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

はじめに

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　**あらかじめ全てのデータが与えられており**、そ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。

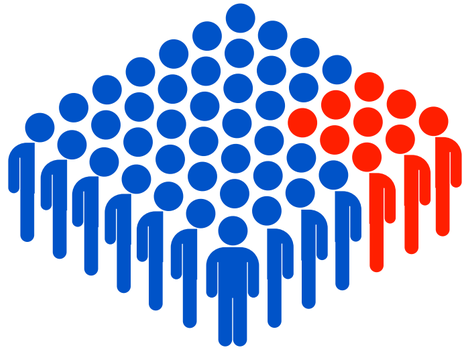
▶ 主な目的は、データを最大値や最小値、中央値などの指標で**整理し、可視化することで、全体像を把握し、意思決定につなげる**こと

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

> 国勢調査など、全数調査に勝る手法はないが、**コストや計算量、欠損等で扱えるデータには限り**があるの実態。

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	データセット全体	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



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)

データの総数を個数で割って求める値。
外れ値の影響を強く受ける特徴がある。

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

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。

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

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

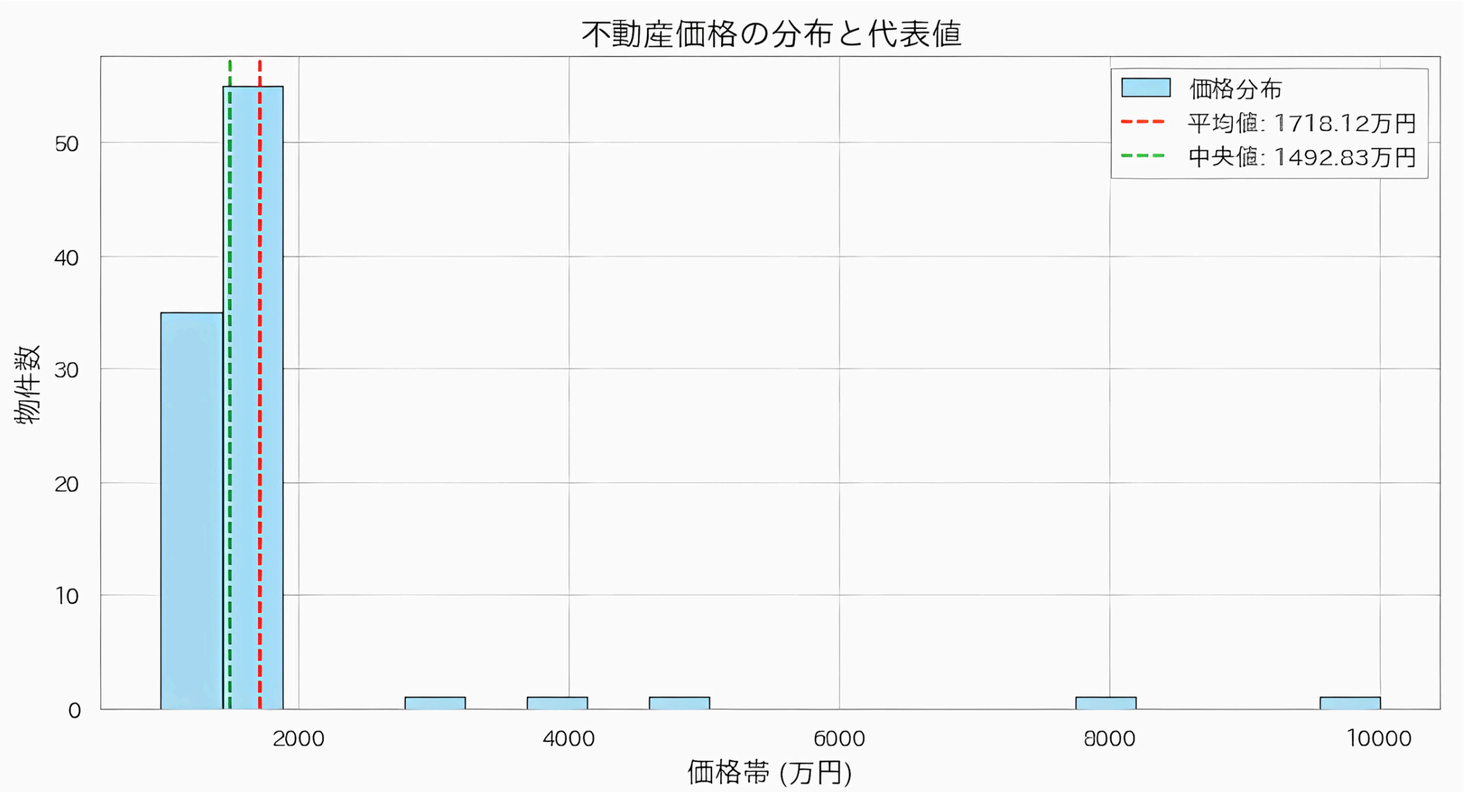
代表値　平均値と中央値

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: 高額な物件が多い場合、平均値が実態を反映しない。
- 中央値: 市場の中心的な価格帯をより適切に反映。

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**

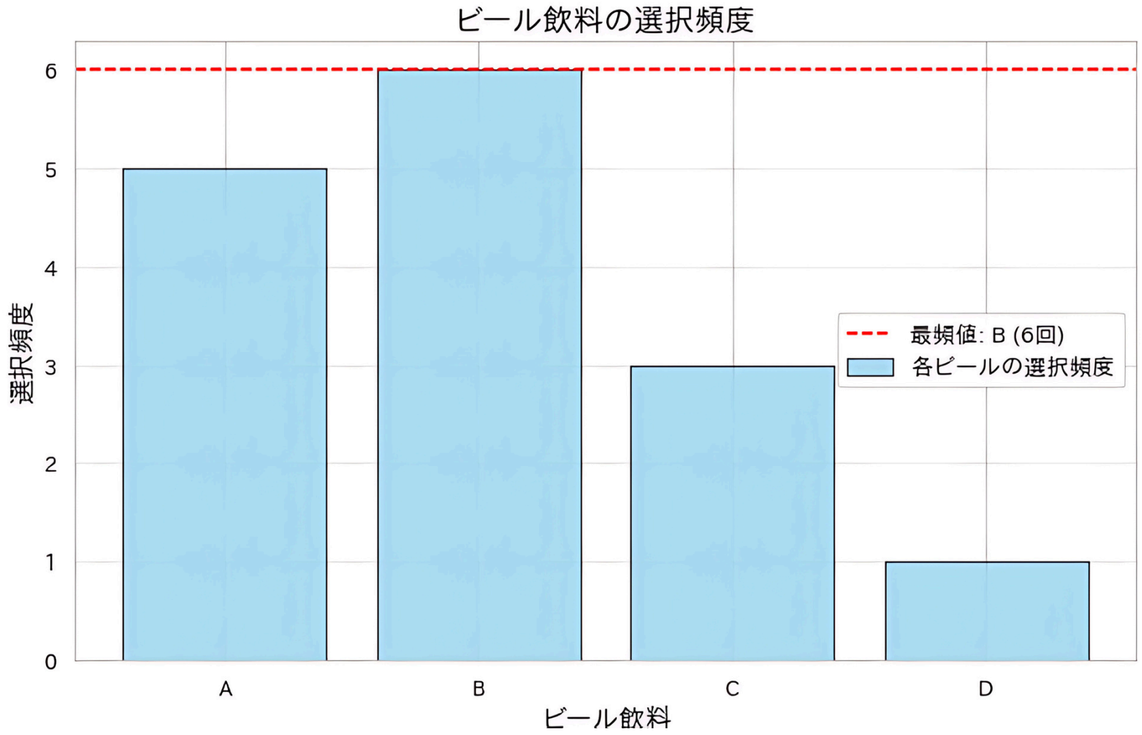


代表値　最頻値

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。

離散データにのみ定義されるのも特徴の一つ。



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）

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

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

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）**
  - データの上位25%に位置する値

具体例

在庫管理

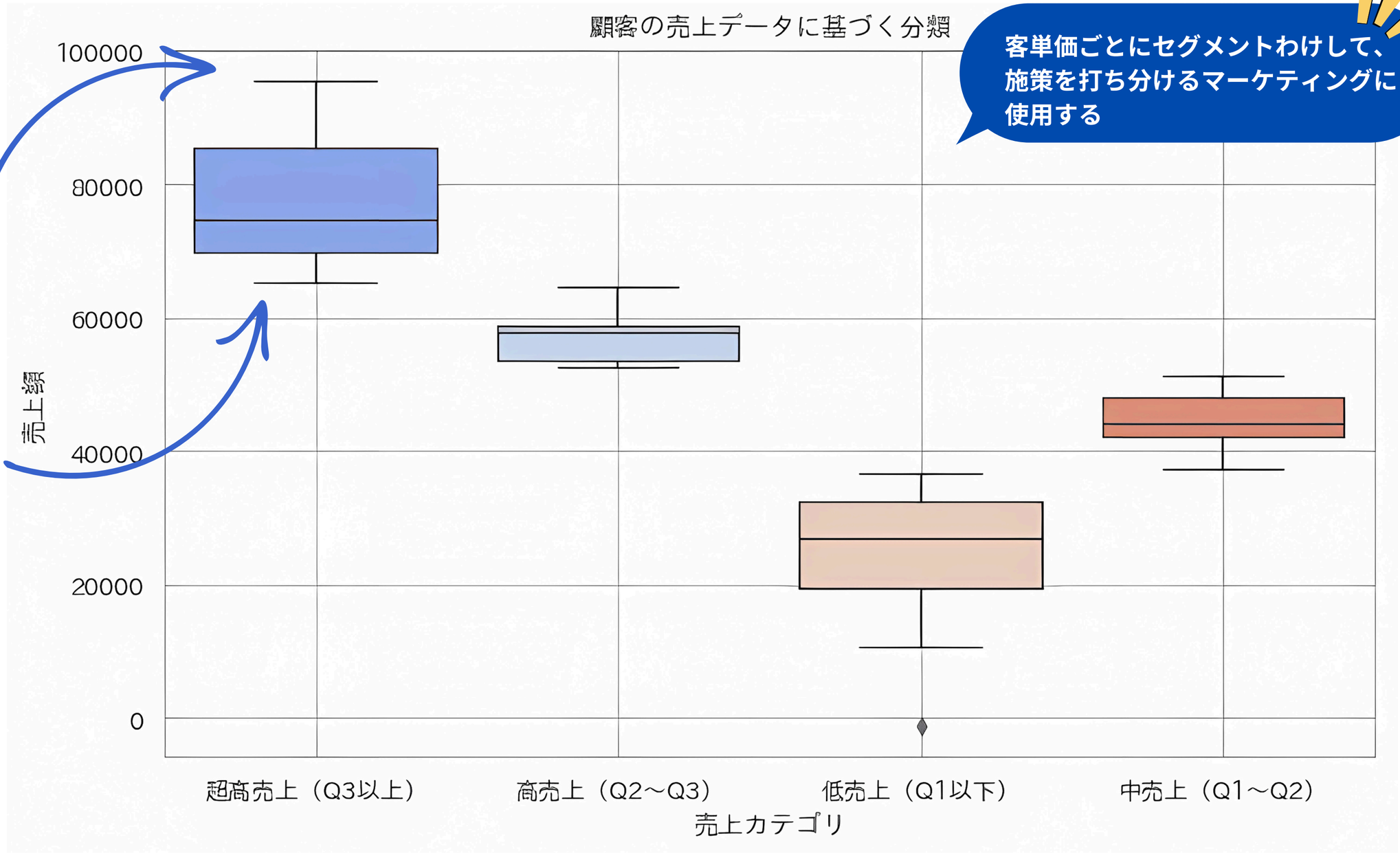
在庫のデータを分析する際、分位数を使用し、どの商品の在庫が過剰または不足しているかを判断できる。
在庫の上位25%の商品はよく売れている商品であり、下位25%の商品は過剰在庫を抱えている可能性が高い

▶ **できるだけ、在庫ロスがない＋売り切れによる機会損失がないように、次回発注数量を調整できる**

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。

分位数　図解



度数　度数の定義と計算方法

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。

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

クラスkの度数は、指示関数I(・)を使って上のように表せる。**あるデータxが階級Cに入っていれば、1カウントする。**

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。

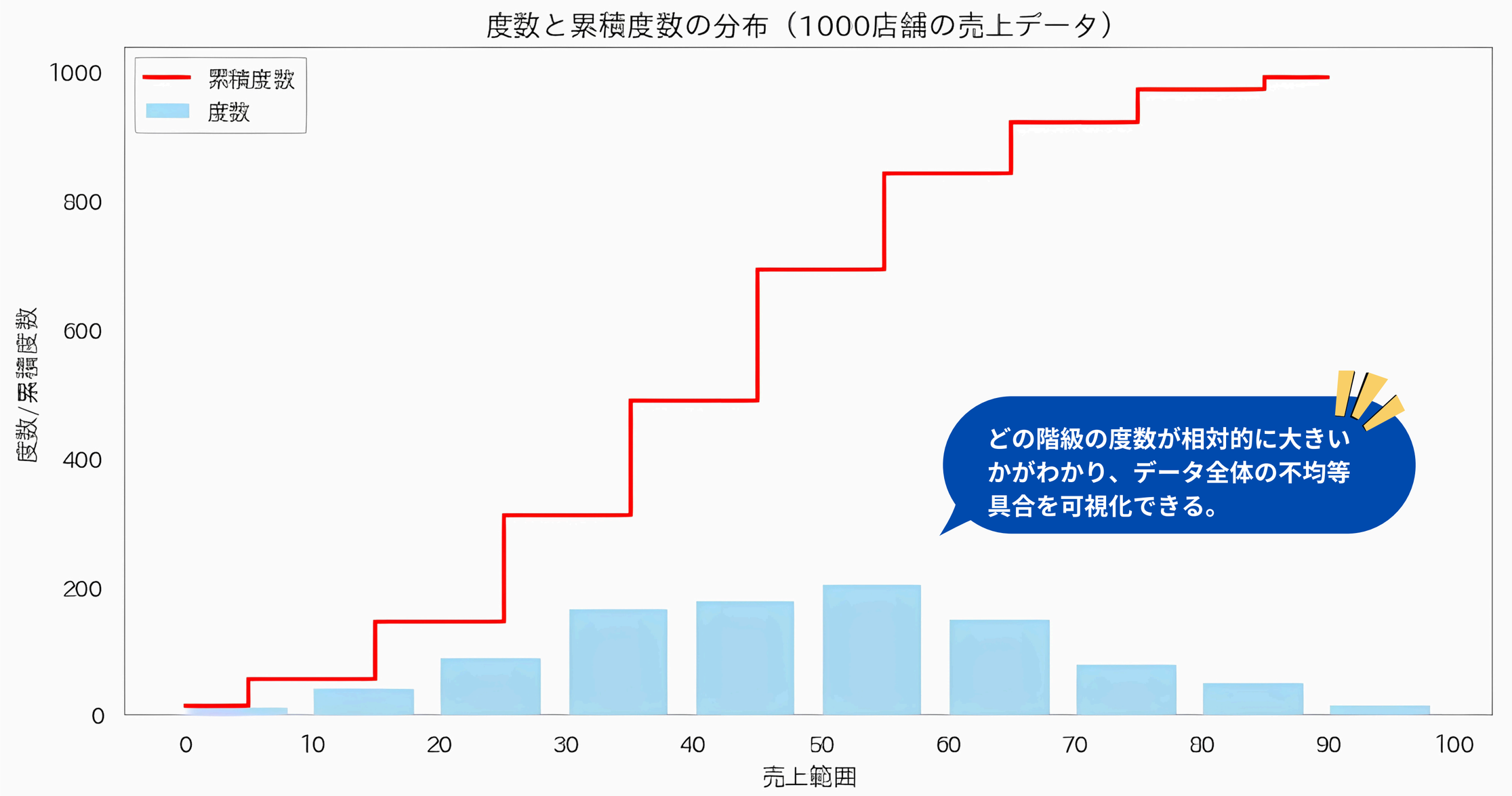
$$F_k = \sum_{j=1}^{n} f_k$$

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　図解



期待値　期待値の定義と計算方法

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

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

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	確率変数の将来の動きを 予測するための理論値
計算方法	$E[X] = \sum_{i=1}^{n} x_i p_i$
特徴	データを収集する前に 確率モデルに基づいて計算
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

平均値

定義	実際に手元にあるデータを 「一つの代表値」にまとめるもの
計算方法	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

平均値は「**過去の実データ**」期待値は「**確率モデル**」に基づいて計算される。
平均値は**実際の代表値**、期待値は**将来の予測値**

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

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

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

▼ より計算しやすい形式にすると・・・

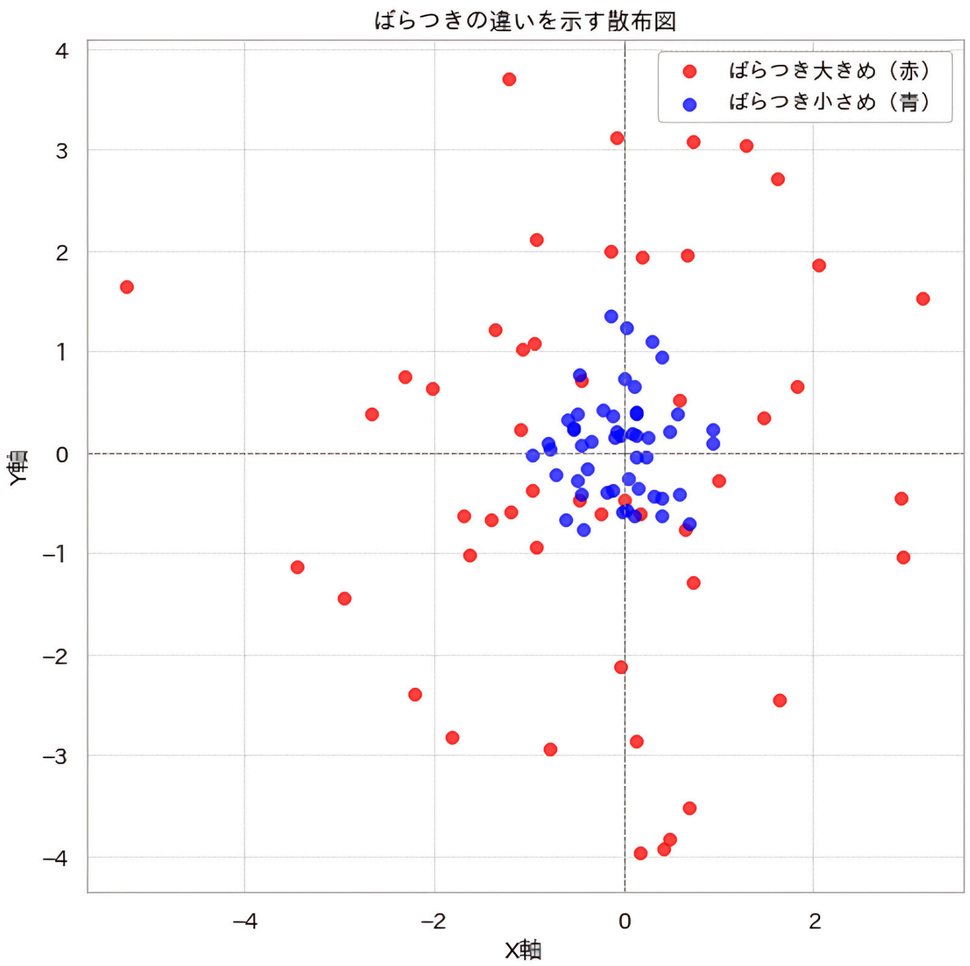
$$V[X] = E[X^2] - (E[X])^2$$

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

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

分散と標準偏差　図解



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。

分散と標準偏差

標準偏差と分散の関係性

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

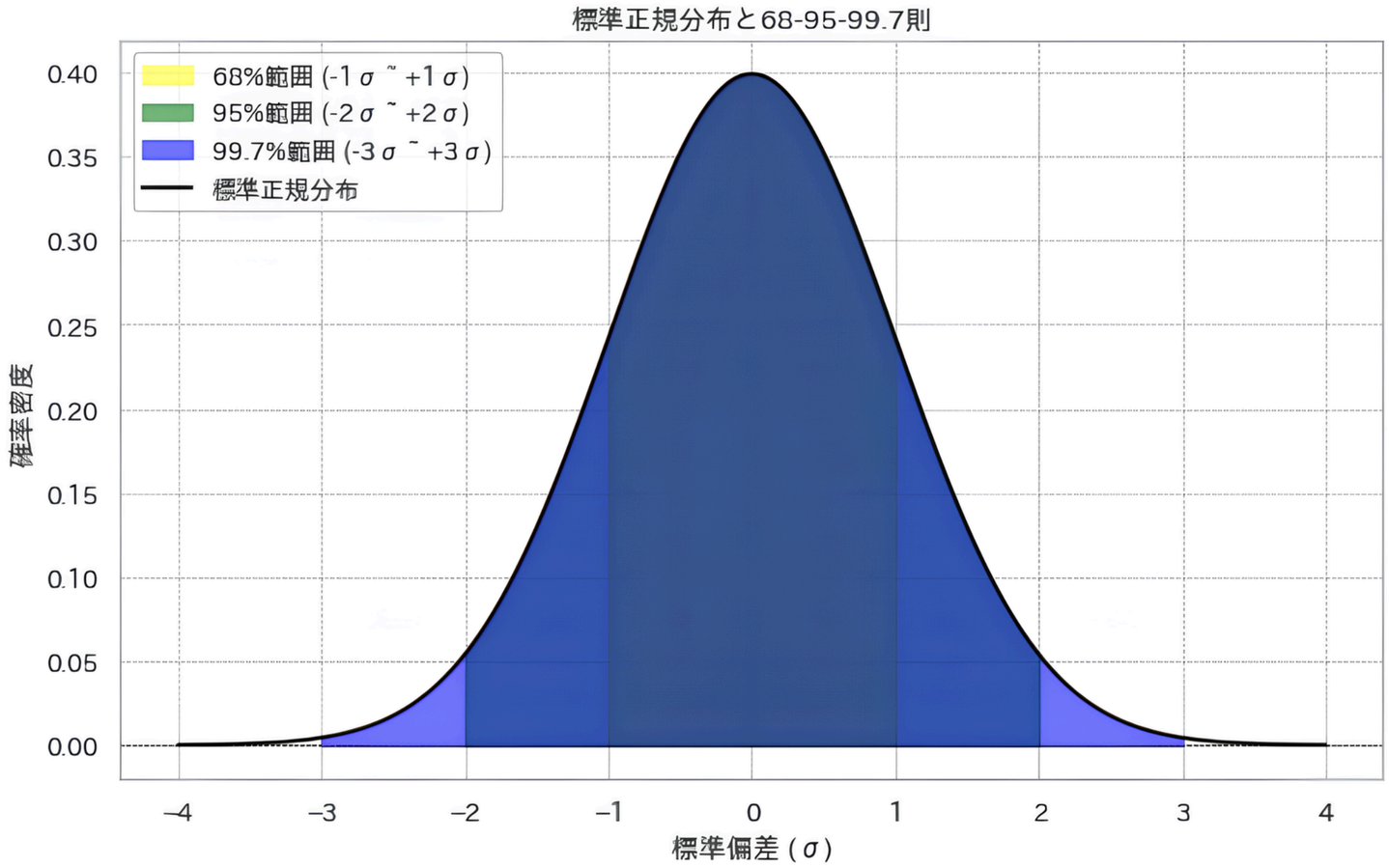
データのばらつきを直感的に理解したり、比較したりするために使用される。**元のデータと同じ単位なので解釈しやすい。**

計算方法

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

分散と標準偏差

正規分布に従う確率変数の標準偏差の範囲



$\mu \pm 1\sigma$ 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。

リターン（R）	確率
-5%	0.2
5%	0.5
15%	0.3

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

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

ちなみに、**標準偏差は7%**なのでかなりボラティリティが大きい銘柄。
期待リターンが吹き飛ぶバラツキがあることがわかった。

$$SD(R) = 0.07$$

分散と標準偏差

標本分散と不偏分散

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

標本データを用いて計算された分散。

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

標本分散だと、不偏性を満たさない。

一致性

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致性と呼ぶ。不偏性と合わせ、推定量の望ましい性質。

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

標本平均が、母平均よりもデータのばらつきを小さく見積もってしまうから。

標本平均 は、**データの中心に「引き寄せられる」傾向**がある。
そのため、データと平均値の差（偏差）が実際より小さくなる傾向があり、結果として**標本分散が母分散よりも小さく見積もられる。**

$$E[(x_i - \overline{x})^2] = \sigma^2 - \frac{\sigma^2}{n}$$

左のように、**母分散よりも標本分散の期待値を低く見積もってしまうのが、**不偏性がない理由。**差分はバイアスと呼ぶ。**

不偏分散

母集団の分散をより正確に推定するため、分母を n−1にする

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

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**

分散と標準偏差

備考 excelでの関数について

Excelでは、どちらの標準偏差が使われているのか

VAR.Pを使う場合

標本分散で計算されている。

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

VAR.Sを使う場合

不偏分散で計算されている

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

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

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

▶ **異なる単位やスケールを持つデータセット同士のばらつきを比較する際に役立つ**

計算例

テストの点数

クラスAのテスト結果: 平均80点、標準偏差10点　　クラスBのテスト結果: 平均100点、標準偏差15点

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

これにより、クラスBはクラスAよりも相対的にばらつきが大きいことが分かる

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**

目次

確率の基本 独立な試行

独立な試行

ある試行の結果が、他の試行の結果に影響を与えない状況を指す。
つまり、各試行が互いに無関係であり、1回目の試行の結果が2回目の試行に影響しない。

ある試行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）

とすると、**各ユーザーのクリック行動は他のユーザーのクリック行動に影響されないため、試行は独立**とみなせる。

ポイント

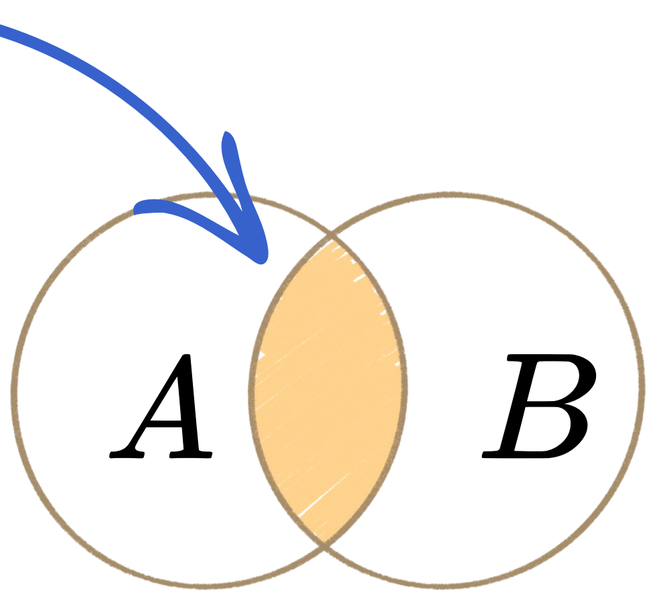
独立な試行では、**各試行の結果が他の試行に影響を与えない**
時系列データでは、独立性の前提が基本的にはない。

確率の基本 条件付き確率

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、試行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ても、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**

確率の基本 条件付き確率の具体例

具体例

Webサービスのログを解析して、ユーザーが 「Aページを訪れた後にBページも訪れる確率」 を求めたい。

データ

- 全体のアクセス数 N は 100,000 とする。
- Aページを訪れたユーザー数 N(A) は 20,000。
- Aページを訪れた後にBページも訪れたユーザー数 N(A∩B) は 5,000。

計算

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

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

▶ つまり、Aページを訪れたユーザーがBページも訪れる確率は25%

確率変数　定義と特徴

確率変数

起こりうる事象に割り当てている値を取る変数。
各事象は確率をもち、その比重に応じて確率変数はランダムに値をとる

離散型

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

離散的な数値を取る確率変数
整数や有限個の値が含まれる

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

$$P(a \leq x \leq b) = \int_a^b f_X(x)$$

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

確率変数は定数ではなく、複数の値を取る可能性があり、それぞれに確率が割り当てられる。

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　期待値と分散の計算方法

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りない。

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

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

独立性　独立な確率変数XとYに対して、

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

分散の性質

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

スカラー倍の変化

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

独立性

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

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

確率変数　期待値と分散の計算方法

分散の加法性

２つの確率変数が、互いに独立か独立でないかで決定する。

確率変数が独立だと共分散はゼロになるので分散同士の和のみ。

独立

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

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

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$$

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。

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。

色々な確率分布　ベルヌーイ分布

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

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

期待値

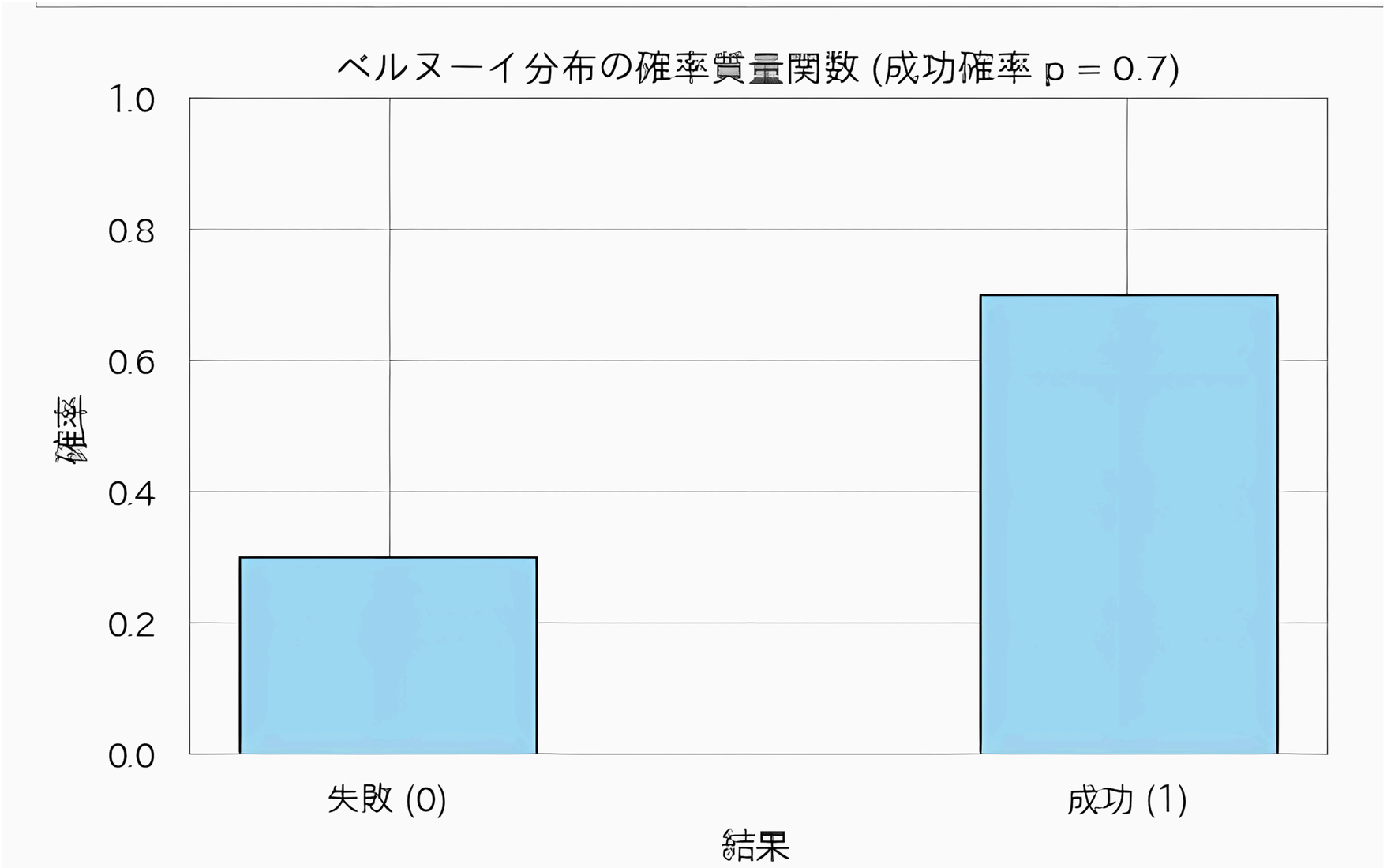
$$E[X]=\sum_{x=0}^{1}xP(X=x)=0*(1-p)+1*p=p$$

分散

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

色々な確率分布　ベルヌーイ分布の図解

確率質量関数



色々な確率分布　二項分布

二項分布

n回の独立したベルヌーイ試行で成功する回数を表す確率分布。
確率変数Xが二項分布に従う場合、確率質量関数は次の通り

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

期待値

$$E[X] = np$$

分散

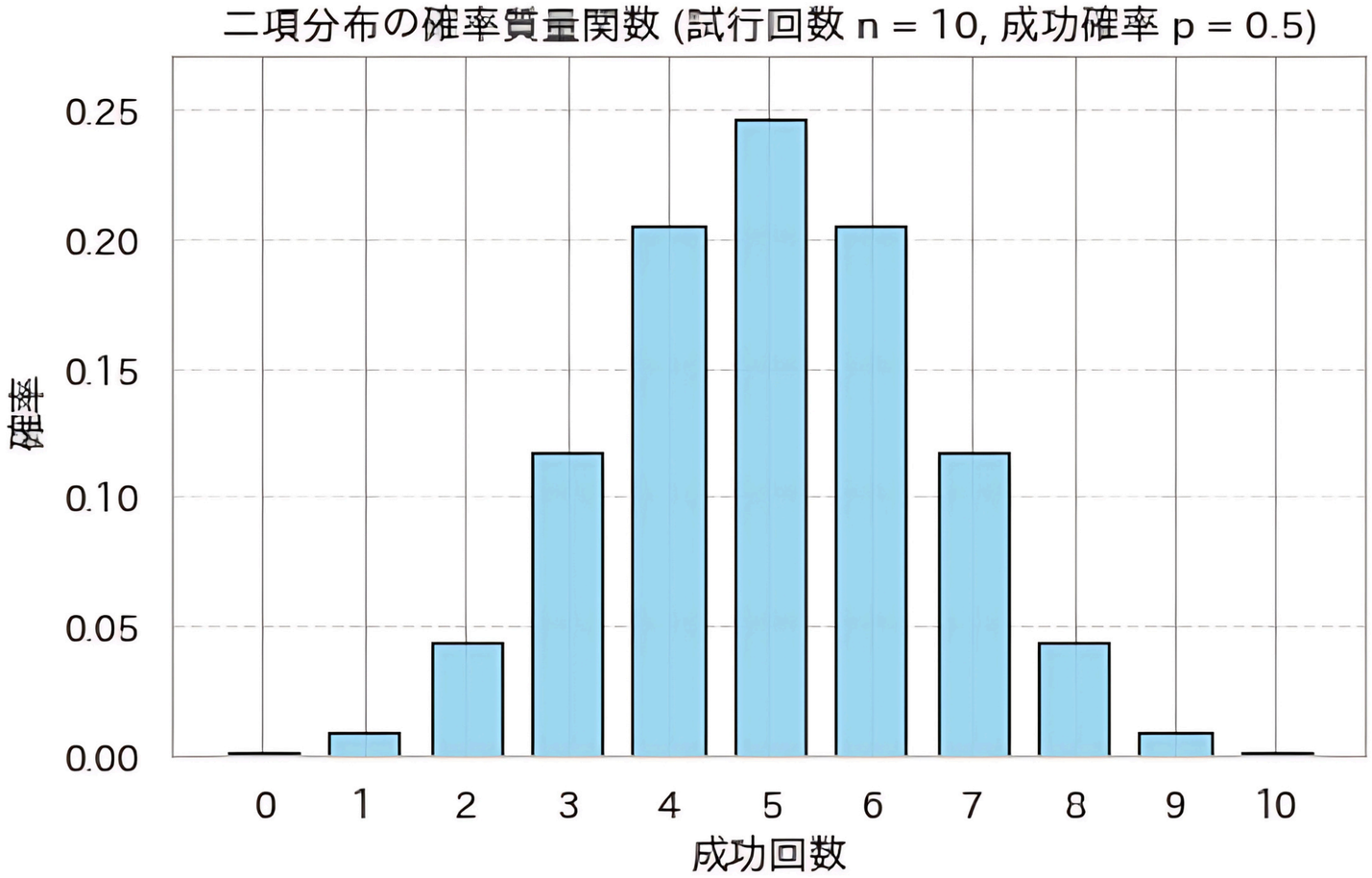
$$Var(X) = np(1-p)$$

$\times n$

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

色々な確率分布　二項分布の図解

確率質量関数



成功回数kの分布は、pとnに依存。pが大きくなると**分布の中心が右に移動し、分散が広がる。**

ポイント

成功確率が0.5の時、サンプリングしてできた分布は左右対称に近くなる。
離散分布だが、正規分布に似ている。

色々な確率分布　正規分布

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

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

期待値

$$E[X] = \mu$$

分散

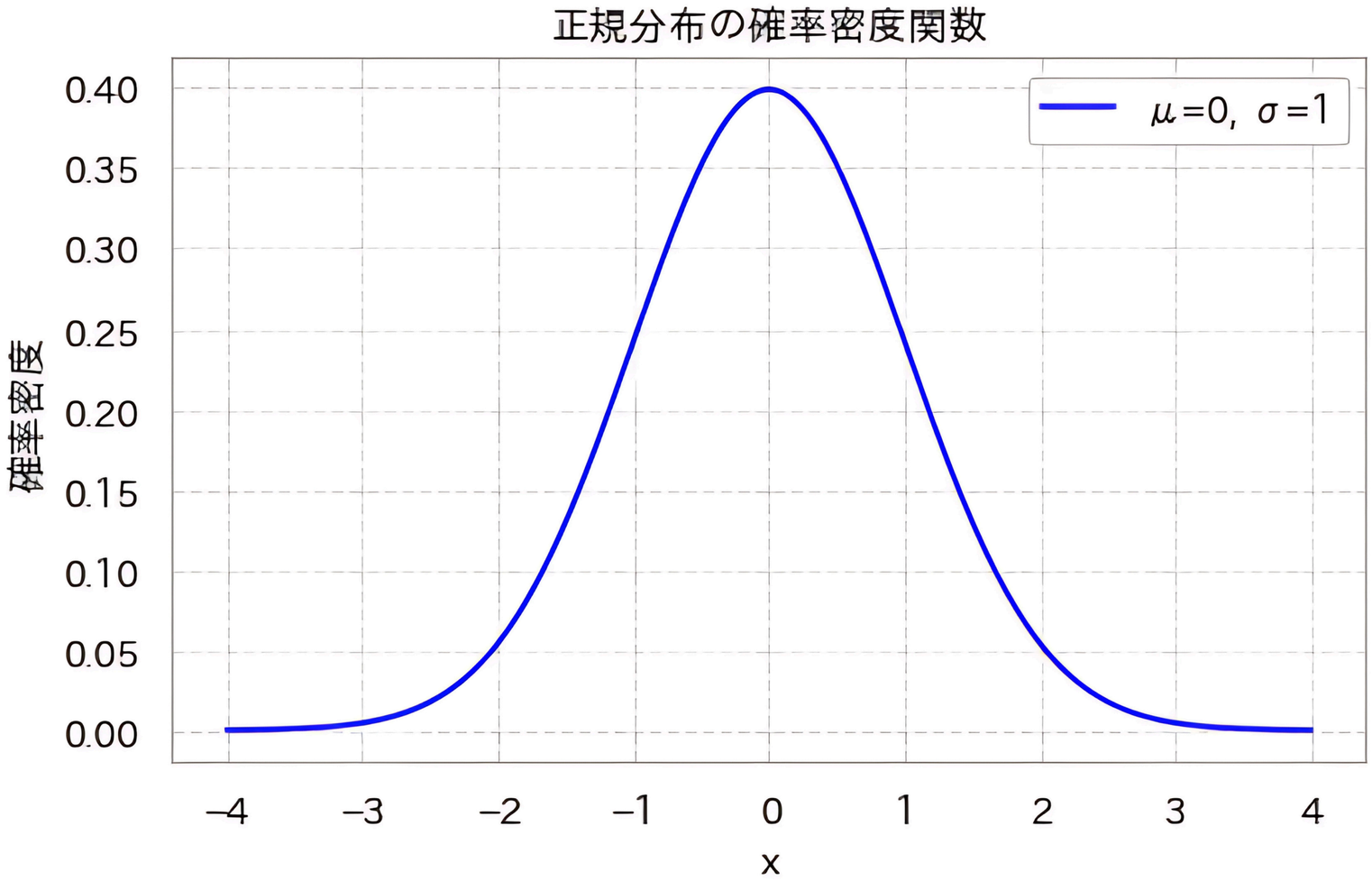
$$Var[X] = \sigma^2$$

確率変数が、ある分布（正規分布）に従うという意味

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

色々な確率分布　正規分布の図解

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。

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

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

> **確率変数同士の和も同じ分布に従う性質を「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

確率変数が独立ではない場合、共分散を考える必要があるので注意。

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

ポイント

加法性：複数の正規分布の和も正規分布に従い、平均と分散はそれぞれの分布の和で計算できる。

色々な確率分布　標準正規分布

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

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

期待値

$$E[X] = 0$$

分散

$$Var[X] = 1$$

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

二項分布は試行回数が多い場合、**正規分布で近似できる。**
計算の簡略化や**統計的推測を容易にするため**に用いられる。

> 今回の近似は、**中心極限定理によるもの。**二項分布限らず、あらゆる分布に従う独立な確率変数の和や平均は、正規分布に分布収束する。

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

▼

$$X_n \sim N(\mu = np, \sigma^2 = np(1-p))$$

二項分布に従う確率変数Xは、期待値と分散をそのまま引き継ぎ、正規分布に従うようになる。

色々な確率分布　二項分布の正規近似

標準化

正規分布に変換するために標準化を行う

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

確率変数が標準正規分布に従うように変換できた

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**

分散をサンプル数で割って、
スケーリングする

ポイント

本来、zスコアは標準正規分布の分位点(例えば、95%信頼区間の場合は1.96程度)
試行回数が大きい場合、**二項分布は正規近似でき、正規分布を用いた信頼区間の計算が可能**

色々な確率分布　二項分布の正規近似

具体例2

z検定を用いた仮設検定では、二項分布で正規近似し、近似分布を使って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

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

$$\hat{p}, p_0$$

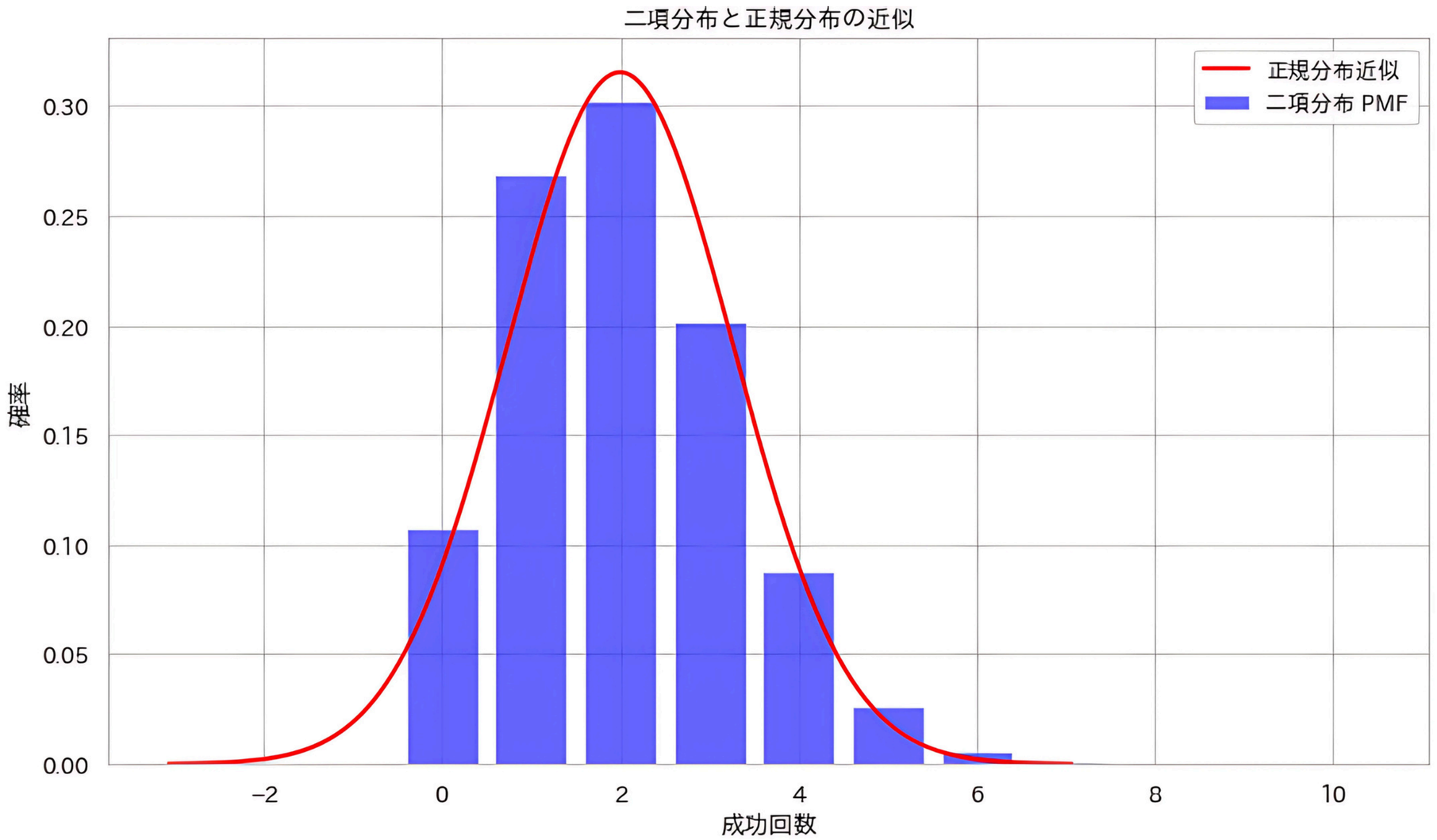
左：サンプルの成功確率
右：帰無仮説での成功確率

ポイント

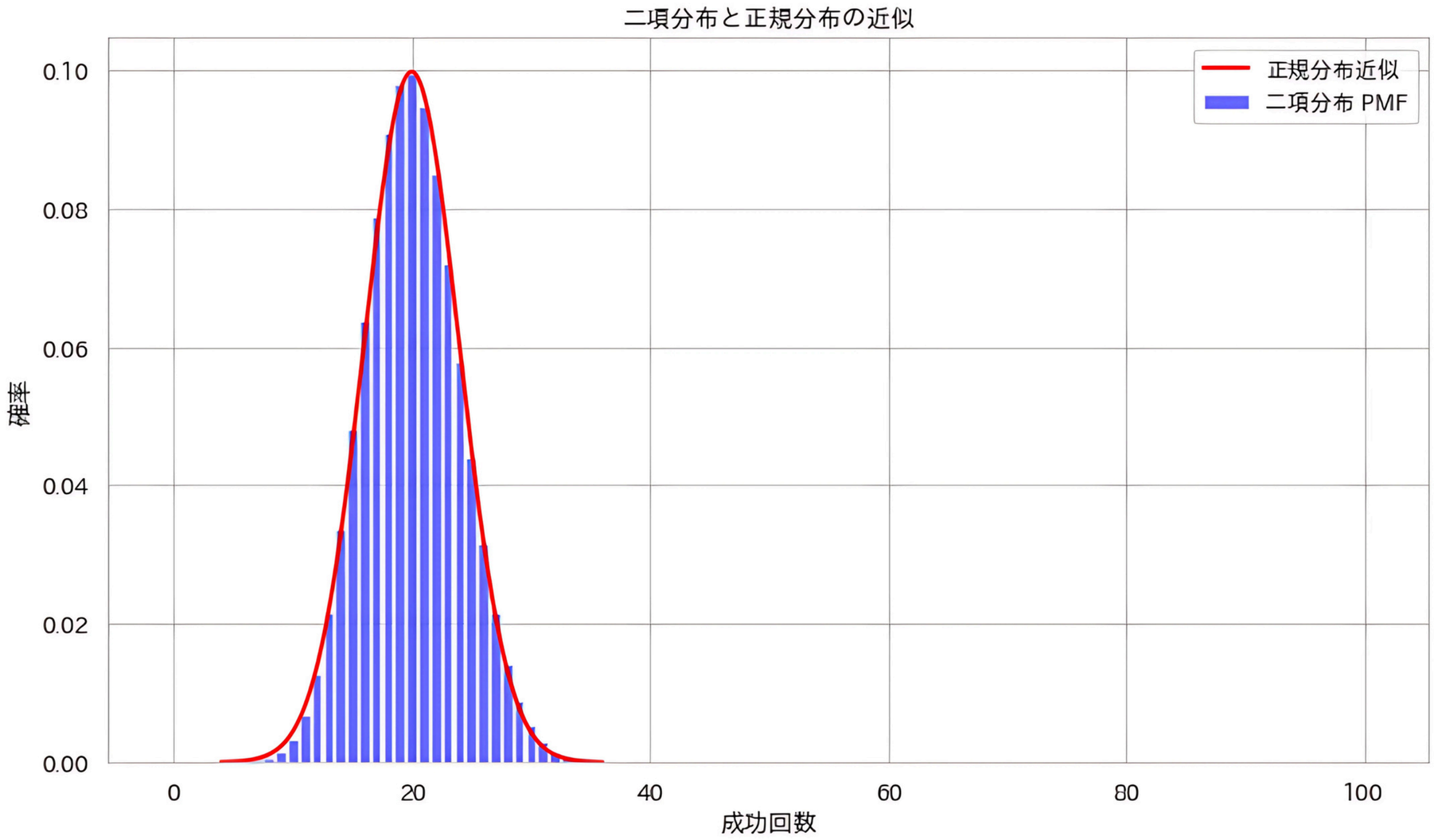
検定統計量 zが標準正規分布に従うため、そのまま有意水準と比較することができる

色々な確率分布　二項分布の正規近似

二項分布：n=10,p=0.2の場合



二項分布：n=100,p=0.2の場合



中心極限定理

定義と成立条件

中心極限定理

独立した同一分布に従う確率変数の和（または平均）は、試行回数が十分に多ければ、正規分布に近づくという性質。

$$\bar{X}_n \xrightarrow{d} N(0, \frac{\sigma^2}{n})$$

標準化

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

サンプル数が大きいとき、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。

中心極限定理の成立条件

確率変数が独立

独立性がなければ、**各試行の結果を単純に足し合わせることができず、和や平均が標準的な正規分布に収束しなくなる**可能性がある。

確率変数が同一の分布に従う

異なる分布に従う確率変数を足し合わせた場合、平均がどこに収束するかが不明瞭になり、正規分布への収束が保証されない。

確率変数の分散が有限

分散が無限大である場合、標本平均の分布が収束しない可能性がある。

ポイント

正規分布では、標本をとってきた元の集団（母集団）がなんであれ、標本の数が大きくなるにつれてサンプル（標本）の平均値は正規分布に近づいていく。

目次

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。

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。

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

より簡単に

$$r = \frac{Cov(X, Y)}{s_x s_y}$$

偏相関係数

変数 x と y の間の相関を、別の変数 z の影響をできるだけ取り除いた状態で測定する指標。
これにより、**zの影響をコントロールした関係を評価できる。**

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

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≤1)

一方の変数が増加すると
もう一方も増加する関係。

例：気温とアイスクリームの売上

負の相関 (−1≤r<0)

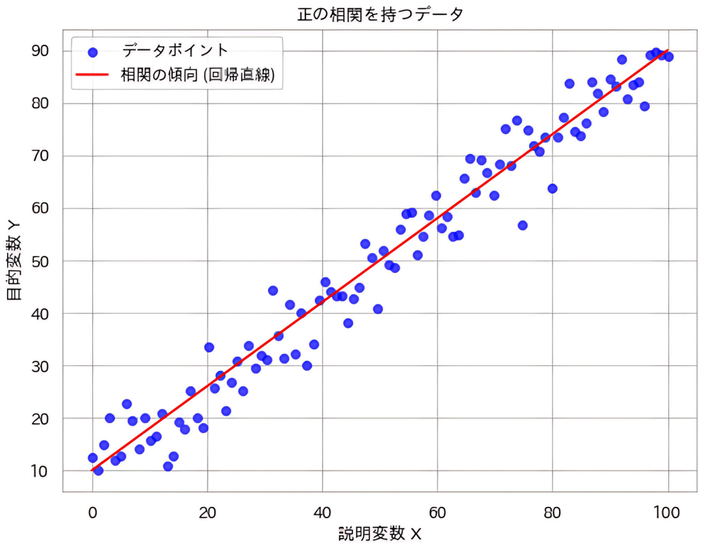
一方の変数が減少すると
もう一方は減少する関係。

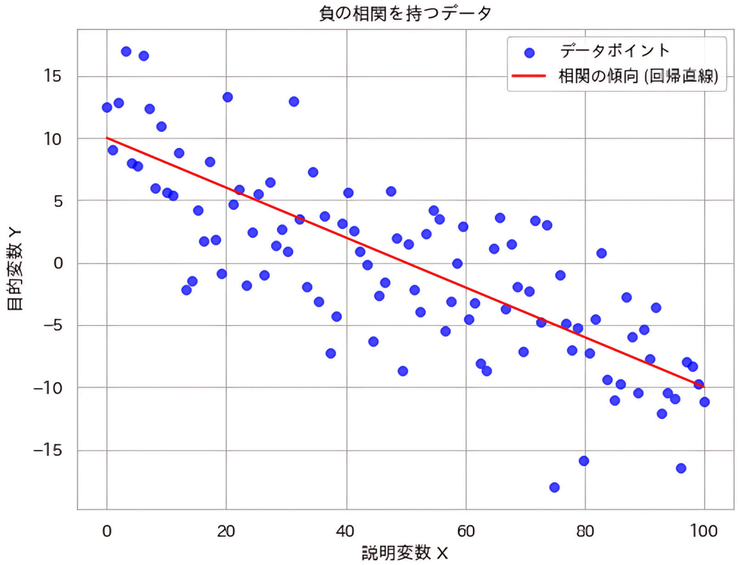
例：年齢と体力テストの成績（一般的に）

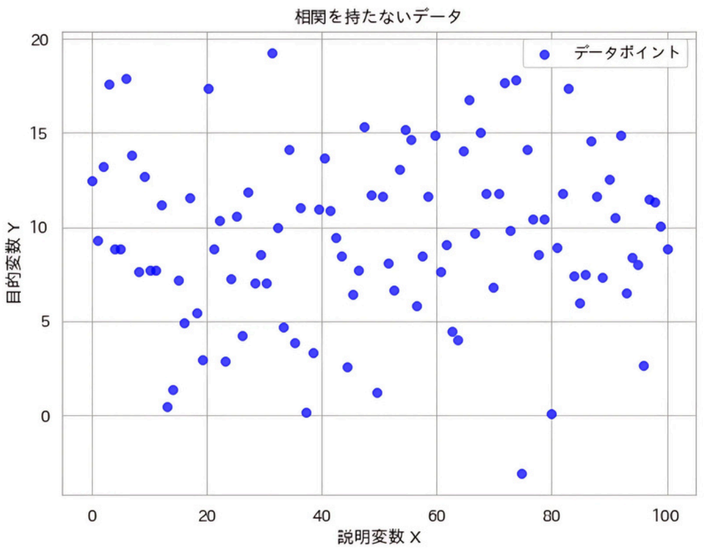
無相関 (r=0)

2つの変数に線形的な関係がない状態。

例：好きな色とテストの点数







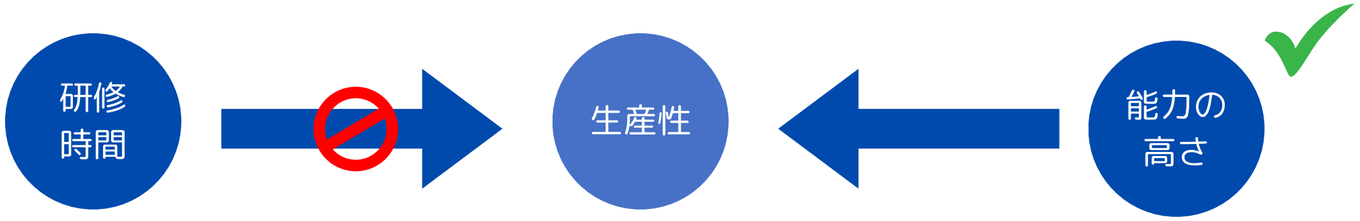
相関係数　相関係数の限界

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。



しかし、これは研修の効果ではなく、**元々能力の高い従業員が研修に積極的に参加していた**ためであり、研修自体が生産性向上の直接的な原因ではなかった。

▶ **「生産性を上げるために、研修時間を増やそう」**という誤った意思決定につながる恐れがある。

相関係数　相関係数の限界

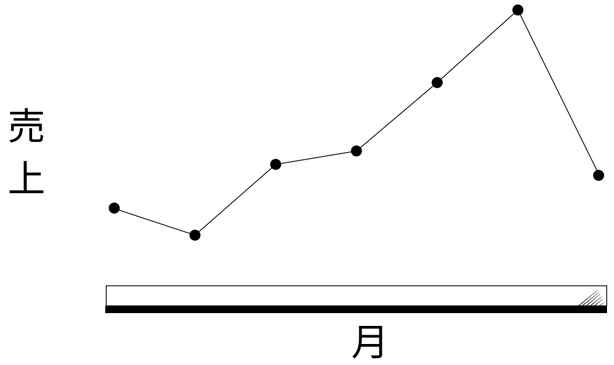
外れ値に敏感

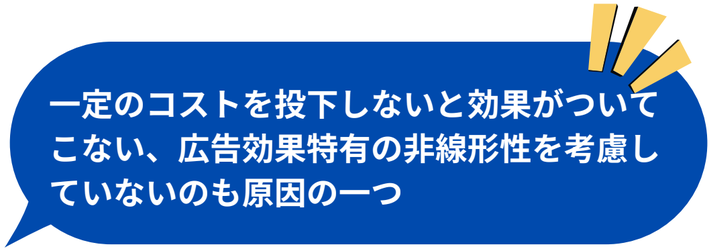
相関係数は、外れ値（極端な値）の影響を受けやすく、結果が大きく歪むことがある。

具体例

マーケティング部門が広告費と売上の相関を分析した際、**特定の月に異常に高い広告費を投じた結果、売上が急増**した。そして、**特定月のデータが全体の相関係数を大幅に引き上げ**、以下のような結果が得られた。

- 特定月を含めた相関係数: r=0.95 （強い相関を示す）
- 特定月を除いた相関係数: r=0.70（適度な相関）





この外れ値が全体の相関係数を高め、**実際の広告効果を過大評価する原因**となった。

▶ データの粒度が粗すぎると、大雑把な傾向しか掴めず、効率の良い施策に繋がらない。

相関係数　相関係数の限界

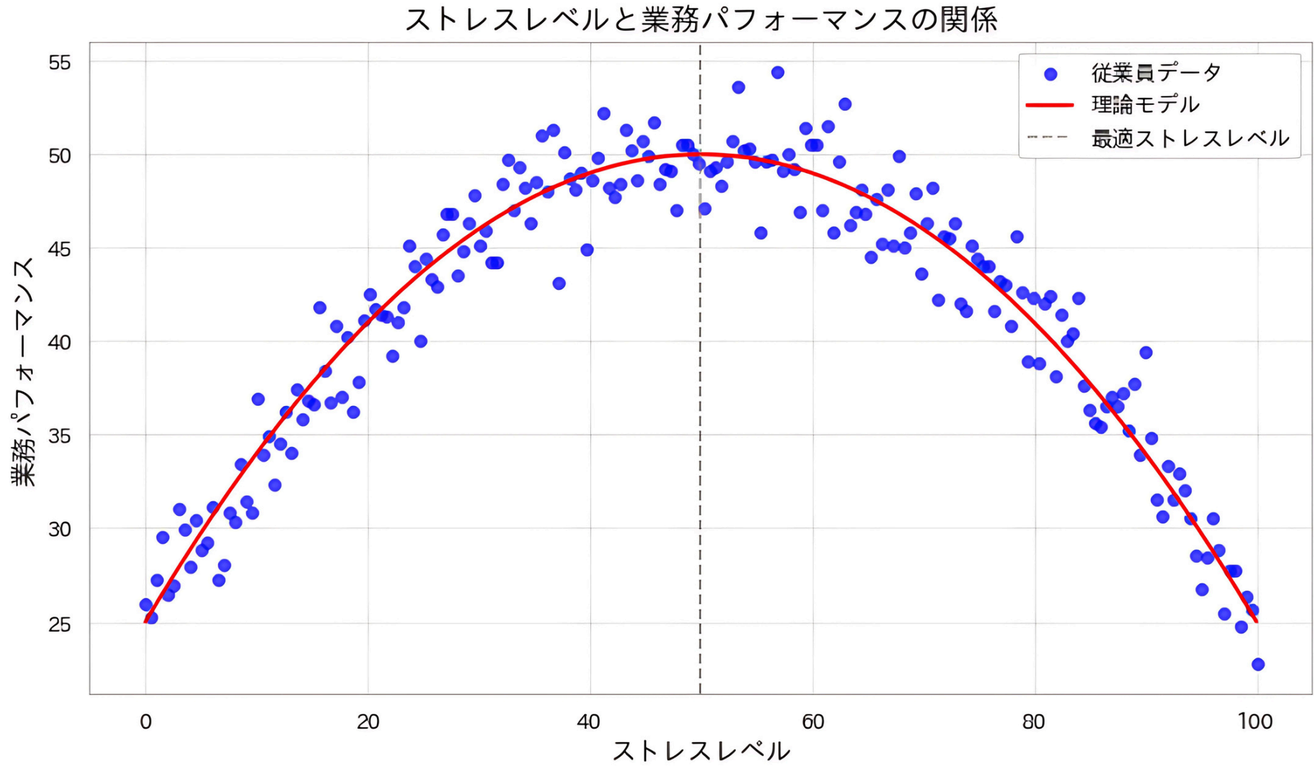
非線形関係を適切に捉えられない

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: 放物線的関係）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** : 説明変数が1単位増加すると、目的変数がどれだけ変化するかを示す。説明される側。
- **定数項** : 説明変数が0のときの目的変数の値

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): 目的変数に影響を与えると考えられる変数。説明する側。
  例: インプレッション、サイト回遊時間

ポイント

回帰式は、**目的変数と説明変数の関係を一次関数の形で表す**。
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。

回帰分析　具体例

具体例

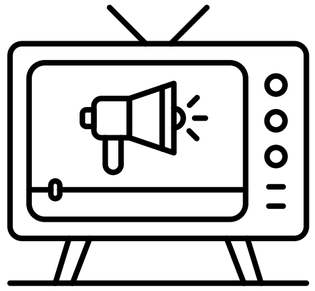
CM投下量(actual GRP)とKPI

目的変数 (Y)：KPI(商材の売上)
説明変数 (X)：CM投下量(actual GRP)

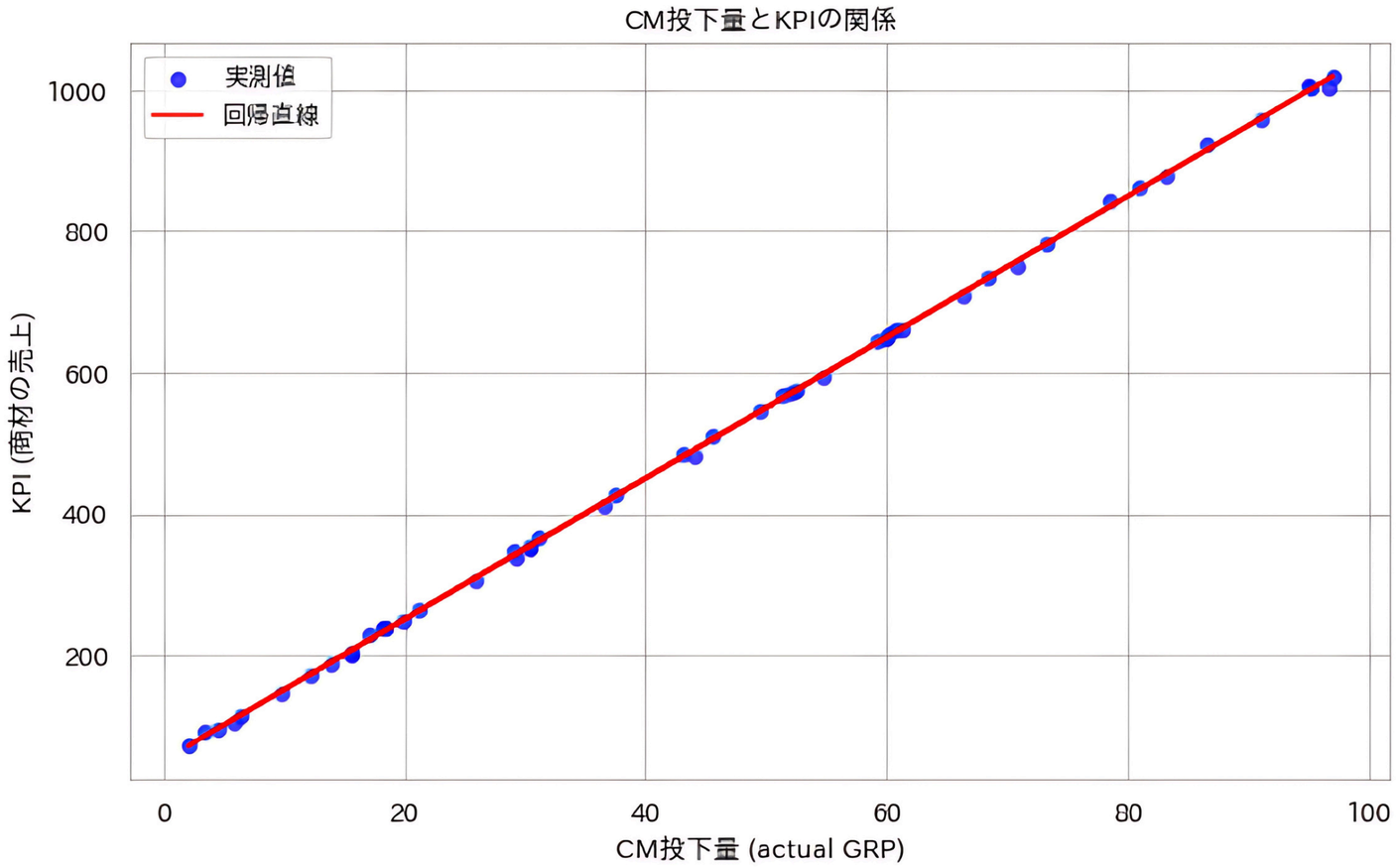
回帰分析を行い、次のような回帰式が得られたとする

$$y = 50 + 10x$$

- CM投下量が(X) が1上がるごとに
  KPI(Y) は10増加する関係
- CM投下がなくても基礎売上として50が見込まれる



実データだと、ここまで綺麗な直線にはならない



回帰分析　単回帰分析のメリット

単回帰の嬉しい点

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

解釈しやすさ

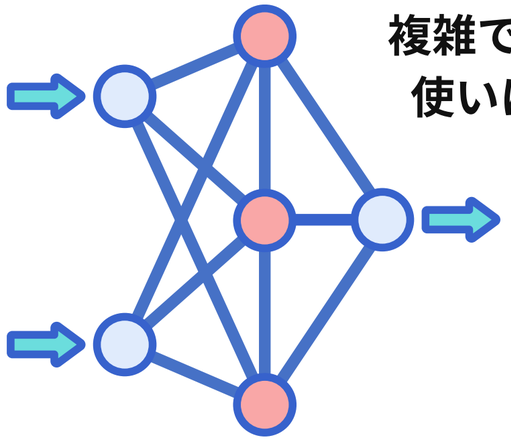
変数間の関係が直線的で理解しやすい

透明性の高さ

各変数の影響度が明確

計算の簡単さ

複雑なモデルに比べて、計算コストが低い



**複雑で実務では
使いにくい…**

例：ニューラルネットワーク

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。

$$\mathbf{y} = \mathbf{X}\beta$$

特徴ベクトル

$$\mathbf{X} = [1, x]^\top$$

重みベクトル

$$\beta = [\beta_0, \beta_1]^\top$$

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**。

$$y = \beta_0 + \beta_1 x + \ldots + \beta_k x_k + \epsilon$$

回帰分析　決定係数

決定係数

回帰モデルの当て嵌まりの良さを表す指標の一つ。
0から1の範囲を取り、1に近いほどモデルの当て嵌まりが良い。

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

> 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

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)

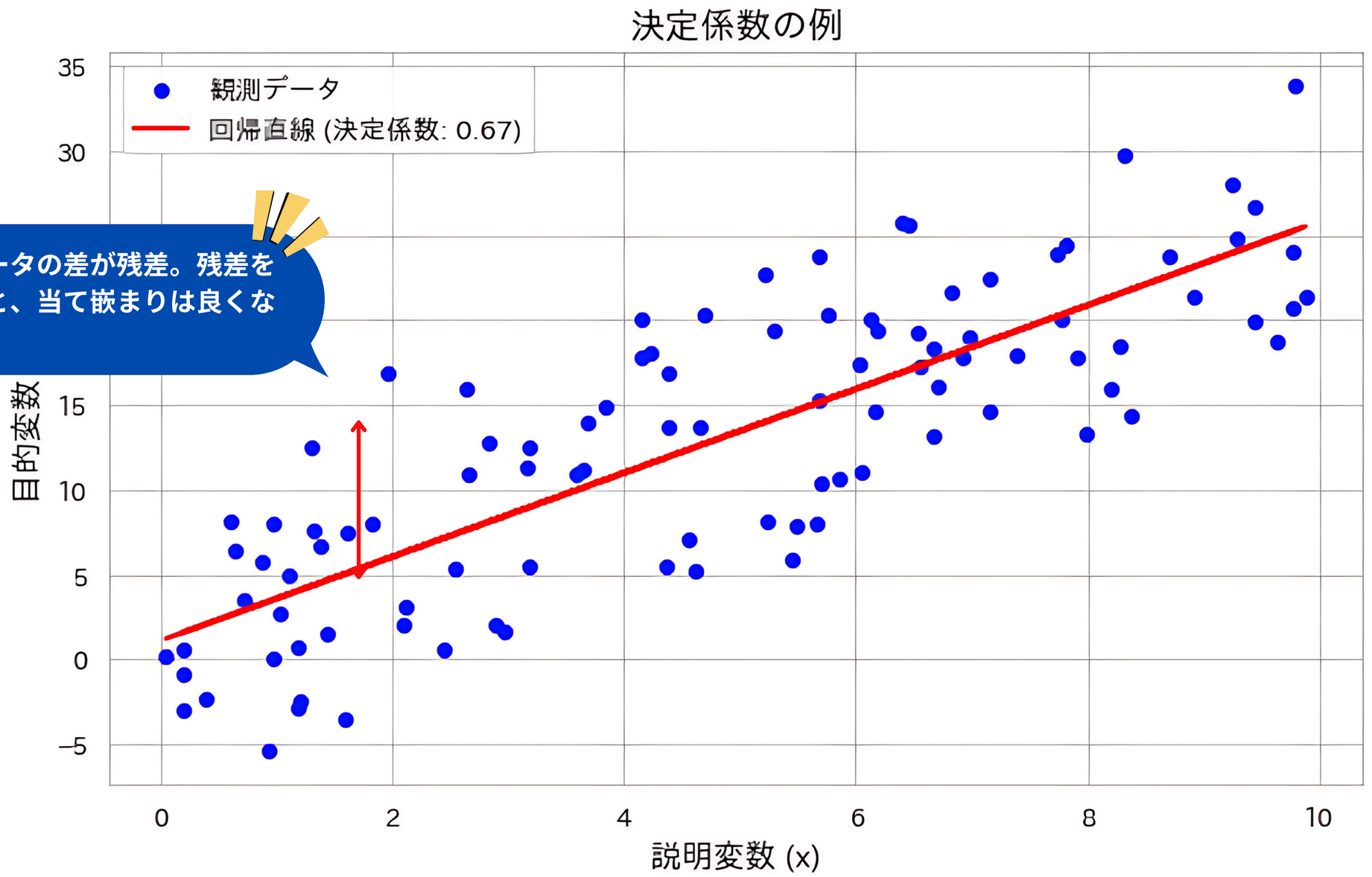
$$\sum_{i=1}^{n}(y_i - \hat{y})^2$$

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。

回帰分析　決定係数の図解



予測と実データの差が残差。残差を小さくすると、当て嵌まりは良くなる。

回帰分析　最小二乗法

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。

$$\beta_1 = \frac{Cov(x,y)}{Var(x)} \qquad \beta_0 = \overline{y} - \beta_1 \overline{x}$$

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

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

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。

回帰分析　最小二乗法

OLSの方針

③偏微分が0になる点を求める

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

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

> 最小値かどうかを厳密に調べるには、**二次微分**して0との大小関係を調べる必要があります。

これを解くと、

回帰係数

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

つまり ▶

$$\beta_1 = \frac{Cov(x,y)}{Var(x)}$$

定数項

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

目次

仮説検定の基礎　帰無仮説と対立仮説

帰無仮説 H_0

定義

特定の統計的検定において「現状に変化がない」または「差や関係がない」と主張する仮説。

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。

対立仮説 H_1

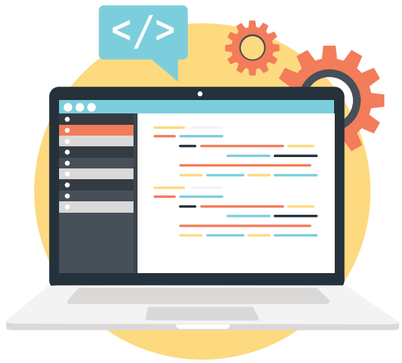
定義

帰無仮説が否定された場合に採択される仮説。

通常、研究者が証明したい主張や予測に対応する。

▶ 仮説検定では、上記の仮説を設定し、データに基づいて**どちらがより妥当かを判断する。**

具体例



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　新しいUIデザインは、CVR向上に寄与しない

H_1　新しいUIデザインは、CVR向上に寄与する

仮説検定の基礎　有意水準と棄却域

有意水準　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。

棄却域　統計量zが一定の基準値（棄却限界値）を超えた場合、帰無仮説を棄却する領域

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。

棄却限界値は　$z_{\alpha/2} = \pm 1.96$　として計算される

両側検定のほうが厳しい条件（片側の棄却域が片側検定に比べて狭い）のため、**片側検定のほうが棄却されやすい**

両側検定　$|Z| > z_{\alpha/2}$

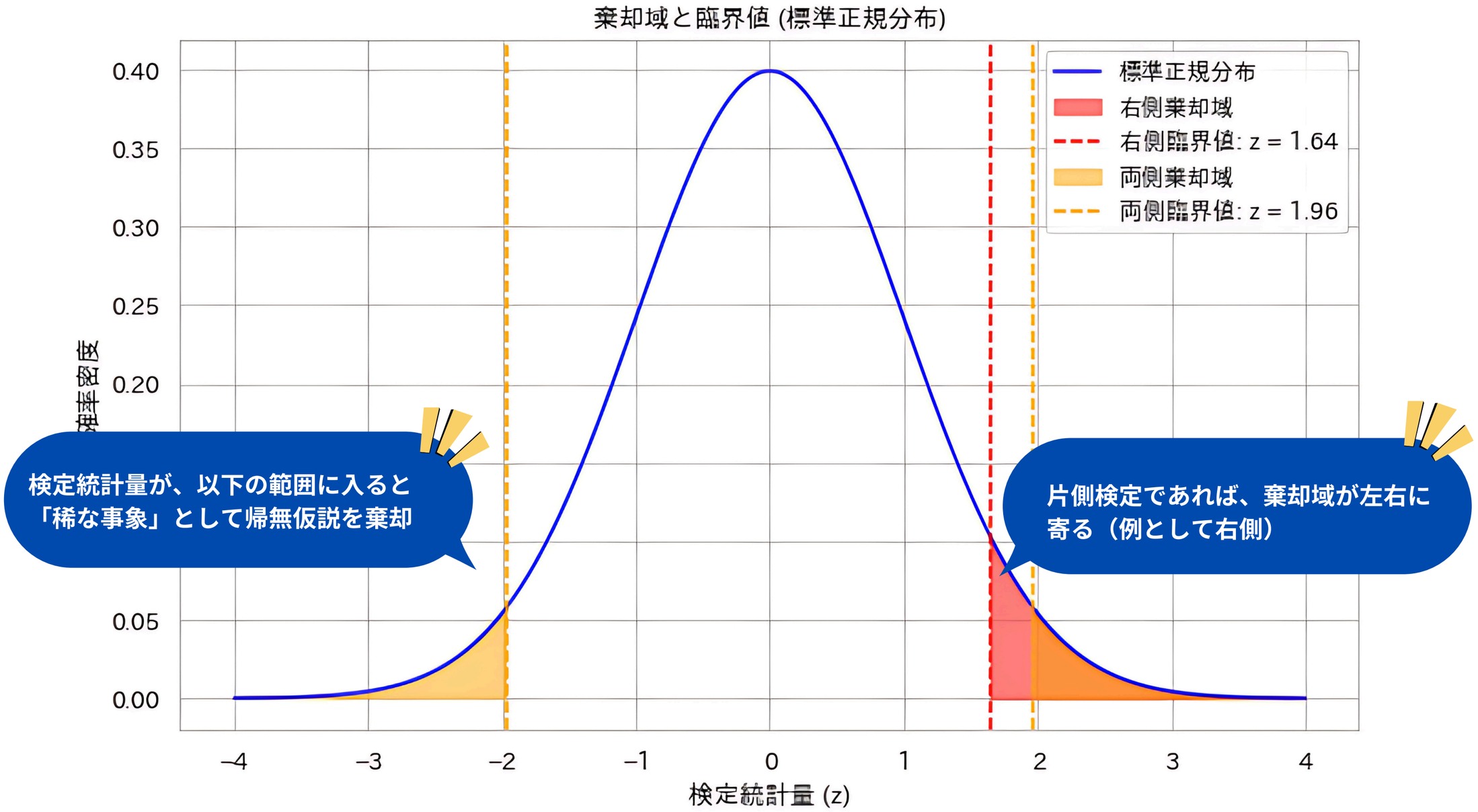
使い方　効果の方向が不明。どちらの方向に対しても関心がある場合。

片側検定(右側)　$Z > z_{\alpha}$

使い方　特定の方向にのみ、効果を期待している場合。

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　棄却域の図解



仮説検定の基礎　p値の意味と解釈方法

p値　帰無仮説が真である、という前提のもとで観測されたデータまたはそれよりも極端なデータが得られる確率

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。

$$p\,value < 0.05$$

帰無仮説の下でデータが観測される可能性が低いことを示しており、帰無仮説を棄却する強い証拠があるとみなされる。

▶ **帰無仮説を棄却する**

$$p\,value \geqq 0.05$$

帰無仮説の下でデータが観測される可能性が高いことを示しており、帰無仮説を棄却する十分な証拠がないとみなされる

▶ **帰無仮説が正しいとも対立仮説が正しいとも言えない**

ポイント　p値は、帰無仮説を棄却するための証拠の強さを定量化する指標として用いられている。**帰無仮説が真という前提に立って**計算される確率。

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。

> **効果量　effect size**
>
> 施策や介入によって、母集団に現れると考えている差分の大きさ。

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。

▶**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）**

仮説検定の基礎　p値の意味と解釈方法

よくある誤解

p 値：「差がないと仮定した中で、標本から得られた差もしくはそれよりも大きな差が得られる確率」

上のような捉え方をすると、p値は
条件付き確率とも考えられる

$$p\,value = P(T > t|H_0)$$

⚠ 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」

仮説検定の基礎　仮説検定の流れ

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：$t = \dfrac{\bar{x_1} - \bar{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

④棄却限界値との比較

- 検定統計量が棄却限界値を超えていれば、帰無仮説を棄却する。
- 超えなければ、帰無仮説を棄却しない。

⑤結論を導出

- 帰無仮説を棄却できれば、対立仮説が支持されることになる

▶ **どう作るのか？**

仮説検定の基礎　検定統計量の定義

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

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

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リングする

分子

$$\overline{x} - \mu_0$$

帰無仮説が正しい場合における「ずれ」を示す。この差が大きいほど、帰無仮説が正しくない可能性が高まる。

分母

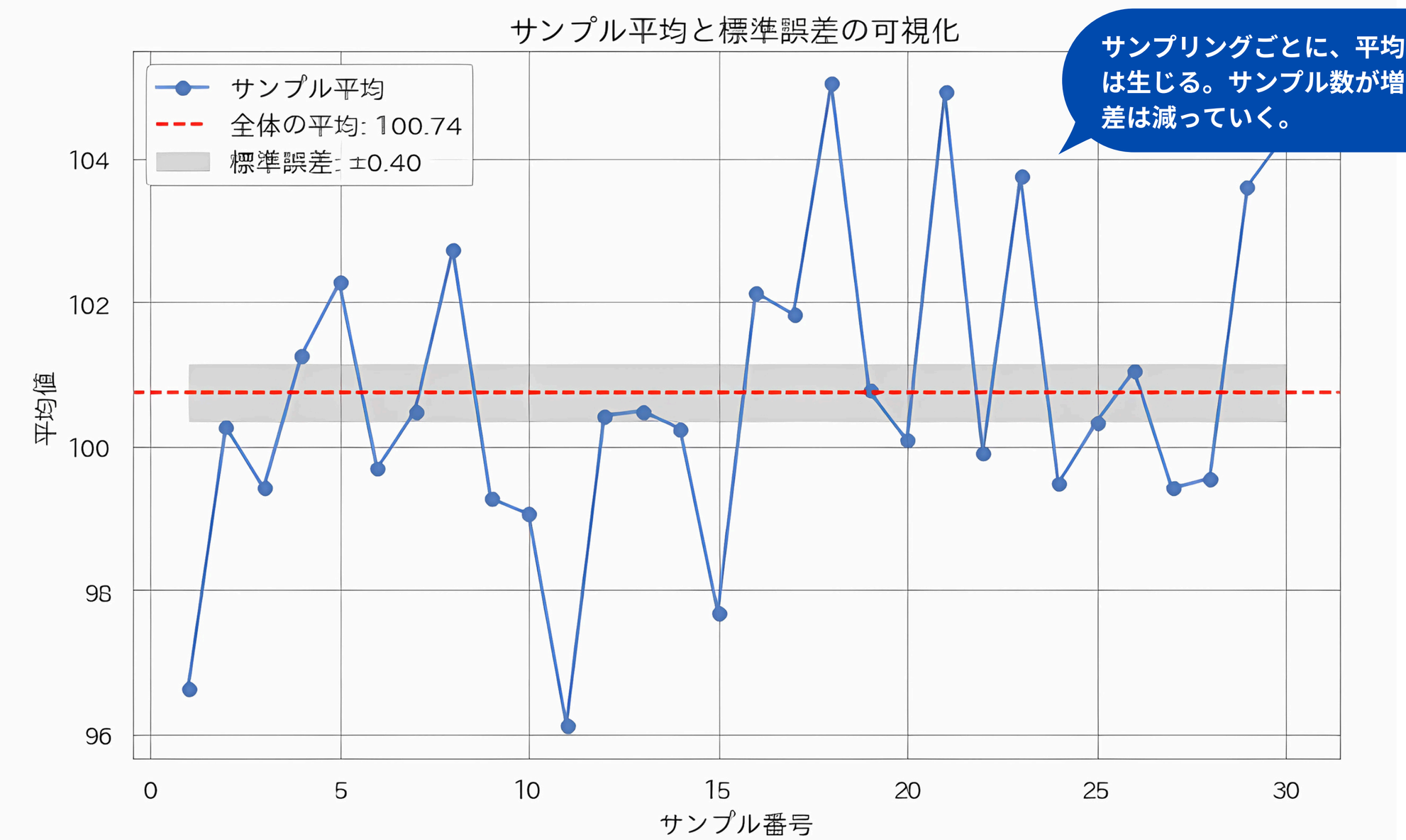
$$SE = \frac{\sigma}{\sqrt{n}}$$

標準誤差は、サンプルごとのばらつきを表す。

ポイント

検定統計量は、帰無仮説と実データの差を表現。分母の標準誤差(standard error)で、**サンプルサイズやデータのばらつきをスケーリング**する。

仮説検定の基礎　標準誤差の図解



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。

点推定と区間推定

特徴と具体例

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

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

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

$$\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \hat{\mu} \leq \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

zは標準正規分布の95%信頼区間に対応する値。今回は約1.96となる

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。

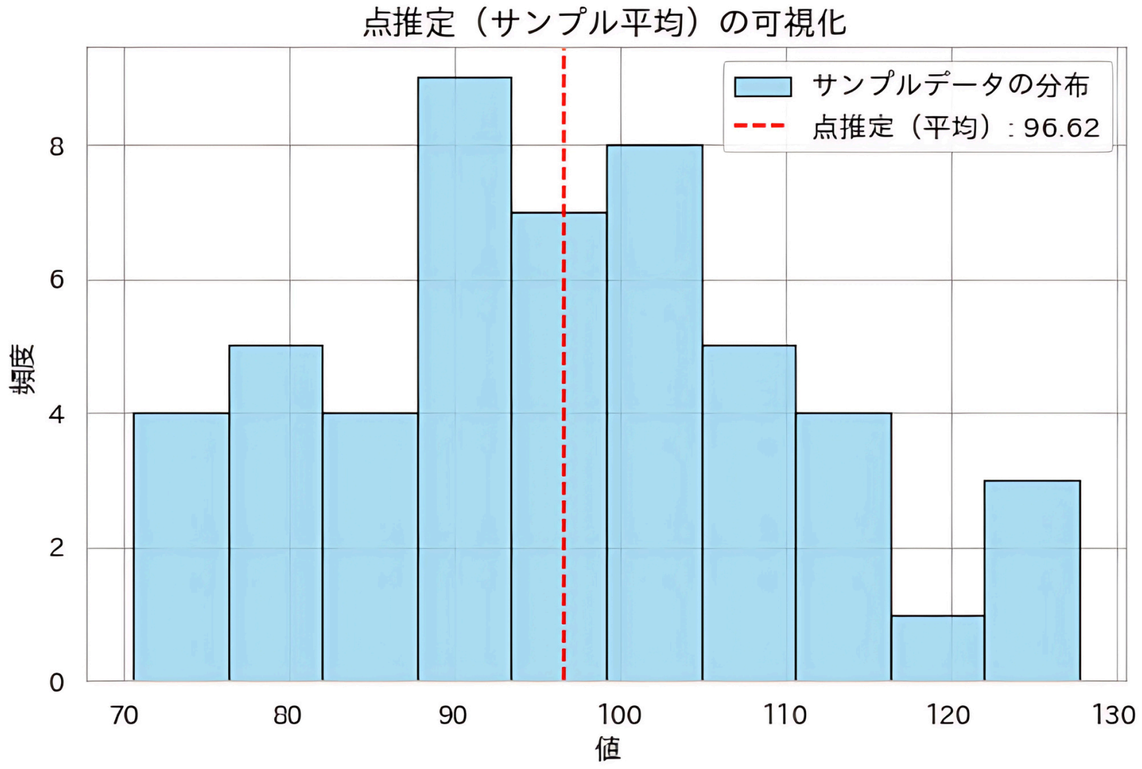
標準正規分布表（左側累積確率）

z（整数部分+小数第1位） \ 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
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。

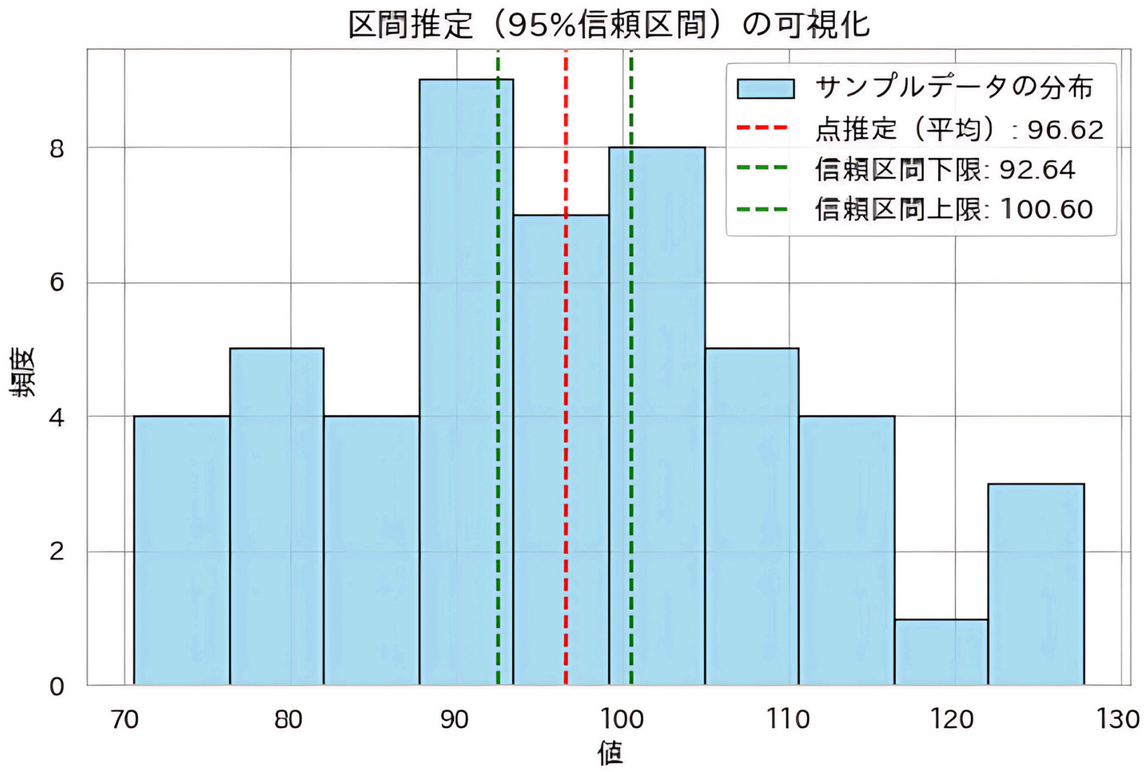
点推定と区間推定　図解

点推定



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。



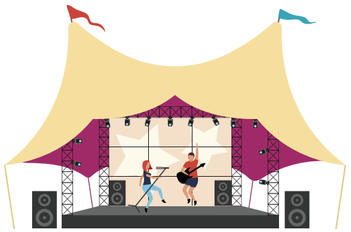
点推定と区間推定　具体例

信頼区間　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

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

具体例

来場者アンケートに答えてくれたカスタマーn人の満足度Xの平均から、今回のイベントの満足度の母平均を推定したい。



$$X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$$

母集団の分散はわからないので、標本分散から標本標準偏差sを算出して、$SE(\hat{X}) = \frac{s}{\sqrt{n}}$

有意水準0.05の時、95%信頼区間は　$\hat{X} \pm 1.96 \frac{s}{\sqrt{n}}$

点推定と区間推定　信頼区間の解釈

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。

✕ **「真の母集団パラメータが信頼区間に含まれる確率が95%」**

✓ **「もし同じ調査を何度も繰り返した場合、そのうち95%の信頼区間が母集団の真の値を含む」**

母平均の推定　パターン分類

前提

- 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**。

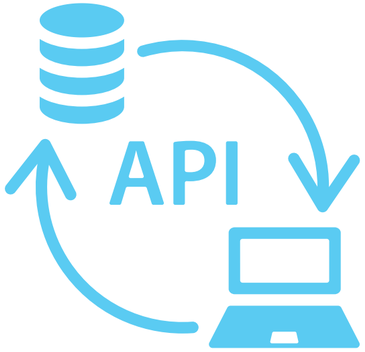
ケース	分散	点推定値	区間推定	使用する分布
母平均の推定	既知	$\overline{X}$	$\overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	標準正規分布
	未知	$\overline{X}$	$\overline{X} \pm t_{n-1} \cdot \frac{s}{\sqrt{n}}$	t分布
母平均の差の推定	既知	$\overline{X}_1 - \overline{X}_2$	$\overline{X} \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1}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

自由度の計算が複雑。別途補足。

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%**

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。

信頼区間を用いた母平均の推定

$$\overline{X} \pm t_{n-1} \cdot \frac{s}{\sqrt{n}}$$

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

$210 \pm 2.01 \cdot \frac{20}{\sqrt{50}}$ より、[204.32ms,215.68ms]となり、**APIの応答時間が目標値以下とは言い切れない。**

母平均の検定

検定の目的と検定統計量の計算方法

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)

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$ **（両側検定）** $$\mu < \mu_0$$ **（左側検定）** または $$\mu > \mu_0$$ **（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

$$t = \frac{\bar{X} - \mu_0}{s / \sqrt{n}} \sim t_{\alpha/2, n-1}$$

母分散が未知で、標本分散を使用する場合に使用。自由度に応じ、分布が変化する。

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。

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）

母平均の検定　t分布

t分布

母集団の分散が未知で**サンプルサイズが小さい場合に利用される確率分布**。
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。

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

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

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。

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

自由度v

制約条件がなければ、データが自由に取れる値の数。
制約は、基本的に平均や分散、合計など計算に利用される統計量の数になる。

t分布の自由度

$$v=n-1$$

通常、標本平均を制約とするため、サンプルから1除いた数が自由度となる。

分散未知の二標本検定の場合

$$v=(n_A+n_B)-2$$

2群の標本平均が制約になるため、制約数は２になる。

回帰分析の自由度

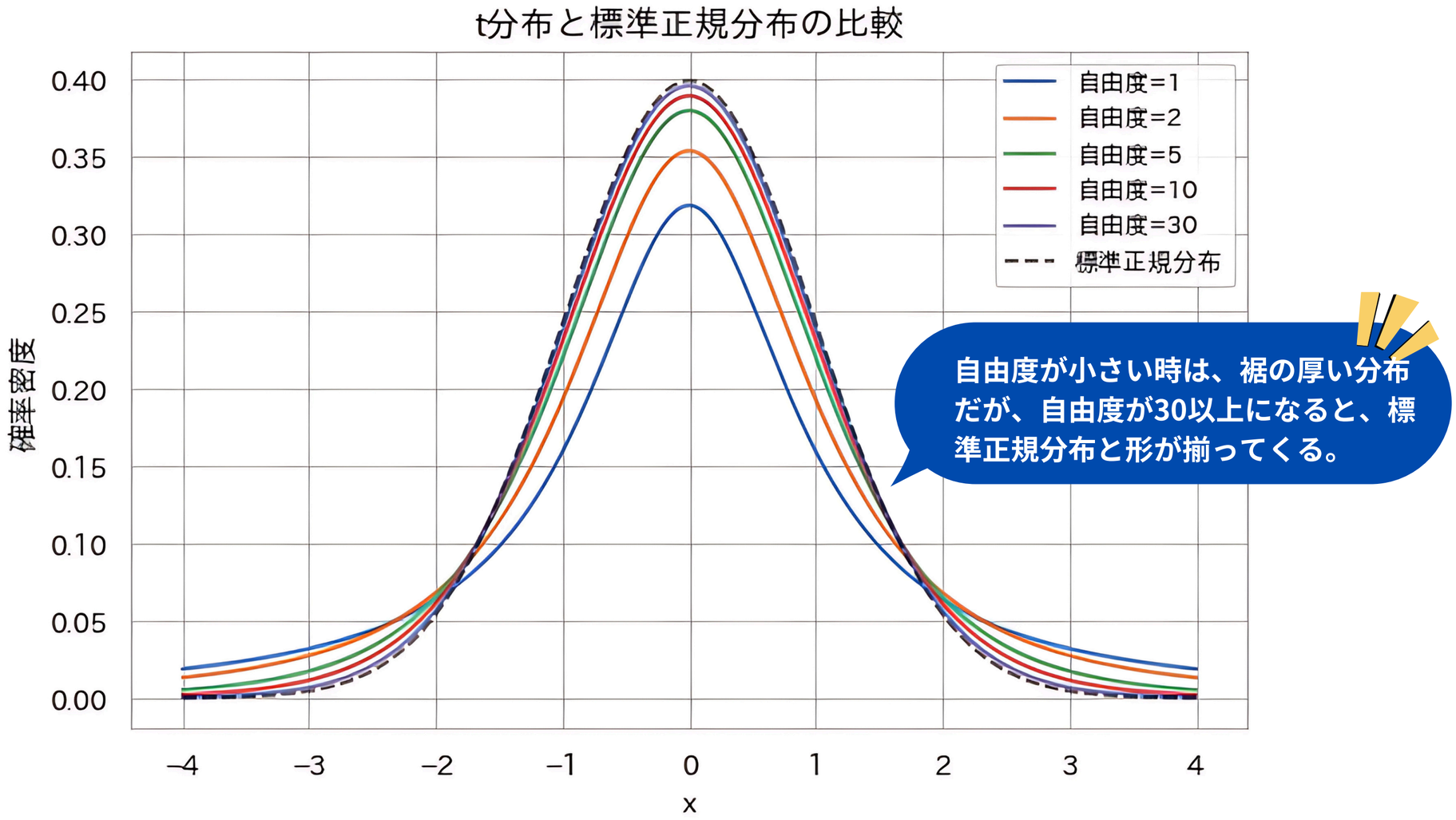
$$v=n-k$$

観測するパラメータの数kによって自由度が変動する。

母平均の検定　t分布

なぜt分布を使うのか

標準正規分布よりも裾が厚い、つまり外れ値が出やすい分布なので、小さなサンプルサイズでの統計量の計算に適している。



自由度が小さい時は、裾の厚い分布だが、自由度が30以上になると、標準正規分布と形が揃ってくる。

母平均の推定と検定　t分布

t分布表　t が自由度mの t 分布に従うとしたとき、t の上側 100α%点を与える。
すなわち、P(t ≧u)=α となる u を与える。

t分布表（自由度1～30）

自由度	片側 α=0.1, 両側 α=0.2	片側 α=0.05, 両側 α=0.1	片側 α=0.01, 両側 α=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)

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{（差がない）}$$

$$H_1 : \mu_1 - \mu_2 \neq 0 \quad \text{（差がある）}$$

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。

分散既知の場合

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}} \sim N(0,1)$$

差の差をとり、標準誤差で割る。

分散未知＆等分散の場合

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t_{\alpha/2, n_1+n_2-2}$$

プールした分散を使い、通常のt検定を行う。自由度は両群の平均がわかるので-2

分散未知＆異分散の場合

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} \sim t_{\alpha/2, df}$$

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。

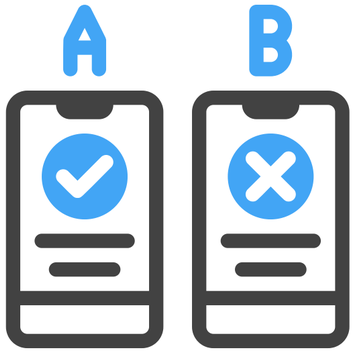
ポイント

- 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。

母平均の差の検定　等分散の場合

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。



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

$$\mu_B = \mu_A$$

新UIと旧UIの平均応答時間に差はない

対立仮説

$$\mu_B < \mu_A$$

新UIの平均応答時間が旧UIよりも短い

▶ 片側検定を行う

母平均の差の検定　等分散の場合

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。

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

$$t = \frac{\overline{n}_B - \overline{n}_A}{s_p\sqrt{\frac{1}{n_B} + \frac{1}{n_A}}}$$

プールした標準偏差は、

$$s_p = \sqrt{\frac{(30_A - 1)25^2 - (30 - 1)20^2}{30 + 30 - 2}} = \sqrt{512.5} \approx 22.65$$

よって、検定統計量は

$$t = \frac{200 - 220}{22.65\sqrt{\frac{1}{30} + \frac{1}{30}}} \approx -3.43$$

自由度は、30+30-2=58で、有意水準は5%なので、臨海値**は-1.67程度**

▶ 検定統計量 t は臨界値 よりも小さいので、帰無仮説を棄却する。

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

プールした標準偏差
両群の標準偏差にサンプル数で重みをつける。

結論

統計的に有意な結果として、**新UIの平均応答時間は旧UIよりも短い** と結論付け、新UIを正式導入する意思決定を下した。

母平均の差の検定　異分散の場合

ウェルチのt検定

母分散が未知かつ異なる場合に、2つの母平均の差を検定するための方法。
特に、**標本サイズや分散が大きく異なる場合に適している**。この検定の自由度は特別な計算を要する。

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

$$\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}$$

標本間の分散の影響を考慮した
「全体の標準誤差の分散」

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。

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}}\sim t_{\alpha/2,df}$$

▶ 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

母比率の推定　二項分布の正規近似

二項分布と比率

比率は、「成功数 / 試行回数」のような割合のこと。この**比率を得るための試行は通常、二項分布に従う**。具体的には、各試行が成功か失敗かのいずれかであり、成功の確率が一定のとき、その試行は二項分布に従い、成功数Xは以下のように表記できる。

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

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。

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。

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

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

分子は、標本比率の分散。

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）**

母比率の推定　パターン分類

前提

標本比率が与えられている場合、二項分布の分散を使って近似する。

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

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

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$$

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

$$z = \frac{\hat{p} - 0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

母比率の近似値である、標本比率で標準偏差を作る。

母比率の検定

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

帰無仮説では、母比率が所与の場合、検定統計量の標準誤差には母比率を使う

やり方を覚えるだけだと、全く気にしない部分ではあるので、注意が必要

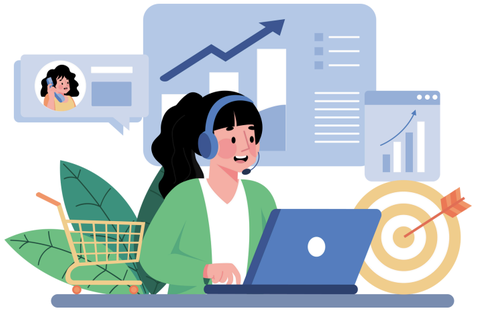
検定では**「仮定された母比率」が所与として与えられている**ため、標準偏差の近似が不要。

なので、信頼区間のような幅を持たせる手続きなしに検定統計量と臨界値をストレートに比較できている。推定との大きな違い。

母比率の検定　具体例

具体例

営業チームが、電話営業を行い、成約率（成約数 / 訪問数）を改善したいと考えている。過去のデータから、これまでの成約率が20%だったとする。最近、新しい営業方法を試しており、この方法が従来の方法よりも成約率を向上させているかどうかを検定で確かめたい。



データ

- 標本比率: 100件の営業活動により、35件の成約が生まれた
- 標本サイズ（n): 100件

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。

仮説設定

帰無仮説

$$p = 0.2$$

新しい営業方法を使っても、成約率は従来と変わらない

対立仮説

$$p > 0.2$$

新しい営業方法を使った結果、成約率は従来より高くなった

▶ 片側検定を行う

母比率の検定　具体例

検定統計量の計算

新しい営業方法が効果的かどうかを確認するために、次の検定統計量を計算する。帰無仮説の下での成約率は0.20と仮定。

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

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（α=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。

> **分散の取り扱いはシンプル**
>
> 二項分布の正規近似であることから考えてもわかるとおり、母比率の検定では、分散は 比率p に基づいて計算され、「分散が未知か既知か」の議論は不要。

結論

新しい営業方法が従来の方法よりも有意に成約率を向上させたと言えるので、営業方法の全面的な切り替えの意思決定が行われた。

母比率の差の検定

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

帰無仮説では、両群の母比率が等しいことを仮定しているため、**両標本の母比率を結合した「共通母比率」**を計算する。
標本数nのうち、成功数がxとすると以下のようになる。

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

共通母比率を使って、検定統計量を作る。

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

母比率の差の検定　具体例

具体例

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。

データ

- 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

帰無仮説

$$p_A \leq p_B$$

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

母比率の差の検定　具体例

検定統計量の計算

帰無仮説の下では、2つの母集団の比率は等しいと仮定するそのため、全体の成功率の合成推定値は以下のように計算する

$$\hat{p}_A = \frac{120}{200} = 0.6, \hat{p}_B = \frac{140}{250} = 0.56, \hat{p} = \frac{120+140}{200+250} \approx 0.5778$$

検定統計量は以下になる

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(\frac{A}{n_A} + \frac{1}{n_B})}}$$

▶ $$z = \frac{0.6 - 0.56}{00.5778(1-0.5778)(\frac{1}{200} + \frac{1}{250})} \approx 0.856$$

Z=0.856 に対して、**片側検定の場合の p値は標準正規分布の上側確率を求めると、約0.195**になる。

▶ p値が 0.05 より大きいため、帰無仮説を棄却できない。

標本比率で見ると、営業手法Aの方が高いが、統計的に見ると有意な差があるとは言えない。

結論

この結果から、営業手法 Aの更新率が B よりも有意に高いとは言えない。標本比率の差分は、サンプル誤差で説明ができる小ささ。

ご紹介　X版「青の統計学」

リンク

特徴

- データサイエンスや生成AI関連の資料を発信
- Web版やnoteの最新投稿の通知
- 開発デモ等のやってみた系投稿

ご紹介　WEB版「青の統計学」

リンク

特徴

- 統計学や機械学習の各論を中心に記事投稿
- 実務に応用する上でのポイントなど
- 直近はマーケティング領域の統計的手法、やベイズ統計学のトピックが多め

ご紹介　note版「青の統計学」

note

[リンク]

特徴

- 統計検定のチートシートなど
- 3万字〜10万字の長めのテキストコンテンツ