統計解析応用研究
推測統計

丸山 祐造 Yuzo Maruyama

神戸大学　大学院経営学研究科

統計的モデル

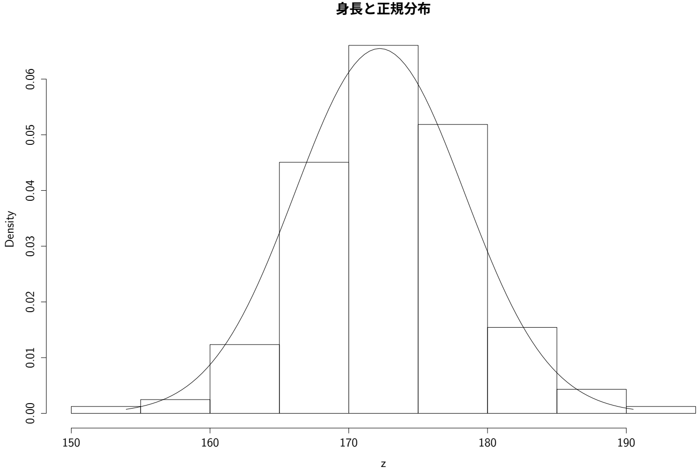
確率論と推測統計学

- 確率論：与えられた確率分布の性質を調べる
- 推測統計学：確率論の応用 $\Leftarrow$ どういう意味？
  $\Uparrow$ データが得られたときに複数の（無限個の）確率分布の候補を用意し，それらの候補の中からデータの分布にあった確率分布を選ぶ
- 統計モデル（統計的モデル，確率モデル）
  - 「確率分布の候補の集合」，あるいは「その作り方」
  - データの分布を説明したり，データから有意義な結論を引き出すための道具

推測統計学の考え方 I

- データは確率変数の実現値とみなす
  実現値 $x_1, \ldots, x_n$
- データがサイコロをふってその実現値を記録した場合，「データは（離散）確率変数の実現値とみなす」ことに異論はないはず
- 身長の分布は正規分布？
  ⇑ csv ファイルにある学生のデータでは，よく合っているように見える

推測統計学の考え方 II



推測統計学の考え方 III

- 様々な異論・反論
  - 人間の身長の決定メカニズムは複雑な現象であるはず．正規分布からの乱数発生と同じであるはずはない
  - 正規分布においては「負の値を取る」確率や「３メートルを超える」確率はゼロではない
  - 分布は連続？小数点第一位までなら離散！
  - 最大値，最小値を制御出来る確率分布に当てはめるか？
  - 個々の遺伝や生育環境などを考慮すべきでは？

推測統計学の考え方 IV

- 現実の単純化としての正規分布モデル

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

- 正規分布モデルを採用する場合には最低限の層別が必要
  - 性別
  - 人種，国（日本人とオランダ人）
  - 年齢（7歳，１８歳，二十歳以上など）

推測統計学の考え方 V

- 身長は正規分布に従う $\Rightarrow$ 身長は正規分布に従わない．ただし，一定の同質性が見込める集団で身長のデータを収集したとき，そのデータが近似的に正規分布に従うと想定する統計的モデルには有用性がある．
  $\Uparrow$ All models are wrong; but some are useful（全てのモデルは間違っている、だが中には役立つものもある）

推測統計学の考え方 VI

- 統計的モデルの導入の有用性
  - 「そのときたまたま測定された $n = 324$ 人」と考えるのではなく，「日本成人男子の中」から $n = 324$ 人をランダムに抽出したという仮定
  - 「日本の成人男子」全体の中で，身長が 190cm の人が上位何 $\%$ なのかを「推定」出来る

$$\int_{190}^{\infty} \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp\left(-\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2}\right) \mathrm{d}x$$

$\Uparrow$ $n = 324$ のデータによる標本平均 $\hat{\mu}$ と標本分散 $\hat{\sigma}^2$

教科書に見る統計的モデル

- 【藪 243p】ある会社が販売している牛乳 1ℓ に含まれる脂質量は正規分布しているとします．
- 【久国 165p】ある町の冬の 1 日当たりの火災発生数が平均 λ のポアソン分布 $Po(\lambda)$ に従うとする．
- 【久国 166p】20 歳代男性の身長が平均 μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ に従うとする.
- 【倉星 208p】大学生男子 20 人の身長を計測する．正規母集団 $N(\mu, \sigma^2)$ からの無作為標本と見なす．

母集団と標本 I

- 最も簡単な確率モデル：壺
- 大きな壷の中にボールが N 個入っていて，ボールには実数が書かれているとする
- 復元抽出：壷の中からボールを一つ取り出して，そのボールに書かれている数を記録し，壷の中に戻す
- これを計 n 回繰り返す
- 用語
  - 母集団：壷の中のボールに書かれた数全体
  - 母集団の大きさ：母集団の総数 N
  - 標本：抜き出されたボールに書かれた数
  - 標本の大きさ：抜き出す回数 n

母集団と標本 II

- ボールを１個だけ抽出．どのボールが抽出される確率も同じ $1/N$
  $\Uparrow$ 単純無作為抽出
- 繰り返す場合には「取り出したボールを壺の中に戻す，戻さない」の区別
  - 復元抽出：戻す場合
  - 非復元抽出：戻さない場合
    $\Uparrow$ i 番目の抽出は，既に抽出された $i-1$ 個以外の $N-i+1$ 個から，確率 $1/(N-i+1)$ で抽出される

母集団と標本 III

- 以下では復元抽出にフォーカス
- 例
  - 福引で球が 100 個入っており，一等賞が 10 本，二等賞が 20 本，はずれ (三等賞) が 70 本
  - 復元抽出で n 個抽出したときの値 $X_1, \ldots, X_n$ は互いに独立で同一の分布

$$P(X_i = 1) = 0.1,\ P(X_i = 2) = 0.2,\ P(X_i = 3) = 0.7$$

  X_i の分布は壺の中の球に書いてある数の相対頻度に一致
  - 母集団分布：壺の中のボールに書かれた数の分布
- 復元抽出では，$X_1, \ldots, X_n$ は互いに**独立で同一の分布**（母集団分布）に従う

母集団と標本 IV

- $X_1, \cdots, X_n$ は，i.i.d.(独立で同一の分布，independently identically distributed) 確率変数
- 壷のモデルでは，母集団の大きさ N が有限の場合
- $N \to \infty$：母集団分布が連続の場合 (分布が密度関数を持つような場合) も極限として扱える
- 用語：無限母集団，有限母集団
  $\Uparrow$ 無限母集団においては，復元抽出と非復元抽出の区別なし

母集団と標本 V

- 極限操作により，任意の分布を壺のモデルで近似できるから，壺のモデルと i.i.d. 確率変数を同一視可能
- 以上の観点から，分布 F に従う n 個の i.i.d. 確率変数を「分布 F からの大きさ n の標本」という

標本調査 I

- ここまでは壺からの抽出を福引のようなランダム化の方法と考えた
- 母集団の様子を調査するための標本抽出を考える
- 例
  - A 社が新製品を売り出す際に，試供品を提供して消費者の反応調査
  - 母集団：その製品を広く売り出したときに関心をもってあろう消費者の全体
  - 試供品を提供して反応を調べることは，母集団から標本を抽出して標本を調べていること
  - この会社の知りたいのは，たとえば母集団においてこの新製品を買ってくれる消費者の割合 p

標本調査 II

- 標本調査：母集団における様々な特性値を知るために標本を抽出して調査をすること
- 記述統計においてもデータの性質，特徴を明らかにするために平均，分散，相関係数などの値 (統計量) を計算する
- 違い
  - 記述統計ではデータを所与としているので，統計量 $t(x_1, \cdots, x_n)$ は一つの値に決まる
  - 推測統計では $X_1, \ldots, X_n$ が確率変数なので，統計量 $t(X_1, \ldots, X_n)$ も確率変数

標本調査 III

- 例：首相支持率
  - 日本国の人口 N 人．各個人に対応するボールが入った壺
    ボール：支持，不支持のラベル（支持 1，不支持 0）
  - 全支持者数は M とする
  - 全て調べるのは手間．n 個抜き出す．$X_1, \ldots, X_n$
  - 支持率 $\sum_{i=1}^{n} X_i/n$．支持率は統計量の一つ
  - 同時期に同じ設定でもう一度調査をしたとき（標本）支持率は違うはず

標本調査 IV

- 上で計算した（標本）支持率は，真の支持率 M/N に近いことを目論んでいるが，通常一致しない
- 繰り返し同じ調査を行ったと仮定すれば，その標本支持率は真の支持率のまわりに誤差を伴って分布
- 標本分布：統計量の分布
- 推測統計学の問題設定
  - 母集団の分布は未知を仮定
  - 得られた標本をもとに適当な統計量を構成 $\Rightarrow$ 母集団分布に関する推測

標本調査 V

- 例：統計的推定
  - 福引担当者しかボールの数, 各等の比率を知らない
  - 福引を引き終わった人 50 人に結果を尋ねて集計する
    一等の相対頻度 4/50，二等の相対頻度 16/50，三等の相対頻度 30/50
  - 各等の比率をこれらの相対頻度であると推測

標本調査 VI

- 例：統計的検定
  - 事前の広告ビラで，「10 人に 1 人が 1 等，各等の比率は $1:2:7$」だと謳っていた．しかし上の例では 50 人中 4 人しか一等がいない．これは嘘ではないか？
  - 嘘かどうかを検証する

分布族とパラメータ I

- 壺の中の確率関数や確率密度関数の想定が余りに柔軟すぎると，結局何も分からない,,,

$$\text{確率関数}\quad p(x) \geq 0, \quad \sum_{x\in\mathbb{R}} p(x) = 1$$

$$\text{確率密度関数}\quad f(x) \geq 0, \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

- 既に学んだ「主要な分布」は有限個のパラメータで確率分布を制御
- 分布族：正規分布や二項分布は，ある特定の分布を指していない．正確には「分布の集合」

分布族とパラメータ II

- **正規分布族：平均 μ と分散 σ^2 の指定で分布が決まる**

$$\{N(\mu,\sigma^2)|\ -\infty < \mu < \infty,\ \sigma^2 > 0\}$$

(μ,σ^2)**：正規分布族の「パラメータ」あるいは「母数」**

- **二項分布族：試行回数 n は与えた下で，成功確率 p の指定で分布が決まる**

$$\{\mathrm{Bin}(n,p)|0 \le p \le 1\}$$

成功確率 p：二項分布族のパラメータ

分布族とパラメータ III

- 統計学では，母集団と標本で対応する特性値について「母」,「標本」をつけて区別
- 母平均：$\mu = E(X)$，標本平均 $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$
- 標本調査では母集団の特性値に興味がある
- 母集団の分布を特定すれば，パラメータの値を知ること $=$ 母集団分布を知ること

分布族とパラメータ IV

- 統計的モデル：母集団の分布として特定化された分布族（限定された集合）
- 重要な仮定：データ $x_1, \ldots, x_n$ は「分布族に含まれるある特定の分布」からの標本
  - その分布を「真の分布」という
  - 真の分布のパラメータを「パラメータの真の値」という
  - 未知母数：パラメータの真の値は未知
- 統計的モデルの仮定の下では，未知母数に関する推測 $=$ 母集団分布に関する推測

分布族とパラメータ V

- より一般的な設定
  - パラメータ θ
  - 確率変数 X の確率関数あるいは密度関数 $f(x, \boldsymbol{\theta})$
  - 母数空間 Θ：θ のとりうる集合
  - 正規分布の場合

$$\boldsymbol{\theta} = (\mu, \sigma^2), \ \boldsymbol{\Theta} = \mathbb{R} \times \mathbb{R}_+ = (-\infty, \infty) \times (0, \infty)$$

  - 密度関数または確率関数を用いて表した分布族

$$\{f(x, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

  - パラメータの真の値：$\theta_0 \in \boldsymbol{\Theta}$

分布族とパラメータ VI

- 分布の特定化の欠点
  - 数学的便宜のため
  - 実際のデータの分布が正規分布族，二項分布族，ポアソン分布族のような数学的に都合の良い分布とは限らない
  - 例えば分布の中心に対して，左右の非対称性があるデータはどのように統計的モデルを構成するのか？

分布族とパラメータ VII

- 分布の特定化の利点
  - 正規分布などの都合の良い分布を仮定すれば，「標準的な」手法を用いてデータが解釈できる
  - 機械学習の最新の手法であっても，単純な統計モデルから出発してその複雑化の道筋を辿ることで理解が深まる

- 「モデル」という言葉は「模型」すなわち現実の現象の単純化という意味合いを含む
  $\Uparrow$ 有限個のデータをもとに確率分布についての推測をしようというのだから，単純化は不可欠！

やや複雑な統計的モデル I

- （単）回帰分析の統計的モデル

$$y_i = \alpha + \beta x_i + \epsilon_i, \ \epsilon \sim N(0, \sigma^2)$$
$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

- 観測値 y は $\alpha + \beta x$ の周りに誤差を伴って発生すると仮定
- $(x_1, \ldots, x_p)$ はランダムでなく定数
- 最小二乗推定量

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

$\Uparrow$ 確率的に変動する

やや複雑な統計的モデル II

- **離散選択モデル（ロジスティック回帰）**
  - **目的変数が２値（ダミー変数** 0 or 1**）．企業の倒産** or **存続．患者の死亡** or **生存．肺がんの発症（罹患）** or not
  - **目的変数に影響があると考えられる説明変数** x **あり**
  - $\alpha + \beta x$ **で** y **を予測したい．**
  - **いくつかの問題：**y **は** $0, 1$ **という２値のみ．しかし，通常の最小二乗法では** $\hat{\alpha} + \hat{\beta} x$ **は** $[0, 1]$ **の外にはみ出るだろう**
  - **例えば** x **の値が大きい（アルコール摂取量）ほど罹患率は高い現実を想定する．しかし，例外もある（大酒飲みだからと言って，必ず肺がんになるわけではない）**

やや複雑な統計的モデル III

- 統計的モデルによる対処法（処方箋）がある！
- 累積分布関数（$F(-\infty) = 0,\ F(\infty) = 1$，$F()$ は増加関数）の利用．
- $F(\alpha + \beta x)$ は $(0, 1)$ の区間に入る．それを成功確率としてベルヌーイ試行．y はその実現値が観測されていると見なす
  $\Uparrow$ 目的変数 y の観測値は，対応する x の値によって，表の出る確率が変化するコインを投げたときの裏表に対応していると見なす
- そのようなモデルで，n 人のデータをもとに $\hat{\alpha}, \hat{\beta}$ を推定
- それを用いて $n+1$ 人目の患者（企業）の x の値に基づいて，罹患率，倒産確率を予測する
- F としてロジスティック分布を使う場合にロジスティック回帰分析と呼ばれる

推定問題の設定 I

- 有限個のデータ $x_1, \ldots, x_n$
- 推測統計：背後に統計的モデルを想定

$$\{f(x,\theta)\,|\,\theta \in \Theta\}$$

複数の要素を持つ集合 $\Uparrow$

- 真の θ_0 が存在すると想定
- $f(x,\theta_0)$ の入った壺から玉を抽出 x_1，玉を戻して壺から玉を抽出 x_2，これを n 回繰り返す
- 確率変数の実現値として得られるのがデータ $x_1, \ldots, x_n$

推定問題の設定 II

- 得られた有限個のデータ $x_1, \dots, x_n$ をもとに
  - 【点推定】θ_0 を一点で言い当てる
  - 【区間推定】θ_0 が入っているであろう区間を構成
  - 【仮説検定】Θ を Θ_0 と Θ_1 に分けたとき

$$\Theta_0 \cup \Theta_1 = \Theta, \ \Theta_0 \cap \Theta_1 = \emptyset$$

θ_0 がどちらに入っているか判定

推定問題の設定 III
入門レベルの統計的モデルの具体例

- **ベルヌーイ分布 成功確率** p

$$P(X=0)=1-p, \quad P(X=1)=p$$

- **カテゴリカル分布 確率** $(p_1,\ldots,p_k)$

$$P(X=1)=p_1, P(X=2)=p_2, \ldots,$$

$$P(X=k)=p_k=1-\sum_{i=1}^{k-1} p_i$$

$\Uparrow$ **多項分布の試行回数が１回の分布**

推定問題の設定 IV

- $\{N(\mu,1)\,|\,\mu\in\mathbb{R}\}$　**未知母数** μ
- $\left\{N(\mu,\sigma^2)\,|\,\mu\in\mathbb{R},\sigma^2\in\mathbb{R}_+\right\}$　**未知母数** μ,σ^2
- $\mathcal{MV}=\{\mathrm{E}[X],\mathrm{Var}[X]$ **が存在する分布**$\}$
  - **離散，連続問わず**
  - **確率関数や確率密度関数の関数形は特定せず**
  - **分布の基本的な指標である期待値と分散が存在．それらの推測に興味がある**
  - **主に点推定において想定される**

推定量

- 有限個のデータ $x_1, \ldots, x_n$ と書いていたが，「『確率変数』の『実現値』」という側面の「確率変数」に注目して標本 $X_1, \ldots, X_n$ と書く
  $\Uparrow$ 全く同じ状況で壺から n 個抽出された $x'_1, \ldots, x'_n$ は，$x_1, \ldots, x_n$ と同じではない
- 統計量：標本 $X_1, \ldots, X_n$ の関数
  未知のパラメータに依存しないことが重要
- 推定量：点推定における言い当てに使われる統計量
  $\Uparrow$ 「推定量」の定義には「良さ」への言及なし

I.I.D. 確率変数と推定論 I

- 第一回講義で学んだ標本平均，標本分散は推測統計の枠組みでどのような意味を持つのか

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

- 統計モデル $\mathcal{MV}$：「期待値 μ，分散 σ^2 の確率分布」
  $\Uparrow$ $\mathrm{E}[X_i] = \mu$，$\mathrm{Var}[X_i] = \sigma^2$
- 統計モデル $\mathcal{MV}$ に含まれる「ある確率分布」からの n 個の独立な実現値として，$x_1, \ldots, x_n$ を得る

I.I.D. 確率変数と推定論 II

- 「確率変数」的な側面を強調して，$X_1, \ldots, X_n$ は，期待値 μ，分散 σ^2 である確率分布からの独立で同一な確率変数（i.i.d. 確率変数）

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

- 既に学んだこと
  - $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ は確率変数
  - その期待値と分散 $\mathrm{E}[\bar{X}] = \mu, \quad \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}$

I.I.D. 確率変数と推定論 III

- $\mathrm{E}[\bar{X}] = \mu$ の推測統計の観点での意味付け
  - $\bar{X}$ は μ の「不偏推定量」
  - 「不偏性」は推定量の良さの指標の一つ．他にもある
- 推定論の２つの内容
  - 推定量の作り方（モーメント法，最尤法）
  - 作った推定量の性質（不偏性，平均二乗誤差，一様最小分散不偏推定量）

I.I.D. 確率変数と推定論 IV

- $X_1, \ldots, X_n$ の関数で

$$\mathrm{E}[T(X_1, \ldots, X_n)] = \sigma^2$$

となるような $T(X_1, \ldots, X_n)$ はあるのか？

- $\mathrm{Var}[\bar{X}] = \dfrac{\sigma^2}{n}$ から $\bar{X}$ が μ の周りにばらつく程度を見積もれる
$\Uparrow$ その意味でも σ^2 を推定したい

I.I.D. 確率変数と推定論 V

- **標本分散の関数形が参考になる** math.pdf 3.1 節

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

- $\mathrm{E}\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = \sigma^2$
- $\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ **は不偏分散と呼ばれる**
- $\mathrm{E}\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}\right] = \frac{n-1}{n}\sigma^2$
  n **で割ると，「平均的には」** σ^2 **を小さめに推定**

平均二乗誤差と不偏推定量 I

- 【点推定】未知パラメータ θ の値を統計量 $\hat{\theta}(X_1, \ldots, X_n)$ によって出来るだけ正確に言い当てる
- 【推定量】 点推定のための統計量
- 【平均二乗誤差】 小さいほど良い

$$\mathrm{E}[(\hat{\theta}(X) - \theta)^2] \quad \text{略記 } X = (X_1, \ldots, X_n)$$

平均二乗誤差と不偏推定量 II

- **その分解** math.pdf 3.4 節

$$
\begin{aligned}
&\mathrm{E}[(\hat{\theta}(X)-\theta)^2] \\
&= \mathrm{E}[\{\hat{\theta}(X)-E(\hat{\theta})\}^2] + \{E(\hat{\theta})-\theta\}^2 \\
&= \mathrm{Var}(\hat{\theta}(X)) + b(\theta)^2
\end{aligned}
$$

- **推定量 $\hat{\theta}(X)$ のバイアス**

$$b(\theta) = \mathrm{E}(\hat{\theta}(X)) - \theta \quad \text{略記 } X = (X_1, \ldots, X_n)$$

平均二乗誤差と不偏推定量 III

- **不偏推定量：バイアスが常に** 0

$$\mathrm{E}[\hat{\theta}(X)] = \theta, \quad \forall\theta \in \Theta$$

$\Uparrow$ $E_\theta(\hat{\theta}) > \theta$ **であれば，**θ **を推定するのに平均的には** θ **より大きい値になる**

平均二乗誤差と不偏推定量 IV

- 不偏性の定義で，「任意の」という部分が必要な理由
  - パラメータ θ には真の値 θ_0 が存在する（というのが統計学の立場）
  - そうすると，$\theta = \theta_0$ のときだけ当たれば良い？
  - もちろん真の値は知らない．だから推定しようとしている
  - 背後にあるパラメータ θ がどんな値であっても，妥当に推し量りたい

平均二乗誤差と不偏推定量 V

- **不偏推定量は一般に複数存在**
  - $\mathrm{E}[X]$ **の推定**

$$\frac{X_1 + \cdots + X_{n-1}}{n-1}, \frac{X_1 + X_2}{2}, \text{etc}$$

  - **さらに** $\mathrm{E}[X]$ **を中心とした対称分布の場合には**

$$\text{median}(X_1, \ldots, X_n)$$

も不偏推定量

平均二乗誤差と不偏推定量 VI

- 不偏推定量であれば，平均二乗誤差は分散に一致

$$\mathrm{E}[(\hat{\theta}(X)-\theta)^2]=\mathrm{E}[(\hat{\theta}(X)-\mathrm{E}[\hat{\theta}(X)])^2]$$

- 不偏性を持つ推定量のクラスに限れば分散を最小にする推定量が望ましい
- 一様最小分散不偏推定量 $\hat{\theta}^*$

$$\mathrm{Var}(\hat{\theta}^*)\leq \mathrm{Var}(\hat{\theta}),\quad \forall\theta,\hat{\theta}\quad (\text{ただし } \mathrm{E}[\hat{\theta}]=\theta)$$

常に存在するわけではないが，いくつかの分布とその母数の推定については，存在が知られている

平均二乗誤差と不偏推定量 VII

- **一様最小 MSE 不偏推定量** $\hat{\theta}^*$

$$\mathrm{MSE}(\hat{\theta}^*) \leq \mathrm{MSE}(\hat{\theta}), \quad \forall \theta, \hat{\theta} \quad (\text{ただし } \mathrm{E}[\hat{\theta}] = \theta)$$

とも言い換えられる

- **一様最小 MSE~~不偏~~推定量** $\hat{\theta}^*$

$$\mathrm{MSE}(\hat{\theta}^*) \leq \mathrm{MSE}(\hat{\theta}), \quad \forall \theta, \hat{\theta} \quad (\cancel{\text{ただし } \mathrm{E}[\hat{\theta}] = \theta})$$

そんな完璧な推定量は存在しない．

フィッシャー情報量と情報量不等式 I

- 与えられた不偏推定量が一様最小分散不偏推定量（UMVUE）であることを示す方法の一つ
  $\Rightarrow$ フィッシャー情報量に基づく情報量不等式
- $X_1, \ldots, X_n$：θ によって制御される分布族（確率関数または確率密度関数として $f(x, \theta)$）からの i.i.d. 標本
- θ の推定問題

フィッシャー情報量と情報量不等式 II

- **定理（情報量不等式）：**$\hat{\theta}$ **を** θ **の不偏推定量ならば**

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$ **不偏推定量の分散の下限**

$\Uparrow$ **フィッシャー情報量** $I(\theta) = \mathrm{E}\left[\left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2\right]$

- **使い方**
  1. **ある不偏推定量** $\hat{\theta}^*$ **が手元にある**
  2. **その分散を計算してみる**
  3. $I(\theta)$ **を計算してみる**

フィッシャー情報量と情報量不等式 III

4. 分散と $\dfrac{1}{nI(\theta)}$ が一致すれば，$\hat{\theta}^*$ が UMVUE

- 具体例 math.pdf 3.2 節
  - 二項分布 $X \sim \mathrm{Bin}(n, p)$．$p$ の推定．$\hat{p} = X/n$ が UMVUE
  - $X_1, \ldots, X_n$ を正規分布 $N(\mu, 1)$ からの $i.i.d.$ 標本とする．μ の推定量 $\bar{X}$ が UMVUE

モーメント法 I

- 適当な母集団からの標本抽出，$X_1, \ldots, X_n$
  $X_1, \ldots, X_n$ は独立で同一な分布に従う確率変数
- 興味あるパラメータ：母集団分布の平均 $E[X]$

$$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$$ 不偏推定量

- 興味あるパラメータ $E[X^2]$

$$\frac{X_1^2 + \cdots + X_n^2}{n}$$ 不偏推定量

モーメント法 II

- **母集団分布の分散 $\sigma^2 = E[(X-\mu)^2]$ の推定**
  - μ **が既知ならば，**

$$\frac{(X_1-\mu)^2+\cdots+(X_n-\mu)^2}{n} \quad \text{不偏推定量}$$

  - **普通は μ を知らないので，$\hat{\mu}$ を代入**

$$\frac{(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n} \ \text{代入をプラグインともいう}$$

  - **ただし，このままでは不偏性を持たない**

$$\frac{(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n-1} \quad \text{不偏推定量}$$

モーメント法 III

- 「モーメント推定量」という場合には「不偏性」を気にせず提案したものを指すことが多い

共分散 $E[(X-\mu_x)(Y-\mu_y)]$ の推定 math.pdf 3.5 節

- 母集団である壺からの２玉串刺し団子の抽出
- 母集団分布
  - $\mathrm{E}[X]=\mu_x,\ \mathrm{E}[Y]=\mu_Y$
  - $\mathrm{Var}[X]=\mathrm{E}[(X-\mu_X)^2]=\sigma_X^2$
  - $\mathrm{Var}[Y]=\mathrm{E}[(Y-\mu_Y)^2]=\sigma_Y^2$
  - $\mathrm{COV}(X,Y)=\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]=\sigma_{XY}$

モーメント法 IV

- 母集団分布からの独立同一な n 個の標本（確率変数ベクトル）$(X_1, Y_1), \ldots, (X_n, Y_n)$ があるとする
- 共分散のモーメント推定量：μ_X，μ_Y 既知

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y) \quad \text{不偏推定量}$$

普通は μ_X，μ_Y 未知

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \text{ プラグイン推定量}$$

$\Uparrow$ $n-1$ で割ると不偏性が満たされる

モーメント法 V

相関係数 $\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}$ のモーメント推定量

- **既に得られたモーメント推定量をプラグイン**

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

モーメント推定量は不偏ではない．不偏性の議論は難しい

モーメント法 VI

モーメント推定の一般論

- 推定したいパラメータが期待値 $E[g(X)]$ で表現可能

$$\frac{g(X_1)+\cdots+g(X_n)}{n}$$ 不偏推定量

- $\begin{cases} \text{推定したいパラメータ } E[g(X,\theta_1)] \\ \theta_2 \text{のモーメント推定量がある} \end{cases}$ ならば

$$\frac{g(X_1,\hat{\theta}_1(X_1,\ldots,X_n))+\cdots+g(X_n,\hat{\theta}_1(X_1,\ldots,X_n))}{n}$$

モーメント法 VII

- 欠点
  - 興味のあるパラメータが簡単な期待値で表現できるとは限らない
  - 場当たり的な手法であることも否めない
  - 平均二乗誤差？不偏性？
- より汎用的な方法として，最尤推定量がある

尤度関数と最尤推定量 I

- **最尤推定量**
  - **不偏推定量が存在しない場合や，存在しても合理的とは言えない場合に推定量を導く一般的な方法**
  - **統計的モデルが複雑であれば，最尤推定量が唯一の実用的な推定量である場合が多い**
- $X_1, \ldots, X_n$ **：** θ **によって制御される分布族（確率関数または確率密度関数として** $f(x, \theta)$**）からの** i.i.d. **標本．同時確率密度 (または同時確率) 関数**

$$f(\boldsymbol{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

尤度関数と最尤推定量 II

- θ の推定問題
- 尤度関数：$f(x, \theta)$ を θ の関数と見たもの

$$L(\theta) = L(\theta, \boldsymbol{x}) = f(\boldsymbol{x}, \theta)$$

- 最尤推定量は尤度関数を最大にする θ の値を推定値とする推定量

$$\max_{\theta} L(\theta) = L(\hat{\theta}),\ \max_{\theta} L(\theta, \boldsymbol{x}) = L(\hat{\theta}(\boldsymbol{x}), \boldsymbol{x})$$

- 同時確率密度 (または同時確率) 関数と尤度関数の読み取り方の対比に注意

尤度関数と最尤推定量 III

- **最尤推定量の良さ：n が大きいときに一様最小分散不偏推定量に相当する性質（詳細略）**
- **具体例（解析的に解ける場合）** math.pdf 3.3 節
  - **ベルヌーイ分布の成功確率 p の最尤推定量は**?

$$\hat{p} = \sum_{i=1}^{n} \frac{Y_i}{n}$$

  - $X_1, \ldots, X_n$ **を正規分布** $N(\mu, 1)$ **からの** $i.i.d.$ **標本**

$$\hat{\mu} = \bar{X}$$

尤度関数と最尤推定量 IV

- **具体例（解析的に解けない場合，数値的最適化）**
  - **コーシー分布の位置が未知の場合**

$$f(x;\mu) = \frac{1}{\pi}\frac{1}{1+(x-\mu)^2}$$

  - $X_1,\dots,X_n$ **が** $f(x;\mu,1)$ **からの独立な標本とするときの** μ **の最尤推定**

$$\text{尤度}\quad L(\mu) = \prod_{i=1}^{n} f(x_i;\mu,1) = \prod_{i=1}^{n}\frac{1}{\pi}\left(1+(x_i-\mu)^2\right)^{-1}$$

解析的な最大化は不可能だが数値的最大化が可能

統計的検定とは I

- 統計的モデルを構成する分布族 $\{f(x,\theta)|\theta \in \Theta\}$
- 母数空間の中の意味を持つ部分集合 Θ_0
  - 回帰分析 $\beta_i = 0$
  - コイン投げ $p = 1/2$
  - 新薬の効果が既存薬と変わらない（臨床試験）

統計的検定とは II

- 帰無仮説：パラメータの真値が Θ_0 に含まれるという想定

$$\mathrm{H}_0 : \theta \in \Theta_0$$

  - Θ_0 が一点集合の場合　単純帰無仮説
  - 二点以上含む集合の場合　複合帰無仮説
- コイン投げ（ベルヌーイ試行）
  - 分布族 $\{\mathrm{Bin}(n, p) | 0 \leq p \leq 1\}$
  - 帰無仮説 $\mathrm{H}_0 : p = 1/2$，$p = 1/2$ に特別な意味

統計的検定とは III

- **回帰分析**
  - $\beta_i = 0$ **に特別な意味**
  - i **番目の説明変数** x_i **が** y **を説明するのに役に立たない**

$$\textbf{帰無仮説} \qquad \mathrm{H}_0 : \beta_i = 0$$

- **推測統計学の回帰分析の設定**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_p x_p + \epsilon$$

注意：β_j **は** x_j **が一単位増加** $x_j^* \to x_j^* + 1$ **するときの** y **の増分**

統計的検定とは IV

- 統計的検定：観測値 $X = (X_1, \ldots, X_n)$ を得たときに，$f(x, \theta)$, $\theta \in \Theta_0$ の分布から発生したと考えて良いか，つまり帰無仮説が正しいかどうかを判断すること
- 棄却 & 受容
  - 棄却：正しくないと判断すること
  - 受容：正しいと判断すること

統計的検定とは V

- 背理法：ある仮説をたてて議論を進めた結果が所与の条件と矛盾する場合に，その仮説を否定する
- 背理法の確率化：観測された事実が帰無仮説のもとでは起こりそうもないこと，つまり確率が非常に小さい事象であった場合に「矛盾」と考えて，帰無仮説を棄却

統計的検定とは VI

- **コイン投げの例** $\mathrm{H}_0 : p = 1/2$
  - $n = 10,\ X = 10 \quad \Pr(X = 10) = 1/1024$
    $\Uparrow$ **十分小さいと考えられるので，仮説を棄却**
  - $n = 3,\ X = 3 \quad \Pr(X = 3) = 1/8$
    $\Uparrow$ **小さいが，仮説を棄却するほど小さいか？「証拠不十分」で仮説を受容**
- **しきい値が問題となる**

統計的検定とは VII

- 「確率が非常に小さい」というときに，どれくらい小さい値をいうのか? 特定の臨界値 α を決めて α 以下の確率を「非常に小さい」と考える
- 有意水準：この α の値（習慣として 5% や 1%）
- 実際に起こった事象の帰無仮説のもとでの確率を計算して，それが α 以下であるとき帰無仮説を棄却

統計的検定とは VIII

- 対立仮説：帰無仮説が成り立たない場合の想定

$$\mathrm{H}_1 : \theta \in \Theta_1 \quad (\text{ここで}\Theta_0 \cap \Theta_1 = \emptyset)$$

  多くの場合，Θ_1 は Θ_0 の補集合
- Θ を二つの部分集合に分割したとき，どちらを Θ_0，どちらを Θ_1 にするかについての数学的基準はない
  - 通常成り立つと考える仮説を帰無仮説．品質管理における不良率
  - データによって反証し，棄却する目的で設定する．新薬の効果の検証

検定のより正確な定式化 I

- 実は，以上の説明で「実際に起こったこと」という部分がまずい
  - 分散 1 の正規分布 $N(\mu, 1)$ から 1 個の観測値を得て，帰無仮説 $\mathrm{H}_0 : \mu = 0$ を検定するとする
  - 確率変数 X の実現値 x がどの値であっても「連続分布」においては特定の値の実現する確率は 0
- 「実際に起こったこと」の全体の中での相対的な起こりにくさを考える

検定のより正確な定式化 II

- 全事象 $=$「実際に起こったこと，あるいはそれ以上に起こりにくいこと」$\cup$「実際に起こったことより起こりやすいこと」

$$\mathbb{R} = \{w : \phi(w) \leq \phi(x)\} \cup \{w : \phi(w) > \phi(x)\}$$

  - ϕ は $N(0,1)$ の確率密度関数
  - $\phi(x)$ は帰無仮説 $\mu = 0$ のもとでの x の起こりやすさ

検定のより正確な定式化 III

- 背理法の確率化の修正版
  - 帰無仮説のもとでの「実際に起こったこと，あるいはそれ以上に起こりにくいこと」が起きる確率
  - この確率が与えられた有意水準以下であれば，帰無仮説を棄却
- 「起きにくさ」の数学的な定式化は後回し
  $\Uparrow$ 当面，起きにくさを測る関数 $T(X_1, \ldots X_n)$ が与えられているとして議論

検定のより正確な定式化 IV

- 検定統計量：検定のための統計量

$$T(X) = T(X_1, \ldots, X_n)$$

- 検定統計量 T に対し，観測値 $x = (x_1, \ldots, x_n)$ を得たときに「実際に起こったことあるいはそれ以上に起こりにくいこと」を表す部分集合

$$R_{T(x)} = \{w | T(w) \geq T(x)\}$$

検定のより正確な定式化 V

- P 値：$R_{T(x)}$ の帰無仮説のもとでの確率
  - 単純帰無仮説 $H_0 : \theta = \theta_0$ の場合

$$\mathrm{Pr}_{\theta_0}(W \in R_{T(x)}) \quad \mathrm{H}_0\text{のもとでの確率} \tag{1}$$

  - 複合帰無仮説 $H_0 : \theta \in \Theta_0$ の場合

$$\sup_{\theta \in \Theta_0} \mathrm{Pr}_{\theta}(W \in R_{T(x)}) \tag{2}$$

$\mathrm{Pr}_{\theta}(W \in R_{T(x)})$ は固定したθのもとでの確率

- P 値が（予め定めた）有意水準 α 以下のときに棄却

起こりにくさを測る T の具体例 I

例１：二項分布 $X \sim \mathrm{Bin}(n, p)$

- 帰無仮説 $H_0 : p = 1/2$
- X の確率関数 $\binom{n}{x} p^x (1-p)^{n-x}$

$\Uparrow$ $H_0 : p = 1/2$ のもとで，

$$P\left(X = x \middle| p = \frac{1}{2}\right) = \binom{n}{x} (1/2)^x (1/2)^{n-x}$$
$$= \binom{n}{x} (1/2)^n$$

起こりにくさを測る T の具体例 II

- **実際に起こった x と同じかそれ以上に起こりにくい事象**

$$
\begin{aligned}
&\left\{y : P\left(X = y \,|\, p = \frac{1}{2}\right) \le P\left(X = x \,|\, p = \frac{1}{2}\right)\right\} \\
&= \left\{y : \binom{n}{y} \le \binom{n}{x}\right\} \\
&= \{y : y!(n-y)! \ge x!(n-x)!\} \\
&= \{y : |y - n/2| \ge |x - n/2|\}
\end{aligned}
$$

起こりにくさを測る T の具体例 III

- **例**
  - $n = 10$，$x = 10$ **に対し，** $y = 0, 10$
  - $n = 10$，$x = 9$ **に対し，** $y = 0, 1, 9, 10$
- **再確認**
  - **検定統計量** $|x - n/2|$，**帰無仮説** $p = 1/2$ **のもとでは大きくなりにくい**
  - $\{y : |y - n/2| \geq |x - n/2|\}$：**実際に起こった** x **と同じかそれ以上に起こりにくい事象**

起こりにくさを測る T の具体例 IV

- P 値
  - $n = 10$, $x = 0$,

$$\sum_{i=0,10} \binom{10}{i} (1/2)^{10} = \frac{1+1}{1024} = \frac{2}{1024}$$

  - $n = 10$, $x = 9$,

$$\sum_{i=0,1,9,10} \binom{10}{i} (1/2)^{10} = \frac{1+10+10+1}{1024} = \frac{22}{1024}$$

  - $n = 10$, $x = 2$,

$$\sum_{i=0,1,2,8,9,10} \binom{10}{i} (1/2)^{10} = \frac{1+10+45+45+10+1}{1024} = \frac{112}{1024}$$

起こりにくさを測る T の具体例 V

例２：正規分布

- 正規分布 $N(\mu, \sigma^2)$ から n 個の i.i.d. 標本 $X_1, \ldots, X_n$ を得て $\mu = 0$ を検定
- σ^2 は既知と仮定
- 正規分布の場合には，n 個の i.i.d. 標本 $X_1, \ldots, X_n$ に対し，標本平均 $\bar{X}$ が μ_0 の推論に関する全ての情報を持つ
  $\Uparrow$ $\bar{X} \sim N(\mu, \sigma^2/n)$ から「１個の標本」（確率変数の実現値）$\bar{x}$ を観測して μ について推論すると考えて良い

起こりにくさを測る T の具体例 VI

- $W \sim N(\mu, \sigma^2/n)$

$$f(w;\mu,\sigma^2/n) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2/n}\right) = \frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{w-\mu}{\sigma/\sqrt{n}}\right)$$

- $\phi(z) = \dfrac{1}{\sqrt{2\pi}} \exp(-z^2/2)$ $N(0,1)$ **の密度関数**
- $\bar{x}$ **は実際に起こった手元にある実現値．混乱を防ぐため** H_0 **のもとでの確率を計算するときには** X **ではなく** W **を使う**

起こりにくさを測る T の具体例 VII

- $\mu = 0$ のもとで $\bar{x}$ の起こりやすさ

$$\frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{\bar{x}}{\sigma/\sqrt{n}}\right)$$

$\Uparrow$ 正規分布は連続確率分布．一点の確率はゼロ

- $\mu = 0$ のもとで $\bar{x}$ が起こるかあるいはそれ以上に起こりにくい事象

$$\begin{aligned}&\left\{w : \frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{w}{\sigma/\sqrt{n}}\right) \leq \frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{\bar{x}}{\sigma/\sqrt{n}}\right)\right\}\\ &= \left\{w : \exp\left(-\frac{n}{\sigma^2}\frac{w^2}{2}\right) \leq \exp\left(-\frac{n}{\sigma^2}\frac{\bar{x}^2}{2}\right)\right\}\\ &= \left\{w : w^2 \geq \bar{x}^2\right\} = \{w : |w| \geq |\bar{x}|\}\end{aligned}$$

起こりにくさを測る T の具体例 VIII

- $\{w : |w| \geq |\bar{x}|\}$：$\mu = 0$ のもとで $\bar{x}$ が起こるかあるいはそれ以上に起こりにくい事象
- 検定統計量 $T(X_1, \ldots, X_n) = |\bar{X}|$
- P 値 $\Pr(|W| \geq |\bar{x}|) \Leftarrow W \sim N(0, \sigma^2/n)$
- $\Pr(|W| \geq |\bar{x}|) = 2\int_{\sqrt{n}|\bar{x}|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z$ math.pdf 3.6 節
  - python: scipy.stats.norm.sf()
  - R: pnorm(, lower.tail = F)

  $\Uparrow$ 正規分布の裾確率を計算する関数

起こりにくさを測る T の具体例 IX

- ところで例２において，P 値

$$2\int_{\sqrt{n}|\bar{x}|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\mathrm{d}z$$

を（数値積分で）計算しなくても，あらかじめ

$$\int_{z_{\alpha/2}}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}}\mathrm{d}z = \frac{\alpha}{2}$$

となるような $z_{\alpha/2}$（上側 $\alpha/2 \times 100\%$ 点）を知っていれば，$\sqrt{n}|\bar{x}|/\sigma$ と $z_{\alpha/2}$ の大小を比較するだけで有意水準 α の検定が実行可能

起こりにくさを測る T の具体例 X

- $H_0 : \mu = 0$ を棄却 $\Leftrightarrow$

$$\frac{\sqrt{n}|\bar{x}|}{\sigma} \geq z_{\alpha/2} \Leftrightarrow 2\int_{\sqrt{n}|\bar{x}|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z \leq \alpha$$

- 例えば $z_{0.05/2} \simeq 1.96 \quad z_{0.01/2} \simeq 2.58$ が知られている

$$\frac{\sqrt{n}|\bar{x}|}{\sigma} \geq \begin{cases} 1.96 \\ 2.58 \end{cases} \Leftrightarrow 2\int_{\sqrt{n}|\bar{x}|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z \leq \begin{cases} 0.05 \\ 0.01 \end{cases}$$

数値積分をしなくても，$\sqrt{n}|\bar{x}|/\sigma$ の大きさから受容棄却の判断だけは出来る $\Leftarrow$ かつては非常に重要

伝統的な検定の手続き（一般）

- 所与の検定統計量 T が実数 c に対して，

$$T(X) \geq c \ \Rightarrow H_0\text{を棄却}$$

  - 棄却点：c の値
  - 棄却域：標本空間の部分集合 $R_c = \{x|T(x) \geq c\}$

- 検定統計量を与えられた棄却点と比較するだけでは，検定統計量がどの程度帰無仮説と乖離しているかを数量的に判断することが困難
- P 値は「実際の観測値が帰無仮説のもとでどれくらい起きにくかったか」を表している分，情報が多い

若干の一般化 I

正規分布 $H_0 : \mu = 0$ の検定 $\Rightarrow H_0 : \mu = \mu_0$ の検定

- **正規分布 $N(\mu, \sigma^2)$ から n 個の** i.i.d. **標本 $X_1, \ldots, X_n$ を得て $H_0 : \mu = \mu_0$ を検定**
- **σ^2 は既知と仮定**
- **再掲**
  - **正規分布の場合には，n 個の** i.i.d. **標本 $X_1, \ldots, X_n$ に対し，標本平均 $\bar{X}$ が μ の推論について全ての情報を持つ**
  - **つまり，$\bar{X} \sim N(\mu, \sigma^2/n)$ から「１個の標本」（確率変数の実現値）$\bar{x}$ を観測して μ について推論すると考えて良い**

後は $\mu = 0$ の場合の単なる平行移動

若干の一般化 II

- $\mu = \mu_0$ のもとで $\bar{x}$ の起こりやすさ

$$\frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)$$

$\Uparrow$ 正規分布は連続確率分布．一点の確率はゼロ

- $\mu = \mu_0$ のもとで $\bar{x}$ が起こるかあるいはそれ以上に起こりにくい事象

$$\left\{w : \frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{w-\mu_0}{\sigma/\sqrt{n}}\right) \leq \frac{1}{\sigma/\sqrt{n}}\phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)\right\}$$
$$= \{w : |w-\mu_0| \geq |\bar{x}-\mu_0|\}$$

若干の一般化 III

- **検定統計量** $T(X_1, \ldots, X_n) = |\bar{X} - \mu_0|$
- P **値** $\Downarrow W \sim N(\mu_0, \sigma^2/n)$

$$\begin{aligned}
&\Pr(|W - \mu_0| \geq |\bar{x} - \mu_0|) \\
&= 2\int_{\sqrt{n}|\bar{x}-\mu_0|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z
\end{aligned}$$

若干の一般化 IV

- $H_0 : \mu = 0$ を棄却 $\Leftrightarrow$

$$\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma} \geq z_{\alpha/2} \Leftrightarrow 2\int_{\sqrt{n}|\bar{x}-\mu_0|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z \leq \alpha$$

- 例えば $z_{0.05/2} \simeq 1.96 \quad z_{0.01/2} \simeq 2.58$ が知られている

$$\frac{\sqrt{n}|\bar{x} - \mu_0|}{\sigma} \geq \begin{cases} 1.96 \\ 2.58 \end{cases} \Leftrightarrow 2\int_{\sqrt{n}|\bar{x}-\mu_0|/\sigma}^{\infty} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \mathrm{d}z \leq \begin{cases} 0.05 \\ 0.01 \end{cases}$$

より現実的に（σ^2 も未知） I

正規分布 $H_0 : \mu = \mu_0$ の検定

- **正規分布 $N(\mu, \sigma^2)$ から n 個の** i.i.d. **標本 $X_1, \ldots, X_n$ を得て $\mu = \mu_0$ を検定　　σ^2 は未知**
- **正規分布の場合には，n 個の** i.i.d. **標本 $X_1, \ldots, X_n$ に対し，標本平均 $\bar{X}$ と（不偏）標本分散**

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

が μ の推論について全ての情報を持つ

より現実的に（σ^2 も未知） II

数学的な事実

- $\bar{X} \sim N(\mu, \sigma^2/n) \Leftrightarrow \bar{X} - \mu \sim (0, \sigma^2/n)$
  $\Leftrightarrow \frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim N(0,1)$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$
- $\bar{X}$ と $\sum_{i=1}^{n}(X_i - \bar{X})^2$ は独立
- σ^2 によらず，$\frac{\sqrt{n}}{\sqrt{S^2}}\left(\bar{X} - \mu\right) \sim t_{n-1}$

  証明：任意の σ^2 に対して

$$\frac{\sqrt{n}}{\sqrt{S^2}}\left(\bar{X} - \mu\right) = \frac{\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}$$

より現実的に（σ^2 も未知） III

- $H_0 : \mu = \mu_0$ のもとで，$\dfrac{\sqrt{n}}{\sqrt{S^2}}\left(\bar{X}-\mu_0\right)$ を指標として，実際に観測した $\bar{x}$ と s^2 の起こりやすさを考える．
- P 値（$H_0 : \mu = \mu_0$ のもとで，実際に観測した $\bar{x}$ と s^2 が起こるかそれ以上に起こりにくい事象が起こる確率）

$$\Pr\left(|T_{n-1}| \geq \frac{\sqrt{n}}{\sqrt{s^2}}|\bar{x}-\mu_0|\right) = 2\int_{(\sqrt{n}/\sqrt{s^2})|\bar{x}-\mu_0|}^{\infty} f_{n-1}(t)\mathrm{d}t$$

$$f_{n-1}(t) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma((n-1)/2)}\left(1+\frac{t^2}{n-1}\right)^{-n/2}$$

$\Uparrow$ $f_{n-1}(t)$ は自由度 $n-1$ の t 分布の密度関数

より現実的に（σ^2 も未知） IV

- P 値が予め与えられた有意水準より小さい場合に $H_0 : \mu = \mu_0$ を棄却
- 予め $t_{\alpha/2,n-1}$，上側 $\alpha/2 \times 100\%$ 点，つまり

$$\int_{t_{\alpha/2,n-1}}^{\infty} f_{n-1}(t)\mathrm{d}t = \frac{\alpha}{2}$$

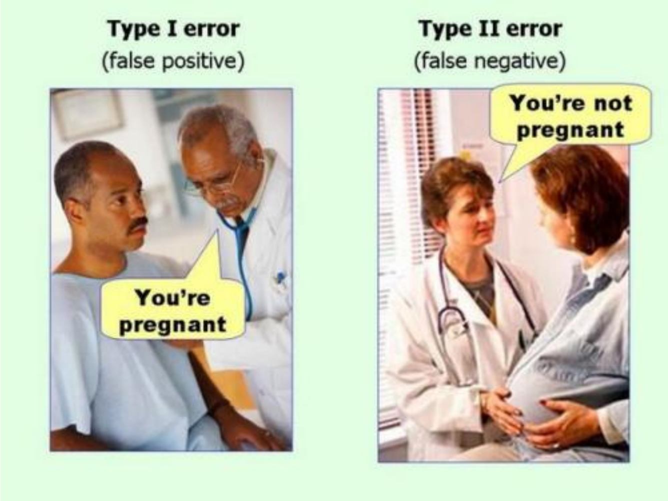
を知っていれば，$(\sqrt{n}/\sqrt{s^2})|\bar{x} - \mu_0|$ と $t_{\alpha/2,n-1}$ の比較で受容，棄却の判断が出来る

$$\frac{\sqrt{n}}{\sqrt{s^2}}|\bar{x} - \mu_0| \geq t_{\alpha/2,n-1} \iff 2\int_{(\sqrt{n}/\sqrt{s^2})|\bar{x}-\mu_0|}^{\infty} f_{n-1}(t)\mathrm{d}t \leq \alpha$$

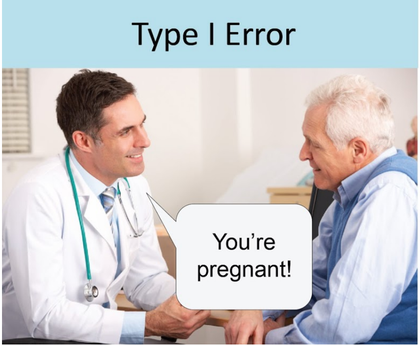
二種類の過誤，理論としての検定論の目的 I

- 以上では適当な検定統計量 T が与えられたとして議論したが，最適な検定統計量とは？
- 検定における二種類の誤り
  - 第一種の過誤：帰無仮説が正しいのに，棄却
  - 第二種の過誤：対立仮説が正しいのに，受容
- 第一種の過誤と第二種の過誤の確率をどちらも小さくしたいが，「常に棄却」「常に受容」するような検定を考えれば，トレードオフの関係にある

二種類の過誤，理論としての検定論の目的 II



二種類の過誤，理論としての検定論の目的 III



二種類の過誤，理論としての検定論の目的 IV

- 最適性：第一種の過誤の確率を α 以下に抑えた上で，第二種の過誤の確率をなるべく小さくする
- 複数の起こりにくさを測る関数

$$
\begin{aligned}
T_1(x_1,\ldots,x_n) &= |\bar{x}-\mu_0| \\
T_2(x_1,\ldots,x_n) &= \left|\frac{x_1+\cdots+x_{n-1}}{n-1}-\mu_0\right| \\
T_3(x_1,\ldots,x_n) &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}} \\
T_4(x_1,\ldots,x_n) &= |\mathrm{median}(x_1,\ldots,x_n)-\mu_0|
\end{aligned}
$$

$\Uparrow$ 推定量も「複数」存在していたことに注意

二種類の過誤，理論としての検定論の目的 V

- P 値

$$\Pr\left(T_i(W_1,\ldots,W_n) \geq T_i(x_1,\ldots,x_n)\right),\ i=1,2,\ldots,$$

$\Uparrow$ $W_1,\ldots,W_n \sim N(\mu_0,1)$ **つまり帰無仮説のもとでの確率**

- **それぞれ** P **値** $\leq \alpha$ **というルールで棄却**
- T_1,T_2,T_3,T_4 **の中では，**T_1 **が良い（第二種の過誤の確率が最小）**

信頼区間の考え方 I

- 点推定，区間推定，検定
  - 点推定はパラメータの値を一点で言い当てようとするもの
  - 区間推定はパラメータの真値を（高い確率で）含むであろう区間を与えるもの
  - 区間推定は論理的には検定の裏返し
- $X_1, \ldots, X_n$ を正規分布 $N(\mu, \sigma^2)$ からの i.i.d. 標本
- 帰無仮説 $H_0 : \mu = \mu_0$，σ^2 は既知
- $z_{\alpha/2}$：標準正規分布の上側 $(\alpha/2) \times 100\%$点

信頼区間の考え方 II

- $H_0 : \mu = \mu_0$ **を受容** if

$$\frac{\sqrt{n}}{\sigma}|\bar{x} - \mu_0| < z_{\alpha/2} \tag{3}$$

- $H_0 : \mu = \mu_0$ **が正しいとき，**(3) **は確率** $1 - \alpha$ **で成立**

$$\Pr_{H_0}\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| < z_{\alpha/2}\right) = 1 - \alpha$$

$$\Uparrow \Pr_{H_0}\left(\frac{\sqrt{n}}{\sigma}|\bar{X} - \mu_0| \geq z_{\alpha/2}\right) = \alpha$$ **の裏返し**

信頼区間の考え方 III

- μ_0 と観測値 $x_1, \ldots, x_n$ が矛盾しない μ_0 の集合
  - μ_0 を予め定められた特別な値と考えず，μ_0 としていろいろな値を同時に考慮
  - 観測値 $x_1, \ldots, x_n$ は既得，$\bar{x}$ は計算済み
  - 帰無仮説 $H_0 : \mu = \mu_0$ が受容される
    $\mu = \mu_0$ のもとでの $x_1, \ldots, x_n$ が「普通の値」
    $\Leftrightarrow$ μ_0 と x が矛盾しないとき
  - 得られた $x_1, \ldots, x_n$ に照らして，どの μ_0 が x と矛盾しないかを考えると，そのような μ_0 の値の集合を構成可能

信頼区間の考え方 IV

- **数学的には，$x_1, \ldots, x_n$ を所与（つまり $\bar{x}$ を所与）として，式** (3) **を μ_0 について解くだけ**
- **式** (3) **の同値変形**

$$\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu_0 < \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

or equivalently $\mu_0 \in I_{x_1,\ldots,x_n}$ where

$$I_{x_1,\ldots,x_n} = \left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

信頼区間の考え方 V

- 「信頼係数 $1-\alpha$ の信頼区間」：任意の μ_0 について

$$\Pr_{H_0}\left(\frac{\sqrt{n}}{\sigma}|\bar{X}-\mu_0| < z_{\alpha/2}\right) = 1-\alpha$$

$$\Pr_{H_0}\left(\mu_0 \in I_{X_1,\ldots,X_n}\right) = 1-\alpha$$

信頼区間の考え方 VI

- **解釈上の注意**
  - **確率的に変動するのは** $I_{X_1,\ldots,X_n}$
  - μ_0 **が真の平均であるときに，** $X_{11},\ldots,X_{1n}$, $X_{21},\ldots,X_{2n}$, $\cdots$, $X_{N1},\ldots,X_{Nn}$ **のように繰り返し** n **組の** i.i.d. **標本を発生**
  - **それぞれ区間** $I_{X_{11},\ldots,X_{1n}}$, $I_{X_{21},\ldots,X_{2n}}$ $,\ldots,$ $I_{X_{N1},\ldots,X_{Nn}}$ **を構成**
  - **このとき，** $i=1,\ldots,N$ **の中で** $\mu_0 \in I_{X_{i1},\ldots,X_{in}}$ **となる相対頻度が** $N \to \infty$ **のとき** $1-\alpha$ **に収束**

二項分布の信頼区間 I

- 試行回数 n 回，成功回数 x が既知
- 成功確率の信頼区間（区間推定）
- 帰無仮説 $H_0 : p = p_0$
- 検定ではある固定した p_0 を考えるが，区間推定では p_0 を動かす
- 有意水準 α の H_0 の検定で H_0 が受容されるような p_0 の集合が信頼区間となる

二項分布の信頼区間 II

- $p = p_0$ のもとで成功回数 x と同じかより起こりにくい成功回数の集合

$$
\begin{aligned}
&Q(p_0, n, x) \\
&= \{y : P\left(X = y \middle| p = p_0\right) \leq P\left(X = x \middle| p = p_0\right)\}
\end{aligned}
$$

  - $Q(p_0, n, x)$ は $\{0, 1, \ldots, n\}$ の部分集合
  - $P\left(X = x \middle| p = p_0\right) = \binom{n}{x} (p_0)^x (1 - p_0)^{n-x}$

二項分布の信頼区間 III

- P 値 $\displaystyle\sum_{y\in Q(p_0,n,x)} P\left(X = y|p = p_0\right)$
- P 値 $> \alpha$ ならば p_0 を信頼区間に含める

$$R(n, x, \alpha) = \left\{ p_0 : \sum_{y\in Q(p_0,n,x)} P\left(X = y|p = p_0\right) > \alpha \right\}$$

「区間」と既に呼んでいるが，結果として $R(n, x)$ は区間になる

$\Uparrow$ $R(n, x, \alpha)$ **：信頼係数 $1 - \alpha$ の信頼区間**

単回帰分析 I

- 正規線形回帰モデル（α, β, σ^2 は全て未知）

$$Y = \alpha + \beta x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
$$Y \sim N(\alpha + \beta x, \sigma^2)$$

- n 個の独立な標本 $(x_1, y_1), \ldots, (x_n, y_n)$ を得た下で $H_0 : \beta = \beta_0$ の検定
- ??ページの設定

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

から独立な $X_1, \ldots, X_n$ を得て，$H_0 : \mu = \mu_0$ を検定する問題の一般化

単回帰分析 II

- 正規線形回帰モデルの場合には，

最小二乗推定量
$$\begin{cases} \check{B} &= \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \check{A} &= \bar{Y} - \check{B}\bar{x} \end{cases}$$

残差平方和
$$\begin{aligned} \mathrm{RSS} &= \sum_i (Y_i - \check{A} - \check{B}x_i)^2 \\ &= \sum_i \{(Y_i - \bar{Y}) - \check{B}(x_i - \bar{x})\}^2 \end{aligned}$$

が α, β, σ^2 の推論について全ての情報を持つ

数学的な事実 I

- $\check{A} \sim N(\alpha, \sigma^2\{1/n + \bar{x}^2/\sum_i (x_i - \bar{x})^2\})$
- $\check{B} \sim N(\beta, \sigma^2/\sum_i (x_i - \bar{x})^2)$

$$\Leftrightarrow \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sigma}(\check{B} - \beta) \sim N(0, 1)$$

- $\dfrac{\mathsf{RSS}}{\sigma^2} \sim \chi^2_{n-2}$, $\left(S^2 = \dfrac{\mathsf{RSS}}{n-2} \Rightarrow \mathrm{E}\left[S^2\right] = \sigma^2\right)$

数学的な事実 II

- $(\check{A}, \check{B})$ と S^2 は独立
- $\dfrac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{S^2}} \left(\check{B} - \beta\right) \sim t_{n-2}$

$$\frac{\dfrac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{\sigma^2}} \left(\check{B} - \beta\right)}{\sqrt{\dfrac{S^2}{\sigma^2}}} = \frac{N(0,1)}{\sqrt{\dfrac{\chi^2_{n-2}}{n-2}}} \sim t_{n-2}$$

検定 I

- $H_0 : \beta = 0$
- H_0 が正しくても，最小二乗推定量 $\check{b}$（小文字にしたのは最小二乗推定量の実現値であるため）はぴったりゼロにならない
- 一方，$\beta \neq 0$ ならば，$|\check{b}|$ はゼロから離れて大きくなる傾向があるだろう．
- どのくらい大きくなれば，$\beta \neq 0$ と判断してよいのか？

検定 II

- $H_0 : \beta = 0$ のもとで，$\dfrac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{s^2}} \left(\check{b} - 0\right)$ を指標として，実際に観測したデータ $(x_1, x_1), \ldots, (x_n, y_n)$ の起こりやすさを考える．
- 回帰分析の t–value

$$t = \frac{\check{b}}{\sqrt{\dfrac{s^2}{\sum_i (x_i - \bar{x})^2}}} = \frac{\text{最小二乗推定量}}{\text{標準誤差}}$$

$\Uparrow$ x の単位のとり方に依存しない

検定 III

- **単位の変換** $x \to cx$
  - **分子** $\check{b} \to \dfrac{1}{c}\check{b}$

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

  - **分母** $\sqrt{\dfrac{s^2}{\sum_i (x_i - \bar{x})^2}} \to \dfrac{1}{c}\sqrt{\dfrac{s^2}{\sum_i (x_i - \bar{x})^2}}$

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{\sum_i \{(y_i - \bar{y}) - \check{b}(x_i - \bar{x})\}^2}{n-2}$$

検定 IV

- P 値（$H_0 : \beta = 0$ のもとで，実際に観測した t が起こるかそれ以上に起こりにくい事象が起こる確率）

$$\Pr(|T_{n-2}| \geq |t|) = 2\int_{|t|}^{\infty} f_{n-2}(t)\mathrm{d}t$$

$f_{n-2}(t)$ は自由度 $n-2$ の t 分布の密度関数

- P 値が予め与えられた有意水準より小さい場合に $H_0 : \beta = 0$ を棄却