回帰と相関

第４章　回帰と相関

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。

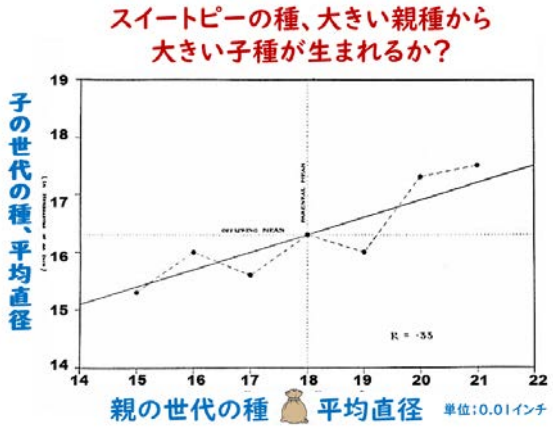
まず進化論を復習します。学生の皆さんは中学校か高校の生物学の時間に進化論を学んでいるはずです。「生物は不変のものではなく、長い年月の間に、確率的な変化が積み重なり、自然選択（自然淘汰）によって、現生の複雑で多様な生物が生じた」という考え方です。

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。

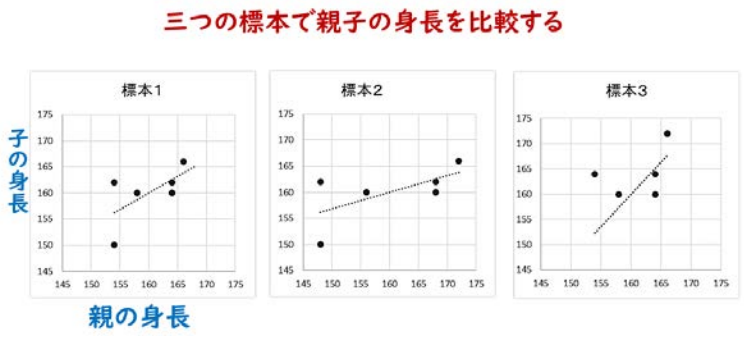
図より、最も小さい親種群の平均直径は 15.0、その親から生まれた子種群の平均直径は 15.2、最も大きい親種群の平均直径は 21.0、そこから生まれた子種群の平均直径は 17.3 などが読み取れます（数値の単位は 0.01 インチ）。

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スピアマ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

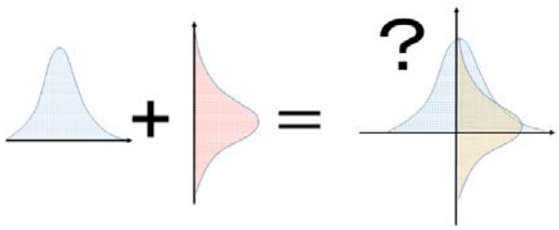
ピアソンはゴルトンの考え方を受け継ぎ、数学的に発展させ、「ピアソンの積率相関係数」の考え方が生まれました。その後、相関係数の考え方は急激に発展し、コンピューターの進歩に伴って現実の世界での統計的な観察を行うときに最もよく使われる方法になりました。

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。

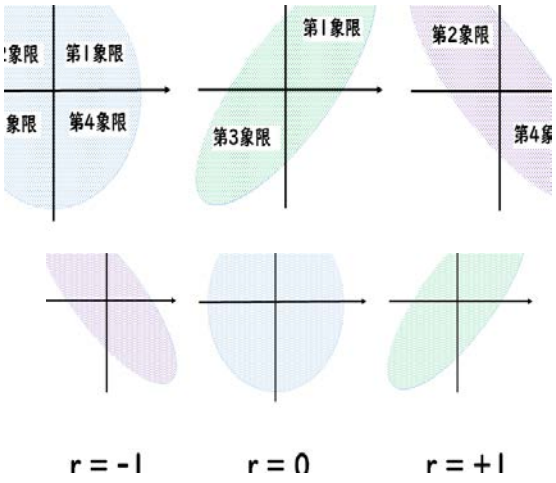
１）個々のデータ（変数）の分布

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＝$\sqrt{\text{Y 分散}}$

＝＿＿＿＿＿＿

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

統計学；COVID-19禍のもとでのオンデマンド授業