【新卒研修】基礎統計学

株式会社ブレインパッド

2023年5月9日・10日

本研修の流れ

1 統計学
統計学の枠組みについて学びます

2 記述統計学
データを解釈する上で重要な記述統計学について学びます

3 確率と確率分布
推測統計学の基礎となる確率の概念について学びます

4 推測統計学
推定、検定などの推測統計学の手法について学びます

5 バイアス
データの解釈の上で注意すべきバイアスについて学びます

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

1. 統計学

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

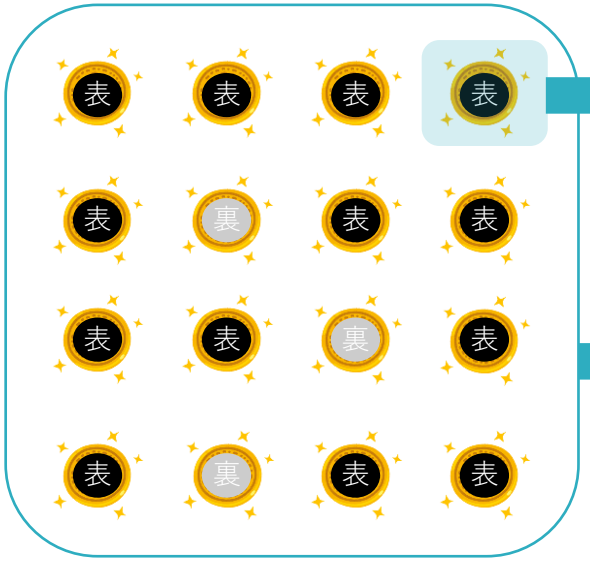
5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

1-1. 統計学とは

統計学とはデータから妥当な結論を導くための論理体系

統計学は不確実性を持ったデータを理解するための方法を与えてくれる。
単一のデータからは何も言えなくても、データを集めることにより、統計学を用いた解釈が可能になる。

例 コインに歪みがないかの検証



1つのデータのみから妥当な結論を導くことは困難

複数のデータを集めると、統計学を用いて
仮説の妥当性の検証が可能になる

1-2. 統計学を学ぶ意義

統計学は客観的な意思決定に活用できる

例えば、統計学は施策の優劣を客観的に判断するための材料として活用できる。

例 Webページの構成の比較

A

購入率：**19**%

B

購入率：**14**%

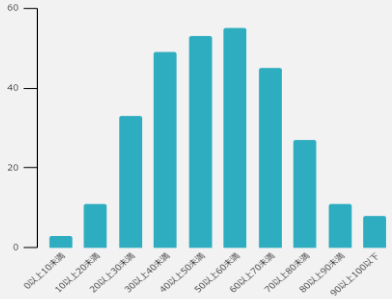
Aの方が購入率が高いが、その差に意味があるかを判断するために統計学の知識を活用し、
客観的な意思決定を支援する。

1-3. 統計学の種類

統計学は記述統計学と推測統計学の2つに大別できる

記述統計学

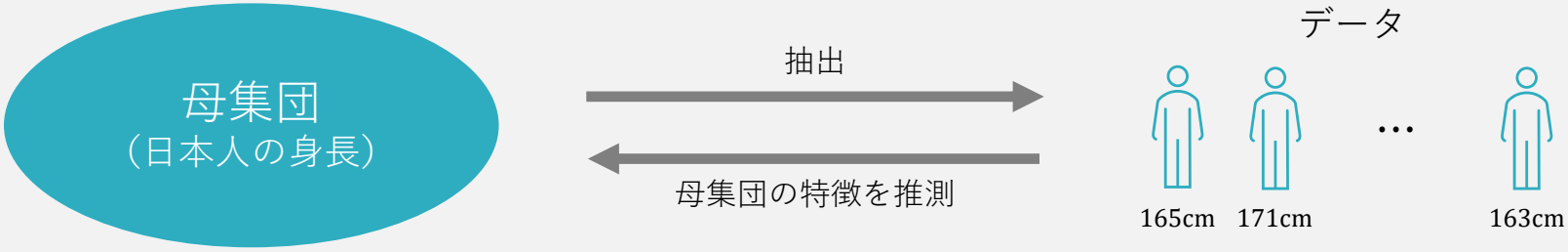
得られたデータをより深く解釈するための整理・要約の方法



平均	49
分散	365.3
標準偏差	19.1
中央値	49
第一四分位点	34
第三四分位点	62

推測統計学

興味の対象である母集団から得た一部のデータから全体の母集団を推測する方法



1-4. データの種類

データはカテゴリ値である質的変数と数量である量的変数に区別され、更に尺度ごとに分類できる

区分	尺度	解釈	例
質的変数	名義尺度	値が同じかどうかのみが意味を持つ	性別（男性、女性）
	順序尺度	値の順序が意味を持つ	成績評価（優、良、可）
量的変数	間隔尺度	値の間隔は意味を持つが比率は意味を持たない ※ 原点0は相対的な意味しか持たない	摂氏での気温 ※ 気温において、10°Cは1°Cの10倍暑いという表現はしない。つまり、比率に意味がない。
	比例尺度	値の間隔、比率ともに意味を持つ ※ 原点0が絶対的な意味を持つ	身長、体重、年齢 ※ 体重において、20kgは10kgの2倍重いという表現ができる。つまり、比率に意味がある。

1. まとめ

- 統計学はデータから妥当な結論を導く論理体系であり、客観的な意思決定に活用できる

- 統計学は次の２つに大別できる
  - 記述統計学：データの整理・要約する方法
  - 推測統計学：データを生成する背後の母集団について推測する方法

- データはその種類に応じて質的変数や量的変数に分類される

2. 記述統計学

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

2-1. 記述統計学とは

得られたデータの特徴を整理・要約するための方法

データ

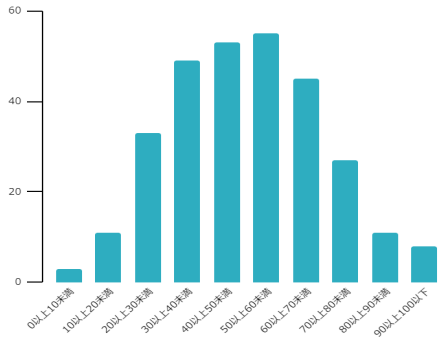
番号	国語の点数
1	56
2	44
3	36
4	63
5	28
6	51
7	74
8	30
9	62
10	65
⋮	⋮
294	44
295	72

整理・要約

度数分布表

階級	度数	相対度数	累積相対度数
0以上10未満	3	0.20	0.20
10以上20未満	11	0.07	0.27
20以上30未満	33	0.07	0.33
30以上40未満	49	0.00	0.33
40以上50未満	53	0.20	0.53
50以上60未満	55	0.20	0.73
60以上70未満	45	0.07	0.80
70以上80未満	27	0.00	0.80
80以上90未満	11	0.13	0.93
90以上100以下	8	0.07	1.00

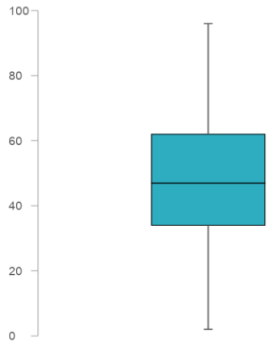
ヒストグラム



代表値の算出

平均	49
分散	365.3
標準偏差	19.1
中央値	49
第一四分位点	34
第三四分位点	62

箱ひげ図



2-2. 1変数データの記述｜度数分布表

度数分布表によりデータの概観を把握できる

度数分布表

データを複数の区間に分割し、各区間にどれほどデータがあるかをまとめた表

データ

番号	国語の点数
1	56
2	44
3	36
4	63
5	28
6	51
7	74
8	30
9	62
10	65
⋮	⋮
294	44
295	72

度数分布表

階級	度数	相対度数	累積相対度数
0以上10未満	3	0.20	0.20
10以上20未満	11	0.07	0.27
20以上30未満	33	0.07	0.33
30以上40未満	49	0.00	0.33
40以上50未満	53	0.20	0.53
50以上60未満	55	0.20	0.73
60以上70未満	45	0.07	0.80
70以上80未満	27	0.00	0.80
80以上90未満	11	0.13	0.93
90以上100以下	8	0.07	1.00

- データの概観を把握できる

- 度数　　：各階級に含まれるデータの数
- 相対度数　　：各度数が全体に占める割合
- 累積相対度数：相対度数を累積したもの

2-2. 1変数データの記述｜ヒストグラム

ヒストグラムによりデータの分布の傾向を把握できる

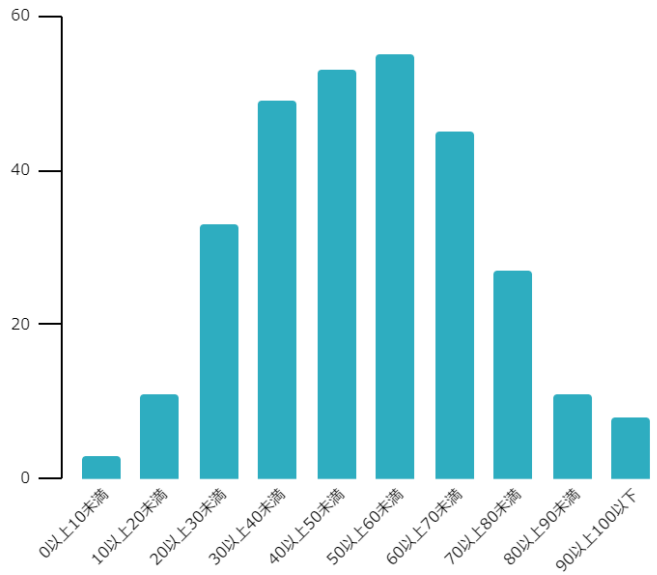
ヒストグラム

度数分布表を可視化したもの

度数分布表

階級	度数
0以上10未満	3
10以上20未満	11
20以上30未満	33
30以上40未満	49
40以上50未満	53
50以上60未満	55
60以上70未満	45
70以上80未満	27
80以上90未満	11
90以上100以下	8

ヒストグラム



- データの分布の傾向を把握できる
- 後述の確率分布に通じてくる

2-2. 1変数データの記述｜要約統計量

要約統計量によりデータを定量的に把握できる

代表値	意味	数式
平均	データの重心	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
分散	データの散らばりの程度	$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$
標準偏差	データの散らばりの程度	$s = \sqrt{s^2}$
最小値	データの中で最も小さい値	—
最大値	データの中で最も大きい値	—
中央値	データを昇順に並べた時に中央にくる値	—
第一四分位点	データを昇順に並べたときに前から25%にくる値	—
第三四分位点	データを昇順に並べたときに前から75%にくる値	—
最頻値	データの中で最も多い度数を示す値	—

2-2. 1変数データの記述｜箱ひげ図

箱ひげ図によりデータのばらつきを視覚的に把握できる

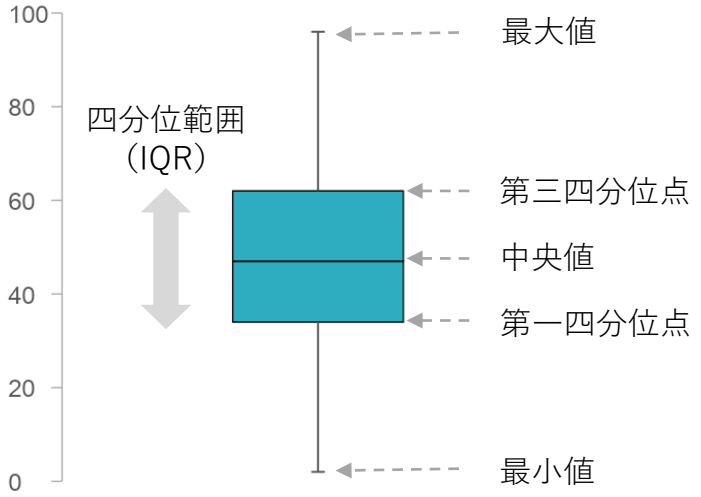
箱ひげ図

データの分位点（最大値、最小値、中央値、第一四分位数、第三四分位数）を可視化したグラフ

要約統計量

最大値	96
第三四分位点	62
中央値	49
第一四分位点	34
最小値	2

箱ひげ図



- データの散らばりを視覚的に把握できる
- 他のデータと分布の比較を容易に行える

* ヒゲの上端を、（第三四分位点 + 1.5 × IQR ）より小さい最大値、下端を（第一四分位点 + 1.5 × IQR ）より大きい最小値で表し、ヒゲの外側に存在するデータ点を「外れ値」としてプロットする場合もある。

散布図により2つの変数の間の関係性を把握できる

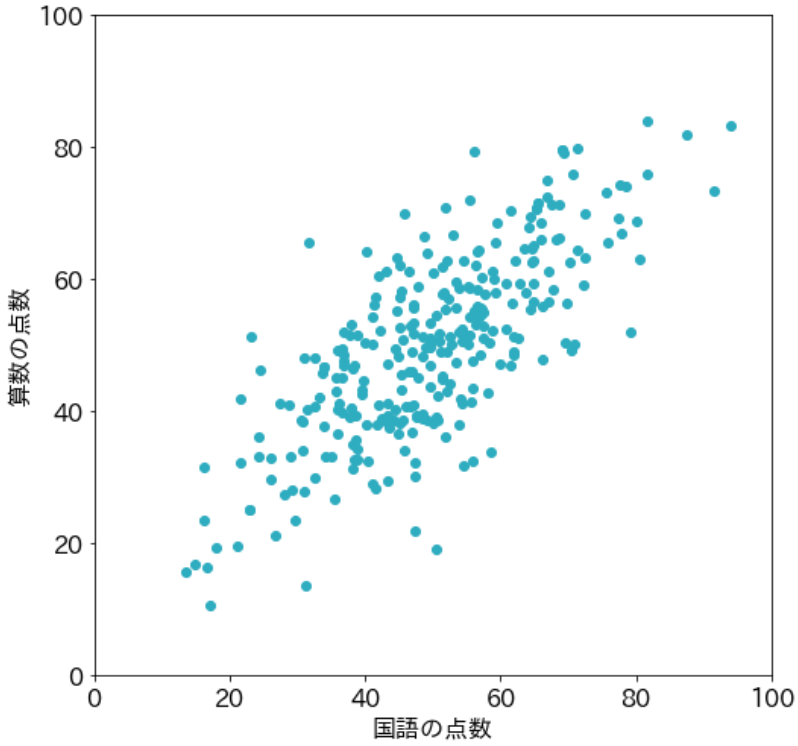
散布図

縦軸、横軸に異なる変数を対応させ、各データ点をプロットしたグラフ

データ

番号	国語の点数	算数の点数
1	56	39
2	44	44
3	36	26
4	63	53
5	28	31
6	51	49
7	74	66
8	30	39
9	62	73
10	65	71
:	:	:
294	44	39
295	72	65

散布図



- 2つの変数の関係性を視覚的に把握できる

2-3. 2変数データの記述｜共分散

2つの変数の間の関係性の強さを表す量として共分散がある

共分散

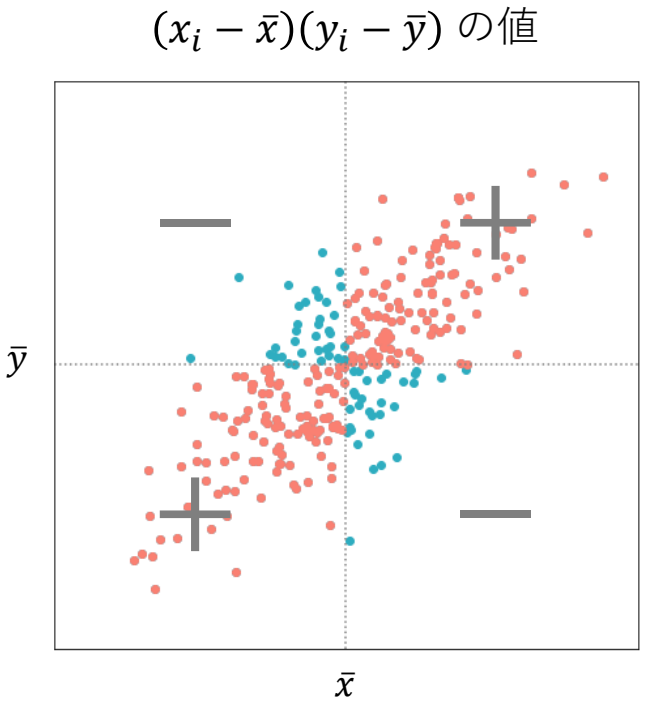
データの関係性の強さを表した量

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

- xが増加するほどyも増加するという関係のとき
  共分散は正の値を取る
- xが増加するほどyが減少するという関係のとき
  共分散は負の値を取る

共分散はデータのスケールに依存する

▶ 定量的な関係の把握のためには相関係数を用いる



2-3. 2変数データの記述｜相関係数

相関係数により2変数の関係性の強さを定量的に把握できる

相関係数

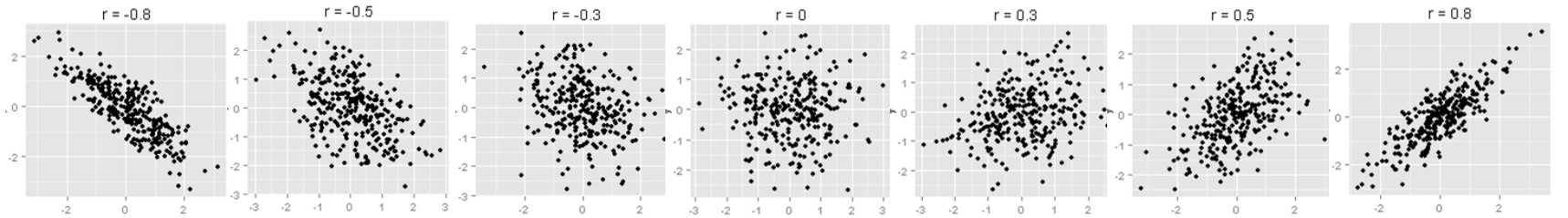
共分散がデータのスケールに依存しないように、それぞれの標準偏差で除した量

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 相関係数は-1から1の間の値を取る（1ほど正の相関が強く、-1ほど負の相関が強い）

※ 相関係数は2つの変数間の線形関係の強さを表す指標

▶ 非線形な関係性は実際に散布図を見て確認することが重要



負の相関が強い　　正の相関が強い

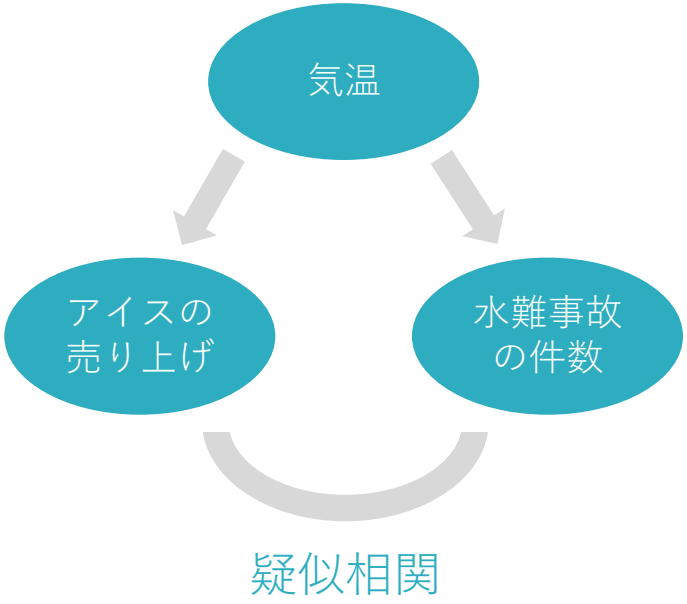
2-4. 相関係数の解釈上の注意｜相関と因果

相関関係は因果関係を意味するとは限らない

相関係数はあくまで、２つの変数を観察したときの関係の強さを測る指標。
相関が大きいことは、必ずしも変数の間に因果関係があることを意味しない。

例 アイスの売り上げと水難事故

アイスの売り上げが伸びると、水難事故の件数も増える。
このことから、アイスが水難事故の原因と推測するのは誤り。
実際は、気温の高さが共通の原因になっていると考えられる。

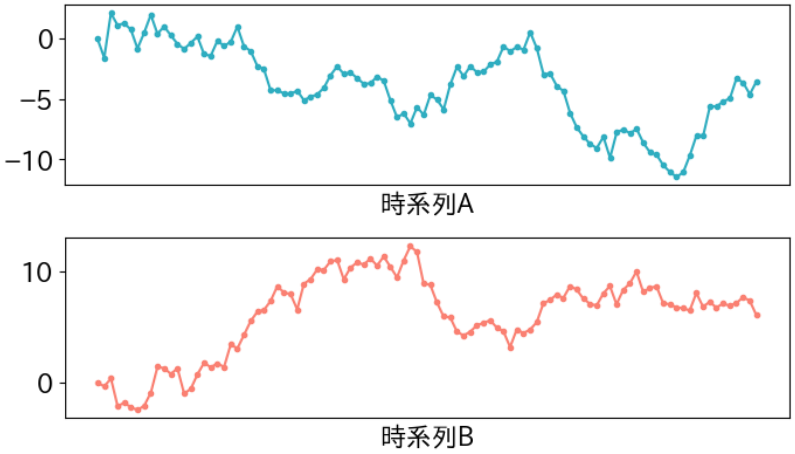


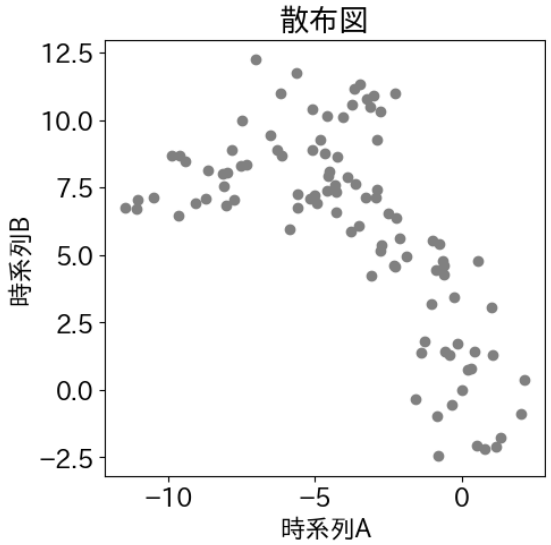
このように、２つの変数に因果関係が無いにも関わらず、背後にある要因によって相関係数が高くなる現象を疑似相関という。

参考：[【統計用語】疑似相関とは - AI Academy Media] https://aiacademy.jp/media/?p=3318

2-4. 相関係数の解釈上の注意｜見せかけの回帰

無関係な時系列データについて相関が高くなることがある





一見２つの系列には負の相関があるように見えるが、実はこれらは全く無関係にランダムに生成した時系列である。

このように、ある特定の時系列データ*に対しては、全く無関係でも相関が高くなる現象を見せかけの回帰という。

時系列データについて相関を確認するときは注意を要する。

* 具体的には、単位根過程と呼ばれる時系列である。
** 見せかけの回帰についてより詳しくは、例えば「経済・ファイナンスデータの計量時系列分析」などを参照してほしい。

2. まとめ

- 記述統計学はデータを整理・要約するための方法である
- データの表現方法には度数分布表、ヒストグラム、箱ひげ図、散布図がある
- １変数データを要約した量としては平均、分散、中央値などがある
- ２変数データの関係性の強さを表す量として共分散、相関係数がある
- 相関関係を因果関係と混同しない、また時系列間で相関を取るときは注意する

3. 確率と確率分布

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

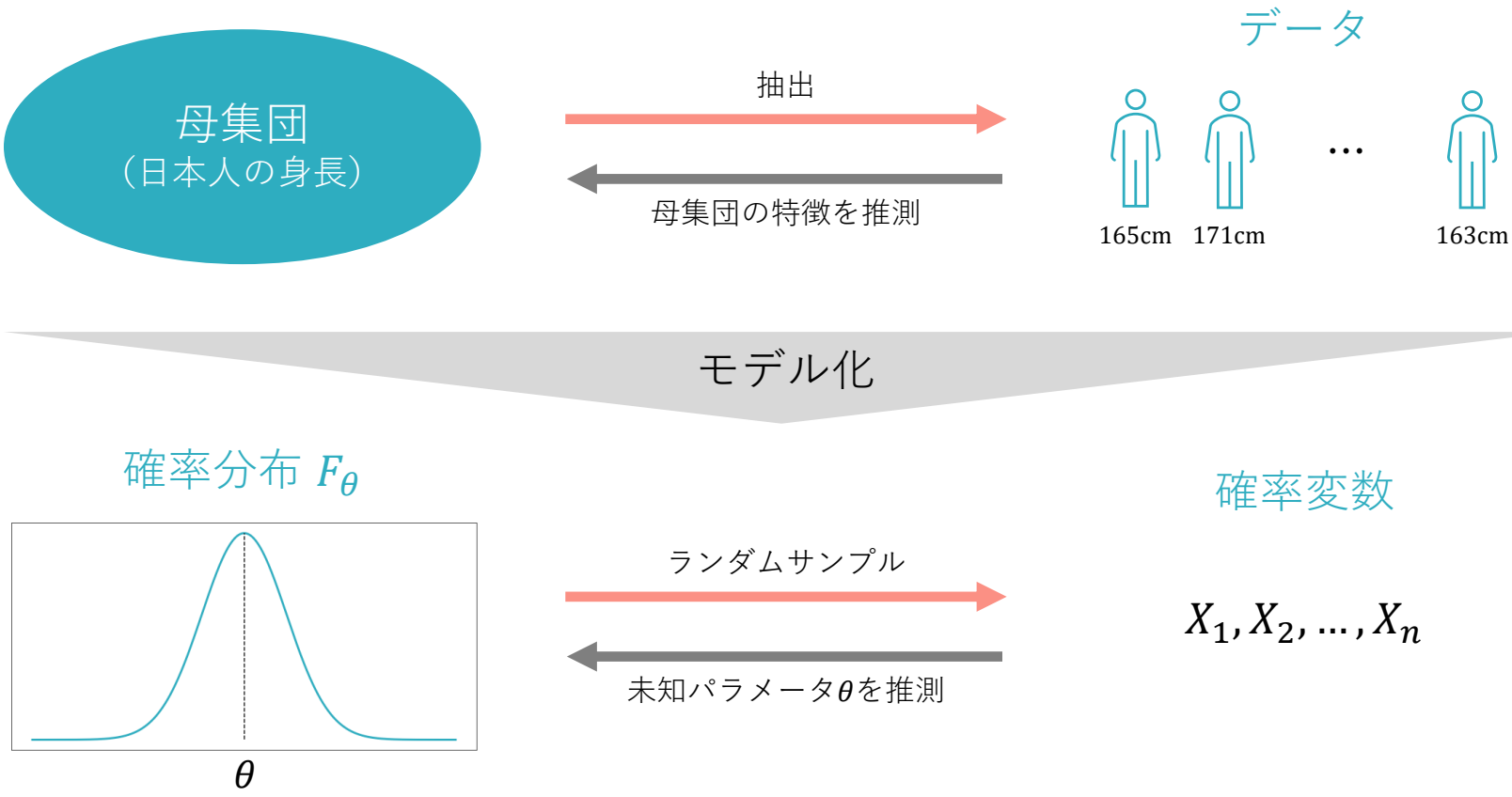
3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

3-1. なぜ確率を学ぶのか

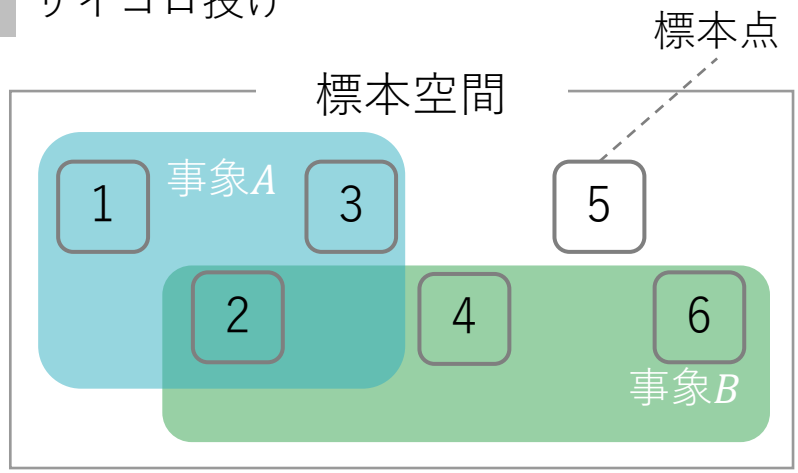
推測統計学では、確率的な概念を利用して母集団やそこから得られるデータをモデル化する



確率とは各イベントの相対的な起こりやすさを表す量である

- 標本空間：全ての起こり得る結果を集めたもの
- 標本点　：起こり得る結果の単位
- 事象　　：起こり得る結果の集まり（イベント）
- 確率　　：事象の相対的な起こりやすさを表す量
  事象Aに対してその確率を$P(A)$と表す

例 サイコロ投げ



- 各標本点がそれぞれの出る目に対応
- 事象A：「3以下の目が出る」
- 事象B：「偶数の目が出る」
- 事象A, Bそれぞれの確率

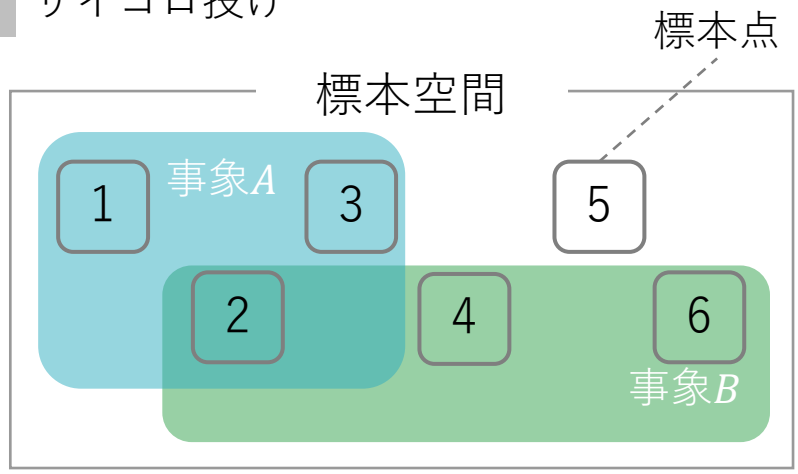
$$P(A) = P(B) = \frac{3}{6} = \frac{1}{2}$$

3-2. 確率｜事象の演算

事象に対しては以下の演算ができる

用語	説明	ベン図	サイコロの例
和事象 $A \cup B$	「AまたはBが起こる」という事象		$\{1, 2, 3, 4, 6\}$
積事象 $A \cap B$	「AかつBが起こる」という事象		$\{2\}$
余事象 A^c	「Aが起こらない」という事象		$\{4, 5, 6\}$
全事象 Ω	起こり得る全ての結果をまとめた事象		$\{1, 2, 3, 4, 5, 6\}$
空事象 $\emptyset$	存在しない事象	―	$\{\}$

例 サイコロ投げ



- 事象A：「3以下の目が出る」
- 事象B：「偶数の目が出る」

※ AとBの積事象が空事象であるとき２つは排反であるという（同時に起きない）

3-2. 確率｜確率の性質

確率は以下のようないくつかの性質を満たす

1. 任意の事象Aに対して次が成り立つ。

$$0 \leq P(A) \leq 1$$

2. 全事象Ωに対して次が成り立つ。

$$P(\Omega) = 1$$

3. 互いに排反な事象の列$A_1, A_2, \ldots$に対して次が成り立つ。

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

上の性質から以下のような基本的な性質が導かれる。

- $A \subset B$ ならば $P(A) \leq P(B)$
- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

* この講義では便宜上、確率の満たす性質として紹介しているが、数学的な立場ではこの3つの性質（確率の公理）を満たすようなものとして確率を定義する。

3-2. 確率｜条件付き確率

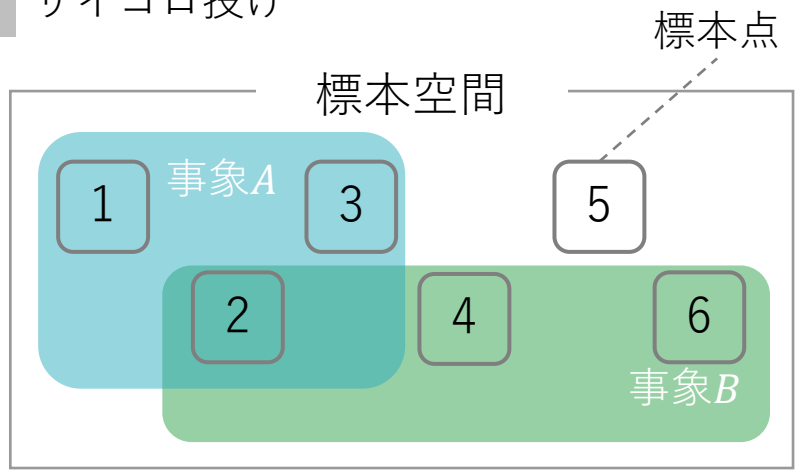
ある事象が起こったという条件の下で他の事象が起こる確率を条件付き確率という

条件付き確率

事象Bが起こった下での事象Aの起こる確率

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

例 サイコロ投げ



- 事象A：「3以下の目が出る」
- 事象B：「偶数の目が出る」
- Bが与えられた下でのAの条件付き確率

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}$$

3-2. 確率｜事象の独立性

ある事象が起こったことが別の事象が起こるかどうかについて情報を与えないとき、２つの事象は独立であるという

２つの事象AとBが独立であるとは、次を満たすこと。

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

これは、条件付き確率を用いると次のように書ける。

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

つまり、事象Bが起こったかどうかによって事象Aが起こる確率は変わらない。

例 サイコロ投げ

大小二つのサイコロを投げたとき、「大きいサイコロの目が偶数である」という事象は
「小さいサイコロの目が奇数である」という事象と独立。

取る値が確率的に決まる変数を確率変数、その値の取り方を確率分布という

例 サイコロ投げ

「2つのサイコロの出た目の和」を確率変数Xとすると、その確率分布は次の表で表される。

x	2	3	4	5	6	7	8	9	10	11	12
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

ある確率変数Xの分布がFであるとき、XはFに従うといい

$$X \sim F$$

と書く。

3-3. 確率変数｜離散型確率変数と連続型確率変数

確率変数は離散型と連続型に分類される

離散型確率変数

離散的な値を取る確率変数

例 サイコロの目、コインの裏表、事故の件数など

連続型確率変数

連続的な値を取る確率変数

例 気温、身長、体重など

離散型確率変数に対する分布は確率関数で表現される

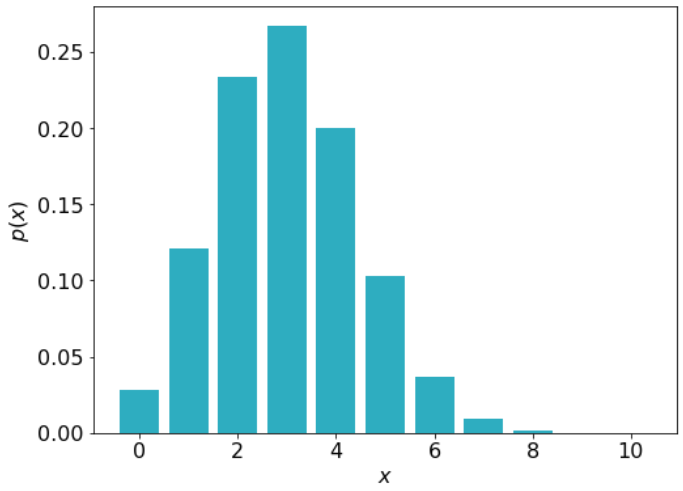
離散型確率変数Xに対して、次の関数を確率関数という。

$$p(x) = P(X = x)$$

確率関数は次の性質を満たす。

$$1.\quad 0 \leq p(x) \leq 1, \qquad 2.\quad \sum_{x} p(x) = 1$$

例 二項分布



連続型確率変数に対する分布は確率密度関数で表現される

連続型確率変数Xに対して、次の性質を満たす関数$f(x)$を確率密度関数という。

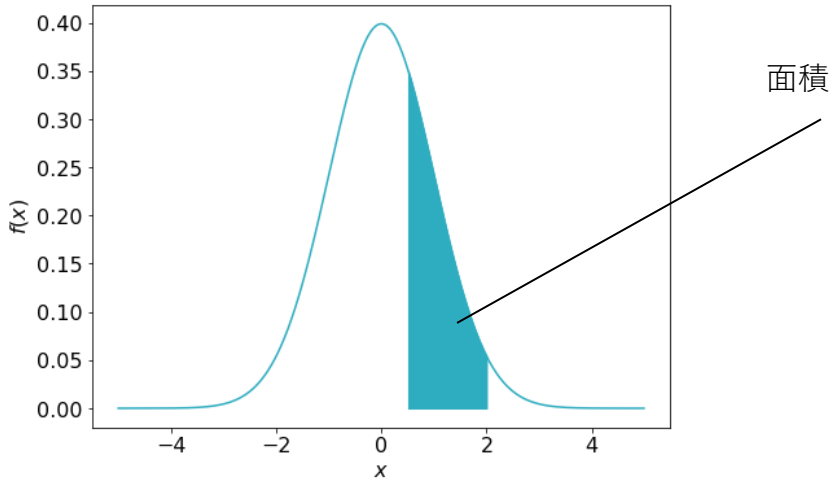
$$P(a \leq X < b) = \int_a^b f(x)dx$$

※ 連続型確率変数ではある1点の値を取る確率は必ず0になるため、このような定義が必要となる。

確率密度関数は次の性質を満たす。

$$1.\quad f(x) \geq 0, \qquad 2.\quad \int_{-\infty}^{\infty} f(x)dx = 1$$

例 正規分布



3-3. 確率変数｜同時分布と周辺分布

複数の確率変数に対する分布も考えられる

同時分布：2つの確率変数を合わせた分布

離散型（同時確率関数）　$p_{X,Y}(x,y) = P(X = x, Y = y)$

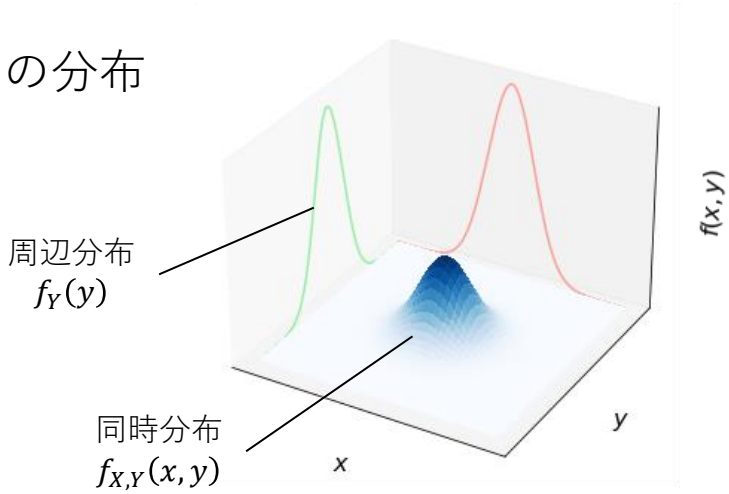
連続型（同時確率密度関数）：次の性質を満たす関数 $f_{X,Y}(x,y)$

$$P(a \leq X < b, c \leq Y < d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

周辺分布：片方の確率変数のみに着目したときの分布

離散型　$p_X(x) = \sum_y p_{X,Y}(x,y)$

連続型　$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$



多次元の時も同様の定義。

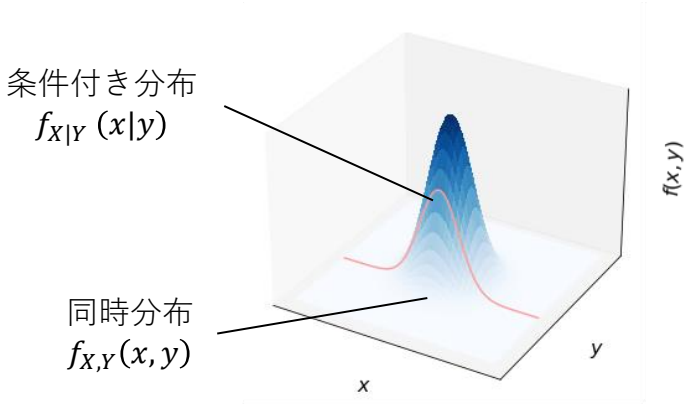
3-3. 確率変数｜条件付き分布と確率変数の独立性

事象と同様に確率変数の条件付き分布・独立性が考えられる

条件付き分布：片方の確率変数の値がわかった下での他方の確率変数の分布

離散型　$p_{X|Y}(x|y) = \dfrac{p_{X,Y}(x,y)}{p_Y(y)}$

連続型　$f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$



２つの確率変数XとYが独立であるとは、次を満たすこと。

離散型　$p_{X,Y}(x,y) = p_X(x)p_Y(y)$　　連続型　$f_{X,Y}(x,y) = f_X(x)f_Y(y)$

これは、条件付き確率を用いると次のように書ける。

離散型　$p_{X|Y}(x|y) = p_X(x)$　　連続型　$f_{X|Y}(x|y) = f_X(x)$

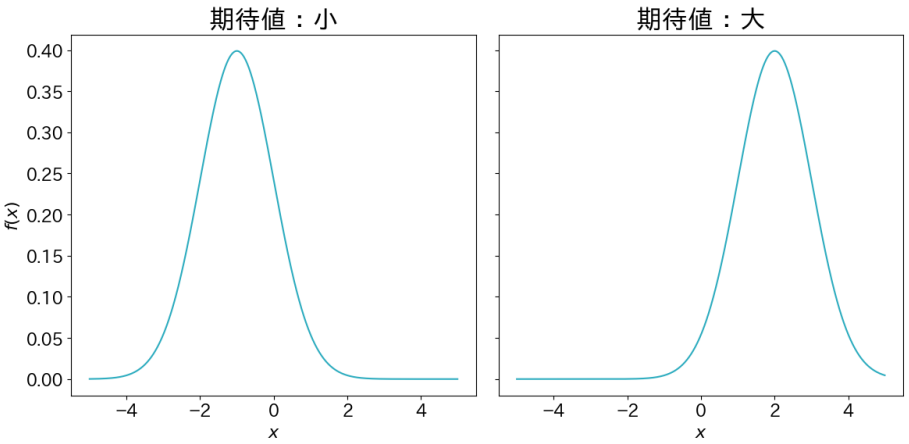
3-3. 確率変数｜期待値と分散

分布を特徴付ける量の一種として期待値、分散がある

期待値

分布の重心を表す（平均ともいう）

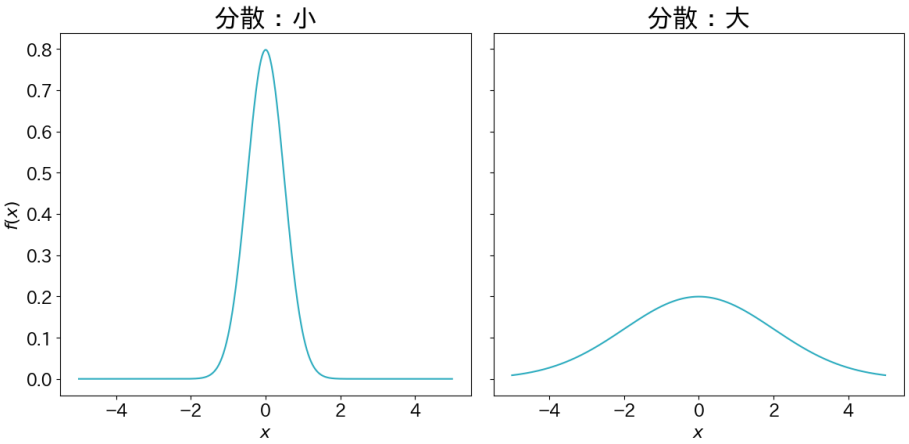
$$E[X] = \begin{cases} \sum_x xp(x) & \text{（離散型）} \\ \int xf(x)dx & \text{（連続型）} \end{cases}$$



分散

分布の散らばりを表す

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



3-3. 確率変数｜期待値と分散の性質

期待値、分散は以下の性質を満たす

期待値の性質

1. （期待値の線形性）

$$E[aX + bY] = aE[X] + bE[Y]$$

2. XとYが独立ならば次を満たす。

$$E[XY] = E[X]E[Y]$$

分散の性質

1. （線形変換に対する性質）

$$V[aX + b] = a^2V[X]$$

2. XとYが独立ならば次を満たす。

$$V[X + Y] = V[X] + V[Y]$$

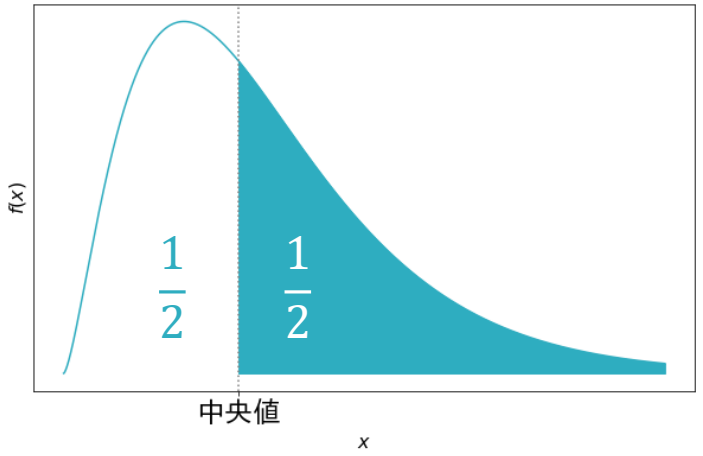
3-3. 確率変数｜中央値と上側α点

確率分布の中での相対的な位置を表す量として、中央値や上側α点がある

連続型確率変数Xに対し

$$P(X \leq x) = P(X > x) = \frac{1}{2}$$

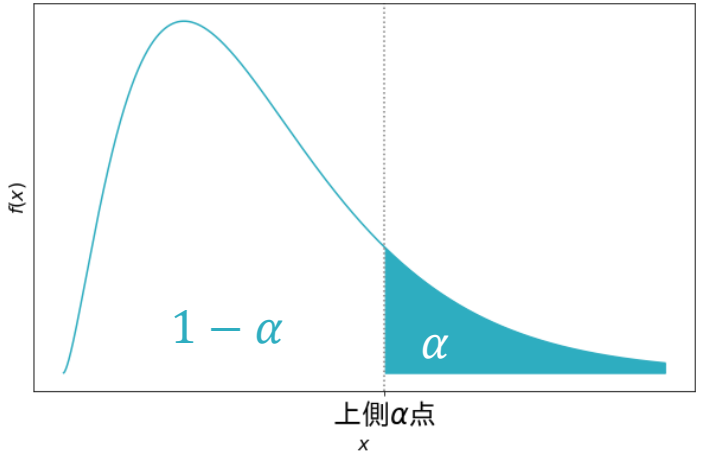
を満たす値xを分布の中央値という。



より一般に、連続型確率変数Xに対し

$$P(X > x_\alpha) = \alpha$$

を満たすような値x_αを上側α点という。



* 離散型確率変数については上記の性質を満たすような値が一意に定まらないため、より厳密な定義の仕方が必要となる。詳細については割愛する。

3-4. 代表的な確率分布

様々な分布を用いて現実の事象をモデル化することができる

以下では代表的な分布を紹介する。

離散分布

- 二項分布
- ポアソン分布
- 負の二項分布
- 幾何分布
- 超幾何分布
- 多項分布

連続分布

- 正規分布
- 指数分布
- ガンマ分布
- t分布
- χ^2分布
- F分布
- 一様分布

分布の集まりの中で、一つの分布を特徴づける量をパラメータ（母数）という。
パラメータの個数は分布の種類によって様々。

例 コイン投げ

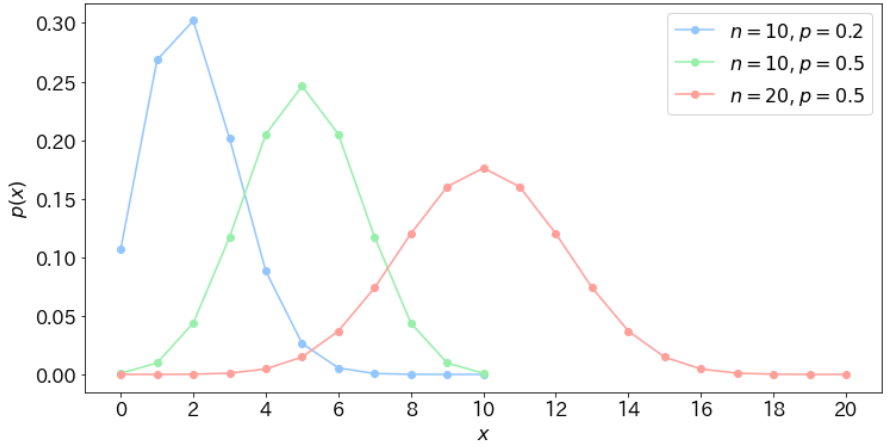
表が出る確率pが分布を特徴づけるパラメータ。

3-4. 代表的な確率分布｜二項分布（離散分布）

二項分布 $Bin(n, p)$

$$p(x) = {}_nC_x p^x (1-p)^{n-x} \quad (x = 0,1,\ldots,n)$$

母数	$0 \leq p \leq 1, n \geq 0$（整数）
平均	np
分散	$np(1-p)$



- 「表が出る確率がpであるコインをn枚投げたときに表が出る回数」が従う分布
- $n \to \infty, p \to 0$の極限でポアソン分布に近づく*
- 多次元に一般化したものを多項分布という

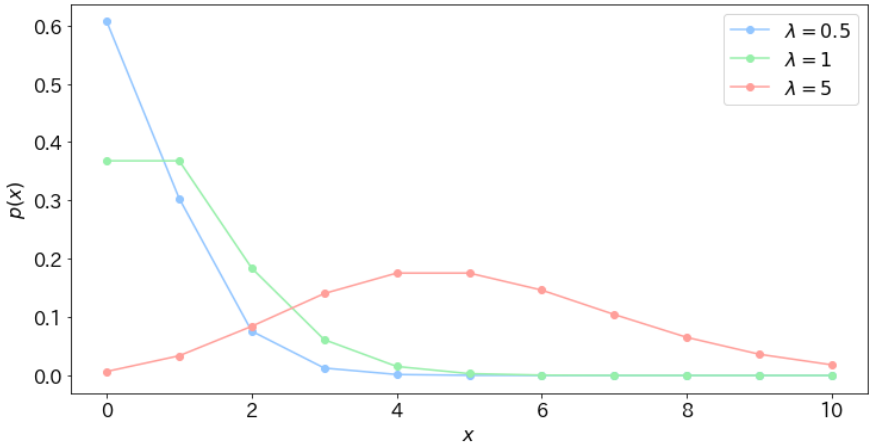
* より正確にはnpを一定に保ったままnを大きくしたときの極限を考える。

3-4. 代表的な確率分布｜ポアソン分布（離散分布）

ポアソン分布 $Po(\lambda)$

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0,1, \dots)$$

母数	$\lambda > 0$
平均	λ
分散	λ



- 「稀にしか起らないイベント」を大量に観測したとき、そのイベントの回数はポアソン分布に従う（典型的には事故の発生件数など）
- 二項分布で$n \to \infty, p \to 0$としたときの極限として得られる

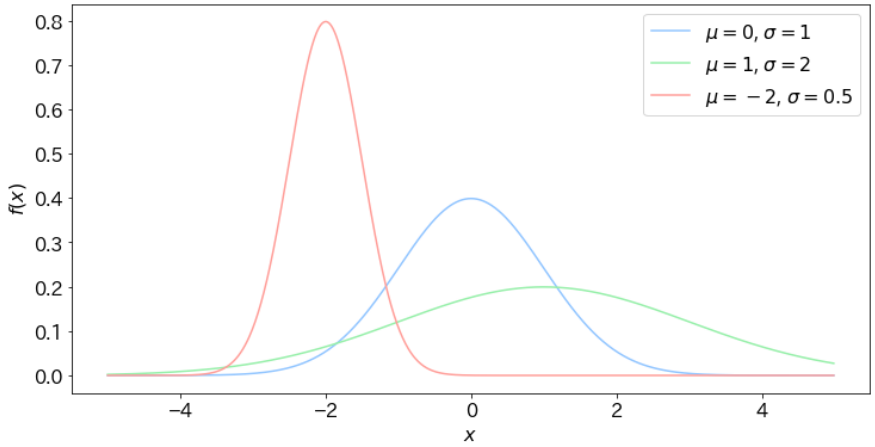
* より正確にはnpを一定に保ったままnを大きくしたときの極限を考える。

3-4. 代表的な確率分布｜正規分布（連続分布）

正規分布 $N(\mu, \sigma^2)$

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

母数	$\mu, \sigma > 0$
平均	μ
分散	σ^2



- 統計における最も基本的な分布
- 様々な不確かさを表現する分布としてよく用いられる（測定誤差など）
- 平均0、分散1の正規分布を標準正規分布と呼ぶ

3-5. 大数の法則と中心極限定理

大数の法則と中心極限定理は、サンプルを大きくしたときの標本平均の振る舞いを説明する

$X_1, \ldots, X_n$を平均μ, 分散σ^2の任意の同一の確率分布に独立に従う確率変数とする。

大数の法則

nを大きくすると標本平均 $\bar{X} = (X_1 + \cdots + X_n)/n$ は真の平均に近づく。

$$\bar{X} \to \mu$$

中心極限定理

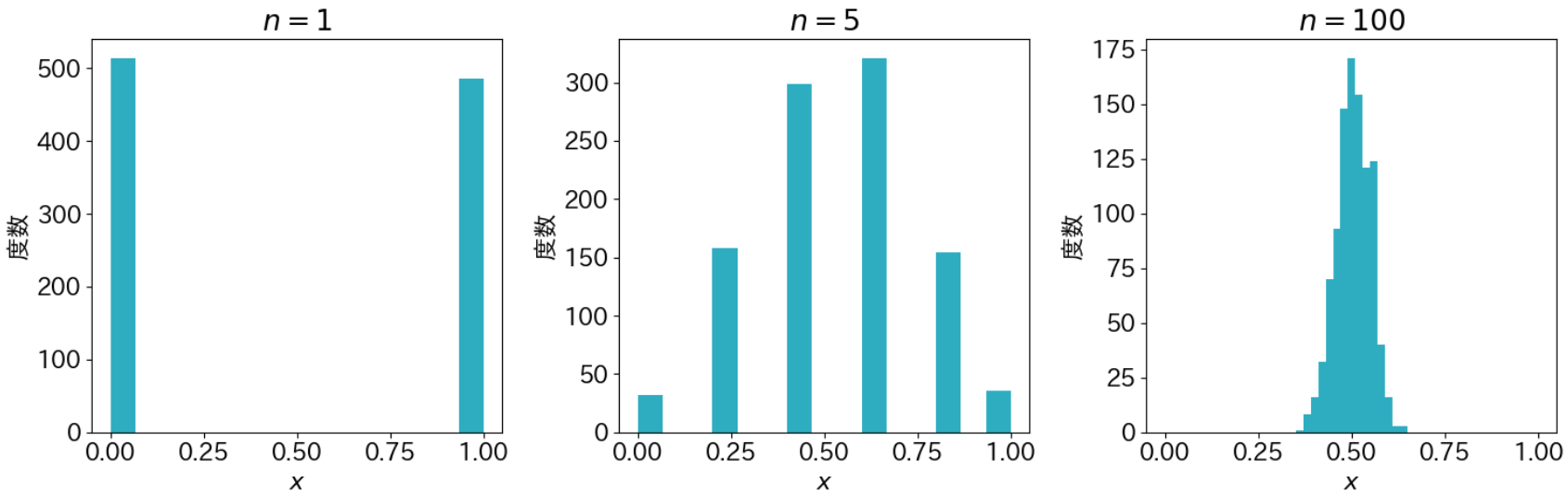
nを大きくすると標本平均は真の平均を中心とした正規分布に近づく。

$$\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2)$$

- いずれの定理も標本平均が真の平均に近づくことを意味するが、中心極限定理はその近づいた時の振る舞いをより詳細に教えてくれている
- 統計では正規分布に近似するテクニックを多用するが、それらの多くは中心極限定理に基づいている

* ここでいう「近づく」とはある意味での収束を意味し、その収束の意味は大数の法則、中心極限定理それぞれで異なる。気になる方は「確率変数の収束」で調べてみて下さい。

3-5. 大数の法則と中心極限定理｜イメージ



$$X_1, \ldots, X_n \sim p(x) \qquad p(x) = \begin{cases} 0.5 & (x = 0) \\ 0.5 & (x = 1) \end{cases}$$

という分布に独立に従う乱数から計算した標本平均$\bar{X}$のヒストグラム（$\bar{X}$は1000回繰り返し生成）

- nが十分大きくなると真の平均0.5の近くに値が集中する（大数の法則）
- また、その分布は正規分布の形に近づく（中心極限定理）

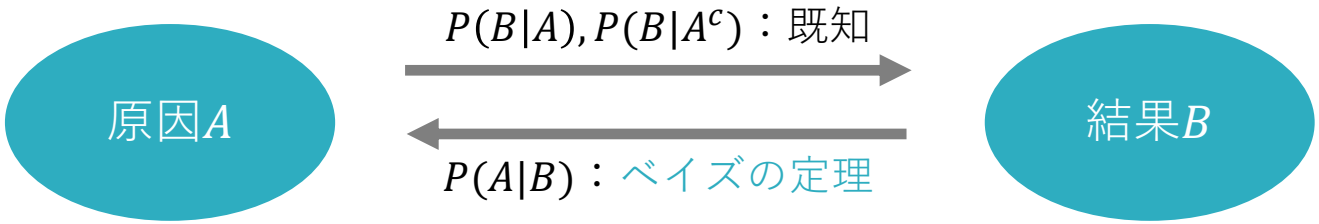
3-6. ベイズの定理

ベイズの定理により、事象に関する事前知識と観測結果に基づき、原因となる事象の条件付き確率を求められる

ベイズの定理

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

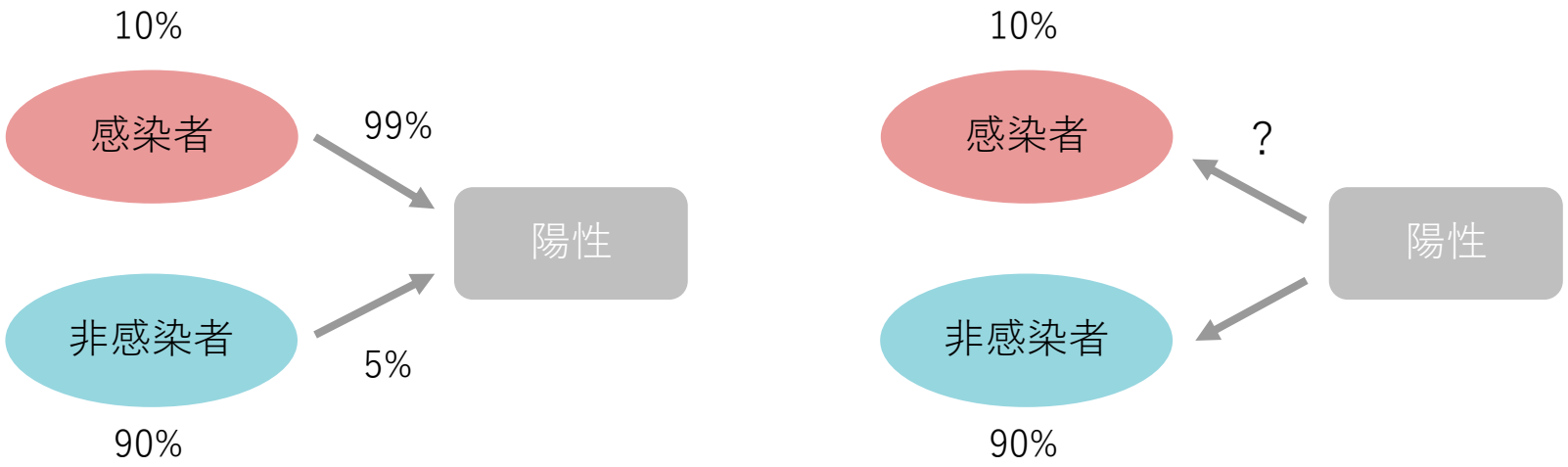
典型的には、原因Aが与えられたときの結果Bが起こる確率がわかっているとき、ベイズの定理を用いることで、結果が与えられたときの原因の確率を求められる。



ベイズの定理に基づく統計学の体系をベイズ統計学という。

3-6. ベイズの定理｜具体例

ある検査は、ある感染症にかかっているときに99%の確率で陽性と判定できるが、
かかっていない場合でも5%の確率で陽性と誤判定してしまう。
感染者の割合が10%のとき、陽性者が実際に感染している確率はどれほどか？



参考：[10-6. ベイズの定理の使い方｜統計学の時間｜統計WEB] https://bellcurve.jp/statistics/course/6448.html

3-6. ベイズの定理｜具体例

ある検査は、ある感染症にかかっているときに99%の確率で陽性と判定できるが、
かかっていない場合でも5%の確率で陽性と誤判定してしまう。
感染者の割合が10%のとき、陽性者が実際に感染している確率はどれほどか？

事象Aを「感染症にかかっている」、事象Bを「陽性と判定される」と置くと、
問題文より事象の確率は次のように求められる。

$$P(A) = 0.1, \qquad P(A^c) = 0.9, \qquad P(B|A) = 0.99, \qquad P(B|A^c) = 0.05$$

以上より、「陽性と判定された下で実際に感染症にかかっている確率」は
ベイズの定理を用いて次のように求められる。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} = \frac{0.99 \times 0.1}{0.99 \times 0.1 + 0.05 \times 0.9} = 69\%$$

3. まとめ

- 確率はランダムなイベントの相対的な起こりやすさを表す量である
- ランダムに値を取る変数を確率変数といい、その値の取り方を確率分布という
- 代表的な確率分布としては次の３つがある。
  - 二項分布：離散型、非負整数値、有限の値を取る
  - ポアソン分布：離散型、非負整数値、無限の値を取る
  - 正規分布：連続型、平均を中心としたばらつきを持つ
- 大数の法則・中心極限定理はサンプル数が増えたときの標本平均の振る舞いを説明する
- ベイズの定理を用いることで、事象に関する事前知識と観測結果に基づき、原因となる事象の条件付き確率を求められる

4. 推測統計学

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

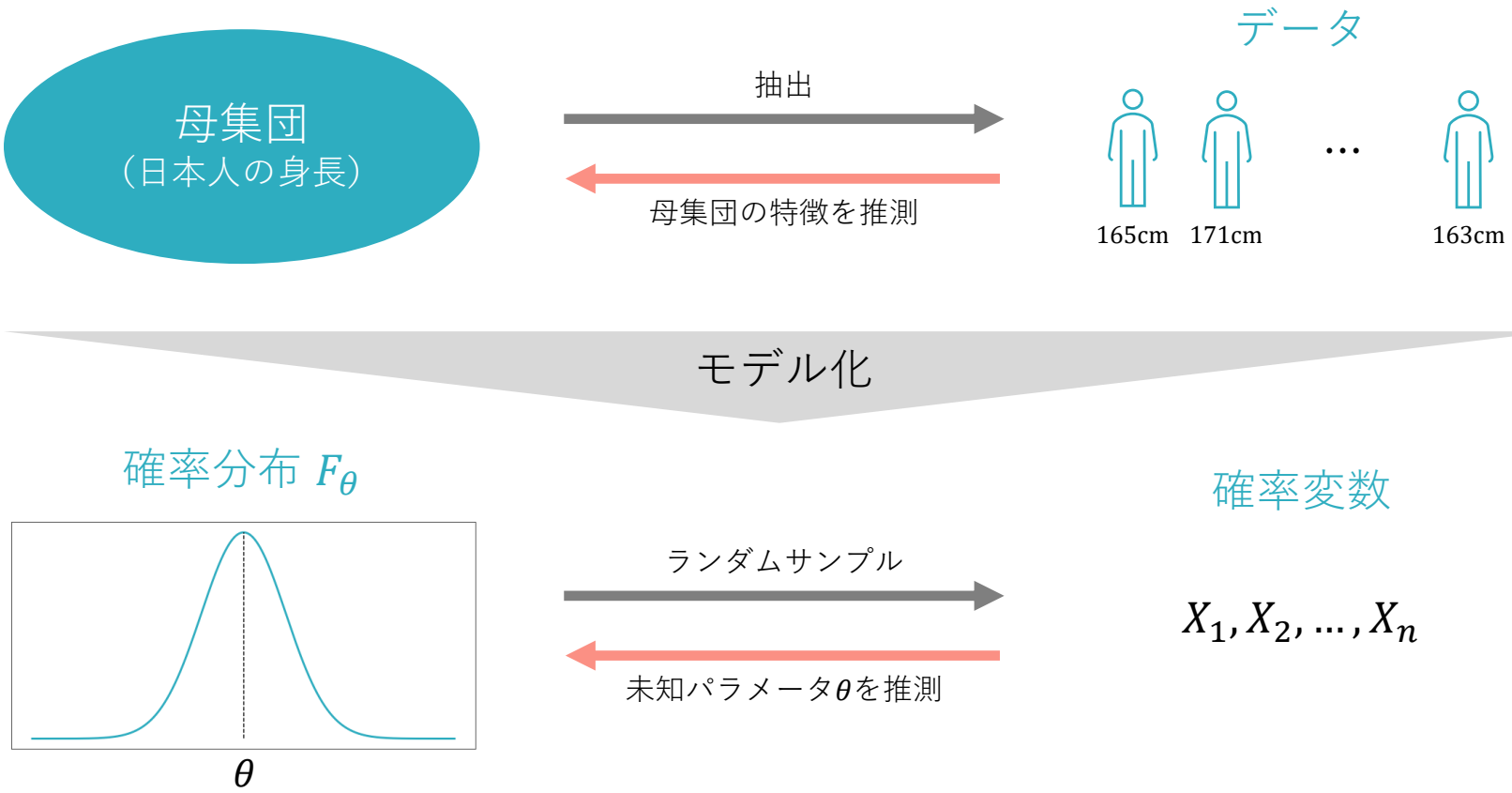
4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

4-1. 推測統計学とは

4-1. 推測統計学とは

推測統計では、一部のサンプルからその背後にある母集団の特徴を推測することを目的とする



* 確率変数$X_1, X_2, \ldots, X_n$は互いに独立に同一の分布に従う（independently and identicaly distributed; i.i.d.）と考える。

4-1. 推測統計学とは｜代表的な手法

推測統計の代表的な手法として推定、検定、区間推定がある

統計的推測

データを元にそれを生成する母集団の未知のパラメータθを推し測ること

統計的推測の手法

- 点推定　　：未知のパラメータθをピンポイントであてに行く
- 検定　　：未知のパラメータθがある仮説を満たすかどうかを検証する
- 区間推定　：未知のパラメータθを高い確率で含むような区間を構成する

※ 区間推定は推定という名前がついているが、手続きとしては検定と近い関係にある。

4-2. 点推定

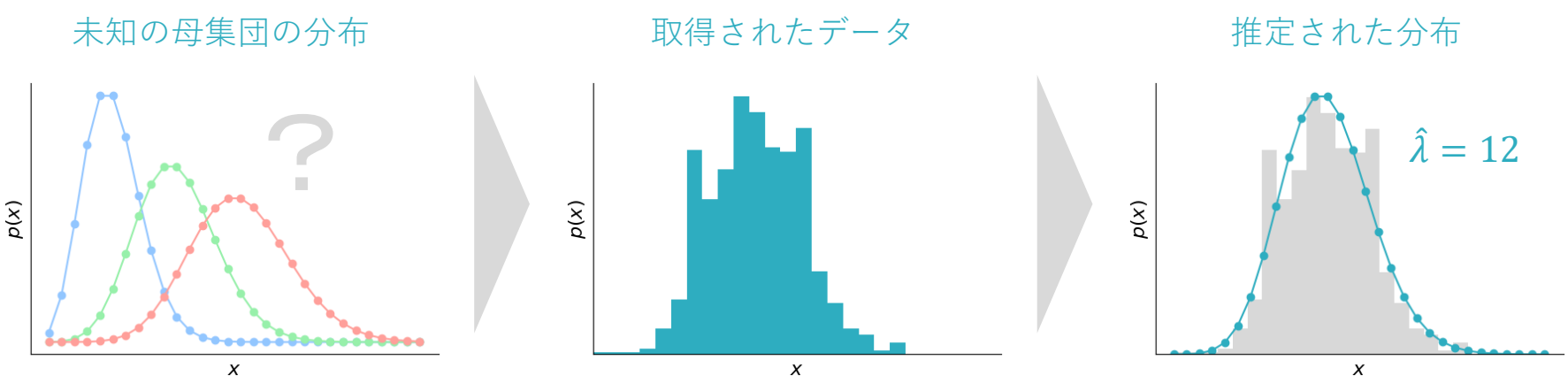
4-2-1. 点推定とは

推定では推定量を用いて未知パラメータを推測する

点推定では、母集団の未知パラメータをピンポイントで当てることを目的とする。
パラメータを当てるためにデータ$X_1, \ldots, X_n$から構成した量を推定量という。
パラメータθの推定量は$\hat{\theta}$で表すことが多い。

例 スーパーのとある商品の売れ行き

あるスーパーで商品Aの一日当たりの販売数は平均λのポアソン分布$Po(\lambda)$に従うとする。
データからλを推定すると$\hat{\lambda} = 12$となり、商品Aの販売数は$Po(12)$に従うことが分かった。



点推定は仮定した分布の下での母集団の特徴を把握するのに役立つ。

4-2-2. 点推定の基礎｜推定量の性質

推定量の満たす望ましい性質として、不偏性、一致性、漸近正規性がある

不偏性：期待値が真のパラメータθと等しい

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

一致性：nを大きくすると真のパラメータθに近づく

$$\hat{\theta} \to \theta$$

漸近正規性：nを大きくすると真のパラメータθを中心とした正規分布に近づく

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \to N(0, A)$$

例 標本平均 $\bar{X}$

標本平均 $\bar{X}$ は母平均 μ の推定量で、不偏性、一致性、漸近正規性を満たす*。

* 標本平均が不偏性を満たすことは期待値の線形性から、一致性を満たすことは大数の法則から、漸近正規性を満たすことは中心極限定理からわかる。

4-2-2. 点推定の基礎｜基本的な推定量

平均、分散の代表的な推定量には以下のようなものがある

標本平均

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

標本平均は母平均の推定量で、不偏性、一致性、漸近正規性を満たす。

標本分散

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

標本分散は母分散の推定量で、一致性、漸近正規性を満たす（不偏性は満たさない*）。

不偏標本分散

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

不偏標本分散は母分散の推定量で、不偏性、一致性、漸近正規性を満たす。

* 推定量が不偏性を満たさないとき「バイアスがある」と表現する。

4-2-3. 最尤推定

最尤推定により汎用的に望ましい推定量を得ることができる

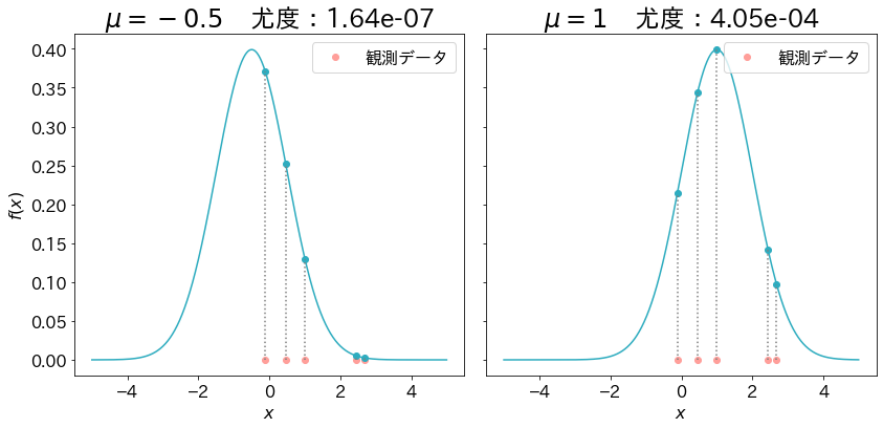
尤度関数：確率（密度）関数をパラメータθの関数として見たもの

$$L(\theta) = p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_\theta(X_i)$$

尤度関数を最大化する値としてパラメータを推定する方法を最尤推定といい、その推定量のことを最尤推定量という。

- 最尤推定量は（適切な条件の下）一致性、漸近正規性を持つ
- 通常の自然な推定量は最尤推定量として得られることが多い

例 正規分布 $N(\mu, 1)$



…

μについて
尤度を最大化

4-2-3. 最尤推定｜ポアソン分布の例

ポアソン分布の尤度関数は次の通り。

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i}}{X_i!} e^{-\lambda}$$

計算の簡便さから通常は次の対数尤度関数を最大化する。

$$l(\lambda) = \log L(\lambda) = \sum_{i=1}^{n} (X_i \log \lambda - \lambda - \log X_i!)$$

対数尤度を最大化する値として、次の尤度方程式の解を求める。

$$\frac{\partial l}{\partial \lambda} = \frac{\sum_{i=1}^{n} X_i}{\lambda} - n = 0$$

尤度方程式を満たす λ は $\hat{\lambda} = \bar{X}$、つまり λ の最尤推定量は標本平均として得られる。

※ このように手計算で求まる場合を除き、一般的には計算機を用いて数値的に算出する。

4-2. まとめ

- 点推定は未知のパラメータをピンポイントであてる推測の方法である
- パラメータをあてるためにデータから構成した量を推定量という
- 推定量の望ましい性質として次の３つがある
  - 不偏性：期待値が真のパラメータと一致する性質
  - 一致性：サンプル数を大きくしたときに真のパラメータに近づく性質
  - 漸近正規性：サンプル数を大きくしたときに分布が正規分布に近づく性質
- 望ましい推定量を得るための代表的な方法として最尤推定がある

4-3. 検定

4-3-1. 検定とは

検定では未知パラメータに関する仮説の検証を行う

検定ではパラメータに関する２つの仮説のいずれが正しいかを推測する。

検証する２つの仮説をそれぞれ帰無仮説、対立仮説と呼ぶ。

- 慣例的に帰無仮説はH_0、対立仮説はH_1という記号で表現される
- 帰無仮説と対立仮説は「両方のどちらか一方のみが成り立つ」という関係にあることが前提

例 クーポンの効果

あるECサイト上でユーザーの購入金額は、クーポンを発行した場合は $N(\mu_1, \sigma^2)$、発行していない場合は $N(\mu_2, \sigma^2)$ に従うとする。この時、２つの群の平均に差があるかどうかを検証したい場合は次の問題を考える。

$$H_0: \mu_1 = \mu_2 \quad \text{v.s.} \quad H_1: \mu_1 \neq \mu_2$$

検定の結果対立仮説 H_1 が正しいことが主張され、クーポンに効果があることが示唆された。

検定はデータに基づいた仮説の検証に役立つ。

4-3-2. 検定の基礎｜第１種の過誤と第２種の過誤

検定における推測には２種類の誤りが存在する

帰無仮説が真のとき、対立仮説を選択してしまう誤りを第１種の過誤
対立仮説が真のとき、帰無仮説を選択してしまう誤りを第２種の過誤という。

検定における２種類の誤りの関係

	帰無仮説を選択	対立仮説を選択
帰無仮説が真	正しい	第１種の過誤
対立仮説が真	第２種の過誤	正しい

これらの誤りを犯すリスクはトレードオフ*

- 第１種の過誤を犯す確率をある小さな値（=有意水準）以下に抑えた上で、第２種の過誤を犯す確率できるだけ減らす、という立場で推測する。

* 例えば、データによらず常に対立仮説を選択するという（不合理な）推測方法では、対立仮説が正しいときに第２種の過誤を犯すリスクはないが、帰無仮説が正しいときには常に第１種の過誤を犯す。

4-3-2. 検定の基礎｜検定の手続き

検定とは「確率的な背理法」である

検定ではデータに基づき構成される検定統計量の取った値に従って推測を行う。

検定における推測は次のような背理法的な手続きに従って行われる。

検定の手続き

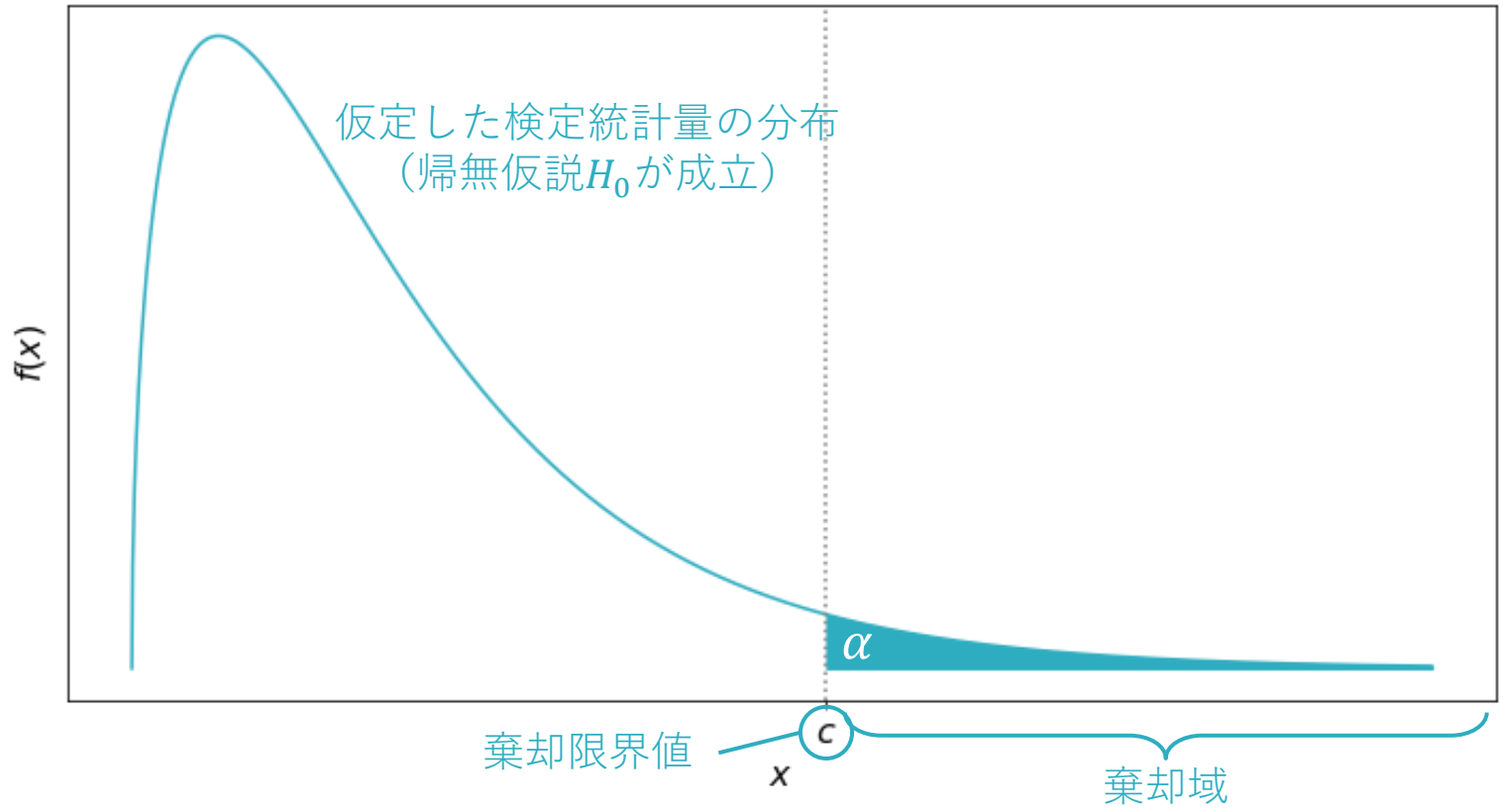
1. 帰無仮説を仮定する
2. 検定統計量を計算する
3. 検定統計量が帰無仮説の下で「確率的にほぼありえない」ような値を取ったとき、帰無仮説を棄却し対立仮説を受容する

- 「確率的にほぼありえない」の程度を定めるのが有意水準
- 帰無仮説を棄却するかどうかは、検定統計量の値が棄却域に入ったかどうかで判断

※ 仮に帰無仮説が棄却できなかったときに帰無仮説を積極的に支持することはできない（背理法では矛盾が生じなかったことは仮定を支持する根拠にならない）。

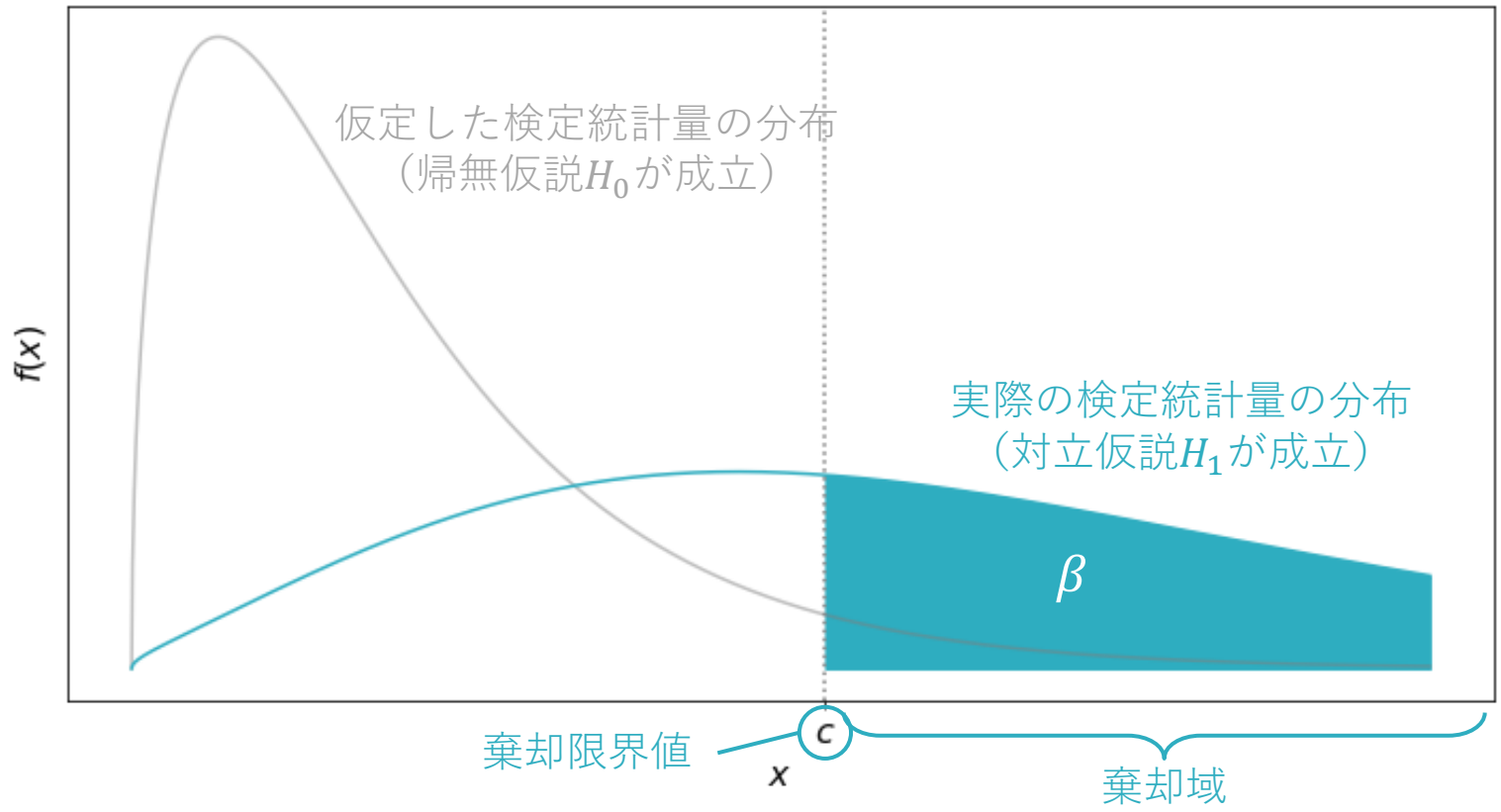
4-3-2. 検定の基礎｜検定の手続き（図解１）

- 検定統計量が帰無仮説の下で低い確率αでしか値を取らない領域として棄却域を設定し、検定統計量が棄却域に入ったときに帰無仮説を棄却する
- 通常、棄却域はある閾値（棄却限界値という）よりも大きい区間として設定される
- αを有意水準といい、この量が第１種の過誤を犯す確率を制御する



4-3-2. 検定の基礎｜検定の手続き（図解２）

- 対立仮説が真となるような状況では、高い確率βで検定統計量が棄却域に入ることが期待されるため、高い確率で帰無仮説を棄却し、対立仮説を支持することができる
- βを検出力といい、$1-\beta$が第２種の過誤を犯す確率に対応する



4-3-2. 検定の基礎｜片側検定と両側検定

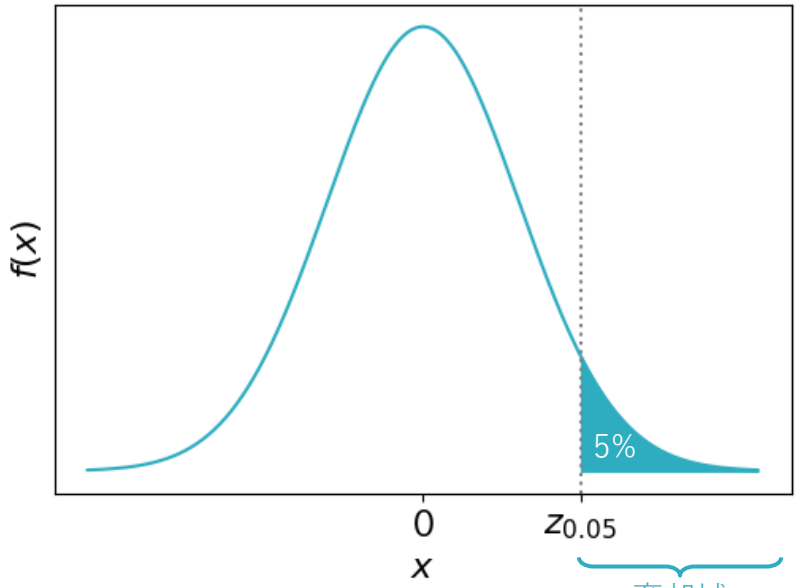
帰無仮説の誤りを検出する方向に応じて、片側検定と両側検定の２種類が考えられる

例 正規分布の平均の検定（有意水準5%）

片側検定：平均が0より大きいことを検出

$H_0: \mu \leq 0$ v.s. $H_1: \mu > 0$

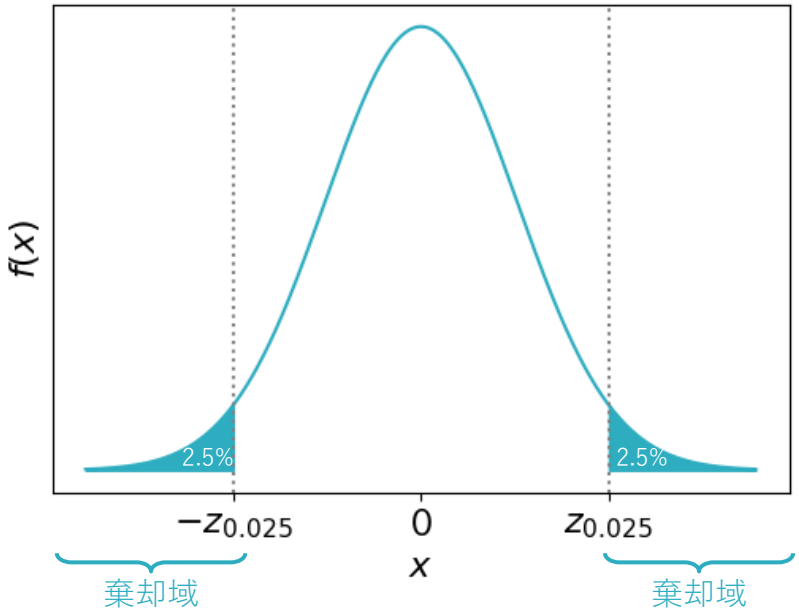
帰無仮説の下での検定統計量の分布



両側検定：平均が0でないことを検出

$H_0: \mu = 0$ v.s. $H_1: \mu \neq 0$

帰無仮説の下での検定統計量の分布



* z_αは標準正規分布の上側α点。

4-3-2. 検定の基礎｜様々な検定手法

前提とする統計モデルや検証する仮説に応じて様々な検定手法が用いられる

一般に、推測対象とする母集団の数に応じて、次のような設定が考えられる。

- 1 標本問題：1 つの母集団に関する推測の問題
- 2 標本問題：2 つの母集団の比較に関する推測の問題

以下では代表的な検定手法を紹介する。

正規分布に関する検定

- 平均の検定
  - 1 標本・分散既知
  - 1 標本・分散未知
  - 2 標本・分散既知
  - 2 標本・分散未知
- 分散の検定

二項分布に関する検定

- 母比率の検定
  - 1 標本
  - 2 標本

分割表に関する検定

- 適合度検定
- 独立性の検定

* 以下では原則的に両側検定に絞って説明する（片側検定は割愛）。

4-3-3. 正規分布に関する検定｜平均の検定（１標本・分散既知）

統計モデル

正規分布１標本問題（分散σ^2既知）

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

検定問題

平均の検定

$$H_0: \mu = \mu_0 \quad \text{v.s.} \quad H_1: \mu \neq \mu_0$$

検定統計量

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$$

は帰無仮説H_0の下で標準正規分布$N(0,1)$に従う。したがって、

$$|T| > z_{\alpha/2}$$

のとき帰無仮説を棄却し、対立仮説を採択する。

* z_αは標準正規分布の上側α点。

4-3-3. 正規分布に関する検定｜平均の検定（１標本・分散未知）

統計モデル
正規分布１標本問題（分散σ^2未知）

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

検定問題
平均の検定

$$H_0: \mu = \mu_0 \quad \text{v.s.} \quad H_1: \mu \neq \mu_0$$

分散既知のときの検定統計量のσを、その推定量で用いて置き換えた

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{s} \qquad \left(s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)$$

は帰無仮説H_0の下で自由度$n-1$のt分布$t(n-1)$に従う。したがって、

$$|T| > t_{\alpha/2}(n-1)$$

のとき帰無仮説を棄却し、対立仮説を採択する。

このようなt分布に基づく検定を総称してt検定という。

* $t_\alpha(k)$は自由度kのt分布の上側α点。

4-3-3. 正規分布に関する検定｜平均の検定（2標本・分散未知）

統計モデル

正規分布２標本問題（分散σ^2共通・未知）

$$X_1, \dots, X_m \sim N(\mu_1, \sigma^2) \quad Y_1, \dots, Y_n \sim N(\mu_2, \sigma^2)$$

検定問題

平均の検定

$$H_0: \mu_1 = \mu_2 \quad \text{v.s.} \quad H_1: \mu_1 \neq \mu_2$$

１標本問題と同様に分散の推定が必要となるが、プールされた推定量

$$s^2 = \frac{1}{m+n-2}\left\{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right\}$$

を用いると、検定統計量

$$T = \frac{\bar{Y} - \bar{X}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

は帰無仮説H_0の下で自由度$m+n-2$のt分布$t(m+n-2)$に従う。したがって、

$$|T| > t_{\alpha/2}(m+n-2)$$

のとき帰無仮説を棄却し、対立仮説を採択する。

* $t_\alpha(k)$は自由度kのt分布の上側α点。
** ここでは２群の分散が共通の値σ^2だと仮定したが、分散が等しくないときはWelchのt検定と呼ばれる方法を用いる。

4-3-4. 二項分布に関する検定｜母比率の検定（１標本）

統計モデル
二項分布１標本問題

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

検定問題
母比率の検定

$$H_0: p = p_0 \quad \text{v.s.} \quad H_1: p \neq p_0$$

母比率の推定量 $\hat{p} = X/n$ を用いると、検定統計量

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

は帰無仮説H_0の下で近似的に標準正規分布$N(0,1)$に従う*。したがって、

$$|T| > z_{\alpha/2}$$

のとき帰無仮説を棄却し、対立仮説を採択する。

* 近似的に正規分布に従うことは中心極限定理から従う。
** z_αは標準正規分布の上側α点。

4-3-4. 二項分布に関する検定｜母比率の検定（２標本）

統計モデル
二項分布２標本問題

$$X \sim Bin(m, p_1) \quad Y \sim Bin(n, p_2)$$

検定問題
母比率の検定

$$H_0: p_1 = p_2 \quad \text{v.s.} \quad H_1: p_1 \neq p_2$$

母比率の推定量 $\hat{p}_1 = X/m$, $\hat{p}_2 = Y/n$、またプールされた推定量

$$\hat{p} = \frac{X + Y}{m + n}$$

を用いると、検定統計量

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)\hat{p}(1 - \hat{p})}}$$

は帰無仮説H_0の下で近似的に標準正規分布$N(0,1)$に従う*。したがって、

$$|T| > z_{\alpha/2}$$

のとき帰無仮説を棄却し、対立仮説を採択する。

* 近似的に正規分布に従うことは中心極限定理から従う。
** z_αは標準正規分布の上側α点。

4-3-5. 独立性の検定

独立性の検定では分割表における２つの変量の間に関係があるかをどうかを検証する

分割表

各変量の値の組み合わせごとに観測された度数を記録した表のこと。
各セルに入る値を観測度数という。

「性別」と「アンケートへの回答の有無」の分割表

	回答	未回答	計
男性	5	35	40
女性	15	45	60
計	20	80	100

独立性の検定では「性別」と「アンケートへの回答の有無」の間に関係があるかどうかを検証する。

4-3-5. 独立性の検定｜多項分布

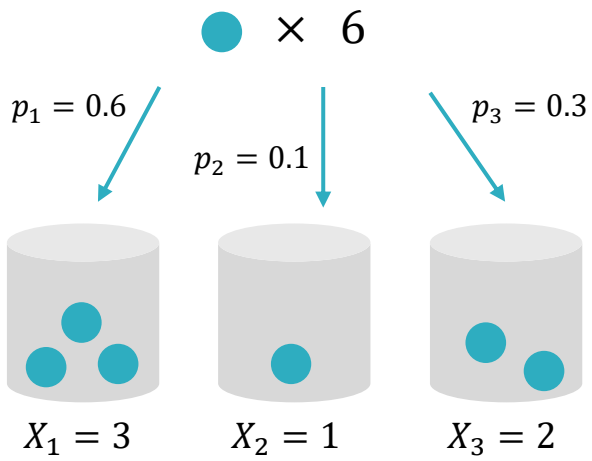
多項分布は二項分布を多次元に一般化した分布である

多項分布 $Mult(n, p_1, \dots, p_k)$

$$p(x_1, \dots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \quad (x_1 + \cdots + x_k = n)$$

母数	$p_i \geq 0\ (i = 1, \dots, k), n \geq 0$（整数） $p_1 + \cdots + p_k = 1$
平均	$E[X_i] = np_i$
分散	$V[X_i] = np_i(1 - p_i)$

- 二項分布の多次元への一般化
- 「n個のボールをk個の箱にランダムに投げ入れたときの各箱の中のボールの数」が従う分布



多項分布のイメージ

4-3-5. 独立性の検定｜分割表の統計モデル

分割表のデータは多項分布を用いてモデル化される

分割表のデータはセルの個数の多項分布としてモデル化される。

$$X = (X_{11}, \ldots, X_{rc}) \sim Mult(n, p_{11}, \ldots, p_{rc})$$

独立性の検定では、各セルの確率が独立な構造を持つかどうかを検定する。

$$H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j} \quad (i = 1, \ldots, r \quad j = 1, \ldots, c)$$

確率

p_{11}	p_{12}
p_{21}	p_{22}

確率（帰無仮説）

$p_{1\cdot} \times p_{\cdot 1}$	$p_{1\cdot} \times p_{\cdot 2}$	$p_{1\cdot}$
$p_{2\cdot} \times p_{\cdot 1}$	$p_{2\cdot} \times p_{\cdot 2}$	$p_{2\cdot}$
$p_{\cdot 1}$	$p_{\cdot 2}$	

観測度数

X_{11}	X_{12}
X_{21}	X_{22}

* r, cはそれぞれ分割表の行数(row)、列数(column)に対応する。

4-3-5. 独立性の検定｜検定の手続き

帰無仮説 H_0 の下では、各セルの観測度数は次の期待度数に近い値を取ることが予想される。

$$\hat{X}_{ij} = n\hat{p}_{i\cdot}\hat{p}_{\cdot j} \qquad \left(\hat{p}_{i\cdot} = \frac{1}{n}\sum_{j} X_{ij}\,, \qquad \hat{p}_{\cdot j} = \frac{1}{n}\sum_{i} X_{ij}\right)$$

そこで、期待度数と実際の観測度数との乖離を表すカイ二乗統計量に基づき検定を行う。

$$T = \sum_{i,j} \frac{\left(X_{ij} - \hat{X}_{ij}\right)^2}{\hat{X}_{ij}}$$

これは、帰無仮説H_0の下で近似的に自由度$(r-1)(c-1)$のカイ二乗分布に従うため、

$$T > \chi^2_{\alpha}\big((r-1)(c-1)\big)$$

ならば帰無仮説を棄却して、対立仮説を採択する。
このようなカイ二乗分布に基づく検定を総称してカイ二乗検定という。

観測度数

	回答	未回答	計
男性	5	35	40
女性	15	45	60
計	20	80	100

H_0の下での確率の推定量

	回答	未回答	計
男性	0.4×0.2	0.4×0.8	0.4
女性	0.6×0.2	0.4×0.8	0.6
計	0.2	0.8	1

期待度数

	回答	未回答	計
男性	8	32	40
女性	12	48	60
計	20	80	100

* $\chi^2_{\alpha}(k)$は自由度kのカイ二乗分布の上側α点。

4-3-5. 検定の多重性

検定を複数回繰り返す際には多重比較法の考え方が必要

例 効果のない薬剤

ある薬剤の効果を検証するために、投薬群とプラセボ群でt検定による2群の比較を有意水準5%で実行する。ただし、実際には薬剤に全く効果がなかったとする（帰無仮説が真）。
この時、データを取り直して検定を実行する、という操作を10回繰り返すと、

$$1-(1-0.05)^{10} \fallingdotseq 40\%$$

の確率で少なくとも1回は薬剤に効果があると主張してしまう。

差が出るまで検定を繰り返すことは「6の目が出るまでサイコロを振る」ことと同じ

- 無暗に検定を繰り返すと、全体として第１種の過誤を犯す確率が高くなり、本来差がないものに対して「差がある」と主張しやすくなる問題を検定の多重性という
- 検定の多重性の問題を回避するためには多重比較法という手法が用いられる
- 多重比較法では、全体としての第１種の過誤を犯す確率をコントロールするために、１つ１つの検定をより厳しい有意水準で実行する

参考：[検定の多重性とは？｜いちばんやさしい、医療統計] https://best-biostatistics.com/multiple/alpha.html

4-3. まとめ

- 検定は未知パラメータに関する２つの仮説のどちらが正しいかを推測する方法
- 検定の対象となる２つの仮説を帰無仮説・対立仮説と呼ぶ
- 帰無仮説が正しいときに対立仮説を採択する誤りを第１種の過誤、
  対立仮説が正しいときに帰無仮説を採択する誤りを第２種の過誤という
- 検定はデータから構成した検定統計量が、設定した棄却域に入ったかどうかで
  どちらの仮説を採択するかを判定する
- 第１種の過誤を犯す確率を制御する基準となる量を有意水準、
  対立仮説が正しいときに正しく帰無仮説を棄却できる確率を検出力と呼ぶ
- 仮定する統計モデル、仮説の種類に応じて様々な検定手法が用いられる
- 検定を繰り返し実施するときには注意を要する

4-4. 区間推定

4-4-1. 区間推定とは

区間推定では未知パラメータに対して幅を持った推測を行う

区間推定では真のパラメータをある一定の確率で含むような区間（信頼区間）を構成することで推測を行う。

信頼区間が真のパラメータを含む確率を信頼係数という。

例 選挙の得票率

とある選挙において、有権者100名に出口調査を行ったところ、60名は候補者Aに投票していることが分かった。この情報から候補者Aの得票率の95%信頼区間を求めると[50.2, 69.0]となり、候補者Aが当選する確度が高いことがわかった。

区間推定は興味のある未知パラメータを信頼度も併せて推測したいときに役立つ。

4-4-2. 区間推定の基礎｜信頼区間の構成法

信頼区間は検定の裏返しとして得られる

例 正規分布の平均の検定（１標本・分散既知）

検定統計量が帰無仮説で満たす式を変形すると、

$$P_\mu\left(\left|\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right| > z_{\alpha/2}\right) = \alpha$$

$$\Leftrightarrow \quad P_\mu\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq z_{\alpha/2}\right) = 1-\alpha \qquad \text{（事象の排反を取る）}$$

$$\Leftrightarrow \quad P_\mu\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad \text{（}\mu\text{について解く）}$$

となり、正規分布の平均μの$1-\alpha$信頼区間 $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ が導かれる。

このように信頼区間と検定は裏表の関係にあり、推論としては本質的には同等*。

* 実際、対応する信頼区間と検定においては「信頼区間に含まれないこと」と「検定で帰無仮説を棄却すること」は同値な関係にある。

4-4-2. 区間推定の基礎｜信頼係数の解釈

信頼係数は、データをサンプルして構成する信頼区間が真のパラメータを含む確率を表す

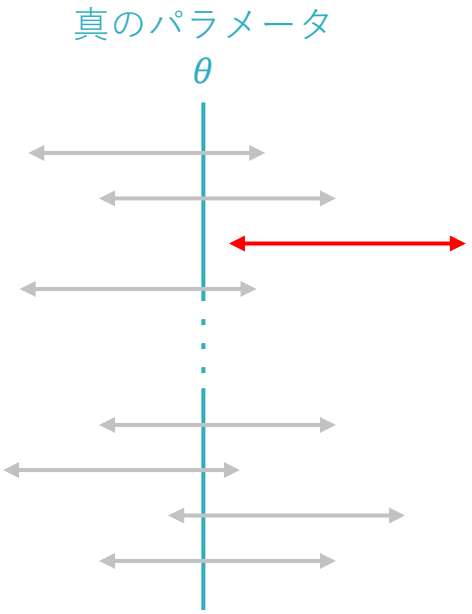
「信頼係数95%の信頼区間」の解釈

× 実際に構成したある信頼区間について、それが真のパラメータを含む確率が95%である

構成した信頼区間に対しては、真のパラメータはその中に含まれるか含まれないかのいずれかであるため、この解釈は誤り。

○ 「データをサンプルして信頼区間を構成する」という手続きを100回を繰り返したときに、概ね95回は真のパラメータを含む

これからデータをサンプルして構成する信頼区間が、95%の確率で真のパラメータを含む、という解釈が適切。



信頼係数95%の
信頼区間のイメージ

4-4-3. 様々な区間手法の手法｜正規分布に関する区間推定

前提とする統計モデル、推測対象のパラメータに応じて
様々な形の信頼区間が用いられる

設定	パラメータ	$1-\alpha$ 信頼区間（上限と下限）
1 標本問題・分散既知 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$	μ	$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
1 標本問題・分散未知 $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$	μ	$\bar{X} \pm t_{\alpha/2}(n-1) \frac{s}{\sqrt{n}}$
2 標本問題・分散未知 $X_1, \ldots, X_m \sim N(\mu_1, \sigma^2)$ $Y_1, \ldots, Y_n \sim N(\mu_2, \sigma^2)$	$\mu_2 - \mu_1$	$\bar{Y} - \bar{X} \pm t_{\alpha/2}(n+m-2) s \sqrt{\frac{1}{m} + \frac{1}{n}}$

- 1 標本問題・分散未知のsは不偏標本分散 $\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$ の平方根
- 2 標本問題のsはプールされた推定量 $\{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\}/(m+n-2)$ の平方根

* z_αは標準正規分布の上側α点。
** $t_\alpha(k)$は自由度kのt分布の上側α点。

4-4-3. 様々な区間推定の手法｜二項分布に関する区間推定

前提とする統計モデル、推測対象のパラメータに応じて
様々な形の信頼区間が用いられる

設定	パラメータ	$1-\alpha$ 信頼区間（上限と下限）
1 標本問題 $X \sim Bin(n,p)$	p	$\hat{p} \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
2 標本問題 $X \sim Bin(m,p_1)$ $Y \sim Bin(n,p_2)$	$p_2 - p_1$	$\hat{p}_2 - \hat{p}_1 \pm z_{\alpha/2}\sqrt{\left(\dfrac{1}{m}+\dfrac{1}{n}\right)\hat{p}(1-\hat{p})}$

- 1 標本問題の $\hat{p}$ は X/n
- 2 標本問題の $\hat{p}_1$, $\hat{p}_2$, $\hat{p}$ はそれぞれ X/m, Y/n, $(X+Y)/(m+n)$

* z_α は標準正規分布の上側 α 点。

4-4. まとめ

- 区間推定は未知パラメータを幅を持って推測する方法である
- 真のパラメータを一定の確率で含むような区間を信頼区間と呼び、
  信頼区間が真のパラメータを含む確率を信頼係数と呼ぶ
- 信頼係数は「データを取得して信頼区間を構成する」という手続きを
  繰り返したときに、信頼区間が真のパラメータを含む確率を表す
- 仮定する統計モデル、推測対象のパラメータに応じて様々な信頼区間が
  用いられる

4-5. 回帰分析

4-5-1. 回帰分析とは

ある変数から他の変数の振る舞いを説明するモデルを
推測する手法を回帰分析という

目的変数：説明する対象となる変数

説明変数：目的変数を説明するための変数

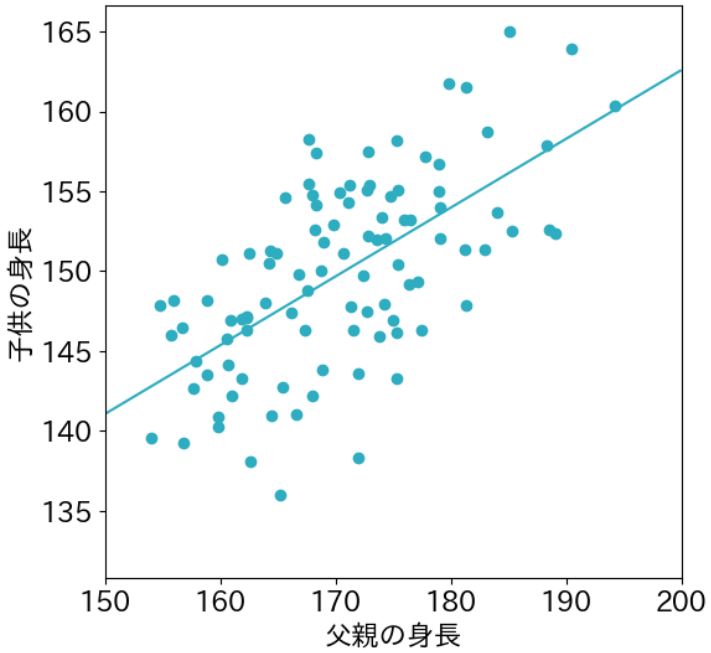
例 親子の身長の関係

- 目的変数 y：子供の身長（cm）
- 説明変数 x：父親の身長（cm）

yの振る舞いはxの一次式

$$y = ax + b$$

でうまく説明できる。



回帰分析は説明変数と目的変数の関係の解釈や、目的変数の予測に役立つ。

4-5-2. 線形回帰

線形回帰では目的変数を説明変数の線形和で表現する

線形回帰モデル

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad (i = 1, \ldots, n)$$

- 推測対象のパラメータは$\beta_k (k = 0, \ldots, p), \sigma^2$
- β_kを回帰係数といい、説明変数$x_{.k}$が1単位変化したときの目的変数の変化量を表す
- 説明変数が1次元（$p = 1$）の場合を単回帰、多次元（$p \geq 2$）の場合を重回帰という

線形回帰モデルは行列形式で次のように表現できる。

$$y = X\beta + \epsilon$$

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

4-5-2. 線形回帰｜最小二乗法

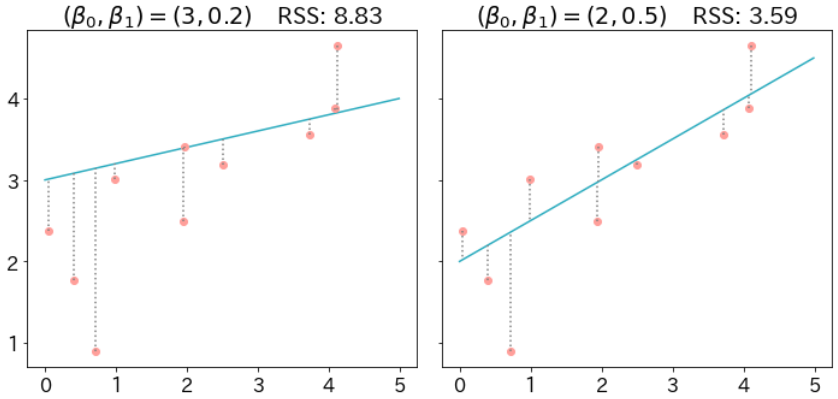
線形回帰のパラメータは最小二乗法によって推定できる

残差平方和（RSS）：予測値と実測値との乖離（残差）の二乗和

$$RSS = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})\}^2$$

残差平方和を最小化する値としてβを推定する方法を最小二乗法といい、
その推定量のことを最小二乗推定量という。

例 単回帰モデル $y = \beta_0 + \beta_1 x_1 + \epsilon$



…

(β_0, β_1)について
RSSを最小化

* RSSはResidual Sum of Squaresの略。
** 誤差分布が正規分布に従うという前提の下で、最小二乗推定量は最尤推定量と一致する。

4-5-2. 線形回帰｜最小二乗推定量の導出

残差平方和は線形回帰モデルのベクトル、行列表現を用いて次のように表される。

$$RSS = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})\}^2 = \|y - X\hat{\beta}\|^2$$

残差平方和を最小化するために、$\hat{\beta}$について偏微分したものを0と置くと、
次の方程式を得る（正規方程式）。

$$X^{\mathrm{T}} X\hat{\beta} = X^{\mathrm{T}} y$$

$X^{\mathrm{T}} X$が逆行列を持つとき、最小二乗推定量は次のように得られる。

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$$

多重共線性

複数の説明変数の間に線形に近い関係が存在するとき、回帰係数の推定が不安定になる*。
この問題を多重共線性という。多重共線性の問題を防ぐ方法の一つとして正則化がある。
（正則化について詳しくはモデリングの講義を参照。）

* （数学的な説明だが）直観的には$X^{\mathrm{T}} X$が特異行列（逆行列が存在しない行列）に近づくことで、逆行列$(X^{\mathrm{T}} X)^{-1}$の計算が不安定になるからであると理解できる。

4-5-2. 線形回帰｜決定係数

決定係数はモデルの当てはまりの良さを表す指標の一つ

線形回帰モデルにおいては、次の平方和の分解が成り立つ。

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{目的変数の変動}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{説明変数で説明される変動}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{説明変数で説明されない変動}}$$

$\bar{y}$：y_iの平均値　　$\hat{y}_i$：y_iの予測値 $\left(=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip}\right)$

目的変数の全体の変動のうち、説明変数により説明される割合を決定係数という。

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

- 0から1の間の値を取り、1に近いほどデータに対するモデルの当てはまりが良いことを表す
- 単回帰の場合には目的変数と説明変数の相関係数の2乗に一致する

※ 決定係数は説明変数を加えるほど1に近づくため、予測の観点からは適切な規準ではない。
予測の観点では情報量規準やクロスバリデーションといった規準を用いることができる。

4-5. まとめ

- 回帰分析はある変数を他の変数から説明するモデルを用いて推測する方法
- 振る舞いを説明する対象の変数を目的変数、目的変数を説明するための変数を説明変数という
- 回帰分析のうち目的変数を説明変数の線形和で表現したものを線形回帰という
- 線形回帰モデルのパラメータは最小二乗法を用いて推定できる
- 線形回帰モデルのデータへの当てはまりの良さを表す量として決定係数がある

5. バイアス

目次

1. 統計学
1-1. 統計学とは
1-2. 統計学を学ぶ意義
1-3. 統計学の種類
1-4. データの種類

2. 記述統計学
2-1. 記述統計学とは
2-2. １変数データの記述
2-3. ２変数データの記述
2-4. 相関係数の解釈上の注意

3. 確率と確率分布
3-1. なぜ確率を学ぶのか
3-2. 確率
3-3. 確率変数
3-4. 代表的な確率分布
3-5. 大数の法則と中心極限定理
3-6. ベイズの定理

4. 推測統計学
4-1. 推測統計学とは
4-2. 点推定
4-3. 検定
4-4. 区間推定
4-5. 回帰分析

5. バイアス
5-1. バイアスとは
5-2. 選択バイアス
5-3. 情報バイアス
5-4. 交絡バイアス

5-1. バイアスとは

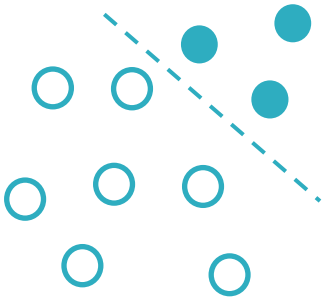
データが母集団の特徴を適切に反映できていない結果、推論結果が歪んでしまうことを「バイアスがある」という

データの分析、解釈にあたってはバイアスに留意することが重要。

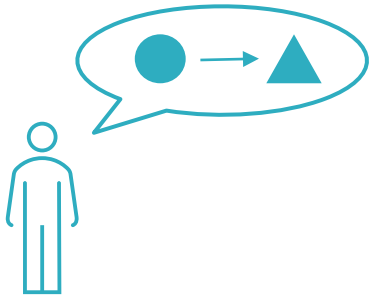
以下の代表的な３つのバイアスについて紹介する。

標本の選択 → **データの収集** → **統計分析**

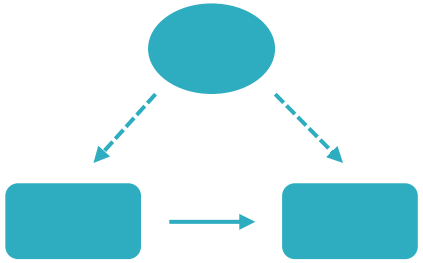
選択バイアス



情報バイアス



交絡バイアス



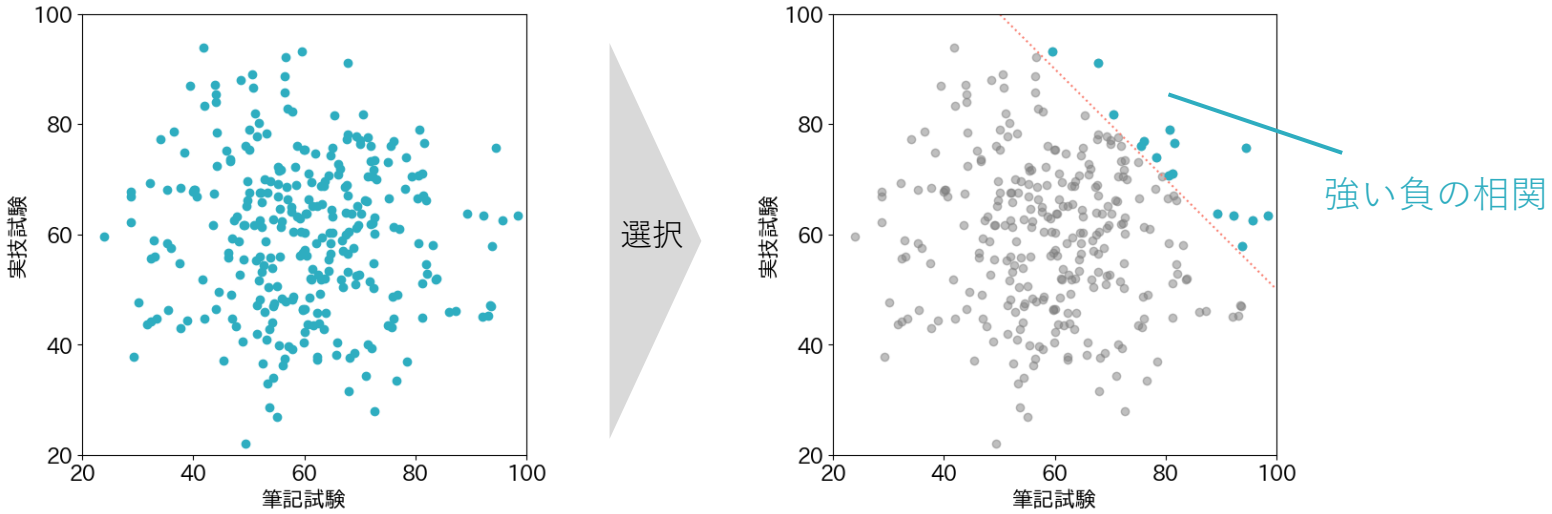
* ここでは「推定量が不偏性を持っていない」という推定量の性質としてのバイアスではなく、より一般的な意味でのバイアスについて取り扱う。

5-2. 選択バイアス

不適切な条件により、母集団の中から偏った標本を選んでしまうことにより生じるバイアス

例 大学入学試験の合格者

ある大学の入学試験では、筆記試験（100点満点）と実技試験（100点満点）の２つの科目の合計点数が150点以上の生徒が合格となる。筆記と実技の点数の関係を調べたい時に、合格者のデータのみから分析を実施すると、本来は存在しない関係性が見出されてしまう。



▶ 分析対象の標本が偏って抽出されたものでないかを事前に確認する

参考：[行政プロセスにデータ分析を取り入れるために知っておきたい知識と事例] https://www.soumu.go.jp/main_content/000675341.pdf

5-3. 情報バイアス

測定方法や情報の取り違いなどの原因からデータ収集過程で生じるバイアス

例 報告バイアス

生活習慣に関するアンケートにおいて、喫煙・飲酒などの習慣は過小に報告されやすくなる。

例 想起バイアス

当人の来歴によって、思い出した情報の正確さや粒度が異なる。
例えば、子供の服薬歴についてのアンケートで、持病のある子供を持つ母親の方が、より鮮明な内容で報告できる。

例 質問者バイアス

アンケートにおいて、本質的には同等の質問であっても、聞き方を変えることで異なる回答が得られる。

▶ データ収集の過程で情報を歪める要因がないかを確認する

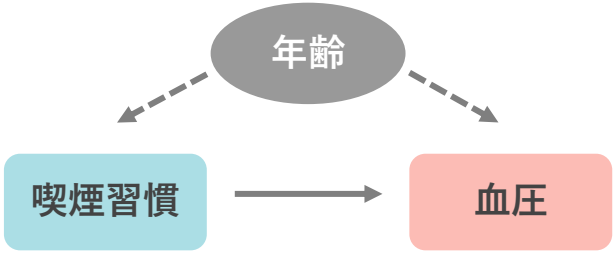
参考：[心理学用語「情報バイアス」とは？意味から具体例までわかりやすく解説 – スッキリ] https://gimon-sukkiri.jp/info/

5-4. 交絡バイアス

処置と結果の双方に影響を及ぼす要因を見逃すことによって生じるバイアス

例 喫煙と血圧の関係

喫煙の血圧に与える影響を調査するために、喫煙者と非喫煙者の集団について血圧の比較を行った結果、有意な差が見られた。
しかし、この分析は喫煙習慣と血圧の双方に影響する年齢という因子を無視しており、影響を過大評価している可能性がある。



処置と結果の双方に影響する要因を交絡因子と呼び、それを無視した解析はバイアスを生む。
交絡バイアスを除くためには以下のような方法がある。

1. 実験デザインを通して交絡因子を排除する

処置を標本にランダムに割り当てて２群を等価な集団にすることで、両者を比較可能にする。
（ランダム化比較試験：RCT）

2. 分析手法を通じて交絡因子の影響を取り除く

交絡因子についての一定の仮定の下で、バイアスの影響を除いた効果を推定する。
（回帰分析、傾向スコア分析など）

* このような、処置の与える効果を統計的に推測する枠組みを統計的因果推論と呼ぶ。

5. まとめ

- データが母集団を反映していないことで推論の結果が歪んでしまうことを「バイアスがある」という

- バイアスには大きく次の3種類がある
  - 選択バイアス：標本が偏って抽出されることで生じるバイアス
  - 情報バイアス：データの収集過程で生じるバイアス
  - 交絡バイアス：処置と結果に与える要因を無視することで生じるバイアス

Appendix

A. 代表的な統計量の従う分布

A. 代表的な統計量の従う分布｜カイ二乗分布

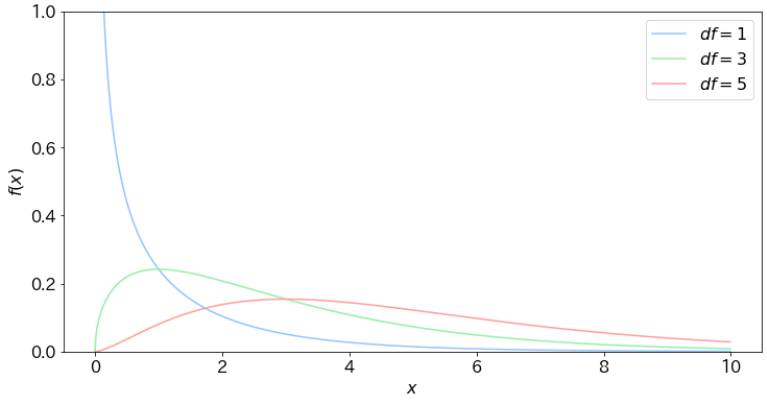
正規分布の分散の推定量は、適切なスケーリングの下でカイ二乗分布に従う

独立に$N(0,1)$に従う確率変数$Z_1, \dots, Z_k$の二乗和

$$Y = Z_1^2 + \cdots + Z_k^2$$

が従う分布を自由度kのカイ二乗分布$\chi^2(k)$と呼ぶ。

- カイ二乗分布はガンマ分布と呼ばれる分布の特殊な場合である



独立に正規分布$N(\mu, \sigma^2)$に従う確率変数$X_1, \dots, X_n$から構成した不偏標本分散

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

について次の性質が成り立つ。

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

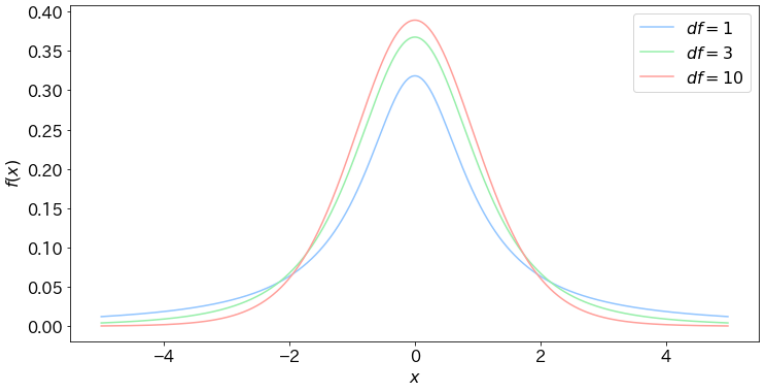
A. 代表的な統計量の従う分布｜t分布

t統計量はt分布と呼ばれる正規分布に近い分布に従う

独立な確率変数 $Z \sim N(0,1), U \sim \chi^2(k)$ に対し

$$T = \frac{Z}{\sqrt{U/k}}$$

が従う分布を自由度kのt分布$t(k)$と呼ぶ。



- 自由度が小さいほど裾が重く*、特に$k = 1$の時の分布をコーシー分布という
- 自由度が大きいほど裾が軽く*、$k \to \infty$の極限で標準正規分布に一致する

独立に正規分布$N(\mu, \sigma^2)$に従う確率変数$X_1, \ldots, X_n$から構成した次のt統計量

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{s} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma}{\sqrt{s^2/\sigma^2}}$$

について、分子は$N(0,1)$、分母は$\chi^2(n-1)/n-1$に従うので、これは$t(n-1)$に従う。

* 分布の端に向かって確率（密度）が急速に減衰する分布を裾が軽い分布、逆に減衰が遅い分布を裾が重い分布と呼ぶ。

BrainPad
Analytics Innovation Company

株式会社ブレインパッド
〒106-0032 東京都港区六本木三丁目1番1号 六本木ティーキューブ
TEL：03-6721-7002
www.brainpad.co.jp　info@brainpad.co.jp