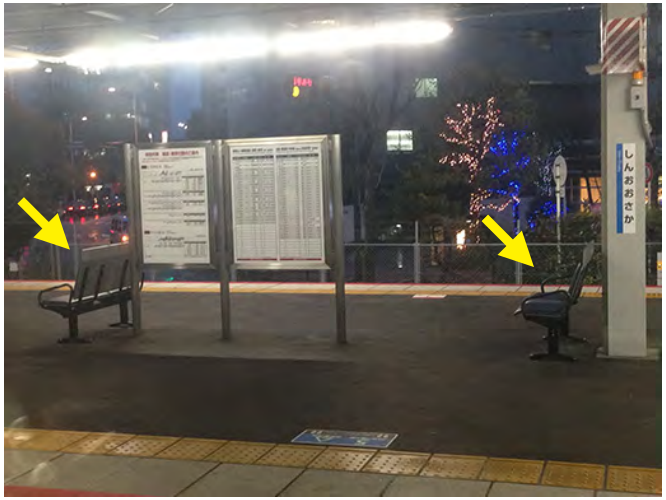
2024. 7. 12 夙川高等学校 出張講義

身近なデータサイエンス

関西大学総合情報学部
浅野 晃



思い込みにとらわれないために

なぜベンチが
「線路に向かって座る」から
「列車の進む向きに座る」に
変わったのだろう？

転落事故56件を調査すると

うち33件(6割弱)は
こうではなく線路に向かって歩いて落ちていた

読売新聞2015. 3. 31

🐻💬思い込みにとらわれず，
きちんとデータを調べよう

では，どうやってデータを調べる？

日本の大学生は世界一？🥇

英国放送協会（BBC） 2016年10月20日

いちばん聡明な学生たちがいるのは，本当はどこの国か？

※英国のオックスフォード大学・ケンブリッジ大学が世界大学ランキング最上位に入ったことに対して書かれた記事

The OECD tested literacy skills among graduates - and the high-flyers were not in the US or UK, but in Japan and Finland.

These figures, based on test results rather than reputation, show a very different set of nationalities from the usual suspects.

The OECD's top 10 highest performing graduates

1. Japan
2. Finland
3. Netherlands
4. Sweden
5. Australia
6. Norway
7. Belgium
8. New Zealand
9. England
10. United States

経済協力開発機構（OECD）諸国の大学卒業生能力トップ10

いったいどこのデータ？元ネタを調べよう

OECD（経済協力開発機構）の「教育概況」2016年版

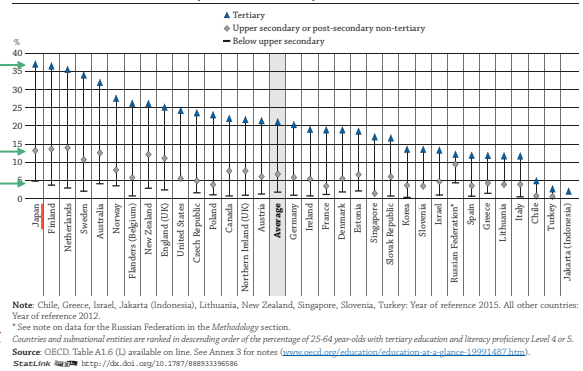
ネットで入手可能です。
講義ウェブサイトで（最後に紹介します）

データを見てみると

教育レベル別　国際成人力調査読解力テストでレベル4,5の人の割合

大卒
（第3期教育）レベル

高卒レベル
中卒レベル



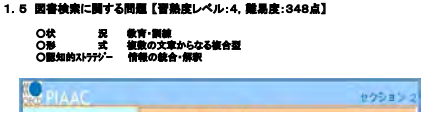
Note: Chile, Greece, Israel, Jakarta (Indonesia), Lithuania, New Zealand, Singapore, Slovenia, Turkey: Year of reference 2015. All other countries: Year of reference 2012.
* See note on data for the Russian Federation in the *Methodology* section.
Countries and subnational entities are ranked in descending order of the percentage of 25-64 year-olds with tertiary education and literacy proficiency Level 4 or 5.
Source: OECD, Table A1.6 (L) available on line. See Annex 3 for notes (www.oecd.org/education/education-at-a-glance-19991487.htm).
StatLink http://dx.doi.org/10.1787/888933396586

※本文の解説をよく読むと，国どうしの比較は目的ではないと書いてある

では，国際成人力調査（PIAAC）とは何？

※文部科学省のサイトに載っています。2022-23年に第2回調査が行われています。

調査の問題例



図書検索に関する問題
習熟度レベル4

図書検索結果を見てください。次の質問の答えとなる書名を選び、その横にある四角をクリックしてください。
遺伝子組み換え食品に賛成の主張と反対の主張のいずれも信頼できないと主張しているのはどの本ですか。

図書検索結果を見てください。次の質問の答えとなる書名を選び、その横にある四角をクリックしてください。
遺伝子組み換え食品に賛成の主張と反対の主張のいずれも信頼できないと主張しているのはどの本ですか。

少々複雑ですが，大学に行かないと習わないことでもない

高等教育で学ぶことは，こういう思考を適切に行えるようになること

現在では，
「元データ」にあたることは
むずかしくありません💡

🐻💬自分でデータを調べよう！

英語🏴は必要です💦
最近はよい翻訳ツールがありますが，
英語ができないとだまされます🪤

統計的見方

コロナ禍は
「終わった」のでしょうか？🦠

公衆衛生学とは

感染症を扱う医学は，「公衆衛生学」👨‍⚕️👩‍⚕️

ほかの医学が扱うのは，目の前のひとりの「人」👴
公衆衛生学が扱うのは，社会を構成する「人々」👫

「人々」の行動を完全にコントロールはできない👫
感染したかどうか，検査で完全にはわからない🦠
ワクチン💉は，感染を完全に防ぐわけではない

感染症と闘うデータサイエンス・統計学

社会を全体として見たときに，
感染の拡がりを抑えなければならない

統計学は，集団を全体として見て，その姿を把握する

「密閉・密集・密接の『三密』を避けよう」
「大人数の会食をやめよう」

- 統計学によって現状を把握して得られた指針
- 感染を社会全体として減らし，医療の逼迫を防ぐため
  （三密や大人数の会食を避けても，絶対に感染しないわけではない）

感染を必ずさけられるのではなくても

「密閉・密集・密接の『三密』を避けよう」
「大人数の会食をやめよう」

感染を必ず避けられるのではないのなら，いったい何のため？

一度に多人数に感染させる「クラスター」を防ぐ

一人の感染者が一人の人にしかうつさなければ，もとの感染者は回復するので，社会全体の感染者の数は増えない

一人の感染者が感染させる人数が「平均して」一人未満なら，
社会全体の感染者数は減っていく（実効再生産数が1未満）

個人ではなく，社会を救う

「平均して」「社会全体の」

というのが，データサイエンス・統計学の発想です

統計学で社会全体のようすを把握し，感染を社会全体で減らすのが↓

「密閉・密集・密接の『三密』を避けよう」
「大人数の会食をやめよう」

あなた個人👨👩を救うのではなく，社会全体🇯🇵🗾を救う

確率的見方

リスクとメリット？

👨‍⚕️💬「ワクチン💉は，リスクとメリットを考慮して，
接種するかどうかを自分で決めてください」

😨「そんなこと，考慮できるの？」

確率が小さいこととは

😨💬「コロナワクチン接種で重篤な副反応が出るのは

10万人に1人の確率だといっても,

その副反応が出た人にとっては100%重篤な事態だ」🤔

それはそのとおりで,「確率が小さいかどうか」と

「事態の重篤さが小さいかどうか」は関係ありません。

くじ引き🎯で,「当たり確率」と「賞金の額」は別の問題なのと同じ

確率とは？

「くじの当たり確率 0.3」とは, 次のような意味です（どちらでも同じ）

- くじを十分多くの回数引くと, そのうち10回に3回の割合で当たる
- 十分多くの人がそれぞれ1回くじを引くと,
  その人たちのうち10人中3人が当たりをひく

いずれにしても,

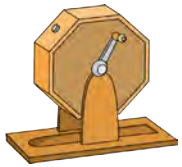
「十分多くの回数」「十分多くの人」について言っていることを

「1回」「ひとり」に当てはめている

確率がわかっても

確率がわかっても,

次の「1回」のくじ引きの結果はわからない。



https://illpop.com/png_season/dec01_a07.htm

※この機械は「新井式廻轉抽籤器」というそうです

確率は, くじ引きのような「ランダム現象」を扱う

ランダム現象とは,「結果に人知の及ばない現象」

確率を云々しても, 人知が及ばないことに変わりはないけれど

「どんな結果になることがどのくらい多いか」を考える

期待値とは

さきほど「別の話」と言った「当たり確率」と「賞金の額」を結びつけて

くじ引きで考えれば, （どちらでも同じ）

- くじを十分多くの回数引いたときの, 1回あたりに得られる賞金の平均
- 十分多くの人がそれぞれ1回くじを引いたとき,
  ひとりが得られる賞金の平均

プロのギャンブラーは

いくらプロのギャンブラーでも

次の1回の賭けに勝てるかどうかはわからない

プロのギャンブラーは

日頃から多くの回数の賭けをする→
賞金の期待値の大きい賭け方を見抜いて賭けることができれば,

1回1回の賭けでは勝ち負けがあっても,
多くの賭けの合計では勝つことができる

リスクとメリットは, 考慮できるか

ワクチン💉の話にもどると

リスクとメリットを考慮して, といわれても

日頃から多くの回数の賭けをするギャンブラーなら
賞金の期待値を問題にすることができるけれど

一生に1度しかしないことの確率や期待値を考えるのはむずかしい

人間の思考の限界？🤔

リスクと統計学

統計的推測とは

もうずいぶん昔ですが, 1994年に
ノルウェー🇳🇴 のリレハンメルで開かれた五輪の開会式で, アナウンサーが

🎤💬「ノルウェー人は背の高い人が多く,
平均身長は男179cm, 女170cmです」

ノルウェー人全員の身長を測ったんですか？🤔

標本調査と統計的推測

当然ながら，身長は人によって違う（「分布している」という）

ノルウェー人全員ではなく，一部の人だけ（標本）を調べて，
ノルウェー人全体のようすがわかるのか？

わかります。かなりの程度わかります。

「一部の人」を選ぶのに，くじびきで選ぶ（無作為抽出）

くじびきで選べば，たいていはいろんな人がまんべんなく選ばれる

「たいてい」と「ほぼ」

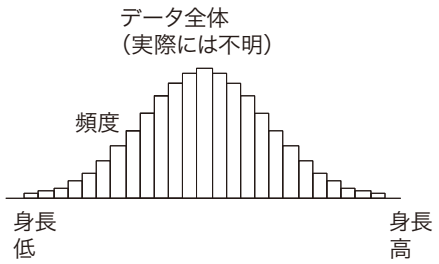
くじびきで選べば，たいていはいろんな人がまんべんなく選ばれる
→選ばれた人の平均は，ほぼ全体の平均に近い

本当？😔
くじびきなんだから，たまにはバレーボール🏐の選手みたいな
大きな人ばかり選ばれることもあるのでは。

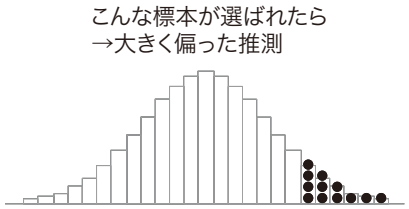
🐻💬そのとおりです。「たまには」そういう失敗をします。
でも，失敗をする確率を計算できます。

無作為抽出すると

分布がこんなようすのとき

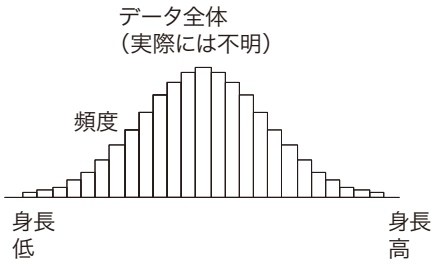


偶然こんな標本（●）が選ばれて
しまう確率は小さい

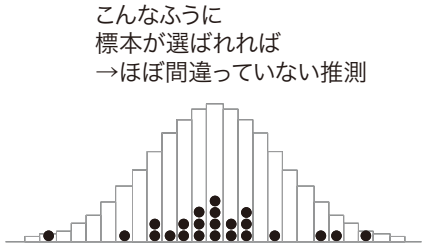


無作為抽出すると

分布がこんなようすのとき



「たいてい」は，
こんなふうに選ばれる



偏って選んでしまうことは…

たくさんの人を抽出すれば，偏らないんじゃないの？🤔

🐻💬 だから，「思い込みはだめ」だって。

無作為抽出なら，そう期待できる。
無作為抽出でなければ，必ずしもそうではない。

SNSから見える世論は
「自分が選んだ世論」
「自分の好みをAIに『おすすめ』された世論」

区間推定

「区間推定」という統計学の方法では，

「ほぼ」
「ノルウェー人男性全体の平均身長は，179cm～182cmの間と推測する。
この推測が当たっている確率は95%」
「たいてい」（失敗の確率5%）

と答える

リスクを（再び）考える

「ほぼ」
「ノルウェー人男性全体の平均身長は，179cm～182cmの間と推測する。
この推測が当たっている確率は95%」
「たいてい」（失敗の確率5%）

「失敗の確率」は
このような統計的推測を「何度も」行う時，
どのくらいの割合の推測が失敗するかを表す

1回だけ推測する時に，それが成功するか失敗するかはわからない

これと同じですから。

リスクを（再び）考える

「ほぼ」
「ノルウェー人男性全体の平均身長は，179cm～182cmの間と推測する。
この推測が当たっている確率は95%」
「たいてい」（失敗の確率5%）

1回だけ推測する時に，それが成功するか失敗するかはわからない

このような統計的推測を何度も行うのなら，
失敗の確率≒失敗の割合　だから，
それに対する備えをする，つまり「リスクを考える」ことができる

人間のデータサイエンスと
人工知能のデータサイエンス

機械のための新しいデータサイエンス

統計学は，人間が集団の姿を把握するためのものだった

統計学（statistics）は，国家（state）と同語源

最近急速に進歩してきた人工知能・機械学習は，
コンピュータが集団の姿を把握するデータサイエンス

人間にわかるかどうかは別問題

コンピュータ棋士は，なぜその手を指すのか，
人間にわかるようには教えてくれない

何を学習してきたかも，教えません。

もしかしたら，「偏った世論」かも…

浅野晃の講義ウェブサイトについて

浅野晃の講義ウェブサイト

http://racco.mikeneko.jp/Kougi/

浅野晃のWebポータル

http://racco.mikeneko.jp/