統計学

兵庫県立大学 社会情報科学部

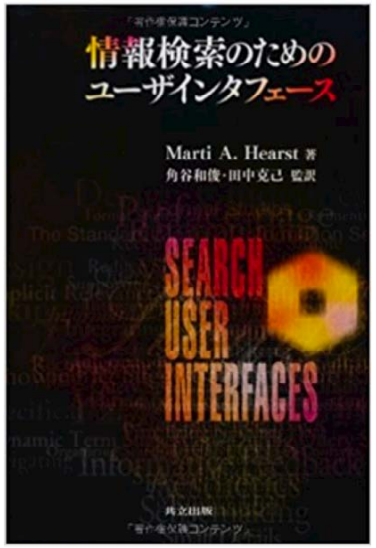
山本 岳洋

t.yamamoto@sis.u-hyogo.ac.jp

平成31年度前期・火曜2限　神戸商科キャンパス 全学共通科目

自己紹介

- **山本 岳洋（やまもと たけひろ）**
  - 社会情報科学部 准教授
  - 1984年生まれ，広島県出身.

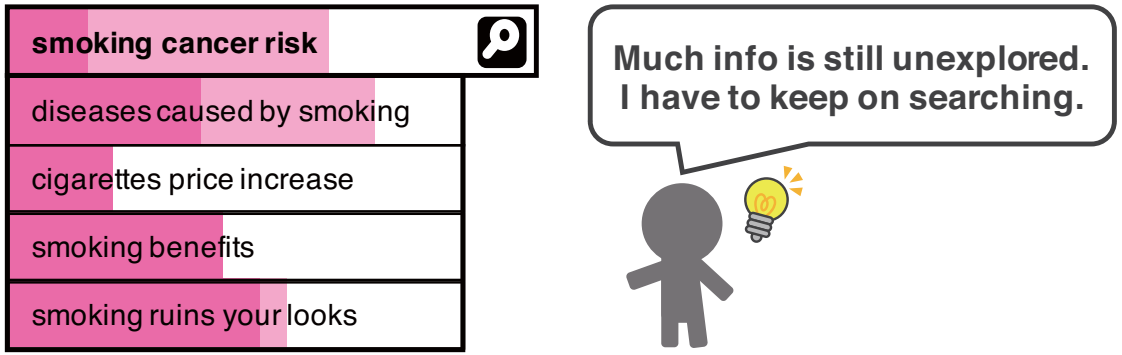


- **専門分野**
  - 情報検索
  - ヒューマンコンピュータインタラクション

こんなことやってます

新しい情報アクセス技術

モバイル協調検索技術

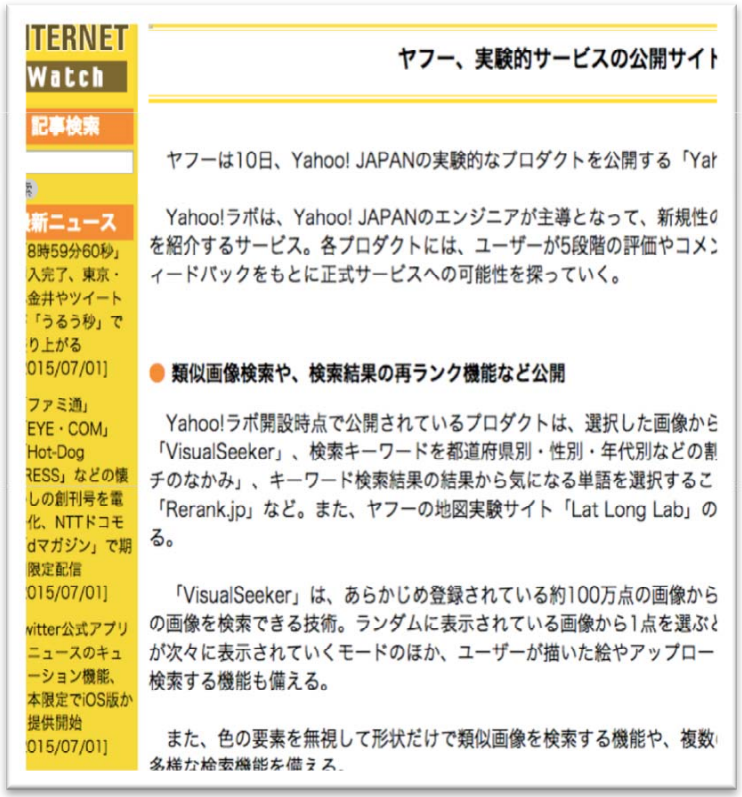


「多様な情報閲覧」を支援するインタフェース

視線情報に基づく検索意図理解

こんなことやってます

企業との共同研究



https://internet.watch.impress.co.jp/cda/news/2009/06/10/23736.html より引用

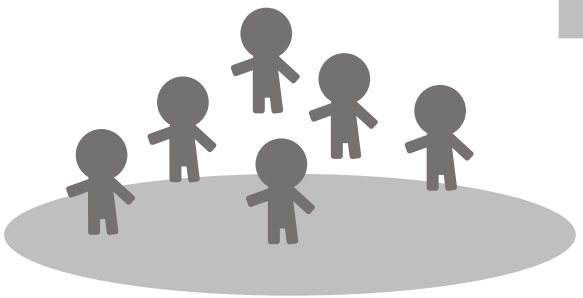
こんなことやってます

Q: 普段のWeb検索で、
正しい情報を取得するよう
どの程度心がけていますか？

こんなことやってます

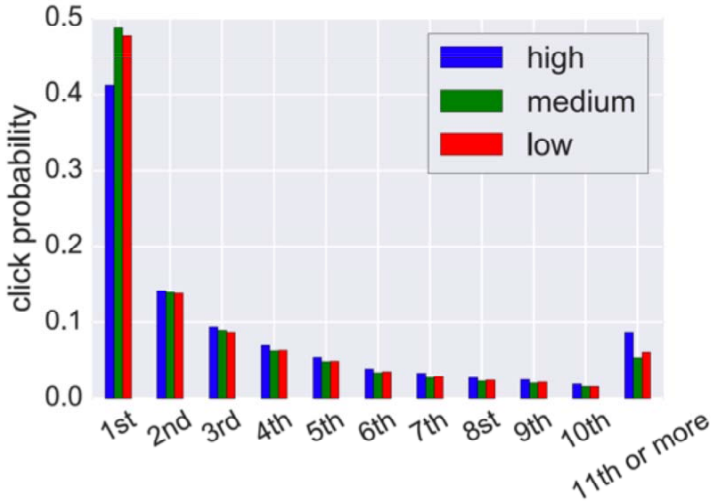
意識の高くない人は高い人に比べて，
1位の検索結果だけをみて検索を終えることが多い

意識の高い人達



そうでない人達

Yahoo! Japan
の検索ログ



何位の検索結果を
よくクリックするか？

なぜ統計学を学ぶか

- **先ほどの研究**
  - 意識の低いグループが高いグループに比べて1位の検索結果だけをよくクリックしている、と主張している

- **本当にそう言える？たまたまでは？**

本日の流れ

- 統計学をなぜ学ぶか
- 本講義の説明
- アンケート
- 本講義で扱う内容の概観

なぜ統計学を学ぶか

- **統計家は今後の**
  最もセクシーな職業だ
  - Hal Varian （Googleチーフエコノミスト）

https://www.amazon.co.jp/dp/4478022216 より引用

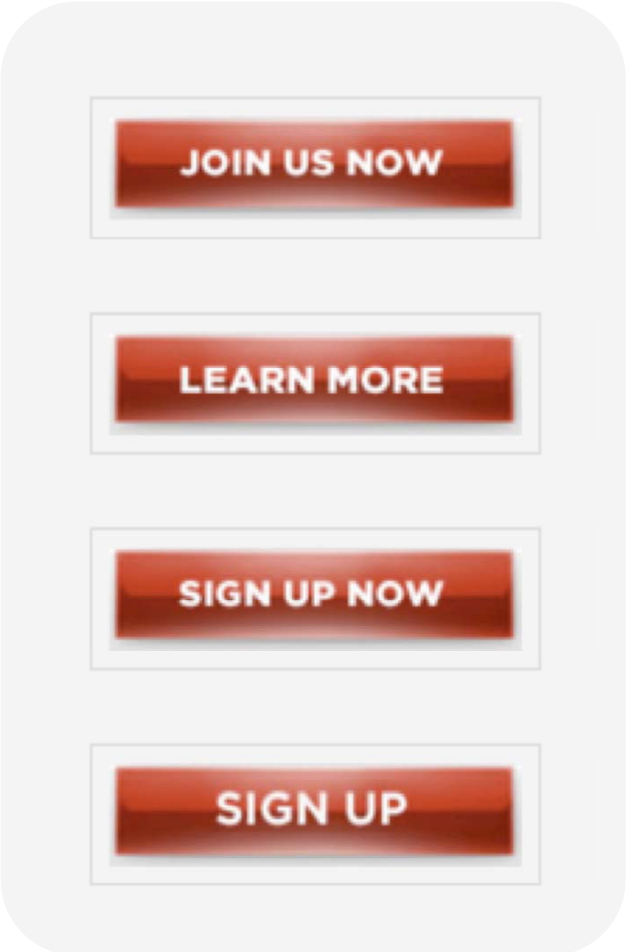
なぜ統計学を学ぶか

- **ビッグデータ時代**
  - 人々の購買データ
  - ウェブ・ソーシャルメディアデータ
  - EBM（エビデンスに基づく医療）
  - EBPM （エビデンスに基づく政策）

- **データに基づいた意思決定**
  - 統計学の知識が不可欠
  - 「経験と勘」からの脱却

どのボタンが効果的か？

どのボタンが最も多くの
ユーザを獲得できるか？



http://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/ より引用

大学生に4人に1人は平均が分かっていない？

問題

> ある中学3年生の生徒100人の身長を測り、その平均を計算すると163.5センチになりました。この結果から確実に正しいといえることには「○」、そうでないものには「×」と答えてください。全問正解で正答。
>
> - 身長が163.5センチよりも高い生徒と低い生徒は、それぞれ50人ずついる
> - 100人の生徒全員の身長をたすと、「163.5×100」で16350センチになる
> - 身長を10センチごとに「130センチ以上で、140センチ未満の生徒」「140センチ以上で150センチ未満の生徒」……というように区分けすると、「160センチ以上で、170センチ未満の生徒」が最も多い
>
> （大学生数学基本調査より）

この問題は、日本数学会が2011年に行った「大学生数学基本調査」で使用されたもの。「小学6年生の教科書の典型的な記述に沿って作問した」とのことで、難しい統計学の知識ではなく、小学校で学ぶべきことがしっかり頭に入っているかどうかが問われています。

さまざまな偏差値の大学、学部に通う学生約6000人に出題したところ、正答率は76%。あなたはちゃんと解けるでしょうか？

https://nlab.itmedia.co.jp/nl/articles/1903/25/news116.html より引用
（一次資料は 大学生数学基本調査 https://www.mathsoc.jp/comm/kyoiku/chousa2011/index.html ）

政府統計不正問題

JIJI.COM

時事ドットコムニュース ＞ 時事ワード解説 ＞ 毎月勤労統計の不正調査

B! コメント 小 中 大

毎月勤労統計の不正調査

2019年01月24日06時59分

毎月勤労統計の不正調査　厚生労働省の毎月勤労統計調査で、従業員５００人以上の事業所を全数調査すべきところ、２００４年から東京都分について約３分の１の抽出調査にしていた問題。０４～１７年は抽出した数値を全数に近づける復元処理を行っていなかったため、給料が高い東京都の大企業の比率が本来より小さくなり、平均賃金などが低くなっていた。

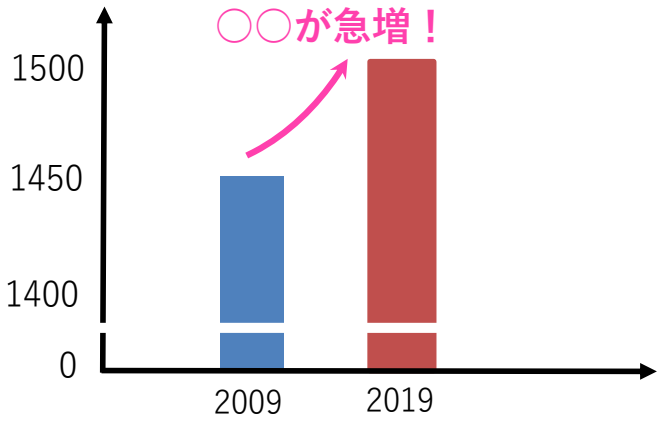
勤労統計の賃金額が給付水準に連動する雇用保険や労災保険などで、約６００億円の支払い不足が発生。延べ約２０００万人に影響が出た。政府は閣議決定をやり直し、１９年度の当初予算案を修正するという異例の対応を余儀なくされた。

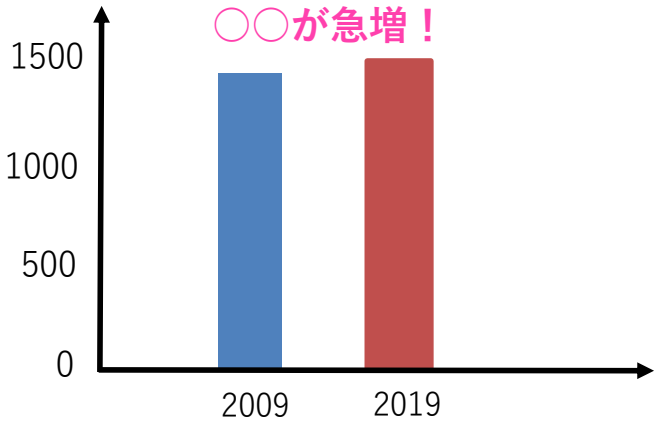
【時事ワード解説記事一覧へ】　【アクセスランキング】

https://www.jiji.com/jc/article?k=2019012400287&g=tha より引用

統計学 を学ぶ必要のある人

- **将来，研究や業務でデータ分析に関わる人**
  - 適切にデータを収集し，適切に判断をする
  - ほぼすべての人？
- **教養としての統計リテラシー**
  - 統計の「嘘」にだまされない，「嘘」をつかない





本講義「統計学」

講義ページ

https://tyamamot.github.io/h31statistics/

必ずブックマークを！

講義目的・到達目標（シラバスより）

データに基づいて現象を客観的に理解し，推測することは現代社会において必要不可欠ともいえる技術である．本講義では，そのための方法論である統計および確率の基礎について学ぶ．　本講義の到達目標は，統計および確率の基礎理論を修得し，「統計的なものの見方」の基礎を身につけることである．

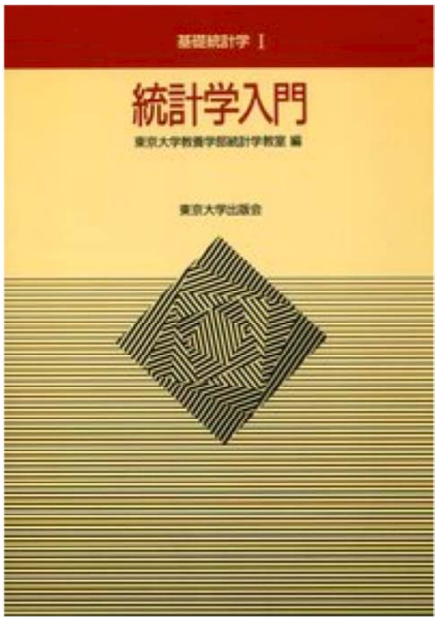
教科書

岡本和夫 著「**新版 確率統計**」実教出版（2012）

http://www.jikkyo.co.jp/book/detail/122514 より引用

参考書

東京大学教養学部統計学教室 編
「**統計学入門（基礎統計学Ⅰ）**」
東京大学出版会（1991）



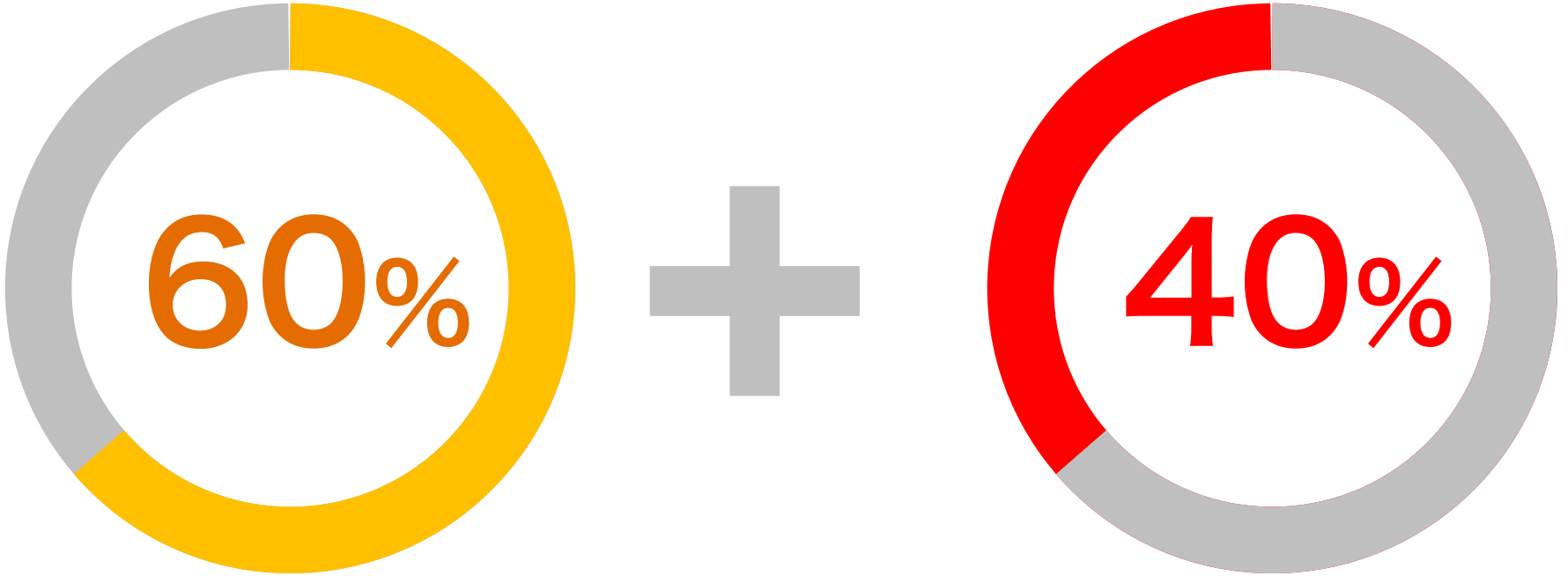
http://www.utp.or.jp/book/b300857.html
より引用

社会情報科学部1回生は
後期「確率・統計」で使用します

日本統計学会 編
「**統計検定2級対応 統計学基礎**」
東京出版（2015）

http://www.tokyo-tosho.co.jp/books/978-4-489-02227-2/
より引用

成績評価（シラバスより）



レポート

講義中の理解度確認テスト
および定期試験

を基準として受講態度（積極的な質問等）を
含めて総合的に評価する.

レポートの中身

- **講義後に課すレポート課題（数回）**

- **講義中に行う小テスト（数回）**
  - こちらを行う場合は事前にアナウンスします

スケジュール（予定）

1.	4月9日	講義概要
2.	4月16日	1変数データの記述と要約1
3.	4月23日	1変数データの記述と要約2
4.	**4月30日**	2変数データの記述と要約1
5.	5月7日	2変数データの記述と要約2
6.	5月14日	確率の基礎1
7.	5月21日	確率の基礎2
8.	5月28日	**理解度確認テストとこれまでのまとめ**
9.	6月4日	確率変数・確率分布
10.	6月11日	正規分布とその他の確率分布
11.	6月18日	母集団と標本
12.	6月25日	推定量の性質と推定量の特性
13.	7月2日	母平均・母比率の推定
14.	7月9日	仮説検定
15.	7月16日	まとめと発展的な話題
16.	7月X日	**定期試験**

アンケート（成績とは一切関係ありません）

https://forms.gle/pRjxvRXab1zd2e3T7

本講義で扱う内容の概観
（ざっくりと）

記述統計

データの整理
- 1変数データ
- 2変数データ

推測統計

母集団と標本

母集団の推定

仮説検定

確率の基礎

集合・確率

確率分布

1変数のデータの記述と要約

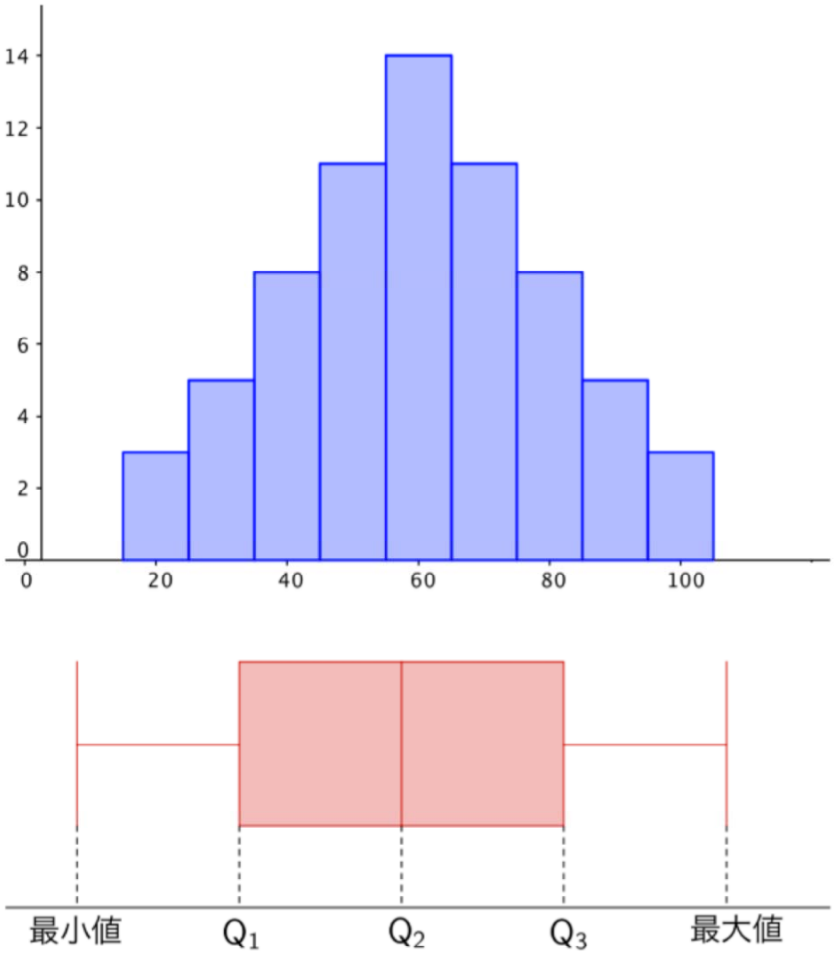
- **記述統計**
  - 収集したデータを**正しく**，**効率的**に把握する

- 日本国民の**所得**は？
- 学生50名の**試験成績**の分布は？
  - **所得**や**試験成績** ＝ **変数**

43	100	35	60	57	68	41	63	100
42	21	28	37	76	45	34	62	59
71	34	57	81	74	90	52	21	100
68	78	78	37	68	70	21	65	71

1変数データの記述と要約

- 平均・中央値
- 分散・標準偏差
- ヒストグラム
- 四分位数
- 箱ひげ図
- *etc.*

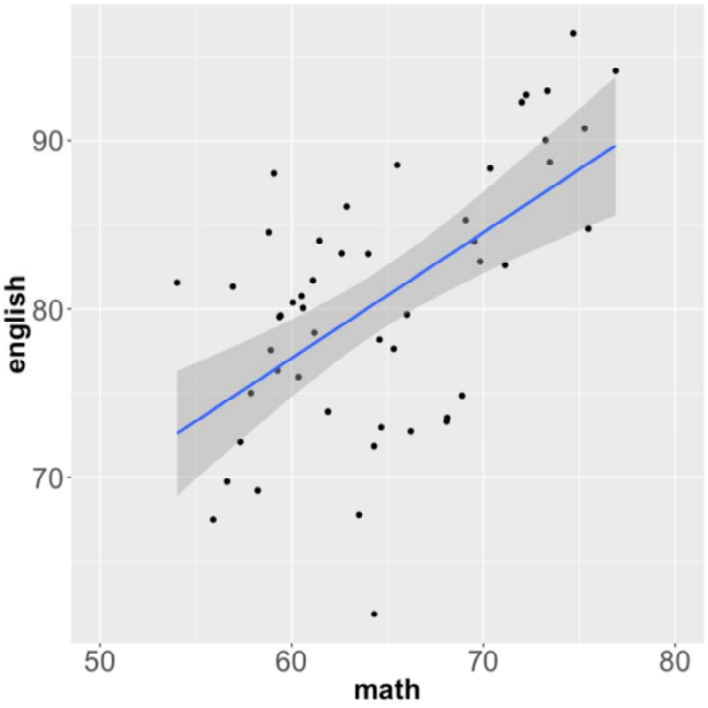


https://高校数学.net/hakohige-hisutoguramu/ より引用

2変数のデータの記述と要約

- **2変数間の関係を知りたい**
  - 学生の**数学**と**英語**の点の関係
  - 賃貸マンションにおける
    家賃と**駅までの距離**の関係
- **散布図**
- **相関係数**
- **回帰**
- *etc.*

学生50名の数学と英語の点数の散布図
(架空データです)



確率の基礎

- **講義後半に扱う推測統計の準備**
  - 確率 = 限られたデータから全体の法則性を理解するための道具

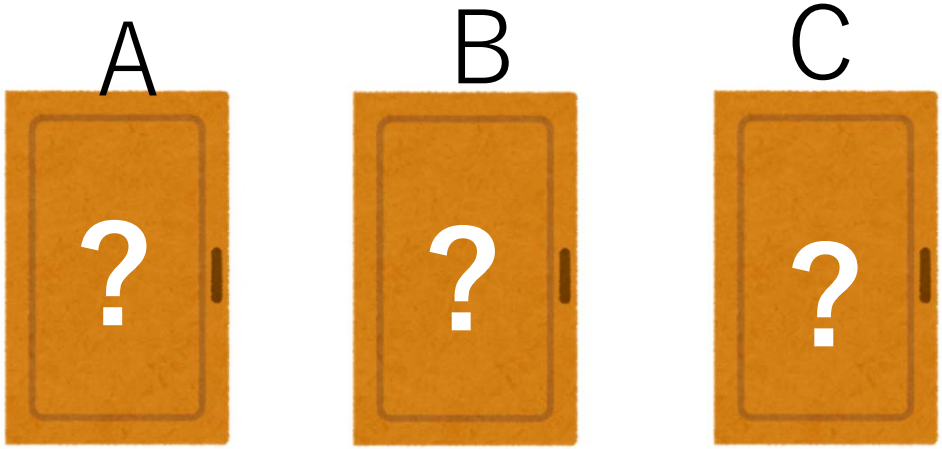
- **確率とは**
- **条件付き確率**
- **ベイズの定理**
- **高校数学の復習**

モンティホール問題

- **3つの扉 A, B, C に正解が1つだけある**

1. 回答者は扉を1つ選ぶ
2. 答えを知っている司会者が，**回答者が選ばなかった扉**で，不正解の扉を1つ開ける
3. 回答者は，始めに選んだ扉を変更できる

Q. **回答者は選択を変更すべきかどうか？**



モンティホール問題

A

B

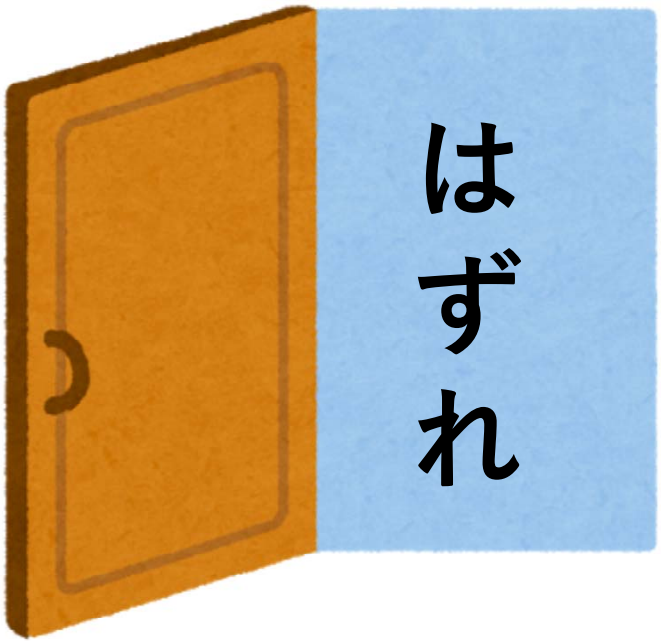
C

モンティホール問題

A

B

C

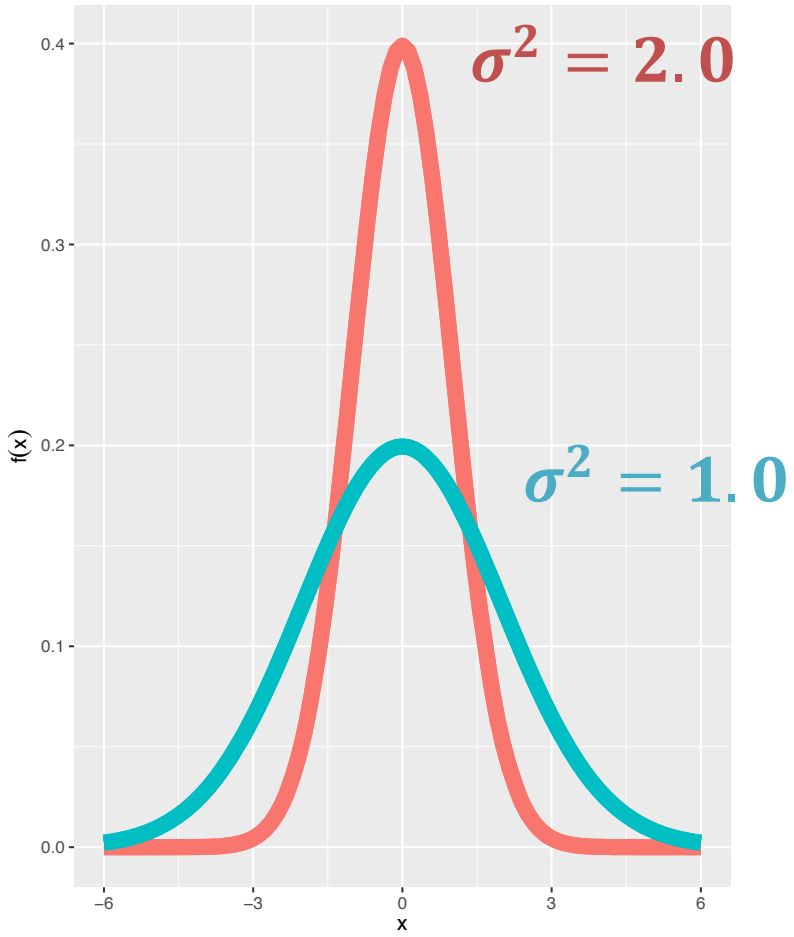


AからBに選択を変更すべきか？

確率分布

- **正規分布**
  - 自然界や人間社会の多くの現象が当てはまる

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



正規分布の例

- **二項分布**
  - コインを n 回投げて表が k 回出る確率

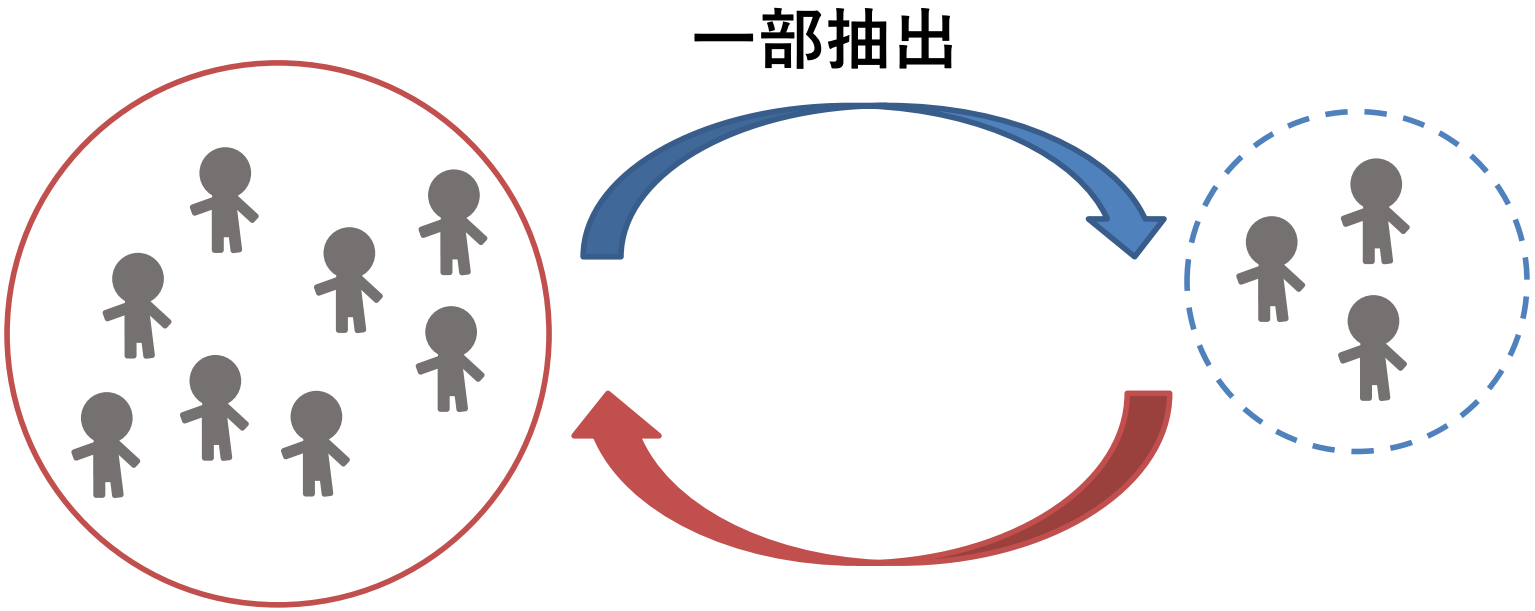
- **その他の分布**

母平均・母比率の推定

- **区間推定**
- **信頼区間**
- **母平均**
  - この区間内に、神戸市全中学生の平均があると95%の確率で言える
- **母比率の信頼区間**
  - 神戸市全中学生のスマートフォン保持率がこの区間内にあると95%の確率で言える

母集団と標本

- **限られたデータから全体を予測したい**
  - 神戸市に住む中学生の100m走の平均タイムは？
  - 神戸市に住む中学生のスマートフォン保持率は？

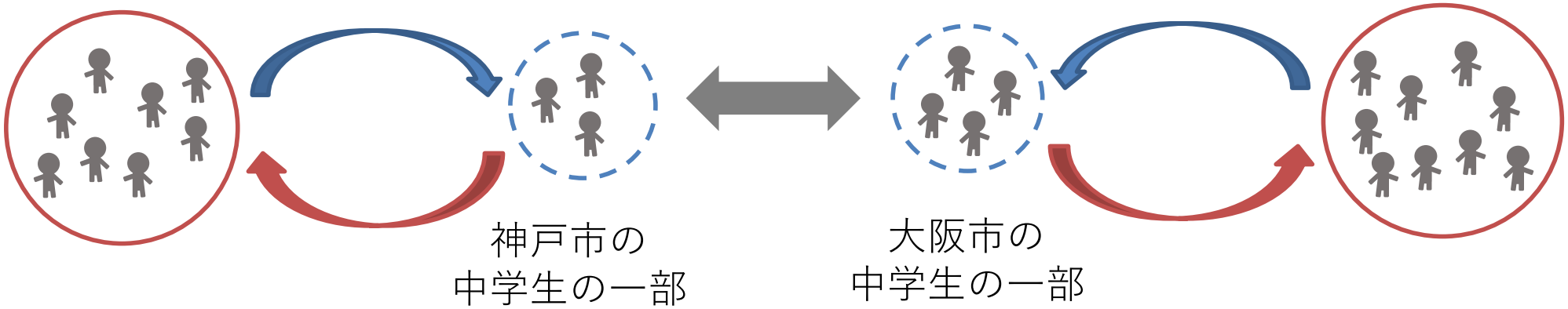


母集団
神戸市に住む
全中学生

標本
神戸市に住む
中学生の**一部**

仮説検定

- **統計的仮説の正しさを判断する方法論**
  - **神戸市の中学生より大阪市の中学生の方が100m走が早い**、という仮説は正しいと言えるか？



※ この講義ではこの問題までは扱わず，もっと単純な問題を扱う予定

次回

- 今週は課題はありません
- 次回: 4月16日（火）