Ⅲ.「学びの杜・学術コース」講義録

統計分析で嘘をつかないために
ーデータの偏りと見せかけの相関ー

名古屋大学
大学院経済学研究科／アジア共創教育研究機構
根本 二郎

講義実施日：2022 年 8 月 26 日

データ、そしてデータを対象にした科学であるデータサイエンスは、今、スポットライトを浴びています。自然現象も社会現象も、データを通して理解できます。理系の人も文系の人も、みんなデータっていうのを扱います。

自然現象も社会現象もデータを通して理解できる。

データを分析していろいろなことを明らかにすることができる。
またその結果を利用して価値を生み出すことができる。

データサイエンス：データを対象にした科学

統計学・情報学など多くの分野の専門知識をあわせて価値を創造する。

参考：竹村彰通「データサイエンス入門」岩波新書

データを分析して、とにかくいろんなことを明らかにすることができます。その結果を利用して、価値というものを生み出すことができます。経済的な価値だったり、人間的な価値だったりと、いろいろありますけれど、とにかく新しい価値を生み出すことができます。統計学とか情報学とか、実に広く多くの専門分野の知識を合わせて価値を創造します。

統計的因果推論。原因と結果の関係を明らかにするためにデータを使うということです。僕の関係してる分野では、すごくよくやるんです。

データを分析していろいろなことを明らかにすることができる。
またその結果を利用して価値を生み出すことができる。

たとえば 統計的因果推論

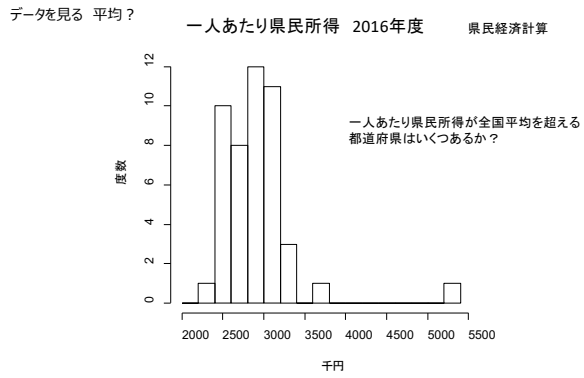
原因と結果の関係を明らかにするためにデータを使う。

- ▶ データを分析して物事の原因を探る。
- ▶ 原因がわかれば、より良い結果を得るための方策を考えることができる。

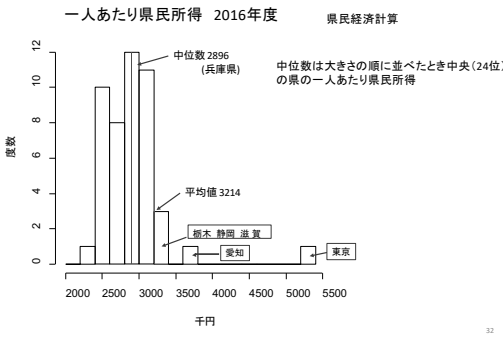
新型コロナウイルスの感染が拡大する原因は何だろうか。経済活動が原因だっていうことになったら、経済活動を抑制しなければいけない。そうすると、みんな困るわけです。経済活動をできなくなると、お金が回らなくなって、みんな大変になる。だけど、実はそうじゃないかもしれない。そうじゃなかったら、経済活動を抑制なんかする必要はなくて、むしろそんなことをするとみんな困っちゃうだけだと。最初の頃は三密と言ってたけれど、三密が原因だったら、三密を避けて経済活動をしましょうというような話になるし。原因がきちんと分からないと、いろんな対策も立てられないということなんです。まず原因を探って、原因が分かれば、よりよい結果を得るための対策とか、いろんな策を考えることができます。

しかし、データを見れば直ちに何か分かるわけではありません。それどころか、データの見方、扱い方を誤れば、嘘をつく道具にもなります。

1 人当たりの県民所得。早速、経済っぽい話になっちゃうんだけど、47 都道府県の 1 人当たり県民所得です。個人の所得だけじゃなくて企業の所得も含んで、それを人口で割ったものが 1 人当たり県民所得です。グラフを見てください。単位は 1,000 円だから、一番右が 550 万円、一番左が 200 万円、この範囲に分布しています。1 人当たり県民所得の高い県が一つあります。これは突出して高いでしょう。どこでしょうか？（生徒：東京です。）東京ですよね。



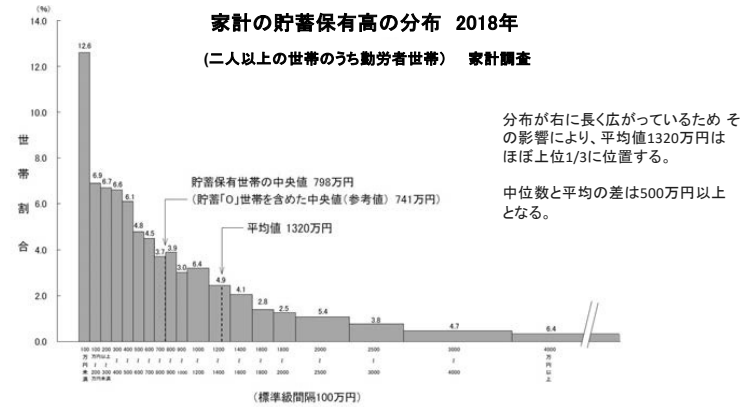
平均を超える都道府県はいくつあるか？東京、愛知、栃木、静岡、の**4つ**は平均以上です。滋賀がぎりぎり 3,214 に届かず、第 5 位で平均以下。あとの 42 都道府県は全部、平均以下です。だから、平均ですよって言われたら、すごくいいってことになるんです。何でそんなことが起きるかっていうと、**東京が突出して大きい**ので、これがものすごく平均値を引っ張り上げているんです。それで、大阪も神奈川も京都も、みんな平均以下になってしまう、ということです。



平均っていろいろ誤解の元じゃないのってことになります。こういう突出したデータがある時は、**中位数**を使います。**メディアン**とも言いますね。47 都道府県の県民所得を大きさの順に並べて 24 番目、つまり真ん中、第 24 位の県の数字、それがメディアンです。それは 2,896 の兵庫県です。

さっき統計は時々うそをつくぞって言ったけど、平均って結構、うそをつきます。

家庭の貯蓄保有高とは、要するにお宅に貯蓄・貯金って幾らありますかって聞いた調査です。そうすると、こういう分布になるんです。世帯の割合が縦軸で、横軸が貯蓄が幾らありますって数字です。見ると分かるんだけど、貯蓄 0 って人がすごく多いんです、20 年前には貯蓄 0 なんてほとんどなかったのに、今は貯蓄 0、100 万円未満しか貯蓄ありませんという



注）標準級間隔100万円（1000万円未満）の各階級の度数は縦軸目盛りと一致するが，1000万円以上の各階級の度数は階級の間隔が標準級間隔よりも広いため，縦軸目盛りとは一致しない。

総務省統計局HP

世帯がすごく増えています。

一方で、億万長者と言われる人たちはいるんです。こっちは際限がないんです。貯蓄なんていくらでもありますよみたいな人たちが、平均値をめっちゃくちゃ引っ張り上げます。なので、**平均値は 1,320 万円**です。皆さんは、ちょっと感覚的に難しいと思うんですけど、政府が、総務省統計局の統計として、日本の平均貯蓄は大体 1,320 万ですと発表するわけです。そうすると、みんながみんな怒るわけです。1,320 万円の貯金があるわけがないだろうと、そんな統計は捏造（ねつぞう）統計じゃないかと。日本はもっと生活、厳しい人が、多いんだぞ、いい加減にしろというふうに言われるんです。しかし、それは違います。平均以上のパーセンテージを足すと大体 3 分の 1 です。**平均値っていったら上位 3 分の 1 ってこと**です。**平均は真ん中ではありません**。平均値は億万長者の人たちが、引っ張り上げるから高くなるんです。

では、中位数、メディアンとも中央値ともいいますが、貯蓄 0 という人を外すと **798 万円**です。0 の人も含めて順位を取って真ん中を取ると **741 万円**です。750 万円ぐらいというと、「まあそんなもんじゃないかな」と思うわけです。「うちはそれよりちょっと下ぐらいだよ」っていう人もいるし、皆さんが納得するんです。実に、**メディアンと、平均値の差が 500 万円以上**あります。これはすごく大きいですよね。平均値ってのが真ん中だと思ってると、こういう誤解の元になります。これを使って、嘘をいろいろつくこともできます。統計を見ても、誤解っていうのが、そうやって生じちゃうんです。

比率の平均。大学入試で合格率を男女別に調べたら、男子が 56%で女子が 46%でした。この大学には学部が二つ、A、B あるんですけれど、男子が 56%で、男子のほうが高い。男子のほうが合格率高いってのは、何か男子が有利になるか、女子が不利になるようなことをしていませんか。それ、一回、調べてくださいということになりました。

大学全体では、女子が 1,000 人受けて 460 人合格、男子は 800 人受けて 450 人が合格しています。A 学部では、女子が 200 人受けて 140 人合格、B 学部は 800 人受けて 320 人が合格しています。そうすると 200 人と 800 人で合計 1,000 人受けて、合格したのは分子の 140 人と 320 人で 460 人です。 一方、男子は 800 人のうち 450 人合格していますが、受験したのは A 学部 700 人、B 学部 100 人、足して 800。合格したのは、A 学部 420 人、B 学部 30 人、足して 450 人でした。

A 学部と B 学部で男女の合格率を計算してみると、A 学部では、女子が 70%、男子 60%で、女子のほうが合格率が高かったです。B 学部は、女子が合格 40%、男子が合格 30%で、女子のほうが高いです。だから、どっちの学部も女子のほうが合格率は高いわけです。でも、大学全体では男子のほうが高いというふうになります。

変じゃないですか、こんなことが起きて。全体で見ると男子のほうが合格率が高いんだけど、学部別で見ると、どの学部でも女子のほうが高いんですね。これ、結構、錯覚するでしょう？ これは起きますよ、こういうこと

データを見る 比率の平均？？

ある大学の入試で合格率を男女別に調べると、男子が56%，女子が44%であった。
この大学では、女子の合格率が低く、優秀な女子が受験していないと考えた。
そこで、優秀な女子を受験させる方策を考えることになった。

この大学にはA,B二つの学部がある。学長はA学部、B学部にそれぞれ女子が入試で不利になっていないか調査を依頼した。

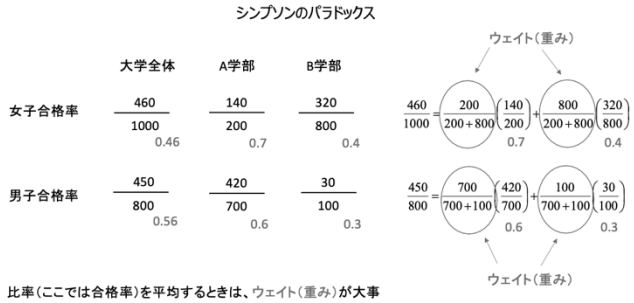
	大学全体	A学部	B学部	
女子合格率	$\frac{460}{1000}$ 0.46	$\frac{140}{200}$ 0.7	$\frac{320}{800}$ 0.4	$\frac{460}{1000}=\frac{200}{200+800}\left(\frac{140}{200}\right)+\frac{800}{200+800}\left(\frac{320}{800}\right)$
男子合格率	$\frac{450}{800}$ 0.56	$\frac{420}{700}$ 0.6	$\frac{30}{100}$ 0.3	$\frac{450}{800}=\frac{700}{700+100}\left(\frac{420}{700}\right)+\frac{100}{700+100}\left(\frac{30}{100}\right)$

が。こういうことは普通に起きます。起きるんですけど、ちょっと不思議ですね。

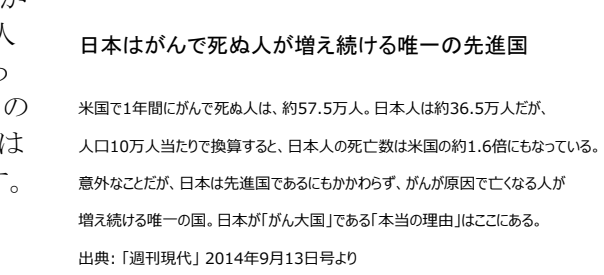
不思議だと思うのはなぜかというと、これは比率ですね。合格率だから、分母と分子がある分数です。0.7 っと 0.4 を平均して 0.46 という数字が出てます。ちょっと直感的に考えると、A 学部では 0.7、B 学部で 0.4 だったら、全体の合格率は、足して 2 で割って 0.5 ちょっとかなという感じじゃないですか。一方、こっちも、0.6 と 0.3 の平均だから 0.45 とか、そのぐらいじゃないかなと思っちゃうでしょう？ そうすると、全然違うんです。

比率の平均は、重み（ウエイト）を考えなきゃいけない。ウエイトって何かっていうと、例えば女子は、この 800 人が 200 人に比べてすごく多い。女子は B 学部を受ける人が多い。男子は A 学部を受ける人が多いんです。ウエートを考慮すると、男子の場合、どうしても A 学部に引っ張られるんです。平均すると 0.6 と 0.3 の間に来ることは間違いないです。女子のほうも 0.7 と 0.4 の間に平均が来ることは、そのとおりです。だけど真ん中、足して 2 で割るっていう値になるかっていうと、ならないです。ウエイトに引っ張られるので、男子は A 学部に引っ張られる、この 0.6 に近いほう。女子は B 学部の 800 が大きいので、この 0.4 に近いほうに数字が引っ張られて、こうなります。

式で書くと、こうなります。1,000 分の 460 っていう女子の合格率 0.46 は、こういうふうに分解できます。結局、0.7 と 0.4 を足して 2 で割ってないことは確かです。1,000 分の 200 と 1,000 分の 800 が、ウエイトで、これは受験者比率のことです。1,000 人のうち 200 人が A 学部を受験し、800 人が B 学部を受けましたっていうのがウエイトです。それがかかって全体を出してます。**比率を平均する時は、実はウエイトが大事**だっていうことです。比率の平均を出す時、平均は真ん中じゃなくて、どっちかに寄ってたりすることがあるのです。



日本はがんで死ぬ人が増え続ける唯一の先進国である。それじゃ、ここからもうちょっと本題に行きます。2014 年 9 月 13 日、「週刊現代」っていう週刊誌にこういう記事が載りました。日本はがんで死ぬ人が増え続ける唯一の先進国である。米国で、アメリカで 1 年間でがんで死ぬ人は 57.5 万人、日本人は 36.5 万人です。だけど人口が違いますから、人口 10 万人当たりで換算すると、実は日本人のがん死亡者の数は米国の 1.6 倍になります。意外なことだが、日本は先進国であるにもかかわらず、がんが原因で亡くなる人が増え続けている唯一の国である。日本ががん大国である本当の理由はここにある。日本の医療は駄目だっていう批判の記事です。


データから原因を探る？

日本はがんで死ぬ人が増え続ける唯一の先進国

米国で1年間にがんで死ぬ人は、約57.5万人。日本人は約36.5万人だが、
人口10万人当たりで換算すると、日本人の死亡数は米国の約1.6倍にもなっている。
意外なことだが、日本は先進国であるにもかかわらず、がんが原因で亡くなる人が
増え続ける唯一の国。日本が「がん大国」である「本当の理由」はここにある。
出典：「週刊現代」 2014年9月13日号より

日本の医療は駄目だ、医療は崩壊しつつあるというのが、この当時には、ニュース、週刊誌や、新聞でも書かれていました。そうしたら、名古屋大学の医学部の先生が怒って、そんなことは絶対にないって言うんです。日本の医療は世界最先端であるとおっしゃるんです。この数字は本当です。しかし、日本の医療が駄目だ、日本はがん大国だ、がんで死ぬ大国というのは違うんです。

どうして違うと思いますか？さあ、誰か？

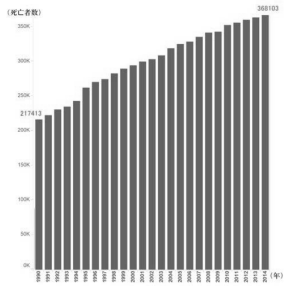
ちなみに、この例でなければ皆さん、気が付くと思う。日本ってどういう国だと思う？だけど、これだけ見せられると、ほんとに日本って大丈夫かって気がしませんか？　がんが増えてる日本はただ一つの先進国である。もうヨーロッパやアメリカでは、がんは減ってるよ、がんで死ぬ人って、もう減ってるよっていうことなんですけど、どうでしょうか？

これは悪用する人もいます。がんが治る特効薬があるといって、悪用している人がいます。

「がんが治る特効薬があるが、厚生労働省が製薬会社か何かに忖度（そんたく）して認可しない。だから、ヨーロッパやアメリカでは、みんながんは治ってるのに、日本のがん患者だけはこの薬を扱えないばっかりに死んでしまう。全くとんでもないことだ。だけど、私たちはひそかに輸入して、安い値段で皆さんに売ってあげることができますよ、買いませんか？」、という商売をネットでしている人がいます。

実は、がんによる死亡者数は増え続けています。1990 年から 2014 年にかけて、日本のあらゆるがんによる死亡者数は、21 万 7,413 人から 36 万人というふうに増えてます。これは市川衛っていう人のニュースブログです。ずっと増え続けてます。今でも増えてます。2022 年になっても、多分、増えて、増え続けてると思います。

人口動態統計による全がん死亡者数の推移（1990年～2014年）



市川衛　Yahoo!Japan　ニュース
2016年4月13日

何か思い付きませんか。日本の特徴は？

（生徒：高齢化？）

（生徒：日本は、すごい医療体制が整っているから、ちゃんと調べてがんが原因とかがわかると思うんですけど、海外とかでは、死亡した原因とかが調べないで・・・。）

ありがとうございます。それはあり得ますよね。日本はちゃんと、人口動態統計とここに書いてありますけど、もう徹底的な統計を取っていて、あらゆる人の死亡原因を記録して統計にしてるんです。これは確かに世

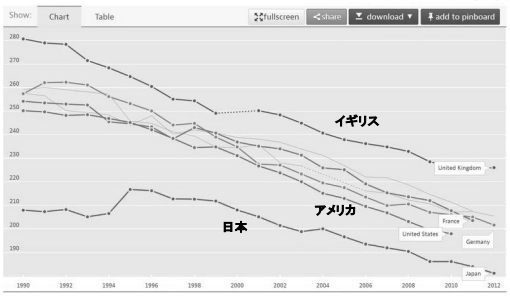
界に誇る統計なんですが、だからもう徹底的に、がんである、がん死亡者ってのを統計にしてるけど、他の国は結構スルーしてないか、死因原因不明とかでスルーしてるから、がんの死亡者って出てこないんじゃないかというのが、それはあり得ると思います。コロナなんかでも同じことが言えると思いますけど、日本は、やるとなったら徹底的にやりますから、それはあり得ます。ただ、がんについては、欧米各国も非常に重要だということで、それと、がんは割合、大きな病院で最期、亡くなる方が多いので、きちんと統計は取られてると思います。うん、まあ、そんなことがあるんですが。

日本って高齢化が進んでいるんですよね。がんって、やっぱり高齢者の亡くなる病気なんで、高齢者が増えるだけで死亡者が増えちゃうんです。人口 10 万人当たり、全がん年齢調整済み死亡率っていう、公衆衛生学の統計があるんですけど、年齢構成を修正して見てやると、人口高齢化の影響を除いてやると、日本はがんの死亡率って減ってるんです。人口当たりの死亡率が、年齢構成の要因を除いてやると減ってるんですね。しかもその死亡率の水準は、先進国の中で一番低い。

だから医学部の先生は怒るわけです。日本が一番いいんじゃないかと言って、医学部の先生は胸を張ります。日本は今、急激に高齢化が進行していて、これは他の国がついてこれないぐらいのスピードで、ものすごく増えてます。それが効いてますね。

G7 人口10万人あたり全がん年齢調整済み死亡率　(1990-2012)
OECD　Deaths from Cancer　https://data.oecd.org/healthstat/deaths-from-cancer.htm

市川衛　Yahoo!Japan ニュース
2016年4月13日



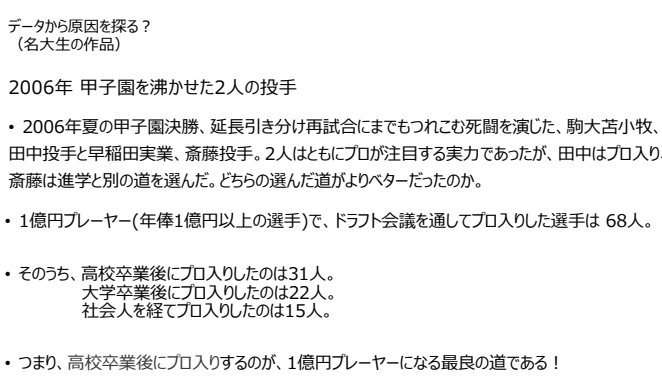
もちろん統計の漏れはあると思います。例えば、**日本はインフルエンザでの死亡者数が多い国**だって知ってます？　コロナになる前は、インフルエンザの死亡者数って、すごく多かったんです。もちろん病院でインフルエンザで亡くなると、インフルエンザで死亡しましたっていうふうに人口動態統計に表れるんですけれど、肺炎で亡くなっただけどインフルエンザかどうかは分かりません。コロナと違って徹底的に PCR 検査とかはしてませんから、肺炎で亡くなったお年寄りがいるんだけど、インフルエンザにかかっていたかどうかは検査しないから分かんないんです。

それを推計する方法がありまして、それが**超過死亡推計**っていう方法です。インフルエンザの時季に死亡者の数が増えてる、その差分を取って、この分がインフルエンザで亡くなった人であろうと推測する方法があるんです。日本は**国立感染症研究所**っていうのがありまして、インフルエンザの死亡者数の推計をしてます。この、超過死亡を推計する方法が、世界と違って日本だけなんですけど、**経済学の確率フロンティア分**

析という方法を使うんです。感染症研究所に経済学出身の人がいて、経済学の方法が採用されて使われています。

「データサイエンス」って、分野はクロスオーバーで、いいものはどこでも使うみたいなことになっています。

16 年前の甲子園で決勝戦、駒大苫小牧高校の田中投手と、早稲田実業の斎藤投手。甲子園で決勝戦で対戦したライバルなんですけれど、田中さんは高校卒業後、すぐプロに入る。斎藤さんは早稲田大学に進んで、その後、プロに入った。やっぱりプロで成功したのは、この田中投手なんですね。それ以来、よく言われてるんだけど、プロに行くんだったら



早く行ったほうがいいと。高卒でプロになったほうが早くプロになじんで成功する。大学に行くと、大学で試合に出て、ピッチャーだったら試合で投げてるうちに肩を壊してプロでは活躍できなくなる。という説があります。それを調べてみました。

「名大生の作品」って書いてありますが、実は 1 年生向けのセミナーで、統計でうそをつけ、うそをつく材料を持ってきて、うそつきプレゼンをやってくださいという課題を出しました。プレゼンを聞くのは名古屋大学の附属中学の生徒さんに来ていただきました。そこで、これをプレゼンしたんですけど、一発でばれました。

年俸 1 億円以上になった選手が何人いるか調べると、2006 年当時、68 人が 1 億円を超える年俸をもらえるような選手に成功しました。その 68 人の内訳は、高卒 31 人、大卒 22 人、社会人 15 人です。68 人のうち一番多いのは高卒で入った人です。社会人で入った人は 15 人しかいない、その倍ですよね、31 人。だから高卒でプロ入りしたほうがいい、確かにデータで証明されてます、ということなんですけれど。これは本当ですか。どうでしょうか。

こういうことです。1 億円プレーヤーは高卒が多いよといっても、元々、高卒って多いんです。高卒後にプロ入りした人って 317 人いて、社会人からプロに行った人って 179 人しかいないんです。大卒はその間の 253 人。**比率**を、見ないといけません。1 億円プレーヤーになった 68 人の内訳じゃなくて、高卒 317 人のうち、このうち何人、1 億円もらえるようになったかってことです。253 人のうち何人が稼げるようなったかってことです。比率を見てると、こうなります。高卒 317 人のうち 1 億円プレーヤーになったのは 9.8%。大卒 8.7%、社会人出身 8.4%です。というわけで、**あんまり差はない**ということですね。あんまり差はないです。

・ドラフト会議を通してプロ入りしたプロ野球選手のうち、

高校卒業後にプロ入りしたのは 317人
大学卒業後にプロ入りしたのは 253人
社会人を経てプロ入りしたのは 179人

・→ 高卒に占める1億円プレーヤーの割合は約 9.8%
大卒に占める1億円プレーヤーの割合は約 8.7%
社会人出身に占める1億円プレーヤーの割合は約 8.4%

ほとんど差はない！

少しの差はあるけれど　統計学を使って仮説検定を行うと、高卒、大卒、社会人卒の間で意味のある差（有意な差）は無いことがわかる。

9.8 と 8.7 と 8.4。でも、差はあるよね。9.8 と 8.7 は、1.1%ポイントというけど、やっぱり高卒のほうが高いでしょう？　やっぱり高卒のほうが少しは有利になるんじゃないの？って気がしちゃうわけです。9.8 と 8.7 っていうのは、これは違いがあると言っていいのかどうか。これ、統計で何か議論をする時、それが必ず問題になります。ぴったり 9.8 と 9.8 というふうに並べば、ああ、同じだねということになるんだけれど、ちょっとだけ差がある、これは意味があるのか、ないのかっていう話です。それで、これは意味がありますよっていうことを統計学で調べる。それが**仮説検定と呼ばれる方法**なんです。その方法を使うと、**意味のある差はない、有意な差はない**ことがわかります。

13 日の金曜日。オランダのロイター通信という通信会社が、オランダ発のニュースとして 2008 年 6 月 13 日に配信したニュースです。多くの国で不吉な日とされる 13 日の金曜日。実際は普通よりも安全な日であることが分かったと、オランダの統計学者が発表した。同国の保険統計センター CVS が発表したんですけど、13 日の金曜日は、他の金曜日と比べ、事故や火事、盗難の件数が少なかった。過去 2 年間でオランダの保険会社が受けた金曜日の交通事故報告件数は平均 7,800 件。一方、13 日の金曜日の平均は 7,500 件だった。CVS の統計学

データから原因を探る？

ロイター通信は2008年6月13日オランダ発で、次のようなニュースを配信した。

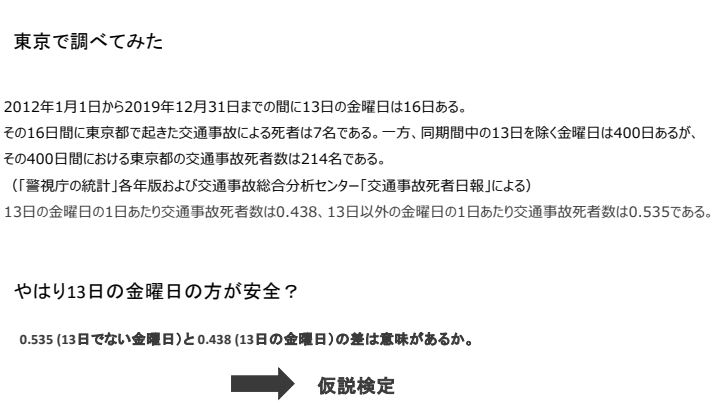
『多くの国で「不吉」な日とされる13日の金曜日だが、実際は通常よりも安全な日であることが、12日のオランダの統計学者らによる発表で分かった。同国の保険統計センターCVSによると13日の金曜日はほかの金曜日と比べ、事故や火事、盗難の件数が少なかった。過去2年間でオランダの保険会社が受けた金曜日の交通事故の報告件数は平均7800件。一方、13 日の金曜日の平均は7500件だった。CVSの統計学者Alex Hoen氏は保険業界誌に対し、事故などの件数が少ないのは人々が不吉なことを避けるために注意深くなったり外出を控えていることが原因とは考えにくいが、「 統計的には、13日の金曜日に運転をすることは（ほかの日と比べて）少し安全だ」と述べた。』

本当か？

者アレックス・ホーンという人が、事故などの件数が少ないのは人々が不吉なことを避けるために注意深くなったり外出を控えるからだ、統計的には 13 日の金曜日に運転することは、他の金曜日より安全であるという結論を出してます。まあ統計学者が言うんだから、この差はきっと有意なんだろうと思うんだけど、これはどうか。これはどうでしょうね。

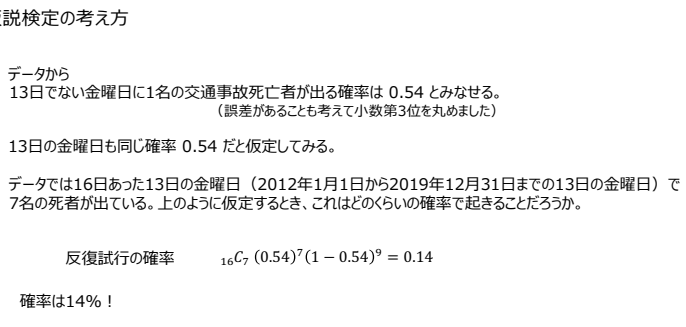
データの数は幾つでしょうか？ 2 年間で、13 日の金曜日は 2 年間に何日ある？ 13 日は 1 年に 12 回しかないでしょう？ 金曜日は、月火水木、7 つあるから、大ざっぱに考えて 7 分の 12 だとして、2 年間で **13 日の金曜日って 2 日か多くても 3 日くらいしかないよ**。それで 7,500 件だった。さすがに金曜日は 1 年間で 50 日ぐらいある。金曜日っていうのは 50 日ぐらいあって、2 年間取ってますから十分にあるんだけど、データとしては。13 日の金曜日に限ると、そんなにデータはないんです。そのデータで物を言うのは無理だと思う、ということです。

東京で調べてみました。



2012 年の 1 月 1 日から 2019 年の 12 月 31 日まで調べました。8 年間で **13 日の金曜日は 16 日**あります。その 16 日間に東京で起きた交通事故による**死者は 7 名**です。一方、同じ期間、この 2012 年から 2019 年までの 13 日を除く**金曜日の 400 日間**における東京都の交通事故死者数は **214 名**でした。1 日当たりにすると、13 日金曜日は 1 日当たりの交通事故死者 0.438、金曜日でない 13 日は、0.535 です。なので、やっぱり 13 日の金曜日は死亡者は少ない。だからちょっとだけ 13 日の金曜日のほうが安全だと言えるか、という話です。言えるかもしれない。数字で見て、1 日当たりの死亡者数は少ないので、言えるかもしれません。が、ここは統計学的に考えると、やっぱりこの差を仮説検定する必要があります。

仮説検定、考え方を紹介します。13 日でない金曜日に 1 名の交通事故死亡者が出る確率を 0.54 と見なしましょう。1 日当たり 0.54 の人が亡くなりますから、これを確率と見なします。そして、13 日の金曜日 1 日に 1 人の人が交通事故に遭って亡くなる確率も 0.54 と仮定してみましょう。つまり、13 日の金曜日も 13 日でない金曜日と同じだと仮定してみましょう。ここで仮定してみるということが重要で、仮にそうだと考えてみますということです。

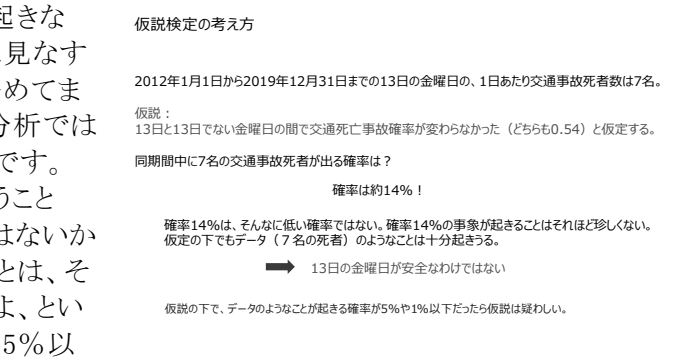


そうしたら、その仮定の下で、2012 年から 2019 年 12 月 31 日までの 13 日の金曜日に 7 名の人が亡くなっているんですけど、この 2012 年 1 月 1 日から 2019 年 12 月 31 日まで 7 年間に、7 名の人が亡くなる確率っていうのは、どのぐらいの確率で起きるんでしょうか。7 名の人が実際に亡くなってるんですけど、13 日の金曜日に。それは、もし死亡確率が 13 日の金曜日は、13 日でない金曜日と同じ危険度、同じ死亡確率だったとした時に、その 7 名の人が死ぬっていうのは、不自然なことなのか、いや、それはあり得ることなのかっていうことを計算しましょう、ということです。

計算ではこんなふうです。16 日で 7 人の死者が出る確率ですね、13 日の金曜日に、16 日間で 7 人の死亡者が出る確率です。これ、反復試行の確率とか独立試行の確率、というふうに言っていますが、数学の授業でもやるかもしれませんね。

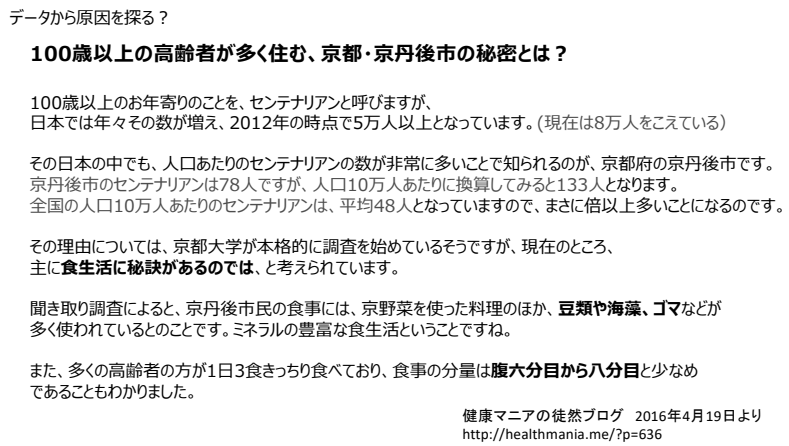
まず、13 日の金曜日も、13 日でない金曜日と危険は同じと仮定してみます。そうすると、過去に実際に起きたことが起きる確率ってのは、実はこういう計算で 14%だということが分かります。14%が高いか低いかなんですけど、14%の確率で起きることって、まあ普通に起きるというふうに考えるわけです。

何%以下だったら、それはもう起きない、あり得ないというふうに見なすかっていうのは、適当に決めてます。大体、統計学、統計分析では 5%か 1%か決めるぐらいです。5%以下、1%以下っていうことは、ちょっと起きないんではないかな。それより高い確率のことは、それは起きて不思議はないよ、というふうに考えます。確率が 5%以



下、1%以下だと、そんな低い確率でしか起きないことが起きたのかっていったら、それはちょっとおかしくないかと、考えます。その差に意味がある、有意という考え方につながるんです。これが、統計の仮説検定の考え方です。1 年生の人は学校でやりますから、ぜひ楽しみにしていてください。これが**仮説検定の考え方**です

次は、京都の京丹後市っていうところのブログです。2016 年 4 月 19 日とありますね。100 歳以上のお年寄りのことをセンテナリアンと言うんだそうですが、2012 年の時点で 5 万人以上、現在は 8 万人を超えてます。そ



データから原因を探る？

100歳以上の高齢者が多く住む、京都・京丹後市の秘密とは？

100歳以上のお年寄りのことを、センテナリアンと呼びますが、
日本では年々その数が増え、2012年の時点で5万人以上となっています。(現在は8万人をこえている)

その日本の中でも、人口あたりのセンテナリアンの数が非常に多いことで知られるのが、京都府の京丹後市です。
京丹後市のセンテナリアンは78人ですが、人口10万人あたりに換算してみると133人となります。
全国の人口10万人あたりのセンテナリアンは、平均48人となっていますので、まさに倍以上多いことになるのです。

その理由については、京都大学が本格的に調査を始めているそうですが、現在のところ、
主に**食生活に秘訣があるのでは**、と考えられています。

聞き取り調査によると、京丹後市民の食事には、京野菜を使った料理のほか、**豆類や海藻、ゴマ**などが
多く使われているとのことです。ミネラルの豊富な食生活ということですね。

また、多くの高齢者の方が1日3食きっちり食べており、食事の分量は**腹六分目から八分目**と少なめ
であることもわかりました。

健康マニアの徒然ブログ 2016年4月19日より
http://healthmania.me/?p=636

の中でも、人口当たりのセンテナリアンの数が非常に多いことで知られるのが京丹後市。京丹後市は 78 人、100 歳以上のご老人がおられるんですけど、人口 10 万人当たりに換算してみると 133 人になります。全国は、10 万人当たりの平均は 48 人だから、京丹後はすごく多い。その理由について京都大学が本格的に調査をしてるんですけれど、食べ物がいいんじゃないか、京野菜を使った料理、豆、海藻、ゴマなんかを食べてるんで、いいんじゃないかなということです。腹六分目から八分目と少なめもいいよ、ということです。

お年寄りが長生きするのに適した環境を発見したら、そこの食生活とか環境や特徴を調べて、それを他の地域でもそういう環境や食生活を実現したら、もっとお年寄りは長生きできるんではないか、こういうことです。これは要するに、100 歳以上の人たちが長生きしている地域が、その原因を調べて、その原因を利用して新たな価値を生み出すという、まさに一番最初に言った、データから価値を生み出すということが、これでできるわけです。皆さん、豆、食べましょう、海藻、ゴマもいいですよ、腹六分目から八分目にしましょう、こういうことができるんですけど。

でも、京丹後市って、本当にこのデータから、お年寄りが長生きするのに快適な、お年寄りが長生きできる環境なのかどうかです。それはほんとか？っていうことですね。これ、ぱっと見て、何か思い付きます、皆さん？

なかなか難しいんですけど、実は、10 万人当たりの 100 歳以上人口が多いって言ってんだけど、それって若者がいないっていうだけじゃないか、という可能性もあるんだね。人口当たり 100 歳以上の人が多いよっていうのは、お年寄りに優しいからそうなってるのか、お年寄りにとって快適だからそうなってるのか、いや、とんでもない田舎で、若者がどんどん逃げていってしまって、もう働ける人はみんな出ていってしまった。最近は、日本は 70 代の人も働いてますから、後期高齢者、75 歳ぐらいから上ぐらいの人しかいない。だから 100 歳以上人口が多いんだ。という可能性も、ないことはないんじゃないですか。分かります？分母と分子があって、100 歳以上人口というのは分子になります。分子が大きいんだけど、比率だから、実は分母が小さいから大きく見えてるだけじゃないか。若者どころか、働く人はみんな出ていってしまうような田舎なんじゃないか、ということなんです。

高齢者が多いのではなくて、若者が少ないのでは？

若者が経済的に豊かな地域に流出してしまうのでは？　その結果、高齢者比率が高くなる。
お年寄りが暮らしやすい。長生きしやすいというのとは違うのでは？

経済学的豊かさ ➡ 一人あたり県民所得で測る。

都道府県別データで
人口10万人あたり100歳以上人口比率 と 一人あたり県民所得 の相関を調べてみる。

そこで、相関をとってみましょう。皆さん、**相関係数**っていうのは、やりました？これ数 I でやるんだよっていうんだけど、学校によるでしょうね。1 学期で教えてもらったかどうかなんですけど、まあいいや。相関係数の説明も後でしてるんですけど、相関というものを測ってみましょう。経済的豊かさと、この 100 歳以上人口って関係があるのか。100 歳以上人口比率と、さっきの 1 人当たり県民所得の相関を見てみましょうと、47 都道府県で。1 人当たり県民所得ってのは豊かさ、経済的豊かさですから、こういうところは何か若者が集まってきそう。それを見てみると、こんな感じになります。

100歳以上人口　10万人あたり
都道府県ランキング

2018年9月　厚生労働省

単位：人

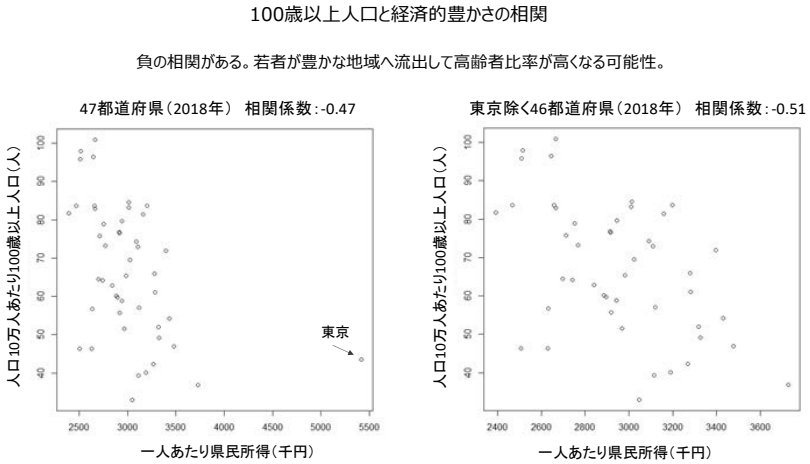
1	島根県	101	25	秋田県	65
	鳥取県	98		北海道	64
	高知県	97		岩手県	63
	鹿児島県	96		群馬県	61
5	香川県	85		福岡県	60
	山口県	84	30	山形県	60
	宮崎県	84		福島県	59
	愛媛県	84		三重県	57
	長野県	83		奈良県	57
10	熊本県	83		岐阜県	56
	沖縄県	82	35	静岡県	54 ←
	山梨県	81		滋賀県	52 ←
	長崎県	80		兵庫県	52
	佐賀県	79		茨城県	49
15	和歌山県	77		栃木県	47 ←
	新潟県	77	40	宮城県	46
	大分県	76		青森県	46
	徳島県	74		東京都	44 ←
	岡山県	73		神奈川県	42
20	広島県	73		大阪府	40
	富山県	72	45	千葉県	39
	石川県	70		愛知県	37 ←
	福井県	66		埼玉県	33
	京都府	65			

2018 年の 100 歳以上人口のランキングですね、これは。さっき、1 人当たり県民所得って、平均以上が 4 県、言いましたが、愛知県、東京都、栃木県、それに静岡県、滋賀県。こういうところ、平均以上の 4 県と平均水準の 5 県で見ると、100 歳以上人口比率ランキングでいうと、下のほうばっかりですよね。若者が多いから、100 歳以上人口比率が大きくならないんじゃないか。逆に 100 歳以上人口比率が高いのは、島根

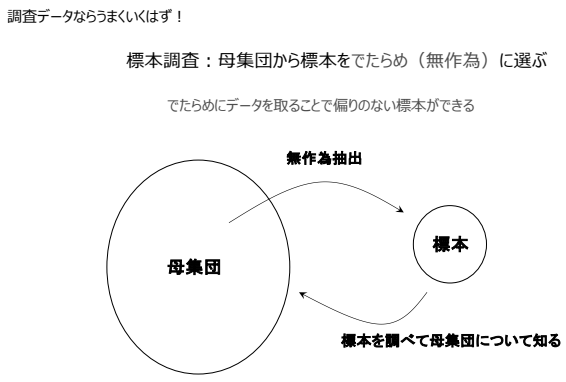
県、鳥取県。これ、島根県は 10 万人当たり 100 人のお年寄りが 100 歳以上。島根県、鳥取県、高知県、鹿児島県っていうとこですね。
経済的な豊かさと 100 歳以上人口の相関がそういうふうにあるんだとすると、実は京丹後市が、100 歳以上のお年寄りがすごく多いといっても、それはやっぱり若者がいなくなっちゃうからではないかなという疑問が起きるんですけれど。

これがプロット図、散布図というのかな、プロット図にしてみると、こういうふうになります。縦軸に人口 10 万人当たりの人口を取り、1 人当たり県民所得を横軸で取ります。

相関係数がマイナス 0.47 になります。だからやっぱり負の相関があるというふうに見える。負の相関っていうのは、片方が増えた時に片方が減ると負の相関になる。片方が増えた時にもう一方も増えていくぞという時には正の相関があるというふうに言います。これを見ると、全体にこういうふうに減ってる感じだから、やっぱり負の相関はあるんだなっていうことです。



いままではいろいろな**データを観察**してきました。**観察データ**ということは受け身だってことです。現れてくるデータを受け取って眺めているのです。一方で、データを取りに行くこともできます。積極的に調査するということです。調査するときは、無作為抽出ということをします。無作為とはでたらめに、ということです。調べたい対象すべての集合を母集団と言います。たとえば日本人について知りたければ、調査する日に生きている日本人全部が母集団です。その一部の人を調査するのですが、調査対象の人たちが偏らないように、たとえば高齢者ばかりだったり、女性ばかりだったりしないようにする必要あります。無作為抽出はでたらめに選ぶということですが、たとえばくじで調査する人を選べば偏ることはありません。調査対象に選んだ人々のことを標本といいますが、でたらめに選ぶことで、標本は母集団の一部で母集団の縮小コピーのようになります。こうしてデータを積極に取れば、知りたいことが正しくわかるのではないでしょうか。



ところが、でたらめに選ぶ、**無作為抽出**をするということは決して**簡単ではありません**。次の例はマーチン・ガードナーという人が作ったパズルです。ある中学校で、その学校の生徒の一部をでたらめに選ぶ、つまり無作為抽出して、数学の成績と足の長さを調べたら正の相関があったというのです。これ、どう思いますか？そんなことってありでしょうか。あるかないか以前に、少し考えればわかると思いますが、無作為抽出の仕方が間違っています。中学だから 3 学年あるわけです。でたらめに生徒を選んだら、1 年生、2 年生と 3 年生が、だいたい 1/3 ずつ選ばれます。1 年生より 2 年生、2 年生より 3 年生の方が体も大きいし、足の長さも長いでしょう。もちろん人によりますけど、全体の傾向としてそうなります。それに上の学年ほど、勉強した数学の範囲も広いのだから、同じ問題でテストすれば 3 年生は 2 年生より、2 年生は 1 年生より成績が良いでしょう。だから足が長いほど数学もできるように見えてしまうのですね。ちょっと同じ問題でテストしたっていうところがわかりにくかったかもしれませんが…で、この場合はどうし

マーチン・ガードナー“数学ゲーム” より

「ある中学校で男子生徒100名をでたらめに選び、足の長さを測ったところ
足の長い生徒ほど数学の成績が良いということがわかった。足が長いと数学ができる。」

せめて学年を固定してくれ。

初歩的ミスです。

たらよかったかというと、学年を決めて生徒をでたらめに選べばよかったのです。それが正しい方法でした。

では次に、実際に NHK が行っている**世論調査**の例を見てみましょう。2014 年の安倍内閣の指持率の調査です。日本全国の 20 歳以上の男女を対象にした電話による聞き取り調査です。無作為抽出をしてます。電話調査ですから、簡単にやろうと思えば、なにかの会員名簿に載っている電話番号、アイドルのファンクラブとか、ゴルフ場の会員名簿とか、老人クラブの連合会とか。こんな名簿使っていたらいくら無作為抽出しても意味ありません。で、**RDD 法**を使います。RDD 法はでたらめに数値を発生させて何桁かの番号を作り、機械的に電話をかけます。でたらめに作った数値を乱数といいます。でたらめに作るので、多くの場合、おかけになった番号は現在使われておりません、となりますが、どんどんかけ続ければつながります。携帯電話にも固定電話にもつながります。NHK の調査は必ず週末を含んでいますね。週末でないと働いている人は職場にいます。職場だとつながっても仕事中なので、調査に協力してくれないでしょう。だから週末みんなが家にいるときをねらって調査します。そこまでしても偏りは完全にはなくならないです。答えたくない人は電話切ってしまいますから。それでも、RDD 法は他の方法に比べれば調査対象に偏りの出ない信頼できる方法です。ネット調査と比べてください。ネットは、積極的に答えたい人、興味があってサイトを訪れてくる人だけが回答します。特に 12 月の調査では大きな差が出てしまうことがわかります。

人間を調査することは難しい

安倍内閣支持率　電話調査 vs ネット調査

NHK放送文科研究所「政治意識月例調査」	Yahoo!ニュース意識調査
対象：全国の20歳以上の男女 調査方法：電話調査（RDD法）	
期間：2014年3月7日（金）～3月9日(日) 回答数：1028人（回答率63.1%） 安倍内閣を支持する：51% 支持しない：30%	期間：2014年3月26日（水）～3月28日（金） 回答数：22904　（男性79.8%，女性20.2%） 安倍内閣を支持する：55.3% 支持しない：41.2%
期間：2014年12月6日（金）～12月8日(日) 回答数：1055人（回答率64.4%） 安倍内閣を支持する：50% 支持しない：35%	期間：2014年12月26日（木）～12月28日（土） 回答数：96330　（男性75.5%，女性24.4%） 安倍内閣を支持する：82.1% 支持しない：16.7%

人間を調査することは難しい

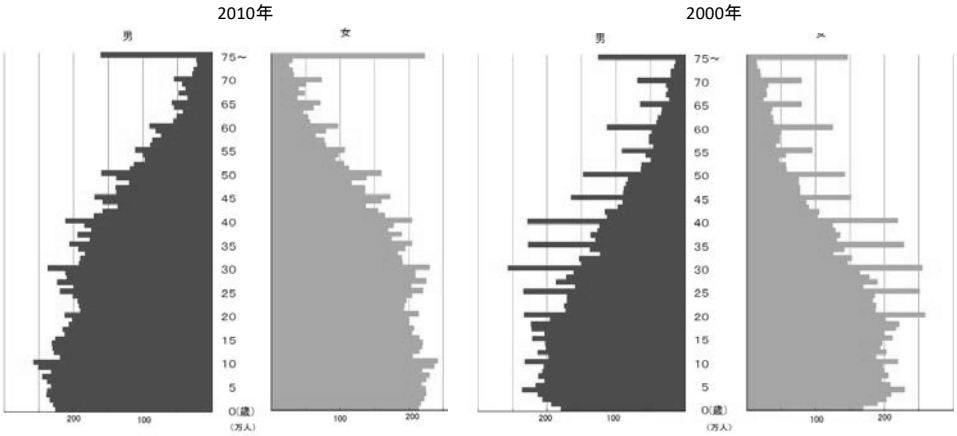
▶ ネット調査は積極的に答えたい人だけが調査対象になる。
同一人の複数回回答もあり得る。

しかし利便性が高く、少ないコストで実施できる。

▶ 電話調査も問題は残る。
RDD法ではでは電話をかける番号は機械によってランダム に作成される。このため、対象の抽出には調査員の主観が入り込む余地がなく、文字どおり無作為（でたらめ）になる。

しかし、電話が固定電話だと… 年齢や性別に偏りが出るおそれがある
例えば、日中の昼間に電話に出るのは主婦や仕事を離れた高齢者が多いなど

人間を調査することは難しい　　インドネシアの人口ピラミッドとAge Heaping
総務省統計研修所　西 文彦　2013年7月24日

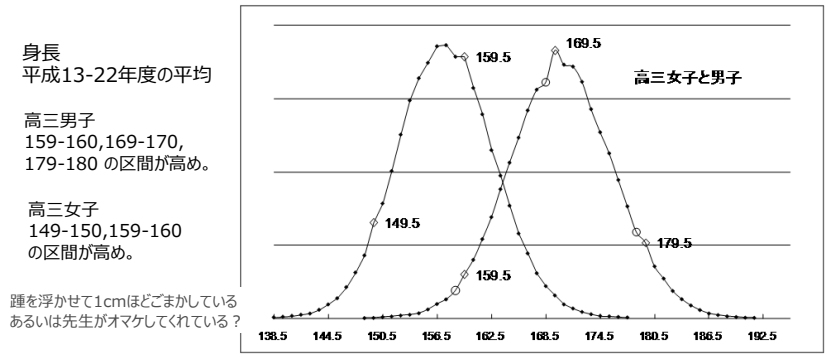


調査対象を無作為に選び、選んだ人たちが回答してくれても、正しいことがわかるとは限りません。**正しく答えてくれるかどうかはわからない**のです。自然現象ではなく人を調査する難しさですね。**インドネシアの人口ピラミッド**を見ましょう。人口ピラミッドっていうのは年齢別の人口の棒グラフです。下から上へ行くと年齢が上がります。右のオレンジ色の棒が女性、左の緑色の棒が男性の人数を表します。下の方が広がっているのは、インドネシアは若い人の方が人数多いんですね。日本の人口ピラミッドって、見たことありませんか。高齢者が多くて少子化で、上が広く下が細いです。ピラミッドになっていませんね。逆ピラミッドです。それはともかく、インドネシアの人口ピラミッド、年齢によって飛び出しているところがありますよね。これ何かというと、インドネシアの人は年齢にこだわりがなくて、自分の年齢をきりのいい数字に丸めて答えちゃっているのです。たとえば、だいたい 25 歳とか 30 歳とか。これは**年齢ヒープ**と呼ばれている現象です。2000 年のグラフと 2010 年のグラフを比べると、2010 年の方が年齢ヒープが少し緩和されています。調査しているインドネシアの統計庁が、なんとか正確な年齢を聞き出そう質問票を工夫した成果なのだそうです。

こんなこと、日本での調査では起こらない、と思うかもしれません。いえいえ、日本の**学校保健統計**で小中高校生の身長と体重を調査していますが、これは**高 3 生の男女別身長の分布**です。きれいな左右対称の分布曲線ですが、よく見るとところどころに凹凸があります。男子は 159cm より大きく 160cm 以下、169cm より大きく 170cm 以下、179cm より大きく 180cm 以下のところが飛び出していて、多くなっていることがわかります。女子も同様で、149cm より大きく 150cm 以下、159cm より大きく 160cm 以下のところが凸になっています。一年だけのデータだと、たまたまということもあるかもしれませんが、平成 13 年度から 22 年度の平均を取っていますので、こんな凹凸ができるのは不思議です。なぜでしょうか？一つの可能性として、本当は 149cm、159cm、169cm等々なんだ

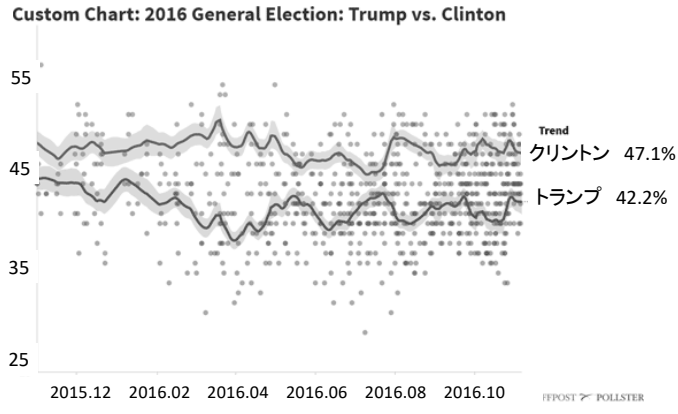
人間を調査することは難しい

森棟公夫「学校保健統計調査の不思議」『統計学会報』148号　2011年7月 より



けれど、測定している先生がオマケして 150cm、160cm、170cm にしてあげているのではないでしょうか。あるいは、本人が頑張って背伸びしているのかもしれません。

最後の例は、**アメリカの大統領選挙**です。2016 年の大統領選挙で、民主党のクリントン候補と共和党のトランプ候補が争った選挙です。投票は 2016 年の 11 月でした。グラフは 2015 年から投票直前まで、どちらの候補に投票するかを尋ねたたくさんの**世論調査**の結果です。グラフの中の一つ一つの点が一つの調査に対応します。青い点がクリントンに投票すると回答した人の割合で、赤い点がトランプに投票すると回答した人の割合です。青い線は各時点ごとの青い点の平均です。つまりクリントンに投票すると回答した人の割合の平均の推移、同じように、赤い線はトランプに投票すると回答した人の割合の平均の推移です。青い点の方が赤い点より上に分布していることがわかると思います。平均の線で見ても、青い線は赤い線よりも上にあります。接近したことあっても、青い線が赤い線の下になったことは一度もありません。つまり、常に、クリントン候補の方がトランプ候補よりも優勢でした。ところが、選挙ではトランプ候補が当選して大統領になりました。世論調査は、投票の 1 年前から直前まで、一貫して大間違いだったのです。どうしてこんなことになってしまうのでしょうか。その理由は政治学の研究者をはじめ、いろいろな分野の専門家が研究していると思いますが、一つ考えられることは、人々が本当のことを回答していない可能性です。



それについて、**ブラッドリー効果**というものがあります。1982 年のカリフォルニア州知事選挙のときのことです。黒人のブラッドリーと白人のデュークメジアンが候補でした。選挙前の世論調査では、ブラッドリーが優勢でしたが、選挙結果はデュークメジアンの勝利でした。白人有権者の多くがデュークメジアンに投票したといわれています。この原因として、世論調査の質問に対して白人であるデュークメジアン候補に投票すると回答することには、人種差別主義者とみなされてしまう不安がある。それを嫌った人達がブラッドリーに投票すると答えたのではないか、というものです。このように回答者が本当のことを言ってくれない現象を、ブラッドリー効果と呼んでいるのだそうです。まさに、トランプとクリントンの大統領選でも、ブラッドリー効果が観察されたのかもしれません。

ブラッドリー効果

ブラッドリー効果（Bradley effect）は、選挙において非白人候補者の得票率が世論調査を下回るとされる説である。

1982年のカリフォルニア州知事選挙で黒人の元ロサンゼルス市長トム・ブラッドリーが白人の共和党候補ジョージ・デュークメジアンと争った。事前に行われた世論調査ではブラッドリーが圧倒的有利な状態で、ほとんどのメディアはブラッドリーの勝利を予想し、サンフランシスコクロニクルは「BRADLEY WIN PROJECTED」の見出しをかかげた。しかし、いざ選挙当日になってみると、それまでブラッドリーを支持していた白人有権者がデュークメジアンに投票し、多くの票がデュークメジアンに流れた結果、当選確実といわれていたブラッドリーは敗れてしまった。これは、白人に投票すると言う意見の表明自体が、調査者に人種差別主義的イメージを以て解されるのを嫌った一部の人が、「ブラッドリーに投票する」と世論調査で答えた結果だと社会心理学的な解釈が行われている[1]。

出典: フリー百科事典『ウィキペディア（Wikipedia）』

以上、きょうはデータが嘘をついてしまう例を見てきました。平均という日常よく使っている指標の落とし穴、何かと何かを比較するときその背後にある条件の違い、たとえば日本と欧米各国の高齢化比率の違いなどですね。それから違いがあってもそれに意味があるか、つまり統計的な有意性のこと、さらに無作為抽出をしている調査データでも、人間を相手にすると難しい問題が出てくることなど、お話ししました。こうしたさまざまな問題をクリアしてはじめて、データから正しい情報を引き出すことができるのです。ただデータがあるから、それを見れば直ちに正しいことがわかる、というのではありません。ただ、このような事例ばかり見せられると、データを使うのはこわいことだ、ちょっと間違うとデータは嘘ばかりついてしまう、と不安になるかもしれません。でも実はここから先に、データ分析がいかに役に立つか、いかに新しい価値が見つかるか、というデータサイエンスの世界が広がっていきます。きょうはそこまでお話しできませんでしたが、これをきっかけにデータに興味を持ってもらえたら、データサイエンスの入口はもうそんなに遠くありません。データはどの分野でも大事です。自身の興味のある分野で、たとえばスポーツなんかデータの宝庫ですけれど、そこからどんなことがいえるのか、あるいはいえないのか、立ち止まって考えてみるとさらに興味が広がると思います。

〈この講義録について〉

「学びの杜・学術コース」は、本学の教員を中心とする研究者が、各専門領域における大学レベルの学びの機会を高校生に提供する無料公開講座である。本学教育発達科学研究科附属高大接続研究センターおよび前身の中等教育研究センターが毎年夏休み期間に開催し、2005 年の開始から 18 年目を迎えた。参加生徒が知の探究のたのしさ厳しさにふれることで自己の興味関心を内省し、将来のキャリアデザインにつなげてゆくことを目的としている。

2022 年度は 23 名の講師による 20 の講義が行われ、高校 1〜3 年生の 83 名が参加した。本講義録は、8 月 26 日に開催された根本二郎教授の講義「統計分析で嘘をつかないためにーデータの偏りと見せかけの相関ー」の録音データを文字化して再構成したものである。