

2021年度秋学期 統計学 第3回
クロス集計と感度・特異度, データの可視化

浅野 晃
関西大学総合情報学部



データの種類～尺度水準～🤔

データは数字だとは言っても

数字は, 必ずしも「数量」を表しているとは限りません

例えば, 選択肢の「1番・2番・3番」は,

数字ではない「a・b・c」でも「イ・ロ・ハ」でも同じだから, 数「量」ではない

尺度水準

統計学では, 数字を「数量」としての
意味をどのくらい持っているかで **これを尺度水準という**
4つのレベルに分けている

比例尺度
間隔尺度
順序尺度
名義尺度

量的データ 足し算引き算ができる

質的データ 足し算引き算ができない

名義尺度

選択肢を区別するための、単なる記号。

1番・2番・3番 さあどれ？

男性:1 女性:2

2番が1番より大きいという意味はない

順序尺度

数字の順番にのみ意味がある

この講義に満足しましたか？

1) 非常に不満・2) 不満・3) 満足・4) 非常に満足

2番は1番より満足度が大きいが、

「2番と1番の満足度の差」と「3番と2番の満足度の差」は同じではない

間隔尺度

数値の間隔にも意味がある

摂氏温度(20°C, -10°C)

「0°Cと10°Cの温度の差」と「-10°Cと0°Cの温度の差」は 同じ

20°Cは10°Cの2倍暖かい？ そんなことはない

20°Cは-10°Cの何倍暖かい？ ???

ところで:マイナスの温度？

摂氏温度には、マイナスはふつうに出てくる



2018年2月28日 フィンランド・タンペレ市

※米国で、頑なに「華氏温度」が使われているのは、「通常の気温ではマイナスが出てこないからわかりやすい」という言い分もある

ところで:マイナスの温度?

※寒い国では、気温が摂氏温度で+か-かは、
雪が融けるか融けないかの違いなので、きわめて重要



「+0°C」は、0°C以上+0.5°C未満であることを示す

2018年2月13日

比例尺度

間隔だけでなく比率にも意味がある

長さ・重さ・年齢など

40歳の人、20歳の人**2倍**の年数を生きている。

マイナスの値は存在しない
(温度なら、摂氏温度ではなく絶対温度がこれにあたる)

※絶対温度とは、これ以上冷やすことができない「絶対零度」を「0度」として表す温度。

簡単に「平均」というけれど

平均できるのは、足し算ができる量的データ(間隔尺度・比例尺度)だけ

この講義に満足しましたか?

1) 非常に不満・2) 不満・3) 満足・4) 非常に満足

こういうのの平均は、本当は意味がない

もし平均を計算していれば、それは間隔尺度だと近似的に考えていることになる。

クロス集計 🤔

質的データの解析について

次回以降は、平均を計算できるデータ=量的データ を扱います

今日は、質的データを扱うクロス集計について

クロス集計

例:商品Aが好きか嫌いか
→好きな人:50%, 嫌いな人:50%

これだけでは大したことはわからない

そこで、回答者が男性か女性かも記録しておく

※最近では、性別を尋ねる質問には注意を要します。
「男性・女性・答えない」という選択肢のものも多くなりました。ここでは、説明を簡単にするために「男性/女性」としておきます。
※関西大学では、2017年度から、受講生名簿に性別欄がなくなりました。統計学の講義を受けるのに、性別は関係ないですものね。

クロス集計

	好き	嫌い	合計
男性	20	30	50
女性	30	20	50
合計	50	50	100

男性は「嫌い」が多く
女性は「好き」が多い

これが[クロス集計]

ひとつのデータ群を2つの項目から見て、項目間の関係を表す

検査の感度

新しい検査法をテスト

	本当に病気である	本当は病気ではない
検査で陽性	A	B
検査で陰性	C	D
合計	A + C	B + D

$A/(A+C)$ [感度] 本当に病気の人のうち、検査で陽性になる人の割合

感度が高ければよいというわけではない

病気であってもなくても「陽性」と答えるなら、 $C=0$ で感度100% ※いわゆる「オオカミ少年」。

検査の特異度

	本当に病気である	本当は病気ではない
検査で陽性	A	B
検査で陰性	C	D
合計	A + C	B + D

$D/(B+D)$ 【特異度】 本当は病気でない人のうち、検査で陰性になる人の割合

特異度が高ければよいというわけでもない

病気であってもなくても「陰性」と答えるなら、 $B=0$ で特異度100%

感度と特異度

	本当に病気である	本当は病気ではない
検査で陽性	A	B
検査で陰性	C	D
合計	A + C	B + D

感度・特異度の両方を同時に100%近くにするのはむずかしい

「感度が90%のとき、特異度はいくら」という言い方で、検査の能力を表す

テキストの例題

ある病気を、感度80%、特異度99%で発見する検査があります。

この病気にかかっている人が検査対象者の1%であるとき、

検査で陽性だった人のうち、本当にこの病気にかかっている人の割合はいくらでしょうか。

テキストの例題

	本当に病気	本当は病気でない	合計
検査で陽性	80	99	179
検査で陰性	20	9801	9821
合計	100	9900	10000

テキストの表3

感度80%、特異度99%

この病気にかかっている人が検査対象者の1%

話を簡単にするため、対象者が10000人いるとする

10000人のうち、
本当に病気の人1%だから100人、本当は病気でない人が9900人

感度80%だから、
病気の人100人のうち 陽性になるのは80人、陰性になってしまう人が20人

特異度99%だから、
病気でない人9900人のうち 陰性になるのは9801人、陽性になってしまう人が99人

テキストの例題

	本当に病気	本当は病気でない	合計
検査で陽性	80	99	179
検査で陰性	20	9801	9821
合計	100	9900	10000

このクロス集計表で、

検査で陽性の人は合計179人　うち、本当に病気なのは80人

検査で陽性の人のうち、本当に病気の人は
 $80 / 179 = 0.447$ つまり44.7%　**半分にも満たない**

これでは検査の意味がありません。ではどうすれば？ それは演習問題で。

データの可視化

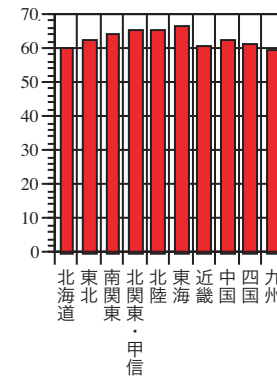
データの可視化

人は、
数字の羅列をざーっと見て即座に意味が理解できるほど
賢くはない

グラフなどの形に「描いて」理解しやすくする

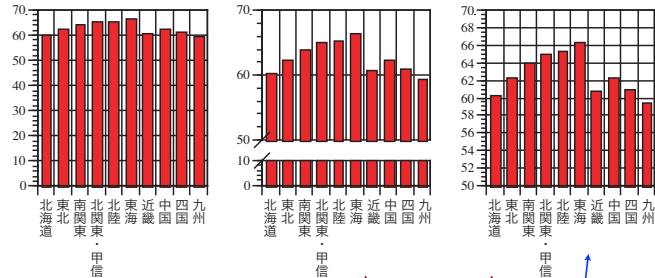
棒グラフ

棒グラフ, はよくご存じだと思いますが



棒グラフでは
横軸は名義尺度でもよい
数字でなくてもよい,　**というのが重要**

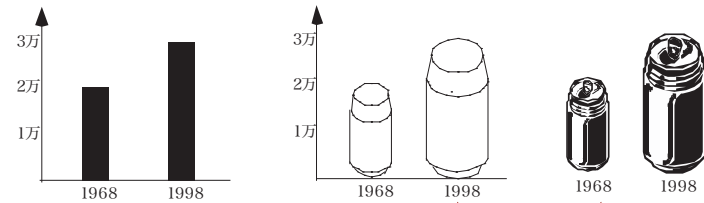
差が際立って見えるのはどれ？



棒の長さが値に比例していない

縦軸の途中を切断(ブレイク)するという「言い訳」すらしていない(ズル)

こんな描き方はあり？



長さが2倍なら、
面積は4倍
体積は8倍
になる

高さで量を表すはずなのに、棒の幅や厚み感も変えて、
面積・体積で表しているかのように印象づけている

縦軸がないから、体積で量を表しているように見える(ズル)

ナイチンゲールのグラフ

フローレンス・ナイチンゲールという人

フローレンス・ナイチンゲールは近代的な看護の先駆者

ナイチンゲールの誕生日の5月12日は
国際ナースデー(日本では「看護の日」とされています)。

※昔の欧州では、看護師は「卑しい職業」とされていたそうです。
ナイチンゲールをはじめとする人々により、現代のように、
看護師は医療の重要な担い手となりました。

やさしいかんごふさん？

※子ども向けのナイチンゲールの伝記では、
なぜか「やさしいかんごふさん」のように描かれているのですが…

ナイチンゲールの業績は

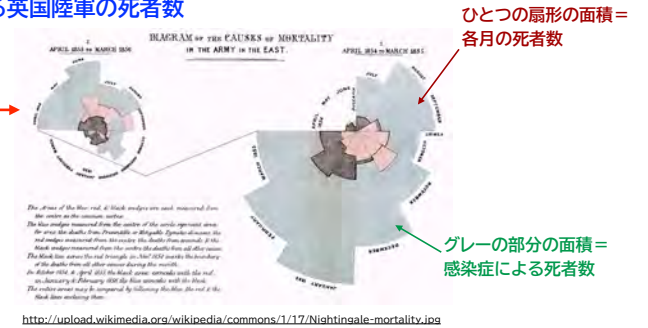
- 戦場の病院での**衛生管理**を徹底することによって、
感染症による死者を大幅に減らしたこと
- その実績をデータとしてまとめ、**グラフによる可視化**を行って
英国の女王や政治家に示すことで、**同国の政策を動かした**こと

ナイチンゲールは統計学の先駆者といわれています

ナイチンゲールのグラフ

クリミア戦争における英国陸軍の死者数

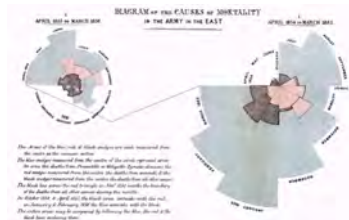
1855年4月以後、
衛生管理によって
各月の死者数が
大きく減ったことを
示している



しかし、この方式のグラフ(coxcomb「鶏のトサカ」といいます)は、現在の水準では問題があります。

ナイチンゲールのグラフの問題点

なにが問題なのか



- 面積で量を表すと、大小の印象はつきやすい。しかし、
長さや角度に目盛りをうつことはできるが、面積に目
盛りをうつことはできないから、面積で表された量を
正確に把握するのはむずかしい。
- とくに、正方形ではなく扇形の面積を把握するのはむ
ずかしく、大きな扇形が誇張されて見える。