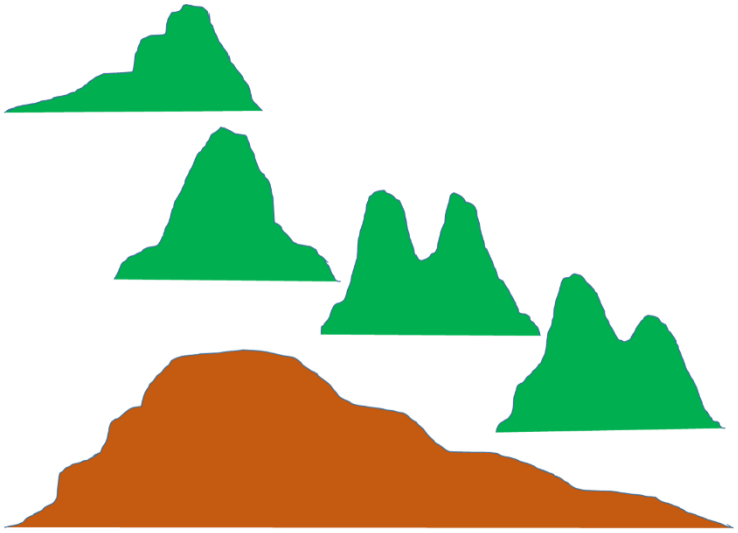
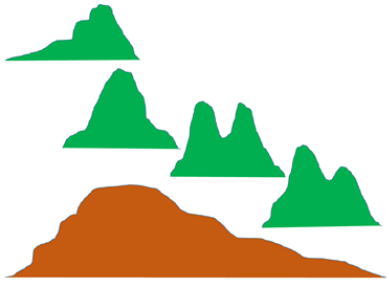
統計学

COVID-19 禍のもとでのオンデマンド授業



日本赤十字九州国際看護大学

守山正樹



目　次

https://jrckicn.repo.nii.ac.jp/?action=repository_opensearch&index_id=49

前書き

自分の専門（ヘルスプロモーション、公衆衛生）とは異なる統計学という科目を教え始めて４年目になります。統計学の教科書や参考書は、看護や保健の分野でも様々なものが出版されています。大学には情報処理実習室もあり、数字が苦手な学生もパソコンの助けを借りると、比較的単純なキー操作でグラフや傾向線を描き、統計的な検定までも行うことができます。教科書や実習室の助けを借りて、これまでは何とか教えることができました。

しかし 2020 年４月、新型コロナウイルス COVID-19 の流行により、大学構内が立ち入り禁止となり、全授業を、対面授業から「自宅にいる学生たちに対して動画を配信する形のオンデマンド遠隔授業」へと切り替えることになりました。３年かけてやっと作り上げた対面授業の流れを、ゼロから考え直さねばならなくなりました。一体どうしたらいいのでしょうか。頭が真っ白になりました。しかし逃げるわけにはいきません。

そして 2020 年５月の連休明けからの遠隔授業に向けて、４月中旬から長いトンネルの中を手探りで歩むような日々が始まりました。その後の３ヵ月間を何とか乗り越えたのは「改めて、どう教えようか？」と問い続ける中で、少しずつ見えて来た統計学の面白さと、それに応えてくれた学生たちの前向きな反応だったように感じています。

統計学マイクロレクチャー　日本赤十字九州国際看護大学リポジトリ URL

https://jrckicn.repo.nii.ac.jp/?action=repository_opensearch&index_id=49

この授業の動画は以下の URL からダウンロードできます。

st01-確率分布　http://id.nii.ac.jp/1127/00000686/

st02-分数と集計　http://id.nii.ac.jp/1127/00000687/

st03-平均と標準偏差　http://id.nii.ac.jp/1127/00000688/

st04-回帰と相関　http://id.nii.ac.jp/1127/00000689/

st05-クロス集計表と行%　http://id.nii.ac.jp/1127/00000690/

st06-2X2表とカイニ乗検定　http://id.nii.ac.jp/1127/00000691/

st07-統計的仮説検定　http://id.nii.ac.jp/1127/00000692/

st09-12 名のデータでも統計学のレポートが書ける　http://id.nii.ac.jp/1127/00000693/

st10-my 標本からクラス全体のデータへ　http://id.nii.ac.jp/1127/00000694/

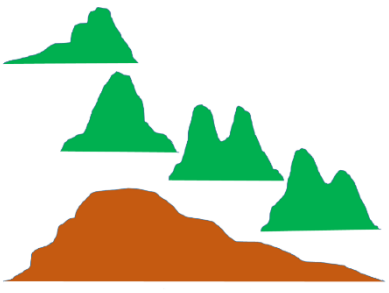
st11-ニ群の比較とt検定　http://id.nii.ac.jp/1127/00000695/

st12(1)-分散分析と F 検定 1　http://id.nii.ac.jp/1127/00000696/

st12(2)-分散分析と F 検定 2　http://id.nii.ac.jp/1127/00000697/

st13-回帰分析　http://id.nii.ac.jp/1127/00000698/

st14-主観と統計　http://id.nii.ac.jp/1127/00000699/



第１章　統計と確率分布

https://youtu.be/vhZk431oWUU

皆さんこんにちは。私が始めて皆さんにお会いしたのは昨年(2019)６月です。覚えていますか。基礎力総合ゼミの時間に対話型の質問系列 Wify(What is important for you?)、さらにプチプチを用いて感覚的な問題提起をしました。さて今日からは、統計学の勉強を始めます。統計学は何かを測定・観察し、結果を数値で表し、集団や社会について考えていく科学です。

1 考え方の特徴

統計学に特徴的な考え方とは何でしょうか。統計学はこの世界の様々な事象を数値で表した上で、事象が「ランダムに・偶然に・確率的に」起きると考え、その起き方の法則を追及します。「全ての事象が偶然性・確率に支配されている」と聞くと「本当？」と疑う人もいるかもしれません。「私自身や私が住む世界は今ここに現存する；私は確率的な存在じゃない！；私も世界も偶然ではなく必然です」と思う人もいるでしょう。他方、今の新型コロナウイルス（COVID-19）の突然の流行を振り返ると「人類とウイルスの偶然の出会い」やその後の「ウイルスの確率的な変異」が世界を揺り動かしていることも事実です。ですから「偶然性・確率を基礎とする統計学の考え方」は今の激動する世界を生きる上でとても大切です。その統計学の中心になるのが、事象を数値（離散量または連続量）で捉え、それらの起こる確率の拡がりを数学的に把握する確率分布の考え方です。

2 確率変数

統計学ではこの世界の「確率的に起こる事象 E」は「①それが取り得る様々な数値（変数；x）を用いて Ex と表せる、②そうした数値には、その数値が現れる確率 P（Ex)が対応する」と考え、その変数（ｘ）を「確率変数」といいます。確率変数は、試験合否やコイン裏表のように 0,1,2 など自然数で表わせる離散型確率変数（離散量）と、身長・体重のように小数点以下何桁までも連続する値で表わせる連続型確率変数（連続量）に分かれます。

この世界の事象→Ex　　ｘ→離散量（離散型確率変数）、連続量（連続型確率変数）

１）　離散型確率変数（離散量）

その値が 1 と 2 だけとか、とびとびの値のみを取り、間の値をとることがない変数が離散量です。このコインには裏と表しかありません。コインを投げると、真ん中で止まることはなく、必ず裏か表が出ます。今度はサイコロです。サイコロは６面があり、投げると１から６のどれかが上になります。1.5 といった値はとりません。これが離散量です。

事象（出来事）がコインの裏表のように、互いに排反する２項目しかない離散量は、私たちの毎日でも、看護や医療の世界でも広く出て来ます。試験の合格／不合格、ヒトの生死、疾患の有無は離散量です。皆さんの健康チェックでも離散量が活躍します；喉の痛みの有無、嗅覚の異常の有無、37.5 度 C 以上の発熱の有無、これらも離散量です。以前は性別も、男性か女性か二つの値の離散量でした。現在は二つ以上の値を持つ離散量と位置付けられます。

２）　連続型確率変数（連続量）

サイコロの目のような、とびとびの値しかとらない離散量に対して、小数点以下何桁までも表すことができるデータが連続量です。

例えば私の手の人差し指の長さを計ってみると 7.1cm あります。7.1cm は連続量ですから７cm と８ｃｍの間にあり、小数点以下をもっと精密に測定しようとすると、理論的には無限に細かくすることが可能です。皆さんの場合はどうでしょうか。自分の指の長さを測り、連続量として表してみてください。自分の体だけでも様々な連続量を見出すことができます。挙げてみてください。

３確率分布

１）事例から確率分布へ

次は確率分布についてお話しします。分布とは「複数の事象が、ある広がりを持って存在するとき、その広がり」を示します。何かが１回・１例だけ存在する事例の場合、分布という考え方は用いません。まず単独事例と分布について、考え方の違いを説明します。

「コインを投げて裏がでた、電車の最初の乗客が女性だった、ある学生の身長を測ったら 160.0 cm だった」などは単独の事例です。こうした事例を出発点として、詳しく聴き取り、記述を大切に進める事例研究は、看護でもよく用いられる研究方法です。一方、統計学で注目するのは、母集団における平均的事象／平均的個体です。１例だけで「コインは裏が出やすい、電車の乗客は女性が多い、学生の身長は 160.0cm！」とは結論しません。２回・３回・４回～ｎ回と投げる試み（試行）、観察の試行を繰り返し重ねることで、初めて「コインは裏と表が同じ確率 0.5 で出る」「電車の乗客は 60%が女性だ」「学生の身長は平均 162.0cm」などと結論できます。

サイコロ（離散量）であれば投げる試み（試行）を繰り返し、身長や体重（連続量）であれば、一人二人と測る試みを増やすことで、いくつもの値が得られ、全体の広がり・分布が見えてきます。それが確率分布です。

２）　離散量に対応する確率分布

離散量は事象（出来事）の起こり方から得られますが、起こり方は一種類ではありません。起こり方に対応して、ここでは二つの確率分布を示します。二項分布とポアソン分布です。

(1) 二項分布

コインの裏表、生死、疾患の有無など、取り得る場合が２項目しかない離散量に対応する確率分布が二項分布です。

コイン投げを例にとります。２回・３回・４回～ｎ回とコインを投げる試み（試行）を増やし、N 回振って、何回表が出たかを横軸に、またその確率を縦軸にとってヒストグラム（棒グラフの一種）を描くと、山の形の確率分布が現れます。これを二項分布といいます。

二項分布の例としては、コイン投げの他に、視聴行動（ある番組を見ない０、見た１）、投票行動（ある候補に投票しない０、投票する１）、治療効果（ある治療が効かない０、効く１）などがあり、何れも選択（どちらかを選ぶ行為）に関連しています。

二項分布を描いてみましょう。以下はアメリカのアイオワ大学によるウェブサイトです。ｎ＝生起確率ｐと試行回数ｎを入力すると、対応する二項分布のグラフを描いてくれます。たとえば、

ある治療が効く確率ｐを 0.6、その治療を試した患者さんの数ｎを 20 などと入力して、その条件に合わせた二項分布を描いてみましょう。

https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html

(2) ポアソン分布

19 世紀に活躍したフランス人の数学者、シメオン・デニス・ポアソンは、ラバ蹴られて亡くなったフランス軍の死亡者の発生の確率分布を研究し、1837 年にポアソン分布を発表しました。事故で亡くなる人の発生は、「どちらかを選ぶ際の離散量」ではなく「自然現象が発生する際の離散量」です。「ある時間内やある領域内で、ときどき発生する自然現象の回数」から求められるのがポアソン分布です。ポアソン分布の例としては「時間内の来客数・来院者数」「時間内の電話相談件数」「空気の体積当たりの特定のウイルス数」などが考えられます。

ポアソン分布もアイオワ大学のウェブサイトで描けます。ポアソン分布は試行回数ｎが十分に大きく、また生起確率 ｐ が非常に小さいときに導かれる二項分布の極限と考えられます。ポアソン分布を計算するときは、ｎとｐとを掛け算した値（λ；ｎ×ｐ）が大切です。このλの値を指定すると、アイオワ大学のウェブサイトから、ポアソン分布のグラフを描けます。

https://homepage.divms.uiowa.edu/~mbognar/applets/pois.html

３） 連続量の確率分布＝正規分布

さて連続量の場合は、観察を重ねると、どのような形のグラフになるでしょうか。

教科書 65 頁の最後を見ると「連続型データの確率変数ｘは（離散型データの場合のような１、２などではなく）どのような値でもとりうるため、確率の計算は簡単にはできない」と書いてあります。しかし理論的な計算は難しくても、その実例は至るところにあります。

連続型確率変数の確率分布がどのような形になるか、実は皆さんは経験的に知っているはずです。健康診断で測定した皆さんの身長や体重、試験の点数など様々な連続量を、たとえば学年単位でヒストグラムに描いてみてください。釣鐘型／ベル型の分布になるはずです。これを正規分布といいます。

アイオワ大学が提供しているサイトで、正規分布の曲線も描けます。試してみてください。

https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

４ 終わりに

さて今日は世界の様々な出来事を統計的に見る考え方の導入として、出来事の起こり方が離散型確率変数または連続型確率変数で表せることをお話ししました。またそれらの値が存在する範囲を目に見える形で示す分布の話をしました。病気の起こり方から身長や体重に至るまで統計的に考える時は、一例の一つの値で「ここだ！」と決めつけるのではなく、試行や観察を繰り返す中で「中心はこのあたり、全体はだいたいこの範囲に分布する」との捉え方をします。

分布としては、代表的な三つ、二項分布、ポアソン分布、正規分布についてお話ししました。中でも最もよく使うのは正規分布です。

さて、世界には様々な出来事・事象があり、それらの分布を全て数式で表わすと、実は代表的な三つでは足りず、多くの数式・分布が必要になります。どのような分布があるか、その全てを示したのが最後の図です。分布はたくさんありますが、心配しないでください。５回目までの授業で実際に用いる

のは、正規分布だけです。６回目以降の授業では正規分布の他にカイ二乗分布、t 分布、F 分布の名前が出て来ます。これらは事象に対応する確率分布ではなく、基本的な統計計算で得られた統計量の存在範囲を示す分布です。名前だけ頭に入れておいてください。では今日はこれで終わります。

演習問題

1. 離散量（離散型確率変数）とはどのような量ですか。あなたの生活に関連して、具体例を挙げてください。

2. あなたの生活に関連して、連続量（連続型確率変数）の具体例を挙げてください。

3. 二項分布とは何ですか？あなたの生活に関連する例を挙げてください。

4. ポアソン分布について、あなたの生活に関連する例を挙げてください。

5. 正規分布について、あなたの生活に関連する例を挙げてください。

6. 二項分布など確率分布曲線を、実際にコイン投げなどを行って描くのは大変です。しかしネットを介し、コンピューターで電子的に曲線を描くのは難しくありません。以下は二項分布を描くサイトです。米国のアイオワ大学が運営しています。説明は英語です。チャレンジしてみて下さい。
https://homepage.divms.uiowa.edu/~mbognar/applets/bin.html

7. 毎日同じような平和な生活が続くと、私たちは世界の事象が偶然性・確率に支配されているなんて、あまり考えません。しかし新型コロナウイルス COVID-19 の流行で明日が見通せない、今のような時代には、確率的に考えることは大切です。今日の講義への感想、質問など、何でも構いませんので、１００文字以内で書いてください。

第２章　分数と統計

https://youtu.be/kCzYbIMBMfY

皆さん、こんにちは。今回は統計学の２回目、テーマは分数です。皆さんはすでに小学校の算数の時間に分数を学んでいます。その一方、統計学でも分数は基本です。まず分数的な発想についてお話しします。

１　分数的な発想

・一人の人に起こる出来事

まず前回の復習、1 人の人における出来事の起こり方は、それが起こって「いない」か「いる」か、0 か 1 かで、離散型確率変数（離散量）表されます。今日は、皆さん自身に起きる可能性のある出来事として「新型コロナウイルス COVID-19 への感染」を考えてみます。既に、新型ウイルスに感染した経験があると考える人は１を､経験がない人と考える人は0を思い浮かべてください。

・集団（例：このクラス 100 人）

皆さんが所属するクラスの総数を 100 名とします。この 100 名に過去 1 カ月のウイルス感染について質問した結果、例えば 7 人から「感染あり」の答えが得られたとします。この感染の発生を数値で表したらどうなるでしょうか。多くの皆さんは、この問題をそれほど難しいとは感じず「100 分の 7 です」などと分数で書き表すと思います。しかしこのように分数で書き表すことは、昔から常識だったわけではありません。

・古代から分数はあった

数の相対的な大きさを表すために分数を使う考え方は古代エジプトやギリシャの時代からあり、5 世紀にはさらに進んだ形がインドでも現れたとされます。古代の分数の考え方は、目に見える物体やお金をいくつかに分配するような状況で使われていました。分数の考え方で集団を捉えていたわけではありません。

・[ものを分配する分数]と[集団の出来事の起こり方を表す分数]との違い

ものを分配する場合の分数は、幾何学的に,視覚的に表現できます。例えば目の前に１個のリンゴがあり、それを 6 人で分けるとしたら 6 分の 1、この 6 分の 1 は視覚的に容易に捉えられます。

一方、このクラスで新型ウイルスへの感染がどのくらい起きたのかを分数で表すとしたら、合計（Σ, シグマ）を求める計算を二回行わなければなりません。分母には、クラス全員の合計（人数）が必要です。分子には、出来事が起こった人の合計（感染者数）が必要です。このように、分数の計算を行うためには、そこにいる人々を集団として捉え、その人数を数える発想が必要ですが、このような発想は 15 世紀までは未成熟でした。

・人間集団に分数を使い始めたのは 17 世紀

歴史的にみると、感染症の流行が分数の考え方を進歩させました。コレラやペストなど感染症の大規模な流行の現状を把握するため、死亡に関連して分数の考え方が導入され、死亡率の計算が実用化したのは 17 世紀と言われています。その代表はイギリス人ジョン・グラウント（1620-74）です。グラウントは当時たびたびペストが流行していたロンドンにおいて、各教区の教会から入手した出生と死亡に関する情報を分析し、1662 年には「死亡調書の自然的および政治的観察」と

題する革新的な本を出版しています。グラウントは、計算で利用した数値が、ロンドンで発生した全出生と全死亡の一部分でしかないことを認識した上で、そこからロンドン全体の状況を推測することも行いました。この考え方「標本のデータから、より大きな母集団の推測・推定を行う」は現在では推測統計（統計的推測）と呼ばれています。

２分数と集計表

１）初めての統計量、分数

では推測統計の考え方を用いて、実際に分数を計算してみます。次のデータは昨年の統計学の全受講者（100 名、操作的母集団）を対象として行った調査結果の一部分です。100 名から 6 名を無作為に選び、得られた標本（a さんから f さんまで 6 名の事例を含む）について、「朝食の摂取」を離散量（離散型確率変数）として示します。（朝食なし０；朝食あり１）

事例	朝食
a	あり
b	なし
c	あり
d	なし
e	なし
f	あり

「朝食あり」の割合を分数で示すにはどうしたらよいでしょうか。まず分母、全体の人数は、a さん b さん c さん...と数えて f さんまでで合計 6 名です。次は分子、「あり」の人数の合計は 3 名、よって「朝食あり」を表す分数は 3／6 となります。この６名のような小さな集団のことを標本、サンプルと呼びます。さて、推測統計の目的は、標本から母集団の様子を推測することです。今回の標本は 6 名と少数ですが、ここで得られた３／６から、母集団（昨年の全受講者 100 名）の様子をどこまで推測できるでしょうか。「推測」とは何らかの根拠をもとに予想をすることを意味します。絶対に正しい結果を導くのではなく、利用できるデータから予想を積み重ねることが大切です。３／６は一つの根拠と位置付けられます。

２）クロス集計表と分数

さて上述の例では一つの離散量（朝食あり・なし）を集計し、分数を計算しました。実際の調査では、離散量がもう一つ増え、二つの離散量が組み合わされた場合が出て来ます。この場合はどうしたらよいでしょうか。組み合わせる（クロスする）場合の集計では、まず表（クロス集計表）による整理が大切です。二つの離散量のカテゴリーごとに分割して集計することから、分割表とも呼ばれます。次に例を示します。

	疲労感あり	疲労感なし	
朝食あり	n11	n12	
朝食なし	n21	n22	

クロス集計表では、横の並びを行、縦の並びを列、行と列の交差するそれぞれの部分をセルと言います。図に示すのは４つのセル（n 11、n 12、n 21、n 22）がある２×２のクロス集計表です。クロス集計表は、どういう条件の人が何人いるかを整理するのに役立ちます。

ではこのクロス集計表を用いて次のデータを集計してみましょう。ここに示すデータはやはり母集団(昨年の全受講者 100 名)から選んだものですが、今回は 6 名ではなく 10 名を抽出しています。また一つ目の離散量、朝食有無に加え、二つ目の離散量として、疲労感ありなしを示します。A さん B さんとデータを見ながら集計表に正の字を書いていきます。

No	朝食	疲労感
a	あり	あり
b	なし	あり
c	あり	なし
d	なし	あり
e	なし	あり
f	あり	なし
g	なし	なし
h	なし	なし
I	あり	なし
j	なし	あり

	疲労感あり	疲労感なし	計
朝食あり	1	3	4
朝食なし	4	2	6
	5	5	10

最後にこれらの値から自分で分数を工夫してそれがいくつになるかを考えてください。

3　分数と母集団と推測統計

今日は事例 10 名の小さな集団（標本）についてクロス集計を行い分数を得ました。何かを知りたい時それに合わせて分数を考え、その分数が計算できるように調査しデータを集めることは統計学でよく行われます。その際あなたが特徴を調べたいと思う集団、母集団は何かを意識することはとても大切です。ジョン・グラウントの場合、実際の計算で用いたデータは各教区の教会から得た標本でした。しかしグラウントが母集団として調べたいと意識していたのは、ロンドン全体でした。

皆さんの場合はどうでしょうか。皆さんが分析した A さんから J さんまでの 10 名のデータは、昨年の受講者全員（１00 名）の名簿から、乱数表（数値がランダムに並んでいる表）を用いて、ランダム（無作為）に抽出した標本でした。標本の分析から得た結果を元に、標本の背後にある母集団の様子を推測することを、推測統計といいます。

操作的母集団とは実際に標本抽出を行うことができる母集団です。では操作的母集団以外に、さらにその元に、調べたい対象全体をあらわす理想的な母集団が存在するのでしょうか。・・・そう考えてくると、実は昨年の受講者 100 名の背後には、同じ県内の他の看護大学の学生も考えることができます。さらに広げていくと、この県だけでなく、隣の県、・・・さらに日本全国まで考えると、さらに多くの学生が視野に入ってきます。そのように対象全体を捉えた時、それを理想的な母集団と呼びます。

4　悉皆調査と分数

さて、最後に悉皆調査、全数調査という言葉に触れておきます。分数を用いて、日本全体の状況を考えることがあるのでしょうか。

たとえば日本全体の人口を分母に、日本全体の死亡数を分子にとると死亡率が、また日本全体の

出生数を分子にとると出生率が計算できます。国全体の出生率や死亡率はとても重要な値ですので、標本から推測するだけでなく、国民全員についての調査から計算することも行われます。このように、対象をもれなく調べる調査を全数調査（悉皆調査）といいます。出生や死亡などの全数調査についての考え方は、皆さんが秋に学ぶ科目、保健統計学の中で詳しくお話しします。

さて、次回は平均値と偏差、標準偏差についてです。これらも中学や高校の時間に学習したテーマですが、皆さん忘れかけているのではないでしょうか。ぜひ復習して、数字が与えられた時に、自分で計算できるようにしておいてください。電卓を使っても構いません。しかし、みなさんは簡単な計算は手で出来るような訓練を、小学校から高校にかけて行って来ています。手計算は大切な基本能力です。せっかく身につけた能力は、忘れないで、活用しましょう。その上でさらに電卓、パソコン、統計パッケージなど、より高度な計算方法にも親しんで行ってください。

演習問題

1. 最近の生活で気になる分数は何ですか。分数の具体例を挙げてください。
2. 昨日、ある保健所で１さんから 6 さんまで６名が、新型コロナウイルス COVID-19 を心配して PCR 検査を受けた結果を以下に示します。PCR 陽性者の割合を分数で表してください。

人	PCR 検査
１さん	陰性
２さん	陰性
３さん	陽性
４さん	陰性
５さん	陽性
６さん	陰性

3. 新型コロナウイルスによる外出自粛の影響調査を行い、二つの離散量（運動ありなし、食欲ありなし）について、10 名から以下の結果を得ました。クロス集計表を作成し、４つのセルがどうなるか、それぞれにどのような数値が入るかを報告してください。

No	運動	食欲
1	なし	あり
2	なし	なし
3	あり	あり
4	なし	なし
5	あり	あり
6	あり	なし
7	あり	あり
8	あり	なし
9	あり	あり
10	なし	あり

4. 前の設問で作成したクロス集計表から「運動なし、食欲あり」の割合を分数で表してください。またこの分数から考えられることを述べてください。

第３章 平均, 偏差, 分散, 標準偏差

https://youtu.be/DrC-0FP9m9Q

今日のテーマは平均値、偏差、標準偏差、分散の考え方です。
どれも中学校あるいは高校の数学ですでに身につけているはずです。以前学んだことを思い出し、さらに考えを深めてください。最初にお話しするのは平均値の歴史です。

１　平均値の考え方;歴史的発展

・一人の特徴の数値化

健康や保健に関連して、人の特徴を数値で表して理解することは、医療従事者は一般的に行っています。看護師が患者さんを見た場合に、その人の体温、身長、体重、血圧、心拍数などをすぐに思い浮かべるでしょう。ここでは特徴として身長を取り上げます。皆さんの目の前にいる一人の友人の身長、例えば 160 センチとします。

・集団（例:10 人の友人）の特徴の数値化

さて目前に、一人ではなく、10 人の友人がいるとします。10 人の身長は同じではありません。一人ひとり値が少しずつ異なります。ではこの 10 人の特徴を何か 1 つの値に代表させて捉える事はできるでしょうか。これは保健統計学の基礎になる考え方で、平均値を用います。そんなの常識だよ！と皆さんは言うかもしれません。しかし平均値の考え方は昔から常識だったわけではありません。

・古代～15 世紀の考え方

15 世紀までは、一人一人身長や胸囲が異なる 10 人の人がいた時、その 10 人の平均値で代表させるというような考え方はなく、「いろいろな身長や胸囲の人がいる」という事実認識にとどまっていました。

・16 世紀以降の, 考え方の転換

状況が変わったのは 16 世紀以降です。統計的な考え方が発展し、個体数を 2 以上、n までの集団に拡張し、その集団に代表値があると考え、代表値の推定値として算術平均を使うと言う考え方が現れてきました。たとえば天文学では「惑星の位置や月の直径を求める際、何回も観測して計測を繰り返し、その平均値を取ると計測の誤差を減らせる」など、平均値の考え方が様々な分野で「計測の誤差を減らす考え方」として 16 世紀以降ヨーロッパに普及し始めました。

・19 世紀，ケトレーによる革新

人間に対して平均値を使うという革新を始めたのは 19 世紀前半に活躍した天文学者・統計学者であるベルギー人アドルフ・ケトレー(1796-1874)です。ケトレーは多くの人々を観察する中で「一人ひとりの人は個々別々であっても、観察の対象となる個人の数を増やしていくと、人（人々）の平均的な特徴がだんだんに明らかになってくる」と考え、そのような特徴を持つ人を平均的人間と呼びました。観察の数を増やして得られた分布がどのようなものになるかに関連して「ケトレーには釣鐘型の正規分布曲線が至るところに見えた」とされています。ある実験でケトレーは 5,738 人のスコットランド人兵士の胸囲を測定し、その値から正規分布図を作成し、得られた結果

と理論から導かれる分布図とがほとんど完璧に対応することを示しました。

ケトレーは身体的データの計測を科学的に発展させたことでも知られ、体重（ｋｇ）を身長（m）の二乗で割ったＢｏｄｙ　Ｍａｓｓ　Ｉｎｄｅｘは、ケトレー指数とも呼ばれ、人の肥満度を表わす体格指数として医療や看護の分野でも広く使われています。

こうしてケトレー以後は、人間に関連した様々な科学において、集団に平均値の考え方さらに標準偏差、また正規分布といった統計学の分野で発展されてきた考え方を当てはまることが一般化しました。

母集団から標本を抽出し、ある 1 時点において横断的な標本調査を行う場合、対象とする集団の様々な健康の特徴をとらえる上で、平均値は最もよく利用される指標の 1 つです。皆さんもこの講義の後半で米標本により計算演習を行いますが、その際も出発点は平均値の計算になります。

２　基本統計量;平均, 偏差, 分散, 標準偏差

さて天才的な数学者ケトレーが観察と推論から見出した「平均的な人間」という考え方は現代の統計学において集団を捉える際の基本です。この考え方は様々な集団に当てはめることが可能です。例えば学生の皆さんが属しているこの大学の二年生という集団、皆さんが将来就職する病院の入院患者という集団、様々な集団が考えられます。皆さんも勉強する時に、なんとなく平均値・標準偏差値などと考えるのではなく、具体的な特定の集団をぜひイメージしてみてください。

・要約統計量（基本統計量）

平均的な人間を数値で要約して示すのが要約統計量（基本統計量）です。特に大切なのは、データ内の特定の位置を示す「代表値」とデータのばらつきを示す「散布度」です。

１）代表値（データ内の特定の位置を示す量）

・平均値 mean

平均値は、データＸの合計（Σ）をデータ数（データの個数、ｎ）で割った数値です。算術平均とよび、Ｘの上にバーをつけて、またはμ（ミュー）で表します。たとえばデータが {1, 2, 4, 6, 9} ならばΣは 22、ｎは 5、平均値は 4.4 です。

・中央値 median

中央値は、データを大きさの順に並べたとき、真ん中の値です。データ数が奇数のときは、ちょうど真ん中の値です。データ数が偶数なら、真ん中の 2 つの値の平均値です。たとえばデータが {1, 2, 3, 4, 5} ならば中央値は 3、データが {1, 2, 3, 4, 5, 6} なら中央値は 3.5 となります。

・最頻値 mode

最頻値とは、データから度数分布表やヒストグラムを作ったとき、最も度数が多い値のことです。たとえば、データが{1, 2, 3, 4, 4, 5}ならば最頻値は 4 です。

・最小値 minimum／最大値 maximum

データの中で最も小さい値が最小値、最も大きい値が最大値です。たとえばデータが {1, 2, 3, 4, 5, 6}ならば最小値１、最大値 6 となります。

２）散布度

データ全体のばらつきを示す値です。

・出発点としての偏差 deviation

あるデータの実際の値と平均値の差が、偏差です。たとえばＡさんの身長が 164cm,Ｂさんの身長が 158cm,クラスの身長の平均値が 160cm とすると、身長の偏差はＡさん＋4cm、B さん-2cm となります。さて偏差は、個々の値が平均値からどれくらい大きいか小さいかを直感的に知るために便利な値なのですが、集団全体について偏差を合計する(偏差和)と、ゼロになってしまいます。

そのため、データ全体のばらつきを示すためには、偏差をさらに加工する必要があります。

・分散 variance

そこで＋やーの値をとる偏差をそのまま用いず、２回掛け算して偏差二乗とするアイデアが出されました。偏差二乗は必ずプラスの値になります。この値（偏差二乗）をＡさん、Ｂさん、Ｃさんの場合・・・と合計（Σ）して偏差二乗和を求め、最後にデータ数ｎで割ると、偏差二乗の平均値が求められます。この値を分散 variance と呼び、VAR またはσ２と書きます。

分散は、確かにデータ全体のばらつきを示す値ですが、二乗したために、元のデータの単位（長さ、重さ）が「長さ×長さ」「重さ×重さ」に変わってしまい、扱い難いとの議論もあります。

・標準偏差 Standard deviation, ＳＤ, σ

標準偏差とは、分散の平方根です。平方根をとることにより、データの単位を元に戻したと考えられます。データが測定値の場合、標準偏差は通常、測定誤差をあらわすとされます。

動画上での計算演習

事例	身長
1	158
2	153
3	162
4	167
5	150

合計Σ＝790.0　　平均値μ＝790.0／5＝158.0

i	xi データ	xi 偏差	xi 偏差^2
1	158	0	0
2	153	-5	25
3	162	4	16
4	167	9	81
5	150	-8	64
合計Σ	790.0	0	186.0（←分散）
平均μ	158.0		37.2

標準偏差＝ $\sqrt{37.2}$ ＝ 6.09

３　平均値と標準偏差値の意味

さて今日は、ケトレーの平均的人間（平均人）という捉え方から出発し、集団の代表値やばらつきを示す値についてお話ししました。平均人の捉え方は、社会に大きな影響を与えています。平均人は建築物や交通機関のデザインをする上でも大切です。皆さんが用いる机や椅子、エレベータなども、平均人の身長や体重に合わせてデザインされています。

平均的な範囲に入っているか、そこから外れているかは、健康や疾病を考える時も大切です。健診で測定した皆さんの体重や血色素の値を思い出してください。自分の値がクラスの平均値より高いか低いかは偏差で捉えられます。一方、クラスの値が集団としてどのくらいバラつくか、自分はクラスのバラツキの中でどの辺りに位置付けられるか、を考えるためには、標準偏差ＳＤ（STANDARD DEVIATION）が大切です。

もう一度、正規分布曲線を確認しましょう。平均値プラスマイナス１ＳＤの間にデータの６８．３％、平均値プラスマイナス２ＳＤの間にデータの９５．５％、平均値プラスマイナス３ＳＤの間にデータの９９．７％が含まれます。

皆さんが将来看護師になったとき、基本的な検査値について、平均値と標準偏差から正規分布曲線をイメージできると、一人の人の値から、その人が医療を必要としているかの概要を判断できます。たとえば皆さんの同級生、Ｂさん２０歳・女性は、先日の健康診断で血色素１０．５Ｇ／ＤＬでした。またクラス全員の女性の血色素は平均値１３．０Ｇ／ＤＬでした。Ｂさんは貧血を心配する必要があるでしょうか。平均値１３．０より２．５低いという情報だけだと、判断できません。しかしクラス全員の血色素の標準偏差ＳＤが１．０Ｇ／ＤＬだと分かっていたらどうでしょうか。Ｂさんは２ＳＤ（標準偏差の二倍）よりも、さらに低い値だと判断できます。Ｂさんは直ぐに健康管理室に相談すべきでしょう。

さて平均値・偏差・分散・標準偏差などが、看護学を学ぶ上で実に大切な考え方であるということは十分に理解できたと思います。計算方法は中学校や高校の数学の時間に身につけたはずですが、思い出せたでしょうか。紙と鉛筆でも計算できるよう、復習しておいてください。

演習問題

１．ある標本６名（ＡさんからＦさんまで）の体重(kg)は以下の値でした。

45, 47, 52, 54, 54, 62kg.

中央値、最頻値、最小値、最大値を答えてください。

２．先ほどと同じ６名の体重についての計算です。

45, 47, 52, 54, 54, 62kg.

平均値、分散、標準偏差を求めてください。

３．あなたのクラスの身長の平均値を 160.0 ㎝、標準偏差を 3.0 ㎝とします。あなたのクラスメートの一人、Ａさんの身長は 163.0 ㎝、Ａさんは自分の身長が高すぎることを気にして、落ち込んでいるようです。あなたはＡさんにどんな言葉をかけますか。

ワークシート：基本統計量計算

（必要に応じてコピーして使ってください）。

変数（　　）

i	データ x_i	偏差 $x_i - \overline{x}$	偏差二乗 $(x_i - \overline{x})^2$
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
合計 Σ		偏差 和	偏差二乗和
平均 Σ/n			(標本)分散 (偏差二乗の平均)

(標本)**分散** =＿＿＿＿　(標本)標準偏差=√= ＿＿＿＿＿＿

変動係数 = 標準偏差／平均値 = ＿＿＿＿＿＿

第４章　回帰と相関

https://youtu.be/Pnv5AdlZ1Jo

皆さんこんにちは。今回は回帰と相関についてお話しします。最初の時間に様々な事象が偶然にランダムに確率的に起きているという考え方に基づいて様々な確率分布を紹介しました。不確定な世の中を生きていくときに確率的な考え方は大切です。その一方、この世界には、安定して、時代を越えて存在し、受け継がれているように見える事象も存在します。個々は偶然に生起すると考えられる事象が、互いに何らかの関連性を持って存在し、それが世界を意味ある存在としているように見えます。そうした関連性を統計的に捉える際に使われるのが、回帰と相関です。以下では、これらの考え方がどう生まれたかをまず紹介します。

1 回帰と相関、考え方の誕生

回帰という考え方は、統計の歴史の中では分数や平均値の考え方よりはかなり新しく、18 世紀後半に生まれました。イギリスの統計学者・博物学者、フランシス・ゴルトンが出発点です。ゴルトンは進化論を提唱したダーウィンの従弟にあたり、進化論から大きな影響を受けて回帰という考え方を導きました。

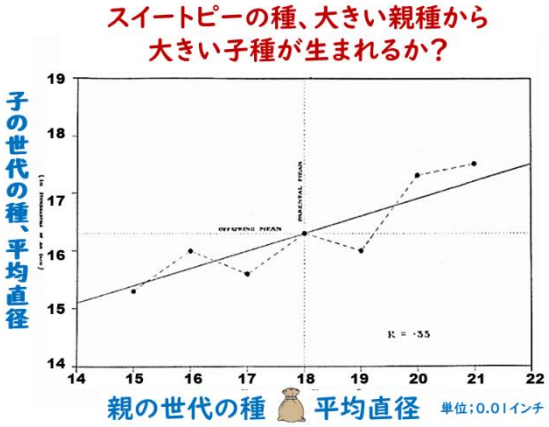
まず進化論を復習します。学生の皆さんは中学校か高校の生物学の時間に進化論を学んでいるはずです。「生物は不変のものではなく、長い年月の間に、確率的な変化が積み重なり、自然選択（自然淘汰）によって、現生の複雑で多様な生物が生じた」という考え方です。

ゴルトンは進化論の影響を受け、様々な出来事が確率的にランダムに起きる一方で、様々な形質が親から子へ孫へと比較的安定して受け継がれている事実に関心を持ち、それを数量的に表わそうとしました。研究を始めるに当たり、人間よりも実験しやすい対象としてゴルトンがまず注目したのが、スイートピーです。

2 ゴルトンの研究

1）回帰の考え方

1875 年にゴルトンが行った実験を紹介します。ゴルトンは、ある時収穫したスイートピーの種 700 個について、種一つずつの大きさ（直径）を測った後、「やや小さめの種の群」から「やや大きめの種の群」まで７群に分け、各群（100 個の種）を袋に入れました。ゴルトンは７人の友人に一人一袋ずつ渡し、各自にスイートピーを育ててもらいました。どの友人にどの大きさの種が入った袋を渡したのか、知っているのはゴルトンだけです。数か月後、ゴルトンは７人の友人から、それぞれに収穫した種を集め、全ての種の直径を測りました。こうしてゴルトンは親種７群と、そこから生まれた子種７群について、直径のデータを得ました。これをグラフに描いたのが次の図です。



横軸；７群の親種、各 100 個につき、直径の平均値（平均直径）を横軸に示す（単位は 0.01 インチ）

縦軸；７群の子種、各 100 個」につき、平均直径を縦軸に示す。

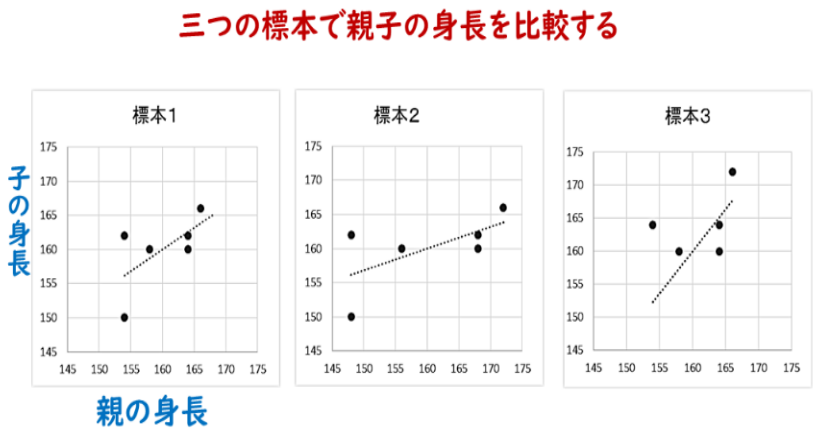
図より、最も小さい親種群の平均直径は 15.0、その親から生まれた子種群の平均直径は 15.2、最も大きい親種群の平均直径は 21.0、そこから生まれた子種群の平均直径は 17.3 などが読み取れます（数値の単位は 0.01 インチ）。

親の平均直径と子の平均直径の間に直線的な関連性があることは、図から明らかです。このデータから、ゴルトンはさらに以下２点に気付きました；１）子の各群の分布のばらつきは、親のばらつきと似た値を取り、どれも正規分布する、２）平均直径が大きい親から生まれた子は平均直径が大きく、平均直径が小さい親から生まれた子は平均直径が小さいが、親世代の平均直径が 15 から 21 の間にあったのに対し、子世代の平均直径は 15.2 から 17.3 と両極端の値が減り、子世代は親世代の全体の平均直径に近づく（親世代の値に戻る）傾向があり、この傾向を線形のグラフ（傾き１以下の直線）で表せる。

この傾向をゴルトンは Regression（平均への回帰）と名付けました。図に示した７個の点の傾向を直線で近似すれば、親種の大きさから子種の大きさを予測できます。ゴルトンの後継者であるスピアマンがこの考え方をさらに発展させ、現代の統計学で重要な回帰分析の考え方に至りました。

2）相関の考え方

ゴルトンは 1870 年代後半から 80 年代にかけてイギリスの南ケンジントンに身体計測研究所を設立し、人間の形質の遺伝について研究を始めました。研究を進める中で、ゴルトンを悩ませた問題の一つが、親子の値をグラフにプロットしたとき、サンプルごとに親と子のデータのバラツキが異なり、異なった傾向線が描ける場合があることです。



図の例は三つの標本における両親と子どもの身長の関連性を示します。何れの標本も６組の親子の身長を示します。標本１は親と子の身長のバラツキが等しくなっています。一方、標本２では子の身長のバラツキが親の場合よりも小さく、また標本３では子のバラツキが親の場合より大きくなっています。親子でバラツキが異なるため、各標本では異なる傾きの傾向線が描けます。しかしバラツキを補正しないと、親と子の身長の関連性の強さを明確に示せません。実はこの３標本は同一の母集団から得られたものであり、親と子の身長の関連性の強さは一定だと考えられました。そこでゴルトンは、計測値の見かけのバラツキを補正し、関連性の強さを直接的に表わす統計的な指標を求めることを試み、その結果、生み出されたのが相関 Correlation の考え方です。

現代の統計学で用いられている相関係数という名前や計算方法は、ゴルトンの後継者であるピアソンがまとめたものですが、元になる相関の考え方はゴルトンによることが知られています。

3 バラツキから相関係数の計算へ

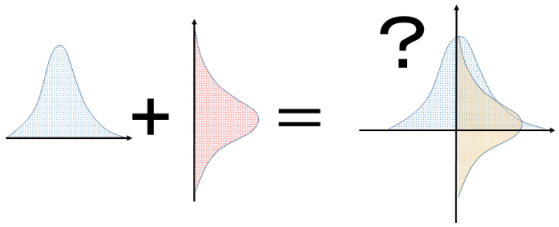
ピアソンはゴルトンの考え方を受け継ぎ、数学的に発展させ、「ピアソンの積率相関係数」の考え方が生まれました。その後、相関係数の考え方は急激に発展し、コンピューターの進歩に伴って現実の世界での統計的な観察を行うときに最もよく使われる方法になりました。

計算方法の原則は、既に前回の授業で学んだデータのバラツキの数値化です。注意すべき点は、データ（変数）を一つひとつ、個々に分布を考えるだけでなく、ＸとＹなど二つのデータが組み合わされた散布図の場合です。こうなると、バラツキの空間的な把握が必要になります。

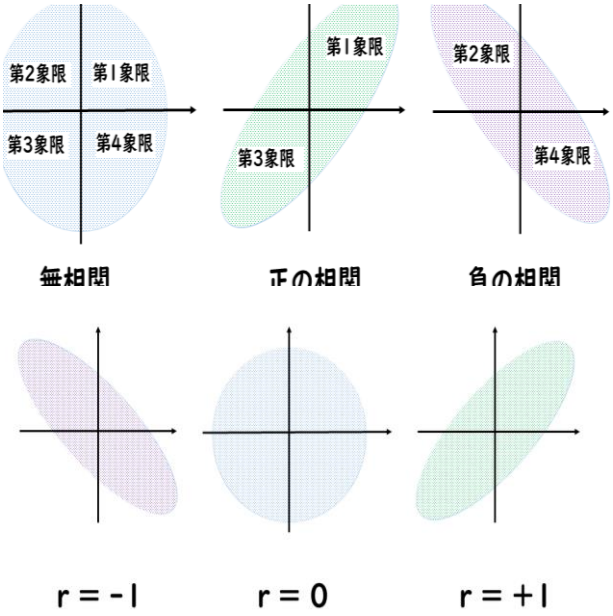
１）個々のデータ（変数）の分布

データが一つの連続量（たとえば身長）の場合、ベル型の分布（正規分布）になることは、前回の授業で学びました。

２）二つのデータが組み合わされたら？



では二つのデータのうち一方をＸ、もう一方をＹとして、散布図（ＸＹ分布図）を描いたら、どうなるでしょうか？



ＸとＹが独立、相互に何の関係もなければ、第１象限から第４象限まで、どの象限にも点が存在する円形の散布図になります。しかし、ＸとＹとの間に関連（相関）がある場合、ＸＹ散布図は第１象限と第３象限を中心にバラつく分布か、あるいは第２象限と第４象限を中心に点がバラつく分布か、どちらかになります。ゴルトンの後継者であるスピアマンが考えたのが、この図の関係（相関）を数値（相関係数）で表わすことです。

4 相関係数の計算方法

相関係数とは、散布図におけるＸＹのバラツキを数値化したものです。まずＸ、Ｙのそれぞれについて、平均・偏差そして標準偏差を計算し，バラツキを数値化します。次に、第１・第３象限へのバラツキが大きければ１に近い値、第２・第４象限へのバラツキが大きければ－１に近くなるような値、共分散を求めます。共分散を二つの標準偏差で割ると相関係数が得られます。

・計算式

$$r = \frac{s_{xy}}{s_x \times s_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

・計算手順

１．二つのデータ（変数；ＸとＹ）それぞれにつき、平均とバラツキ（偏差、分散、標準偏差）を求める。

２．二つのデータの共通するバラツキを求める。

１）偏差積；Ｘ偏差とＹ偏差を掛け算する。

２）共分散；偏差積の平均値を求める（偏差積の合計をデータの個数ｎで割る）

３）相関係数；共分散をＸ標準偏差とＹ標準偏差で割り算する。

・では実際に計算してみましょう。

動画上での計算演習

ステップ１、Ｘの標準偏差を求める。

	HT	ht偏差	ht偏差^2
1さん	168	7	49
2さん	154	−7	49
3さん	158	−3	9
4さん	160	−1	1
5さん	165	4	16
合計	805		124
平均	161		24.8
		標準偏差＝	4.9799598

ステップ２；Ｙ（体重）の標準偏差を求める。

	WT	wt偏差	wt偏差^2
1さん	60	4	16
2さん	48	−8	64
3さん	52	−4	16
4さん	62	6	36
5さん	58	2	4
合計	280		136
平均	56		27.2
		標準偏差＝	5.215362

ステップ３；　偏差積、分散を求め、最後に相関係数を得る。

	HT	ht偏差	ht偏差^2	WT	wt偏差	wt偏差^2	**偏差積**
1さん	168	7	49	60	4	16	**28**
2さん	154	−7	49	48	−8	64	**56**
3さん	158	−3	9	52	−4	16	**12**
4さん	160	−1	1	62	6	36	**−6**
5さん	165	4	16	58	2	4	**8**
合計	805		124	280		136	**98**
平均	161		24.8	56		27.2	**19.6**
		標準偏差＝	**4.9799598**		**標準偏差＝**	**5.215362**	

相関係数＝ 0.75465

5 相関係数の理解と利用

１）図と関連した理解

ポイントは、二つの連続量（変数）、X と Y の相関（相互の関連性）を見ることです。ゴルトンのように散布図からＸとＹの相関を直感的に判断することが大切です。

左の図ではＸが増えると Y は減る関係が明らかで、傾向を右下がりの直線で示せます。

真ん中の散布図は座標の中央に分布し、相関はゼロです。右の図ではＸが増えると Y も増える関係が明らかで、傾向を右上がりの直線で示せます。このように、直線で関係を示せることを、線形関係といい、線形関係の強弱を示す値が先ほど計算した「ピアソンの積率相関係数（相関係数）」です。相関係数はマイナス１からプラス１までの値をとります。

２）相関係数と言葉の表現

相関係数と共に、よく用いられるのが相関係数を２乗した値、決定係数です。Ｘ軸の変数の変化が、Ｙ軸の変数の変化を説明する割合と言われます。教科書１５３頁の図には、相関係数や決定係数の数値と、それをどう言葉で表現するかの対応表があるので、参照してください。

３）回帰と相関をどう組み合わせるか

歴史的にはまず回帰の考え方が生まれ、そこからばらつきを補正した考え方として相関が生まれたことを、お話ししました。一方、現実に統計を利用する場合は、まず相関係数を計算して相関があるかどうかを観察し、相関があるとわかったら、次に回帰式を求めて予測するような使い方が多く行われています。教科書の 152 から 153 頁を参照してください。

４）離散量と相関

今回は X も Y も連続量の場合の相関を扱いました。相関の考え方は非常に強力で便利なためピアソンの相関係数の後さらに研究が進み、順位などの離散量も変数に含める相関の考え方が出てきています。

６　まとめ

相関は基本的な考え方ですが、使い方によっては、事象の意味を深く分析することができます。たとえば遺伝や進化という問題に立ち向かうとき、学生の皆さんが思いつくのはどのような方法でしょうか。たとえば現在問題となっている新型コロナウイルス COVID-19 の変異や診断のためのＰＣＲ検査は、全て遺伝子を操作する技術を用いています。一方、ゴルトンの時代は、遺伝子の構造が解明されるはるか前の時代です。しかしゴルトンはスイートピーの種の大きさとか身の回りの人々の身長とか体重など、身近な現象に注目し、二つの変数をグラフに描き、二つの量が関連するとはどういうことか、その意味を考えぬき、進化や遺伝の考え方とも結びつけていきました。

相関はそれを出発点にして、人間のあり方や社会のあり方まで分析することができる方法論です。人間の知性や感情や行動など、把握が難しい現象についても、相関の考え方を通して捉える試みが進んでいます。新型コロナウイルスの流行に伴って、ビッグデータから携帯電話の位置情報と人々の行動の相関を求め、さらに人々の気のゆるみなど心理的な側面を分析することも普通に行われています。皆さんも身の回りに様々な相関を見いだすことができるはずです。ゴルトンやピアソンのように、相関を通して人間や社会の有様を考え始めてください。

演習問題

１．相関とはどのようなことですか。思いつく具体例を挙げてください。

２．昨年の受講生調査（100 名）から無作為抽出した標本 5 名（ＡさんからＥさん）について、通学時間と予習復習時間のデータを示します。単位は分です。

i	通学	予習復習
A さん	50	30
B さん	20	80
C さん	30	70
D さん	120	10
E さん	80	10

通学時間の平均と標準偏差を求めなさい。（参考；平方根はスマートフォンで計算できます。すぐに画面が現れない場合、スマホを 90 度回転すると、画面が現れます！）

３．上述のデータにつき、予習復習時間の平均と標準偏差を求めなさい。

４．上述のデータにつき、共分散と標準偏差を求めなさい。（動画中で用いたのと同様のワークシートは、講義資料の最後にあります。必要であれば、利用してください。）

５．昨年の調査時は、通常の対面授業が行われており、COVID-19 禍の下での現在の皆さんの状況とは異なります。上記の計算結果から推測される昨年の状況と今のあなたの状況を比較して、100 字以内で考察してください。

ワークシート：相関係数計算

i	データ Xi (　　)	Xi 偏差	Xi 偏差^2	データ Yi (　　)	Yi 偏差	Yi 偏差^2	XiYi 偏差積
合計 Σ		X 偏差 和	X 偏差二乗和		Y 偏差 和	Y 偏差二乗 和	XY 偏差積 和
平均 Σ/n	X 平均		X 分散	Y 平均		Y 分散	共分散(偏差積の平均)

⇩

X 標準偏差＝$\sqrt{\text{X 分散}}$
＝＿＿＿＿＿

⇩

Y 標準偏差＝$\sqrt{\text{Y 分散}}$
＝＿＿＿＿＿

$$\text{相関係数}=\frac{\text{共分散}}{\text{X 標準偏差} \times \text{Y 標準偏差}}$$

第５章　クロス集計表と行%

https://youtu.be/xaPyBTDUEYo

皆さんこんにちは。すでに分数の授業で、二つの離散量（離散型確率変数、変数）で世界を分割して捉えるクロス集計表を取り上げ、その最も単純な形、２Ｘ２のクロス集計表も紹介しました。今回は２Ｘ２のクロス集計表をさらに詳しく学びます。

１　世界を分割する考え方

世界を二つに分割する考え方がダイコトミーDICHOTOMY、二分法です。世界を渾沌とした連続した存在として捉えるのではなく、二分法のように、幾つかの状態がある離散量（変数）で捉える発想は、ギリシャ時代のアリストテレスにまで遡ると言われます。また事象を、表か裏か、勝ちか負けかなどの二分法で捉える数学は、ギャンブルの発達と共に理論化が進みました。

二分法は便利な考え方ですので、私たちはあまり意識せずに二分法を用いています。例を挙げると、たとえば正規労働と非正規労働、検査正常と検査異常、富裕層と貧困層など、いろいろありますね。どれも一種の変数（離散量）です。対立する二つの部分に分けて捉えるため二項対立ともいいます。各部分を合わせた全体は何かと考えると、正規と非正規は「雇用状態」、検査正常／異常は「健康状態」、富裕／貧困は「経済状態」などとなります。

二つの変数を組み合わせ、全体を４分割して捉える２Ｘ２クロス集計表（２Ｘ２表）の考え方も、古くから存在したとされますが、いつを起源とするかは議論があるようです。

２　四分表（２Ｘ２表）の作成と記述的分析

１）利用可能な全ての変数（離散量）を見渡す

作表と分析の第一歩は、意味のある表を作ることです。そのための第一歩が、調査から得られた全ての変数（離散量）を見渡すことです。皆さんの先輩が行った調査では、どのような変数が得られたでしょうか。以下に昨年の例を示します。

- 出身と生活；出身地（大都市／それ以外）住まい（単身／同居）ペット（いる／いない）睡眠（６時間未満／６時間以上）通学（１時間未満／以上）
- 身体の状態；風邪（引きやすい／引きにくい）食（好き嫌い多い／少ない）アレルギー（なし／あり）
- 心の状態；性格（悩む／楽天的）気分（安定／不安定）人好み（ない／ある）
- 学生生活；勉強（１時間以内／以上）講義（楽しい／楽しくない）実習（楽しい／楽しくない）バイト（する／しない）部活（する／しない）
- 将来のこと；卒後希望（看護のみ／他の職業も考える）親介護（家族／施設に任せる）高齢期仕事（70 歳以下／以上も）

昨年の調査では、上記以外にも調査項目があり、変数（離散量）は全部で 36 個得られました。

２）二つの変数を選び、関連性を意識する。

２Ｘ２表では二つの変数（離散量）の関連性が明らかになります。前述の例のように 30 個以上の変数がある場合、２つずつを組み合わせるとすると、数百もの組み合わせが可能ですが、全てを試すわけにはいきません。意味を考えて組み合わせる必要があります。どうしたらよいでしょうか。

どの変数を組み合わせるか迷う時は、調査の目的を再確認します。明らかにしたいこと、調べたい関連性、それに対応した変数はどれでしょうか。「○が原因らしい、その結果が○○らしい」と仮説を意識できるでしょうか。大切なのは「原因の可能性がある」変数と「結果の可能性がある」変数とを区別して整理することです。昨年の履修生が考えた仮説の例を以下に示します。

原因らしい変数⇒結果らしい変数

・住まい（一人暮らし／親と同居）⇒アルバイト（する／しない）
・通学時間（１時間以内／以上）⇒勉強時間（１時間以内／以上）
・性格（悩み多い／楽天的）⇒親の介護（家族がする／施設に任せる）
・親の仕事（医療系／非医療系）⇒卒後の希望（看護のみ／他の職業も考える）
・食（好き嫌い多い／少ない）⇒風邪（引きやすい／引きにくい）
・睡眠（６時間未満／６時間以上）⇒風邪（引きやすい／引きにくい）
・人の好み（人見知りする／しない）⇒実習（楽しい／楽しくない）
・気分（安定／不安定）⇒部活（する／しない）

３）２Ｘ２表を作り、集計する。

以上のように整理できたら、「原因らしい変数」を行に「結果らしい変数」を列にして、集計表の枠組みを作ります。

	風邪引きやすい	風邪引きにくい
睡眠短い	？	？
睡眠長い	？	？

集計表の枠組みができたら、実際に集計します。A さん B さんとデータを見ながら 集計表に正の字を書いてデータの数を数え、実測度数を得たことを思い出してください。表の４つのセルの全てに、当てはまるデータの数（実測度数）を書き込みます。昨年の全受講者 100 名のデータを集計した結果、以下の表になりました。こうしてできたのが、実測度数の２Ｘ２表です。

	風邪引きやすい	風邪引きにくい
睡眠短い	40	20
睡眠長い	10	30

注；一般のアンケート調査は皆さんが後期に学ぶ疫学調査とは異なり、厳密に原因と結果との関連性を調べることはできません。しかし統計的な関連性は検討できます。よってある程度、原因的な要素と結果的な要素を頭に入れておくと集計を意味あるものとして進めることができます。

４）２Ｘ２表で、周辺度数を計算する。

２Ｘ２表に示された実測度数は、全体に対する割合、％として表わすことで、関連性が考えやすくなります。そこで、まず行の計、列の計、全体の合計など周辺度数を計算しておきます。

	風邪引きやすい	風邪引きにくい	計
睡眠短い	40	20	60
睡眠長い	10	30	40
	50	50	100

このように行や列の合計とさらに全体の合計をまとめて周辺度数と言います。　周辺度数の中でも 右下に来るのが 全ての合計 全体の度数です。

５）２Ｘ２表で、行%を計算する。

２Ｘ２表が示す傾向を観察し、考察するためには、行%が役立ちます。行における%、行%は、行の周辺度数（行の計）を分母にした分数として計算します。

	風邪引きやすい		風邪引きにくい		計	
睡眠短い	40	66.67%	20	33.33%	60	100.0%
睡眠長い	10	25.00%	30	75.00%	40	100.0%

“睡眠短い”の場合は 40 を 60 で割って 66.67%、22 を 60 で割って 33.33%です。“睡眠長い”の場合は 10 を 40 で割って 25.00%、30 を 40 で割って 75.00%です。

さてこの%の値からは、何が結論できるでしょうか。睡眠が短い場合は、風邪を引きやすい人の割合が高い傾向がある、とか、睡眠が長い場合は、風邪を引きにくい人の割合が高い、などが読み取れます。

２Ｘ２表を作り、行%を観察するのは２Ｘ２表による統計分析の第一段階です。行%を観察することで、二つの変数（離散量）の関連性を記述することができます。

まとめ

さて、ここまでで２Ｘ２表を使った記述統計分析の考え方をお話ししました。

記述統計は調査した実測値（実測度数）をもとに平均値を計算したり相関係数を計算したりまた今回のように行パーセントを計算し、そこからデータの示す割合（%）の大小に注目して様々な考察を行えます。ここまでの方法を皆さんが身につけることで基本的な統計が使えるようになります。

さてこれで統計学が終わるかと言うと実はここまでは基本的な統計学の第一部、次に出てくるのが、統計における仮説検定という考え方です。

言葉だけ聞くと難しそうに思えるかもしれませんが、皆さんはすでに四分割表を作る時に様々な仮説を用い行パーセントを計算していました。もう皆さんはすでに仮説検定の考え方を使い始めているわけです。次回、さらにお話しします。

演習問題

1. 新型コロナウイルス COVID-19 流行下での学生生活につき、新調査を行うことになりました。あなたなら何を質問したいですか。以下に一つ例を示します。

・外出自粛中のオンデマンド授業は（１楽しい　２楽しくない）

あなたも新たに質問を一つ考えてください。ただし回答は「１はい、２いいえ」など、二つの離散量のどちらかを選ぶ形式とします。

2. ２Ｘ２表を作り、記述的分析を行うためには、二つ以上の質問項目（離散量、変数）と、どちらが「より原因らしい」、どちらが「より結果らしい」の仮説が大切です。動画の中には「睡眠時間⇒風邪の引きやすさ」という仮説が出て来ました。あなたも、新たに仮説を一つ考えて下さい。内容は自由です。矢印などの記号を用いても、全て文章で表しても構いません。

3. 昨年の受講者が立てた仮説から作った２Ｘ２表を以下に示します。

	・アレルギー	
・ペット	あり	なし
いる	12	18
いない	8	62

周辺度数を計算してください。結果は「○○の行の計が xx, YY, △△の列の計が aa, bb, 全ての合計が zz」など、文章で回答してください。

4. ペットとアレルギーに関する上記の２Ｘ２表から行%を計算し、その値から何が考えられるかを、50 文字以内で考察してください。

第６章　クロス集計表とカイ二乗検定

https://youtu.be/XQ231BNwIzc

みなさん、こんにちは。今回は２X２表を用いた仮説検定についてお話します。

１　仮説検定の考え方

１）仮説検定とは

まず仮説検定とは何でしょうか。

辞書を引くと「検定」は「一定の基準に基づいて検査し、合格・不合格、等級などを決めること」と定義されています。この検定を統計的な仮説について行うのが「統計的な仮説検定」です。

学生の皆さんは、既に前回の授業で「原因らしい変数⇒結果らしい変数」として仮説を立てています。また２Ｘ２表で行%を計算し、その値から何が考えられるかを考察しています。この考察はとても大切ですが、その一方で、みなさんの主観も含んでいるかもしれません。この主観をできるだけ取り除き、誰もが認める一定の基準（統計的な基準）に基づいて、仮説を受け入れるかどうかを決めることが、統計的な仮説検定です。

２）帰無仮説とは

統計的な仮説検定を行うためには、出発点になる仮説を統計的な発想で、改めて定義する必要があります。これが帰無仮説です。

帰無仮説は「帰」と「無」という二つの文字からいうと、「無に帰すことを前提とした仮説」です。私たちが普通に立てる仮説は「〇が原因で、◇が起こるのではないか？」「△と◇の間には、何か差があるのではないか？」など、観察の結果、事象の相互の関連性や違いに関心を持ったときに、その関連性や違いを具体的に知りたくて立てます。他方、「帰無仮説」というのは「関連性や違い」を否定する、「関連性や違い」は「無である」とする仮説です。

例を挙げると、前回の授業で、みなさんは「睡眠の長さ（６ｈｒ未満／以上）は、風邪への罹患（引きにくい／引きやすい）と関連するかもしれない」「ペットの存在（いる／いない）は、アレルギーへの罹患（あり／なし）と関連するかもしれない」などの仮説を考えました。このように、通常は、差や関連性を疑って仮説を立てます。他方、帰無仮説は「睡眠の長さと、風邪への罹患は、関連しない（両者は独立である）」「ペットの存在と、アレルギーへの罹患は、相互に関連しない（両者は独立である）」となります。

３）帰無仮説を立てる理由

なぜ知りたいことと反対の仮説を、わざわざ立てるのでしょうか。理由としては「両者に関連性がある」「両者に差がある」などの通常の仮説を立てた場合は、小さな関連性から大きな関連性まで、わずかな差から大きな差まで、あらゆる場合について、考えなければならなくなるからです。他方、帰無仮説の場合は「関連性がない」「差がない」という状態だけ検討し、帰無仮説が否

定（棄却）されれば「『関連性がない・差がない』という仮説が棄却された」との判断（検定）を行えるからです。

２　２Ｘ２表における独立性のカイ二乗検定

概要；

カイ二乗検定とは相関係数のときも出てきたピアソンが 1900 年に発表した方法です。２Ｘ２表における帰無仮説は、「二つの変数（行に示す離散量、および列に示す離散量）の間に何の関係もない」「二つの変数は独立である」となります。

検定の手順としては「集計して実際にセルに記入した値（実測度数）」と「帰無仮説による値；独立を仮定した場合に、各セルに期待される値（期待度数）」とを比較し、両者がどれだけ乖離しているか（はなれているか）をカイ二乗値という検定統計量で表わし、カイ二乗値の大きさから、帰無仮説を検定します。

１）２Ｘ２表で実測度数と周辺度数を整理する

ここまでは前回の授業で行っています。思い出してください。正の字を書いて集計し、各セルに書き込んだ４つの値が実測度数です。また行の計、列の計、全体の合計などが周辺度数です。

	B1	B2	計
A1	n_{11}	n_{12}	
A2	n_{21}	n_{22}	
計			

	B1	B2	計
A1			N_{A1}
A2			N_{A2}
計	N_{B1}	N_{B2}	N

	風邪引きやすい	風邪引きにくい	計
睡眠短い	40	20	60
睡眠長い	10	30	40
計	50	50	100

２）２Ｘ２表で期待度数を計算する

帰無仮説が成立している状態、「独立の状態」を数値で表わすのが期待度数です。二つの離散量が独立であれば、その二つの離散量を組み合わせても何の関連性もないわけですから、期待度数は、各離散量を単独で観察した場合の確率を単に掛け算した値で計算できます。

まずワークシートで計算方法を説明します。

計算例；

	風邪引きやすい	風邪引きにくい	計	
睡眠短い	?	?	60	
睡眠長い	?	?	40	
	50	50	100	100.0%

以上の計算を記号で示すと、次のようになります。

	B1	B2	計
A1	$E_{11}=N_{A1}\times N_{B1}/N$	$E_{12}=N_{A1}\times N_{B2}/N$	N_{A1}
A2	$E_{21}=N_{A2}\times N_{B1}/N$	$E_{22}=N_{A2}\times N_{B2}/N$	N_{A2}
計	N_{B1}	N_{B2}	N

ようするに、期待度数の計算には、各セルの実測度数は必要なく、周辺度数の割合さえあれば計算できます。　一般的な式は

セルの期待度数

＝（セルの行の合計度数／全体度数）X（セルの列の合計度数／全体度数）X（全体度数）

＝（セルの行の合計度数）X（セルの列の合計度数）／（全体度数）

	風邪引きやすい	風邪引きにくい	計
睡眠短い	30	30	60
睡眠長い	20	20	40
	50	50	100

3）実測度数と期待度数の差を計算する

すでに実測度数が得られており、新たに期待度数が得られました。実測度数から期待度数を引くと、理論的な独立の状態から現実の数値がどのくらいずれているか、乖離しているかが数値化できます。

	風邪引きやすい	風邪引きにくい	計
睡眠短い	40−30= 10	20−30= −10	
睡眠長い	10−20=−10	30−20= 10	

4）カイ二乗値を計算する

さて、実測度数と期待度数のずれはプラスとマイナスの両方があり、どんな２Ｘ２表でも合計するとゼロになってしまいます。そこで、二回目の授業で偏差からバラツキを検討したのと同じ論理に従って、実測度数と期待度数の差を二乗し、すべてプラスの値にします。この二乗した値は、標本数によって大きく異なります。よって標本数による影響を少なくするために、実測度数と期待度数の差の二乗の値を、期待度数で割り算します。こうして得たセルごとの値を足し合わせたものがカイ二乗値です。

各セル；χ^2 $(実測度数-期待度数)^2/期待度数$

	風邪引きやすい	風邪引きにくい	計
睡眠短い	$(40-30)^2/30$	$(20-30)^2/30$	
睡眠長い	$(10-20)^2/20$	$(30-20)^2/20$	
計	クロス表全体のχ^2値　16.667		

各セルχ^2 ＝

$$\frac{(実測度数-期待度数)^2}{期待度数}$$

$$\chi^2=\sum\frac{(実測度数-期待度数)^2}{期待度数}$$

３　カイ二乗検定による判断

さてこれで検定統計量カイ二乗値の値を計算することができました。

実測度数と期待度数との乖離が少なく帰無仮説が成立している場合、すなわち二つの離散量の間に関連性が認められない場合は、カイ二乗値は総体的に小さな値をとります。他方、実測度数と期待度数との乖離が大きくなり、二つの離散量の間に何らかの関連性を否定するのが難しい状況に近づくと、カイ二乗値は総体的に大きな値をとります。

では２Ｘ２表の場合、カイ二乗値がどのくらい以上に大きくなったら、帰無仮説を棄却できるのでしょうか。

一つの目安としては 3.84 が用いられます。これはカイ二乗分布での有意水準 5%自由度 1 における値です。この詳しい意味については次回以降の授業で説明します。

皆さんはこれまで行パーセントからの考察を経験していますが、そこには皆さんの主観が入っていました。

今回、みなさんはカイ二乗値を得たことで、主観ではなく、統計的な客観性に基づいて、帰無仮説を棄却し、「二つの離散量は独立ではない、何らかの関連性がある」という判断を行うことができます。先ほどの例でいえば、カイ二乗値 16.667 は明らかに 3.84 より大きいので、有意水準５％で「睡眠時間が６時間未満の場合は、風邪をひきやすい人の割合が、有意に高くなる」などと結論できます。

４　まとめ

さてこれまで、確率分布分数の考え方、クロス集計表、平均値・分散・標準偏差、共分散・相関係数などを学び、ワークシートで計算演習を重ねてきました。今回の 2X2 表によるカイ二乗検定は、皆さんが将来何らかの調査を行うときに、最も役立つ検定統計量です。

プリントの最後にはワークシートもありますので是非自分で実測値、期待値、セルごとのカイ二乗値、そして全体のカイ二乗値などを計算してみてください。

なおこれまで授業を重ねるごとに少しずつ複雑な計算を手でしてきましたが、手による計算は今回でほぼ終わりです。これ以降の授業では、手で経験した計算の考え方を、コンピューターを用いて実行することに焦点を移していきます。計算が苦手な人も計算はコンピューターに任せると割り切れば、楽しく学ぶことができると思います。ではまた次回にお会いしましょう。

演習問題

1. 帰無仮説について理解できましたか。あなたは帰無仮説のような考え方をすることがありますか。帰無仮説について、思うことを 40 字以内で書いてください。何を書いても構いません。

2. 前回の授業で出てきたのと同じ２Ｘ２表です。この表について、帰無仮説を立て、40 字以内の文章で示してください。

	・アレルギー	
・ペット	あり	なし
いる	12	18
いない	8	62

3. 上記の２Ｘ２表について、期待度数を計算してください。ワークシートは資料の末尾にあります。答えはセルの順番に、４個の数値で示します。

4. 上記の２Ｘ２表について、各セルのカイ二乗値、および表全体のカイ二乗値を計算してください。答えはセルの順番に数値で示し、最後に表全体のカイ二乗値を示してください。

5. 上記の２Ｘ２表のカイ二乗値から、あなたはどのような結論を出しますか。40 字以内の文章で示してください。

ワークシート：カイ二乗値計算

0 仮説

1 仮説に従って実測度数から２Ｘ２表を作成

原因らしいもの ＼ 結果らしいもの			計
計			

2 行%を計算

			計
			100%
			100%
計			100%

3 期待度数(二変数が独立と仮定)を計算

期待度数＝(そのセルの行の計)×(そのセルの列の計)÷合計

			計
計			

4 実測度数と期待度数の差を計算

			計
計			

5 各セルの χ^2 (実測度数-期待度数)^2/期待度数 を計算

			計
計			

クロス表全体の χ^2 値

6考察

・行%の観察から
　感じること、考えること
・実測と期待の度数を比べ感じること、考えること
・得られた χ^2 値から感じること、考えること

0 仮説

1 仮説に従って実測度数から２×２表を作成

2 行%を計算

3 期待度数(二変数が独立と仮定)を計算

期待度数＝(そのセルの行の計)×(そのセルの列の計)÷合計

4 実測度数と期待度数の差を計算

5 各セルの χ^2 (実測度数-期待度数)^2/期待度数 を計算

クロス表全体の χ^2 値

6考察

・行%の観察から
　感じること、考えること
・実測と期待の度数を比べ感じること、考えること
・得られた χ^2 値から感じること、考えること

第７章　統計的仮説検定

https://youtu.be/PjsCnLEft-Q

皆さんこんにちは。前回、統計的仮説検定の導入、帰無仮説についてお話ししました。今回は仮説検定の全体像をお話しします。

１　帰無仮説による検定の考え方

１）概要

統計的仮説検定では「2 つの変数の間に関連性がある」という仮説を最初から証明しようとせず、その逆に、まず「2 つの変数の間に関連性がない」という仮説（帰無仮説）を検討します。関連性を示す証拠として用いるのが検定統計量です。カイ二乗（$\chi 2$）値も検定統計量の一つです。変数相互に関連がない場合、帰無仮説が成立している場合、検定統計量は小さな値を取ります。他方、変数相互の関連性が無視できないくらい強くなったとき、「この関連はたまたま、偶然に生じたとするには、あまりにも希少な、小さな確率(例えば 0.05、0.01）で起っている」「偶然とはいい難い」との証拠が固まった時、帰無仮説を棄却（否定）します。では、ずっと「関連性が無い」と言い続け、帰無仮説を保持して来たのに、最後に保持しきれなくなり、帰無仮説を棄却した状態を、どう位置付けたらよいでしょうか。この場合、何も仮説が無くなったのではなく、帰無仮説と反対の仮説が採択されたと考え、その反対の仮説、つまり「2 つの変数の間に関連性があるという仮説」を「対立仮説」といいます。

２）どのようなときに統計的仮説検定を行うか

統計的な仮説検定について教科書は多くのページを割いて述べています。もちろん看護師の国家試験にも出題されます。それくらい大切な考え方ですが、それほどしばしば使うものではありません。例えば今新型コロナウイルスの流行で第２・３波が来ると様々なニュースが報じ、テレビや新聞には毎日、患者数や死亡者数など統計の数字が出てきます。平均値、標準偏差、相関係数などの基本統計量は、現状を知るのに便利です。しかし教科書があれほど重視している仮説検定は、話題になりません。一体どうなっているのでしょうか。どこに仮説検定の話題があるか見渡すと、新型ウイルスに関連して言えば、治療薬やワクチンがあります。

たとえば、治療薬として既に数か月前から幾つもの名前が上がっていますが、実際に用いられるのは、まだ先の話になりそうです。なぜ時間がかかるのでしょうか。理由の一つは、効果をチェックするための仮説検定を丁寧に行う必要があるからです。「良さそうだから、すぐに使う！」という仮説・行動を取るわけにはいきません。被験者を二群にわけ、投与群と非投与群とで効果を比較する統計的検証、「投与群に効果がある」とする仮説ではなく、「投与群と非投与群とで差が無い」「薬物の投与は疾病の治癒と関連しない」という帰無仮説を検証します。なぜ効果があることを期待しているのに、「効果が無い」という帰無仮説を立てて、慎重に検証を続けるのでしょうか。なぜなら、統計的な検定を行い、帰無仮説を棄却し、効果があると判断すると、簡単には後戻

りできないからです。多くの場合、帰無仮説を棄却すると、その方向で、さまざまな社会的対応が後に続きます。帰無仮説を棄却するというのは、社会的な責任を伴う重い判断だと言えます。

2　数表を用いた仮説検定の進め方

検定を行う場合は、その検定統計量の表を読み取ることが必要です。の表は、教科書の末尾にあります。表を読むための基礎知識を以下に示します。

1）主な検定統計量

検定統計量とは、検定の対象となる複数の事象の「関連性や相違の程度」を 1 つの数値に代表させたもので、与えられた標本のデータから計算されます。

どの課題にどの検定統計量をもちいるか、例を以下に示します。

・課題		検定種類	検定統計量	分布と数表
二群で平均を比較	⇒	t 検定	t 値	t 分布
二群でバラツキを比較	⇒	F 検定	F 値	F 分布
三群以上で平均を比較	⇒	F 検定など	F 値	F 分布
期待値と実測値を比較	⇒	χ^2検定	χ^2値	χ^2分布

まず教科書の 195 頁から数頁をチェックし、数字の並んだ表があるのを確認してください。

2）検定統計量の表の見方

（1）表の構造

197 頁のカイ二乗分布表を例に説明します。

表の左端で縦に並ぶ数値、ギリシャ文字でニュー、英語の V の字に似ています、1 から 30 まで、これが自由度です。

表の上端で横に並ぶ数値が有意水準、ギリシャ文字、アルファで表示し、検定統計量に対応する確率を示します。小数点以下の数値が並び、右に行くほど 0.05、0.01 など小さくなっています。

そして表の中に並ぶ数値が、検定統計量、カイ二乗値です。

（2）カイ二乗検定における表の用い方

- あなたが、観測度数と期待度数から計算したカイ二乗値（４つのセルのカイ二乗値の合計）を用意します。
- 自由度と有意水準を決めます。普通は、自由度＝１、有意水準＝0.05 を用います。
- カイ二乗分布表（197 頁）を見て、自由度１、有意水準 0.05 のカイ二乗値をメモします。（＝3.84）
- 上記であなたが計算したカイ二乗値と、表から得た「有意水準に対応するカイ二乗値＝3.84」を比較します。この 3.84 は、帰無仮説の棄却を判断する限界の値ということで、限界値とも呼ばれます。

- あなたが計算したカイ二乗値が 3.84 未満であれば、帰無仮説は棄却できません。「行に示す離散量と列に示す離散量の間には、有意な関連が認められない」と結論します。
- あなたが計算したカイ二乗値が、3.84 以上であれば、帰無仮説は棄却されます。「二つの離散量の間には、有意水準 0.05%で関連性がある」と結論されます。

３　もう一歩詳しく

・自由度；　自由度とは、さらに詳しくいうと、自由に決められる値という意味で統計的には検定しようという標本の複雑さを表します。カイ二乗検定では表の複雑さに対応し、M×N 表など複雑なクロス集計表の自由度は（ｍ－１）（ｎ－１）で計算します。２X2 表はクロス集計表の中では最も単純ですので、自由度は（２－１）（２－１）＝１となります。

・有意水準；　有意水準はある事象の起こる確率が偶然とは考えにくい、とする判断基準です。統計的検定では、帰無仮説を検討し、どこかの水準で帰無仮説を破棄するという重い判断をする分けですが、その際、この確率が 0.2（２０％）、0.1（10%）など高めの値だと、帰無仮説が正しいのにそれを棄却してしまう誤り（第一種の過誤）を犯す可能性があります。そこでこの値を普通は 0.05（５％）さらに慎重に判断する場合は 0.01（１％）などと設定し、第一種の過誤を少なくすることが一般的です。

４　表からコンピューターへ

さて数表は昔から統計計算に欠かせないものとされ、どの統計の教科書にも巻末に数表があり、様々な統計的仮説検定を行う際に活用できます。特にコンピューターの助けを借りることが難い時代、数表の作成は高度に専門的な作業でした。例えば私が大学院に入った 45 年前は、まだネットもパソコンも存在せず、ソロバン・真空管で動く初歩の計算機・歯車の動きを組み合わせて計算する機械的計算機などが使われていた時代です。自分でカイ二乗分布の曲線を計算することなど、考えられませんでした。

他方、現在はコンピューターが進歩し、実は数表を使わなくても、統計的な数値を直接に計算できます。たとえば、以前紹介した米国アイオワ大学のサイトを使って、教科書の数表にあるような値を計算してみることができます。

まずパソコンやスマートフォンでアイオワ大学のサイトにアクセスします。このサイトでは様々な確率分布曲線が描けます。

https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

今、開いているのはカイ二乗分布曲線を描く頁です。教科書を持っている人は 72 頁の図をみて下さい。曲線を描くためには、条件の指定が必要です。そこで自由度に対応する空欄（ＤＦ）に１を入力すると、赤い矢印で示す曲線が現れました。これが自由度１のカイ二乗曲線です。この自由度として２、３、４などの数字を入れると、教科書 72 頁のような図が現れます。

今度は、有意水準Ｐの空欄に 0.05 と入力してください。カイ二乗値が 3.84146 と計算されます。教科書 197 頁の数表にあるカイ二乗値 3.84 と比べると、いま得た数値の方がより正確で桁数が多い

ことが分かります。

最後は、前回の授業で計算したカイ二乗値 16.667 を入力してみてください。Ｐの欄に 0.00004 という数値が現れます。これがカイ二乗曲線から直接に計算した確率、有意確率です。この有意確率が簡単に計算できるなら、実はもう伝統的な数表は必要ありません。有意確率が有意水準より小さな値を取ることが明らかなので、帰無仮説は棄却されます。

５　まとめ

以上でカイ二乗検定の基本は終わりです。

統計的な考え方は、帰無仮説の考え方にも現れているように、確率分布を基礎にした様々な数学的な考え方が組み合わさって出来ています。統計は一度、理論や方法が確定しても、それで終わりではなく、考え方をより洗練させ、より正確な判断が出来るように、検討が続けられています。

今回、動画の中ではお話しませんが、資料の中には、みなさんがよりよく理解できるように、説明を追加しています。余裕があれば、資料にも目を通しておいてください。

（以下、動画ではお話していない部分を、参考資料として示します。）

６　参考

１）カイ二乗値について

ピアソンが確立したカイ二乗検定は、離散量から計算されるカイ二乗値を、連続的なカイ二乗分布で近似するため、セルに入る数値が小さいと、近似が不正確になることが指摘されました。

・イェーツの補正：　これを補正するために、イェーツは 2X2 表の各実測度数と期待度数の差の絶対値から 0.5 を差し引くという簡単な補正法を提案しました。これがイェーツの補正です。

・フィッシャーの直接確率：　2×2 表のカイ二乗検定で、実測度数が５以下、場合によってはゼロなど、とても小さな値を取ると、より厳密な補正方法が必要になります。そこでフィッシャーは、順列組み合わせに基づいたより正確な有意確率の計算方法を提案しました。これがフィッシャーの直接確率法です。

２）自由度（Degree of freedom）

比較する群に含まれる標本数が多くなると、標本のバラツキが増え、そこに含まれる情報量も増えます。この情報量を表わす目安が自由度です。自由度は検定統計量にも影響を与えます。主要な検定統計量は、自由度別に数字が並んでいます。たとえば 196 頁には平均値の検定に用いるｔ分布表が、197 頁にはカイ二乗検定に用いるカイ二乗分布表があります。これらの表では、自由度はギリシャ文字ニューで示されています。英語の V の字に似ていますので、見つけてください。

改めて定義すると、自由度とは「変数のうち独立に（自由に）選べるものの数」を意味します。たとえば、A さんから D さんまでの４人（ｎ＝４）について、体重が 46、48、51、53kg とすると平均は 49.5kg、自由に値を取れるデータは４人分の体重ですから、自由度は４、よってデータの個数（ｎ）がそのまま自由度になります。しかし統計学で、平均値から出発し、さらに様々な統計量を計算していく場合、自由度はｎではなく、ｎ－１となります。なぜｎ－１になるのか、先ほどの例で言えば、平均値 49.5 という情報を使ってさらに分散などの計算をするとき、平均値 49.5

に加えて、A さんからＣさんまで３人分の体重の情報があると、４番目のＥさんの体重は、情報として必要なくなるからです。（４人分の平均が 49.5、A さんからＣさんまでが、46、48、51 であれば、４人目の体重は既に決定されたことになります。）このような理由で、統計的仮説検定を行うときには、自由度はｎ－１を使うのが一般的です。

・クロス集計表の自由度

さて、自由度はｎ－１と言いましたが、クロス集計表の自由度は、ちょっと独得なので、追加して説明します。皆さんがこれまで集計した２×２表は、クロス集計表の中でも、もっとも単純なもので自由度＝1 でした。では、より複雑なクロス集計表とはどのようなものでしょうか。また表が複雑だと、自由度はどうなるでしょうか。

２×２表の場合、行に示した離散量も、列に示した離散量も、それぞれ二つの値（１／０、はい／いいえ、あり／なし）しか取りませんでした。しかし二つ以上の値を取る離散量も多く存在します。たとえば「意思表示；はい／いいえ／どちらでもない」「満足度；とても満足／やや満足／どちらとも言えない／やや不満／とても不満」「回数；０回／１回／２回／３回／・・ｎ回」などです。こうした離散量をクロス集計する場合は２×２表では足りず、２×N 表、M×N 表などが必要になります。M×N のクロス集計表の自由度は（M−1）×（N−１）となります。

３）有意水準

帰無仮説が成立している状態、行と列に示した二つの離散量が、互いに独立で、両者に何の関連性もない場合、つまり、実測度数が期待度数と一致する場合、カイ二乗値はゼロになります。

他方、実測度数と期待度数の差が大きくなると、カイ二乗値も大きくなります。カイ二乗値がどこまで大きくなったら帰無仮説を棄却するかの基準は有意水準という確率値で示されます。

これが自由度１のカイ二乗曲線です。横軸がカイ二乗値、縦軸はそのカイ二乗値が出現する確率、例えば３は 0.3 に対応します。この曲線の下の面積は合計すると、つまり積分すると１になります。

さて、カイ二乗値が段々に大きくなったとき、どこかで帰無仮説を棄却するかの判断をしなければなりません。たとえばカイ二乗値が１とか２とかで棄却すると、帰無仮説が正しいのに、二つの項目が本当は無関係なのに、その仮説を棄却する間違いを侵す可能性が高いです。棄却域かどうか、その境目のカイ二乗値が棄却限界値（限界値）です。グラフに示した限界値よりも右側の曲線下の領域が棄却域です。統計的な仮説検定では、かなり慎重に棄却域を設定します。曲線下の全面積を１としたとき、棄却域の占める面積の割合を確率、ｐ値で表わし、有意水準と呼びます。有意水準として通常用いられるのは５％、または１％の値です。自由度１のカイ二乗曲線では、有意水準５％のカイ二乗値は 3.84、実測値から計算したカイ二乗値がこの値よりも大きいとき、前回の授業中に出てきた例でいえば、カイ二乗値 16.667 は明らかに 3.84 より大きいので、有意水準５％で「睡眠時間が６時間未満の場合は、風邪をひきやすい人の割合が、有意に高くなる」などと結論できます。

以上の判断は、有意水準５％ですが、帰無仮説を棄却するかどうかの判断をより厳しくする場合は、有意水準１％、0.1%などを使うこともあり得ます。では様々な有意水準に対応した棄却限界値を知るにはどうしたらよいでしょうか。また自由度が 1 より大きい場合はどうしたらよいでしょうか。コンピューターがあれば、先ほどのアイオワ大学のサイトでのように、直接に棄却限界値などを計算できます。また教科書の最後には、予め計算した表が載っていますので、参照してください。

演習問題

1. 動画では前回に続き帰無仮説に触れています。新型コロナウイルス COVID-19 に関連して、何かあなたらしい帰無仮説を立ててください。40 字以内で書いてください。

2. ２X2 表からの帰無仮説を検定する場合、自由度と有意水準の設定が必要です。有意水準として 0.05 を用いる場合は既に練習しました。では有意水準を 0.01 に変えたとき、検定統計量としてのカイ二乗値は、どのような値になるでしょうか。教科書の表から読み取って、以下に記してください。

3. 以前出てきたのと同じ２Ｘ２表です。

	・アレルギー	
・ペット	あり	なし
いる	12	18
いない	8	62

あなたは既にこの場合のカイ二乗値を、前回の授業で計算しています。有意水準を 0.01 としたとき、この表からの帰無仮説について、あなたはどう判断しますか。帰無仮説を棄却しますか、それとも維持しますか。あなたの判断とその理由を、以下に 30 字以内で書いてください。

4. 数表や確率分布曲線を作る作業は、以前はとても難しく、数学者が時間をかけておこなっていました。でも今はコンピューターの助けを借りて自分で分布曲線を描けます。以下アイオワ大学のサイトを利用し、自由度に様々な数値（整数）を入力し、曲線を描いてみてください。

https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

自分で曲線を何本か描いたら教科書 72 頁図 3 - 21 と比較してください。ほぼ同じ？それとも違いがありますか。結果や感想を 40 字以内で書いてください。

第 8 章　調査票の観察と集計

統計学の学習はもう後半に入りました。まず期末レポートの概要を説明します。

「文字数 1000 字以上。図表や数値は文字数に含めない。締切 7 月 22 日。pdf ファイルで提出」

レポートを書くために必要なデータ、レポートの細かい形式、提出方法などは、次回、第 9 回目の授業で示します。

・今回の授業は、レポートを書くための考え方を整理します。

・今回の授業では動画は用いません。各設問についている画像の中に、必要な情報を示しています。資料中にも同じ画像があります。

課題 1

図に示すのは 3 年前の統計学の時間に、皆さんの先輩が回答した調査票の一部です。ここには連続量あるいは離散量で表せる質問が並んでいます。あなたが関心を持つ項目を、そのデータ形式（連続量か、離散量か）も含めて、三つ挙げてください。

課題 2

上記の調査票で、あなたが相関を調べたい項目がありますか。組み合わせを二つ考えてください。

回答を入力してください。

課題 3

調査票を集計するときは仮説が大切です。もし上記の調査票をあなたが集計するとしたら、どのような仮説を立てますか。まず「原因らしいもの」はどれか、次に「結果らしいもの」はどれかを述べます。また、なぜその二つを選んだのか、理由も記してください。50 文字以内です。

課題 4

上記の調査票で調査し、得られた結果の一部（8 名分)を図に示します。８名分のデータを観察して何か気付いた点があれば、述べてください。このようにデータを観察し見解を述べることは、調分析の出発点として大切です。50 文字以内で書いてください。（新型コロナウイルス COVID-19 が流行している現在の生活と比較すると、気づくことが多いと思います。）

課題 5

上記の調査結果から二つの離散量を選んで２×２表を作成し、行%を計算し、得られた値と考察を 50 文字以内で記してください。

課題 6

期末レポートは、データを観察して分かったこと、立てた仮説、仮説に基づいた計算結果、考察など、順序立てて書くことが大切です。その練習として、以上の問（１から 6 まで）で考えたこと、感じたことをまとめ、分かりやすく 200 文字以内で書いてください。

2017年　調査票・保健統計学質問　自分の記号 __ __ __

A 基本: ・性別 ①女 ②男　・年齢 __ __歳　・身長 __ __ __cm

1 親の職業	①医療関係	②医療以外
3 生育環境	①田舎・小都市	②大都市で育つ
4 片づけ・整理	①得意だ	②苦手だ

B 現在の生活

1 住まい	①一人暮らし	②親元から通う
2 バイト	①しない	②する
3 大学部活	①しない	②する
4 ボランティア	①しない	②する
5 片道通学時間	0---10---20---30---40---50---60---70--	
6 予習.復習時間	0---10---20---30---40---50---60---70--	
7 ﾈｯﾄ.ﾃﾚﾋﾞ時間	0---10---20---30---40---50---60---70--	
8 バイト時間	0---10---20---30---40---50---60---70--	

C 大学に関連

1 講義	①楽しくない	②楽しい
2 実習	①楽しくない	②楽しい

D 健康状態

1 花粉症等ｱﾚﾙｷﾞｰ	①ｱﾚﾙｷﾞｰあり	②ｱﾚﾙｷﾞｰなし
2 食べ方	①食べ過ぎ多い	②食べ過ぎはない
3 風邪ひき	①すぐ風邪引く	②風邪引きにくい

E 生活習慣

1 歩くこと	①一日１時間以内	②一日１時間以上
2 食の好き嫌い	①食の好き嫌い多い	②何でもよく食べ

ID	性別	年齢	身長	体重	睡眠時間	A1親職	A3生育	A4片づけ	B1住まい	B2バイト	B3部活	B4ボランティ	B5通学時間	B6予複時間(分)	B7ネットテレビ時間	B8バイト時間(分)	C1講義	C2実習
1	1	20	153	45	6	2	1	1	2	2	2	2	60	0	200	200	1	1
2	1	19	160	57	6	2	1	1	1	2	1	2	20	60	90	200	1	1
3	1	19	154	53	6	2	1	1	2	2	2	2	20	50	80	200	1	1
4	1	19	153	45	5	2	2	2	2	2	2	2	180	80	60	200	2	1
5	1	19	157	45	7	1	1	2	1	1	1	1	20	120	200	0	1	1
6	2	20	172	53	6	2	1	1	1	2	2	2	10	0	80	200	1	2
7	1	20	153	42	6	2	1	2	1	2	1	2	90	90	90	200	1	1
8	1	20	168	55	5	1	1	2	2	1	2	2	50	30	140	0	1	2

第９章　12 名のデータでもレポートが書ける

https://youtu.be/hfvXDZ_rpyA

みなさん、こんにちは。今回は少数データでもレポートが書けることをお話しします。

統計学ではさすがに数名程度のデータでは、何も統計的な計算ができないため、レポートを書くことはできません。では何名以上のデータがあれば、統計学を使ってレポートを書けるのでしょうか。絶対的な基準を示すのは難しいのですが、昨年までの統計学、特に前半では 12 名の標本を用いたレポート作成を行っていました。

１　なぜ 12 名が意味を持つのか。

・12 人なら統計が使える

12 名のデータがあれば、特に問題なく基本的な統計計算が行えます。第２回目から４回目までの授業での例題の標本数を思い出してください。

・12 人なら手が使える

12 人程度であればコンピューターがなくても、ワークシートを使って、手で計算できることも重要です。計算をすべてコンピューターに任せるのではなく、どのようにして基本的な統計量が得られるかを、手で体験することは、意味があります。

・12 人なら質的観察ができる

皆さんは看護研究の授業を受けたことがあるでしょうか。質的研究では、少数の人から詳しく話を聞き、事例的・質的に考えることが求められます。他方、量的研究では、多数の人々から多くのデータを集め、統計的・数量的に考えることが求められます。100 名の人のデータは、もちろん量的な研究の対象です。100 名のデータが並んでいる表を見ると、データの多さに圧倒され、ひとり一人の、ひとつ一つの数値を丁寧に見ようという気持ちは出て来ません。では 12 名のデータならどうでしょうか。12 名だと一人一人の人のことも気になります。だから人数が少なく、統計的な数値の計算に限界があることは、欠点ばかりではありません。統計学だから数字だけが大切と思わず、物の見方を一歩質的な方向に近づけることで、「たった 12 名のデータ」ではなく「12 名の個性が反映した数値」と感じられます。そう思うことで数値の向こう側に、その数値が表す人間が感じられます。

２　私だけの 12 名のデータにどう出会うか？

・昨年まで

さて、ではどうやって私だけの 12 名のデータに出会ったらよいでしょうか。昨年までは、統計学の時間に、まずクラス全員（ほぼ 100 名）の協力を得て生活調査を行った後、全員（100 名）から 12 名を無作為に抽出し、その抽出標本（my 標本）を統計計算の演習用として、クラスの各学生の皆さんに渡していました。学生 A さんの my 標本を得るためにまず一度無作為抽出、学生 B さんのために二度目の無作為抽出、・・・と、私は無作為抽出を 100 回行いました。だから全員の皆さんに、それぞれの異なる my 標本を用意することができました。

・今回は？

今回はオンデマンド授業のため、昨年までのやり方が使えません。そこでみなさん自身が、自分の my 標本を選んでください。皆さんの今日の資料の中に皆さんの先輩 150 人分のデータを並べた一覧表が入っています。一覧表のデータは、既に私が無作為に並べ直し、先頭の 1 から順番に番号を付けています。あなたは、この一覧表から 12 人分のデータを選び、あなたの my 標本としてください。選び方は、あなたの出席番号の末尾 3 桁と同じ番号を出発点として、一覧表でその番号から連続して 12 人を選ぶやり方です。例えばあなたの出席番号が 10 番であるなら、一覧表の 10 番から 21 番までの 12 人を選んでください。出席番号が 110 番の人は、リストの 110 から始め 121 までを選んで、ださい。このようにすることで、皆さん一人一人が、それぞれに異なるマイ標本を得ることができます。

・12 名の my 標本で何をするか

今日の授業の表題にもなっているように、あなたはあなたの my 標本を使ってレポートを書くことができます。

３　カイ二乗値計算方法の補足

すでに皆さんはカイ二乗値をワークシートから求める方法を学びました。カイ二乗検定はアンケート調査を集計しそこから考えていく際にとても役立つ方法です。しかしどの統計の計算方法も標本数が少ないと誤差が大きくなったり不正確になったりする傾向があります。

・イェーツの補正

たとえばカイ二乗検定は、離散量から計算されるカイ二乗値を、連続的なカイ二乗分布で近似するため、セルに入る数値が小さいと、近似が不正確になることが指摘されました。これを補正するために、イェーツは 2X2 表の各実測度数と期待度数の差の絶対値から０．５を差し引くという簡単な補正法を提案しました。これがイェーツの補正です。

・フィッシャーの直接確率

２×2 表のカイ二乗検定で、実測度数が５以下、場合によってはゼロなど、とても小さな値を取ると、より厳密な補正方法が必要になります。そこでフィッシャーは、順列組み合わせに基づいたより正確な有意確率の計算方法を提案しました。これがフィッシャーの直接確率法です。

・今後のカイ二乗検定、計算法

さて、せっかくカイ二乗検定を理解したのに、新たな補正方法が出てきて、なんだか複雑だな、と思う人もいるかもしれません。しかしもう皆さんは基本的な計算方法は習得しているので、新しく出てきた補正については、もう自分で計算する必要はありません。コンピューターを使ってください。パソコンからもスマートフォンからも使える計算のサイトを紹介します。

まずここにアクセスしてください。

左の画面を見るとメニューがあります。ここで M×N のカイ二乗検定を選びます。

さて複雑なものが出てきました。これは複雑なカイ二乗検定を行うためです。皆さんの 2×2 表の場合は、ここにそれぞれ２の値を入れます。

さてこれで２×２のカイ二乗検定を行う準備ができました。

ここに入れる数値ですが、以前、第６回目の授業でで来た睡眠の長さと風邪の引きやすさの数値を入れてみます。

	風邪引きやすい	風邪引きにくい
睡眠短い	40	20
睡眠長い	10	30

さてこのサイトが便利なのは、皆さんが既に習得した基本的なカイ二乗値に加えて、イェーツの補正もフィッシャーの直接確率も、自動的に計算してくれる点です。

４　最後に

さて今日はデータがたった 12 人しかなくても、データをしっかり観察し、また外部のサイトを使うことで、データ数が少なく補正を要する場合でも、楽にカイ二乗値を計算することができるとわかりました。レポートを書くことはそんなに難しいことではありません。あなたらしい仮説を立て、どんな計算をするか、考え始めてください。

演習問題

1. 動画の中にある方法に従い、あなたの my 標本を選んでください。元になる 150 人分のデータは、資料フォルダ中にあります。選んだら、その 12 人分のデータを紙に書き写すかコピーするかして、観察してください。気付いた特徴を 50 字以内で記してください。

2. 動画に出てきた js-STAR を利用し、２ x ２表のカイ二乗検定を行い、結果を 50 字以内で記してください。

   http://www.kisnet.or.jp/nappa/software/star/

   分析する数値は、あなたの my 標本から得てください。イエーツの補正を適用するかは、自分で判断してください。

3. 今回の資料の最後に、皆さんの先輩二人が書いたレポートを参考資料として示しています。二人のレポートを読んで、感じたこと、気づいたことを 50 字以内で書いてください。

第 10 回　ｍｙ標本からクラス全体のデータへ

https://youtu.be/7eugz5sZOew

さて前回の授業では、12 人だけのデータからも統計学のレポートが書けるという話をしました。しかし皆さんが利用可能なのは、my 標本だけではありません。my 標本を抽出した母集団（操作的母集団）、皆さんの先輩 150 人分のデータについても、利用できます。今回は my 標本から、その母集団（操作的母集団）に目を向けて、統計的な用語の説明を中心に行います。

１　my 標本から母集団へ

・標本

まず皆さんが前回救出した小さな集団選んだ 12 人のデータセット、my 標本、これを標本と言います

・操作的な母集団

実際に標本抽出を行うことができる母集団のことです。2017 年あるいは 2018 年に、この統計学を履修した皆さんの先輩が、この操作的な母集団です。

・概念的な母集団

教科書によると、調べたい対象全体を表す理想的な母集団と定義されています。皆さんの先輩が操作的母集団だとすると、その背景にある、より大きな、理想的な集団といえば、「他の大学の看護学生も含めた、日本全体の看護学生の集団」といえます。

２　標本抽出と乱数

母集団から標本を選ぶ際、適当に選ぶわけにはいきません。私たち人間が適当にこの選ぶ作業を行うと主観が入る可能性があります。主観を取り除き、無作為にランダムに選ぶ方法が求められます。全ての標本に番号が振られた母集団の名簿をもとに、乱数表などを使ってランダムな順番で番号を選ぶ方法が単純無作為抽出法です。今回の 12 人は皆さん自身が乱数表を使って選んだのではなく、私が乱数を使って、すでに皆さんのために順番をランダムに並び替えていたデータを利用しました。乱数表は、数字を無作為に並べている表のことです。昔の統計学の教科書には必ず乱数表が付いていました。最近はエクセルや、またスマートフォンでも iPhone であれば、簡単に乱数を発生させることができます。iPhone の計算機は横にすると関数電卓になります。Rand キーを押すごとに、１未満の乱数が発生します。

インターネット上にも乱数を発生できるサイトがあります。

https://keisan.casio.jp/exec/system/1425449868

他方、印刷された乱数表は見かけることが少なくなりました。

３　推定

推定とは、標本での計算から得た標本の特徴（標本の統計量）から、母集団の特徴（母集団の統計

量）を推し量ることを、統計学的な推定といいます。

・抽出と推定の関連

抽出と推定は統計学ではセットにして使うことが多い考え方です。教科書の始めの方、5 頁から 8 頁の記述を読み直してください。母集団を大きなスープの鍋、標本をそこから取ったスプーン一杯として説明されています。どのようにして大きなスープ鍋からスプーン一杯を取るかの方法が「標本抽出」、スプーン一杯の味からスープ鍋全体の味付けを予測するのが「推定」という関係が示されています。

・点推定

具体例で説明します。皆さんが本当に知りたいのは操作的母集団。皆さんの先輩におけるアルバイトの時間だとします。しかし皆さんの先輩は人数が多すぎて、全員のデータから計算するのは困難だとします。でも my 標本 12 名であれば、すぐに計算でき、アルバイト時間の平均がある値（たとぇば 70 分）になったとします。この値（70 分）をもって、皆さんの先輩のアルバイト時間を 70 分と推測したとき、その値を求める行為を「点推定」といいます。推定した結果を、一点の値として表わすのが、点推定です。教科書の 91 ページを見ると「適切な標本抽出法を用いて、標本平均は、母集団の平均の点推定値として最も優れているはずである」とあります。最も優れていると言っても、皆さんはそれぞれに my 標本を持っているわけですから、点推定値も my 標本によって、同じではありません。my 標本ごとに、やや異なる点推定値が得られることになります。

・標準誤差

皆さんの標本から得た平均の点推定値は、操作的母集団の平均に対して、どのくらいの誤差を持っているでしょうか。この誤差を標準誤差といいます。標準誤差の計算式は以下のとおりです。

$$\text{標準誤差} = \frac{\text{標準偏差}}{\sqrt{n}}$$

・区間推定

点推定によって平均を推定したとき、その推定値は、推定の誤差を考慮しなければならないために確率的な表現が必要です。つまり点推定のように特定の点を指定するのではなく、真の平均値は値 A と値 B の間におそらく含まれている、という考え方です。これを区間推定といいます。区間推定、特に 95%信頼区間は以下の式で表わされます。

$$\text{95\%信頼区間} = \text{標本平均} \pm 1.96 \times \text{標準誤差}$$

4　操作的母集団をどう分析するか？

さてここまでは、12 名からなる標本（my 標本）を中心に分析を考えてきました。この my 標本でレポートを書くことも出来ます。

しかし、皆さんは my 標本だけでなく、すでに操作的母集団全体のデータも利用することができます。12 人ではなく、150 人のデータ全てを使ってレポートを書くことも可能です。

しかし、標本から操作的母集団へと視点を移して分析するためには、それなりの方法を用いる必要があります。

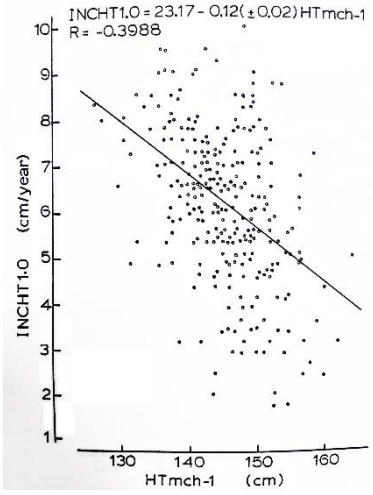
my 標本は 12 名と数が限られたデータでしたので、じっくりと一人ひとりのデータを観察し、質的な考察も可能でした。しかし 150 人以上ものデータだと、じっくりデータをひとり一人見る方法は時間がかかりすぎます。もっと早く、データの全体像を把握する必要があります。この目的で、もっともよく使われるのがエクセルです。皆さんは昨年、情報学でエクセルを使っているので、基本は理解しているはずです。

エクセルを使う場合、特に便利なのは、ピポットグラフの方法です。これを上手に使うと、簡単に度数分布表や棒グラフやクロス集計表を作ることが出来ます。

さて、皆さん全員がエクセルを使える環境にあるのであれば、エクセルの話をするのですが、どうでしょうか。現実には、パソコンやエクセルが利用できず、スマートフォンやタブレットで学んでいる皆さんも多いと思います。

そこで、次回からのオンデマンド授業でも、スマートフォン中心に話をしたいと思います。スマートフォンだけで、適切なサイトを使えば、様々な統計計算が可能です。しかし、もしエクセルが利用可能であるなら、昨年学んだ知識を活用して、ぜひエクセルを用いた計算にもチャレンジしてください。

5　手書きして考えることの大切さ



それから、ネットやコンピューターに頼るだけでなく、自分でグラフ用紙に手書きで点をプロットして考えることも大切です。

画面に示すのは、私が 40 年前、ネットやパソコンが無い時代に、手で書いた散布図です。１枚の図を描くのに何日もかかりましたが、図を描きながら、いろいろなことを考えました。現在は、エクセルで一瞬にして図が描けてしまうため、あのときほどは、深く図を見て考えることをしなくなったかもしれません。それから、回帰と相関の授業のときに紹介したゴルトンの図も思い出してください。ゴルトンの時代は、コンピューターはもちろん、機械的な計算機も存在しない時代です。あのとき、ゴルトンは多くのスイートピーの種の重さを測り、手書きでグラフを書きました。そして回帰という現象を発見したのです。

そういえば、皆さんの先輩にも、レポートにグラフを手書きした人がいました。資料の中に入れていますので、参考にしてください。

演習問題

1. 標本、操作的母集団、概念的な母集団について動画とは別な例を考え、50 字以内で書いてください。

2. 自分でいくつか乱数を発生させた上で、どのような乱数が現れたか、それを見て何を感じたか、何に用いたいかなど、80 字以内で書いてください。あなたのスマホで乱数が発生できない場合は、以下のサイトを利用してください。

   https://keisan.casio.jp/exec/system/1425449868

3. あなたの my 標本 12 名のデータには、身長、体重、睡眠時間など、いくつかの連続量が含まれています。どれか一つ連続量を選んで区間推定を行い、得られた 95%信頼区間を以下に記してください。

4. あなたはレポートを書く際に、どのような手段で計算を行う予定ですか？またどのような計算手段に関心がありますか？もし将来、今回のレポートを書く際にも、本格的な統計パッケージを利用したいなら、JASP がお勧めです。

   https://jasp-stats.org/

   しかし JASP を使うには、あなたの自由になるパソコンが必要で、また基本操作は英語で行う必要があります。いったん使用方法を習得したら、エクセルよりもずっと容易に高度な統計計算ができます。（JASP に関心のある人が多ければ、次回以降の授業で基本を説明します）

   以下、複数回答が可能です；ワークシートで手計算したい／電卓で計算したい／js-STAR などネット上の計算サイトを利用したい／パソコンでエクセルを使いたい／パソコンで JASP など本格的統計パッケージを利用したい。

第 11 回　二群の比較と t 検定

https://youtu.be/nR2eD1pfIBw

皆さんこんにちは。すでに第５回目の授業で「世界を二つに分割する考え方;ダイコトミーDICHOTOMY、二分法」についてお話ししました。この二分法の考え方と関連して使われることの多い検定 T 検定についてお話しします。

１　t 検定の発想

t 検定は皆さんのマイ標本についても様々な形で適用できます。何れかの離散量（アレルギー有り／アレルギー無し、好き嫌い多い／好き嫌い少ない、スポーツ苦手／スポーツ好き）などで二群に分けた上で、何れかの連続量（身長、体重、睡眠時間）の平均値に差があるかどうかを検定する方法です。「平均値の差の検定」といいます。

t 検定ではこれまで学んだ様々な統計の考え方を総合的に用いますので、ここで復習しておきます。

・平均を比べるとはどういうことか

t 検定では二群の平均を比べます。しかし、みなさんのｍｙ標本 12 名について t 検定を行うとして、12 名だけの中で、ただ比較するわけではありません。12 名中、アレルギーありが６名、アレルギー無しが６名だとして、６名と６名の平均を、ただ差をとって比べるだけなら、小学校の算数計算です。他方、統計的な検定で、常に考えているのは、母集団のことです。統計的に比較するとは「ある６名から推定される母集団（アレルギーありの人々）の平均」と「別の６名から推定される母集団（アレルギー無しの人々）の平均」の比較を意味します。みなさんの my 標本はたった 12 名で構成されていますが、そこから得られる二つの平均は、それぞれ母集団を想定した際の推定値（点推定の値）であることを、忘れないでください。

２　t 検定の歴史

t 検定についてよりよく理解できるように歴史をお話しします。 t 検定といえば、普通は「スチューデントの t 検定」を意味します。この名称は 1908 年に t 検定の論文を書いたウィリアム・ゴセットのペンネーム student に由来します。当時、ゴセットはアイルランドにあるギネスのビール醸造所で研究者として働いていました。会社の方針で実名による研究発表ができず student というペンネームを使いました。当時のゴセットの仕事は黒ビールの質をモニターするための経済的な方法を見つけることで、t テストの考案に至りました。その発想は、どこまで標本数を小さくできるか、例えば一つのビール醸造タンクから抽出する標本数をわずか 3 としても、そこから計算される平均などの値から、タンク全体のビールの品質が推定可能か、といったものでした。そして「正規分布する母集団からいくつもの標本を抽出したときに、その標準偏差はどのような分布になるか」といったテーマで論文を書きました。この試みからスチューデントの t 分布が得られました。 t 分布や t 検定は、その後、統計学者ロナルド・フィッシャーの紹介により世界に広まりました。

「スチューデントのｔ分布を用いたスチューデントのｔ検定」はとても広く使われる方法となり、スチューデントを省略して、単にｔ分布、ｔ検定といわれることも増えました。また、当初スチューデントのｔ検定が扱ったよりも複雑な条件にまでｔ分布を適用して検定を行う場合も現れました。その一例としてウェルチのｔ検定があります。

３　ｔ検定、計算の考え方

・定義；　２つの母集団がいずれも正規分布に従うと仮定したうえでの、平均が等しいかどうかの検定。

・帰無仮説；「二群の標本から推定される母平均に差がない」

・分類；　ｔ検定は比較する標本の在り方によって、以下の場合に分かれます。

１）二つの標本がペア（対）の場合；　同じ人に前後２回調査など。

２）二つの標本が独立で、等分散の場合；　二つの標本の分散が等しいと仮定できる。

３）二つの標本が独立で、等分散性ではない場合；異分散。（この場合はウェルチのｔ検定）

・計算方法概要；

２群の標本から計算した母平均の推定値の差が、その標準誤差の何倍離れているかを計算します。計算式は教科書 125 から 130 頁を参照してください。

計算演習　以下、エクセルと js-STAR についてお話しします。

４　エクセルについて

１）計算の準備

エクセルではさまざまな計算ができますが、分析ツールを使うにはまずその設定が必要です。

・まずエクセルを開き、画面一番上の左にある「ファイル」タブをクリックします。

・次は左下に表示される「オプション」をクリックします。

・エクセルのオプションという頁が表示されたら、左にあるメニューの下の方にある「アドイン」をクリックします。

・次の画面で現れる一番下の管理ボックスで「Excel アドイン」を選び、「設定」をクリックします。

・アドインのボックスが現れますので「分析ツール」にチェックを入れＯＫをクリックしてください。

・以上で、分析ツールの読み込みが完了します。

２）データの準備と整理

計算の事例は、あなたのｍｙ標本から選んでください。ここではｍｙ標本の以下のデータを用います。

ｉｄ	体重	バイト
1	49	2
2	48	2
3	44	1

4	42	2
5	49	1
6	50	2
7	46	2
8	56	1
9	40	2
10	50	1
11	54	1
12	52	2

帰無仮説は「バイトする；１」群と「バイトしない；２」群との間で、体重の平均値に差がないです。

マイ標本における実際の値は「バイトする人」と「しない人」が混在して並んでいます。T 検定で二群を比べる場合には「バイトする群」と「しない群」をはっきり二つに分けなければなりません。

このような時使うエクセルの基本操作として「データ並べ替え」があります。並べ替えるのは、群分けの基準にした変数「バイトするかしないか」です。バイト１の人が最初に、バイト２の人はその後に来るように、データを並べ替えます。並べ替える時は忘れずに「選択範囲を拡張する」を選んでから行なってください。選択範囲を拡張するとは、そのデータだけではなく関連するデータも一緒に並べ替えるという意味です。並べ替えることで、二群を整理できました。

ここまで作業をした上で、先ほど設定した分析ツールを使います。

ID	体重	バイト
3	44	1
5	49	1
8	56	1
10	50	1
11	54	1
1	49	2
2	48	2
4	42	2
6	50	2
7	46	2
9	40	2
12	52	2

・すでに設定したエクセルの分析ツールを使うためには、まずエクセルの画面上部にあるメニューからデータのタブを選びます。
・すると、上の右端に分析というタブが現れるので、それをクリックします。
・すると、分析ツールのボックスが現れるので、メニューから「ｔ検定；等分散を仮定した 2 標

本による検定」を選びＯＫを押します。

・するとｔ検定のボックスが現れるので「変数１の入力範囲」およ「び変数２の入力範囲」を指定します。変数１の入力範囲とは、バイト１の条件の人の「体重」が入っているカラムです。

・変数２の入力範囲とは、バイト２の条件の人の「体重」が入っているカラムです。

・両方を指定してＯＫを押すと、直ぐに計算結果が表示されます。結果を見ると、まず両群の平均値・分散・標本数が現れます。その後統計量が並びます。変数１の自由度は 5 から 1 を引いて 4、変数 2 の自由度は 7 から 1 を引いて 6、このｔ検定の自由度は変数１と変数２の自由度を合計して 10 となります。

・帰無仮説は「両群の標本から推定される母平均に差がない」です。

・ｔの値は１．４８１５５と計算されました。

５　片側・両側について

初めて出てきた言葉があるので説明しておきます。表の中に両側、片側との記述があります。

両側検定とは二群の平均値の差を比較する時に、どちらが大きい可能性があるか、事前に全くわからない時に使う検定の考え方です。他方、二つの平均値を比較するといっても「もし差がある場合は「必ずａ群の方が大きくなる」など事前に「差の方向性」を予測できる稀な場合があります。このような時に片側検定を使います。

今回はどちらが大きいかなど予想できませんので、両側検定を使います。

この表には両側検定と片側検定のとちらにも対応できるように、同じｔ値に対して二つの有意確率が示されていますが、今回は０．１６９２を採用します。

さてコンピューターで検定をした場合は数表を見るまでもなく直接にｔ値と対応する有意確率が計算されます。有意確率が０．０５以下であるなら、帰無仮説は棄却できます。しかし今回計算した有意確率は０．０５より遥かに大きい値です。よって帰無仮説は棄却されず「二つの平均値の間には有意な差が認められない」と結論されます。

６　ｊｓ－ＳＴＡＲによる分析

js-STAR はパソコンがなくてもスマートフォンから使えることはお話ししました。

http://www.kisnet.or.jp/nappa/software/star/

js-STAR のサイトを開くと左側にメニューが出てきます。ｔ検定には二つ選択肢がありますが、その最初の方を選びます。

まず現れるのは、データ数が二つだけの単純な表です。この表は必要に応じて縦方向に拡張して使うようになっています。２と出ている部分に、実際の比較したい標本の人数を入力します。先ほどの例で言えば第１群の参加者数は５、第２群の参加者数は 7 です。二つの数値を入力すると、表が縦に伸びて値が入力できるようになります。ここに比較したい変数の実測値、体重の数値を 12 人分入力します。両群を分け、まず最初に群１の値、続いて群２の値を入力ます。

入力を終えて「計算のキー」を押すと結果が現れます。

この結果で注意すべきは、ウェルチのｔ検定を行っている点です。Student のｔ検定は、本来、等分散を仮定した条件で生まれたのですが、その後、等分散の仮定を設けなくても適用できるウェルチの T 検定が生まれました。ウェルチのｔ検定は計算がやや複雑だったのですが、コンピュータの

発達で簡単に計算できるようになったため、使用が増えています。js-STAR ではこのような現状を考慮して最初からウェルチのｔ検定を計算してしまいます。ウェルチのｔ検定は教科書１２９頁に説明があるので、チェックして下さい。

さて今日は二つの平均を比較するｔ検定の考え方をお話ししました。T 検定には様々な統計の考え方が反映されています。是非復習しておいてください。

演習問題

1. 二群を比較するｔ検定は、とてもよく使われる検定の方法です。あなたのｍｙ標本を観察し、どのデータでｔ検定をしてみたいかを、50 文字以内で書いてください。

2. エクセルを用い、あなたのｍｙ標本で何かｔ検定を行ってください。結果は 50 文字以内で書いてください。エクセルが利用できない場合は、動画中のエクセルの説明を見て、感じたことを 50 文字以内で書いてください。（今エクセルを使えなくても、登校禁止が解除されて大学に来られるようになったら、ぜひ情報処理室でエクセルに触れてください。）

3. js-STAR を用い、あなたの my 標本データでｔ検定を行ってください。
すでにエクセルで計算済みの場合、同じデータで構わないので、ぜひ js-STAR でも計算を試みてください。同じデータを用い、複数の方法で計算を行ってみることで、各方法の特徴を把握でき、また各方法の限界も理解できます。結果や気づいた点を 50 字以内で書いてください。

第 12 回　分散分析と F 検定

https://youtu.be/t8xnSNeeg04

皆さんこんにちは。今回は分散分析についてお話しします。

１　分散分析の考え方

１）データのばらつき・変動から出発

分散分析は「偏差二乗和（ＳＳ）と分散（ＶＡＲ、σ２）」が大活躍する分析方法です。偏差二乗和や分散って何か覚えていますか。データの「ばらつき・変動」を示す基本的な値です。第 3 回目の授業でお話しました。実測値と平均の差、偏差もデータのばらつきを表わすのですが、集団全体について偏差を合計するとゼロになってしまいます。そこで、ゼロにならないように二乗して合計したのが偏差二乗和（ＳＳ）、その平均が分散（σ２）です。ワークシートで計算したことを思い出してください。

ただ一つ重要な相違点があります。それは、今回の分散分析で用いる分散は、実は第 3 回目の授業での分散とはやや異なり、偏差二乗和を割り算するとき、データ数 N ではなく自由度 n-1 を用いるという点です。

２）分散分析の種類

検討する要因が一つだけの場合が一元配置分散分析です。同時に検討する要因の数が増えると、二元配置分散分析、多元配置分散分析などになります。

以下では、もっとも一般的な「一元配置分散分析」を取り上げます。

３）何に使うか

分散分析は統計学を使って本格的に研究をするときに役立つ方法です。探索的な使い方と実験計画的な使い方とがあります。

・探索的な使い方 ; 何らかの観測データが得られた場合、そのデータのばらつき・変動は、どのような要因によって引き起こされているか、可能性のある要因を探索するような使い方です。例えば新型コロナウイルスによる発症の重症度というデータがあったとします。それはどのような要因で影響を受けるのか、年齢・性別・栄養状態・民族的などの要因から、一つまたは二つを選び、要因でグループ分けし、グループ間やグループ内で、要因の効果をデータの平均や分散を用いて比較します。

・実験計画的な使い方 ; あらかじめいくつかの処理・条件を設定した実験計画を立て、実行したときに、処理群間や処理群内で効果を、平均や分散を用いて比較します。たとえば、新たな肥料Ａ、Ｂ、Ｃについて、また新たな治療薬Ａ、Ｂ、Ｃの効果を比較するなどの使い方です。

・コンピューターが現在ほど進歩するまでは、分散分析は基本の方法でした。その後コンピューターが進歩し計算が容易になるにつれて、分散分析は、他のより大量な計算を必要とする方法と組み合わせて使われることが増えています。その結果、最近の統計学の教科書では分散分析という項目が登場しない場合もあります。

４）なぜ分散に注目するか

なぜデータのばらつき・変動を示す偏差二乗和や分散が、分析や研究の出発点になるのでしょうか。それはデータ全体を何かの要因でグループ分けしてグループ別（群別）の分散を求めた時、その要因によるグループ分けが意味を持っている程度に合わせて、「データ全体の分散」に対し「要因によって説明できる分散」の占める割合が、変化するからです。他方、何らかの要因でグループ分けをしたとしても、そのグループ分けによって説明できる分散の部分が少なければ、グループ分けしたことに意味がない、ということになります。

２　グラフで考える

分散分析の意味を直感的に把握できるよう、グラフで説明します。「COVID-19 禍のもとでのオンデマンド授業についての学生の皆さんの学習意欲」を出発点となるデータ全体として考えてください。まず、学習意欲を０から１００までの数値として、その分布を茶色のヒストグラムで表わします。学習意欲は低い人から高い人まで様々です。この学習意欲はどのような要因によって説明されるでしょうか。いくつかの要因でグループ分けして考えてみます。

図１では、学生の皆さんがオンデマンド学習を行う主な時間帯で、早朝型・午前型・午後型・深夜型とグループ分けしました。時間帯が早い方が遅い方に比べて、学習意欲の平均が高い傾向が認められます。他方、どの型でもデータのばらつき・変動の幅が広めで、４つの分布は相当に重なっています。

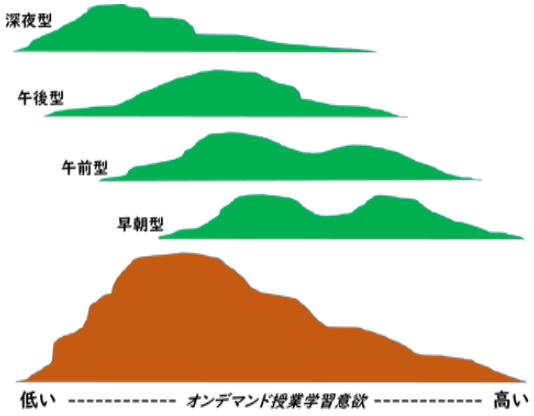


図２では、オンデマンドの動画を見ている方法でグループ分けしました。パソコンよりもタブレット端末が多少学習意欲の平均が高いようです。スマートフォンを使う人と、幾つかの方法を複合して使う人とでは、差がはっきりしません。他方、どの方法でもばらつき・変動の幅が広く、全域にわたって重なっています。

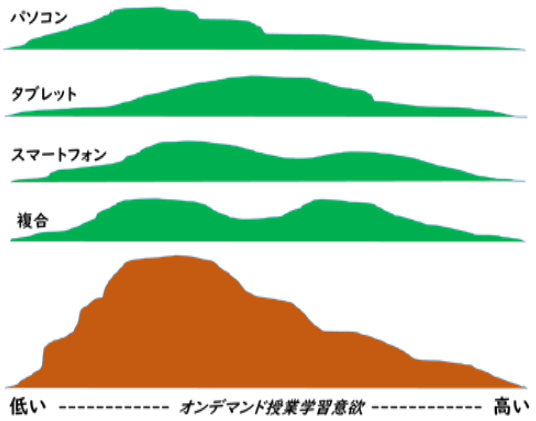
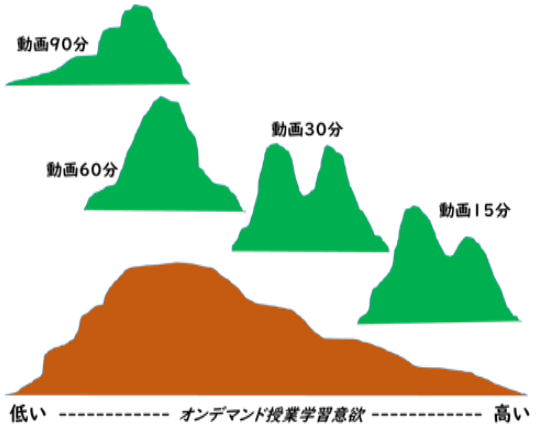


図３では、動画の時間の長さでグループ分けしました。ここで大きな差が出てきました。長い動画を見ている人は、動画が途中で途切れたり長すぎて注意が散漫になるのか学習意欲が低くなっています。他方、動画の時間が 15 分と極めて短いと、見る時の集中力が高まるせいか、学習意欲の平均が高くなっています。



ここで平均に明らかな差があるということも大切ですが、それよりも注目すべきは、四つのグループがそれぞれの平均と狭い範囲でのバラツキ・変動を示し、明らかな特徴を持つグループとして、際立っていることです。学習意欲という元データの全変動（ＳＳＴｏｔａｌ）が、主として、動画の時間という要因による変動（SS Treatment）で説明できることが、読み取れます。特に短い動画で学習意欲が高いということであれば今後より短時間に集中して学生の皆さんに多くのことを考えてもらうような形の動画を作成することが大切であるとわかります。

３　計算演習

前項の動画の長さを例にした実験計画の例です。15 名の学生をランダムに 3 群に分け、各群に異なる長さの動画を視聴してもらい、視聴後に学習意欲（1～100 点）を測定し、表の結果を得た。動画の長さが学習意欲に影響を与えているか、平均に差があるかを、有意水準 0.05 で検定しなさい。

https://youtu.be/Ok36LwINy5w

動画 10 分	動画 30 分	動画 90 分
80	60	50
85	75	55
70	52	70
78	58	40
60	68	45

１）エクセルを用いる場合

・すでに設定したエクセルの分析ツールを使うためには、まずエクセルの画面上部にあるメニューからデータのタブを選びます。
・すると、上の右端に分析というタブが現れるので、それをクリックします。
・すると、分析ツールのボックスが現れるので、メニューから「分散分析：一元配置」を選びＯＫを押します。
・すると「分散分析：一元配置」のボックスが現れるのでまず「入力範囲」を指定します。
・各群のデータは縦方向に並んでいますので、「列」を選びます。
・入力範囲の先頭行はデータではなく、「動画 90 分」など群の名前ですので、「先頭行をラベルとして使用」もチェックします。
・指定してＯＫを押すと結果が二つの表で示されます。
・概要の表には、三つあった群別の標本数、合計、平均、分散が示されます。
・次の分散分析表で、まず変動とあるのは、偏差二乗和（ｓｓ）です。データ全体の SS が。グループ間（群間）の SS とグループ内（群内）の SS に分けて示されます。
・表の自由度の部分を見ると「２、１２、１４」と数値が三つ並んでいます。１４は「全体の自由度」、「全標本数－１」で計算されます。他方、２は要因の群分けに関する自由度で、「要因の自由度」とよび「グループ数－１」で計算されます。「全体の自由度」から「要因の自由度」を引き算したものが「残差の自由度」です。
・F 分布の形は二つの自由度で規定され、印刷された F 分布表を見るときも、二つの自由度を用意する必要があります。

・F 値とは「グループ間の分散」を「グループ内の分散」で割り算したのが「観測された分散比」、これが F 値です。
・F 値 6.20446 に対応する有意確率（P 値）は 0.014119 です。よって、今回の計算例ではｐ値（有意確率）は、有意水準を５％（０．０５）と設定した場合には、それより低い値をとるため、帰無仮説は棄却されます。

２）ｊｓ－ＳＴＡＲによる場合

・js-STAR のサイトを開くと左側にメニューが出てきます。
・js-STAR には分散分析として多くの選択肢が用意されていますが、その最初、Ａｓ（１要因参加者間を選びます。
・まず現れるのは、群が二つ、各群の参加者二名だけの単純な表です。この表は条件に応じて縦方向に拡張して使うようになっています。２と出ている部分に、実際の比較したい群の数、各群の参加者人数を入力します。先ほどの例で言えば第１群から第３群まで、各群の参加者は５名です。
・条件を入力すると、表が縦に伸びて値が入力できるようになります。ここに比較したい各群のデータ（学習意欲）を群別に入力します。
・入力を終えて「計算のキー」を押すと結果が現れます。
・js-STAR の計算結果は、表示窓には計算結果の一部分しか出ていませんので、カーソルを動かして最初の部分から結果を見ていきます。
・先ほどのエクセルでの計算とほぼ一致する結果が得られたことを確認してください。
・エクセルでは「観測された分散比」となっていた項目は、F（F 値）として示されています。
・F 値６．２０の横に＊印がついています。これはその下の説明によると「＊P＜．０５」つまりこのＦ値（６．２０）に対応する有意確率（ｐ値）は０．０５（有意確率）より小さいこと、よって有意水準０．０５で帰無仮説は棄却されることを意味します。

４　分散分析の背景

１）分散分析の歴史

分散分析が形になったのは 1918 年にロナルド・フィッシャーが分散という用語を導入してからです。その後分散分析はフィッシャーが 1925 年に書いた本を介して、広く世界に知られるようになりました。しかし分散分析に至る考え方は何世紀にもわたって育まれてきたとされ、統計学に関する様々な考え方「仮説検定、二乗和の分割、実験手法、加法モデル」などが含まれています。

特に「全体の変動（偏差二乗和、SS）をグループ内 SS とグループ間 SS に分解する」、「各 SS を自由度で割り算して分散を求める」、「二つの分散の比を F で表わす」という論理は、計算方法がシンプルで分かりやすく、多くの場面で用いられました。

２）F 分布の歴史

F 分布は分散分析を行うときの統計量として知られています。F 分布においては、この考え方への貢献が大きい 2 人の統計学者、スネディガーとフィッシャーの名前をつけて、F-distribution、Snedecor's F distribution、the Fisher-Snedecor distribution などと呼ばれています。

「全体の変動（偏差二乗和、SS）をグループ内 SS とグループ間 SS に分解する」を数式で書くと以下のようになります。

ＳＳtotal = SSError + SStreatments

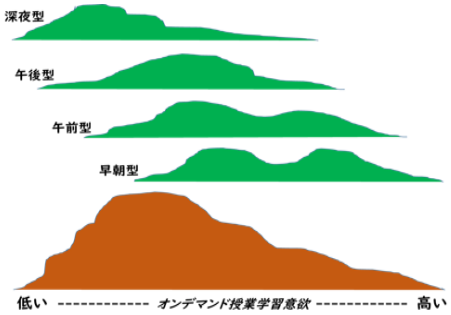
なお、元データの全変動中、要因による変動を除いた残りの変動がＳＳerror です。

また分散分析で使うＦ値とは、要因による変動の分散を、残りの変動の分散で割り算したものです。

３）質的研究と分散分析、発想の違い

看護研究では研究方法として質的な研究の方がよく使われる傾向にありますが、特にこの分散分析の方法は、質的な研究の研究者にとっても興味深い研究方法だと考えられます。研究方法としての特徴は、名前にも現れている通り、分散に注目して研究する方法です。

質的な研究の場合は少数の人々の思考や行為における意味に注目します。他方、分散分析では、授業第 2 回目で平均値や標準偏差などと共に説明した基本的な統計量「分散」に注目します。質的な研究が個人的な現象の意味から入ることが多いのに対し、分散分析では意味よりも、社会集団における現象の分布の形に注目します。

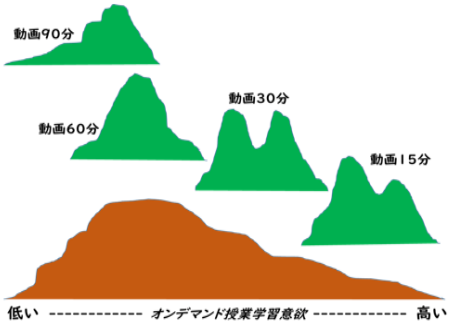


先ほどの分布図を思い出してください。

Ｆ値や有意確率は計算しないと求められません。

しかし図を見れば、どちらの群分けの方が現象を解明する手がかりが得られるか、明白ですよね。生活時間帯で分けた上の図では分布の重なりが多く、全体の変動から各要因の変動を明らかに取り出すのは困難です。

他方、下の図では、動画の長さで群分けすることで、各動画の変動がはっきりと分離できます。この場合にＦ値を計算したら、おそらく、確実に帰無仮説は破棄され、４群の間に差があると結論できるでしょう。



５　まとめ

さて今日は分散分析についてお話ししました。

分散分析は統計学の考え方が最もよく現れた分析方法です。みなさんもぜひ試してみてください。

演習問題

1. 分散分析は統計学の代表的な分析方法です。あなたはどのような課題に分散分析を使ってみたいですか。50 字以内で書いてください。

2. 動画で分散分析の計算方法を学んでください。エクセルを用い、my 標本で何か分散分析を行い、結果を 50 文字以内で書いてください。（my 標本ではなく 150 名データを用いても構いません）エクセルが利用できない場合は、動画中のエクセルの説明を見て、感じたことを 50 字以内で書いてください。（今エクセルを使えなくても、登校禁止が解除されたら、ぜひ情報処理室でエクセルに触れてください。）

3. js-STAR で分散分析を行ってください。

http://www.kisnet.or.jp/nappa/software/star/

すでにエクセルで計算済みであっても、js-STAR でも計算を試みてください。同じデータを用い、複数の方法で計算してみることで、各方法の特徴を把握でき、また各方法の限界も理解できます。結果や気づいた点を 50 字以内で書いてください。

第 13 回　回帰分析

https://youtu.be/mCIKpeYqf0w

皆さんこんにちは。今回は回帰分析についてお話しします。

二つの変数 X と Y の回帰や相関の考え方、ワークシートを用いたピアソンの積率相関係数の計算については、すでに第 4 回目の授業でお話ししました。今回の回帰分析は、目的変数 Y と一つ又はそれ以上の説明変数 X の間の関係を推定するための統計的な考え方です。最も一般的な形は線形回帰と呼ばれます。

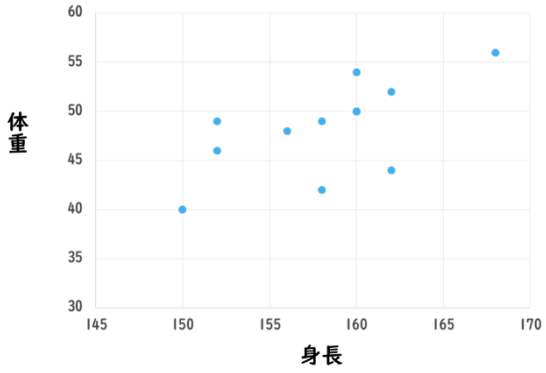
1　回帰分析の目的

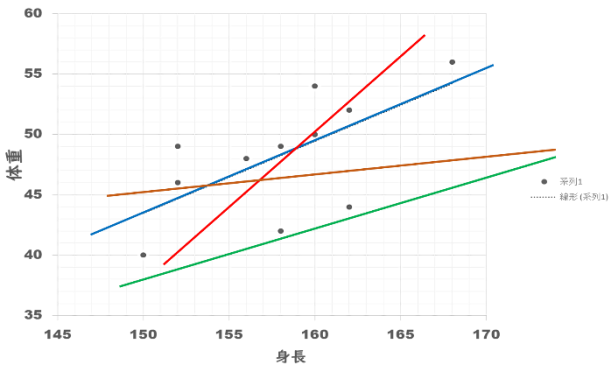
回帰分析は主に二つの目的で使われます。1 番目は予測です。線形回帰の考え方で y と x の間を数式で表すことができれば、x から Y の値を予測できます。Y は「予測したい変数」であり「目的変数」「従属変数」と呼びます。x は「予測に用いる変数」であり「説明変数」「独立変数」と呼びます。学生のみなさんはｍｙ標本を用いて、たとえば身長から体重を予測するとか、勉強時間をアルバイトの時間から予測するとか、関心のあるテーマで回帰分析を行ってみてください。

2 番目の目的は、因果関係の推測です。X と Y の間に相関関係が認められる場合、それが直接に因果関係を意味することはありません。しか相関関係は、因果関係の可能性を示します。y を「疾病」、x を「疾病の原因」と考えて、x と y の関連性をより具体的に考えていくのは、皆さんが後期に学ぶ疫学の中心です。

2　回帰式の求め方

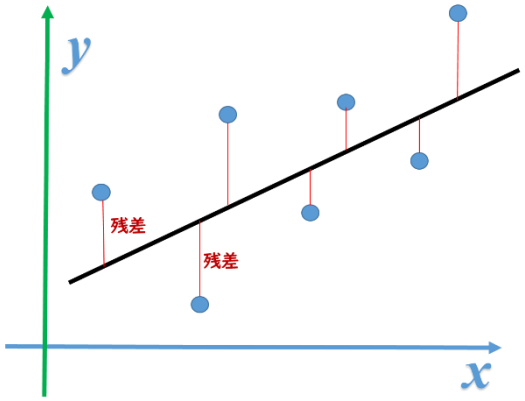
さて、観測されたデータに基づいて、ｘからｙを予測する回帰式はどのように導いたらいいでしょうか。あるｍｙ標本での身長と体重の関連性を図に示します。身長 X が増えると体重 Y が増える関連性が認められます。ではこの X と Y との関連性を代表する一本の直線の存在を決定するにはどうしたらよいでしょうか。このような時に使うのが最小二乗法という考え方です。最小二乗法は、回帰分析の中でも最も基本的なものです。さらに考えを進めるために、図に５本の直線を引いてみました。この５本の直線では、どの直線が最もふさわしいでしょうか。





最小二乗法では、測定データＹは、モデル関数と誤差（残差）の和で表わされます。モデル関数が測定データにどのくらい適合するかは、残差で判断できます。

残差は測定データからモデル関数の値を引いたものです。残差は正、または負の値を取るため、負の値を取らないように残差を二乗し、残差二乗（残差平方）とした上で、その合計、つまり残差二乗和が最小になるように連立方程式を解いてモデル関数を決定するのが、最小二乗法です。



3　回帰分析の歴史

最小二乗法は回帰分析の原型とされる重要な計算方法です。18 世紀の間に天文学や数学の分野で行われた進歩がこの方法に結集しました。最小二乗法の論文は 1805 年にフランスの数学者ルジャンドルが公開し、それより前に最小二乗法を見つけたというドイツの数学者ガウスと論争になりました。ガウスはさらに考えを発展させて確率の原理に結びつけ、またガウスはこの過程で正規分布を発明しました。19 世紀の初めに話題になっていた小惑星セレスの軌道をガウスはこの最小 2 乗法で正確に予測したことが知られています。

19 世紀にフランシス・ゴルトンがスイートピーの親種と子種の直径を比較する研究から見出した傾向を表わすために、回帰、regression という用語を用いたことは、すでに第 4 回目の授業でお話ししました。この回帰の概念がその後発展し、最小二乗法と結びつき、回帰分析として知られるようになりました。

回帰分析は 20 世紀半ばまでは膨大な計算を必要とするため、コンピューターが一般化する前、1970 年代以前は機械式計算機で一つの結果を得るまでに、24 時間かかる場合もあったとされます。その後コンピューターの普及とともに回帰分析は発展し、今では経済学・天文学・医学など様々な分野で使われています。

4　エクセルでの計算

計算の事例はあなたのｍｙ標本から選んでください。ここではｍｙ標本から取った以下のデータを用います；

Y;ネットテレビ時間、一日当たりのネットやテレビ視聴にかける時間（分）

X;予習復習時間、一日当たりの予習復習にかける時間（分）

予習復習	ネットテレビ
60	90
70	60
0	120
60	60
80	70
30	100

一次回帰の予測式は Y＝α＋β x です。Y は「目的変数（従属変数）」です。 x は「説明変数（独立変数）」です。 α と β を回帰分析で求めます。

・すでに設定したエクセルの分析ツールを使うためには、まずエクセルの画面上部にあるメニューからデータのタブを選びます。
・すると上の右端に分析というタブが現れるので、それをクリックします。
・すると分析ツールのボックスが現れるので、メニューから「回帰分析」を選びＯＫを押します。
・すると回帰分析のボックスが現れるので「入力 Y 範囲」および「入力 X 範囲」を指定します。
・入力 Y 範囲とは、回帰分析で予測したいｙ、今回の例ではネットテレビ時間が入っているカラムです。入力ｘ範囲とは、回帰分析で予測に用いる X、予習復習時間が入っているカラムです。
・両方を指定してＯＫを押すと直ぐに計算結果が表示されます。結果は三つの表に分かれて示されます。

1）最初の表、回帰統計では相関係数が 0.8907 と示されています。相関係数の前に「重」という文字がついています。その理由は、回帰分析で説明変数の数が複数になると、相関係数も複数の変数から計算することになり、その際は単に相関といわず、重相関というからです。回帰分析を行う前に、説明変数ｘと目的変数 Y との間に相関があることを確認するのは重要です。X と Y の間に相関が認められなければ、回帰式を立てることが意味を持ちません。

2）二番目の表は分散分析です。前回の授業で学んだ分散分析は、実はこのように、回帰分析の中にも組み込まれています。データ Y の全変動を示す偏差二乗和 2933.33 に対し、回帰式で説明できる変動（偏差二乗和）が 2327.27、それを引き算した残差の変動が 606.06、そして「観測された分散比」である F 値が 15.36、この F 値に対応するｐ値（有意確率）が 0.01726 です。有意水準を５％（0.05）と設定した場合、ｐ値はそれより低い値をとるため、帰無仮説は棄却されます。棄却される帰無仮説とは「回帰式の予測による誤差（残差）の減少は無い」です。帰無仮説を棄却すると、「回帰式による予測が誤差を減少させる」と認めることになり、回帰式の予測が意味あるものと位置付けられます。

3）三番目の表では、回帰式のαとβを具体的な数値として求め、またその区間推定値も示します。最終的に得られる予測式は Y=119.697-0.7273X となります。以上がエクセルによる回帰分析です。エクセルではこの他に散布図や回帰直線を描くこともできます。その方法をお話しします。

- まず先ほど回帰分析を行ったデータのあるセルをもう一度範囲選択します
- エクセルの画面上部にある挿入のタブを選びグラフのグループにある散布図のボタンをクリックします。すると散布図が描けます
- 散布図を描いたらエクセルの画面上部「デザイン」のタブを選び、上部左のクイックレイアウトを選ぶと散布図に様々な情報を付け加える選択肢が画像ボタンで表示されます。
- 一番上の行の 3 番目にある画像ボタン（直線を描く）を押すと、先ほど散布図に回帰分析で得た直線を描き込むことができます。

５　Casio、Linear regression Calculator

エクセル以外の計算方法として今回は Casio、Linear regression Calculator を紹介します。

https://keisan.casio.com/exec/system/14059929550941

- サイトにアクセスしてすぐに現れる表にはすでに X と Y の値が例として書き込まれています。
- まずこれを下のクリアキーを押して消去します。
- 空白の表にしたら、そこに X と Y の値を入力します。

- 入力が終わったら、計算実行を意味する EXECUTE のキーを押します。
- そうするとXとYそれぞれの平均値、相関係数そして回帰式のアルファとベータの値が表で表示されます。
- Excel の場合とほぼ同様の式 Y=119.6970ー0.7273ｘが得られたことを確認してください。このサイトでは同時に直線のグラフも描いてくれます。

６　まとめ

さて今回の授業までで、学生の皆さんにお伝えしたい統計学の基本的な考え方は、ほぼお話ししたように思います。統計学の教科書を見ると、私がこれまでにお話しした以外にも様々な方法が示されています。様々な方法の名前を知ることも無駄ではありません。でも統計学は暗記科目ではありません。統計は、皆さんがこれから日々の生活や仕事や研究の中で、それらを実際に使って物の見方考え方を広げていくための道具です。

これまで皆さんにお話しした基本的な方法、特にクロス集計表とカイ二乗検定、二つの平均の差のｔ検定、分散分析と F 検定、回帰分析などに親しみ、それらを用い始めれば、まだお話ししていない他の統計の方法についても、徐々に理解し使いこなせるようになると思います。

私が自身の仕事の中で、これまで最もよく用いていたのは回帰分析です。「あることから別なことを予測する」というのは興味深く、特に間の発育に関連して回帰分析を使ってきました。人は子ども時代から思春期に入った時に、身長や体重が突然増加速度を速めたり、また思春期を過ぎて逆に増加速度が遅くなったりします。そのような時に、1 年前の身長や体重の増加から 1 年後の増加をどのくらい予測できるかなどに、関心を持ちました。また人の代謝に関連した基本的な物質としてクレアチニンがあります。小児において、尿中のクレアチニン排泄量を身長や体重で予測することも、私がかって関心を持ったテーマです。皆さんも、皆さんらしい課題を見つけ、回帰分析を使ってみてください。

演習問題

1. あなたはどのような課題に回帰分析を使ってみたいですか。50 字以内で書いてください。

2. エクセルを用い、my 標本で何か回帰分析を行い、結果を 50 字以内で書いてください。（my 標本ではなく、150 名データを用いても構いません。）エクセルが利用できない場合は、動画中のエクセルの説明を見て、思ったことを 50 字以内で書いてください。（登校禁止が解除されたら、ぜひ情報処理室でエクセルに触れてください。）

3. カシオ Linear regression Calculator で何か回帰分析を行ってください。すでにエクセルで計算済みであっても、こちらでも計算を試みてください。異なる手段で計算してみることで、手段の特徴や限界を理解できます。結果や気づいた点を 50 字以内で書いてください。

第 14 回　主観と統計

https://youtu.be/jGuwEr4WgwY

皆さん、こんにちは。今回は第 14 回目、実質的には最後の授業です。そこで今回は主観と統計についてお話しします。

１　なぜ主観か？？

この科目、統計学では、第 1 回目から「大切なのは客観」と言ってきました。主観は出来るだけ排除し、「何か影響がありそうだ、効果がありそうだ」と感じても、それを表には出さず「効果がない、影響がない」とする帰無仮説を立て、検証を続け、ｐ値（有意確率）が 0.05 とか 0.01 まで小さくなった時に、帰無仮説を棄却する、と話して来ました。

この考え方は 17 世紀以降、300 年の歳月をかけて成立してきた古典的な統計学の考え方の基本です。ただこの考え方だけで押し通すと、先に進むことが難しい場合も起きてきます。例えば「どの統計ソフトが学生の皆さんにとって最適か？」という課題を考えてみます。

コンピューターの発達とともに様々な計算手段が生まれ、ネット上にも多くの計算サイトがあり、そのいくつかは皆さんに紹介してきました。では授業を終わるにあたり、学生の皆さんに今後もお勧めできる統計のソフトは何でしょうか？

統計学の授業ですから、皆さんにお勧めする統計のソフトも、本来は統計学的エビデンスに基づくべきでしょう。全ての統計ソフトは「無効である」と帰無仮説を立てた上で、様々な統計ソフトを比較し、最後は帰無仮説を棄却して、意味のあるソフトを選ぶことも考えられます。しかし実際にはそのような検討は困難です。

ネット上のアクセス数やランキングで上位の人気ソフトにするのも一つの可能性ですが、そうする気がおきません。なぜかと考えた時に、学生の皆さんに私が自信を持ってお薦めするのだとすれば、何といっても私自身の経験と信頼がとても重要だ！と思い至りました。主観が大切なのです。

２　統計ソフトの主観的な判断基準

では私が皆さんにお勧めできる統計ソフトに求められる主観的条件とは何でしょうか。まず次の５点を考えました；１信頼できる、２無料、3 分かりやすい、４夢と発展性、５日本語。

さて残念ながら全ての条件を満たす統計ソフトは存在しません。しかし一つ条件を減らし、４条件としてよければ、皆さんにぜひ推薦したいソフトがあります。そこで 5 番目の日本語という条件は外したいと思います。日本人である私たちにとって、日本語で利用できることは大切です。しかし日本語にこだわると、本当に良い統計ソフトに出会うことが困難になります。これを機会に、学生の皆さんには、ぜひ英語の統計ソフトにチャレンジしていただきたく思います。

３　統計ソフト JASP

四条件を満たすソフトとして私が今日皆さんに推薦するのは JASP です。JASP はパソコン上で動く無料の統計解析ソフトウェアで、説明は英語です。

・JASP は信頼できる

まず信頼性について、JASP はオランダのアムステルダム大学・心理学部が中心になり、欧米のいくつかの大学が協力して開発している統計解析ソフトウェアです。オープンソースであり多くの利用者がコミュニティを通して改善に協力して、進化を続けており、とても信頼できます。

・JASP は無料

JASP は無料です。誰でもが自由に最新版をダウンロードして使うことができます。

・JASP は分かりやすい

JASP はグラフィカルユーザーインターフェースを採用しているため、ボタンを押すような感じで簡単に直感的に操作できます。内容は英語ですが、この授業でこれまで皆さんに紹介してきたどのソフトウェアよりも使いやすいと思います。図には JASP のグラフィカルインターフェースを示します。これまでの授業で皆さんにお話ししてきた全方法が画像ボタンに示されています。

最も左にある descriptive は第 3 回目の授業でお話しした平均値・偏差・分散・標準偏差などの説明的な基本統計量を示します。

その隣 T-tests は第 11 回目の授業でお話しした 2 群の比較や t 検定などに対応しています。

三番目の ANOVA は第 12 回目の授業でお話しした分散分析です。

そして 4 番目がリグレッション、第 4 回目の授業の回帰と相関、また第 13 回目の授業の回帰分析に対応しています。そして 5 番目 Frequencies これは第 6 回目の授業でのクロス集計表とカイ二乗検定に対応します。要するにこれまでの授業で皆さんにお話ししてきた主な統計計算が、すべてこの画像ボタン、グラフィカルインターフェースに集約されています。

なお６番目の Factor これは主成分分析・因子分析などの多変量解析に対応したメニューです。多変量解析はリグレッションの方法をさらに発展させたもので、特に心理学などの分野で広く使われる方法です。今回の授業の中ではお話ししませんが、興味深い方法で、JASP を使えば簡単に計算できますので是非機会があれば試してみてください。

・JASP の夢と発展性

さて、以上で私の主観で決めた条件のうち、三つについてお話ししましたが、もうひとっ、４番目、夢と発展性があります。JASP のおそらく隠れた最大の特徴としては、通常の古典的な統計の方法とは異なるベイズ統計が使えるという点です。ベイズ統計は授業でこれまで皆さんにお話ししてきた通常の統計学とは異なる発想の統計学です。

４　JASP の基本

ベイズ統計は今お話すると混乱する心配があります。以下 JASP の通常の使い方をお話しします。

１）JASP のインストール

・まず JASP のホームページを開きダウンロード JASP のボタンをクリックします。

https://jasp-stats.org/

・するとダウンロードのページが開きますので JASP Windows をクリックします。

・ダウンロードが終わるとセットアップのウィザードが開きますので、ダウンロード先のフォルダを指定します。

・インストールを押すとインストールが始まります。

・終了したらフィニッシュを押します。すると JASP の最初のページが開きます。

・この状態でデータさえあればすぐに計算が始められます。

２）JASP のデータ読み込み

皆さんの先輩 150 人分の JASP で読めるデータは既に統計学のフォルダに入っていますので利用してください。ファイル名は JASP-data-150people です。JASP の画像ボタン一番左端ファイルのタブを選ぶとファイルを指定できます。皆さんの先輩のデータファイル名を選び“開く”を押します。さてもう画面には 150 人のデータが表示されているはずです。

データの名前や順番や数字の意味は皆さんのマイ標本と同じです。ただし変数名だけは日本語ではなく英語にしています。変数名の日本語をどのように英語に変換しているかは、統計学のフォルダ中の JASP-日英-変数名対応表を見てください。

５　JASP による計算の実際

では早速 JASP でいくつか計算をしてみます。

１）クロス集計とカイ二乗検定

まず第 6 回目の授業のテーマクロス集計表とカイ二乗検定を JASP で行います。

画像ボタンメニューから Frequencies を選びさらにメニューの上から 3 番目 Contingency Tables を選びます。これがクロス集計表です。選ぶとすぐに変数指定画面が出てきます。クロス集計表の行と列にどの離散量を入れるかを設定します。行には Food（食の好き嫌い）、列には Cold（風邪の引きやすさ）を指定します。

行と列を指定するとその瞬間に 150 人のクロス集計が行われ、カイ二乗値 1.175 が計算されます。JASP は特に指定しないと最小限の計算結果しか出てきません。もう少し詳しく、たとえばイェーツの補正を行いたい場合は、変数指定メニューの下にある Statistics というタブを選び、カイ二乗値の他に χ^2 continuity correction （連続性の補正）を選ぶとイェーツの補正まで計算してくれます。

２）回帰分析

13 回目の授業で取り上げた回帰分析の場合は Regression の画像ボタンからさらに Linear Regression 線形回帰を選ぶと条件指定画面が現れます。まず Dependent Variable とは Y 目的変数あるいは従属変数のことです。Covariates とは X 説明変数、独立変数のことです。体重を身長で予測する場合は Dependent Variable に Weight(体重)、Covariates に Height(身長)を指定します。指定するとすぐに分析結果が表示されます。

３）相関分析

回帰分析とともによく使われるのは第 4 回目の授業で取り上げた相関分析です。二変数の相関を求めるだけであれば、回帰分析と似た結果が出ます。行ってみましょう。

Regression の画像ボタンを押し、Correlation Matrix を選ぶと、条件指定画面が現れます。変数の指定は先ほどの Linear Regression よりは簡単です。どちらが X でどちらが Y といった区別をする必要はありません。関連性を見たい変数をどんどん指定していけば、計算し、図を描いてくれます。まず身長 height と体重 weight を指定するとこの図が出てきます。さらに三番目の変数として Sleep 睡眠時間を加えると次の結果が現れます。指定した変数のどの組み合わせについても相関係数を計算し、散布図と回帰直線を描くことができます。多くの変数があり、どの関係が強いかどれは弱いかなどをまとめて検討するときに、とても便利な方法です。

以上で JASP の基本的な使い方を終わります。

演習問題

1. 動画では統計解析ソフトウェア JASP について説明しています。JASP はパソコンが無ければ使えませんので、今は無理だと感じる皆さんも多いでしょう。しかし JASP はこれから必ず皆さんの役に立つ方法です。JASP について思うことを、50 字以内で記してください。

2. この授業では時々英語のサイトを紹介してきました。統計の考え方は国際的です。簡単な英語で統計を学んでおくと、皆さんが将来国際的に活躍する際も役立ちます。以下はベトナムのナムディン大学での授業に関連して以前作成した動画です。今回の授業２回目と３回目くらいの内容です。https://youtu.be/MeUDkXtVNiE
   同じ内容でも言語が異なると、気づく点があるかもしれません。視聴して気付いた点、感想など 50 字以内で記してください。

3. 最初の動画の中でベイズ統計について一言述べました。ベイズ統計は皆さんの教科書には一言も書いてありません。今のところ、国家試験に出る可能性はゼロです。この授業で、最後に、もう少しベイズ統計についてお話ししようと考えて準備して来ましたが、時間がかかりすぎるため、この授業での説明は断念します。以下のサイトには、ベイズ統計の短い説明があります。
   https://www.otsuka-shokai.co.jp/words/bayesian-analysis.html
   あなたはベイズ統計に関心がありますか。ひとこと、20 字以内で書いてください。

4. パソコンが使える人は、余裕があれば、ぜひ JASP を使ってみてください。
   https://jasp-stats.org/
   JASP 用のデータ、変数の説明など参考情報は、統計学のフォルダに入れておきます。

終わりに

新型コロナウイルス COVID-19 禍で対面授業ができない状況下、このオンデマンド教材によるここまでの統計学の学習、本当にご苦労様でした。Forms を介して、皆さんからは様々なコメントや質問をいただきましたが、結局、どれにもお答えできませんでした。教材の動画作りに追われる日々の中で、動画を一方的に流すだけの授業になってしまったことを、この場を借りてお詫びします。ここまでのこの授業について、感じたこと、気づいたことなどありましたら、以下に自由にお書きください。（レポート提出は忘れないでくださいね）

参考文献

Alibali M.W. & Nathan M.J. (2012) Embodiment in mathematics teaching and learning: evidence from learners' and teachers' gestures. The Journal of the learning sciences, 21: 247-286. https://alibali.psych.wisc.edu/wp-content/uploads/sites/371/2018/02/AlibaliNathan2012.pdf

Goss-Sampson M. (2020) JASP; Jeffrey's Amazing Statistics Program. https://jasp-stats.org/

文部科学省（2017）小学校学習指導要領解説；算数編。https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387017_004.pdf

文部科学省（2017）中学校学習指導要領解説；数学編。https://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2019/03/18/1387018_004.pdf

守山正樹（2019）講義室での体験を出発点として公衆衛生学を学ぶ；指先から世界の有様に近づく試み. 感性と対話, 2(2): 49-64. https://narrativesenses.files.wordpress.com/2020/01/wpp16-moriyama.pdf

長与専斎(1902) 松香私志. 松本順自伝･長与専斎自伝/小川鼎三, 酒井シヅ校注(1980). 東洋文庫386, 東京:平凡社.

難波修史ほか（2016）JASPによる心理学者のためのベイズ統計. 広島大学心理学研究,16: 97-108. https://ir.lib.hiroshima-u.ac.jp/files/public/4/42606/20170323110508926377/HPR_16_97.pdf

日本学術会議（2014） 提言「ビッグデータ時代における統計科学教育・研究の推進について」 http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t197-1.pdf

Pearson, K. (1911) The grammar of science (3rd edition, revised and enlarged) pp.1-600. London; Adam and Charles Black. http://sarkoups.free.fr/pearson1911.pdf

Piovani, J.I. (2008) The historical construction of correlation as a conceptual and operative instrument for empirical research. Quality & Quantity, 42: 757–777.

Stanton, J.M. (2001) Galton, Pearson, and the Peas: a brief history of linear regression for statistics instructors. Journal of Statistics Education, 9:3 DOI: 10.1080/10691898.2001.11910537

Stigler S. (2002) The missing early history of contingency tables. Annales de la faculte des sciences de Toulouse. 11(4) 563-573.

杉亨二著（河合利安編）杉亨二自叙伝、杉八郎（発行）佐脇印刷所、1918年 https://dl.ndl.go.jp/info:ndljp/pid/98078

高木晴良（2017）系統看護学講座 基礎分野 統計学,第7版第2刷. 医学書院, 1-202.

後書き

2020 年 5 月初めにオンデマンド形式の遠隔授業で統計学を教え始めたばかりのときは、「これ以上“数値嫌い、統計苦手”を増やさない授業をしよう、基本のキを分かりやすく教えよう」と考えるのが精一杯でした。いちおう教え終わった今、やっと私自身が改めて統計学の面白さに気付き始めたためか、新たな疑問も出て来ました。遺伝と統計に関連する疑問です。

統計は、実は、遺伝と深くかかわっています。特に回帰や相関などの考え方は、ゴルトンやピアソンといった先人が、遺伝という現象を数理的に説明しようとして生まれました。こうした遺伝に関連する話は、それを意識し始めると、統計学の理解に必須だと感じられるのですが、一般的な教科書にはほとんど出てきません。興味深い歴史的な背景に触れずに、計算方法だけを教えるのであれば、学生たちが統計学に興味を失うのはあたりまえです。

江戸時代末期から明治初期において、杉亨二はバイエルンの文献からスタチスチックに目覚め、長与専斎は欧米の視察中にヒギエーネの存在に気づきました。その目覚めや気づきが元になって統計学や衛生学・公衆衛生学が生まれ、その実利性のゆえに、小学校から大学まで、半ば義務的に、私たちはこれらの科目を学び（学ばされ？）また教えて来た現状があります。最初の時点での目覚めや気付きを大切にし続けたら、現状とは異なる形の教育がなされていたかもしれません。

でも、まだ遅くはないと考えられます。新型コロナウイルス COVID-19 禍は、本来の統計学の教育の方向性に気付く機会を与えてくれています。

索引

統計学; COVID-19 禍のもとでのオンデマンド授業

2020 年 9 月 1 日　発行

編　集　守山正樹
発行者　NPO 法人ウェルビーイング
　　http://www.well-being.or.jp/
住　所　福岡市中央区大名 1-15-24　Well-Being BLDG. 2 F
電　話　092-771-5712
印刷所　プリントパック

ISBN 978-4-904997-03-1 C

ISBN 978-4-904997-03-1 C

2020年9月1日　第1刷発行
発行者　　NPO法人ウェルビーイング
　　http://www.well-being.or.jp/
〒810-0041 福岡市中央区大名1-15-24　Well-Being BLDG. 2F