4-1 統計的検定とは

これから学ぶ統計学の「検定」は自分の主張が正しいことを、確率の考え方を使って説得する論法です。考え方は極めて常識的ですが、「能力検定」、「技能検定」……などの「検定」をイメージして統計学における「検定」を理解しようとすると混乱が起きますので、検定という概念を白紙に戻してスタートしてください。

●検定の考えはきわめて常識的

統計学における検定の考え方は極めて常識的です。つまり、「ある主張のもとでは、起こりにくいことが起きたときには、その主張を否定する」ということです。ここでは「主張」と言いましたが、統計学では「**仮説**（hypothesis）」という言葉を使います。すると「**ある仮説のもとでは起こりにくいことが起きたときには、その仮説を棄てる（棄却する）**」と言い換えられます。つまり、「ある仮説が正しいとして実験や観察をしたら、その仮説のもとでは、起こりにくいこと（確率が極めて小さいこと）が起きてしまったときには、その仮説を認めない」という考え方です。

逆に、ある仮説が正しいとして実験や観察をしたら、その仮説のもとでは、起きて不思議でないこと（確率が小さくないこと）が起きたときには、その仮説は否定しません。しかし、このとき、**積極的に仮説が正しいと認めるわけではなく、棄却できるほどの理由がなかったという考え方**をします。このような考え方を**統計的検定**（略して「**検定**」）といいます。もう一度、プロローグの §0−3 を参照してください。

●帰無仮説と対立仮説

検定では**検定者**（検定をする人）が「怪しいから棄てた方がよい」と思っている仮説を、無に帰したいので「**帰無仮説**（きむ）」といいます。これに対して、検定者が正しいと主張したい仮説を「**対立仮説**」といいます。たとえば、検定者が「小学生の家庭でのスマホの利用時間は増えている」と主張したいのであれば、対立仮説は「利用時間は増えている」であり、帰無仮説は「利用時間は変わらない」となります。仮説は英語で hypothesis なので、対立仮説には H_1 を、帰無仮説には H_0 という名前を付けます。「帰無仮説は無に帰したいから 0」と覚えておくといいでしょう。

帰無仮説、対立仮説という言葉を使って検定の流れをまとめると次のようになります。

ある仮説 H_0 について疑問をもち、それと反する仮説 H_1 が正しいと確信した人がいるとき、この人が自分の正しいと思う仮説 H_1 を第三者に認めてもらうために検定者になり、次の「検定」という手続きを踏みます。

①相手の主張を認め帰無仮説 H_0 が正しいとする。

②標本調査などを行ないデータを得る。

③データが、

（イ）帰無仮説 H_0 のもとで起こりにくいことであれば H_0 がおかしいとして仮説 H_0 を**棄却**し、対立仮説 H_1 を採択する。

（ロ）帰無仮説 H_0 のもとで起こりにくいことでなければ仮説 H_0 を棄却しない（棄てられない）。このとき、帰無仮説 H_0 を**受容する**という。

このようにして、**自分が正しいと思っている仮説（つまり対立仮説）を立証しようとするのが検定**なのです。能力検定などとは大きく違います。

4-2 仮説の採否を決める棄却域

検定では、ある仮説のもとで起こりにくい（ 稀（まれ）な）ことが起きればその仮説を棄てるわけですが、この起こりにくいと見なす範囲について調べてみましょう。ここでは、右のコインは表が出やすいと主張したい人の立場に立って説明しましょう。

●仮説を立て棄却域を設定する

まず、次の仮説を立てます。

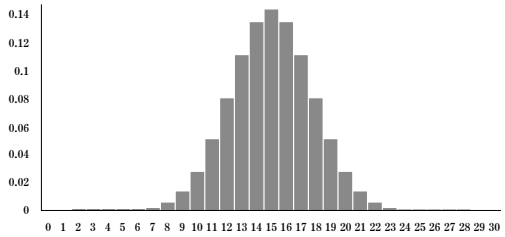
帰無仮説 H_0：このコインは表と裏が同じ確率で出る。

対立仮説 H_1：このコインは表が裏より出やすい。

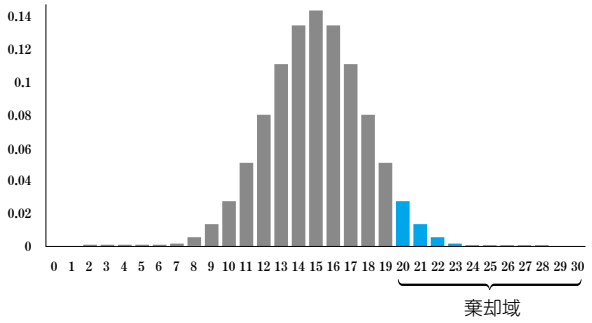
この帰無仮説が正しいとしたとき、コインを 30 回投げると表の出る回数 X の確率分布は次のような分布になります（節末〈Note〉参照）。

X	確率
0	0.000000
1	0.000000
2	0.000000
3	0.000004
4	0.000026
5	0.000133
6	0.000553
7	0.001896
8	0.005451
9	0.013325
10	0.027982
11	0.050876
12	0.080553
13	0.111535
14	0.135435
15	0.144464

X	確率
16	0.135435
17	0.111535
18	0.080553
19	0.050876
20	0.027982
21	0.013325
22	0.005451
23	0.001896
24	0.000553
25	0.000133
26	0.000026
27	0.000004
28	0.000000
29	0.000000
30	0.000000



表を見ると、20 回以上表の出る確率は 0.05（厳密には 0.049368…）ぐらいしかないことがわかります。そこで、このコインを 30 回投げて表が 20 回以上出たら帰無仮説のもとでは起こりにくいこと（稀なこと：確率 0.05）が起きたことになるので、帰無仮説は棄てることになります。この 20 回以上となる範囲を**棄却域**と呼んでいます。また、このときの確率 0.05 は「極めて稀」の基準を具体的に表現したもので**有意水準**と呼んでいます。**「その確率より小さいことが起これば、それは偶然ではなく、必然的な意味が有る」という意味**で**有意**と呼ぶのです。通常は 5% または 1% が採用されます。これは、常識的に「稀」と見なされる、小さくて切りのいい数値と思われるからです。



●反論対策

この検定で、もし、実験の結果「表が 20 回以上」出て帰無仮説を棄ててしまったら次のような反論があるかも知れません。

「帰無仮説が正しいときにも、**表が 20 回以上**となることが確率 0.05 で起きることがあるのだ。今はたまたまそれが起きたのだ」

と。正論です。そこで、「表が 20 回以上」出て帰無仮説を棄却するときには、但し書きを付けることにします。つまり、帰無仮説を棄却したけれど、そのことが「誤りである確率は 0.05 だけあります」と。この 0.05 は、先ほどは有意水準と言いましたが、誤りを犯す危険の度合いと考えることができるので**危険率**とも呼ばれます。そこで、「誤りである確率は 0.05 だけあります」を簡潔に「危険率 5%」と表現することにします。

8-1 二変量の関係を座標平面で視覚化 ～相関図（散布図）

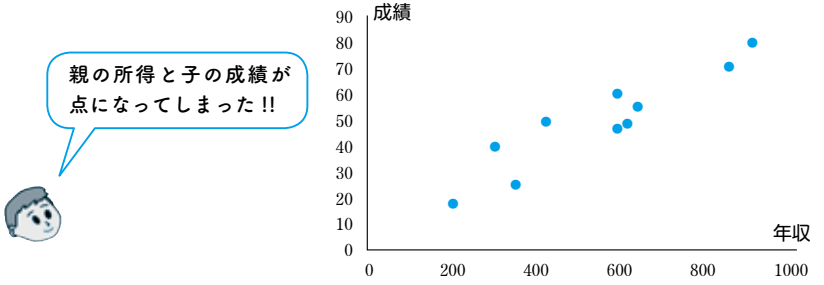
この章では「身長と体重」「子供の成績と親の収入」というように、変量が二つある場合に、これらの関係を解明する方法を調べてみましょう。

まずは、次の資料を例として二つの変量の関係を視覚化してみます。

世帯番号	1	2	3	4	5	6	7	8	9	10	平均
親の所得（万）	350	851	589	201	634	588	905	611	302	420	545.1
子の成績（点）	25	70	47	18	55	60	80	49	40	50	49.4

●座標平面を活用して点を描く

親の所得を横軸、子供の成績を縦軸にとった座標平面上に 10 人分の点（所得成績）をプロットした図を描きます。この図を二つの変量の**相関図**（**散布図**）といいます。

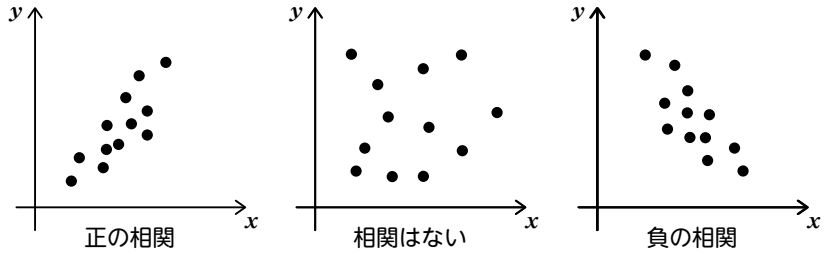


この所得と成績の相関図を見ていると、親の所得が高くなるにつれて子供の成績が良くなっているという関係がわかります。表からでは見えにくかったことが、データを座標平面上に点として図示することによって 2 変量の関係が一目瞭然となります。

（注）表から相関図を手作業で作成するのは大変です。しかし、Excel などの表計算ソフトや統計解析ソフトを使えば簡単です。

●正の相関、負の相関

二つの変量 x、y を相関図で表わすと、大まかに次の 3 つのパターンに分類できます。左側の図は変量 x が増加すれば変量 y も増加するという関係を表わし、このとき 2 変量の間には「**正の相関**」があるといいます。右側の図は、変量 x が増加すれば変量 y は減少しているので「**負の相関**」があるといいます。真ん中の図は「**相関はない**」といいます。つまり、2 変量 x、y の間には、とりたてて関係があるとはいえないということです。



先の所得と成績の例では、相関図より判断すると正の相関があるといえそうです。ただ、相関図で判断すると見方に個人差が現れるかも知れません。そこで、相関の度合いを客観的にわかるように数値化することも考えられています。それが §8−3 で紹介する相関係数です。

Note 偽相関とは

相関関係はあるが因果関係がない相関のことを**偽相関**（ぎ）といいます。たとえば、「町にあるポストの数と風邪を引いた人の数」などがそうです。人口の多い町は、一般に郵便ポストも多く、風邪を引く人も多いわけで、「ポストが風邪の原因」になっているわけではありません。

8-2 二変量の相関関係を正負で判断 ～共分散

二つの変量の関係を視覚化したのが**相関図（散布図）**ですが、二つの変量の関係を図ではなく、一つの数値で表わす工夫をしてみましょう。

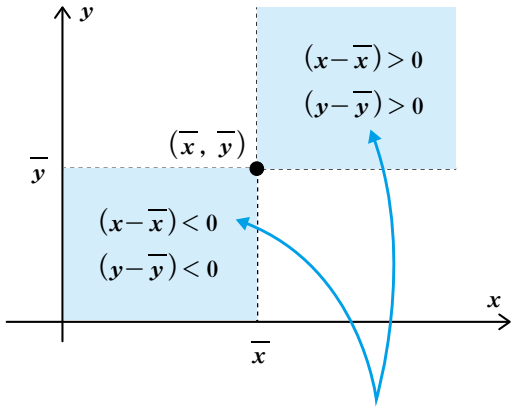
まずは、正の相関があるときは正の数を、負の相関があるときには負の数を対応させる**共分散**というものについて調べてみることにします。

●偏差の積に着目

二つの変量 x、y の関係を一つの数値で表現するために偏差に注目してみます。偏差というのは変量の値から平均値を引いたものです。

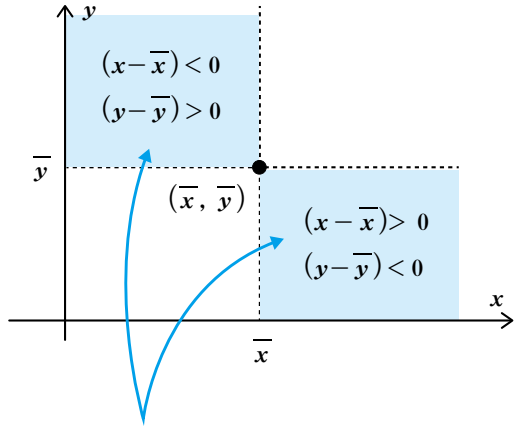
偏差＝変量の値－変量の平均値

したがって、変量の値が平均値より大きければ偏差は正の数、小さければ負の数になります。すると下の相関図において、網掛け部分のデータは二つの変量 x、y の偏差の積はいずれも正になります。



偏差の積 $(x-\overline{x})(y-\overline{y})$ が正

また、下の相関図において、網掛け部分のデータは二つの変量 x、y の偏差の積は負になります。したがって、偏差の積 $(x-\overline{x})(y-\overline{y})$ の総和が正の数であることは正の相関と、負の数であることは負の相関と、また、0 に近いときは相関なしと判定できそうです。



偏差の積 $(x-\overline{x})(y-\overline{y})$ が負

●共分散――個体数による変動を抑える

このようにして、2 変量の偏差の積 $(x-\overline{x})(y-\overline{y})$ の総和と相関を関係づけましたが、困ったことがあります。それは、偏差の積 $(x-\overline{x})(y-\overline{y})$ の総和は資料の個体数が大きいときはいくらでも大きな値、または、いくらでも小さな値をとり得るということです。

その結果、偏差の積の総和が 0 に近いと思われた 2 変量も、個体数を増やして調べれば総和は大きな数（または、小さな数）になってしまい、相関の有無がわからなくなってしまうことがあります。

そこで、個体数による偏差の積の総和の変動を押さえるため、偏差の積の総和を個体数で割ったものを考えることにします。つまり、偏差の積の平均値です。この値を**共分散**と呼ぶことにします。

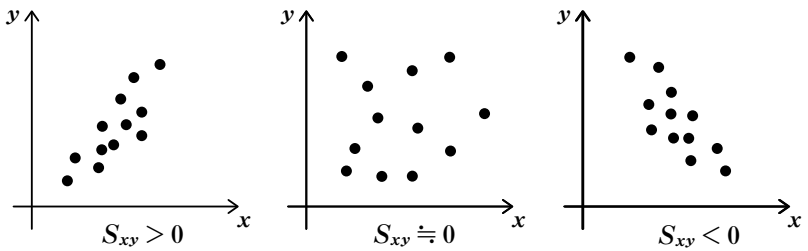
二つの変量 、$\{x_1, x_2, x_3, \cdots, x_n\}$、$\{y_1, y_2, y_3, \cdots, y_n\}$に対して**共分散**$S_{xy}$を一般的にまとめると次のように書き表わされます。

$$S_{xy}=\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$$

ここで、$\overline{x}$、$\overline{y}$はそれぞれの変量の平均値を意味します。このとき、共分散S_{xy}の符号と二つの変量x、yの相関については次のことがいえます。

個体番号	変量 x	変量 y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
…	…	…
n	x_n	y_n
平均値	$\overline{x}$	$\overline{y}$

$S_{xy}>0$のとき正の相関、$S_{xy}<0$のとき負の相関、$S_{xy}\fallingdotseq 0$のとき相関なしとなります。図示すれば次のようになります。



〔例〕下記の資料における親の所得と子どもの成績の共分散を求めてみると、次のようになります。

$$\frac{(350-545.1)(25-49.4)+(851-545.1)(70-49.4)+\cdots+(420-545.1)(50-49.4)}{10}$$
$$=3591.1$$

世帯番号	1	2	3	4	5	6	7	8	9	10	平均
親の所得（万）	350	851	589	201	634	588	905	611	302	420	545.1
子の成績（点）	25	70	47	18	55	60	80	49	40	50	49.4

8-3 相関の度合いを −1 以上 1 以下で表現 ～相関係数

共分散を利用すれば、それが負ならば負の相関、正ならば正の相関が二つの変量の間にあることがわかります（§8−2）。しかし、共分散は相関の強さまで表現することができません。つまり、共分散は相関の強さの客観的な指標にはなりえません。たとえば、同じ資料でも測定した単位によって、共分散は大きく変化してしまうのです。

●共分散に客観性をもたせた相関係数

試験の得点を見ただけでは、その点数がよい点数なのかそうでないのかを識別するのは困難でした。ところが、得点を偏差値に換算してみると客観的に善し悪しが識別できるようになりました。つまり、標準化することによって客観性を確保したわけです。

共分散についても一種の標準化を行なうことによって、相関の程度の客観性を確保できます。その換算式は次のようになります。

$$\frac{S_{xy}}{S_xS_y}$$

つまり、共分散を「各々の変量の標準偏差の積」で割ってあげるのです。この数値は**相関係数**と呼ばれています。これをr_{xy}と書けば、

$$r_{xy}=\frac{S_{xy}}{S_xS_y}=\frac{\text{変量 } x,\ y \text{ の共分散}}{\text{変量 } x \text{ の標準偏差}\times\text{変量 } y \text{ の標準偏差}}$$

となります。

（注）相関係数は厳密には**ピアソンの積率相関係数**と呼ばれています。