大学院　ライフサイエンス統計学講義
2019年度

第1回
統計データの分類

授業の進め方（授業計画，課題など）
イントロダクション
1．統計データの分類

授業の進め方

- 担当者：稲葉由之　明星大学経済学部
- yoshiyuki.inaba@mesei-u.ac.jp
  - 欠席時に提示した課題の確認などは，他の受講生に聞くか，メールで私まで連絡してください．
- 注意点：次の授業までに内容をスライドで復習してください．
  - 理解不足の箇所が1か所でもあると，それ以降の内容を理解できなくなる可能性があります．

2

授業の目的

- 本授業では，基本的な記述統計学から推測統計学における推定・検定までの統計学の基礎的な内容を習得することを目的とする．可能ならば，統計的検定の詳細について習得することも含める．
- 目標：統計検定2級程度の知識を得る．
  - 統計検定2級の出題範囲すべてを説明することはできません．
- 統計検定2級の範囲を網羅的に説明するよりも，統計学の基本的な考え方を理解することを目的とします．

3

授業日程

- 以下の月日の4時限（15:00-16:45）12回を予定している．
  - 6月
    - 10日（月），17日（月），19日（水），24日（月），26日（水），
  - 7月
    - 1日（月），3日（水），8日（月），17日（水），22日（月），29日（月），
  - 8月
    - 5日（月），予備日7日（水）

4

参考書

- 教科書ではなく参考書とした理由（口述します）
- 『統計学【第2版】』刈谷武昭，勝浦正樹，東洋経済新報社
  - 統計学の基礎知識に不安のある方におすすめ
  - 授業内容の項目1～16に該当する
- 『医学への統計学【第3版】』丹後俊郎，朝倉書店
  - 辞書的な使い方も可能
  - 授業内容の項目10，13～に該当する
  - 授業の進度により，この部分に到達しない可能性もあるため，はじめから購入しないで結構です．

5

授業の内容（1/4）

1．データの分類
  - 4つの尺度，量的変数と質的変数，データの分類とグラフ表現との関係
2．中心の位置の統計量
  - 平均値，中央値，最頻値，様々な平均値（幾何平均など），平均値と中央値の性質
3．散らばりの統計量
  - 分散の考え方，標準偏差，分位数，四分位範囲，範囲
4．標準偏差の活用
  - チェビシェフの不等式，変動係数，標準化
5．グラフ表現
  - 箱ひげ図，度数分布，ヒストグラム，モザイク図

6

授業の内容(2/4)

6. 2変数の関連性
   - 散布図，共分散，相関係数，尺度に対応した様々な相関係数
7. 統計的記述に関する話題
   - 幹葉表示，変化の文章表現，寄与度，統計データの入手方法，ソフトウェアR
8. 母集団と標本
   - 母集団，標本，無作為抽出，標本誤差の考え方，母数，実験計画
9. 確率と確率変数
   - 事象と確率，確率の公理，条件付き確率，ベイズの定理，離散確率変数，連続確率変数
10. 確率分布
    - ベルヌーイ分布，二項分布，超幾何分布，一様分布，ポアソン分布，確率分布の平均値と分散，確率変数の期待値

7

授業の内容(3/4)

11. 大数の法則
    - 二項分布による大数の法則，標本平均による大数の法則
12. 中心極限定理
    - 標本平均の分布，中心極限定理，実験
13. 正規分布とカイ二乗分布
    - 標準正規分布，確率分布の再生性，カイ二乗分布
14. 母数の点推定
    - 不偏推定量，不偏分散，t分布，点推定
15. 母数の区間推定
    - 信頼区間の考え方，母平均の区間推定，母比率の区間推定，F分布

8

授業の内容(4/4)

16. 統計的仮説検定　平均値に関する推測
    - 仮説検定の考え方と流れ，帰無仮説と対立仮説，第I種の過誤と第II種の過誤，検出力，検定統計量，棄却域，両側検定と片側検定
17. 頻度に関する推測
    - 適合度検定，罹患率に関する推測，疫学研究における検定
18. 分散分析
    - 分散分析，線形モデル，ノンパラメトリックな検定，多重比較
19. 標本の大きさの決定
    - 推定精度，ネイマンの最適配分
20. いくつかのトピック

9

授業の到達目標

「授業の内容」16の内容の理解を本授業における第一の到達目標としたい．

16. 統計的仮説検定　平均値に関する推測
    - 仮説検定の考え方と流れ，帰無仮説と対立仮説，第I種の過誤と第II種の過誤，検出力，検定統計量，棄却域，両側検定と片側検定

10

平均値に関する推測の例(1/3)

- いま，全国におけるコンビニエンスストアの1回当たりの平均購買金額(以降，平均客単価という)が600円であるとする．
- ある店舗のポイントカード所有者の客単価は，全国平均よりも高いと言えるのか否かを知りたいため，統計的仮説検定を実施する．
- 調査の結果，標本50人の平均客単価は687(円)，標本不偏分散は46,225(円^2)だった．

① 帰無仮説の設定

- 帰無仮説を $H_0: \mu_X = 600$ とおき，対立仮説を $H_1: \mu_X > 600$ とした片側検定を考える．
  - 主張したいことは，ある店舗のポイントカード所有者の客単価は全国平均の600円よりも高いということのため，片側検定にした．

11

平均値に関する推測の例(2/3)

② 検定統計量の作成

- 標本の大きさ n が50であるため，帰無仮説が真である下で検定統計量 T_0 は自由度 49 の t 分布にしたがう．

$$T_0 = \frac{\bar{X} - \mu_0}{\sqrt{\frac{\widehat{\sigma_X^2}}{n}}} \sim t(49)$$

③ 危険率 α の棄却域 R の設定

- 危険率5%の片側検定における棄却域 R は，$R = \{T_0 : T_0 > t_{0.05}(49) = 1.68\}$ である．

12

平均値に関する推測の例(3/3)

④ 仮説検定の判断

- 検定統計値 t_0 は以下のように計算する.

$$t_0 = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\widehat{\sigma_X^2}}{n}}} = \frac{687 - 600}{\sqrt{\frac{46225}{50}}} \approx 2.86$$

- 検定統計値 t_0 は棄却域 R に含まれたため, 危険率(有意水準)5%で帰無仮説 (H_0: $\mu_X = 600$) を棄却し, 対立仮説を採択する.
- したがって, ある店舗のポイントカード所有者の客単価は全国平均よりも高いと判断する.
  - ただし, この判断を誤る可能性は5%あることに気を付けなければならない.

13

目標に到達するために必要な知識

- 以下の内容を理解して他者に説明できるようになることが本授業の目的です.
  - 平均値の性質(2)
  - 標準化(4)
  - 母集団と標本の理解(8)
  - 確率, 確率変数, 確率分布の知識(9,10)
  - 中心極限定理の理解, 標本平均の分布(12)
  - 正規分布, t分布の理解(13,14)
  - 統計的仮説検定における仮説の設定, 検定統計量, 棄却域の設定, 判断に関する理解(16)

14

評価方法など

- 授業への参加(出席状況によっては小テストのような参加チェックを実施することになるかもしれません), 課題(4, 5回を予定)の提出状況とその評価による.
- 授業の2, 3回に一度くらいの割合で, 授業日に課題を提示します(課題内容を配付).
  - 提出日は1週間後, または1週間経過後はじめの授業日
  - 8月5日まで, 課題は受け取りますが, 提出日に遅れた場合は評価を一定割合下げますので, 出来る限り提出日を守ってください.
  - 欠席時に課題が提示された場合, 他の受講生に聞くか, 私までメールで連絡してください.
  - 欠席により課題が提出できない場合, 他の受講生に渡して提出しても結構です.

15

近年のデータサイエンティスト等統計分野への注目

- 西内啓(2013)『統計学が最強の学問である』, ダイヤモンド社.

- 1980年代以降, 統計学やその関連分野であるデータサイエンス(機械学習, ニューラルネット等)は何回か注目されてきた.
- 主流にならない理由(私の意見)
  - ソフトウェア開発や手法として統計学を利用している.
  - 理論に基づいた利用ではなく, 操作手順に基づいた利用であるため, 応用がきかない.
    - 新たな発見を導くことが難しい.
    - 分析方法の前提などを理解していないのではないか.

16

イントロダクション

(1)統計データと集計データ
(2)統計データのグラフ表現
(3)統計学の歴史
(4)「統計」訳語論争
(5)記述統計学と推測統計学

17

(1)統計データと集計データ

- 公表統計のほとんどは集計値である.

統計データ

世帯番号	世帯主年齢	世帯人員
1	52	4
2	38	3
3	27	2
:	:	:

世帯番号2の世帯における統計データ

集計値/集計データ

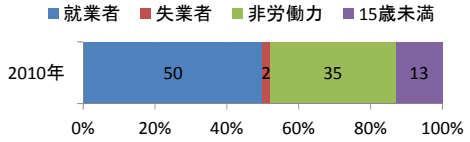
世帯主年齢＼人員	1人	2人
20～29歳	54	24
30～39歳	43	67
:	:	:
60～69歳	31	54
70歳以上	42	65

世帯主年齢が30～39歳の1人世帯に該当する世帯数(集計データ)

18

各種の統計数値(1/2)

- 日本の人口(2010年10月)
  - 1億2780万人 うち日本人 1億2618万人
- 日本が100人の村だったら(2010年労働力調査)
  - 就業者が50人，非労働力(高齢者や主婦など)が35人，15歳未満が13人，失業者が2人です．



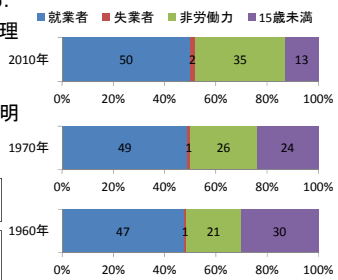
19

各種の統計数値(2/2)

- 統計数値で社会を語る．
  - 単純化した方が問題を理解しやすい．
  - AとBのどちらが問題を明確に表現しているか？

A 世界には，きれいで安全な水をのめない人がいます．

B 世界がもし100人の村だったら，17人の村人はきれいで安全な水を飲めません．



20

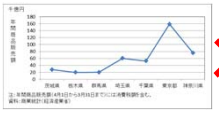
(2)統計データのグラフ表現(1/2)

- グラフの選択：棒グラフ or 折れ線グラフ
- 統計データの分類に基づいてグラフを選択して，作成しなければならない．
- 折れ線グラフでは，横軸には順序関係がないと線で結ぶことはできない．
  - 都道府県別人口は都道府県に順序がないため，折れ線グラフで表現しない．
  - 時間は順序が意味をもつ統計データのため，年別人口は折れ線グラフで表現する．

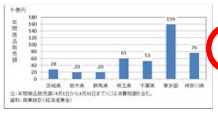
21

(2)統計データのグラフ表現(2/2)

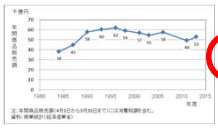
- 横軸：順序関係なし



- 横軸：順序関係なし



- 横軸：順序関係あり



22

(3)統計学の歴史

- 3つの源流
  - 政治算術学
    - ジョン・グラント 「死亡表」の収集：出生は男14，女13
  - ドイツ国勢学
    - 人口と土地に関する統計調査
    - 「Statistik」という名称が与えられる．
  - 古典確率論
- 3源流の統合(確率を含めた社会統計学)
- 数理統計学
  - 標本から母集団を推定する．

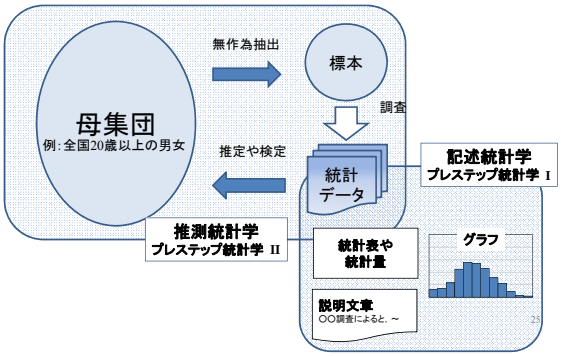
23

(4)「統計」訳語論争

- Statistics の訳語
  - スタチスチック：今井武夫
    - スタチスチックの役割は社会法則の発見にある．
  - 統計：森林太郎(森鷗外)
    - 統計の役割は事実を表現することにあり，因果関係を探求することにはない．
- Statisticsを「統計」と訳したことで，統計は方法論の1科目となり，他の専門分野と分離された．
  - ただし，この分類により，経済学や社会学，心理学，経営学，工学，医学などの実証分析をともなうさまざまな専門分野において，統計学は必要不可欠な科目となった．

24

(5) 記述統計学と推測統計学



1. 統計データの分類

1.1 概念：統計データ，変数，観測値
1.2 統計データの4つの尺度
- 2種類の統計データ：質的データと量的データ
- 4つの尺度への分類

1.3 表記方法
1.4 グラフ表現との関連

26

1.1 統計データ，変数，観測値

- データ（文字や数値など）を符号化（置き換えて区分に分類する）したものを統計データと呼ぶ．
  - 食事に関する感想は言葉や順位などさまざまな表現がある．→ 5段階の評価に置き換える．
    → $\{5, 4, 3, 2, 1\}$ の統計データとなる．
- 統計データは情報を符号化したデータの総称である．

27

変数と観測値の定義

- 変数：出席日数 x
- 観測値：番号2の出席日数 $x_2 = 13$
- 1つの観測値でも変数全体でも統計データ

i				x
番号	性別	生年	成績	出席日数
1	女	1994年	B	10
2	男	1993年	A	13
⋮	⋮	⋮	⋮	⋮

28

1.2 統計データの4つの尺度

- 和（＋）や差（－）が意味をもつ変数
  - 生年，出席日数 → 量的データ，量的変数
- 和（＋）や差（－）が意味をもたない変数
  - 性別，成績 → 質的データ，質的変数

番号	性別	生年	成績	出席日数
1	女	1994年	B	10
2	男	1993年	A	13
⋮	⋮	⋮	⋮	⋮

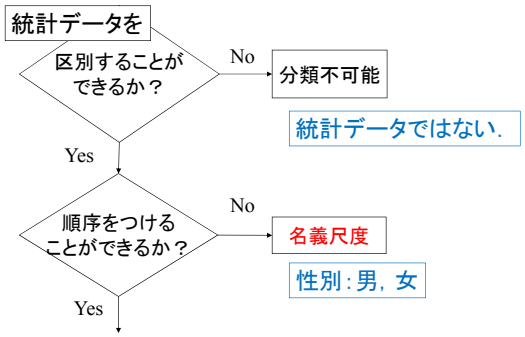
29

4つの尺度

- 名義尺度，順序尺度，間隔尺度，比率尺度
  - 統計データならば，4つの尺度のいずれかに分類される．
- 名義尺度と順序尺度は質的データ
- 間隔尺度と比率尺度は量的データ
  - 名義尺度：名前が意味をもつ
  - 順序尺度：順序が意味をもつ
  - 間隔尺度：値の間隔が意味をもつ
  - 比率尺度：値の比が意味をもつ
    - 比率尺度は比尺度や比例尺度ともいう．

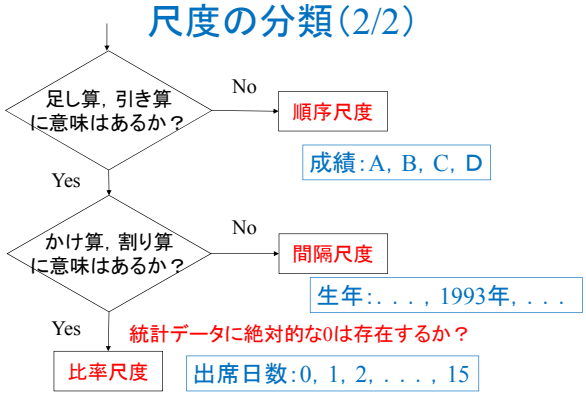
30

尺度の分類(1/2)



31

尺度の分類(2/2)



32

問題1

- つぎの変数の尺度は何か？
  ＜判断のコツは観測値を想定すること＞

(1) 満足度
- 満足； やや満足； やや不満足； 不満足

(2) 西暦
- 2013年； 2012年，．．．

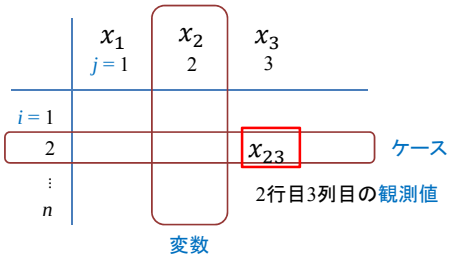
(3) 年間収入
- A: 200万円未満； 200～400万円未満；
- B: ○×万円

33

1.3 表記方法

統計データ x_{ij}　i：番号，　j：項目

記号はイタリック体(斜体)を用いる．



34

多変数データ，時系列データ，横断面データ

- 多変数データ
  2変数以上の変数から成る統計データ

x_{ij}
i は個人番号，$i=1,2,\ldots,n$,
j は項目，1：性別，2：年齢，3：年間収入

- 時系列データ
  時間を経るごとに観測した統計データ

x_t ：東京の日最低気温の平均値
t は年月を表す．$t=1$ は2008年1月．

- 横断面データ
  時間を固定して観測した統計データ

x_i ：年齢
i は個人番号を表す．$i=1,2,\ldots,n$.

35

問題2

- i 国の t 年におけるGDPを x_{it} とする．
  - i を固定して(ある国家に限定する)，1970年から2010年まで1年ごとの統計データを得た場合，〈　　〉データという．
  - t を固定して(2010年に限定する)，OECD加盟国の統計データを得た場合，〈　　〉データという．

x_{it}：i 国の t 年におけるGDP
i は国を，t は年を表す．$t=1$ は1970年．

36

1.4 グラフ表現との関連
棒グラフ，折れ線グラフ(1/2)

- グラフの種類による縦軸と横軸において，表現可能な尺度を限定される．
- 棒グラフ
  - 縦軸：間隔，比率； 横軸：名義，順序，(間隔，比率)
- 折れ線グラフ：横軸方向の関連性が意味をもつ
  - 縦軸：間隔，比率； 横軸：順序，(間隔，比率)
    - 例：名義尺度である都道府県を線でつないでも意味はない．

37

棒グラフ，折れ線グラフ(2/2)

- i 国の t 年におけるGDP x_{it} をグラフに表現する．
  - 縦軸にGDP，横軸を年
    - 棒グラフ，折れ線グラフの両方で表現可能
  - 縦軸にGDP，横軸に国
    - 棒グラフでは表現可能，折れ線グラフでは表現しない．

38

まとめ1

- 統計データは質的データと量的データに分けられる．
- 統計データの分類(名義尺度，順序尺度，間隔尺度，比率尺度)を明らかにしてから，集計や分析を行う．
  - なぜならば，統計データの分類は，集計やグラフ表現，統計値の計算の可否などさまざまな統計処理の前提になるからである．

39