

# データ解析

# ガイダンス

---



**創域理工学部**

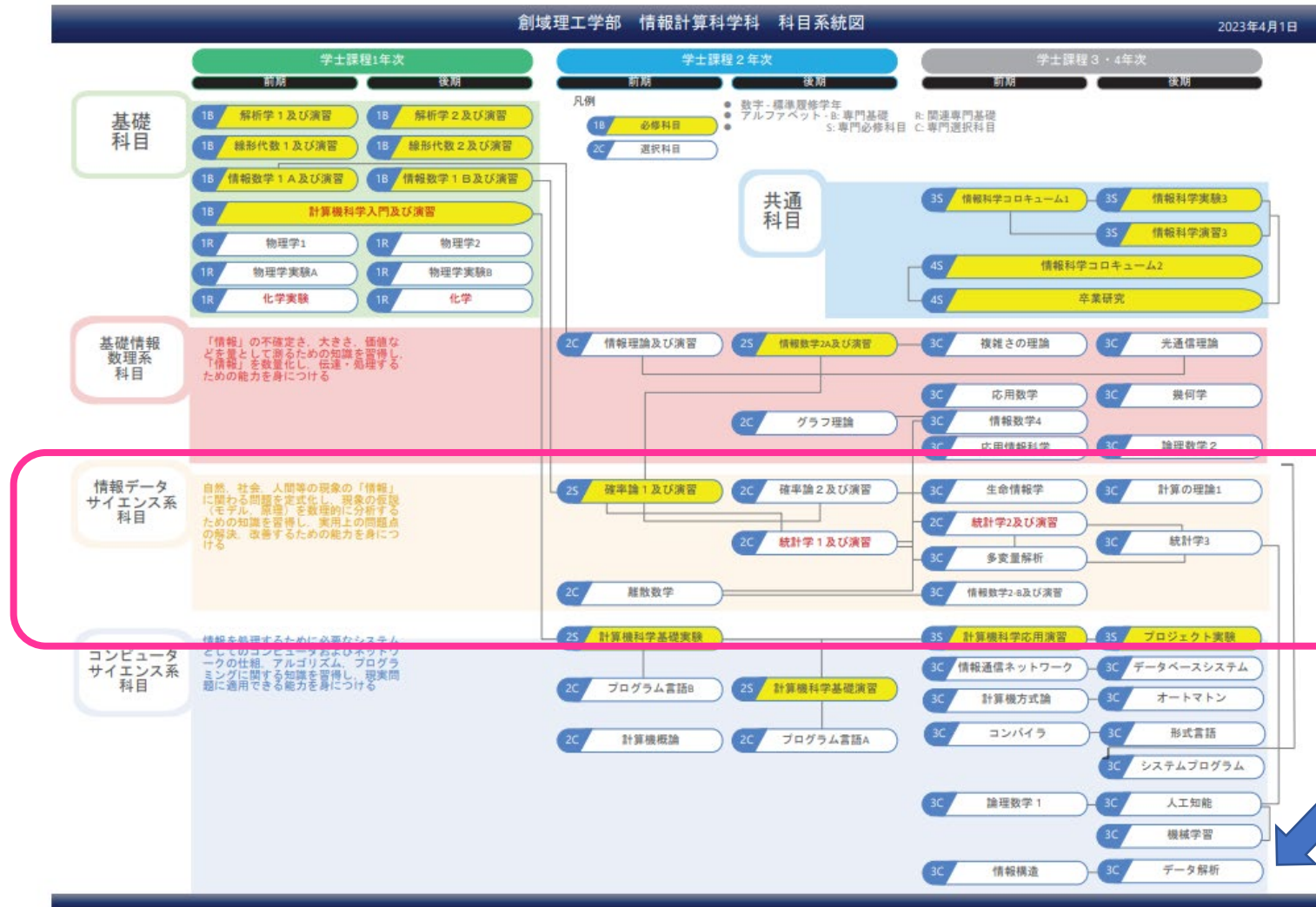
Faculty of Science and Technology

東京理科大学  
創域理工学部情報計算科学科  
安藤宗司

---

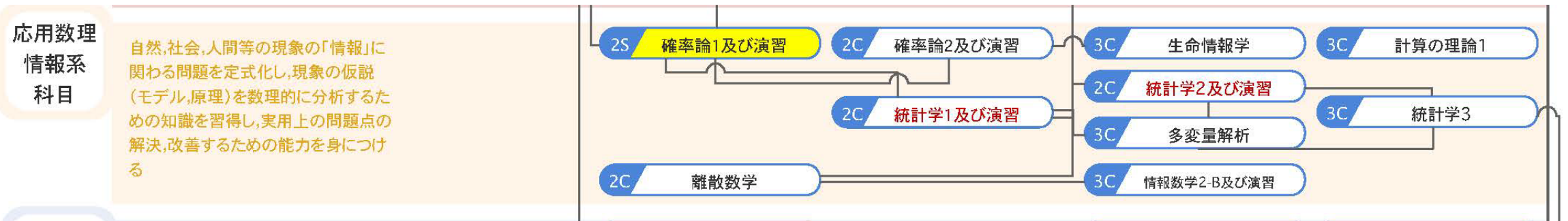
2023年9月14日

# 情報計算科学科の科目系統図



データ解析  
はここ！

# 情報データサイエンス系科目に注目すると



- 統計学1及び演習 (2年前期)
  - 統計的推測の推定を扱った
- 統計学2及び演習 (2年後期)
  - 統計的推測の検定を扱った
- 多変量解析 (3年前期)
  - 互いに従属する一群の変量間の関係を扱う手法を扱った

## データ解析

これら科目の後続する科目として位置付けられています

# 統計学3のカリキュラム

---

	日程	内容
第1回	9/14	ガイダンス, 統計学1, 統計学2及び多変量解析の復習 (記述統計)
第2回	9/21	統計学1, 統計学2及び多変量解析の復習 (推定)
第3回	9/28	統計学1, 統計学2及び多変量解析の復習 (検定)
第4回	10/5	統計学1, 統計学2及び多変量解析の復習 (線形回帰モデル)
第5回	10/12	一般化線形モデル1 (指数分布族)
第6回	10/19	一般化線形モデル2 (推定)
第7回	10/26	一般化線形モデル3 (区間推定, 仮説検定)
第8回	11/2	学習内容の点検と確認 (中間試験)

# 統計学3のカリキュラム

---

	日程	内容
第9回	11/9	一般化線形モデル4（正規線形モデル）
第10回	11/16	一般化線形モデル5（一元配置分散分析）
第11回	11/30	生存時間データ解析1（生存関数，ハザード関数）
第12回	12/7	生存時間データ解析2（生存関数の推測，Cox回帰分析）
第13回	12/14	サンプルサイズ設計1
第14回	12/21	サンプルサイズ設計2
第15回	1/18or25	到達度評価試験

# 成績評価方法

---

□ 到達度評価試験（60%）

□ 中間試験（40%）

# データ解析

## 統計学1, 統計学2及び 多変量解析の復習

---



**創域理工学部**

Faculty of Science and Technology

東京理科大学  
創域理工学部情報計算科学科  
安藤宗司

---

2023年9月14日

# Contents

---

- 統計学
- 記述統計
  - データの尺度と種類
  - データ尺度別の集計方法
  - グラフ表示
- 統計的推測
  - 母集団と標本
  - 母集団分布とパラメータ
  - 推定
  - 検定



# 統計学

---

- データ解析のための方法を研究する学問
- 多くの分野で広く用いられる手法
- 記述統計
  - 調査や実験によって得られたデータを整理
  - 解釈を助ける統計的手法
  - グラフ, 観測値の平均や標準偏差
- 統計的推測
  - データが誤差あるいは確率的な変動を多く含む
  - データの背後にデータ発生の方に関する確率的なモデルを想定
  - データから確率モデルの推定や検定

# 統計的推測の考え方

---

## □ 新薬の効果を評価したい

- 薬の効果は人によってさまざま（すべての人に効果があるとは限らない）
- 「統計的」に効果があるかを検討

確率的な変動を多く含むデータ

データのモデル化が容易

## □ 統計的推測の手法は広く用いられ、その有効性が認識されている

# データの種類

---

## □ 質的データ

- 数値として観測することに意味がない
- あるカテゴリに属していること，またはある状態にあることがわかるデータ
- 性別，出身地，癌のstage，重症度

## □ 量的データ

- 定量的な値をもつデータ
- 気温，西暦，身長，体重

# データの尺度（質的データ）

---

## □ 名義尺度

- データの値が同一かどうかの区別のみが意味をもつ  
質的データ
- 例えば、性別、血液型

## □ 順序尺度

- データの値と並び順が意味をもつ質的データ
- 例えば、癌のstage、重症度

# データの尺度（量的データ）

---

## □ 間隔尺度

- データの **間隔** に意味がある量的データ
- 和や差の演算が可能
- 0はひとつの状態を表す
- 例えば，気温や西暦など

## □ 比尺度

- データの **間隔と比** に意味がある量的データ
- 和差積商の演算が可能
- 0は何もないことを表す
- 例えば，身長，体重，球速

# データの種類のまとめ

---

種類	尺度	演算
質的データ	名義尺度	できない
	順序尺度	順序の比較は可
量的データ	間隔尺度	和差 (+, -)
	比尺度	和差積商 (+, -, ×, ÷)

# データの尺度に対応した集計方法

---

- データの尺度によって集計方法は異なる
- なんでも平均値を求めればいいわけではない
- 情報量の観点から次の関係が成り立つ

比尺度 > 間隔尺度 > 順序尺度 > 名義尺度

# 量的データの要約

---

□ 平均値  $\frac{\text{観測値の合計}}{\text{サンプル数}}$

- 最も用いられる要約指標の一つ
- 解釈が容易

□ 中央値

- 観測値を昇順に並び替える

$n$ : サンプル数

$$\text{中央値} = \begin{cases} \frac{\frac{n}{2} \text{番目の観測値} + \frac{n}{2} + 1 \text{番目の観測値}}{2} & (n \text{ が偶数}) \\ \frac{n}{2} + 1 \text{ の整数番目の観測値} & (n \text{ が奇数}) \end{cases}$$



# 平均値と中央値の違い

---

## □ 厚生労働省「2022年 国民生活基礎調査の概況」

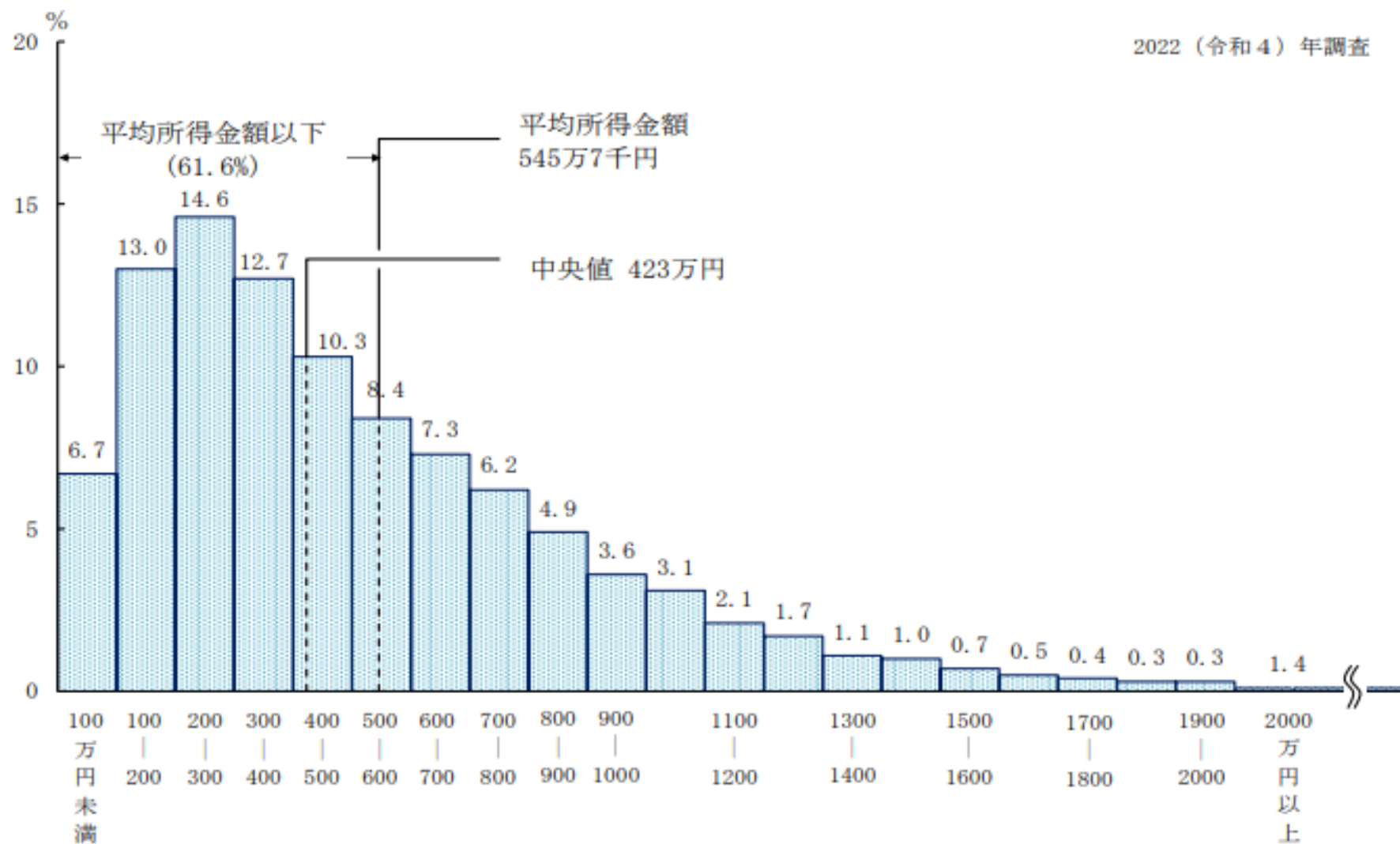
- 日本人の年収の平均値は約545万円
- 日本人の年収の中央値は約423万円

## □ 平均値は「外れ値」が観測値に含まれると、その影響をうけてしまう

- データの特徴に合わせて、平均値と中央値を適切に使い分ける必要がある

# 所得の分布

出典：2022（令和4）年 国民生活基礎調査の概況



# ばらつきを表す指標

---

2つのデータの平均値（中央値）が等しい

✖ II

2つのデータの特徴が同じ

□ ばらつきを測らないと判断できない

- 平均値（中央値）が同じでもばらつきが異なる状況がある
- データの特徴を示すには、代表値とばらつきを表す指標の両方を示す必要がある

平均値と分散（標準偏差）

中央値と四分位範囲

# 分散（標準偏差）

---

## □ 分散の定義

$$\frac{(\text{症例1の観測値} - \text{平均値})^2 + \dots + (\text{症例}n\text{の観測値} - \text{平均値})^2}{n}$$

- 平均値と比較して観測値がどの程度ばらついているかを表現した指標
- 分散の値が小さいならば、観測値が平均値に近い傾向がある
- 観測値の単位と異なるため解釈が難しい

## □ 標準偏差の定義

$$\sqrt{\frac{(\text{症例1の観測値} - \text{平均値})^2 + \dots + (\text{症例}n\text{の観測値} - \text{平均値})^2}{n}}$$

- 観測値と同じ単位であるため解釈が容易

# 正規分布と標準偏差

---

## □ 医学論文

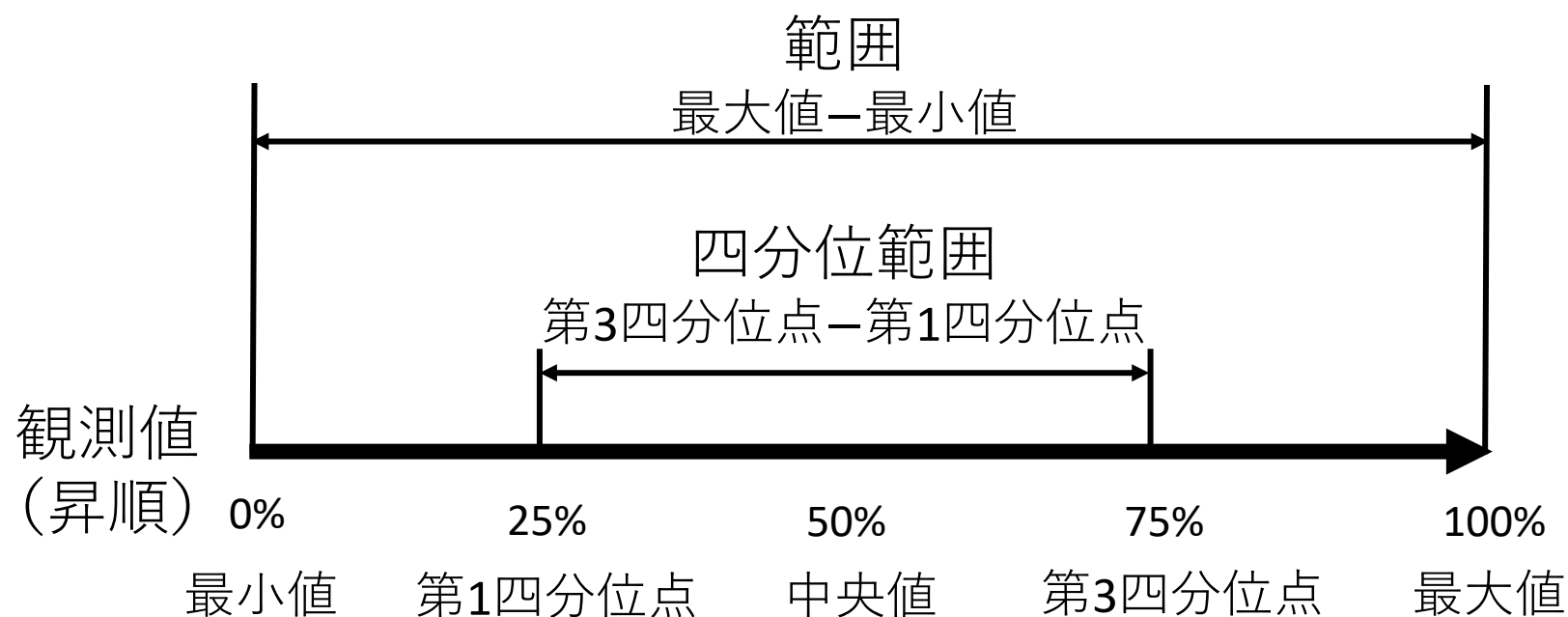
- 患者背景の量的データに対して、平均 $\pm$ 標準偏差を算出していることが多い

## □ 量的データが正規分布に従っていると仮定

- 平均 $\pm$ 標準偏差の範囲：データの約70%が含まれている
- 平均 $\pm 2 \times$ 標準偏差の範囲：データの約95%が含まれている
- 平均 $\pm 3 \times$ 標準偏差の範囲：データの約99%が含まれている

# 四分位範囲

## □ 定義



四分位範囲の値が小さいならば，観測値が中央値に近い傾向がある

# グラフ表示

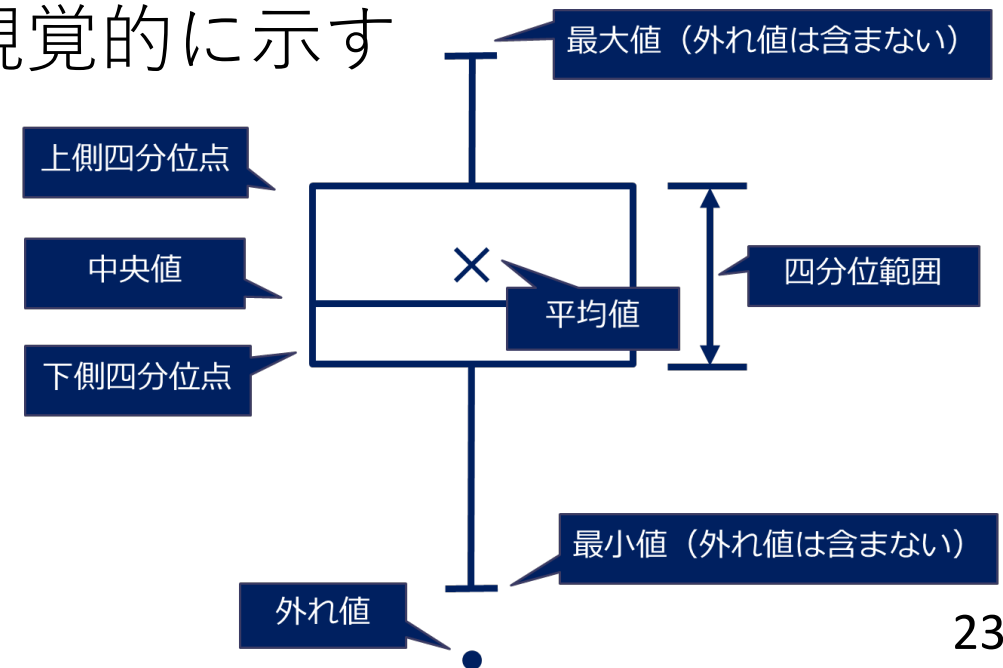
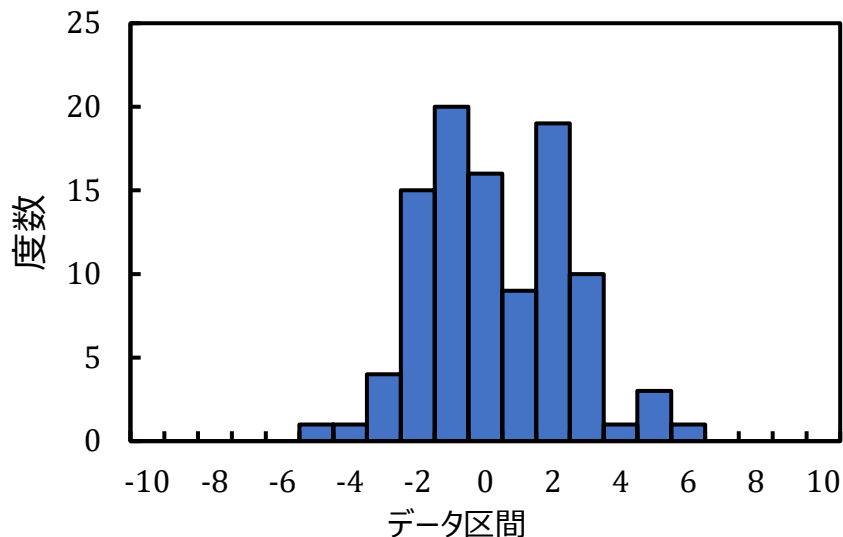
- データの特徴を視覚的に把握することは重要

- ヒストグラム

  - 観測値の分布を視覚的に示す

- 箱ひげ図

  - 観測値の代表値とばらつきを視覚的に示す



# 質的データの要約

---

## □ 度数

- 対象のカテゴリーに属する症例数

## □ 比率

- 対象のカテゴリーに属する症例の比率（度数/全体の症例数）

重症度	度数	比率
軽度	120	0.60
中等度	56	0.28
重度	24	0.12
計	200	1



# Contents

---

## □ 統計学

## □ 記述統計

- データの尺度と種類
- データ尺度別の集計方法
- グラフ表示

## □ 統計的推測

- 母集団と標本
- 統計的モデル（母集団分布とパラメータ）
- 推定
- 検定

# 導入

---

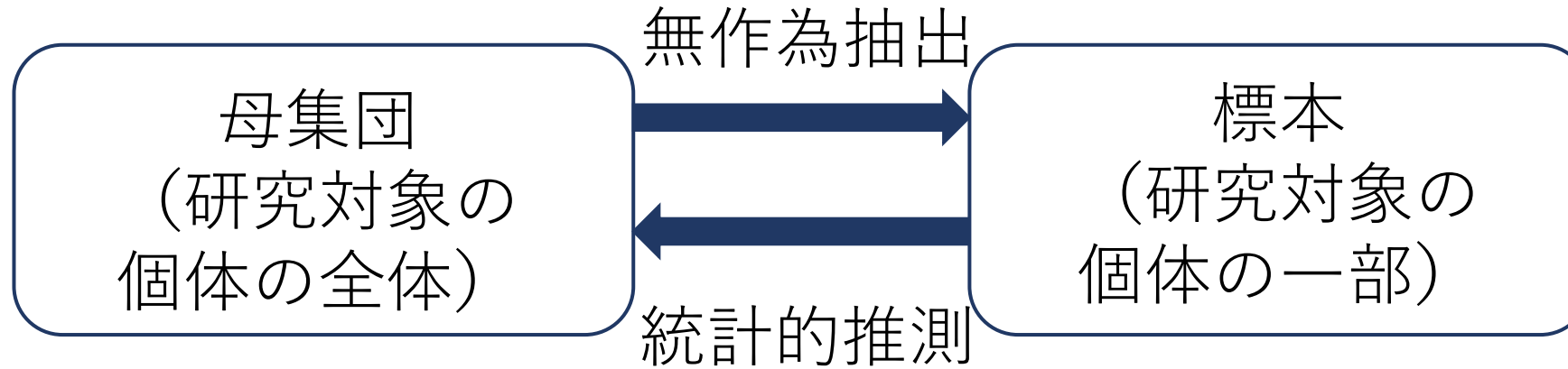
- がん患者**100**名に対して新規抗がん剤を投与したところ、効果があった患者は**60**名だった
  - 新規抗がん剤は**60%**の患者に有効
- 別のがん患者**100**名に対して新規抗がん剤を投与したところ、効果があった患者は**55**名だった
  - 新規抗がん剤は**55%**の患者に有効

確率的な変動を含むデータ

新規抗がん剤の**真の有効率**を知るためには、**がん患者全員**に新規抗がん剤を投与して結果を得るしかない

# 母集団と標本

---



母集団から標本を無作為抽出できているかどうかが重要である

# 統計的モデル

---

## □ 母集団分布

- 母集団から無作為に抽出した標本が取り得る値とその確率の対応関係を表すもの
- 正規分布，二項分布

## □ パラメータ

- 母集団分布の関数を決定付けるもの
- 統計的推測（推定と検定）では，パラメータが推測対象になる

# 確率変数

---

- 観測値はばらつく
- 統計モデルではばらつきの扱いが重要

ばらつきを表現・定式化するにはどうしたらよいか？

- 確率変数 $X$ の定義
  - $X$ が取り得る値はある範囲に定まっている
  - $X$ はある時点が過ぎると値が定まる
  - $X$ の取り得る値について確率分布は定まっている

# 確率変数の例

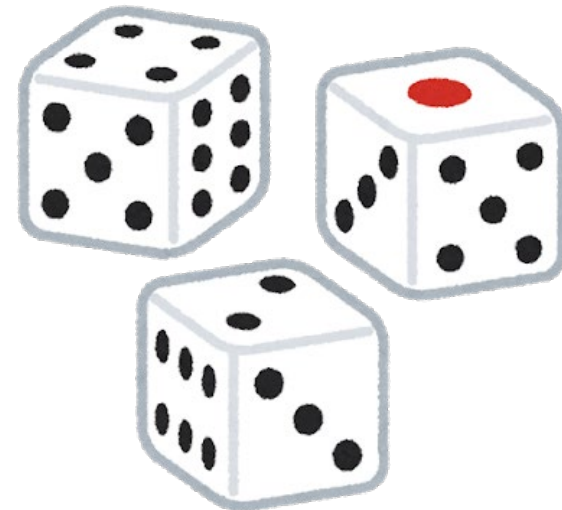
---

## □ 確率変数 $X$ の定義

- $X$ が取り得る値はある範囲に定まっている
- $X$ はある時点が過ぎると値が定まる
- $X$ の取り得る値について確率分布は定まっている

## □ サイコロの出る目 : $X$

- 1, 2, 3, 4, 5, 6 の 6 つ
- サイコロを投げると目は定まるが  
投げるまでは値が不確定である
- $\Pr(X = i) = \frac{1}{6} \quad (i = 1, 2, 3, 4, 5, 6)$



# サイコロの出る目 $X$ の母集団分布

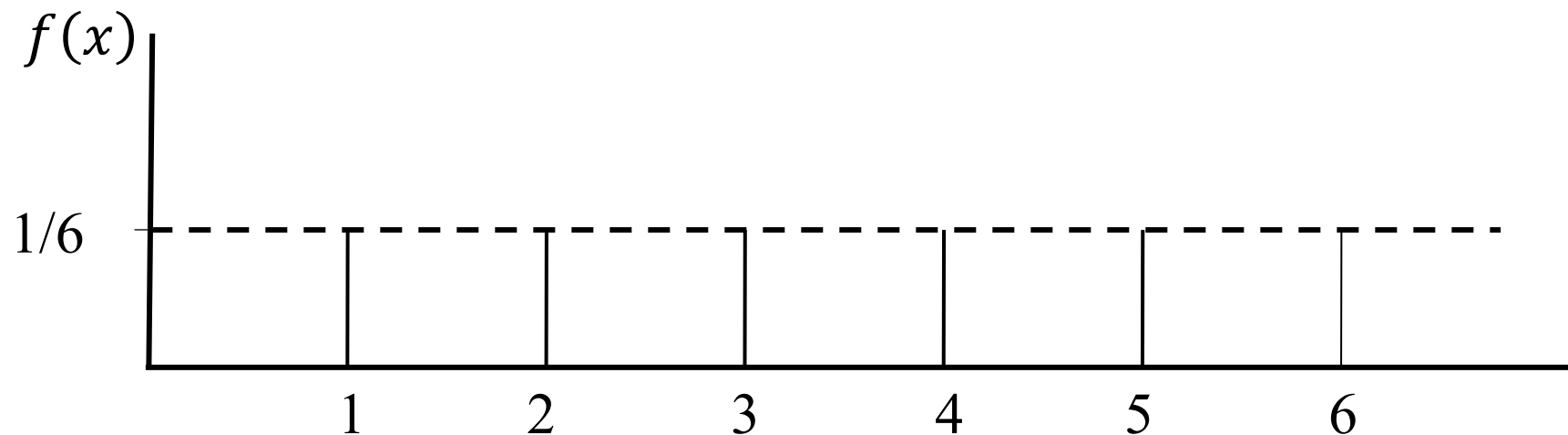
---

## □ 確率関数 $f(x)$

- 確率変数  $X$  の観測値  $x_i$  に対する確率  $p_i$  を対応させたもの
- $X = 2$  に対する確率は  $\frac{1}{6}$ , つまり  $f(2) = \frac{1}{6}$

## □ 母集団分布

- 確率変数を取り得る値の確率を記述・表現したもの



# 母集団分布の種類

---

## □ 離散分布

- 二項分布  
ポアソン分布  
多項分布など
- 確率変数の取り得る値が  
離散的
- 質的データ
- 確率関数

$$f(x_i) \geq 0$$

$$\sum_{i=1}^n f(x_i) = 1$$

## □ 連続分布

- 正規分布  
指数分布  
t分布など
- 確率変数の取り得る値が  
連続的
- 量的データ
- 確率密度関数

$$f(x) \geq 0$$

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



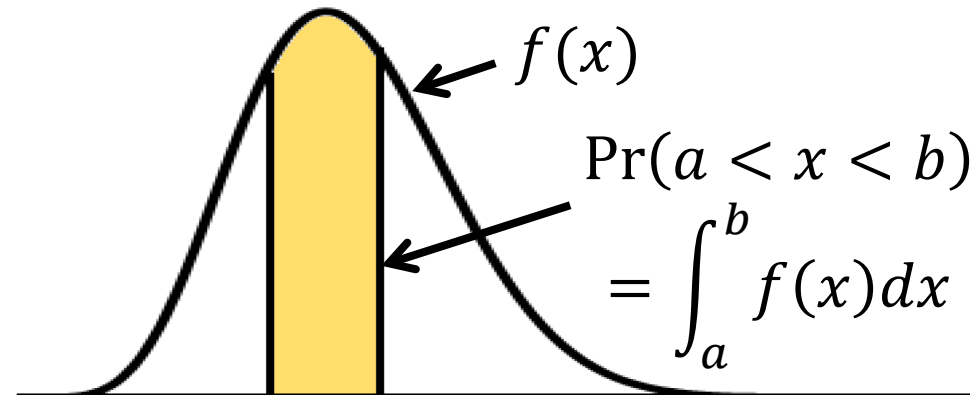
# 連続分布

## □ 確率変数が連続値をとる時の確率分布

- 取る値が連続であるため、1点に対して確率を定められない
- 確率関数の代わりに確率密度関数を用いる

## □ 確率密度関数 $f(x)$

- $x$ のある区間に対して確率を定める
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$



# 正規分布

---

## □ 確率密度関数

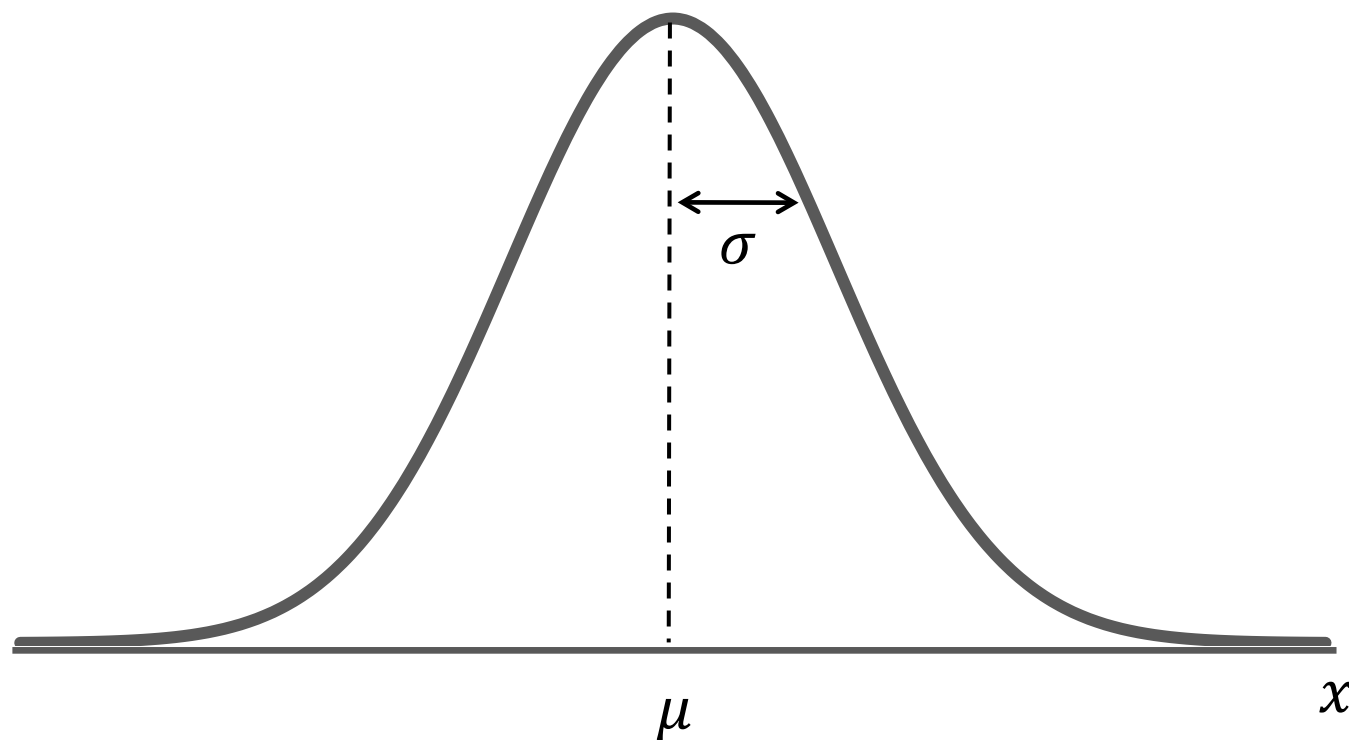
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (-\infty < x < \infty)$$

- パラメータ： $\mu, \sigma$ ，それ以外は定数
- $\mu$ と $\sigma$ の値が定まれば， $f(x)$ の値が定める
- 一般的に，正規分布は $N(\mu, \sigma^2)$ と表される

# 正規分布の概形

---

- $\mu$ と $\sigma$ の値によらず，左右対称の単峰性の分布
- $\mu$ は分布の平均， $\sigma$ は分布の標準偏差



# 標準正規分布

---

□  $\mu = 0, \sigma = 1$ をもつ正規分布

□ 標準正規分布と正規分布の関係

$$\begin{array}{ccc} & X_2 = X_1 \times \sigma + \mu & \\ X_1 \sim N(0,1) & \begin{array}{c} \xrightarrow{\hspace{1cm}} \\ \xleftarrow{\hspace{1cm}} \end{array} & X_2 \sim N(\mu, \sigma^2) \\ & X_1 = \frac{X_2 - \mu}{\sigma} & \end{array}$$

# 二項分布

---

## □ 確率関数

$$f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

- パラメータ： $\pi$ ，それ以外は定数
- $\pi$ の値が定まれば， $f(x)$ の値が定める
- 一般的に，二項分布は $\text{Bin}(n, \pi)$ と表される

# 点推定

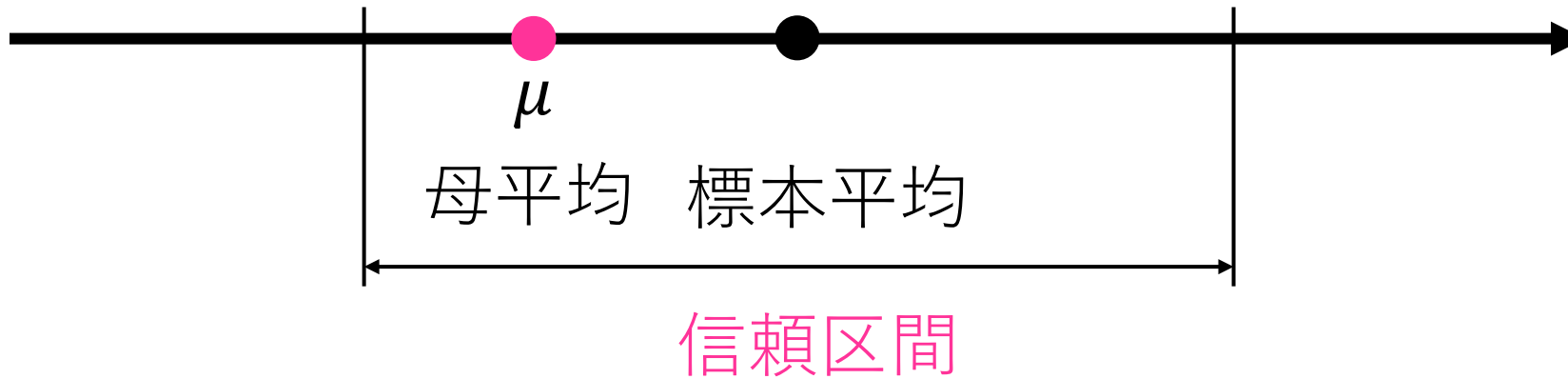
---

- パラメータを観測値から1つの値で推定する方法
  - がん患者100名に対して新規抗がん剤を投与したところ、効果があった患者は60名だった
    - 真の有効率（パラメータ $\pi$ ）の点推定値は60%
  - 別のがん患者100名に対して新規抗がん剤を投与したところ、効果があった患者は55名だった
    - 真の有効率（パラメータ $\pi$ ）の点推定値は55%

母集団から標本を無作為抽出するたびに、  
真の有効率の点推定値は変わってしまう

# 区間推定

- 母集団パラメータが含まれるであろう区間
- 95%信頼区間は，母集団パラメータが95%の確率で含まれる区間として定義される



- 信頼区間は，母集団パラメータを含む確率（信頼係数）を高くするほど，区間幅が広くなる

# 信頼区間の意味

---

- 正規母集団  $N(\mu, \sigma^2)$  から無作為に20個の個体を抽出し、95%信頼区間を構成する作業を100回繰り返すとする
- 100回のうち、95回はパラメータ $\mu$ を含んでいることが期待される
- 95%信頼区間とは、母集団から何度も標本を抽出し、その都度信頼区間を構成したとき、そのうちの95%がパラメータ $\mu$ を含んでいることが期待される区間である



# 標本分布

---

- 母集団からある確率で個体が抽出される
- 抽出された個体から計算される推定量（標本平均など）も，ある確率によって変動する
- 推定量の取り得る値とその確率の対応関係を表すものを標本分布という
- 量的データ
  - 母集団分布：正規分布 $N(\mu, \sigma^2)$
  - 標本分布
    - $\sigma^2$ を**既知**として $\mu$ を推測：正規分布
    - $\sigma^2$ を**未知**として $\mu$ を推測：t分布

# ばらつきを表す指標

---

## □ 標準偏差

- データのばらつきを表す指標

## □ 標準誤差

- 推定量（標本平均など）のばらつきを表す指標

$$\text{標準誤差} = \frac{\text{標準偏差}}{\sqrt{n}}$$

- サンプル数が増えれば，標準誤差は小さくなる

# 信頼区間の構成

---

□ 確率変数 $X$ の母集団分布は $N(\mu, \sigma^2)$ に従う

■  $\sigma^2$ を既知として $\mu$ を推測

■ 母集団分布からの独立な標本:  $X_1, X_2, \dots, X_n$

■ 標本平均 $\bar{X}(= \frac{X_1 + X_2 + \dots + X_n}{n})$ は $N(\mu, \frac{\sigma^2}{n})$ に従う

■  $u = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$ は $N(0, 1)$ に従う

$Z_{\alpha/2}$ : 標準正規分布の上側 $\alpha/2$ 点

$$\Pr(-Z_{\alpha/2} \leq u \leq Z_{\alpha/2}) = 1 - \alpha$$

$$\Leftrightarrow \Pr\left(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Leftrightarrow \Pr\left(\bar{X} - Z_{\alpha/2} \sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + Z_{\alpha/2} \sqrt{\sigma^2/n}\right) = 1 - \alpha$$

# 信頼区間の構成

---

□ 確率変数 $X$ の母集団分布は $N(\mu, \sigma^2)$ に従う

■  $\sigma^2$ を未知として $\mu$ を推測

■ 母集団分布からの独立な標本:  $X_1, X_2, \dots, X_n$

■ 標本平均 $\bar{X}(= \frac{X_1 + X_2 + \dots + X_n}{n})$ は $N(\mu, \frac{\sigma^2}{n})$ に従う

■ 不偏標本分散:  $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n - 1$

■  $\frac{(n-1)s^2}{\sigma^2}$ は自由度 $n - 1$ のカイ二乗分布に従う

■  $t = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} / \frac{s}{\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)}{s}$ は自由度 $n - 1$ のt分布に従う

$t_{\alpha/2}(n - 1)$ : 自由度 $n - 1$ のt分布の上側 $\alpha/2$ 点

$$\Pr(-t_{\alpha/2}(n - 1) \leq t \leq t_{\alpha/2}(n - 1)) = 1 - \alpha$$

$$\Leftrightarrow \Pr\left(\bar{X} - t_{\alpha/2}(n - 1)\sqrt{s^2/n} \leq \mu \leq \bar{X} + t_{\alpha/2}(n - 1)\sqrt{s^2/n}\right) = 1 - \alpha$$

# 同時確率（密度）関数と尤度関数

---

## □ 同時確率（密度）関数

$$f(x_1, x_2, \dots, x_n; \theta)$$

- $\theta$ が与えられたとき，どのような $x_1, x_2, \dots, x_n$ が得られやすいかを示す関数

## □ 尤度関数

$$L(\theta; x_1, x_2, \dots, x_n)$$

- $x_1, x_2, \dots, x_n$ が得られたとき，それが出現しやすい $\theta$ の値を示す関数

$$f(x_1, x_2, \dots, x_n; \theta) = L(\theta; x_1, x_2, \dots, x_n)$$

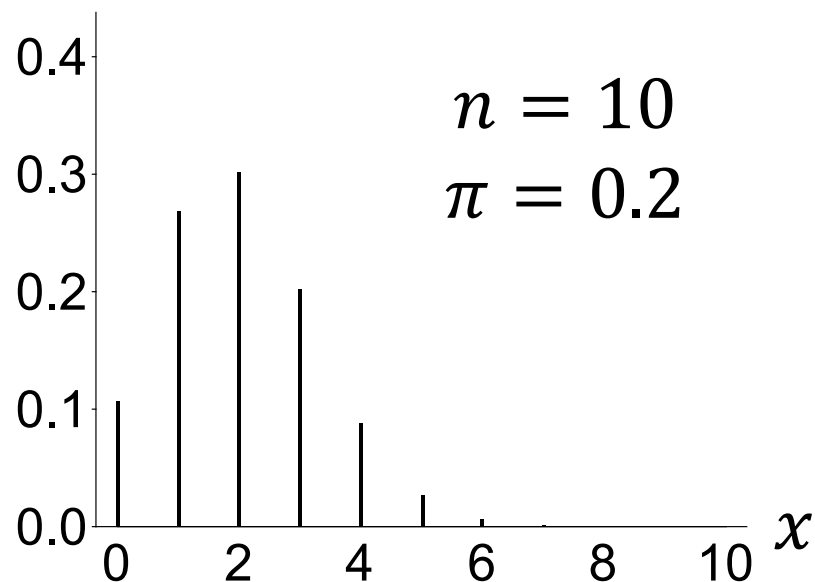
# 二項分布の母数の推測

確率関数

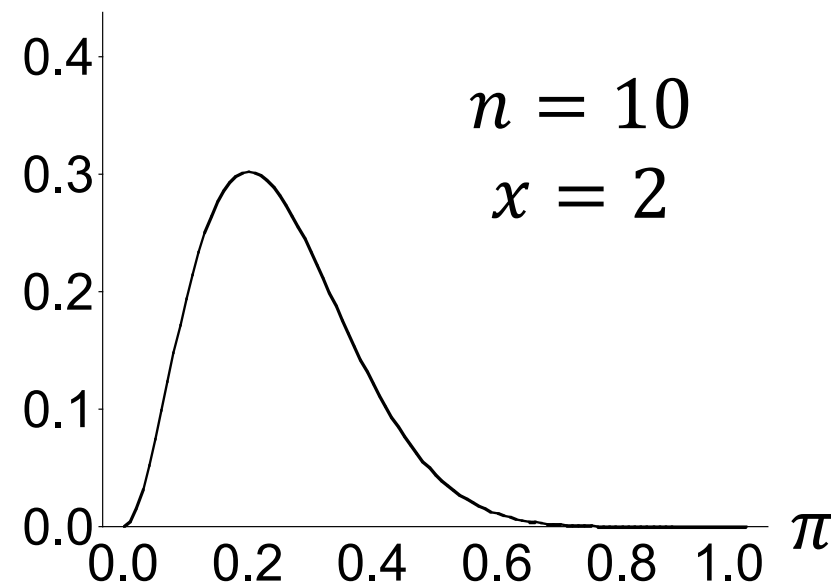
尤度関数

$$f(x|\pi) = \binom{n}{x} \pi^x (1 - \pi)^{n-x} = L(\pi|x)$$

$f(x|\pi)$



$L(\pi|x)$



# 最尤推定

---

## □ 対数尤度関数

$$l(\pi|x) = \log L(\pi|x) = \log \binom{n}{x} + x \log \pi + (n - x) \log(1 - \pi)$$

## □ 尤度方程式

$$\frac{\partial}{\partial \pi} l(\pi|x) = \frac{x}{\pi} - \frac{n - x}{1 - \pi} = 0$$

## □ 最尤推定量

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

# 最尤推定量の分散

---

## □ 分散

$$V[\hat{\pi}] = V\left[\frac{x}{n}\right] = \frac{V[x]}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}$$

## □ 分散の推定量

$$\hat{V}[\hat{\pi}] = \frac{\hat{\pi}(1-\hat{\pi})}{n} = \frac{x/n(1-x/n)}{n} = \frac{x(n-x)}{n^3}$$



# 二項分布の母数の信頼区間

---

## □ 複数の信頼区間構成法がある

- Wald型
- スコア型
- Agresti and Coull の信頼区間
- 正確な信頼区間
- 他にもいろいろ

## □ 性能評価の指標

- 被覆確率
- 信頼区間幅
- 扱いやすさ（計算の容易さ）

# Wald 型の信頼区間の構成

---

□  $X \sim \text{Bin}(n, \pi)$

□  $\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$  ( $n$  が十分大きい場合: 近似)

$$\Pr\left(-z_{\alpha/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

最尤推定量  $\hat{\pi} = \frac{x}{n}$

$\pi \rightarrow \hat{\pi}$  に置き換える (近似)

$$\Pr\left(\hat{\pi} - z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \leq \pi \leq \hat{\pi} + z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right) = 1 - \alpha$$

# スコア型の信頼区間の構成

---

□  $X \sim \text{Bin}(n, \pi)$

□ 帰無仮説  $H_0: \pi = \pi_0$

□  $\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$  ( $n$  が十分大きい場合: 近似)

$$\Pr\left(-z_{\alpha/2} \leq \frac{\hat{\pi} - \pi}{\sqrt{\pi(1-\pi)/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$\pi \rightarrow \pi_0$  に置き換える

$$\Pr\left(-z_{\alpha/2} \leq \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

# スコア型の信頼区間の構成

---

$$-z_{\alpha/2} \leq \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} \leq z_{\alpha/2}$$

$$\Leftrightarrow (n + z_{\alpha/2}^2)\pi_0^2 - (2n\hat{\pi} + z_{\alpha/2}^2)\pi_0 - n\hat{\pi}^2 \leq 0$$

$\pi_0$ に関してこの二次不等式を解くと

$$\frac{2n\hat{\pi} + z_{\alpha/2}^2 \pm \sqrt{z_{\alpha/2}^2 + 4n\hat{\pi}(1 - \hat{\pi})}}{2(n + z_{\alpha/2}^2)}$$

$$\Pr\left(\frac{2n\hat{\pi} + z_{\alpha/2}^2 - \sqrt{z_{\alpha/2}^2 + 4n\hat{\pi}(1 - \hat{\pi})}}{2(n + z_{\alpha/2}^2)} \leq \pi \leq \frac{2n\hat{\pi} + z_{\alpha/2}^2 + \sqrt{z_{\alpha/2}^2 + 4n\hat{\pi}(1 - \hat{\pi})}}{2(n + z_{\alpha/2}^2)}\right) = 1 - \alpha$$

# Agresti and Coullの信頼区間の構成

---

スコア型の信頼区間を変形すると

$$\frac{2n\hat{\pi} + z_{\alpha/2}^2 \pm \sqrt{z_{\alpha/2}^2 + 4n\hat{\pi}(1 - \hat{\pi})}}{2(n + z_{\alpha/2}^2)}$$

$$= \hat{\pi} \left( \frac{n}{n + z_{\alpha/2}^2} \right) + \frac{1}{2} \left( \frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \right)$$

$\hat{\pi}$  と  $\frac{1}{2}$  の重み付き平均と解釈可能

$$\pm z_{\alpha/2} \sqrt{\frac{1}{n + z_{\alpha/2}^2} \left\{ \hat{\pi}(1 - \hat{\pi}) \left( \frac{n}{n + z_{\alpha/2}^2} \right) + \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) \left( \frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \right) \right\}}$$

$n = 1$  のときの  $\pi = \hat{\pi}$  の分散と  $\pi = \frac{1}{2}$  の分散の重み付き平均

# Agresti and Coullの信頼区間の構成

---

信頼区間の中点

$$\tilde{\pi} = \hat{\pi} \left( \frac{n}{n + z_{\alpha/2}^2} \right) + \frac{1}{2} \left( \frac{z_{\alpha/2}^2}{n + z_{\alpha/2}^2} \right)$$

「2つの分散の重み付き平均」を「重み付き平均  $\tilde{\pi}$  の分散  $\tilde{\pi}(1 - \tilde{\pi})$ 」に置き換えた次式の信頼区間

$$\tilde{\pi} \pm z_{\alpha/2} \sqrt{\frac{\tilde{\pi}(1 - \tilde{\pi})}{n + z_{\alpha/2}^2}} = \tilde{\pi} \pm z_{\alpha/2} \sqrt{\frac{\tilde{\pi}(1 - \tilde{\pi})}{\tilde{n}}}$$

スコア型の性能とWald型の計算の容易さを兼ね備えた方法

95%信頼区間の構成では、試行回数に4を加え、成功回数に2を加えるという単純な置き換えに相当する

ベータ(2, 2)事前分布の下での $\pi$ の事後期待値に相当する

# 手元にあるコインはいかさまコインかどうか

---

- 表が出る確率  $\pi$  は  $1/2$  かどうか
- 実際にコインを  $N$  回投げて、確かめる実験を考える

表 表 表 表 表 表 ... 表  $n$  回

裏 裏 裏 裏 裏 裏 ... 裏  $N - n$  回

- 表が出た割合  $p = \frac{n}{N}$

# コイン投げの実験を仮説検定で検証

---

## □ 仮説の設定

- 表が出る確率は $1/2$ かどうか

## □ 設定した仮説を評価するためのデータを収集

- 実際にコインを $N$ 回投げる

↳  $N$ を設定することをサンプルサイズ設計という

## □ 事前に設定した判定基準に基づき判断

- 表が $m$ 回以下（または以上）のとき、いかさまコインと判断

↳ この判定基準を確率論的に設定する



# 仮説の設定

---

- 「表が出る確率  $\pi$  は  $1/2$  かどうか」を検証するために2つの仮説を設定する
  - 帰無仮説  $H_0: \pi = 1/2$
  - 対立仮説  $H_1: \pi \neq 1/2$
  
- 帰無仮説が成り立つと仮定する
  - 手元にあるコインはいかさまコインではないと仮定する
  - 収集したデータに基づき帰無仮説が成り立つかどうかを判断する

# 設定した仮説を評価するためのデータを収集

---

- 「帰無仮説」と「対立仮説」のどちらが正しいかを判断するために、コインを  $N$  回投げる
  - $N$ （サンプルサイズ）はどう設定すればいいのだろうか？
  
- 検出力に基づいてサンプルサイズを設計する
  - 検出力は検定の精度を表す指標
  - 詳しくは後ほど紹介

# 事前に設定した判定基準に基づき判断

---

## □ 判断基準の考え方

- いかさまコインではないとき  
10回コインを投げれば常に表が5回出るとは限らない
- 表が出る回数は確率的に変動する
- 偶然に出る可能性のある「表の回数」の範囲を考える

## □ この範囲を確率論的に設定する

# 偶然に出る可能性のある「表の回数」の範囲

□ いかさまコインではない（帰無仮説  $H_0: \pi = 1/2$ ）と仮定

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

確率の計算式  ${}_{10}C_x \left(\frac{1}{2}\right)^x \left(1 - \frac{1}{2}\right)^{10-x}$

表の回数3から7である確率は85%以上

表の回数が1以下, または9回以上である確率は2.16%  
表の回数が2以下, または8回以上である確率は10.94%

# 有意水準

---

- 帰無仮説のもとで、5%（または1%）未満でしか起きない事象は偶然ではないと考える
  
- 10回コインを投げた結果
  - 表の回数が1以下，または9回以上の場合 ➡ 偶然ではない
  - 表の回数が2以上，または8回以下の場合 ➡ 偶然である
  
- 棄却域
  - 偶然ではないと考える範囲

# 検定結果の解釈

---

## □ 10回コインを投げた結果

### ■ 表の回数が1以下，または9回以上の場合

- 統計学的に有意と判定
- 帰無仮説を棄却して，対立仮説を採択する
- 「表が出る確率  $\pi$  は1/2ではない」と判断する

### ■ 表の回数が2以上，かつ8回以下の場合

- 統計学的に有意でないと判定
- 帰無仮説を採択する
- 「表が出る確率  $\pi$  は1/2ではない」とはいえないと判断する

「表が出る確率  $\pi$  は1/2である」とは判断できないことに注意！

# 仮説検定には誤りが存在する

- 帰無仮説のもとで5%未満の確率でしか起きない事象は偶然ではないと考えて有意水準を設定
- 裏を返せば，帰無仮説のもとでも，5%未満の確率で生じる事象ということになる
- 第1種の過誤
  - 帰無仮説が正しいときに，誤って帰無仮説を棄却する誤り
  - 第1種の誤りを起こす確率を第1種の過誤確率という

		検定結果	
		帰無仮説が正しいと判断	対立仮説が正しいと判断
真実	帰無仮説が正しい	正しい	第1種の誤り
	対立仮説が正しい	第2種の誤り	正しい

## 2種類の誤り確率

---

- 仮説検定では、第1種の誤りと第2種の誤りが存在
- 有意水準を設定することで第1種の過誤確率を制御している
- 第2種の過誤確率はどのように制御するのか？



# いかさまコインであると仮定

---

- これまでは，帰無仮説（いかさまコインではない）が成り立つと仮定して議論してきた
- いかさまコインであると仮定して，表が出る回数の確率を求める
  
- 検出力
  - 対立仮説が正しいとき，対立仮説が正しいと判断する確率
  - $1 - \text{第2種の過誤確率}$

# 表が出る確率が70%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の  
過誤確率

いかさまコインの場合

表の回数	確率
0	0.0006%
1	0.01%
2	0.15%
3	0.90%
4	3.68%
5	10.29%
6	20.01%
7	26.68%
8	23.35%
9	12.11%
10	2.82%

検出力  
(無視可能)

第2種の  
過誤確率

検出力

# 表が出る確率が80%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の  
過誤確率

いかさまコインの場合

表の回数	確率
0	1.024e-05%
1	0.0004%
2	0.007%
3	0.08%
4	0.55%
5	2.64%
6	8.81%
7	20.13%
8	30.20%
9	26.84%
10	10.74%

検出力  
(無視可能)

第2種の  
過誤確率

検出力

# 表が出る確率が90%のいかさまコイン

いかさまコインではない場合

表の回数	確率
0	0.1%
1	0.98%
2	4.39%
3	11.72%
4	20.51%
5	24.51%
6	20.51%
7	11.72%
8	4.39%
9	0.98%
10	0.1%

第1種の  
過誤確率

第1種の  
過誤確率

いかさまコインの場合

表の回数	確率
0	1.024e-05%
1	9e-07%
2	3.645e-05%
3	0.0009%
4	0.01%
5	0.15%
6	1.12%
7	5.74%
8	19.37%
9	38.74%
10	34.87%

検出力  
(無視可能)

第2種の  
過誤確率

検出力

# 12回コインを投げた結果

いかさまコインではない場合

表の回数	確率
0	0.02%
1	0.29%
2	1.61%
3	5.37%
4	12.08%
5	19.34%
6	22.56%
7	19.34%
8	12.08%
9	5.37%
10	1.61%
11	0.29%
12	0.02%

第1種の  
過誤確率

第1種の  
過誤確率

表が出る確率が80%のいかさまコインの場合

表の回数	確率
0	4.096e-07%
1	1.96608e-05%
2	0.0004%
3	0.006%
4	0.05%
5	0.33%
6	1.55%
7	5.32%
8	13.29%
9	23.62%
10	28.34%
11	20.62%
12	6.87%

検出力  
(無視可能)

第2種の  
過誤確率

検出力

# 検出力の特徴

---

□ 帰無仮説からの乖離の程度に依存する

■ コインのいかさまの程度（表の出る確率）に依存する

表の出る確率	検出力
70%	14.93%
80%	37.58%
90%	73.61%

□ サンプルサイズ  $N$  に依存する

コイン投げの回数	表の出る確率	検出力
10	80%	37.58%
12	80%	55.83%

# 検定結果の解釈

---

- 帰無仮説を棄却し，対立仮説を支持
  - この判断が間違っている確率は $\alpha$ 以下であることを保証
  
- 帰無仮説を採択し，帰無仮説を支持
  - この判断が間違っている確率は制御されていない
  - 積極的に帰無仮説が正しいことを主張することは危険
  - 帰無仮説が正しいことを主張するには，  
データ収集前にサンプルサイズ設計を行い  
第2種の過誤確率を制御する必要がある

# 検定の手順

---

- Step 1: 帰無仮説と対立仮説を設定する
- Step 2: 帰無仮説が成り立つと仮定する
- Step 3: 得られた標本（データ）の結果が，帰無仮説のもとで矛盾するか否かを判断する
- Step 4:
  - 標本の結果と帰無仮説が矛盾する場合，帰無仮説を棄却（対立仮説を主張）
  - 標本の結果と帰無仮説が矛盾しない場合，**帰無仮説を否定する根拠がない**

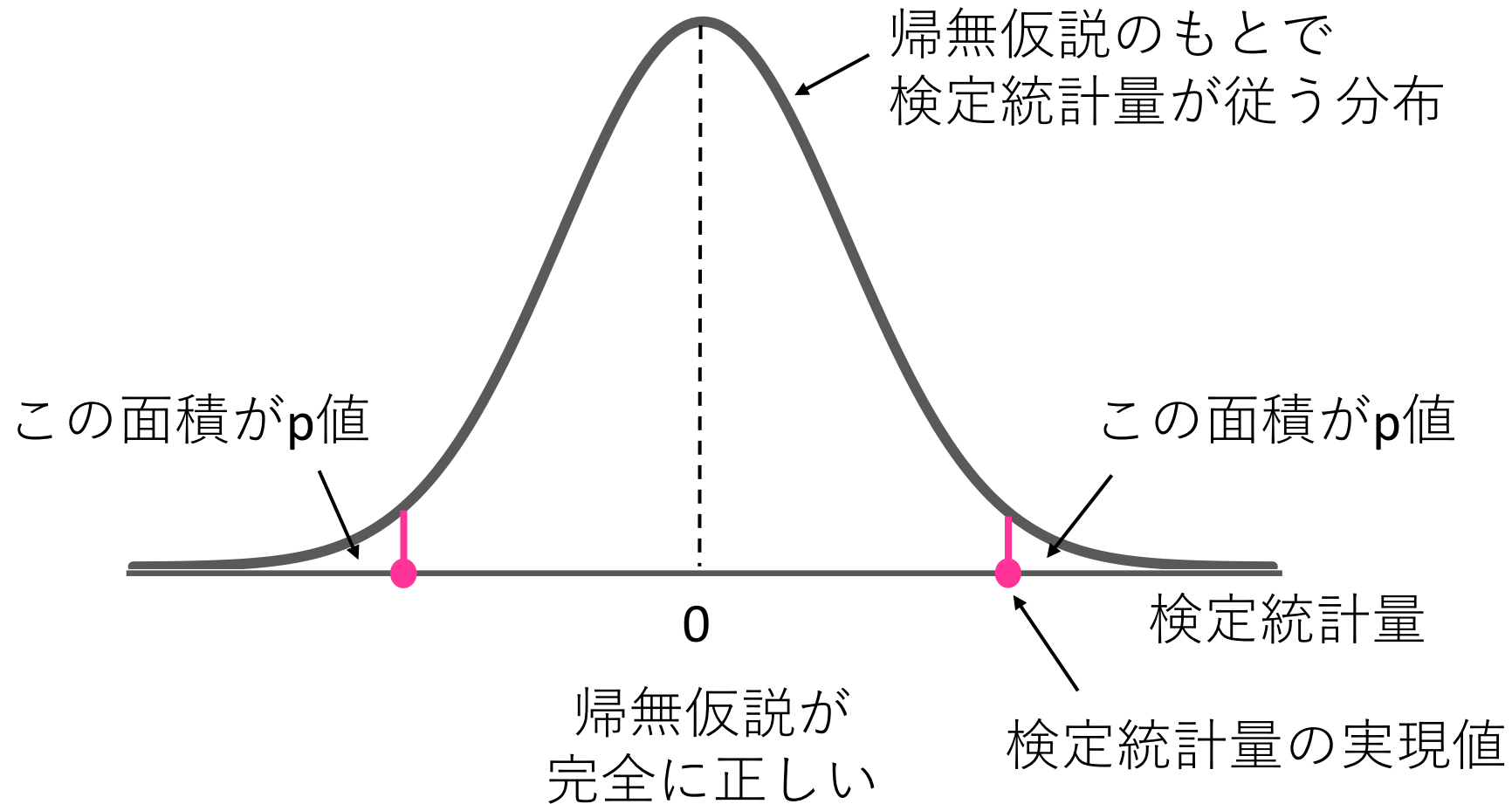


# p値

---

- 帰無仮説が正しいときに、標本から得られた結果が偶然生じたものであるかを表す確率のような指標
- 帰無仮説が正しいときに、検定統計量がどの程度の確率で得られるものを計算している
- p値が設定した有意水準よりも小さいとき、帰無仮説を棄却し、対立仮説を支持する

# 検定統計量とp値の関係



# 信頼区間と仮説検定の関係

---

## □ 信頼区間

- 仮説検定で棄却されない母数の集合

## □ 帰無仮説 $H_0: \pi = \pi_0$

## □ 対立仮説 $H_1: \pi \neq \pi_0$

## □ 検定統計量 (Wald 型) $Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})/n}}$

## □ 棄却限界値: $z_{\alpha/2}$

## □ 帰無仮説が棄却されない条件

- $|Z| \leq z_{\alpha/2} \ (\Leftrightarrow -z_{\alpha/2} \leq Z \leq z_{\alpha/2})$

# Example. 1

---

- 試行回数  $n = 10$
- 成功回数  $x = 2$
- 帰無仮説  $H_0: \pi = \pi_0 = 0.5$
- 対立仮説  $H_1: \pi \neq 0.5$
- SAS データセット
- FREQ プロシジャ

```
data data;  
  input x n;  
  cards;  
  1 2  
  0 8  
run;
```

```
proc freq data = data order = data;  
  weight n;  
  table x / binomial;  
run;
```

# 出力結果

x = 1 の二項分布の比率

比率	0.2000
漸近標準誤差	0.1265
95% 信頼下限	0.0000
95% 信頼上限	0.4479

Wald 型 (棄却される)

正確な信頼限界	
95% 信頼下限	0.0252
95% 信頼上限	0.5561

棄却されない

H0: 母比率 = 0.5 に対する検定

帰無仮説が正しいもとの漸近標準誤差

Z

片側  $Pr < Z$

両側  $Pr > |Z|$

0.1581

-1.8974

0.0289

0.0578

$$\sqrt{\frac{0.5(1 - 0.5)}{10}}$$

スコア 型  
(棄却されない)

## Example. 2

---

- 試行回数  $n = 10$
- 成功回数  $x = 2$
- 帰無仮説  $H_0: \pi = \pi_0 = 0.5$
- 対立仮説  $H_1: \pi \neq 0.5$
- SAS データセット
- FREQ プロシジャ

```
data data;  
  input x n;  
      cards;  
      1 2  
      0 8  
run;
```

```
proc freq data = data order = data;  
  weight n;  
  table x / binomial (var = sample);  
run;
```

# 出力結果

x = 1 の二項分布の比率

比率	0.2000
漸近標準誤差	0.1265
95% 信頼下限	0.0000
95% 信頼上限	0.4479


Wald 型 (棄却される)

正確な信頼限界	
95% 信頼下限	0.0252
95% 信頼上限	0.5561

棄却されない

H0: 母比率 = 0.5 に対する検定

ASE (Sample)	0.1265
Z	-2.3717
片側 Pr < Z	0.0089
両側 Pr >  Z	0.0177


$$\sqrt{\frac{0.2(1 - 0.2)}{10}}$$

Wald 型 (棄却される)

# Example. 3

---

- 試行回数  $n = 10$
- 成功回数  $x = 2$
- 帰無仮説  $H_0: \pi = \pi_0 = 0.5$
- 対立仮説  $H_1: \pi \neq 0.5$
- SAS データセット
- FREQ プロシジャ

```
data data;  
  input x n;  
      cards;  
      1 2  
      0 8  
run;
```

```
proc freq data = data order = data;  
  weight n;  
  table x / binomial (all);  
run;
```



# 出力結果

タイプ 95% 信頼限界

Wald	0.0000	0.4479
Wilson	0.0567	0.5098
Agresti-Coull	0.0459	0.5206
Jeffreys	0.0441	0.5028
Clopper-Pearson (Exact)	0.0252	0.5561

← スコア型  
(棄却されない)

H0: 母比率 = 0.5 に対する検定

帰無仮説が正しいもとの漸近標準誤差	0.1581
Z	-1.8974
片側 $\Pr < Z$	0.0289
両側 $\Pr >  Z $	0.0578

← スコア型  
(棄却されない)