

データ解析の基礎

高木 廣文
東邦大学医学部看護学科
国際保健看護学研究室

e-mail: halwin@med.toho-u.ac.jp
http://homepage2.nifty.com/halwin/takagi.html

1

統計学と統計について

統計学 statistics とは何か？

- ・統計： 統計をとる（？）
- ・統計学： 統計学を使う（？）

2

「統計をとる」とは？

- ・アンケート調査で学生のアルバイト実施を調べる。
- ・ある病院の診療科別外来患者数を調べる。

データを収集する（データをとる）。
集計をする。人数を数える。等々。

3

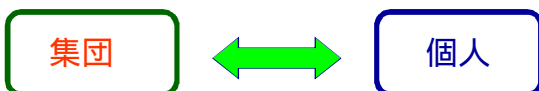
統計学とは？

- ・数理統計学
- ・生物統計学
- ・経済統計学
- ・

データに内在する傾向を明らかにするための科学的方法論を与える

4

統計学の対象は何か？



集団がもつ各項目や特性などの傾向についてデータから検討するための方法を提供

5

統計学の立場

統計学の特徴

- ・現象の数量化（データ化）
 - ・各種検査値
 - ・臨床的な症状や患者の性格特性
 - ・QOLなど
- ・「質」を「量」として把握
 - ・客観性を高める。
「再現性」, 「比較可能性」

科学的な研究に不可欠

6

統計学的なものの捉え方

現象を統計学ではどのように把握するか？

- (1) 具体性, 現実性
現実的な具体的な現象のみを扱う。
- (2) 操作性
具体的に扱うために数字で表現する。
- (3) 変動性
対象を常に変動するものとする。
- (4) 傾向性
変動性の中に傾向性が存在する。

7

統計学を使う目的

統計学を目的により大きく2つに分けて考えることがある。

- (1) 特定の事象の「記述」
「記述統計学」 descriptive statistics
- (2) 調査結果や研究結果の「一般化」
「推測統計学」 inferential statistics
検定, 推定

8

記述統計学の方法

データをまとめて, ある特性を示す。

- 図や表を用いて示す方法
- ひとつの数値で示す方法

9

データを図示する手法

「円グラフ」, 「帯グラフ」, 「棒グラフ」,
「ヒストグラム」, 「折れ線グラフ」,
「幹葉表示 stem-and leaf display」,
「箱ヒゲ図 box and whisker plot」,
「相関図(散布図)」など

10

データを要約するための指標

- 「代表値」 - 平均値, 中央値, 最頻値
- 「散布度」 - 分散, 標準偏差, 変動係数
- 「相関係数」, 「割合」, 「クロス表」 など

11

推測統計学の方法

データから得た結果を一般化, 普遍化する。

推定 estimation

データ(標本)から一般集団(母集団)の特性を求める:

母平均値, 母比率の信頼区間など

検定 test

データから一般集団の特性に関する仮説を検証
独立性の検定, 無相関の検定など

12

データ解析について

従来の多くの研究:

統計的検定の多用 - **確証的解析**
confirmative analysis

記述的方法を多用, データに依存した解析:
探索的データ解析 exploratory data analysis

幅広い知識や教養, また洞察力・直観力も極めて重要。方法の選択や結果の解釈を正しく行うには, 統計学に関する正確な知識も必要

1

データとは何か

例1 . A子さんの身長は160cmです。

- ・ A子さんの身長の測定結果。
- ・ A子さんの身長の「**データ**data」。

例2 . 学校で行われる身体計測

- ・ あるクラス30人の身長の一覧表。
- ・ そのクラスの「**身長**の**データ**」。

14

ラベリングについて

例3 . A子さんの血液型はA型です .

- ・ 血液型のデータを示している点は, 身長の場合と全く同じである。
- ・ 血液型は, ある特定の反応の有無により A, B, O, AB の4タイプに分類

標識付け

(ラベリング labeling, ラベル付け)

15

データの定義

データとは

個体のある特性について測定を行い, 適当なラベルを付けたもの。
もしくは, その全体, およびそれらをまとめたもの

16

データの分類

1) 身長や体重のデータ

- ・ ある「物差し」を用いて測定: 数値で表現
- ・ データを足したり引いたりできる。

量的データ quantitative data

2) 血液型のデータ

- ・ ある特性に名称をつけたもの。
- ・ それぞれを足したりすることは不可能。

質的データ qualitative data

17

量的データの分類

例1 . A子さんの体重は50Kg,
K子さんの体重は60Kgです。

(1) K子さんはA子さんより10Kg体重が重い。

- ・ データの「**差**」の計算ができる。

(2) K子さんはAさんの1.2倍の体重がある。

- ・ データの「**比**」の計算ができる。

「**比尺度** ratio scale」によるデータ

18

間隔尺度

例2 . 一日の最高気温20 , 最低気温10 。

(1)一日の気温の差は10 である。

差の計算可能である。

(2)最高気温は最低気温の2倍である？

× 比の計算不可。

「間隔尺度 interval scale」によるデータ

原点0のもつ意味による相違：

- ・ 負のデータの存在。
- ・ 比尺度に負のデータはない！

19

質的データの分類

例1 . 血液型のデータの場合：

A子さんはA型 , K子さんはA B型。

・ 差の計算： $A - AB = A(1 - B)$ ？

・ 比の計算： $AB / A = B$ ？

- ・ A型とB型の差や比を取ることは不可能。どちらが大きいともいえない。

「名義尺度 nominal scale」によるデータ

20

順序尺度

例2 . あなたは寝起きはよいですか」の
ような質問項目への回答

1.非常によい 2.よい 3.悪い 4.非常に悪い

- ・ 各カテゴリについての数値1~4の差や比は計算できない。
- ・ 数値が大きくなるにつれ、寝起きが悪くなるという、順序がある。

「順序尺度 ordinal scale」によるデータ

21

データの分類再考

データの測定尺度によるまとめ

(1)量的データ：

(A)比尺度によるデータ

(B)間隔尺度によるデータ

(2)質的データ：

(C)順序尺度によるデータ

(D)名義尺度によるデータ

変更可能

22

データの「質」と基本的統計手法

(1)量的データ：

平均値 , 分散 , 標準偏差 , 相関係数 , など

(2)質的データ：

人数 , 割合 , クロス表など

求められる基本統計量が異なる！

23