統計学メモ

@mathnuko

2020/02/12 〜

目 次

第 I 部　統計学 **7**

第 1 章　統計学の目的 **8**
1.1　統計学とは何のための学問か？ 8
1.1.1　「データ」とは？ 8
1.1.2　統計学とは？ 8
1.2　統計学の体系：記述統計学と推測統計学 9
1.3　記述統計学，推測統計学の基本的な流れ 10
1.3.1　データ分析における問題設定 10
1.3.2　データ分析の流れ 11
1.4　補足 データの解釈：「定性的な解釈」と「定量的な解釈」 11

第 2 章　記述統計学 **12**

第 3 章　推測統計学 **13**

第 II 部　数理統計学 **14**

第 4 章　正規分布の導出 **15**

第 III 部　生物統計学 **16**

第 5 章　生物統計学：基本 **17**
5.1　理解度チェックリスト 17
5.2　標準偏差の種類 18
5.3　標本平均の分布 18
5.3.1　補足：期待値，分散の性質 19
5.4　標本平均の分布による区間推定 20
5.5　正規分布のグラフの概形 21
5.6　パラメトリック検定，ノンパラメトリック検定 21
5.7　2 群間比較の統計手法 21
5.8　3 群感比較の統計手法 21

第 6 章　生物統計学：応用 **22**
6.1　多変量解析 22
6.2　クラスタリング 22
6.3　主成分分析 22

第 7 章　オミクス解析 **23**
7.1　オミクス解析の基本的な流れ 23
7.2　トランスクリプトーム 23

7.2.1 Enrichment 解析 . . . 23
7.2.2 Enrichment 解析の注意点 . . . 23
7.2.3 メタボローム . . . 23

第 IV 部 統計学インデックス 24

第 8 章 統計学とは 26
8.1 統計学の分類 . . . 26

第 9 章 記述統計学 27
9.1 代表値 . . . 27
9.1.1 平均値 . . . 27
9.1.2 中央値 . . . 27
9.1.3 最頻値 . . . 27
9.2 データのばらつき . . . 27
9.2.1 分散，標準偏差 . . . 27
9.2.2 分位数 . . . 27
9.2.3 変動係数 . . . 27
9.3 変量の関連性 . . . 27
9.3.1 相関係数 . . . 27
9.3.2 順位相関係数 . . . 27

第 10 章 確率分布 28
10.1 確率変数，確率分布 . . . 28

第 11 章 誤差論 29
11.1 有効数字 . . . 29
11.2 系統誤差と偶然誤差 . . . 29
11.3 誤差の伝搬 . . . 29

第 12 章 推定 30
12.1 何を「推定」するのか：母数 . . . 30
12.2 統計量 . . . 30
12.3 統計的推測における「正規分布を仮定する」という前提 . . . 30
12.4 点推定 . . . 30
12.4.1 推定量 . . . 30
12.4.2 最尤推定 . . . 30
12.4.3 モーメントを用いた推定 . . . 30
12.4.4 推定量の満たすべき性質 . . . 30
12.5 区間推定 . . . 30
12.5.1 標本分布 . . . 30
12.5.2 母平均の信頼区間 . . . 30
12.5.3 母比率の信頼区間 . . . 30
12.5.4 母分散の信頼区間 . . . 30
12.5.5 母相関係数の信頼区間 . . . 30
12.5.6 母相関係数の信頼区間 . . . 30
12.5.7 シミュレーションで母数を推定する：ブートストラップ法 . . . 30

第 13 章 仮説検定 **32**
13.1 仮説検定とは 32
13.2 仮説検定の分類：パラメトリック検定とノンパラメトリック検定 32
13.3 帰無仮説と対立仮説 32
13.4 仮説検定の手順 32
13.5 仮説検定におけるリスク：第一種の過誤，第二種の過誤 33
13.5.1 第一種の過誤，第二種の過誤を図表で理解する 33
13.5.2 「第一種の過誤＝有意水準」の意味するところ 35
13.6 母比率の検定 35
13.7 母分散の検定 35
13.8 相関の有無の判定：無相関の検定 35
13.9 平均値の差の検定：対応のない 2 群の検定 35
13.10平均値の差の検定：対応のある 2 群の検定 35
13.11比率の差の検定：対応のない 2 群の場合 35
13.12非劣性試験 35

第 14 章 分散分析 **36**
14.1 一元配置分散分析：実験で効果を確かめる 36
14.2 多群の等分散の検定：Bartlett 検定 36
14.3 対応のある一元配置分散分析：個体差を考慮する 36
14.4 二元配置分散分析：交互作用を見つけ出す 36

第 15 章 検定の多重性 **37**
15.1 検定の多重性とは 37
15.2 多重検定補正 37
15.2.1 Bonferrorni 法 37
15.2.2 Benjamini-Hochberg 法 37
15.3 Storey 法 37
15.4 Tukey 法，Tukey-Kramer 法 37
15.5 Dunnett 法 37

第 16 章 ノンパラメトリック検定 **38**
16.1 ノンパラメトリック検定：分布によらない仮説検定 38
16.2 独立性の検定（ピアソンの χ^2 検定）：質的データの検定 38
16.3 フィッシャーの正確確率検定：2×2 分割表の検定 38
16.4 対応のない 2 群の順序データの検定：マンホイットニーの U 検定 38
16.5 対応のある 2 群の順序データの検定：符号検定 38
16.6 対応のある 2 群の量的データの検定：ウィルコクソンの符号付き順位和検定 38
16.7 対応のない多群の順序データの検定：クラスカル・ウォリス検定 38
16.8 対応のある多群の順序データの検定：フリードマン検定 38

第 17 章 実験計画法 **39**
17.1 フィッシャーの 3 原則：1 反復 39
17.2 フィッシャーの 3 原則：2 無作為化 39
17.3 フィッシャーの 3 原則：3 局所管理 39
17.4 色々な実験配置 39
17.5 直交計画法：実験を間引いて実施する 39
17.6 直交計画法の応用 1：品質工学（パラメータ設計） 39

17.7 直交計画法の応用 2：コンジョイント分析 39
17.8 検出力分析：標本サイズの決め方 39
17.8.1 検出力とは 39
17.8.2 検出力分析 39
17.8.3 補足：サンプルサイズとサンプル数 39

第 18 章 回帰分析 **40**
18.1 回帰分析：原因と結果の関係を探る 40
18.2 最小二乗法 40
18.3 決定係数：回帰線の精度を評価する 40
18.4 t 検定：回帰線の傾きを検定する 40
18.5 残差分析：分析の適切さを検討する 40
18.6 重回帰分析：原因が複数あるときの回帰分析 40
18.6.1 多重共線性：説明変数間の問題 40
18.7 変数選択法：有効な説明変数を選ぶ 40
18.8 質の違いを説明する変数 1：切片ダミー 40
18.9 質の違いを説明する変数 2：傾きダミー 40
18.10プロビット分析：2 値変数の回帰分析 40
18.11イベント発生までの時間を分析する 1：生存曲線 40
18.12イベント発生までの時間を分析する 2：生存曲線の比較 40
18.13イベント発生までの時間を分析する 3：Cox 比例ハザード回帰 40

第 19 章 多変量解析 **41**
19.1 多変量とは 41
19.2 主成分分析：情報を縮約する 41
19.3 因子分析：潜在的な要因を発見する 41
19.4 多次元尺度構成法 MDS 41
19.5 クラスタリング 41
19.5.1 非階層的クラスタリング 41
19.5.2 階層的クラスタリング 41
19.5.3 重要：クラスタリングの結果の解釈 41
19.6 構造方程式モデリング（SEM）：因果構造を記述する 41
19.7 コレスポンデンス分析：質的データの関連性を分析する 41
19.8 数量化 1 類 41
19.9 数量化 2 類 41
19.10数量化 3 類 41

第 20 章 統計モデリング **42**
20.1 統計モデルとは 42
20.2 統計モデルの種類と発展 42
20.3 線形モデル 42
20.3.1 最小二乗法 42
20.4 一般化線形モデル 42
20.4.1 最尤推定法によるパラメータ推定 42
20.5 一般化線形混合モデル 42
20.6 階層ベイズモデル 42

第 21 章 統計的因果推論 **43**

第 22 章 ベイズ統計学：基本 **44**

第 V 部　日本統計学会 - 統計検定に関するまとめ **45**

第 23 章 統計検定の情報の整理 **46**
23.1 統計検定 2 級 46
23.1.1 試験要旨 46
23.2 統計検定 1 級 47
23.2.1 試験要旨 47

はじめに

この文書は統計学のポイントをまとめるためのノートである.[1)]hoge

[1)]脚注サンプル

第Ⅰ部

統計学

第1章　統計学の目的

まず,「統計学とは何のための学問か」についてまとめる．目的を理解した上で統計学の体系について説明する．

1.1　統計学とは何のための学問か？

1.1.1　「データ」とは？

統計学は「データの要約，解釈を行うための学問」であるが，そもそも「データ」とは何かを定義しておく．ざっくり言うと「データ = 事実，資料」である．

- データ … 立論・計算の基礎となる，既知のあるいは認容された事実，資料．

自然科学の分野では，しばしば「データ」,「事実」は,「現象（研究対象）を観測した際に得られる数値」，つまり「数値データ」を指すことが多い[1)]．

また,「データ」は数値データに限らず，文字，文書なども立派な「データ」である．数値データではないデータを統計学で扱う場合には，前処理として数値データへの変換などが行われた後に利用される．

1.1.2　統計学とは？

「統計学とはどのような学問か」という問いに対するいくつかの説明を挙げておく．

- 数値データの要約，解釈を行う上での理論的根拠を提供するための学問
- （赤本[2)]の説明を要約）数値データをどのように分析し，どのような判断を下したら良いかを論ずるための学問
- 対象とする集団，現象の数値データからその性質，法則性を導き出すための学問

[1)]数値データとしては個数，長さ，体積，重さ，特定の事柄が起こった回数，時刻，続いた時間の長さなどがある．
[2)]統計学入門（東京大学出版）

1.2 統計学の体系：記述統計学と推測統計学

「データに対する解釈を与える」という統計学の目的を達成するためにはどうすれば良いか？

ある現象の法則性を導くためには，まず，現象に関するデータを観測，取得する必要がある．ここでは,「ある現象に関するデータを観測・取得済み」の状態を考える．それらのデータから現象の法則性を導き出すためには，以下の 2 つの手順が考えられる．

- データから現象の法則性を道きび出すための方法
  - 手元のデータを丹念に調べ，その特徴，性質を調べ，現象の規則性，法則を見出す．
  - 手元のデータを用い,「論理性のある」推測により現象の規則性，法則を見出す．

上記手順はどちらも「統計学」が扱う範疇であるが，これらはそのまま統計学の分類に対応する．前者は「**記述統計学**」[3]，後者は「**推測統計学**」[4]と呼ばれる学問に分類される．ざっくり説明すると以下のようになる．

- 統計学
  - 記述統計学 … 手元のデータの説明
  - 推測統計学 … 手元のデータから推測[5]

記述統計学と推測統計学の関係が非常に分かりやすく描かれているのが以下の図．後に出てくる統計学の用語が挙げられているが，ここでは雰囲気を掴む．

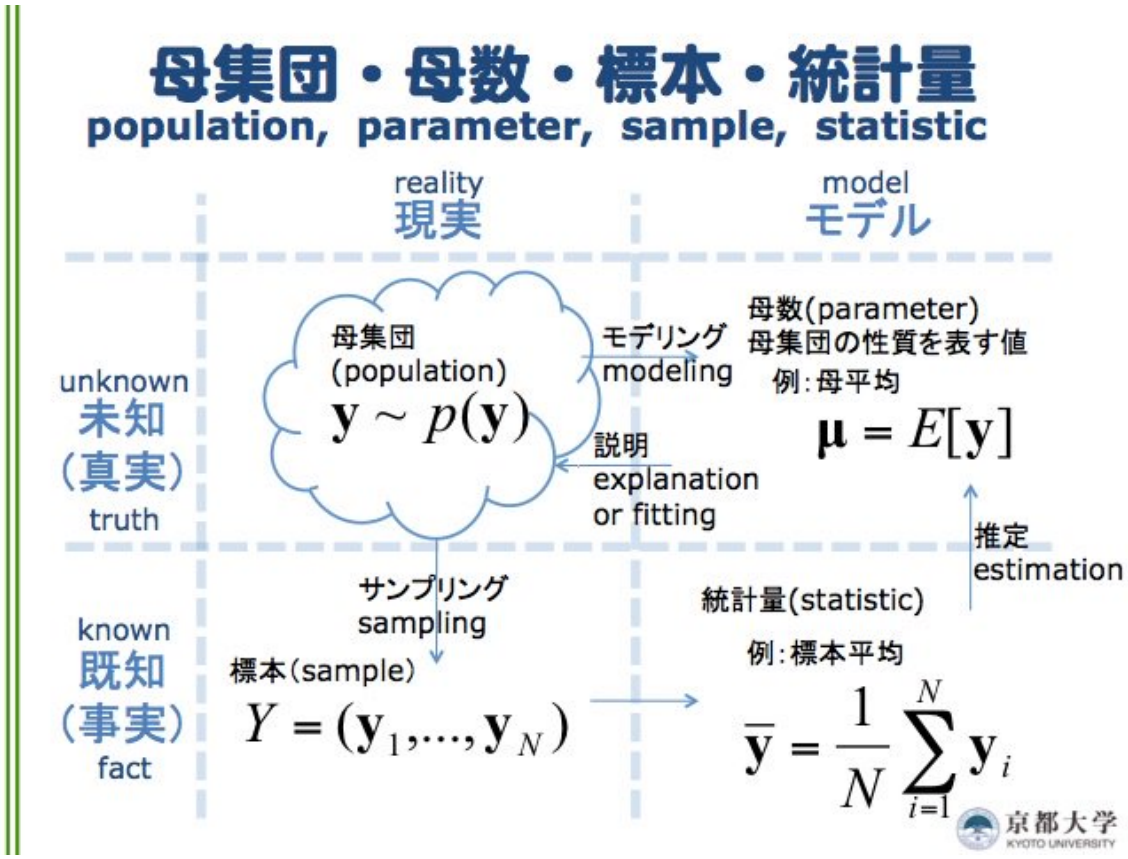

図 1.1: 統計学の体系（hoge より引用）

[3] 記述統計学，descriptive statistics
[4] 推測統計学，inferential statistics，inductive statistics
[5] 正確には「手元のデータから『母集団の性質』の推測」

1.3 記述統計学，推測統計学の基本的な流れ

記述統計学，推測統計学の手法等を用いてデータ分析行う際の基本的な流れを説明する．

1.3.1 データ分析における問題設定

まず，データ分析の際に必ず行わなければならないのが「問題設定」である．「○○という現象を理解したい」など，データ分析を行った結果何を得たいかという目的が必ず設定されているべきである．

また，**統計学はあくまでも「データに対する解釈を与える」ための手段の 1 つである**，ということは認識しておくべきである．手段が目的となってはならない．つまり，「統計学の○○という手法を使って△△という現象を理解したい」，「機械学習の△△という手法を使って何かできないか」ではなく，「△△という現象を理解したい → その現象から得たデータに適した統計学（もしくは機械学習）の○○という手法を用いる」という順番であるべきである[6)]．

まとめると，データ分析を行う際は，闇雲に統計学の手法を用いるのではなく，

1. 「どんな現象を理解したいか」という**問題設定**
2. 必要なデータの取得
3. データの解釈，データから推測（必要であれば統計学の手法を用いる）

という手順を踏むべきである．「問題設定 → データに適した統計学の手法の選択」という流れを意識したい．問題設定には例えば以下のようなものが考えられる．

- 問題設定の例
  - ある現象のデータからその傾向を見たい（記述統計学）
  - ○○という仮説を検証したい（記述統計学，推測統計学）
  - 手元のデータから未来のデータを予測したい（推測統計学）

[6)] 現実のデータ分析でこのような綺麗な流れがあるかは分からないが，少なくとも問題設定は必ず為されるべきである

1.3.2 データ分析の流れ

前節を踏まえ，記述統計学，推測統計学の手法等を用いてデータ分析行う際の基本的な流れをざっくりまとめてみる．両者の最終的なアウトプットの違いに着目すると良い．

記述統計学の流れ

1. 問題設定，必要なデータを決める
2. データの取得
3. データの要約
4. データの解釈

推測統計学の流れ

1. 問題設定，必要なデータを決める
2. データの取得
3. データから母集団の性質に当てはまる数理モデルを構築[7)]
4. 数理モデルを使った推定・推測・推論（データを生成する未知の真の確率分布）
5. 数理モデルを使った推定・推測・推論の妥当性を考察・検証

- 記述統計学のアウトプット：対象となる現象，集団の性質，特徴の説明，解釈
- 推測統計学のアウトプット：対象となる現象，集団の性質，特徴の予測．例えば,「予測を行うための数理モデル」がアウトプットになる．この場合,「数理モデルの構築 → 検証」まで行って初めて意味を成す．

データ分析においては，上記の流れは上から下に綺麗に流れていくものではなく，上記の流れを回すである．必要に応じて再度データ収集を行ったり，数理モデルの構築 → 検証のサイクルを回したり，といった

1.4 補足 データの解釈：「定性的な解釈」と「定量的な解釈」

統計学，または統計学だけでなくもっと広い意味で「データに対する解釈」を行う際に，解釈の結果として何らか（統計学で使われる指標など）の数値に落とし込むことが当たり前のように感じられるが，必ずしも数値に落とし込むことだけが「データに対する解釈」ではない[8)]．データに対する特徴・性質を「言葉」で表すこともできる．

[8)]現実に統計学が使われる場面（データ分析など）では，アウトプットとして何らかの数値（指標）に落とし込むことが求められるのがほとんどだとは思いますが...

第2章　記述統計学

第3章　推測統計学

第II部

数理統計学

第4章　正規分布の導出

第III部

生物統計学

第5章　生物統計学：基本

生物統計学の基本を扱う.

5.1　理解度チェックリスト

生物統計学の基本の理解度チェックリスト [2] を以下に示した.

- 基本
  - □ 記述統計学と推測統計学の違いは？
  - □ 母集団を意識して研究しているか？(研究者の基本)
  - □ 母集団と標本の違いは？
  - □ 母数とは？
  - □ 統計量とは？
  - □ 3 種類の標準偏差の違いは？
    - – 母標準偏差（母分散）
    - – 標本標準偏差（標本分散）
    - – 不偏標準偏差（不偏分散）
- 標本分布，推定，検定
  - □ 標本分布とは？
  - □ 標準偏差（SD）と標準誤差（SE）はどう違うか？
  - □ SD と SE はそれぞれどんな意図を持って用いられるか？
  - □ 正規性の検定，糖分酸性の検定は何のためにあるのか？必要か？
  - □ パラメトリック検定，ノンパラメトリック検定とは何か？どう使い分ける？
  - □ 対応のある（関連した）検定と対応のない（独立した）検定はどう違うのか？[1)]
  - □ 片側検定，両側検定はどう違うのか？
  - □ 有意差とはどう言う意味か？どのようにして有意差を決めるのか？
  - □ 帰無仮説，対立仮説，危険率，有意水準の意味とは？
- 分散分析，実験計画
  - □ 一元配置分散分析で何が分かるのか？
  - □ 3 群以上ではなぜ t 検定ではなく，多重比較なのか？
  - □ 二元配置分散分析で何が分かるのか？どのような実験に使えるのか？
  - □ どうすれば有意差の得られやすい実験計画が立てられるのか？
  - □ 実験計画の立案に統計の知識はどうかかわるのか？

[1)] 対応のある検定 paired test，対応のない検定 unpaired test

5.2 標準偏差の種類

母標準偏差 … 母集団の平均からのずれ（ばらつき）を表す値.

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

標本標準偏差 … 標本の平均からのずれ（ばらつき）を表す値.

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}}$$

不偏標準偏差 … 母標準偏差を推定するための値.

$$u = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

5.3 標本平均の分布

母集団が $N(\mu, \sigma^2)$ から標本抽出を行った時，サンプルサイズを n，各観測値を X_i，観測値の平均値を $\bar{X}$，とする．各観測値は独立とする.

各観測値は $N(\mu, \sigma^2)$ に従うため，以下のように表せる.

$$X_1 \sim N(\mu, \sigma^2) \tag{5.1}$$
$$X_2 \sim N(\mu, \sigma^2) \tag{5.2}$$
$$\vdots \tag{5.3}$$
$$X_n \sim N(\mu, \sigma^2) \tag{5.4}$$

各観測値は独立であるため，以下が成り立つ.

$$X_1 + X_2 + \cdots + X_n \sim N(\mu + \mu + \cdots + \mu, \sigma^2 + \sigma^2 + \cdots + \sigma^2) \tag{5.5}$$
$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2) \text{（「観測値の和」が従う確率分布）} \tag{5.6}$$

上記より，標本平均 $\bar{X}$ の従う分布が導ける.

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \sim N(\mu, \frac{\sigma^2}{n}) \ (\because V(cX) = c^2 V(X)) \tag{5.7}$$
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{5.8}$$

5.3.1 補足：期待値，分散の性質

標本平均の従う分布を導く際に利用した期待値の性質をまとめておく．

- 期待値

$$E(X) = \sum_{i=1}^{n} x_i f(x) \text{（離散型）} \tag{5.9}$$

$$E(X) = \int_{-\infty}^{\infty} x f(x) \text{（連続型）} \tag{5.10}$$

- 期待値の演算

$$\begin{aligned} E(c) &= c \text{（定数）} \\ E(X+c) &= E(X) + c \\ E(cX) &= cE(X) \text{（期待値のスカラー倍）} \\ E(X+Y) &= E(X) + E(Y) \text{（期待値の加法性）} \end{aligned} \tag{5.11}$$

続いて分散の性質を列挙する．

- 分散

$$V(X) = E\{(X-\mu)^2\} \text{（定義）} \tag{5.12}$$

$$V(X) = \sum_{i=1}^{n} (x_i - \mu)^2 f(x) \text{（離散型）} \tag{5.13}$$

$$V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) \text{（連続型）} \tag{5.14}$$

- 分散の定義から導かれる性質

$$V(X) = E(X^2) - (E(X))^2 \tag{5.15}$$

証明

$$\begin{aligned} V(X) &= E\{(X-\mu)^2\} \text{（}\because \text{分散の定義）} \\ &= E(X^2 - 2\mu X + \mu^2) \text{（}\because \text{展開）} \\ &= E(X^2) - 2\mu E(X) + E(\mu^2) \text{（}\because \text{期待値の加法性，期待値のスカラー倍）} \\ &= E(X^2) - 2\mu^2 + \mu^2 \text{（}\because E(X) = \mu\text{，定数の期待値）} \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - (E(X))^2 \text{（}\because E(X) = \mu\text{，定数の期待値）} \end{aligned}$$

- 分散の演算

$$\begin{aligned} V(c) &= 0 \\ V(X+c) &= V(X) \\ V(cX) &= c^2 V(X) \end{aligned} \tag{5.16}$$

証明 $V(cX) = c^2 V(X)$

$$\begin{aligned} V(cX) &= E\{(cX - E(cX))^2\} \text{（}\because \text{分散の定義 } V(X) = E\{(X-E(X))^2\}\text{）} \\ &= E\{(cX - cE(X))^2\} \text{（}\because \text{期待値のスカラー倍）} \\ &= E\{c^2(X-E(X))^2\} \\ &= c^2 E\{(X-E(X))^2\} \text{（期待値のスカラー倍）} \\ &= c^2 V(X) \text{（}\because \text{分散の定義 } E\{(X-E(X))^2\} = V(X)\text{）} \end{aligned} \tag{5.17}$$

5.4 標本平均の分布による区間推定

前節では，標本平均 $\bar{X}$ が従う分布を求めることができた．これにより,「標本平均の」標準偏差が $\frac{\sigma}{\sqrt{n}}$ であることが分かった．この「標本平均の分散」のように，ある母集団から標本抽出を行い，その観測値から得た**統計量のばらつきを標準偏差で表したものを「標準誤差 standard error（SE）」**という．

統計量を指定せずに単に「標準誤差」と言った場合，**「標本平均の」標準誤差 stardard error of the mean（SEM）**のことを指す．

- 標準誤差 … 統計量のばらつきを標準偏差で表したもの．統計量を指定せずに単に「標準誤差」と行った場合，標本平均の標準偏差（標本平均の「標準誤差」）を表すことが多い．

ここでは，標準誤差を「標本平均の標準誤差」として扱う．標準誤差は標本平均の標準偏差であるため，

- 区間 $\mu \pm 1SE$ … 標本抽出を行って標本平均を算出した時，67% の確率でその区間に収まる
- 区間 $\mu \pm 1.96SE$ … 標本抽出を行って標本平均を算出した時，95% の確率でその区間に収まる

5.5 正規分布のグラフの概形

正規分布 $N(\mu, \sigma^2)$ に従う母集団[2]から以下の条件で標本抽出を行った場合を考える．

- $N(\mu, \sigma^2) = N(50, 20^2)$
  - 母平均 50，母標準偏差 20（母分散 400）
- サンプル数：10[3]
- サンプルサイズ：30[4]

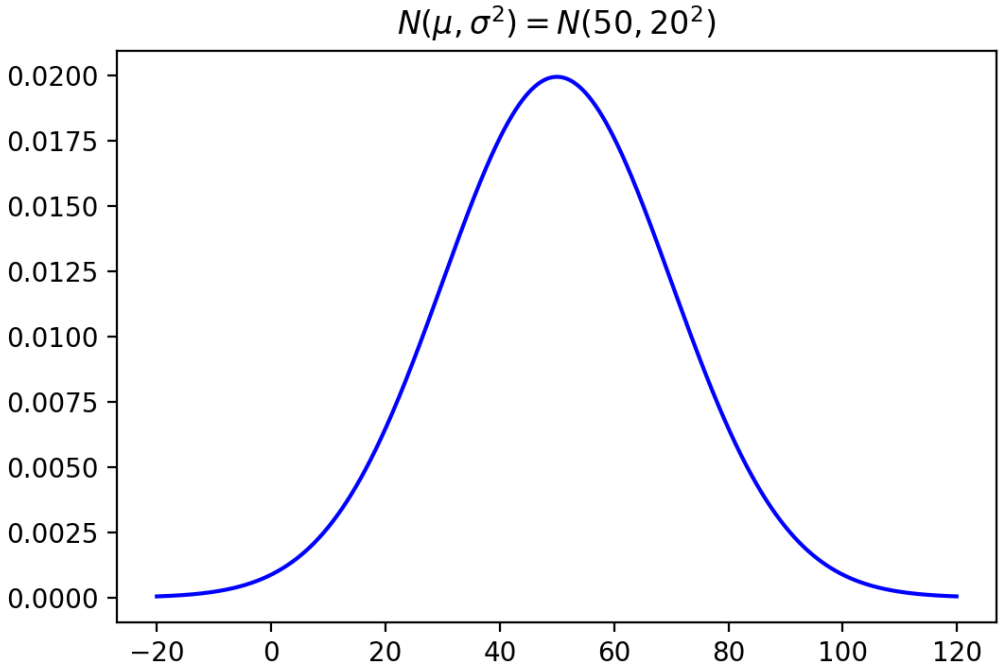

図 5.1: $N(\mu, \sigma^2) = N(50, 20^2)$ の正規分布

一つのサンプルを $X_i = \{x_1, x_2, \ldots, x_{30}\}$ としたとき，

5.6 パラメトリック検定，ノンパラメトリック検定

5.7 2 群間比較の統計手法

5.8 3 群感比較の統計手法

[2] 正規分布に従う母集団を「正規母集団」という

[3] サンプル数 the number of samples … 群数，サンプルの数（**「サンプル ＝ 群」と覚えておけば間違えない．**）．詳しくは「サンプル数とサンプルサイズは意味が違う」を参照.「サンプル」の誤用として多いのが,「個々の測定値（観測値）をサンプルと思っている」である．

[4] サンプルサイズ sample size … 1 サンプルの大きさ．

第6章　生物統計学：応用

6.1　多変量解析

6.2　クラスタリング

6.3　主成分分析

第7章　オミクス解析

7.1　オミクス解析の基本的な流れ

7.2　トランスクリプトーム

7.2.1　Enrichment 解析

基本,「発現が上がっている遺伝子と下がっている遺伝子の組」という発想が Enrichment 解析の精神.「(有意に) 上がっている，下がっている」をどう決めるかは関連分野の論文を参考にして定める，もしくは独自の指標で定める.

7.2.2　Enrichment 解析の注意点

GO(Gene Ontology) によって大規模データの Enrichment 解析をする時，多重検定補正をしてもなお，補正後 p-value（q-value）は実際より高めに見積もられてしまう．これは，各々の遺伝子に割り当てられた annotation term が重複してる場合があるためである.

「重複を免れない」というデータ構造上の問題まできちんと明示的に考えて Enrichment 解析を使ってる生物系の論文ってあまり見ない気がする．もちろん，予め有意水準低めに設定して検定を厳しくするとかはできる.

7.2.3　メタボローム

第IV部

統計学インデックス

統計学図鑑 [1] を参考に，統計学のインデックスを作成する．統計学の基礎部分の体系がどのようになっているかをざっと確認できるようにするのが目的.

第8章　統計学とは

8.1　統計学の分類

統計学の分類を確認する.

- 統計学 … 対象集団，現象の特徴を把握する方法を体系化した学問
  - 記述統計学 … 手元のデータの特徴を捉える（平均，ばらつき，相関係数など）
  - 推測統計学 … 手元のデータの特徴を捉える（点推定，区間推定，仮説検定）
    - 頻度論的統計学 … 一般的な統計学
    - ベイズ統計学（ベイズ統計学を推測統計学に含めないとする考え方もある）

第9章　記述統計学

9.1　代表値

9.1.1　平均値

9.1.2　中央値

9.1.3　最頻値

9.2　データのばらつき

9.2.1　分散，標準偏差

9.2.2　分位数

9.2.3　変動係数

「相対標準偏差」と言われれば一発で理解できる．

- 変動係数 coefficient of variattion，CV
  - 標準偏差を相対化した数値．標準偏差を平均値で割ったもの.「**相対標準偏差**」ともいう．
  - 各群の標準偏差を平均値で補正することで，**単位，平均値が異なる群間のばらつきを比較することができる**．

$$CV = \frac{\sqrt{s}}{\bar{x}} \tag{9.1}$$

9.3　変量の関連性

9.3.1　相関係数

ピアソンの積率相関係数

一般に,「相関係数」と言った場合「ピアソンの積率相関係数」を指すことが多い．

9.3.2　順位相関係数

- 順位相関係数
  - 2 つの順序変数間の相関の強さを測る指標.「スピアマンの順位相関係数」と「ケンドーコバヤシの順位相関係数」がある．どちらの指標を用いるかに明確な基準はない（要出典）．

スピアマンの順位相関係数

順位データに対して「ピアソンの積率相関係数」を計算した数値が「**スピアマンの積率相関係数**」である．データが連続的な値を取る場合，まず順位データに変換する．

ケンドールの順位相関係数

第10章　確率分布

10.1　確率変数，確率分布

第11章　誤差論

11.1　有効数字

11.2　系統誤差と偶然誤差

11.3　誤差の伝搬

全ての計算を終えてから計算を行うか，間に平均などの集計を挟んで計算を行うか．後者の場合，誤差の扱いが難しい．

第12章　推定

12.1　何を「推定」するのか：母数

推定の対象は，母数と呼ばれる母集団分布の特徴を表す定数である．

- 母数 parameter
  - 母集団分布の特徴を表す定数．
  - 母平均，母分散，母標準偏差など．

12.2　統計量

12.3　統計的推測における「正規分布を仮定する」という前提

流れの中で，今どんな前提条件，または仮定の元で論理を進めているか意識する必要がある．

12.4　点推定

母数を「1 点の数値」として推定する．幅を持たない推定．

12.4.1　推定量

12.4.2　最尤推定

12.4.3　モーメントを用いた推定

12.4.4　推定量の満たすべき性質

12.5　区間推定

母数を「取りうる値の区間」として推定．幅を持たせた推定．

12.5.1　標本分布

統計量が従う確率分布のこと．

12.5.2　母平均の信頼区間

12.5.3　母比率の信頼区間

12.5.4　母分散の信頼区間

12.5.5　母相関係数の信頼区間

12.5.6　母相関係数の信頼区間

12.5.7　シミュレーションで母数を推定する：ブートストラップ法

ブートストラップ法

- ブートストラップ法
  - 手元にある n 個のデータから同じサイズの再標本を何度も復元抽出し，その再標本の統計量から母数を推定する統計手法．

- 手元のデータから復元抽出を繰り返して（リサンプリング）たくさんの再標本を生成し，その統計量から母数を推定する方法.
- 小標本の場合など，母集団に確率分布が仮定できなくても母数の推定を可能にする.
- 統計学における「モンテカルロ法（コンピュータ・シミュレーション法）」の 1 つだが，乱数を用いず，実際にあるデータを使って分布を推定する.

手元にある少ないデータだけで母数を高い精度で推定するにはどうすればいいか考えた時,

母集団分布に正規性を過程できない

区間推定では母集団分布が正規性であることを仮定することが多いが，正規性を無理に仮定して t 分布を用いて推定しても，誤差が多すぎて実用に絶えない推定になる（信頼区間が広すぎるなど）.

第13章　仮説検定

13.1　仮説検定とは

帰無分布 … 帰無仮説が正しいものとして考えた時の標本分布.

13.2　仮説検定の分類：パラメトリック検定とノンパラメトリック検定

13.3　帰無仮説と対立仮説

13.4　仮説検定の手順

どのタイミングで「母集団分布」を仮定するのかに注意する.

13.5 仮説検定におけるリスク：第一種の過誤，第二種の過誤

13.5.1 第一種の過誤，第二種の過誤を図表で理解する

どっちがどっちかややこしいけど，以下の画像を見ればすっきりする．

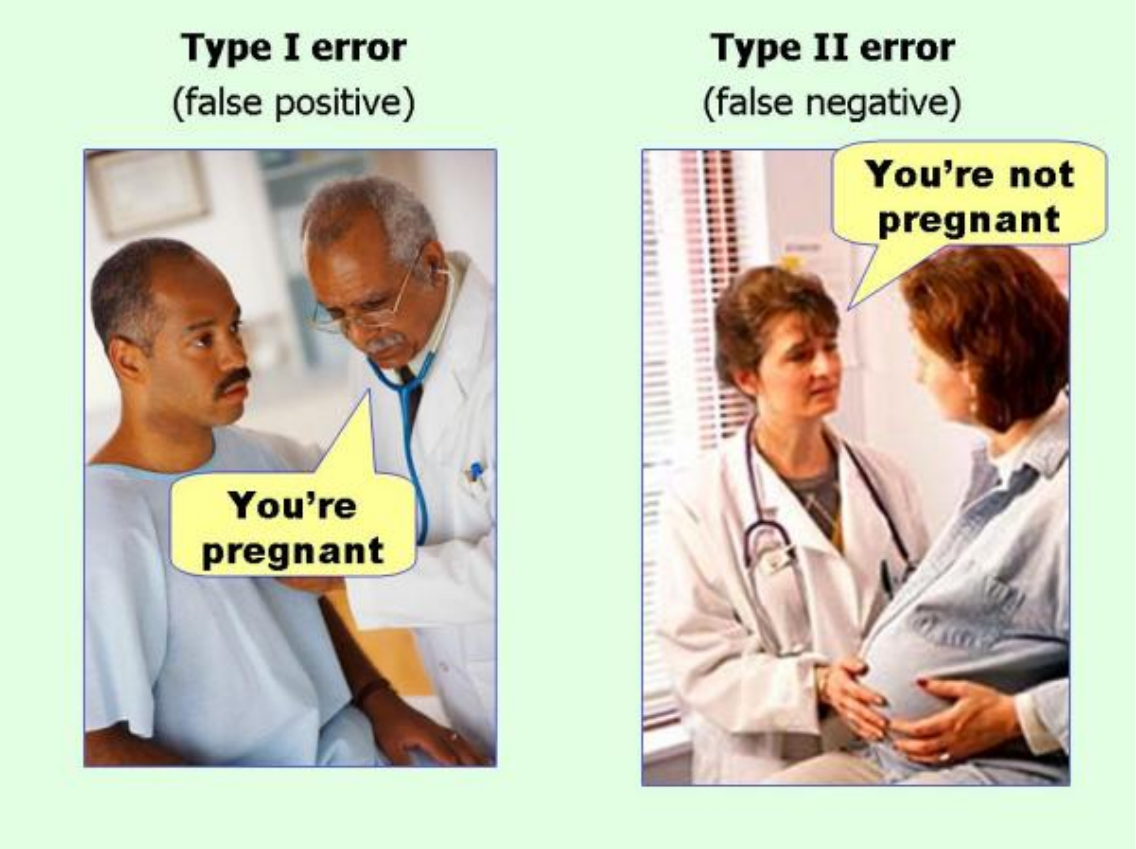

図 13.1: 第一種の過誤（偽陽性），第二種の過誤（偽陰性）

（引用元：I always get confused about Type I and II errors. Can you show me something to help me remember the difference?）

ただ，上記画像だけでは理解が不十分．第一種の過誤，第二種の過誤を犯してしまう際，それぞれを犯す確率というのが算出できる．それを確率分布の図で表したものが以下．

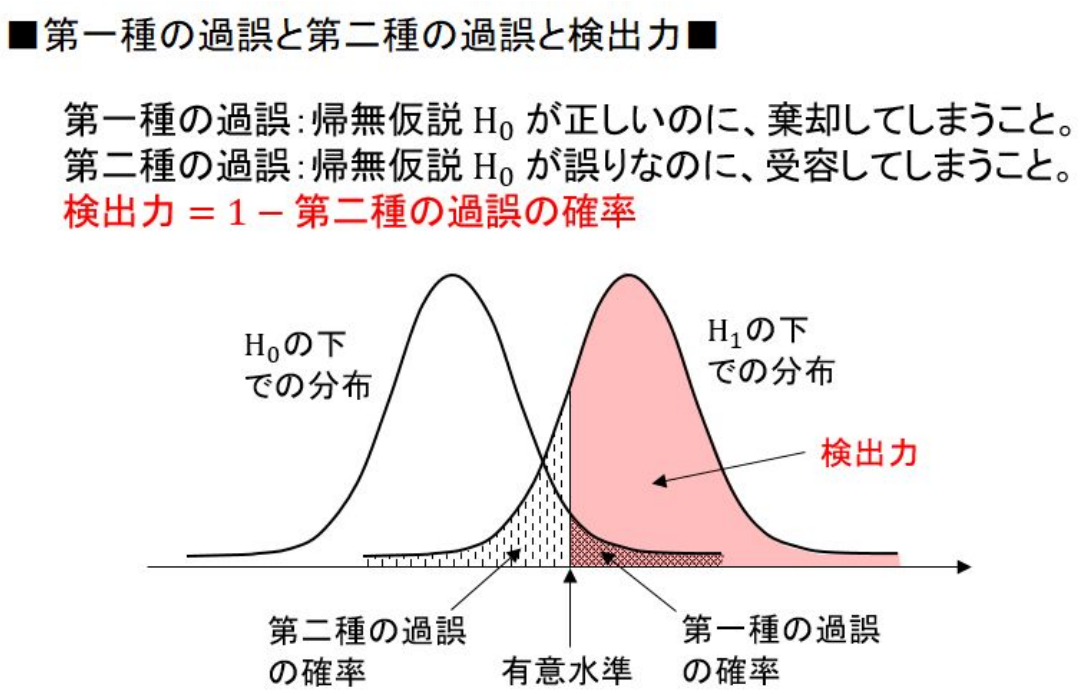

図 13.2: 確率分布で見る第一種の過誤（偽陽性），第二種の過誤（偽陰性）

（引用元：Twitter より）

よく見る，第一種の過誤，第二種の過誤を整理した表は以下．この表だけに頼ると永遠に忘れ続ける．ただし，下記画像の「差があると判断」は正しくない．帰無仮説を「差が無い」としたとき，帰無仮説を棄却した結果主張できるのは「差が無いとは言えない」である（結果的に「差がある」と判断するだけれど）である．

（この主張↑間違っている可能性あり.「判断」だから間違いでは無い？）

表4. 第1種の過誤と第2種の過誤

		実際の状態：偽陽性	実際の状態：偽陰性
		偶然が原因（本当は差がない）	偶然は原因ではない（本当は差がある）
帰無仮説に関する判断	帰無仮説を採択（差がないと判断）	適切な採択（1−α）	第2種の過誤（不適切な採択）（β）
	帰無仮説を棄却（差があると判断）	第1種の過誤（不適切な棄却）（α）	適切な棄却（1−β）

第1種の誤りの確率=α，第2種の誤りの確率=β，1−βを検定力という

図 13.3: 表で見る第一種の過誤（偽陽性），第二種の過誤（偽陰性）

（引用元：Twitter より）

13.5.2 「第一種の過誤＝有意水準」の意味するところ

ここまでで，なぜ「第一種の過誤＝有意水準（危険率）」なのかと疑問に思った．それに対する回答を以下に述べる．

有意水準は,「帰無分布に従うある事象が起こる確率が，有意水準より低かったら確率的に起こりにくいことであるとみなす（帰無分布から取り出される確率が低いとみなす）」という意味の（人為的な）基準である．しかし，ある事象が起こる確率が有意水準より低くても，実は「その事象が（起こる確率は低くとも）帰無分布から発生した事象である」ということがありうる．

つまり，実際には帰無仮説が正しいのに,（人為的に決めた）棄却域に入っている値を帰無分布からサンプリングしてしまったが故に，仮説検定によって有意だと検出されてしまったということである．

有意水準 $\alpha = 0.05$ のとき，**帰無分布からサンプリングすると 0.05 の確率で棄却域に存在する標本を抽出してしまう**．そして，この棄却域から抽出した標本に対して仮説検定を行うと有意だと判定されてしまう.「実際には違うのに，有意だと判定されてしまう」，つまり，これが「偽陽性」であり，サンプリングしたときに第一種の過誤が起こる確率が有意水準である 0.05 そのものなのである.「有意水準」は,第一種の過誤を起こしてしまう危険性を数値的に示す値として「**危険率**」とも呼ばれる．

13.6 母比率の検定

13.7 母分散の検定

13.8 相関の有無の判定：無相関の検定

ただし，相関係数は「線形な」相関関係を示すものであり,「非線形な」関係の相関は表現できない．

13.9 平均値の差の検定：対応のない 2 群の検定

ほとんどの生物実験．

13.10 平均値の差の検定：対応のある 2 群の検定

マウスの血圧測定．

13.11 比率の差の検定：対応のない 2 群の場合

13.12 非劣性試験

第14章　分散分析

ANOVA（ANalysis Of VAriance）

14.1　一元配置分散分析：実験で効果を確かめる

14.2　多群の等分散の検定：Bartlett 検定

14.3　対応のある一元配置分散分析：個体差を考慮する

14.4　二元配置分散分析：交互作用を見つけ出す

第15章　検定の多重性

15.1　検定の多重性とは

15.2　多重検定補正

15.2.1　Bonferrorni 法

かなり厳しい多重検定補正の手法.

15.2.2　Benjamini-Hochberg 法

Benjamini-Hochberg 法：FDR を制御して多重比較検定補正を行う方法

15.3　Storey 法

15.4　Tukey 法，Tukey-Kramer 法

15.5　Dunnett 法

第16章　ノンパラメトリック検定

16.1　ノンパラメトリック検定：分布によらない仮説検定

- パラメトリック検定 … 母集団分布を仮定した仮説検定.
- ノンパラメトリック検定 … 母集団分布を仮定しない仮説検定.

16.2　独立性の検定（ピアソンの χ^2 検定）：質的データの検定

16.3　フィッシャーの正確確率検定：2 × 2 分割表の検定

16.4　対応のない 2 群の順序データの検定：マンホイットニーの U 検定

16.5　対応のある 2 群の順序データの検定：符号検定

16.6　対応のある 2 群の量的データの検定：ウィルコクソンの符号付き順位和検定

16.7　対応のない多群の順序データの検定：クラスカル・ウォリス検定

16.8　対応のある多群の順序データの検定：フリードマン検定

第17章　実験計画法

17.1　フィッシャーの 3 原則：1 反復

17.2　フィッシャーの 3 原則：2 無作為化

17.3　フィッシャーの 3 原則：3 局所管理

17.4　色々な実験配置

17.5　直交計画法：実験を間引いて実施する

17.6　直交計画法の応用 1：品質工学（パラメータ設計）

17.7　直交計画法の応用 2：コンジョイント分析

17.8　検出力分析：標本サイズの決め方

17.8.1　検出力とは

17.8.2　検出力分析

17.8.3　補足：サンプルサイズとサンプル数

- サンプルサイズ，標本サイズ，sample size
  - 1 標本に含まれる観測データの個数.
- サンプル数，標本数，the number of samples
  - 抽出した標本の数．つまり，何回標本抽出したかを表す数.

第18章　回帰分析

18.1　回帰分析：原因と結果の関係を探る

18.2　最小二乗法

18.3　決定係数：回帰線の精度を評価する

18.4　t 検定：回帰線の傾きを検定する

18.5　残差分析：分析の適切さを検討する

18.6　重回帰分析：原因が複数あるときの回帰分析

18.6.1　多重共線性：説明変数間の問題

略して「マルチコ」と呼ばれることもある.

18.7　変数選択法：有効な説明変数を選ぶ

18.8　質の違いを説明する変数 1：切片ダミー

18.9　質の違いを説明する変数 2：傾きダミー

18.10　プロビット分析：2 値変数の回帰分析

18.11　イベント発生までの時間を分析する 1：生存曲線

18.12　イベント発生までの時間を分析する 2：生存曲線の比較

18.13　イベント発生までの時間を分析する 3：Cox 比例ハザード回帰

第19章　多変量解析

19.1　多変量とは

多変量データの意味から言えば,「単回帰分析」,「重回帰分析」も多変量解析に入るが，前章で扱ったためこの章では扱わない.

19.2　主成分分析：情報を縮約する

次元削減，次元圧縮

19.3　因子分析：潜在的な要因を発見する

19.4　多次元尺度構成法 MDS

19.5　クラスタリング

神嶌先生のリンクが非常に分かりやすい.

- クラスタリングの参考リンク
  - 神嶌 敏弘：クラスタリング Clustering
  - クラスタリング（クラスター分析）

19.5.1　非階層的クラスタリング

- kmeans 法

19.5.2　階層的クラスタリング

- 最短距離法 nearest neighbor method
- 最長距離法 furthest neighbor method（完全連結法 complete linkage method）
- 群平均法 group average method
- ウォード法 Ward's method

19.5.3　重要：クラスタリングの結果の解釈

最も重要な点は，クラスタリングは探索的（exploratory）なデータ解析手法であり，クラスタリングによる分割は分析者のなんらかの主観や視点に基づいている，ということである．よって，クラスタリングした結果は，データの要約などの知見を得るために用い，客観的な証拠として用いてはならない.

19.6　構造方程式モデリング（SEM）：因果構造を記述する

19.7　コレスポンデンス分析：質的データの関連性を分析する

19.8　数量化 1 類

19.9　数量化 2 類

19.10　数量化 3 類

第20章　統計モデリング

20.1　統計モデルとは

20.2　統計モデルの種類と発展

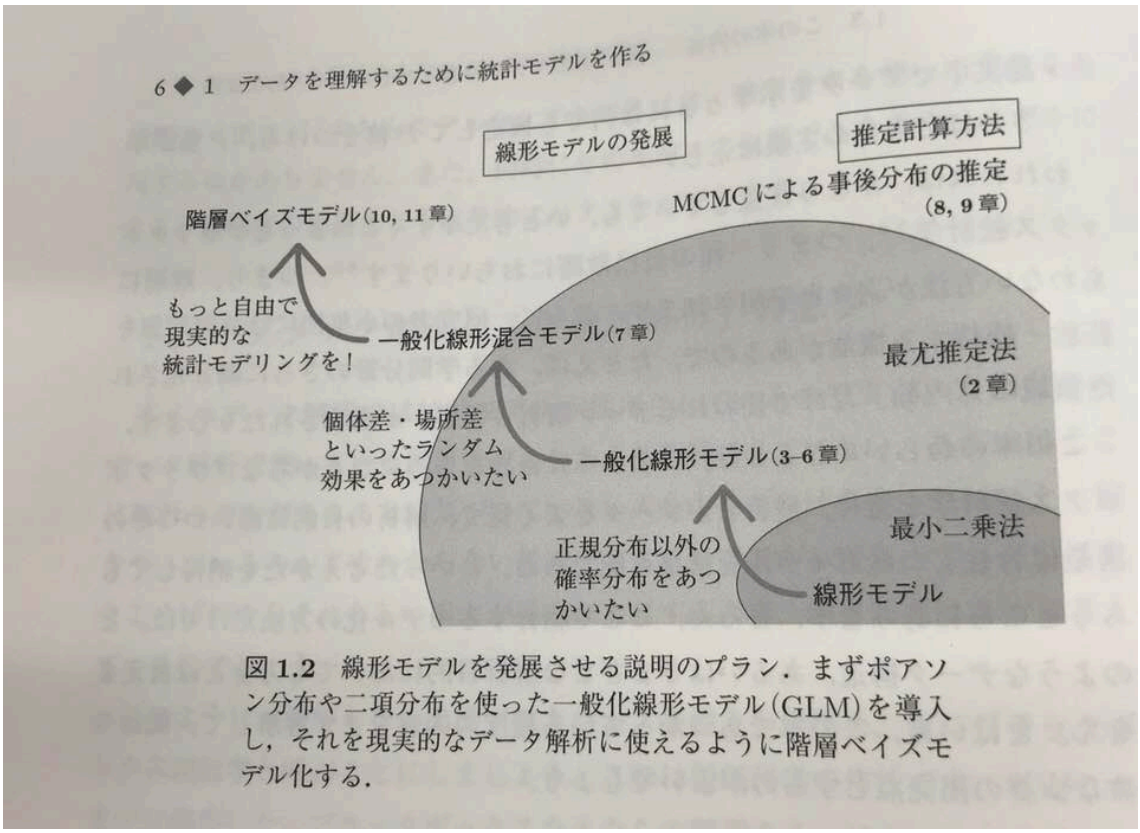

図 20.1: 統計モデルの発展

20.3　線形モデル

20.3.1　最小二乗法

20.4　一般化線形モデル

20.4.1　最尤推定法によるパラメータ推定

20.5　一般化線形混合モデル

20.6　階層ベイズモデル

第21章　統計的因果推論

第22章　ベイズ統計学：基本

第Ⅴ部

日本統計学会 - 統計検定に関するまとめ

第23章　統計検定の情報の整理

23.1　統計検定 2 級

23.1.1　試験要旨

- 公式 HP：統計検定 2 級
- 出題範囲

23.2 統計検定 1 級

23.2.1 試験要旨

- 公式 HP：統計検定 1 級
- 出題範囲

関連図書

[1] 栗原伸一, 丸山敦史. 統計学図鑑. オーム社, 2020.

[2] 池田郁男（東北大学大学院農学研究科）. 実験で使うとこだけ 生物統計 1 キホンのキ 改訂版. 羊土社, 2017.