



統計学



第1回 統計の基礎1

授業内容

- ▶ ガイダンス
- ▶ 資料の整理
- ▶ 度数分布表とヒストグラム
- ▶ 統計図表



ガイダンス

この授業の目的

- ▶ 統計学の意義を理解し、統計の活用・調査・研究の基本を理解する
- ▶ 看護のみならず、日常生活においても頻繁に使われる『統計』について、数学理論をなるべく使わずに、実践的に講義する



授業内容

- ▶ 第1部 統計の基礎(2回)
- ▶ 第2部 確率と確率分布(1回)
- ▶ 第3部 推定と検定(3回)
- ▶ 第4部 相関と回帰(1回)



教科書・参考書

▶ 教科書

- ▶ 涌井良幸・涌井貞美「文系のための統計学の教室」SBクリエイティブ

▶ 参考書

- ▶ 金森雅夫「系統看護学講座 基礎4 統計学」医学書院
- ▶ 田栗・藤越・柳井・ラオ「やさしい統計入門」講談社ブルーバックス



評価方法

- ▶ レポート(4種類): 55点
- ▶ 期末試験: 45点



注意事項

- ▶ 授業中に計算してもらうので、電卓(ルート計算が出来るもの)は必須
 - ▶ 試験でも必要
- ▶ その他、授業に関する事は、<http://www.f-edu.u-fukui.ac.jp/~c-matumo/>を参照



レポートの作成

- ▶ レポート1以外のレポートは、レポートは、一人で作成しても、グループで協力して作成しても良い
- ▶ ただし、きちんと「協力」すること
 - ▶ 他人が解いたものを写すのはNG
 - ▶ グループに所属する人を連名で書いておく
- ▶ レポートは、個人ごとに、レポート1～4をまとめて提出をお願いします
 - ▶ グループで作成した場合も、コピーの上、個人個人で提出すること



レポート1

- ▶ 統計関連の本(ハードカバーでもソフトカバーでも新書でも文庫でもよい)を読み、面白かったところをA4用紙1枚以内にまとめて提出しなさい
 - ▶ ちゃんと本を読むこと
 - ▶ 何の本を読んだか、どのように面白かったかを具体的に書くこと
- ▶ 締切:「統計学」試験終了時
 - ▶ どんな本を読めばよいか迷う場合には、『参考文献の本棚』(<http://booklog.jp/users/cxi8912>)を参考にしてください



教員情報

- ▶ 名前：松本智恵子
- ▶ 所属：福井大学教育学部
- ▶ 専門分野：数理統計学



この授業におけるルール

- ▶ 私語をしない
- ▶ レポートの提出日を厳守する





資料の整理

統計とは？

- ▶ 「統計」は現在生活のあちこちで使われている
 - ▶ 新聞に載っている世論調査
 - ▶ 雑誌やテレビでも数値を使ったニュースがある
 - ▶ 看護の現場にも、患者の病状を知る為の様々な数値が存在する
- ▶ でも、統計って何？
 - ▶ 計算？
 - ▶ 数値？
 - ▶ どちらにしる面倒そうだなあ.....



統計学の種類

- ▶ 記述統計学(データの整理)
 - ▶ 統計図表
 - ▶ 統計量
 - ▶ 数理統計学(データの解析)
 - ▶ 推測統計学(母集団のパラメータ(母数)に着目)
 - ▶ 頻度論的統計学(従来の統計学)
 - 推定、検定
 - ▶ ベイズ論的統計学(経験を生かす)
 - ベイズ推定、ベイズ検定
 - ▶ 多変量解析(データに着目)
 - ▶ 回帰分析
 - ▶ 数量化理論
 - ▶ Etc...
-

記述統計と推測統計

記述統計

▶ 内閣支持率の調査

- ▶ RDD調査によると、「支持する」と答えた人の割合は、前回の34.5%から29.1%になった
- ▶ これだけの情報から、「内閣支持率は下落した」といえるか？
 - ▶ 5%以上下がっているのだから、明らか？
 - ▶ 有権者全員に調査しているわけではない
 - その所為で下落しているようにみえるだけかもしれない

推測統計



「統計調査」を図で表すと



- ▶ 記述統計：得られたデータの特徴を調べる
 - ▶ 推測統計：得られたデータから推測する
-



統計学（統計科学）とは

- ▶ いくつかの対象について、観測、調査、実験などを行って得た結果を、数字として表したものを「データ」という
 - ▶ **（不確実性を含む）** データから必要な情報を引き出すことを「統計データ解析」といい、そのための方法を「統計データ解析法」という
 - ▶ 「統計学」とは、統計的方法の体系化ならびに統計的概念の本質に関する研究を課題にしている学問分野である
-



データに「不確実性」が生じる理由

- ▶ 本質的に備わっているため(人間のやること)
- ▶ 情報の欠如のため
- ▶ 入手可能な情報の不正確さのため
- ▶ 必要な測定を行うことができないため
- ▶ 方法論として、ランダム性(無作為抽出)を利用するため

「統計学」とは、**不確実性**を考慮して
「間違う度合い」を数値化する学問



不確実性と間違ら度合

- ▶ 統計: 仮定を立てた上で推測を行う
- ▶ 「正しい統計的推測」や「正しい統計モデル・事前分布」は存在しない
 - ▶ 統計モデルの決定や統計的推測を行うために必要な「真の分布(データのふるまい)」は不明である
- ▶ データの「真の」ふるまい(真の分布)が分からなくても、推測法を固定すれば、得られた結論がどのくらい間違っているか(誤差)を数量的に示すことができる



「バラツキ」を調べる

- ▶ 集めたデータが全て同じ値であれば、その値を知れば分析は終わり
- ▶ でも、実際のデータは色々な値を取っている
 - ▶ 身長、テストの点、居住地、好き嫌い
- ▶ この「色々な値を取っている」ことを「バラツキ」という
- ▶ データ分析では「バラツキ」を調べていく

不確実性＝バラツキ



統計学で何ができるか？

- ▶ 「統計的推定」ができれば
 - ▶ 標本調査より母集団の平均や分散がどのくらいかが分かる
- ▶ 「統計的検定」ができれば
 - ▶ 標本調査より、仮定が正しいかどうかをチェックできる
- ▶ 「回帰分析」ができれば
 - ▶ 複数の項目があるデータを用いて、ある変数から他の変数を予測することができる
- ▶ 「数量化理論」ができれば
 - ▶ データが数値じゃなくっても多変量解析ができる

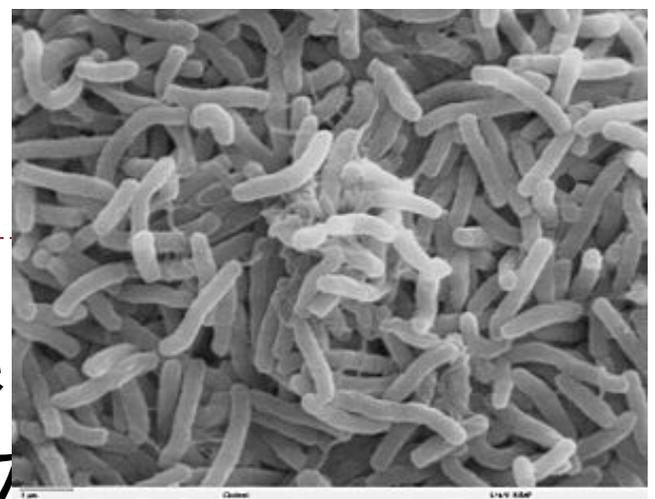


統計学は最強の武器になりうる

- ▶ ちゃんと使いさえすれば、どんな分野の議論においても、データを集めて分析することで、最速で最善の答えを出すことができる
 - ▶ 「理由」や「原因」を突き止めることはできない
 - ▶ でも、きちんと仮説を立てて調査して分析すれば、解決策を示すことは可能



例



▶ 19世紀 ロンドン

▶ 急激な都市化、工業化が原因で むようになり、伝染病で多くの人が

▶ 伝染病の原因が分からなかった

- ▶ 細菌が病気の原因であることが分かったのは19世紀末(コッホやパスツールの時代)
- ▶ 瘴気という「ウイルスを含んだ悪臭」によって病気になるという説が支持を得ていた
- ▶ 「人を介してうつる」という伝染説もあったが、伝染性が弱い(接近しないとうつらない)病気では中々感染が証明できなかった

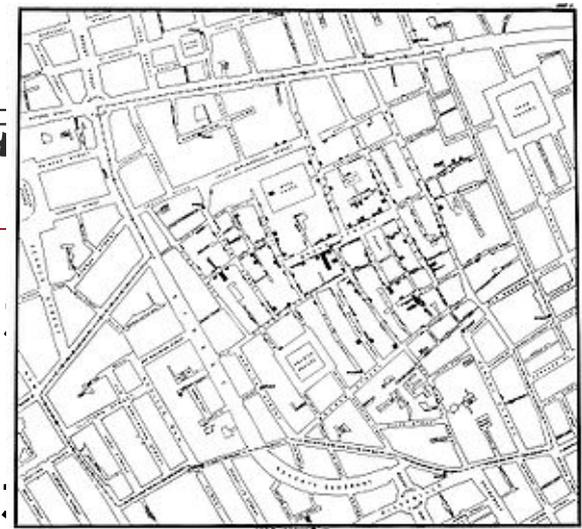
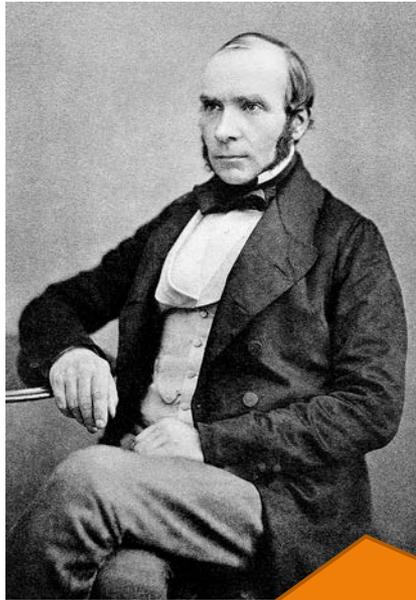
▶ 1854年8月末～9月にかけて、ロンドン、ソー ホーにあるブロード・ストリートでコレラが発生した

▶ コレラ:コレラ菌によって発症する病気、

- ▶ 急激な下痢による脱水症状を起こす→電解質、特にカリウムが失われて心臓の電気インパルスが弱まり、死に至る
- ▶ 糞口感染



「疫学の父」 ジョン・スノウの



- ▶ 同じような状態でコレラに罹った人の違いを調べ、
- ▶ 仮説を得てから大々的に発症／非発症と関連していると考えられる「違い」がどれくらい確からしいか検討する

ある水道の水を飲んでいて137人のうち80人がコレラにかかった。

空気 水の所為？

→同 →消化器系が最初にダメージをくらう

空 同じ水道の水を飲んでいる人がコレラに罹っている

飲んで
人は20

9月8日に、原因となっているっぽいポンプの柄を取り外す
→コレラの発症が止まる

比較をして「違い」を見つける

▶ 対照実験

- ▶ 他の条件を全て同じにして実験を行い、結果を比較する
- ▶ 何度か実験をして、その平均値を比較することが多い

▶ 問題:どのくらいの差があれば「違いがある」といえるのか？

- ▶ 「感覚的」な基準ではなく、「客観的」な基準が必要



「違い」を比較する例

- ▶ さいころを1000回振ったとき、出た目の平均は3.319だった
 - ▶ ゆがみのない、正確なさいころでの平均は3.5なのだが、このさいころは正確ではないのだろうか？
 - ▶ 0.2くらいずれているだけだから、まあ正確？
 - ▶ 基準が感覚的だと、判断しづらい
 - ▶ 標準偏差も加味して、正確かどうかを考える
-



数学と統計の本質的な違い

▶ 数学

- ▶ 定義から定理や計算を行う
- ▶ 100%正しい

▶ 統計

- ▶ データから推定する
- ▶ 誤差を含んでいる
- ▶ ……の割に誤差の話は省略した結果ばかりでてくる
 - ▶ ~を行うと痩せる
 - ▶ ~をやっている人は一だ
- ▶ しかも、「断定調」な統計結果の方が面白がられる
- ▶ 数学の理論を使っている



「統計学」の簡単な歴史

- ▶ 古代：記述統計（徴税、軍役の為の人口調査）
- ▶ 17世紀以降
 - ▶ 観測された数値から法則を見つける
 - ▶ 天文（ケプラーの第3法則）
 - ▶ 物理（ガリレオの落下の法則）
 - ▶ 生命科学（ハーヴェーの血液循環説）
 - ▶ 数を基礎とした新しい「国政術（シュタティスティク）」
 - ▶ 人口表（グラント）、国勢調査
 - ▶ 確率論（パスカル）



データを元にして推論する（推測統計学）



昔の統計学、今の統計学

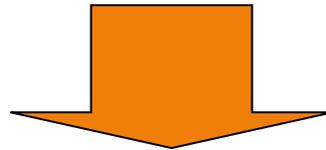
- ▶ 統計解析にはデータが必要
- ▶ 昔は、コンピュータがなかったので、大きなデータは扱えなかった
 - ▶ データをまとめるのも手作業
 - ▶ 計算も手作業
- ▶ しかし現在では、ITという強力なパートナーを得て、全ての学問分野を横断し、あらゆる問題に最善の答えを出せる環境になっている



統計の「嘘」

「嘘には3種類ある。嘘、見え透いた嘘、そして統計」
There are three kinds of lies: lies, damned lies,
and statistics. (ディズレーリ)

統計を使えば、幾らでも「騙せる」。
騙されてても分からない。

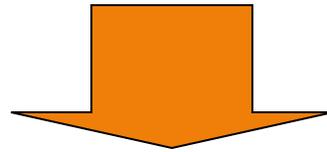


統計を「正しく」利用、使用しなければならない



統計の限界

データを用いて、様々な問題を解決しようとするのが「統計」



データが無いと何もできない.....



統計の誤差と「許容度」

どんな種類の統計にも、殆ど例外無しに「誤差」がある

なるべくなら誤差の小さい統計が望ましい

誤差を小さくするためには、より多くの費用や労力や時間が必要

統計の目的、作成に関わる制約、データ獲得の迅速性などを考慮して、許容される誤差がどの程度までであるかをつかむことが大切(誤差が大きすぎると役に立たないし、必要以上に小さすぎるのは資源の無駄遣い)



集団の観察

- ▶ 「統計」を扱う時にはいつも、何らかの「集団」を考察している
 - ▶ 統計は、集団の特徴を数量的に表す役割を持つ
 - ▶ 考察される集団は、一定の共通した目印を持つ個体の集合として構成されている



データとは？

- ▶ 数値化された情報の集合
- ▶ 基礎的な事実や資料をさす言葉
 - ▶ 情報処理や考察によって付加価値を与える前提で集められている
 - ▶ 基本的に、複数個の事象や数値の集合



データの構成要素

- ▶ **個体**: 統計データを構成する単位
 - ▶ 個人、学級、学年、学校、自治体、市町村、都道府県、国、etc...
 - ▶ **変量(変数)**: 個体によってその値が変化するもの
 - ▶ 身長、体重、血液型、etc...
 - ▶ **値**: それぞれの変数を取る具体的な数値
 - ▶ 血液型: A, B, O, AB
 - ▶ 身長: 様々な数値
- まとめた表: 個票



統計データの色々

- ▶ 時系列データ: 時間的な流れの中でデータの収集が行われ、その時間に添った形で整理されたデータ
 - ▶ 毎日の入場者数、おこづかいの収支
 - ▶ ある患者の血圧を毎日測る
- ▶ クロスセクションデータ(切断データ、横断型データ): ある時点を固定して整理されたデータ
 - ▶ 入場者の属性、おこづかいを何に使ったか
 - ▶ ある日の来院者の血圧



質的データと量的データ

- ▶ 量的データ(数量データ): 数値で表される、量的な尺度で計測されたデータ(物差しとか計りとかで計測されたデータ)
 - ▶ 連続データ: 連続的な実数値を取る
 - ▶ 離散データ: 飛び飛びの値(整数値)しか取らない
- ▶ 質的データ(分類データ): カテゴリで表される定性的なデータ、類別する為のデータ、カテゴリデータ



何が違うか？

- ▶ 平均やバラツキの尺度の求め方
 - ▶ 量的データ: 簡単に計算できる
 - ▶ 質的データ: 計算できる?
 - ▶ 適当な数値に置き換える(数量化理論)
- ▶ 集計
 - ▶ 量的データ: 自分でグループを作る必要
 - ▶ 血圧: 低血圧グループ、普通の血圧グループ、高血圧グループ.....
 - ▶ 質的データ: 元々グループになっている



連続型データの分析のポイント

要素	指標
中心	平均値, 中央値, 最頻値 等
広がり	範囲, 四分位範囲, 分散, 標準偏差, 外れ値 等
形状	分布の形状(左右対称, 歪んでいる, 峰が複数等), 歪度 等
密度	度数, 相対度数, 累積相対度数, 四分位数, 外れ値 等

データの尺度

- ▶ 名義尺度：区別だけを行う尺度
 - ▶ 性別、学部、天気
- ▶ 順序尺度：区別に順序が付く尺度
 - ▶ 好き嫌い、統計検定の級
- ▶ 間隔尺度：差（間隔）が意味を持つ尺度
 - ▶ 気温
- ▶ 比例尺度：差と比が意味を持つ尺度
 - ▶ 重さ、長さ



実際の分析における尺度の取り扱い

- ▶ 一般に、名義尺度と順序尺度で計測されたデータは質的データ、間隔尺度と比例尺度で計測されたデータは量的データ
- ▶ 但し、順序尺度データを間隔尺度データとみなすばあいがある
 - ▶ 好き嫌い: 普通の感覚では「好き」と「普通」と「嫌い」との間には等間隔でない差があるけど、それを等間隔の差とみなす



度数分布表とヒストグラム

教科書

▶ P40～47



女子大生の身長

151	154	158	162
154	152	151	167
160	161	155	159
160	160	155	153
163	160	165	146
156	153	165	156
158	155	154	160
156	163	148	151
154	160	169	151
160	159	158	157
154	164	146	151
162	158	166	156
156	150	161	166
162	155	143	159
157	157	156	157
162	161	156	156
162	168	149	159
169	162	162	156
150	153	159	156
162	154	164	161

量的データ



このデータから何が分かるか？

- ▶ 女子大生の身長は、皆同じではなく、ばらばらな数値を取っている
 - ▶ 統計の専門用語で「分布する」という
 - ▶ 分布が生じるのは、その数値が決まる背後に何らかの「不確実性」が働いているから
- ▶ 生データの数値をじっと睨んでいるだけでは、何がなんだかさっぱり



縮約(データとして並んでいるたくさんの数値を、何らかの基準で整理整頓して、意味のある情報を抽出する)



度数分布表とは何か？

自分でグループを作る

- ▶ データを5～15程度のグループに分け、データの分布傾向を把握するための表
- ▶ 階級、階級値、度数、相対度数、累積度数、累積相対度数からなる表

階級	階級値	度数	相対度数	累積度数
141～145	143	1	0.0125	1
146～150	148	6	0.075	7
151～155	153	19	0.2375	26
156～160	158	30	0.375	56
161～165	163	18	0.225	74
166～170	168	6	0.075	80

度数分布表における用語集

- ▶ 階級：データの範囲をいくつかの区間（小グループ）に分けた、その区間
 - ▶ 度数：階級の中に入っているデータの個数
 - ▶ 相対度数：データ数を1とした時の、階級の中に入っているデータの個数を割合で表記したもの
 - ▶ データの個数が異なる二集団を比較する際に役に立つ（ex.外来患者と入院患者の比較）
 - ▶ 累積度数：度数を小さい方から次々と足していったもの
 - ▶ 累積相対度数：累積度数を相対度数化したもの
 - ▶ 階級値：階級を代表する値
 - ▶ 階級幅：階級の幅
-



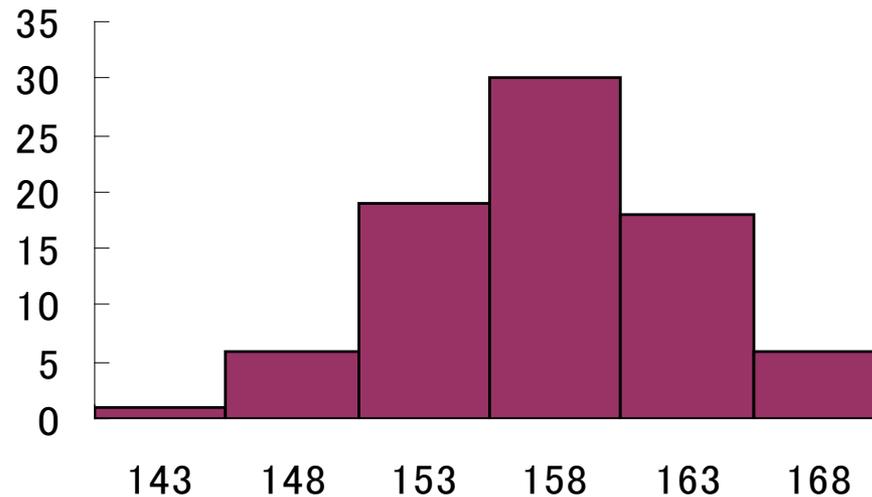
度数分布表の作り方

- ▶ データの中から、最も大きな数値(最大値)と最も小さな数値(最小値)を見つける
 - ▶ 範囲が最大値から最小値になるような区切りのいい範囲を作り、その範囲を5~8程度の小範囲(階級)に切る
 - ▶ 各階級を代表する値「階級値」を決める(一般には、階級の真ん中の値が選ばれる)
 - ▶ 各階級に入るデータ数「度数」を数える
 - ▶ 各階級の度数の、全体の中に占める割合を計算する(相対度数、全部足すと1になる)
 - ▶ その階級までの度数を合計したものを計算する(累積度数)
-



ヒストグラムとは何か？

- ▶ データを、その数値の大きさに分類し、分類グループの中に入っているデータの数(度数)を数え、度数の面積で棒グラフを作る



ヒストグラムの作り方

- ▶ 数直線を描く(横軸)
- ▶ 横軸に階級を刻む
- ▶ 各階級の上に棒をのぼす
 - ▶ 棒の面積が階級の度数になるように(相対度数でも可)

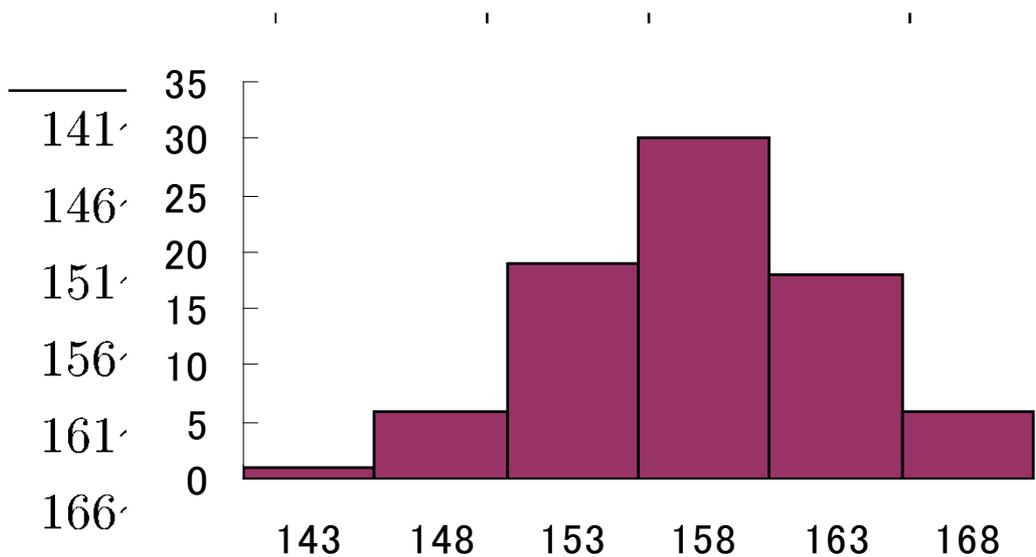


ヒストグラムの種類

- ▶ データ内の数値には、色々な種類がある
 - ▶ 量的データ
 - ▶ 質的データ
- ▶ 量的データには、飛び飛びの値を取るもの（離散量）と連続した値を取るもの（連続量）がある
- ▶ 「ヒストグラムっぽいヒストグラム」が描けるのは連続型数量データ



女子大生の身長

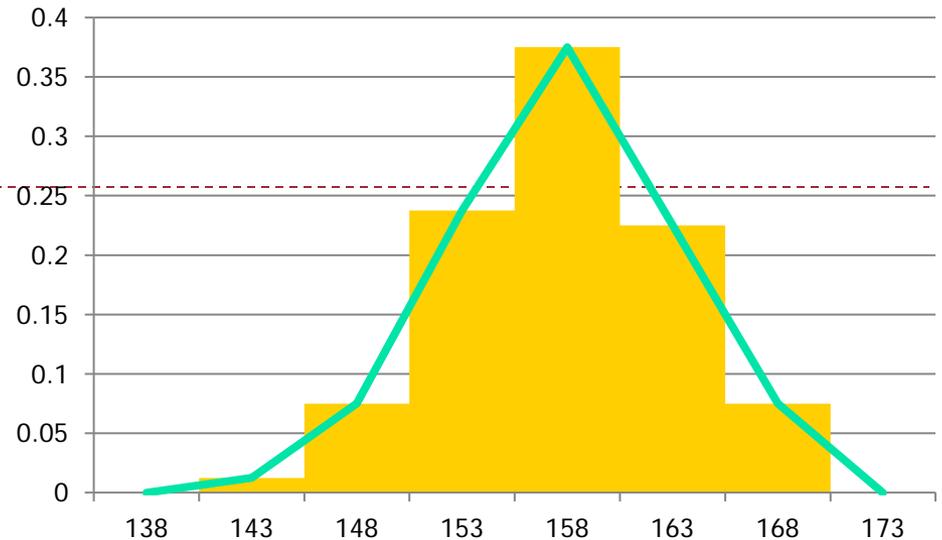


151	154	158	162	
154	152	151	167	
160	161	155	159	
160	160	155	153	
163	160	165	146	
156	153	165	156	
8	155	154	160	
1	6	163	148	151
7	4	160	169	151
26	0	159	158	157
56	4	164	146	151
74	2	158	166	156
80	6	150	161	166
102	155	143	159	
157	157	156	157	
162	161	156	156	
162	168	149	159	
169	162	162	156	
150	153	159	156	
162	154	164	161	



度数折れ線

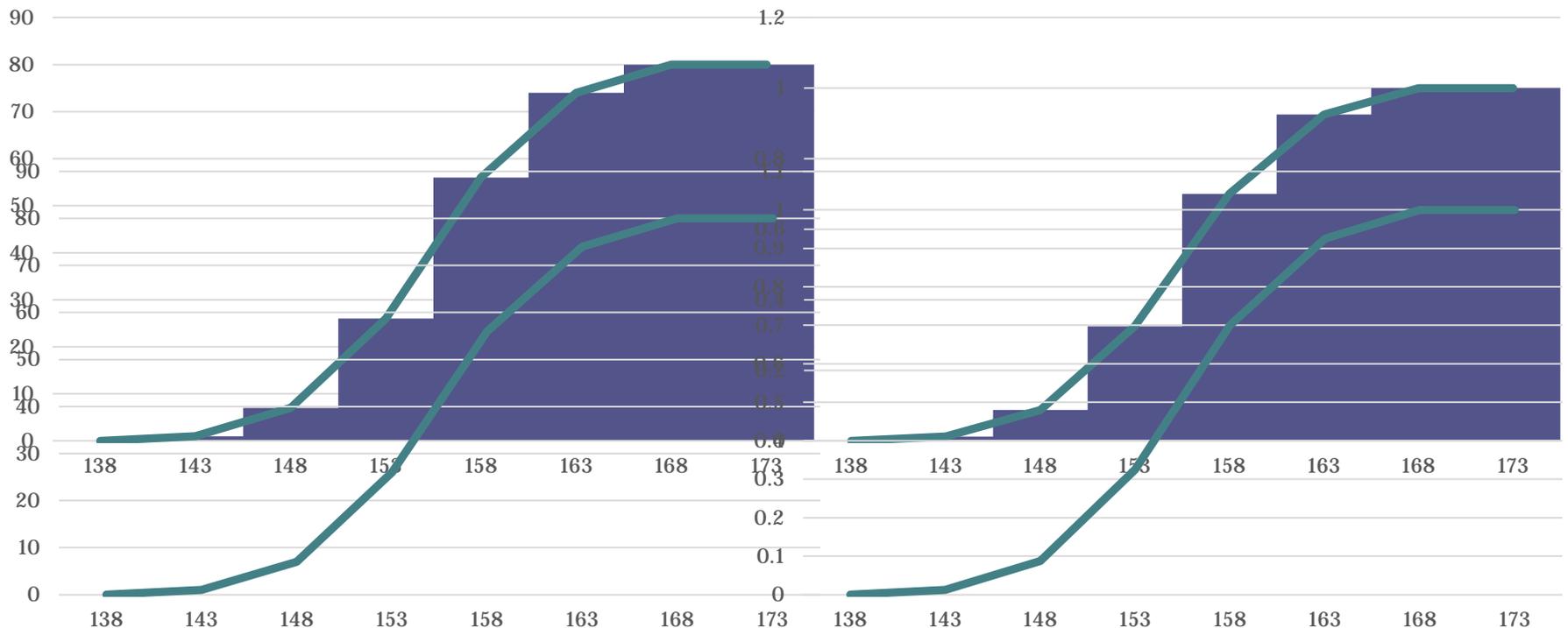
棒の面積(=度数)と折れ線で囲まれた部分の面積は同じ



- ▶ ヒストグラムのそれぞれの長方形の上辺の中点を取り、結んだもの
- ▶ 但し、左端は一つ手前の階級の度数を0とした場所、右端は一つ先の階級の度数を0として結ぶ
- ▶ 縦軸を相対度数とするヒストグラムでは、このような折れ線を相対度数折れ線という(確率・確率分布を考える時に出てくる)
- ▶ 連続データの場合、階級幅を縮めると、度数折れ線は曲線になる(度数分布曲線、相対度数分布曲線)

累積度数と累積相対度数

- ▶ 累積度数: その階級までの度数を足し合わせたもの
- ▶ 累積相対度数: 累積度数を総度数で割ったもの、あるいは相対度数の累積度数



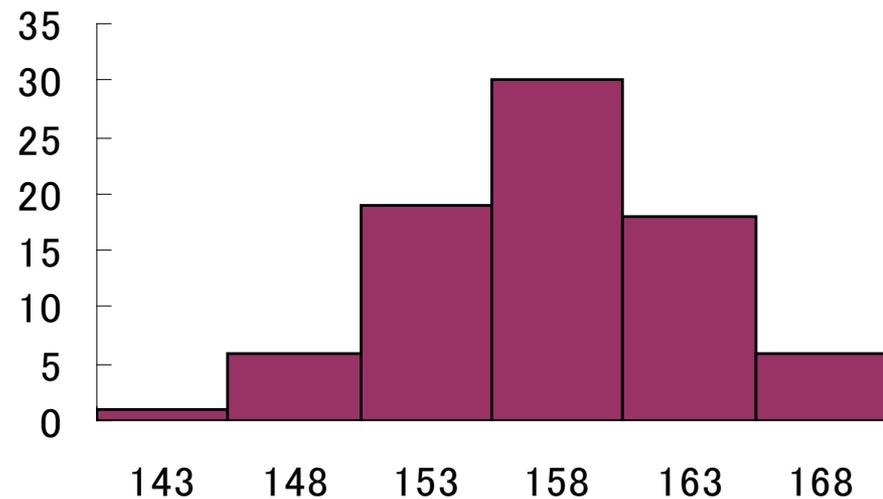
パソコンによる度数分布表とヒストグラム作成

- ▶ EXCELを使えば(一応)描ける
 - ▶ FREQUENCY関数(ある値以下であるデータの個数を表示)を使う
 - ▶ 分析ツール(データ分析)を使う
 - ▶ 階級は決めておく必要がある
 - ▶ グラフの種類は縦棒、オプションで「棒の間隔」を0にすればよい
 - ▶ EXCEL2016のグラフ作成機能の中にはヒストグラムが入っている
- ▶ SimpleHist (<http://www.cc.miyazaki-u.ac.jp/yfujii/histogram/>)を用いる



ヒストグラムから何が分かるか？

- ▶ 単峰性
- ▶ 対称性
- ▶ 外れ値の有無



ヒストグラムのチェックポイント

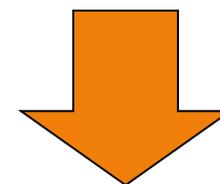
- ▶ 単峰(山が一つ)か、複峰(山が2つ以上)か？
 - ▶ 山が2つ以上→二つ以上のデータが混ざっている？
- ▶ 左右対称か？ どちらかに偏ってないか？
 - ▶ 左右対称→平均値、中央値、最頻値が同じ階級にある
 - ▶ 左右非対称→平均値、中央値、最頻値が異なる階級にある
- ▶ ぽつんと一つだけ離れたところにある棒はないか？
 - ▶ 外れ値の可能性
- ▶ その他、「変な山型」になってないかどうか
 - ▶ 平たい、のっぺり(分散・標準偏差の違い)
 - ▶ 観察や調査の打ち切り、不良品の切り捨てなどにより、変な形のヒストグラムが出てくる場合もある



世帯収入の分布のヒストグラム

金融資産保有額（階級）	階級値	世帯数（度数）
～100 万円未満	50 万円	229
100～200 万円未満	150 万円	186
200～300 万円未満	250 万円	160
300～400 万円未満	350 万円	147
400～500 万円未満	450 万円	108
500～700 万円未満	600 万円	250
700～1000 万円未満	850 万円	245
1000～1500 万円未満	1250 万円	291
1500～2000 万円未満	1750 万円	188
2000～3000 万円未満	2500 万円	215
3000 万円以上	4000 万円	306
合計		2325

幾つかの、級間隔の
異なった級が混合し
ている

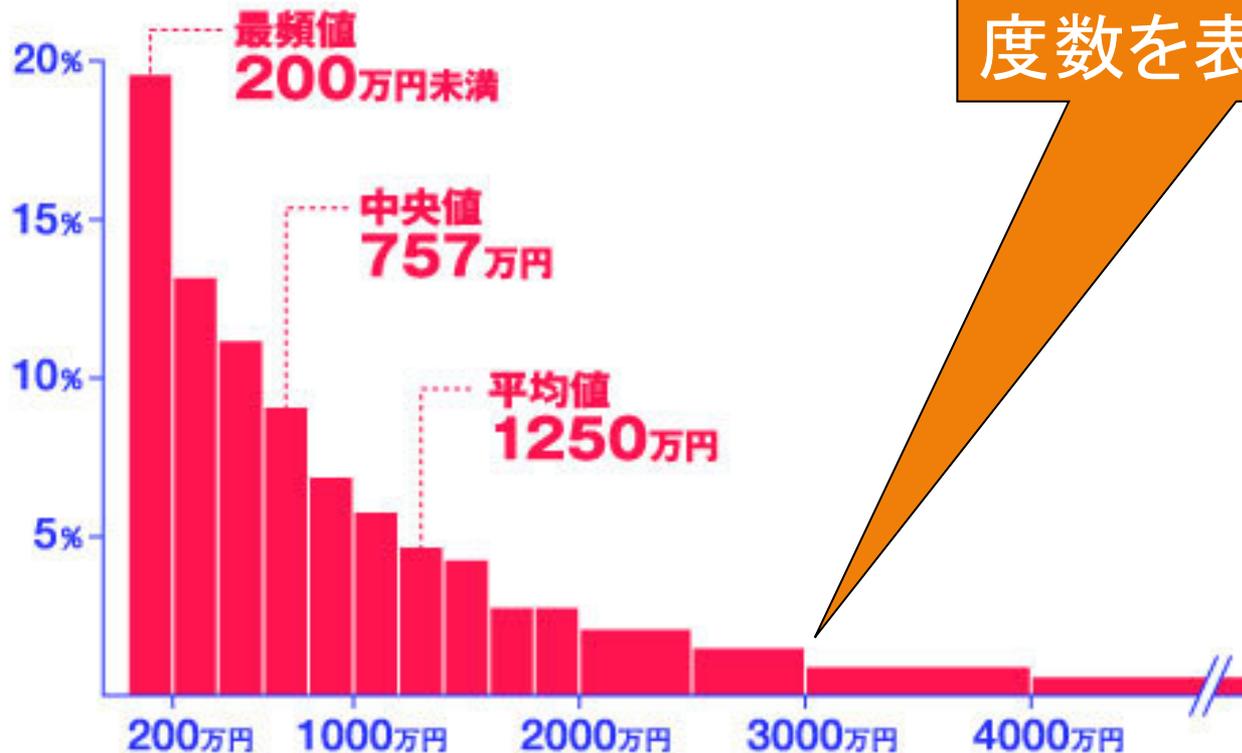


ヒストグラムの各柱の
面積が度数に比例す
るように書けばよい。



階級の間隔が異なる場合の ヒストグラム 1

勤労者世帯の貯蓄額の



出典:総務省「家計調査」勤労者世帯の貯蓄現在高階級別世帯分布(2008年)

階級の間隔が異なる場合の ヒストグラム2

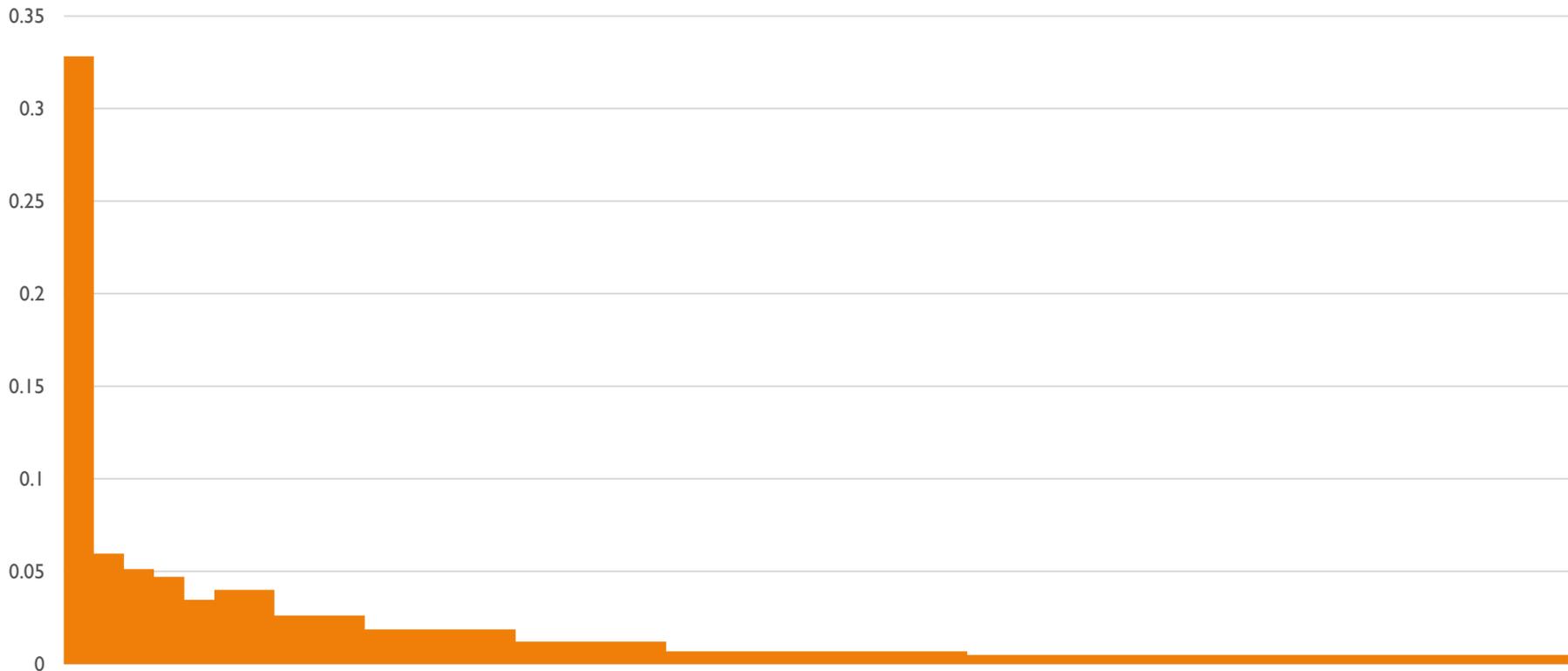
	A	B	C	D	E	F	G	H	I	J	K	L
1	金融資産保有額			度数	相対度数	累積相対度数						
2	万円以上		万円未満									
3	0 ~		100	1024	0.328205	0.328205						
4	100 ~		200	186	0.059615	0.387821						
5	200 ~		300	160	0.051282	0.439103						
6	300 ~		400	147	0.047115	0.486218						
7	400 ~		500	108	0.034615	0.520833						
8	500 ~		700	250	0.080128	0.600962	125	0.040064		2		
9	700 ~		1000	245	0.078526	0.679487	82	0.026175		3		
10	1000 ~		1500	291	0.093269	0.772756	58	0.018654		5		
11	1500 ~		2000	188	0.060256	0.833013	38	0.012051		5		
12	2000 ~		3000	215	0.06891	0.901923	22	0.006891		10		
13	3000 ~			306	0.098077	1	15	0.004904		20		
14	合計			3120	1							

階級の幅が200万
→100万の2倍だから、
度数、相対度数を1/2すれば
うまく表せるのでは？

階級の幅が基準の何倍か？



階級の間隔が異なる場合の ヒストグラム 3



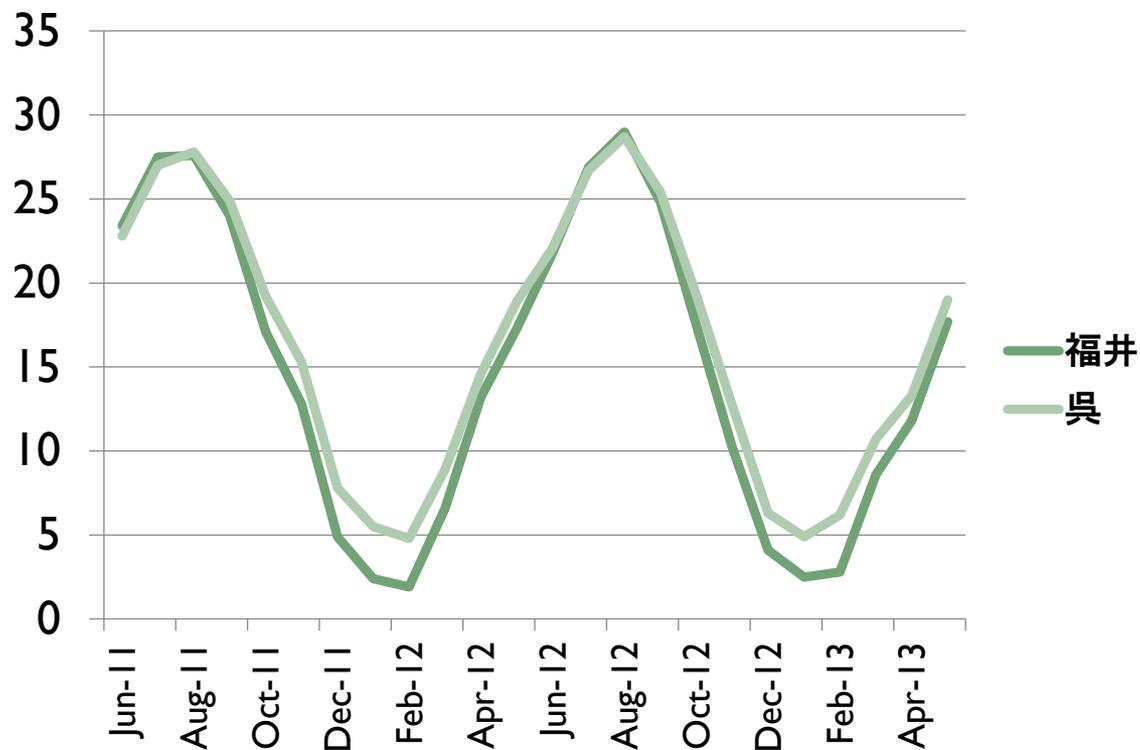
統計図表とは

- ▶ 統計調査によって得られたデータを、分かりやすいように表記する為の図や表
- ▶ 「グラフ」は「図」と言い表すことが多い
- ▶ 適切で見やすく、分かりやすい統計図表を作成することが重要
- ▶ データに合った図を描く



例

平均気温		
年月	福井	呉
Jun-11	23.4	22.8
Jul-11	27.5	27
Aug-11	27.6	27.8
Sep-11	24	24.9
Oct-11	17.1	19.2
Nov-11	12.8	15.3
Dec-11	4.9	7.8
Jan-12	2.4	5.5
Feb-12	1.9	4.8
Mar-12	6.6	8.9
Apr-12	13.2	14.6
May-12	17.3	18.9
Jun-12	21.8	22.1
Jul-12	26.9	26.7
Aug-12	29	28.7
Sep-12	24.8	25.4
Oct-12	17.5	19.3
Nov-12	10.2	12.7
Dec-12	4.1	6.3
Jan-13	2.5	4.9
Feb-13	2.8	6.2
Mar-13	8.6	10.7
Apr-13	11.8	13.3
May-13	17.7	19



統計図表の提唱者 1

- ▶ 昔の統計データ: 生データor集計データ
 - ▶ 数字の方が信頼されていた
 - ▶ 写本の時代: 誤記、省略、写す人の癖
 - ▶ 印刷技術の発展→銅版画を中心に図版の精度の向上
 - ▶ 数字ばかりだと見にくい(例: Petty「政治算術」)
- ▶ 自然科学の発達に伴い、「定量化」と「視覚化」が志向される(17世紀)
 - ▶ デカルト座標
 - ▶ ランベルトのグラフ
 - ▶ 光の強さ、地球の地表面の気温、磁気変動などを折れ線グラフで表し、物理学を数学と同じように分析
 - ▶ 社会・経済現象は対象外



統計図表の提唱者 2

- ▶ ウィリアム・プレイフェア：金銭や人口、面積などといった社会・経済現象を棒グラフや折れ線グラフなどの「図」で表した
 - ▶ 1786年「商業と政治のアトラス」
 - ▶ 棒グラフはこの書物で初めて書かれた
 - ▶ プレイフェアは、乗法を視覚化し理解を促進するインフォメーション・デザインにとどまらず、意図した印象を醸成するイメージ・デザインとしての図も開発している



プレイフェアのグラフ

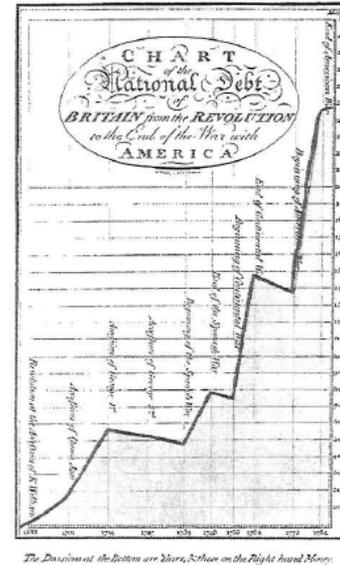
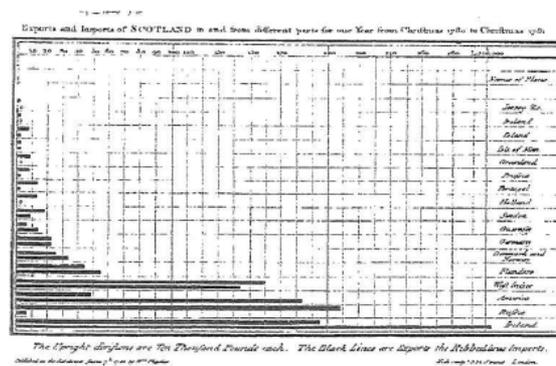
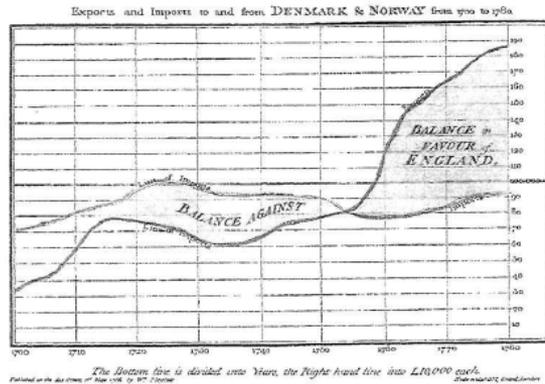
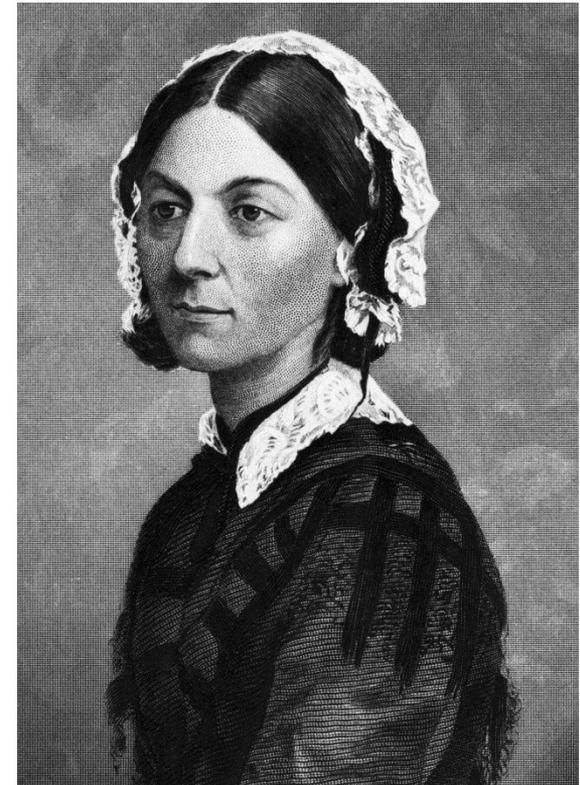


図1 『アトラス』(初版)のチャートイギリスと海外諸国との輸出入額の推移と入超の多寡が、所定のカラーリングで示されている(左上)。スコットランドの過去のデータがないので、やむを得ず開発した世界初の「棒グラフ」(左下)。イギリス国債の推移を政治上の出来事と関連させたチャート(右上)。

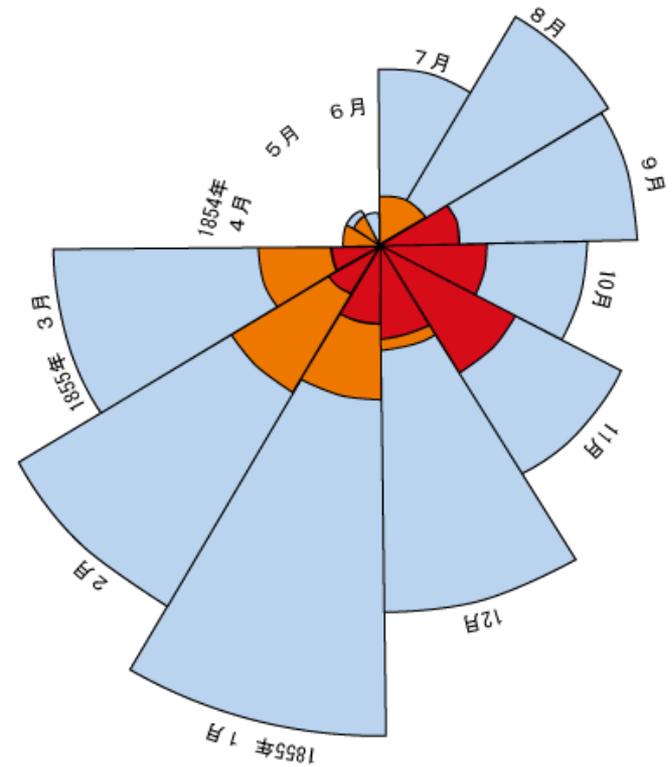
統計図表の提唱者 3

- ▶ Florence Nightingale: 生データを、分かりやすく分類して報告に利用
- ▶ 公衆衛生の必要性を訴える為に、Nightingaleは統計図表を利用した



Nightingaleのグラフ

- 赤：戦場での負傷により亡くなった兵士の数
- 青：病気（感染症）で亡くなった兵士の数
- 橙：その他の原因で亡くなった兵士の数



統計図表の種類

- ▶ 表
- ▶ 棒グラフ、折れ線グラフ、円グラフ
- ▶ ヒストグラム
- ▶ 散布図
- ▶ 箱ひげ図
- ▶ ボックスプロット、デンドログラム(樹上図)
- ▶ Etc...



表の形式

全国世帯の1ヵ月平均消費支出，特に食料費の内訳（人口5万以上の都市）
 ——昭和47年～昭和51年——

(4)
単位金額 円

	消費支出 総 額	食 料 費				
		総 額	主 食	副 食	嗜好食品*	外 食
昭和47年	96,026	31,357	4,629	16,497	6,962	3,269
48	112,116	35,789	4,995	19,333	7,630	3,831
(8)→49	136,024	44,350	6,062	24,252	9,381	4,655
50	157,982	50,479	7,081	27,344	10,621	5,433
51	174,790	55,308	7,868	30,054	11,188	6,198

全国世帯とは、農林漁業を営む世帯，単身者世帯を除く一般世帯をいう。なお，年次は年平均を示す。

(注)* 菓子類，果物類，酒類，飲料を示す。

(5)→(資料) 総理府統計局「家計調査報告」

- ▶ 表題
- ▶ 表頭
- ▶ 表側
- ▶ 計測単位の表示
- ▶ 脚注



適切な表を作成する為に

- ▶ 表頭と表側にどの項目をどのように置くか？
- ▶ 計測単位の表示(どこに書くか？)
- ▶ 脚注を表示する時には分かりやすく
- ▶ 数字の表し方(どの桁まで出すか？)
- ▶ 報告書などのように、表がたくさん出てくるときには、その大部分に共通する注意は一ヶ所にまとめて凡例として出した方がすっきりする



適切なグラフを作成する

- ▶ グラフ表現は、データを様々な角度で観察したり、多種多様な方法で説明や自分の理解を裏付けたり、理解を促したりするためのもの
- ▶ 説明の為のグラフを作る場合は、自分が表現したい内容と、グラフを見る人が感じる印象が大きく異ならないように作成する必要がある



グラフ作成の前に

- ▶ データの尺度をチェック
 - ▶ 比尺度: 0が重要だから、グラフでも0を入れる
 - ▶ 間隔尺度: 0が重要ではなくなるので、棒グラフより折れ線グラフの方が適切(例: 気温)
 - ▶ 名義尺度、順序尺度: 集計の必要
- ▶ 時系列データか横断型データなのかチェック
 - ▶ 時系列: 折れ線グラフを用いるとより良く推移が分かる
- ▶ 項目の並びをチェック



グラフ作成の手順

- ▶ グラフ作成の目的を明確にする
- ▶ グラフ作成の為にデータを加工する
 - ▶ 集計
 - ▶ 円グラフ用に%で表す
 - ▶ ヒストグラム用に度数分布表を作る
- ▶ グラフ構成案を考える
- ▶ グラフのデザインを決定する



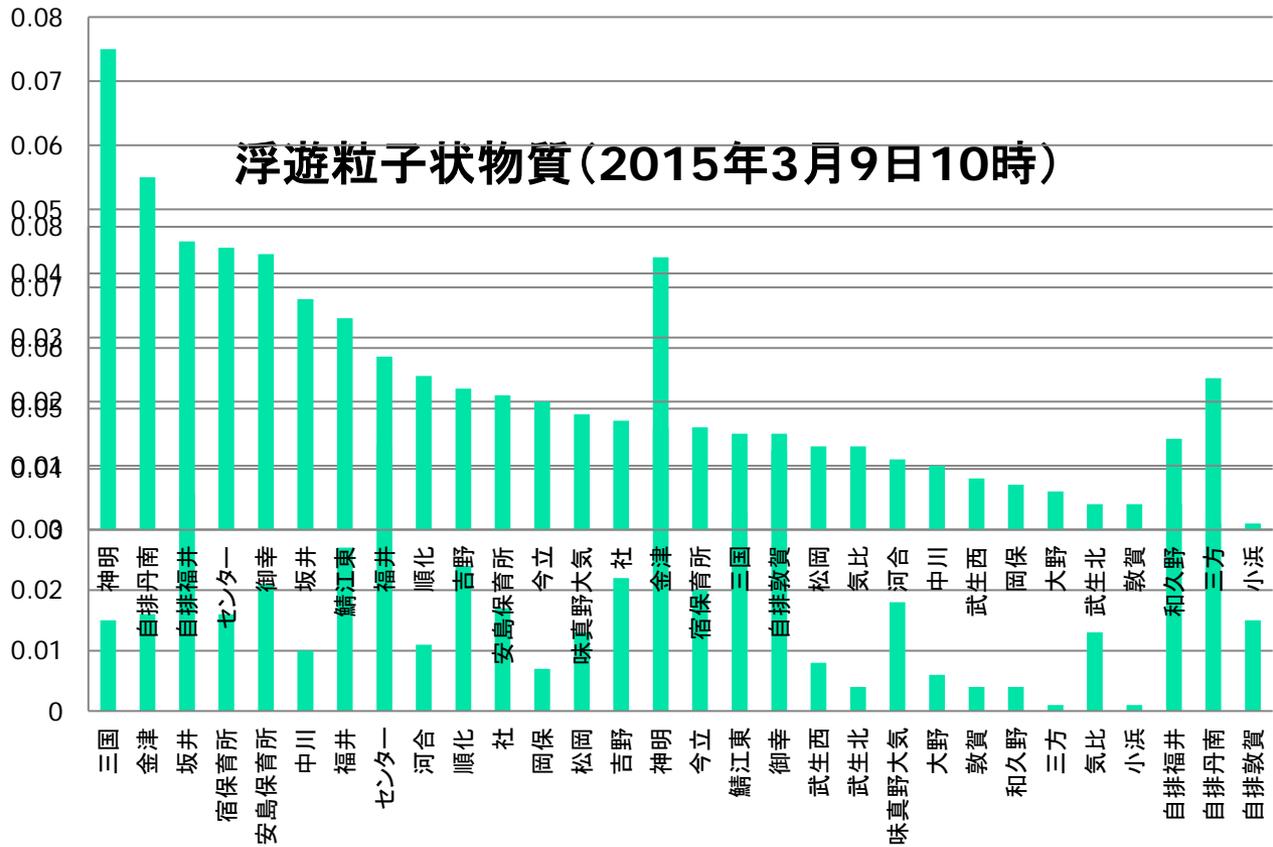
目的別グラフ表現

	比較	内訳	推移	分布	相関
単位グラフ・面積グラフ	◎				
棒グラフ	◎	○	○	△	△
帯グラフ	○	◎	○		
円グラフ	○	◎			
面グラフ	○	◎	○		
折れ線グラフ	△		◎		
レーダーチャート	◎	○	○		
散布図			○	◎	◎



棒グラフ例

浮遊粒子状物質(2015年3月9日10時)



局名	浮遊粒子状物質
	mg/m ³
三国	0.015
金津	0.016
坂井	0.036
宿保育所	0.016
安島保育所	0.021
中川	0.01
福井	0.027
センター	0.044
河合	0.011
順化	0.024
社	0.017
岡保	0.007
松岡	0.013
吉野	0.022
神明	0.075
今立	0.02
鯖江東	0.033
御幸	0.043
武生西	0.008
武生北	0.004
味真野大気	0.018
大野	0.006
敦賀	0.004
和久野	0.004
三方	0.001
氣比	0.013
小浜	0.001
自排福井	0.045
自排丹南	0.055
自排敦賀	0.015

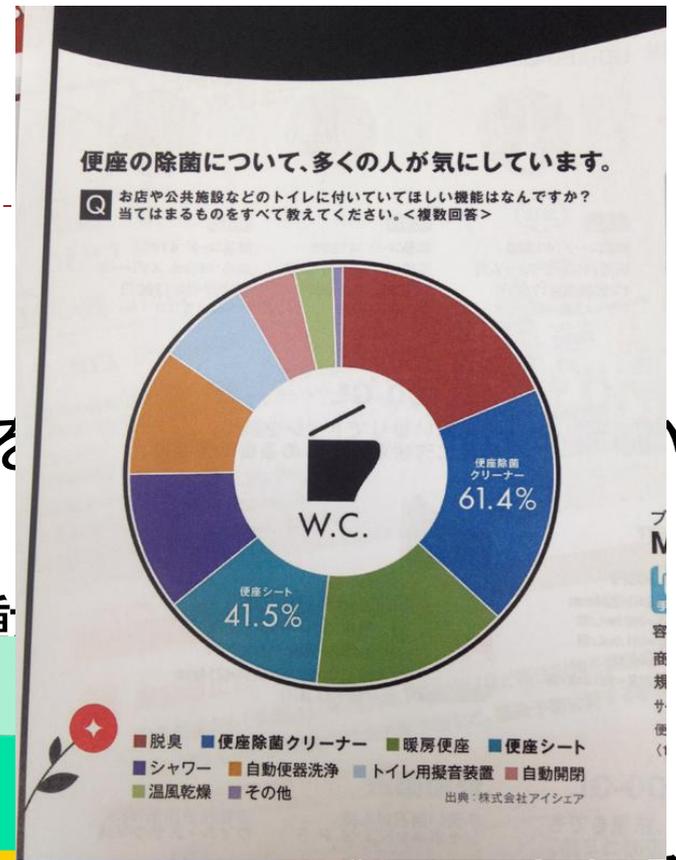
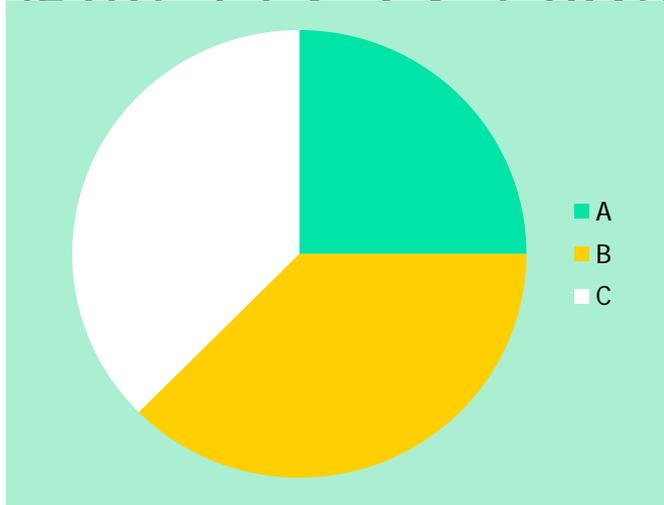
項目の並べ方の問題

- ▶ データの分類項目に序列がない場合、棒を数量の降順に並べると分かりやすいグラフになる
- ▶ データの分類項目に序列がある場合は、序列の通りに
 - ▶ 年齢順
 - ▶ 順序尺度
 - ▶ 地形、地名？



円グラフの誤用例

- ▶ 3Dグラフ: 楕円表示になるので誤解を
（けど使ってはいけない）
- ▶ 歪んだ円グラフ: 円の中心の位置を重



0%」となる

構成比率

- ▶ 一つの集団を特定の属性について観察すると、属性に関する集団の構造が得られる
 - ▶ 職業分類
 - ▶ 学部、課程、学年
- ▶ その構造を分かりやすく示すには、集団全体の大きさを基準とした比率で表すのが便利である
 - ▶ この教室には女性が～%、男性が～%.....
- ▶ 2つ以上の属性系列の間で構成の比較を行うためには欠くことのできない手段



構成比率の定義

- ▶ 集団の総数を1としたときの、各部分集団の大きさ
 - ▶ 全体に対する「パーセント」とか「割合」とか、そんな感じ
- ▶ 計算すると小数で出てくるが、日常の使用に便利ないように、100倍して%表記にする



構成比率の2つの型

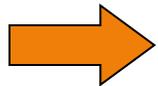
- ▶ 各部分集団のパーセントを足すと100になる
 - ▶ 属性の分類区分が排反である、すなわち、一人の人の持つ属性が重複しない
 - ▶ 構成棒グラフや構成パイグラフで示せる
- ▶ 各部分集団のパーセントを足すと100より大きくなる
 - ▶ 2つ以上の分類区分に入るものがある場合(アンケートの複数回答など)
 - ▶ パーセントを足すと100を越すことと、その理由を説明する必要



比率における注意点

比率：構成単位の総数を基準にしている。

（総数が分母）



分母は何か？

