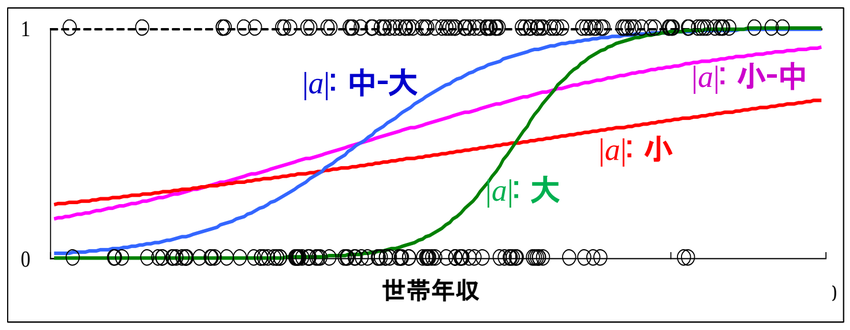
# 大学の統計授業で使うデータについて

島谷　健一郎　データ科学研究系　准教授

統計の授業で使うデータは、大きく3つある。
1. インターネットで公開されているデータ
2. 人工的に作った数値
3. 授業時に学生自身が調査して作ったデータ




$$\frac{1}{1+e^{ax+b}}$$

3が好ましいように思えるが、専門科目の実習や実験で3を実践する専攻では、3を統計の授業で行う必要性は必ずしも高くない。
1は時流に沿っているように思えるが、自身でデータ収集を行う専攻では、完成されたデータやフォーマットが規格統一されたデータを扱う意義は、必ずしも高くない。
2でできることは計算練習くらいで、統計解析結果についての考察ができないから不適切と考えられがちだが、はたしてそうだろうか？
人工データで統計モデルと計算法を学び、同じようなデータを「想像」して考察を行う作文課題という統計教育を、0-1データに関するロジスティック回帰モデルで実践している。

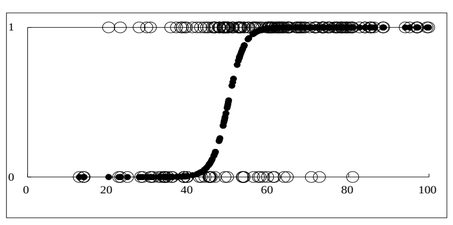
作文課題
1. 0-1データの例を考えてください。
2. 連続的なxが変化すると, 0-1のyが、0が多いから1が多いに変化するような例を考えてください。
3. そんなデータについてロジスティック回帰モデルを適用すると、どんな知見（予測）が得られるでしょう。

青枠は大学生の作文例（一部改変）　赤枠は講師（本発表者）が大学生に与えたコメント

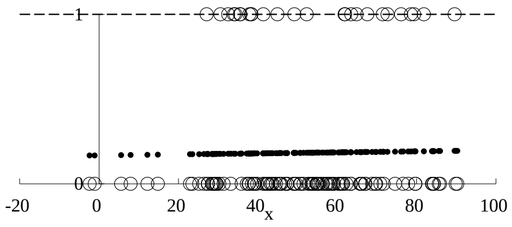
## サッカーにおける枠外シュートと勝敗

枠内シュートを打たないとゴールは決まらないから積極的にシュートを打つことを求められるが、結果として枠外シュートに終わっていいのだろうか。正の相関が見られれば枠内、枠外を問わずシュートを打つことは試合の勝利につながる。

予測可能




予測できない




## 女性の肌の露出度と痴漢の被害

女性が性被害に遭うと女性の危機管理能力の欠如を責められる二次被害が多い。「露出の多い服装だから被害に遭う」という主張と「地味でおとなしそうな服装のほうが狙われやすい」という主張が論争を繰り広げている。こうしたデータにより論争を終わらせ二次被害も減らせられる？

## スマートフォンやゲームなど画面を見る時間と、メガネやコンタクトの使用

画面を長く見ていると視力が落ちるかどうか検証できないか。

## 世帯年収と男性の育児休暇の取得

世帯年収が増加すると育児休暇取得も増加
>>男性の育児休暇取得率が低い理由として収入面での不安がある
そのような結果が得られない>>上司や同僚の理解が得られないことなど他の要因

## 被害女性だけ見ての意見が多いが、被害のない女性との比較は必須。実験では必ず 対照区(control ) を設けるが、野外調査や社会調査では、「起こった」データだけ見て考察する過ちを犯す。しばしば「なかった」データのほうが収集も分析も難しい

総括
実際のデータ収集及び解析で重要な視点や、どういう課題とデータにどういう統計手法が適切か等々を学ぶ機会を「想像データ」はもたらしてくれる。
人工データでも、「想像データ」と組み合わせることで、案外と統計教育効果があるのかもしれない

## テストの点数と朝食を食べるか

「朝食を食べるかとテストの点数」と考え、t検定などのほうが望ましい？