均 μ と分散 σ^2 が存在し、 $0 < \sigma^2 < \infty$ である。分散が無限の場合、この標準化では結論が成り立たないことがある。



ポイント

元の分布が正規分布でなくても、上の条件の下では、大きな標本の平均の分布を正規分布で近似できる。必要な標本サイズは元の分布の歪みなどに依存する。



練習問題

Web版「青の統計学」で腕試し

学んだ内容を問題で確認

この章に対応する演習問題を3問選びました。リンクまたはQRコードから挑戦できます。

問題 1

レベル2

条件付き確率

2つのサイコロで、少なくとも1つが6のとき、両方が6である確率を求める。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 2

レベル2

二項分布の期待値と分散

$p=0.3$ 、 $n=10$ の二項分布の期待値と分散を求める。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 3

レベル2

正規分布の標準化

平均50、標準偏差10の正規分布で、 X が65を超える確率を求める。

▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます



ポイント

この章のほかの問題にも、Web版「青の統計学」で挑戦できる。

統計検定2級・3級対策の演習問題と模擬試験を用意しています。 app.statisticsschool.com



1. 記述統計
2. データの散らばりの指標
3. 確率と確率分布
- 4. 相関と回帰**
5. 統計的推測

相関係数

相関係数の定義と計算方法

相関係数

2つの変数間の 線形関係の強さと方向 を示す指標。
-1から+1の範囲 をとる。

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

▼ より簡単に

$$r = \frac{s_{xy}}{s_x s_y}, \quad s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

共分散は確率変数のスケールに依存するため、使いづらい。

s_x 、 s_y は分母 $n-1$ で求めた標本標準偏差、 s_{xy} は同じ分母で求めた標本共分散。両変数の標準偏差が0でない場合に定義される

偏相関係数

x と y をそれぞれ z に線形回帰し、その残差同士の相関を求める指標。
 z との線形な関係を調整するが、 **z のあらゆる影響や因果的な交絡を取り除けるとは限らない。**

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$



ポイント

相関係数は、共分散を x と y の標準偏差で正規化したもの。
スケールに依存せず、評価しやすい。

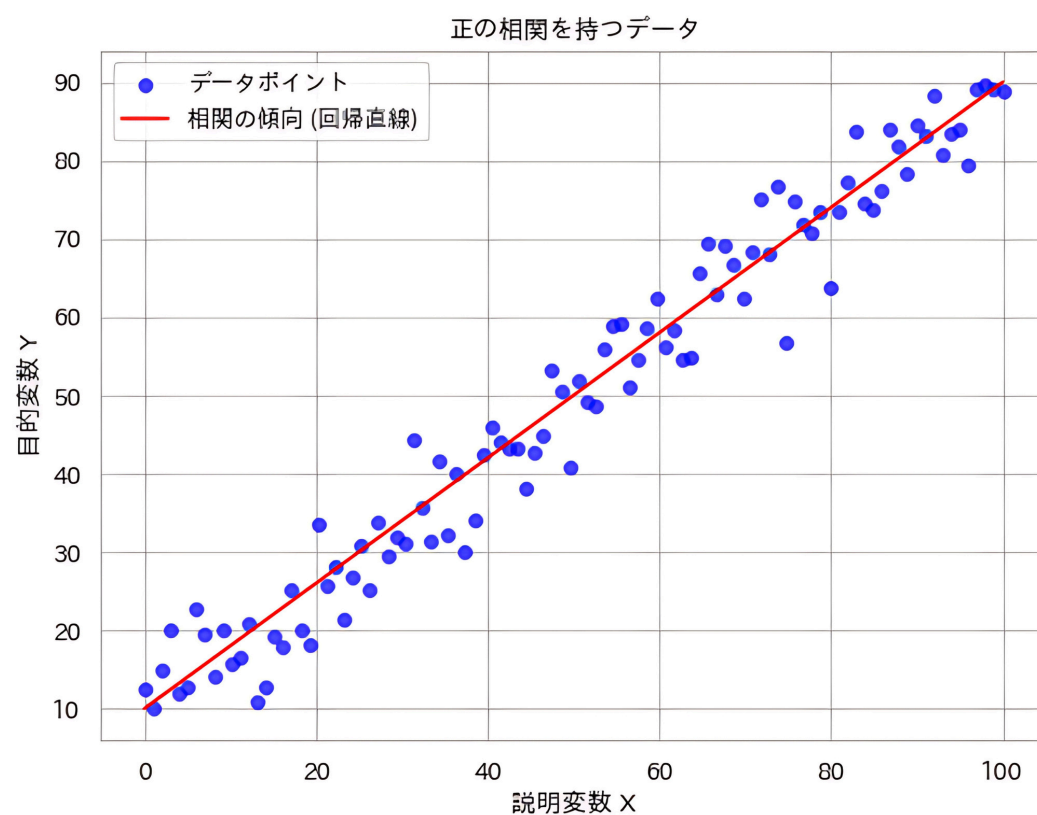
相関係数

正の相関・負の相関・無相関の違い

正の相関 ($0 < r \leq 1$)

一方の変数が増加すると
もう一方も増加する傾向。

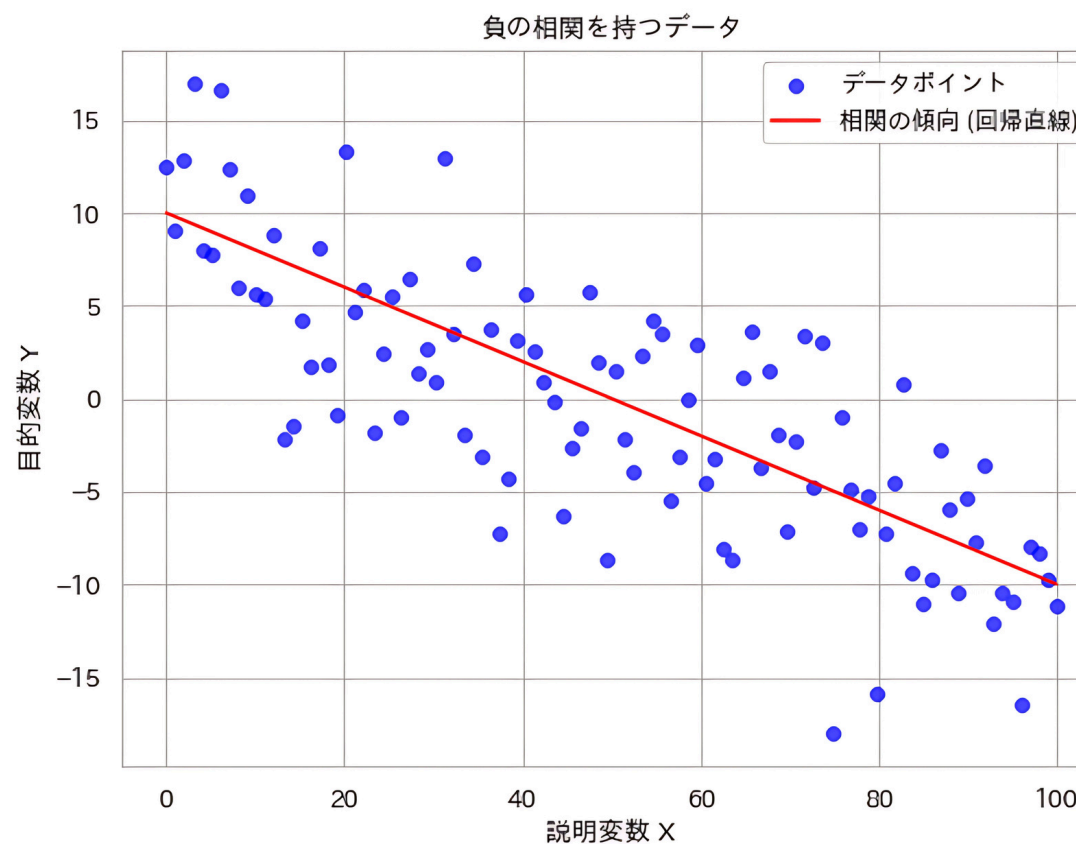
例：気温とアイスクリームの売上



負の相関 ($-1 \leq r < 0$)

一方の変数が増加すると
もう一方は減少する傾向。

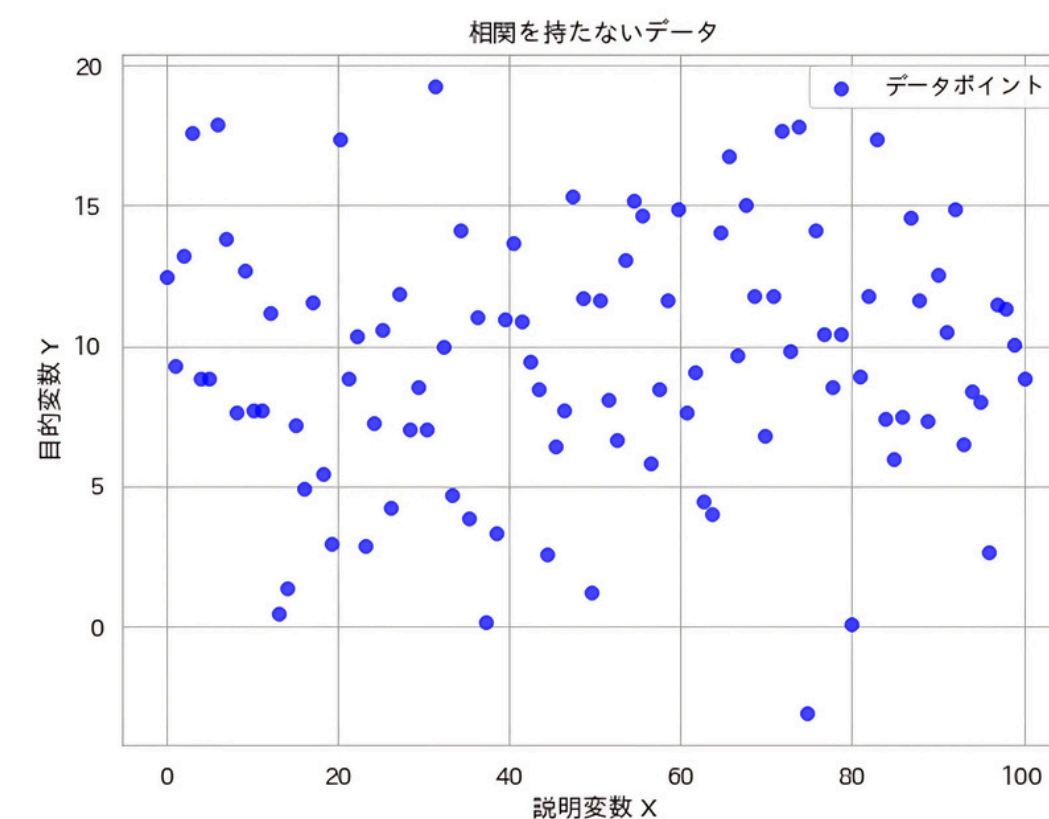
例：年齢と体力テストの成績（一般的に）



無相関 ($r=0$)

2つの変数に線形的な関係がない状態。

例：独立な2つの数値変数から生成した模擬データ



母相関が0でも、有限の標本から求めた相関係数は通常、厳密には0にならない。また、無相関でも非線形な関係がある場合がある。

相関係数

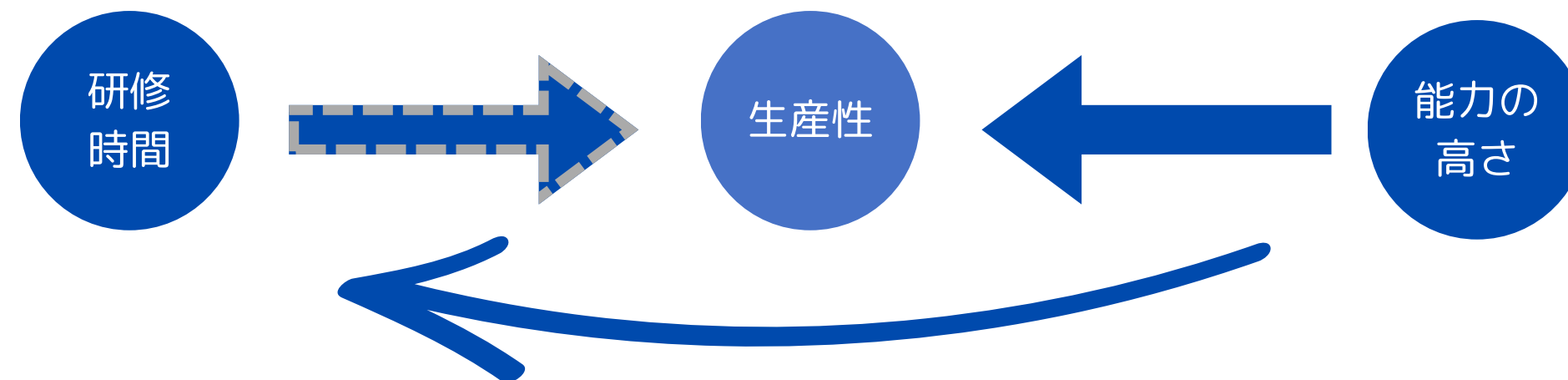
相関係数の限界

因果関係を証明するものではない

相関係数が高いからといって、一方が他方に影響を与えているとは限らない。第三の要因や偶然による「見かけの相関」（擬似相関）が存在する可能性がある。

具体例

ある会社が、従業員の研修時間と生産性の相関を調査したところ、高い相関が見られた。



例えば、元々能力の高い従業員ほど研修に積極的に参加し、生産性も高い可能性がある。この場合、能力の高さが共通の要因となる。相関だけでは、研修による効果と元々の能力の差を区別できない。

▶ 研修の効果を調べるには、無作為割付など、比較条件を揃えた評価が必要になる。



相関係数

相関係数の限界

外れ値に敏感

相関係数は、外れ値（極端な値）の影響を受けやすく、結果が大きく歪むことがある。

具体例

マ広告費と売上の相関を調べたところ、広告費も売上も特に大きい月があった。その月を含めると相関係数は0.95、除くと0.70だったとする。

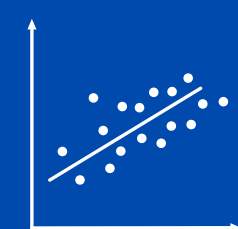
集計対象	相関係数（説明用の設定）
特定の月を含める	0.95
特定の月を除く	0.7

外れ値は自動的に削除しない。入力ミスか、実際に起きた重要な現象かを確認し、除外の根拠を明示する

これは、相関係数が一部の観測値に強く左右され得ることを示す。どちらの数値も、広告の因果効果を直接表すものではない。

○

外れ値を含めた結果と除いた結果を比較し、結論がどの程度変わるかを確認する。相関の強さと施策の効果は区別する。



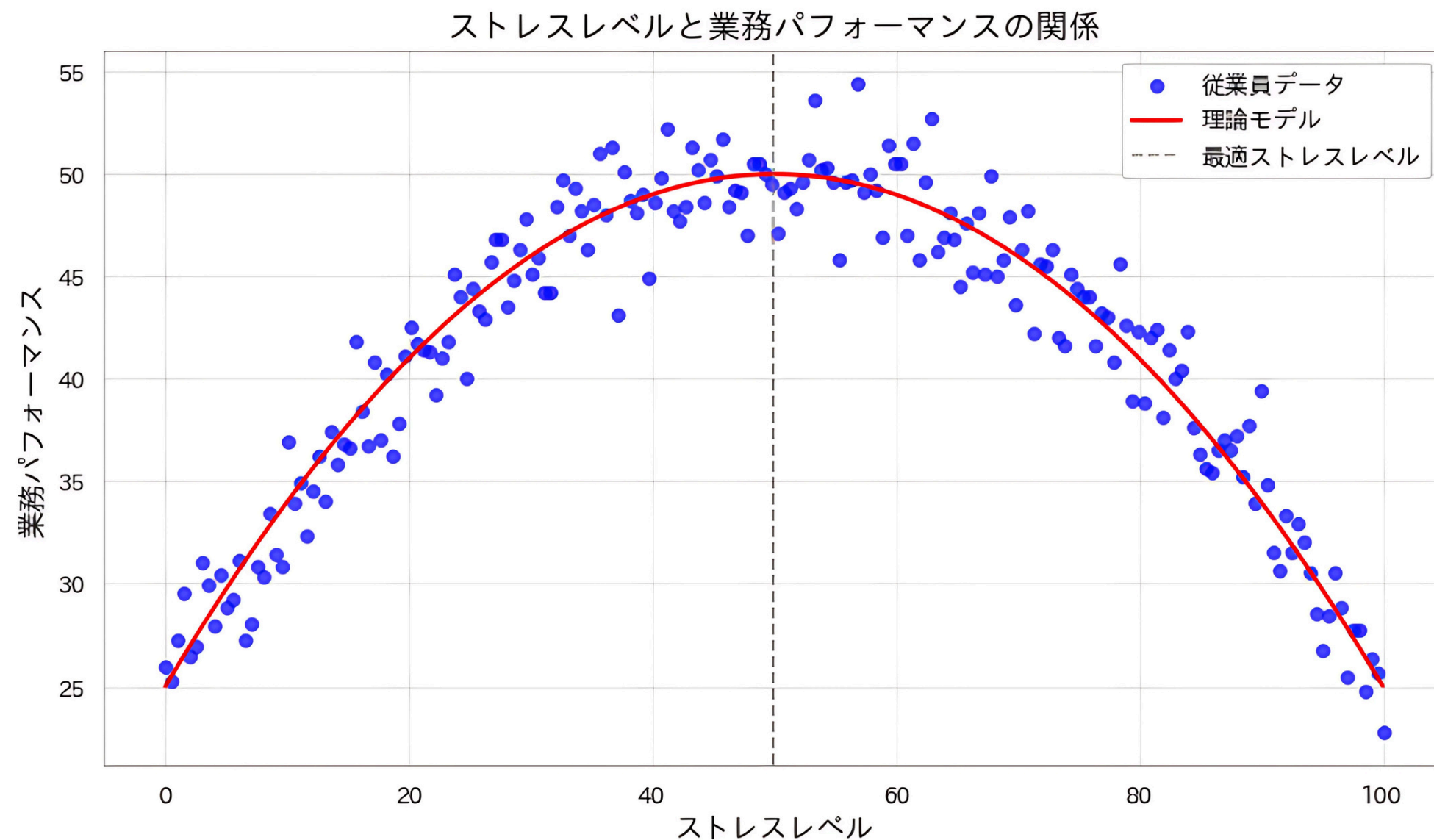
非線形関係を適切に捉えられない

相関係数は、直線的な関係を前提としているため、曲線的な関係（例: 放物線的關係）を示すデータでは、相関係数が0に近くなる場合があるが、実際には強い関係が存在することもある。

具体例

人事部が従業員のストレスレベルと業務パフォーマンスの関係を調査したところ、相関係数はほぼゼロだった。

しかし、詳細に分析すると、**ストレスが低すぎても高すぎてもパフォーマンスが低下する**という曲線的な関係が存在していた。



▶ 相関分析だけでは、変数間の関係性を捉えられない場合がある。

回帰分析 目的変数と説明変数の関係性

回帰直線

2つの変数の関係を 一次関数の形で数式化したもの。

単回帰

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

β_1 回帰係数 β_0 定数項 ϵ 誤差項

- **回帰係数**: 説明変数 x が1増えるとき、目的変数の条件付き平均が β_1 だけ変わることを表す。因果効果を意味するには別の仮定が必要。
- **定数項**: $x=0$ のときの目的変数の条件付き平均。

$$E[\epsilon \mid x] = 0, \quad E[Y \mid x] = \beta_0 + \beta_1 x$$

目的変数と説明変数

回帰分析では、ある変数（目的変数）を別の変数（説明変数）を用いて説明または予測する。

- **目的変数** (Y): 分析の対象となる変数。説明される側。
例: 売上、KPI、CV、予約数
- **説明変数** (X): 目的変数を説明・予測するために用いる変数。説明変数に含めただけでは因果関係は保証されない。
例: インプレッション、サイト回遊時間



ポイント

ここで扱う単回帰では、目的変数の条件付き平均を説明変数の一次関数で表す。
切片(定数項)と回帰係数を与えられたデータから推定する。

具体例

CM投下量(actual GRP)とKPI

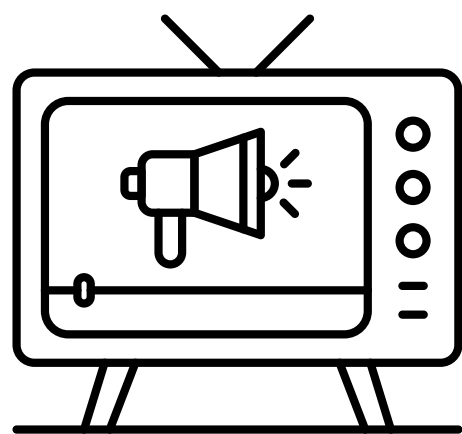
目的変数 (Y) : KPI(商材の売上)

説明変数 (X) : CM投下量(actual GRP)

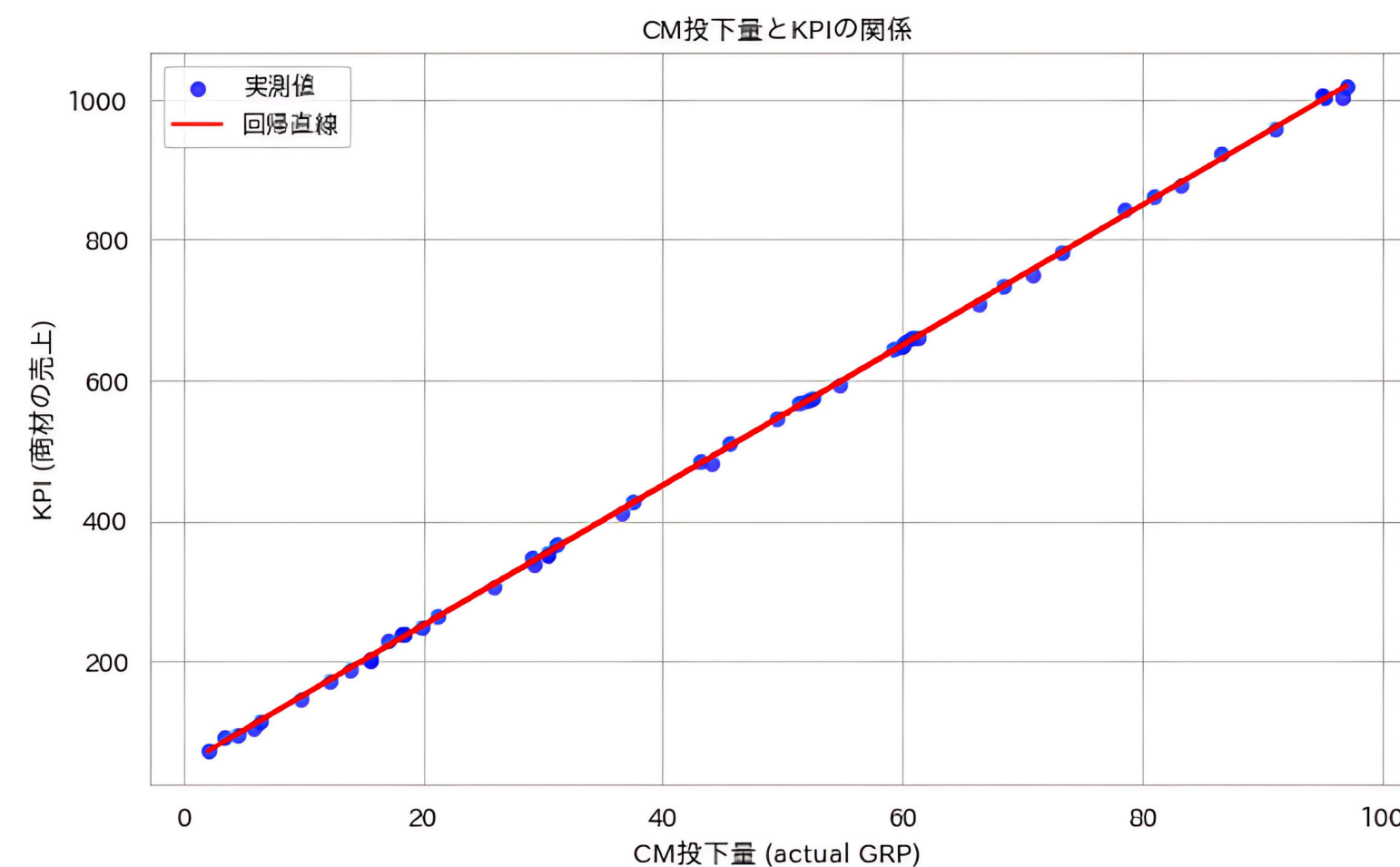
回帰分析を行い、次のような回帰式が得られたとする

$$\hat{y} = 50 + 10x$$

- CM投下量が1増えると、モデル上のKPIの予測値は10増える。
- x=0での予測値は50。ただし、観測範囲の外への予測には注意が必要。



実データだと、ここまで綺麗な直線にはならない



* この回帰結果だけでは、CMを増やすことによる売上の因果的な増加量は分からない。

単回帰の嬉しい点



解釈しやすさ

変数間の関係が直線的で理解しやすい



透明性の高さ

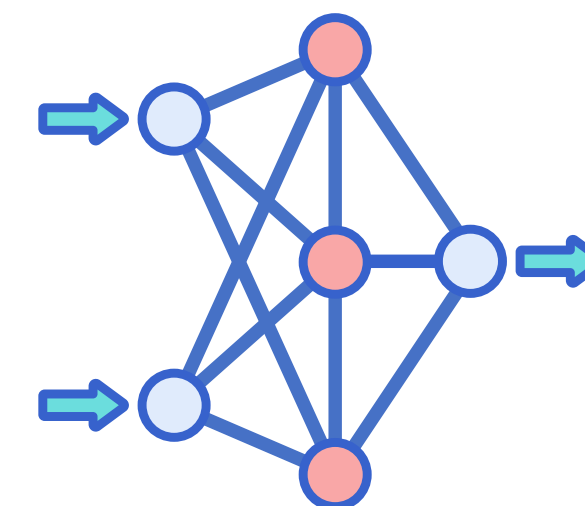
係数から、説明変数と目的変数の平均的な関連の方向と大きさを読み取れる。



計算の簡単さ

複雑なモデルに比べて、計算コストが低い

ビジネス現場においては、手法が高度だと説明が難しく、進めづらい



例：ニューラルネットワーク

非線形な関係を表現できる一方、説明方法や計算量に工夫が必要。

回帰分析

ベクトル表記と重回帰モデルの紹介

ベクトル表現 ベクトル表記を用いることで、説明変数が増えた場合でも数式を簡潔に記述できる。
下は1観測の場合。

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

特徴
ベクトル

$$\mathbf{x}_i = (1, x_i)^T$$

重み
ベクトル

$$\boldsymbol{\beta} = [\beta_0, \beta_1]^T$$

重回帰

説明変数が一個だと単回帰。説明変数が二個以上になると、重回帰と呼ぶ。
例えば、売上（目的変数）を「広告費」「販売員数」「季節」など**複数の要因（説明変数）**で説明できる。

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

決定係数

回帰モデルの当てはまりを表す指標の一つ。

以下は、切片を含む最小二乗回帰を、学習に使った同じデータで評価する場合の式。

この場合、 R^2 は0～1の範囲になる。

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

- 同じデータに対する最小二乗回帰では、説明変数を増やしても R^2 は下がる。
- 未使用データでの予測性能なども確認する。切片なしのモデルや未使用データでの評価では、 R^2 が負になる場合がある。

全変動(SST)

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

回帰変動と残差変動の和。

回帰変動(SSR)

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

回帰モデルによって、どれほどデータを説明できているか。

残差変動(SSE)

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

回帰モデルによって、どれほどデータを説明できていないか。

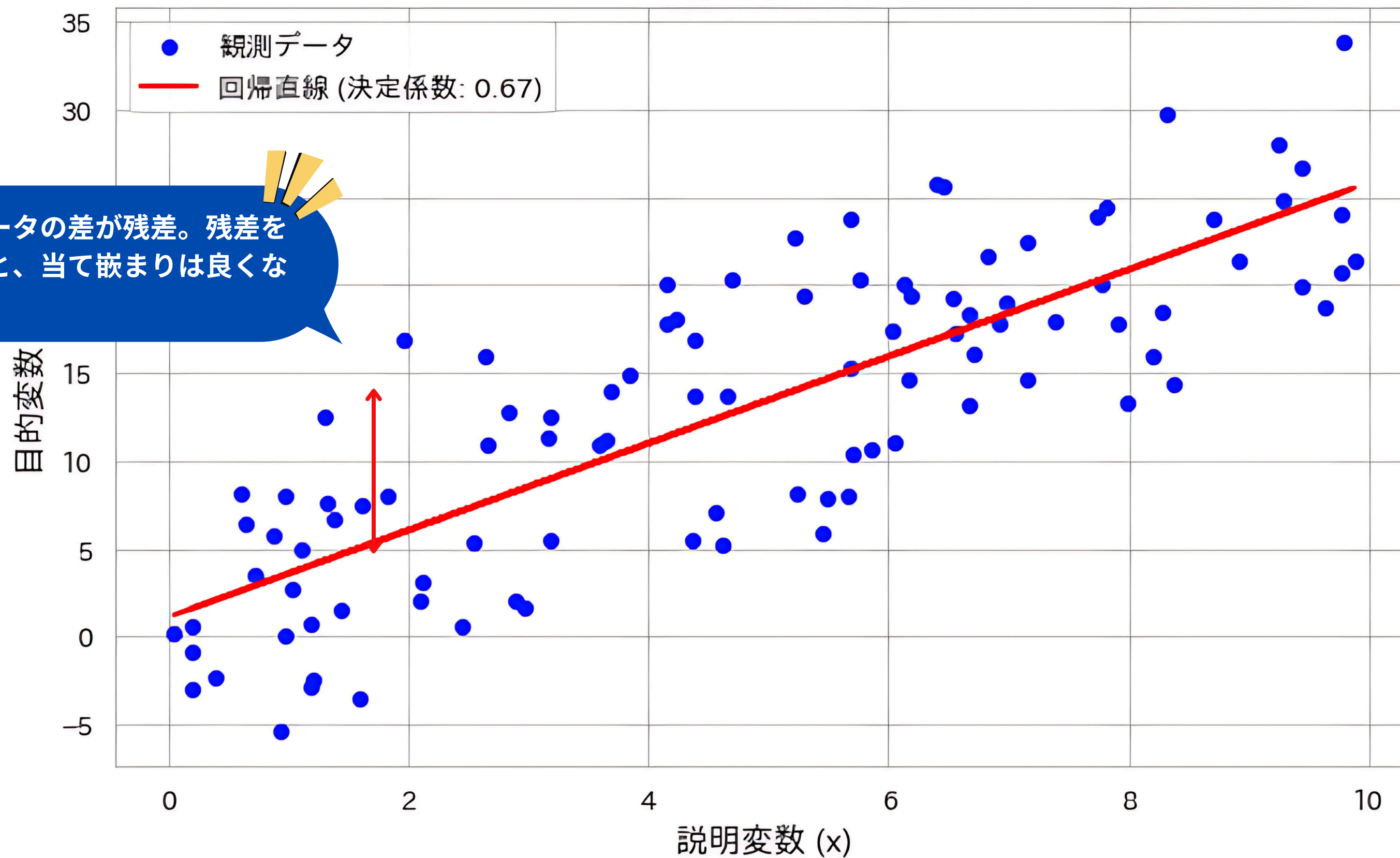


ポイント

決定係数は、**モデルの当て嵌まり具合を見る指標**。説明変数を増やせば、決定係数は上昇するので、モデルの良さを決定係数だけでは決めない。

説明変数を増えるとペナルティを与える「**自由度調整済み決定係数**」がモデル選択ではよく使われる

決定係数の例



実際の計算 回帰係数や定数項は、通常**最小二乗法(OLS)**を用いて解析的に算出する。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

OLSの方針 実データと予測値の誤差、つまり残差を最小化するようなパラメータを選ぶ。

①残差の平方和 (RSS) を最小化する

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

②パラメータごとに偏微分する

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i)$$

決定係数で紹介した「残差変動」を小さくすると、回帰変動で説明できる範囲が広がる、というイメージ。

偏微分をなぜ使うか

目的関数(残差平方和)の最小値を求めるために、傾きがゼロになる点を見つけるため。

OLSの方針 ③偏微分が0になる点を求める

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$
$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

これを解くと、

回帰係数

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

定数項

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

この目的関数は凸な二次関数。xがすべて同じ値でなければ、偏微分を0とする解が一意的な最小値を与える。

学んだ内容を問題で確認 この章に対応する演習問題を3問選びました。リンクまたはQRコードから挑戦できます。

問題 1

レベル2

相関係数

2変数のデータから相関係数を計算する。



▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 2

レベル2

回帰係数の推定

単回帰分析で、回帰直線の傾きを求める。



▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

問題 3

レベル2

決定係数

単回帰分析の決定係数を求める。



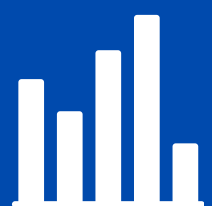
▶ 問題に挑戦する

クリック、またはQRコードを読み取ると問題が開きます

**ポイント**

この章のほかの問題にも、Web版「青の統計学」で挑戦できる。

統計検定2級・3級対策の演習問題と模擬試験を用意しています。 app.statisticsschool.com



1. 記述統計
2. データの散らばりの指標
3. 確率と確率分布
4. 相関と回帰
- 5. 統計的推測**

$$H_0$$

検定の基準として置く仮説。例えば、母平均が基準値に等しい、2群の母比率が等しい、などを表す。

検定の出発点となる仮説で、最初に疑うべき前提条件として採用される。

$$H_1$$

帰無仮説と対比して検討する仮説。差の有無や方向など、調べたい問いに応じて分析前に設定する。

通常、研究者が証明したい主張や予測に対応する。

▶ データが、帰無仮説および検定の前提とどの程度整合するかを調べ、帰無仮説を棄却するか判断する。

A stylized illustration of a laptop. The screen shows a code editor with a dark sidebar on the left containing several colored squares (red, orange, grey). The main area of the screen displays lines of code in various colors (red, orange, grey, blue). Above the laptop, there is a blue speech bubble containing the code symbols '</>'. To the right of the speech bubble are two interlocking gears, one red and one grey. The entire scene is set against a yellow circular background.

CVR(サイト訪問者の予約率など)向上のために、サイトのUIを変更したい。
A/Bテストにより、現行デザインと新デザインでCVRを比較する。

$$H_0 : p_{\text{new}} \leq p_{\text{old}},$$

新しいUIデザインは、CVR向上に寄与しない

$$H_1 : p_{\text{new}} > p_{\text{old}}$$

新しいUIデザインは、CVR向上に寄与する

右片側検定では、帰無仮説の境界 $p_{\text{新}} = p_{\text{旧}}$ を用いて検定統計量を計算する。

有意水準と棄却域

帰無仮説が正しい場合に、誤ってそれを棄却する確率の上限

通常は5% ($\alpha = 0.05$) や1% ($\alpha = 0.01$) が選ばれる。

帰無仮説を棄却するとあらかじめ定めた、検定統計量の値の範囲。
右側・左側・両側のどこに置くかは、対立仮説によって決まる。

例えば、標準正規分布を用いた検定では、有意水準が5%の場合、棄却域は分布の両端に位置し、それぞれ2.5%の範囲を占める。

両側5%検定では、上側臨界値 $z_{0.025} \approx 1.96$ を用い、 $Z < -1.96$ または $Z > 1.96$ を棄却域とする。

同じ有意水準なら、片側検定は事前に指定した方向の差を検出しやすい。その反対方向の差は検出対象にしない。**データを見てから片側・両側を選ばない。**

$$|\mathbf{Z}| > z_{\alpha/2}$$

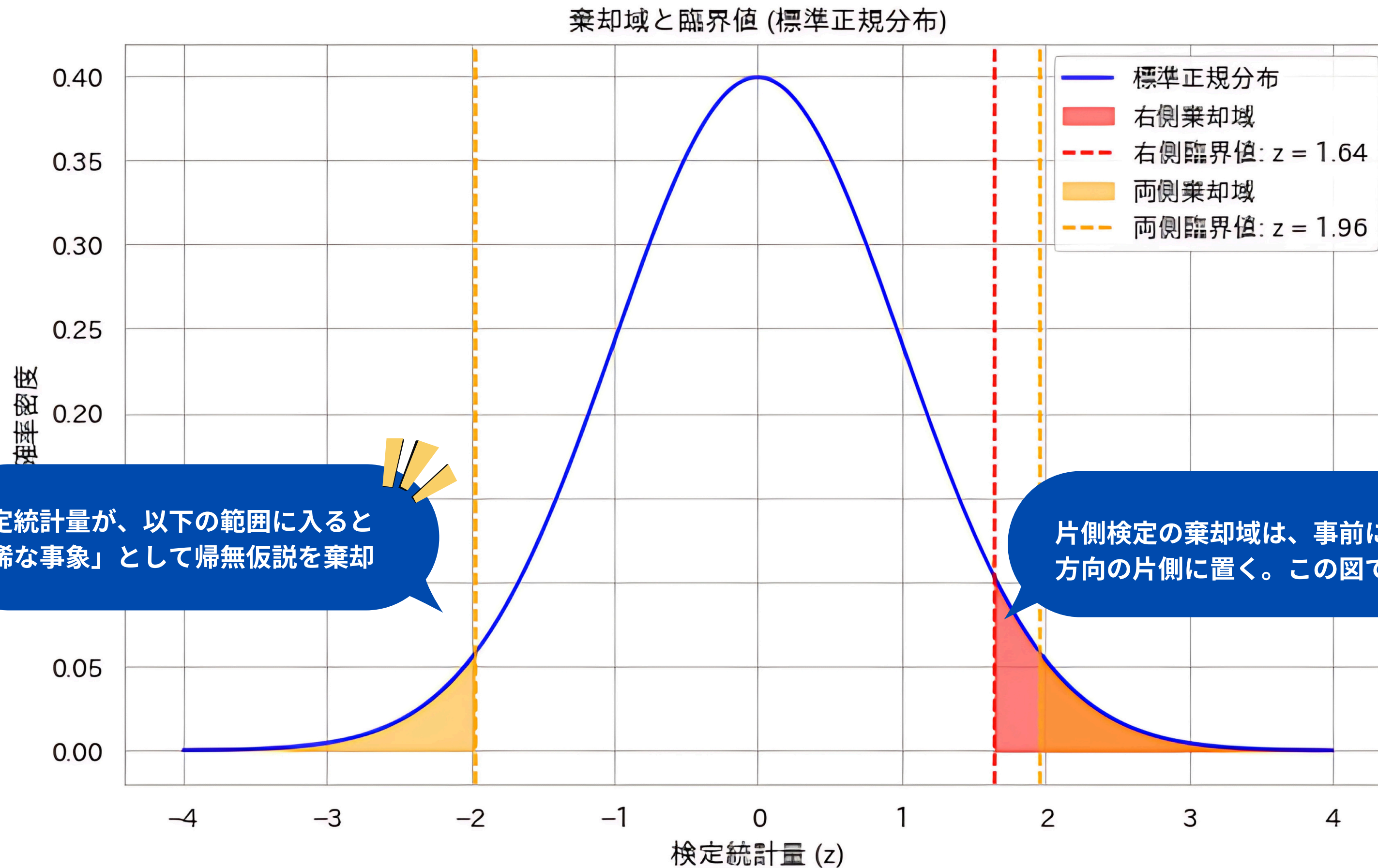
効果の方向が不明。どちらの方向に対しても関心がある場合。

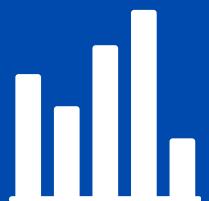
$$Z > z_\alpha$$

調べる方向が分析前に定まっており、反対方向の差を棄却対象にしない場合。



前ページの「サイトUI変更」：「現行デザインよりCVRが向上する」仮説なので、片側検定になる。





仮説検定の基礎

p値の意味と解釈方法

p値

帰無仮説と検定の前提の下で、観測された検定統計量と同じか、それ以上に極端な値が得られる確率。

解釈方法

例として、有意水準は5%としている。
この時の、p値と0.05の大小関係における解釈は以下になる。

$$p\text{ value} < 0.05$$

p値が0.05未満なら、有意水準5%で帰無仮説を棄却する。帰無仮説が誤っている確率を表しているわけではない



帰無仮説を棄却する

$$p\text{ value} \geq 0.05$$

p値が0.05以上なら、有意水準5%では帰無仮説を棄却しない。
帰無仮説が正しいことや、実質的な差がないことが確認されたわけではない。



帰無仮説を棄却する十分な証拠が得られていない



ポイント

p値は、帰無仮説とデータの整合性を評価する材料の一つ。
効果の大きさ・推定の精度・調査設計も合わせて判断する。

p値の注意点



帰無仮説が真であるという前提のもとで計算

p値が小さいことは帰無仮説が誤りであることを直接的に示すものではない



効果の大きさを表すものではない

p値が小さいからといって、効果が大きいとは限らない。



サンプルサイズに依存

サンプルサイズが大きいほど、小さな効果でも統計的に有意になる可能性が高くなる。

▶ 独立な観測が増え、ばらつきが同程度なら、
サンプルを増やすほど標準誤差が小さくなるから（サンプルごとの平均値のばらつきが小さい）

効果量 effect size

群間差や変数間の関連の大きさを表す指標。因果効果として解釈できるかどうかは、研究・実験の設計と仮定による。

よくある誤解

p値：帰無仮説と検定の前提の下で、観測値と同じか、それ以上に極端な検定統計量が得られる確率。ここでは、大きい値ほど対立仮説を支持する右片側検定を考える。

$$p\text{値} = P_{H_0}(T \geq t_{\text{obs}})$$

t_obsは実際に観測された検定統計量。連続分布では、上式の $\geq$ を $>$ としても値は同じ。



混同しやすいのが、 α エラー

α エラー：「本当は差がないのにも関わらず、差があるとする確率」

第1種の誤り：帰無仮説が真なのに、それを棄却すること。有意水準 α は、その確率を制御する上限として設定する。

①帰無仮説 H_0 と対立仮説 H_1 を設定



②有意水準 (α) を決定

通常 $\alpha = 0.05$ や $\alpha = 0.01$ を設定



③検定統計量を計算

検定統計量は、データから計算される

例：
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$



④棄却域に入るか判定する

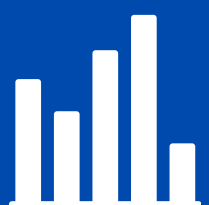
- 検定統計量が事前に定めた棄却域に入れば、帰無仮説を棄却する。
- 入らなければ、帰無仮説を棄却しない。



⑤結論を導出

- 検定結果とともに、差の推定値や信頼区間も報告する。
- 棄却しない場合も「差がないと証明された」とは解釈しない。

▶ どう作るのか？



仮説検定の基礎

検定統計量の定義

青の統計学

検定統計量

サンプルデータに基づいて計算される値であり、母集団のパラメータに関する仮説を評価するために用いられる

(例)母平均の推定

$$T = \frac{\bar{x} - \mu_0}{SE}$$

差が大きいか小さいかの基準は、データのばらつき（分散や標準偏差）に依存するため、**分子の差だけでは判断できない。**

▶ 分母の標準誤差でスケールする

分子

$$\bar{x} - \mu_0$$

標本平均と、帰無仮説で置いた母平均 μ_0 との差を表す。差の大きさは、その推定に伴うばらつきと比較して評価する。

分母

$$SE = \frac{\sigma}{\sqrt{n}}$$

標準誤差は、同じ条件で標本を取り直したときの推定量の標準偏差。**独立・同一分布の標本平均では $\sigma/\sqrt{n}$ となる。 σ が未知なら $s/\sqrt{n}$ で推定する。**



ポイント

検定統計量は、帰無仮説からのずれを標準誤差で割って評価する。
同じ差でも、標本サイズやデータのばらつきにより結果は変わる。

- 母標準偏差 $\sigma = 20$ の母集団から、独立に n 個のデータを取る。
- 同じ方法で標本を取り直したとき、標本平均の標準偏差は $20/\sqrt{n}$ となる。

サンプリングごとに、平均等のずれは生じる。サンプル数が増えると誤差は減っていく。

1回の標本サイズ n	標本平均の標準誤差 $20/\sqrt{n}$
25	4
100	2
400	1

標本サイズを4倍にすると、標本平均の標準誤差は半分になる。「1回の標本サイズ n 」と「同じ実験を繰り返す回数」は別のもの。

点推定

特徴

- 母集団パラメータを単一の値で推定する
- 不確実性の情報はない

例

標本平均を用いて、母平均を推定する

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

区間推定

特徴

- 推定値に信頼区間という形で幅を持たせる
- サンプリングを繰り返した場合の信頼区間のブレを確率的に表現。

例

母平均の95%信頼区間の例

(独立な正規標本・母標準偏差 σ が既知)

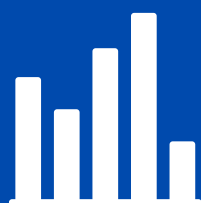
$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

95%信頼区間では $\alpha=0.05$ 、上側 $\alpha/2$ 点は $z_{0.025} \approx 1.96$ 。



ポイント

点推定は、単一の値で母集団のパラメータを表し、簡潔だが不確実性を示さない。
区間推定は、推定量に一定の幅を持たせ、**不確実性を信頼区間という形で表現する。**



標準正規分布表

標準正規分布においてその値以上の値を取る確率を表している

1.96以上の値を取る確率を知りたい場合は、左見出しにある1.9と上見出しにある0.06の交差点見ると、0.025と分かる。

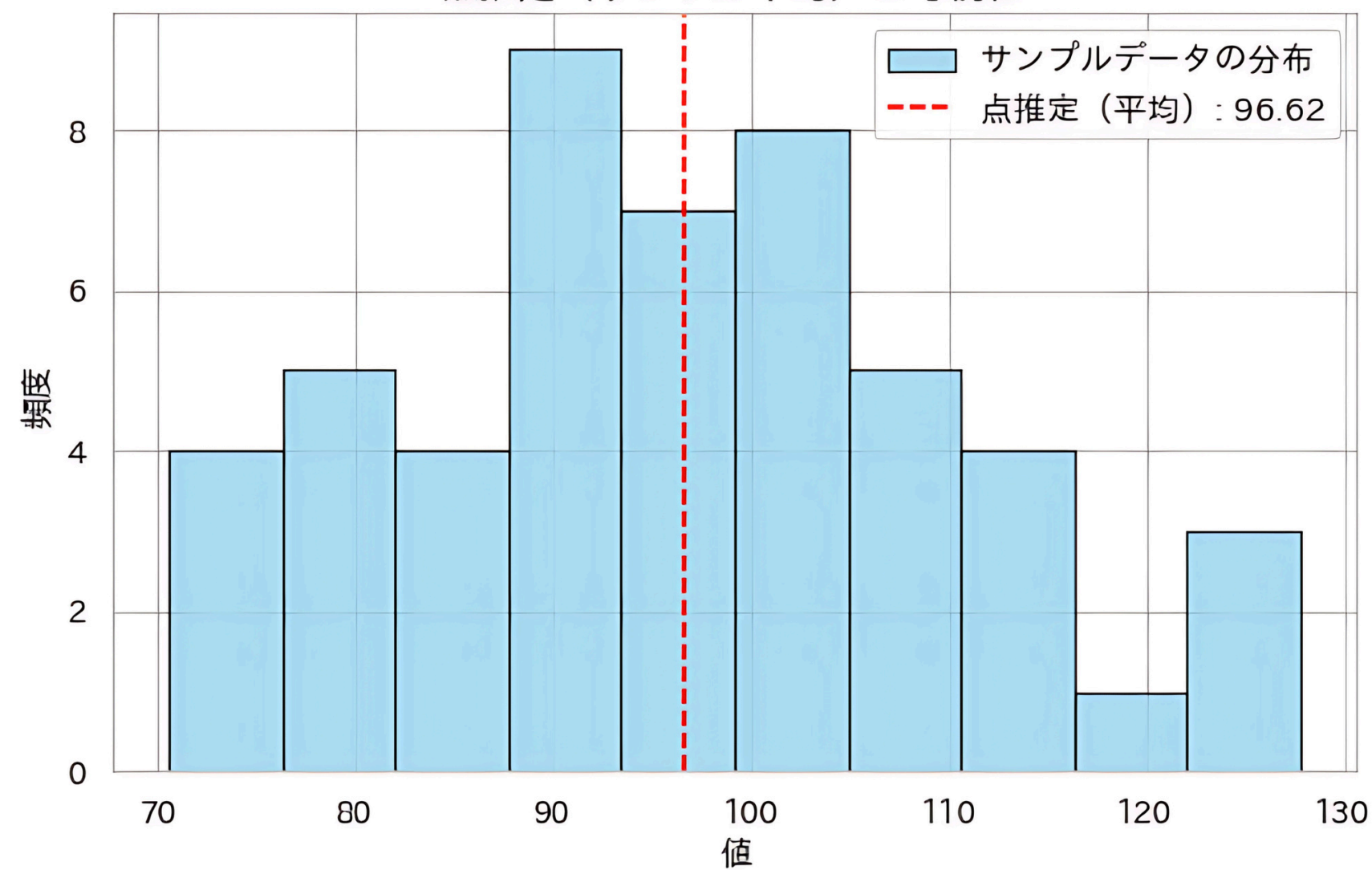
標準正規分布表 (左側累積確率)

小数第2位	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
z（整数部分+小数第1位）										
0.0	0.500000	0.496011	0.492022	0.488034	0.484047	0.480061	0.476078	0.472097	0.468119	0.464144
0.1	0.460172	0.456205	0.452242	0.448283	0.444330	0.440382	0.436441	0.432505	0.428576	0.424655
0.2	0.420740	0.416834	0.412936	0.409046	0.405165	0.401294	0.397432	0.393580	0.389739	0.385908
0.3	0.382089	0.378280	0.374484	0.370700	0.366928	0.363169	0.359424	0.355691	0.351973	0.348268
0.4	0.344578	0.340903	0.337243	0.333598	0.329969	0.326355	0.322758	0.319178	0.315614	0.312067
0.5	0.308538	0.305026	0.301532	0.298056	0.294599	0.291160	0.287740	0.284339	0.280957	0.277595
0.6	0.274253	0.270931	0.267629	0.264347	0.261086	0.257846	0.254627	0.251429	0.248252	0.245097
0.7	0.241964	0.238852	0.235762	0.232695	0.229650	0.226627	0.223627	0.220650	0.217695	0.214764
0.8	0.211855	0.208970	0.206108	0.203269	0.200454	0.197663	0.194895	0.192150	0.189430	0.186733
0.9	0.184060	0.181411	0.178786	0.176186	0.173609	0.171056	0.168528	0.166023	0.163543	0.161087
1.0	0.158655	0.156248	0.153864	0.151505	0.149170	0.146859	0.144572	0.142310	0.140071	0.137857
1.1	0.135666	0.133500	0.131357	0.129238	0.127143	0.125072	0.123024	0.121000	0.119000	0.117023
1.2	0.115070	0.113139	0.111232	0.109349	0.107488	0.105650	0.103835	0.102042	0.100273	0.098525
1.3	0.096800	0.095098	0.093418	0.091759	0.090123	0.088508	0.086915	0.085343	0.083793	0.082264
1.4	0.080757	0.079270	0.077804	0.076359	0.074934	0.073529	0.072145	0.070781	0.069437	0.068112
1.5	0.066807	0.065522	0.064255	0.063008	0.061780	0.060571	0.059380	0.058208	0.057053	0.055917
1.6	0.054799	0.053699	0.052616	0.051551	0.050503	0.049471	0.048457	0.047460	0.046479	0.045514
1.7	0.044565	0.043633	0.042716	0.041815	0.040930	0.040059	0.039204	0.038364	0.037538	0.036727
1.8	0.035930	0.035148	0.034380	0.033625	0.032884	0.032157	0.031443	0.030742	0.030054	0.029379
1.9	0.028717	0.028067	0.027429	0.026803	0.026190	0.025588	0.024998	0.024419	0.023852	0.023295
2.0	0.022750	0.022216	0.021692	0.021178	0.020675	0.020182	0.019699	0.019226	0.018763	0.018309
2.1	0.017864	0.017429	0.017003	0.016586	0.016177	0.015778	0.015386	0.015003	0.014629	0.014262
2.2	0.013903	0.013553	0.013209	0.012874	0.012545	0.012224	0.011911	0.011604	0.011304	0.011011
2.3	0.010724	0.010444	0.010170	0.009903	0.009642	0.009387	0.009137	0.008894	0.008656	0.008424
2.4	0.008198	0.007976	0.007760	0.007549	0.007344	0.007143	0.006947	0.006756	0.006569	0.006387
2.5	0.006210	0.006037	0.005868	0.005703	0.005543	0.005386	0.005234	0.005085	0.004940	0.004799
2.6	0.004661	0.004527	0.004396	0.004269	0.004145	0.004025	0.003907	0.003793	0.003681	0.003573
2.7	0.003467	0.003364	0.003264	0.003167	0.003072	0.002980	0.002890	0.002803	0.002718	0.002635
2.8	0.002555	0.002477	0.002401	0.002327	0.002256	0.002186	0.002118	0.002052	0.001988	0.001926
2.9	0.001866	0.001807	0.001750	0.001695	0.001641	0.001589	0.001538	0.001489	0.001441	0.001395
3.0	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
3.1	0.000968	0.000935	0.000904	0.000874	0.000845	0.000816	0.000789	0.000762	0.000736	0.000711
3.2	0.000687	0.000664	0.000641	0.000619	0.000598	0.000577	0.000557	0.000538	0.000519	0.000501
3.3	0.000483	0.000466	0.000450	0.000434	0.000419	0.000404	0.000390	0.000376	0.000362	0.000349
3.4	0.000337	0.000325	0.000313	0.000302	0.000291	0.000280	0.000270	0.000260	0.000251	0.000242
3.5	0.000233	0.000224	0.000216	0.000208	0.000200	0.000193	0.000185	0.000178	0.000172	0.000165

大学の試験問題や統計検定などでは、付表として見れるので覚える必要などはない。

点推定

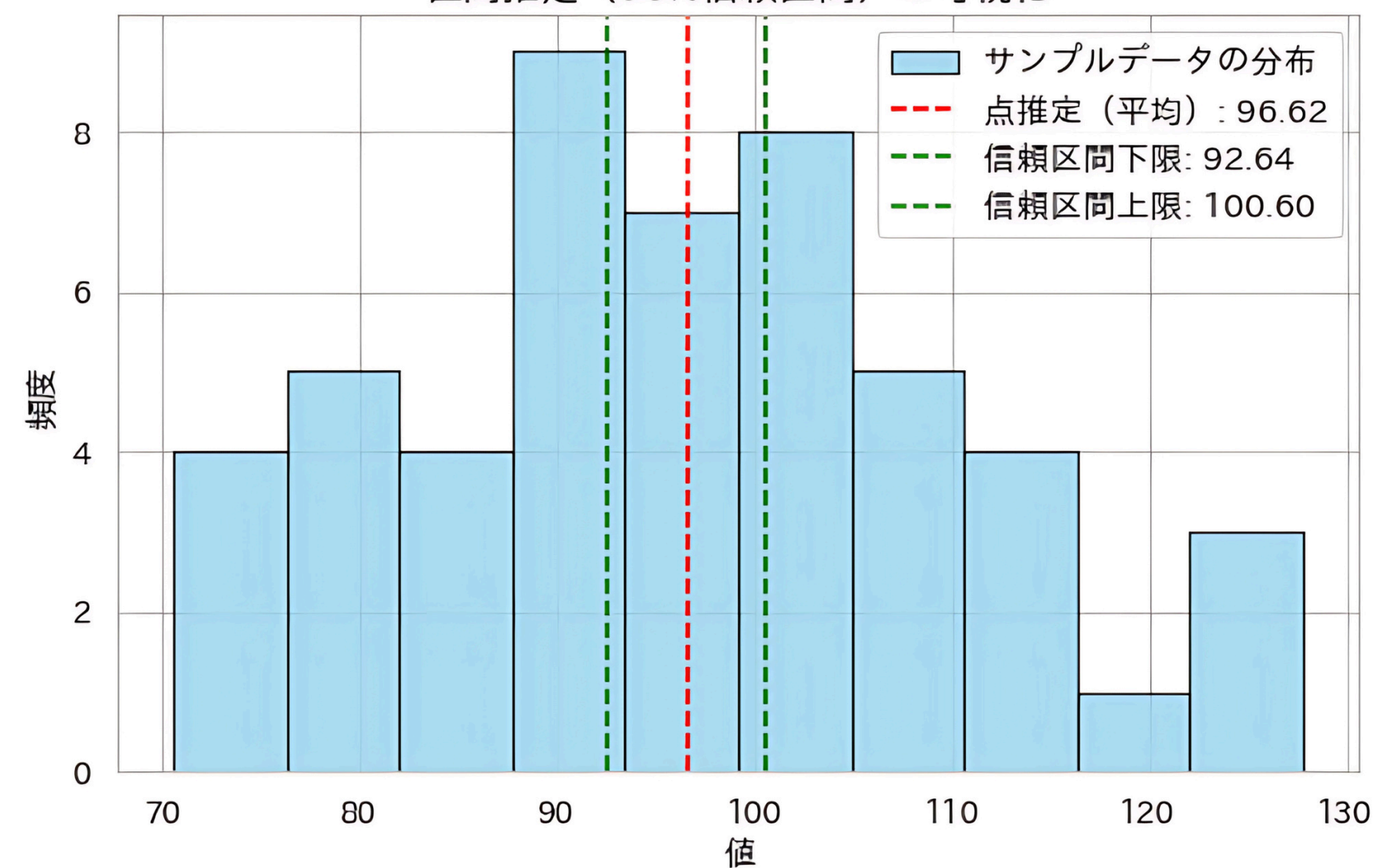
点推定（サンプル平均）の可視化



区間推定

施策リリース後の切り戻しラインなどの設定には、区間を伴った形で検証設計に導入されることが多い。

区間推定（95%信頼区間）の可視化



母集団のパラメータを推定するために、標本から点推定量を計算し、推定量の周りの範囲を提供する

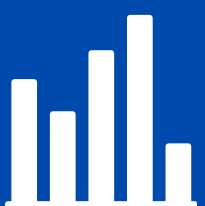
来場者全体から無作為に選んだn人について、満足度の回答が得られたとする。来場者全体の平均満足度を推定したい。

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$$

ここでは正規母集団を仮定し、母分散は未知とする。
不偏分散から求めた標本標準偏差sを用いると、母平均の95%信頼区間は次のようになる。

$$\bar{x} \pm t_{0.025, n-1} \frac{s}{\sqrt{n}}$$

満足度が段階評価の場合、正規性はモデル上の近似。任意回答のアンケートでは回答者の偏りにも注意する。大標本では t の臨界値は約1.96に近づく。



前提

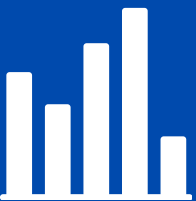
信頼区間は、実験を何度も繰り返すという仮想的状況に基づいて解釈されるものであり、毎回のサンプリングで異なる標本が得られる

▶ あくまで確率の対象となるのは、「信頼区間」であり、「母集団パラメータ」ではない。

パラメータが不確実性をもつのは、ベイズ統計の考え方。信用区間での解釈になる。

✗ 「今回計算して確定した区間について、母平均がその中にある確率を95%とする。」

✓ 「同じ条件で標本抽出と区間計算を繰り返すと、作られる区間の約95%が固定された真の母平均を含む。」



母平均の推定

パターン分類

- ポイント**
- いずれも、分布の臨界値と標準誤差をかけて信頼区間を算出するという構造は同じ。
 - 点推定値には変わりはないが、「分散がわからない」「等分散の仮定を置けない」という場合だと、**信頼区間が変わる**。

ケース	分散	点推定値	区間推定	使用する分布
母平均の推定	既知	$\overline{x}$	$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	標準正規分布
	未知	$\overline{x}$	$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$	t分布
母平均の差の推定	既知	$\overline{x}_1 - \overline{x}_2$	$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	標準正規分布
	未知（等分散）	$\overline{x}_1 - \overline{x}_2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	t分布
	未知（異分散）	$\overline{x}_1 - \overline{x}_2$	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t分布

各群内の独立な正規標本を仮定する。2群の比較では群間も独立とする。

(A blue arrow points from the text "vはWelch-Satterthwaiteの近似自由度。p.85参照" to the symbol ν in the formula of the last row of the table.)

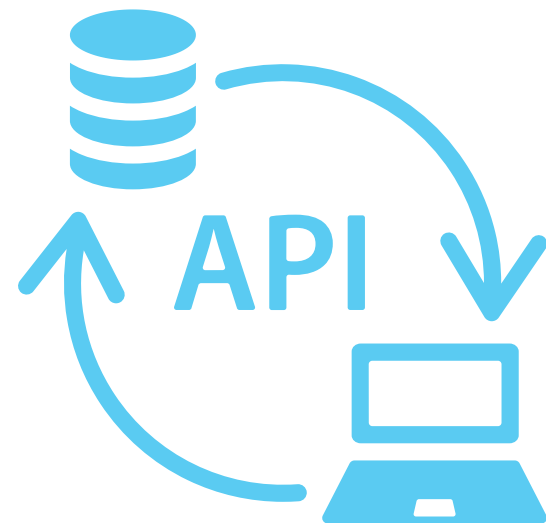
vはWelch-Satterthwaiteの近似自由度。p.85参照

母平均の推定

具体例

具体例

WebサービスのAPI応答時間を監視しています。サービスレベル目標は、APIの平均応答時間を 200ms以下 にすること。全リクエスト（母集団）の応答時間を測定するのは非現実的なため、サンプルデータを基に母平均を推定することになった。



データ

- サンプルサイズは、**n=50**
- サンプルの平均応答時間は、**210ms**
- サンプルの標準偏差は**20ms**
- 信頼水準**95%**

応答時間は独立な標本として扱い、**t区間**による近似が妥当と仮定する。

信頼区間を用いた母平均の推定

$$\bar{x} \pm t_{0.025, 49} \frac{s}{\sqrt{50}}$$

今回、母分散がわからずサンプルから標準偏差を使っているためt分布を使う。

自由度は、50-1より49。95%信頼区間の場合、t分布の値はおよそ2.01

$$210 \pm 2.01 \cdot \frac{20}{\sqrt{50}}$$

95%信頼区間は約[204.32 ms, 215.68 ms]。

この仮定の下では、母平均応答時間が目標の200 msを上回ることを示す結果であり、性能改善を検討する。

検定の目的 母集団の平均値が特定の値と異なるかどうか

仮説(例)

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0 \text{ (両側検定)} \quad \mu < \mu_0 \text{ (左側検定)} \quad \text{または} \quad \mu > \mu_0 \text{ (右側検定)}$$

検定統計量

母分散が検定統計量に使えるかどうかにより、統計量と従う分布が変わる。

分散既知の場合

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)$$

分布の形状は固定で、標準偏差が1、平均が0となる。

分散未知の場合

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t_{n-1}$$

母標準偏差 σ を S で推定することによる不確実性を、自由度 $n-1$ の t 分布で考慮する。

母分散がわかっている場合はほぼ無いので、 t 検定を使うことが現場では多い。

いずれも、独立な正規標本で帰無仮説が真のときの分布。 S は分母 $n-1$ の標本標準偏差を表す。観測後の値は小文字 s で表す。



ポイント

- 母分散が既知 → 標準正規分布 (Z分布)
- 母分散が未知 → t 分布 (自由度に依存)
- 正規性の仮定を外す場合は、**大標本近似の妥当性を確認する**。

母平均の検定

t分布

t分布

正規母集団の平均を、母分散が未知の状態では推定・検定するときなどに用いる分布。小標本専用ではなく、自由度が大きくなると標準正規分布に近づく。

F分布やカイ二乗分布と同様に、「標本に関する分布」と呼ばれる。

$$f(t; v) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi} \cdot \Gamma(\frac{v}{2})} \left(1 + \frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

自由度 ν

データからパラメータを推定した後に残る、独立に変動できる情報の数。計算式は、モデルと推定するパラメータによって異なる。

対応のあるt検定では、 n 組の差を1標本として扱い、自由度は $n-1$ 。

1標本t検定の自由度

$$\nu = n - 1$$

1つの標本平均を推定するため、偏差には和が0という制約が1つ生じる。

独立2標本・等分散のt検定

$$\nu = n_1 + n_2 - 2$$

各群で平均を推定する。Welch検定の自由度は別式で求める。

切片あり重回帰の残差自由度

$$\nu = n - (k + 1)$$

k は説明変数の数。切片を含む $k+1$ 個の係数を推定。計画行列がフルランクの場合。

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

ガンマ関数。階乗を一般化した関数。自然数を代入すると、以下になる。

$$\Gamma(n) = (n - 1)!$$

母平均の検定

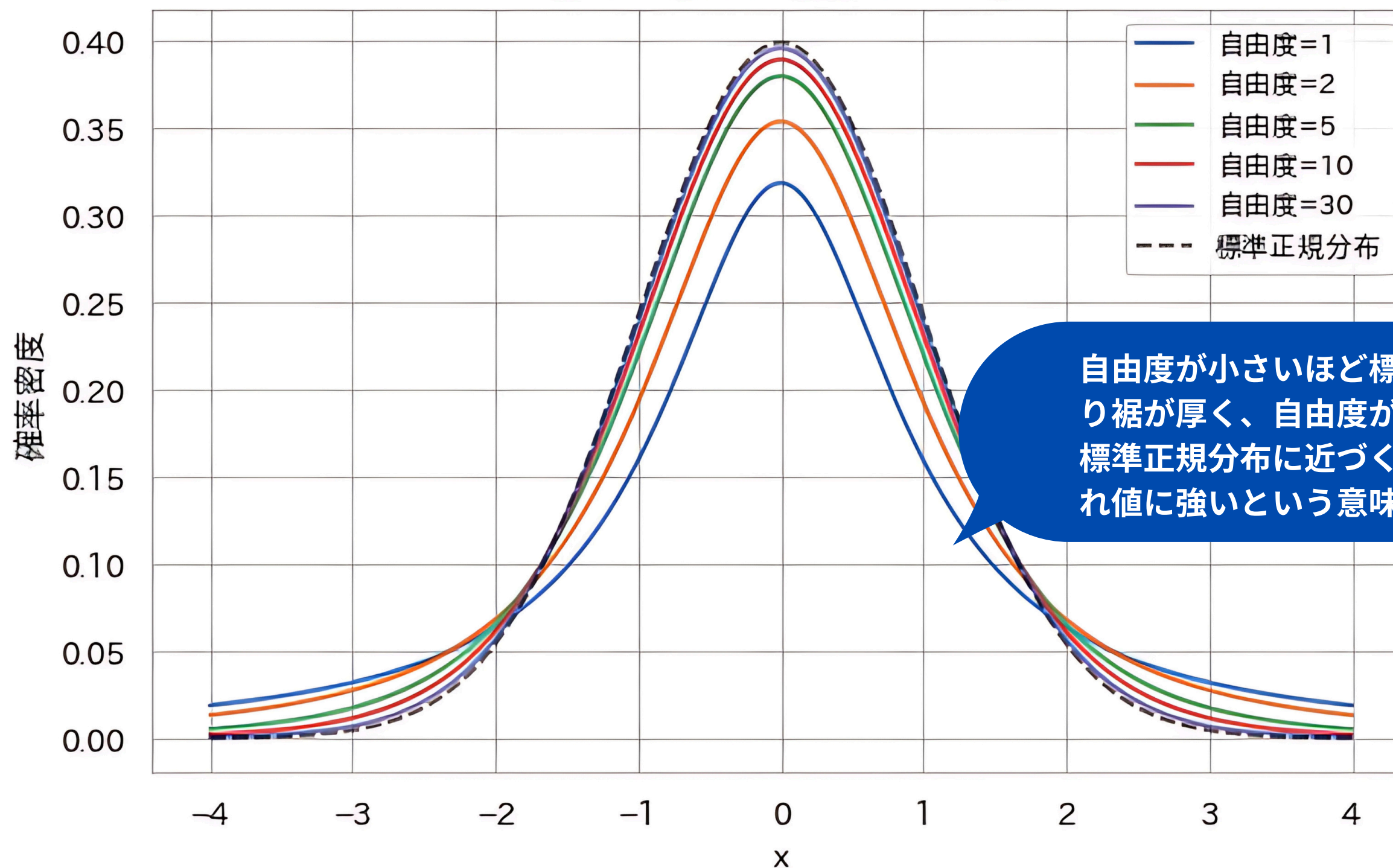
t分布

なぜt分布を使うのか

母標準偏差 σ が未知なので、標本標準偏差 S で推定する。

その推定の不確実性も含めると、正規標本の標準化した平均は標準正規分布ではなくt分布に従う。

t分布と標準正規分布の比較



t分布の上側確率aに対応する臨界値 $t_{a,v}$ 自由度vのt分布について、 $P(T > t_{a,v}) = a$ となる値を示す。
95%の両側信頼区間では、 $a = 0.025$ の列を使う

t分布表（自由度1～30）

自由度	片側 $\alpha=0.1$, 両側 $\alpha=0.2$	片側 $\alpha=0.05$, 両側 $\alpha=0.1$	片側 $\alpha=0.01$, 両側 $\alpha=0.02$
1	3.077684	6.313752	31.820516
2	1.885618	2.919986	6.964557
3	1.637744	2.353363	4.540703
4	1.533206	2.131847	3.746947
5	1.475884	2.015048	3.364930
6	1.439756	1.943180	3.142668
7	1.414924	1.894579	2.997952
8	1.396815	1.859548	2.896459
9	1.383029	1.833113	2.821438
10	1.372184	1.812461	2.763769
11	1.363430	1.795885	2.718079
12	1.356217	1.782288	2.680998
13	1.350171	1.770933	2.650309
14	1.345030	1.761310	2.624494
15	1.340606	1.753050	2.602480
16	1.336757	1.745884	2.583487
17	1.333379	1.739607	2.566934
18	1.330391	1.734064	2.552380
19	1.327728	1.729133	2.539483
20	1.325341	1.724718	2.527977
21	1.323188	1.720743	2.517648
22	1.321237	1.717144	2.508325
23	1.319460	1.713872	2.499867
24	1.317836	1.710882	2.492159
25	1.316345	1.708141	2.485107
26	1.314972	1.705618	2.478630
27	1.313703	1.703288	2.472660
28	1.312527	1.701131	2.467140
29	1.311434	1.699127	2.462021
30	1.310415	1.697261	2.457262

大学の試験問題や統計検定などでは、付表として見れるので覚える必要などはない。

母平均の差の検定

検定の目的と検定統計量の計算方法

検定の目的 2つの独立な母集団の平均値の差 $\mu_1 - \mu_2$ に統計的な差があるかを判断

仮説(例) $H_0 : \mu_1 - \mu_2 = 0$ (差がない)

$H_1 : \mu_1 - \mu_2 \neq 0$ (差がある)

検定統計量の計算

母平均の差の検定においては、母分散が既知か未知かに加えて、**母集団の分散が等しいかどうか**も検定統計量を決定する分岐になる。各群が独立な正規標本で、群間も独立、かつ H_0 が真のときの分布。右のWelch検定はt分布による近似。vの計算はp.85参照。

分散既知の場合

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

差の差をとり、標準誤差で割る。

分散未知 & 等分散の場合

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

プールした分散を使い、通常のt検定を行う。各群の平均を推定するため、自由度は $(n_1-1) + (n_2-1)$ 。

分散未知 & 異分散の場合

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \overset{\text{approx}}{\sim} t_v$$

ウェルチのt検定と呼ばれる。**プールした分散を使わずに、個別の標本分散を使う。**



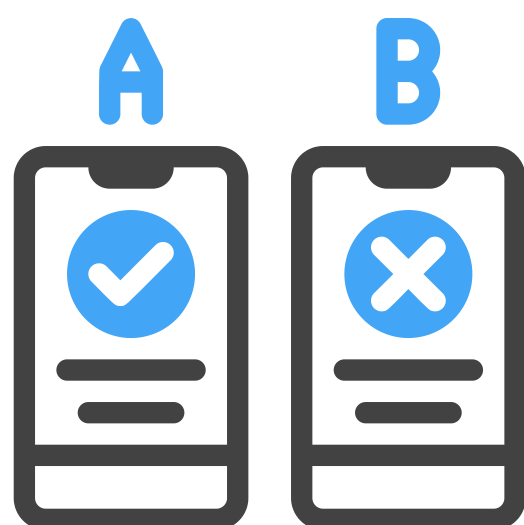
ポイント

- 等分散の場合：プールされた標準偏差を使い、標準t検定を行う。
- 異分散の場合：個別の標準偏差を使い、ウェルチのt検定を行う。

具体例

あるWebサービスのUX改善のため、新しい検索機能（新UI）を導入した。この機能が、従来の検索機能（旧UI）より平均応答時間を短縮できるかを評価したい。新旧UIでの応答時間の母平均に差があるかを有意水準を5%として検定し、効果を統計的に判断する。

旧UIと新UIの観測は独立で、比較可能な条件で収集したものとする。ここでは各群の正規性と等分散を仮定する。



データ

	旧UI (A)	新UI (B)
サンプルサイズ	30	30
標本平均(ms)	220	200
標本標準偏差(ms)	25	20

仮説設定

帰無仮説

$$H_0 : \mu_B \geq \mu_A,$$

新UIの母平均応答時間は、旧UI以上
検定統計量は境界 $\mu_B = \mu_A$ で計算する

対立仮説

$$H_1 : \mu_B < \mu_A$$

新UIの平均応答時間が旧UIよりも短い

▶ 片側検定を行う

検定統計量の計算

両群の母分散は等しいと仮定し、t検定を行う。

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_p \sqrt{1/n_A + 1/n_B}}$$

プールした標準偏差は、

$$s_p = \sqrt{\frac{(30-1)25^2 + (30-1)20^2}{30+30-2}} = \sqrt{512.5} \approx 22.64$$

よって、検定統計量は

$$t = \frac{200 - 220}{\sqrt{512.5} \sqrt{1/30 + 1/30}} \approx -3.42$$

- ▶ 自由度は58。左片側5%検定の臨界値は $-t_{0.05,58} \approx -1.672$ 。
 $t \approx -3.42$ は臨界値より小さいため、帰無仮説を棄却する。左片側p値は約0.00057。

結論 この仮定の下では、**新UIの母平均応答時間が旧UIより短いことを示す統計的証拠が得られた**。標本平均の差は-20 ms。
導入判断では、短縮幅の実務上の価値や、安定性・導入コストも合わせて評価する。

等分散の仮定を置かない場合は、
Welchのt検定という別の
検定となり、検定統計量も変わる

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

プールした分散

各群の不偏分散を、自由度 n_a-1 、 n_b-1 で重み付けして平均する。

母平均の差の検定

異分散の場合

ウェルチのt検定

母分散が未知の独立な2群について、**等分散を仮定せずに母平均の差を検定する方法**。
各群の分散と標本サイズを用いて、近似自由度を計算する。

等分散の仮定を置いている時とは異なり、サンプル数だけでは自由度は決められない、ということ。

自由度

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

分子

$$\hat{V} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}, \quad \text{分子} = \hat{V}^2$$

$\hat{V}$ は標本平均の差の分散の推定値。自由度の式の分子には、この二乗を使う。

分母

標本サイズや分散の違いに基づく自由度の調整項を計算

検定統計量

プールした分散は使わず、それぞれの標本分散を検定統計量に使う。

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \quad H_0, \underset{\sim}{\text{approx}} t_\nu$$

▶ $H_0: \mu_1 = \mu_2$ の下でt分布を使って近似する。両側有意水準 α では、 $|t| > t_{\alpha/2, \nu}$ なら棄却する。
自由度計算が複雑だが、これにより標本の分散やサイズの不均衡を反映し、より適切な検定が可能

二項分布と比率

1回の試行の結果を、成功なら1、失敗なら0で表すと、ベルヌーイ分布に従う。
成功確率 p が共通の独立な試行を n 回行くと、**成功数の合計 X が二項分布に従う。**

$$X \sim B(n, p)$$

母比率の推定

母比率とは、ある集団（母集団）における特定の事象が発生する割合のこと。
点推定値は、単純にサンプルから得られる比率（標本比率）を使って推定する。

点推定値

$$\hat{p} = \frac{x}{n}$$

標本のサイズで、事象の発生数を割った値。

区間推定

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

標本比率の分散は $p(1-p)/n$ 。 p を $\hat{p}$ で置き換え、平方根を取ったものが標準誤差の推定値。

この区間はWald近似。小標本や成功・失敗が極端に少ない場合は精度が悪くなる。

ただ、比率の推定や検定が二項分布に基づくのに対して、なぜ標準正規分布が登場するのか？

▶ 中心極限定理を応用している。（前章参照）

各試行が独立な二値データで、成功確率が群内で共通とする。
2群の差では群間も独立とする。下表は各群の成功数・失敗数が十分多い場合に用いるWald近似の信頼区間。

ケース	点推定値	区間推定	使用する分布
母比率の推定	$\hat{p}$	$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$	標準正規分布
母比率の差の推定	$\hat{p}_1 - \hat{p}_2$	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$	標準正規分布

*小標本や比率が0・1に近い場合、この近似は不正確になりやすく、区間が比率の範囲を外れることもある。

母比率の検定

検定の目的と検定統計量の計算方法

検定の目的 標本データを使って母集団の比率について仮説を検証する

仮説(例) $H_0 : p = p_0$ (母比率 p は特定の値)

$H_1 : p \neq p_0$ (母比率 p は特定の値ではない。または、他の値である)

検定統計量の計算

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \stackrel{H_0}{\sim} \text{approx } N(0, 1)$$

推定と検定で使う値の違い

推定と検定で、同じように標準誤差を求めがちだが、重要な違いがある。

母比率の推定

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

母比率 p が未知なので、標本比率 $\hat{p}$ を代入して標準誤差を推定する。

母比率の検定

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

$H_0: p = p_0$ の下での標準誤差は、仮説で指定した p_0 を使って計算する。

p_0 を用いても、正規分布による検定は近似。 np_0 と $n(1 - p_0)$ が十分大きいことを確認する。

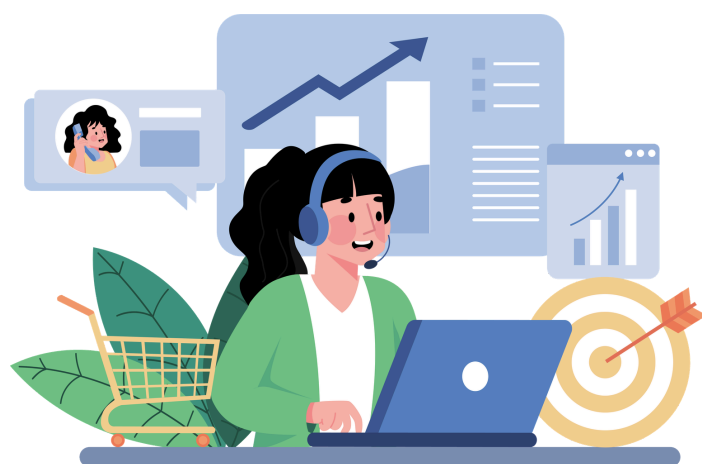
▶ ポイントは、標準誤差に代入する値。
ここで示す推定では $\hat{p}$ 、検定では帰無仮説の p_0 を用いる。

母比率の検定

具体例

具体例

営業チームが、事前に定めた基準成約率20%を上回っているか確認したい。一定の営業方法で100件の独立な営業機会を観測し、35件の成約が得られた。ここでは20%を、標本から推定した値ではなく固定の比較基準として扱う。



データ

- 成功数： $x=35$ 件
- 標本サイズ： $n=100$ 件
- 標本比率： $\hat{p}=35/100=0.35$

仮説設定

帰無仮説

$$H_0 : p \leq 0.20$$

母成約率は基準20%以下。

対立仮説

$$H_1 : p > 0.20$$

母成約率は基準20%を上回る。

比率の分析は、UIUX改善施策などでも行われる。エントリー率やCVRなど、応用する機会は現場でも多い。

▶ 片側検定を行う

右片側検定を分析前に選び、帰無仮説の境界 $p_0=0.20$ で計算する。

母比率の検定

具体例

検定統計量の計算

母成約率が固定の基準20%を上回るか、帰無仮説の境界 $p_0=0.20$ を用いて検定する。

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

標準誤差は、

$$SE = \sqrt{\frac{0.2(1-0.2)}{100}} = 0.04$$

よって、検定統計量は

$$z = \frac{0.35 - 0.2}{0.04} = 3.75$$

5%の有意水準（ $\alpha=0.05$ ）を使用すると、**片側検定の臨界値は1.645**

▶ 検定統計量は臨界値よりも大きいので、帰無仮説を棄却する。

結論

有意水準5%で帰無仮説を棄却し、母成約率が基準20%を上回ることを示す統計的証拠が得られた。ただし、この結果だけで営業方法の変更に
よる因果効果は分からない。方法の採用は、**比較実験の結果や費用対効果も合わせて判断する。**

分散の取り扱いはシンプル

ベルヌーイ分布の分散は $p(1-p)$ 。この検定では $p_0=0.20$ を代入するため、別の分散パラメータを推定する必要はない。ただし、**検定統計量の正規分布による扱いは近似。**

母比率の差の検定

検定の目的と検定統計量の計算方法

青の統計学

検定の目的 2つの母集団における比率が等しいかどうかを検定する

仮説(例) $H_0 : p_1 - p_2 = 0$ (2つの母集団の比率が等しい)

$H_1 : p_1 - p_2 \neq 0$ (2つの母集団の比率が等しくない)

当然一方向の検定として、片側検定を行うことも可能。

検定統計量の計算

$H_0: p_1 = p_2$ の下では、2群に共通の母比率を仮定する。

そこで、両群の成功数を合計し、両群の標本サイズの合計で割ったプール推定値 $\hat{p}$ を使う。

$$\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2} \quad \text{とすると} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

共通母比率のプール推定値を使い、帰無仮説の下での標準誤差を計算する。

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}} \stackrel{H_0}{\sim} N(0, 1)$$

2群の二値データは独立とする。帰無仮説の下で各群の成功・失敗の期待度数が十分多い場合に正規近似を使う。

ある保険会社では、顧客が契約更新をする割合（更新率）について、2つの異なる営業手法 A と B を比較したいとする。新しい営業手法、Aの方が更新率が高いかを検定したい。
A群とB群は重複のない独立な顧客群とし、同等の条件で観測したものとする。Aの更新率が高いかを調べる右片側検定を、分析前に設定する。



- 営業手法 A: 200人にアプローチし、そのうち120人が契約を更新した。
- 営業手法 B: 250人にアプローチし、そのうち140人が契約を更新した。

仮説設定

説無帰

$$p_A \leq p_B$$

営業手法 A の更新率は B と同じかそれ以下

対立仮説

$$p_A > p_B$$

営業手法 A の更新率が B より高い

▶ 片側検定を行う

学んだ内容を問題で確認 この章に対応する演習問題を3問選びました。リンクまたはQRコードから挑戦できます。

問題 1

レベル2

p値の解釈

p値0.03という結果の、最も適切な解釈を選ぶ。

**▶ 問題に挑戦する**

クリック、またはQRコードを読み取ると問題が開きます

問題 2

レベル2

信頼区間の解釈

95%信頼区間の正しい解釈を選ぶ。

**▶ 問題に挑戦する**

クリック、またはQRコードを読み取ると問題が開きます

問題 3

レベル2

母平均の検定（両側検定）

25個の標本から、母平均が10cmかを有意水準5%で検定する。

**▶ 問題に挑戦する**

クリック、またはQRコードを読み取ると問題が開きます

**ポイント**

この章のほかの問題にも、Web版「青の統計学」で挑戦できる。

統計検定2級・3級対策の演習問題と模擬試験を用意しています。 app.statisticsschool.com