統計学メモ

@mathnuko

2020/02/12 〜

目 次

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．

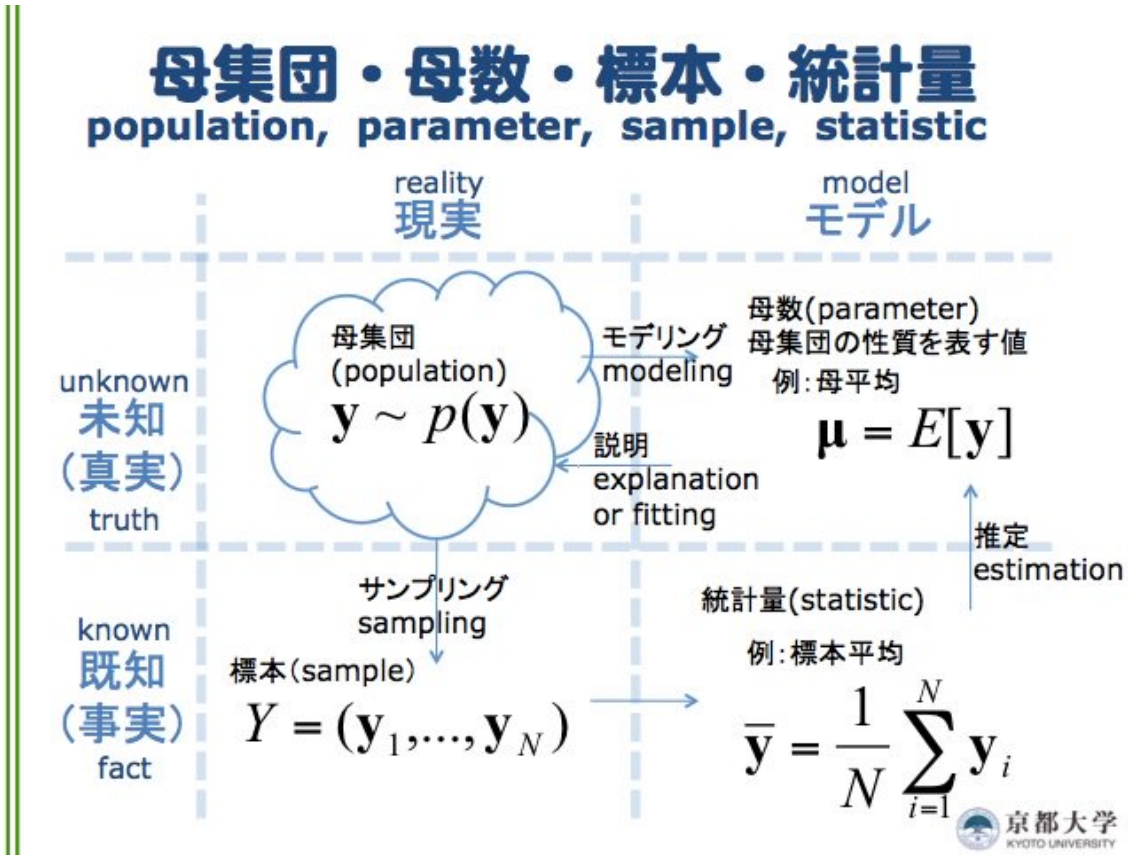

図 1.1: 統計学の体系（hoge より引用）

[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 [1] 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]

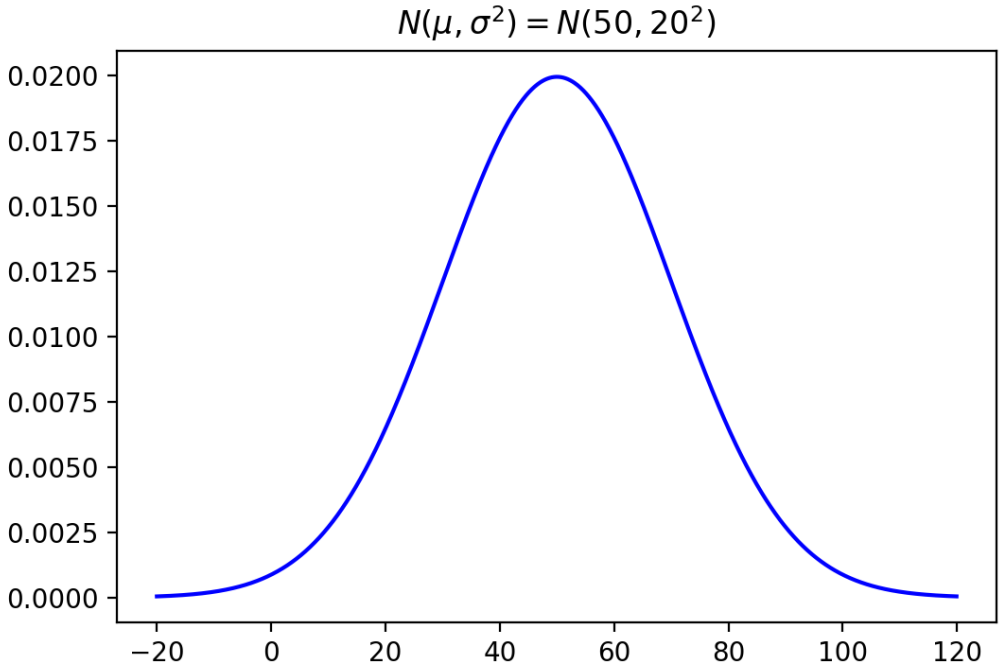

図 5.1: $N(\mu, \sigma^2) = N(50, 20^2)$ の正規分布

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

第Ⅴ部

日本統計学会 - 統計検定に関するまとめ

第8章　統計検定の情報の整理

8.1　統計検定 2 級

8.1.1　試験要旨

- 公式 HP：統計検定 2 級
- 出題範囲

8.2 統計検定 1 級

8.2.1 試験要旨

- 公式 HP：統計検定 1 級
- 出題範囲

関連図書

[1] 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.