確率と統計(2)

佐賀大学　新井康平

データ標準化

- 標準化したデータの平均は0に、分散（標準偏差も）は1になります。
- これにより、異なる項目のデータであってもその大小を比較できるようになります。
- すなわち、大きければ大きいほど成績が良いことを表します。

競技名	平均	標準偏差
ボール投げ	4	2
走り幅跳び	6	1
50m走	5.5	1.2
高跳び	5	0.8
木登り	4.5	3

レーダーチャート

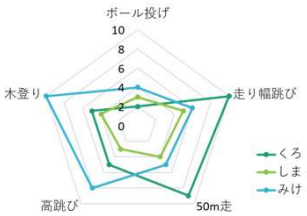
- レーダーチャートとは、いくつかの項目の大小を1つのグラフで表したグラフのことです。
- レーダーチャートを使うと、同時に別のデータとの比較も簡単にできます。

競技名	くろ	しま	みけ
ボール投げ	2	3	4
走り幅跳び	10	5	6
50m走	9	4	5
高跳び	5	3	8
木登り	5	4	10

- 標準化は次の式から行います。は元のデータを、は平均値を、は標準偏差を表します $\frac{x-\overline{x}}{s}$
- 例えば、「くろ」の「ボール投げ」のデータは次のように標準化できます。$\frac{2-4}{2}=-1.0$
- 同様にしてすべてのデータを標準化すると次のようになります。

競技名	くろ	しま	みけ
ボール投げ	-1.0	-0.5	0
走り幅跳び	4.0	-1.0	0
50m走	2.9	-1.3	-0.4
高跳び	0	-2.5	3.8
木登り	0.2	-0.2	1.8

- 他のデータもレーダーチャートに加えることで、それぞれのデータの特徴を比較することができます。
- このグラフを見ると、「くろ」は走るのが得意、「みけ」は飛んだり登ったりが得意、「しま」は体を動かすのがあまり得意ではなさそうということが読み取れます。



偏差値

- 「偏差値」を計算する場合があります。偏差値は、元のデータを平均が50、標準偏差が10となるように変換した値のことです。具体的には、標準化した値を10倍して50を足すことで求められます $\frac{x-\overline{x}}{s}\times 10+50$
- は元のデータを、は平均値を、は標準偏差を表します。例えば、「くろ」の「ボール投げ」のデータから偏差値を計算すると次のようになります。 $\frac{2-4}{2}\times 10+50=40$
- 同様にしてすべてのデータから偏差値を求めると次のようになります。

競技名	くろ	しま	みけ
ボール投げ	40	45	50
走り幅跳び	90	40	50
50m走	79	37	46
高跳び	50	25	88
木登り	52	48	68

偏差値の特徴

- 偏差値が高いほど成績が良いことを、低いほど成績が悪いことを表します。
- 平均点と同じ点数だった場合、偏差値は50になります。これは、平均点と同じ点数の場合には標準化した値が0となるためです。
- 異なる科目や競技で同じ点数だった場合でも、偏差値は等しくなるとは限りません。その科目や競技の平均値や標準偏差が異なる場合には、偏差値は異なったものになります。
- 偏差値は100以上の値やマイナスの値をとる場合があります。標準偏差が小さく、平均値からかけ離れた点数を取った場合には偏差値は非常に大きな値や小さな値をとります。

- 2つの要素xとyからなるn個のデータが得られたとき、その相関関係の強弱を表す相関係数は次の式から求められます。

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- nはサンプルの数を、とはi番目のデータの値を、とはxとyのデータの平均値を表します。また、はi=1からi=nまでの値をすべて足したものということを意味します。

相関係数

- 相関係数を求めるにあたって、まず2種類のデータのみをグラフに表してみます。

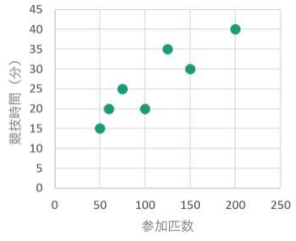
競技	参加匹数	競技時間（分）	怪我した猫（匹）
玉入れ	100	20	4
綱引き	150	30	8
リレー	50	15	1
毛糸玉ころがし	75	25	5
にぼし食い競走	60	30	5
ムカデ競争	125	35	9
騎馬戦	200	40	15

外れ値

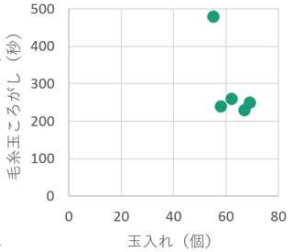
- 「外れ値」とは、他のデータと比べて大きく外れた値のことです。例えば、次の「玉入れで入れた玉の数」と「毛糸玉ころがしでかかった時間」のデータをプロットしてみます。

チーム	玉入れ（個）	毛糸玉ころがし（秒）
赤組	69	250
青組	50	240
黄組	55	480
緑組	67	230
白組	62	260

- 各競技の参加匹数と競技時間をプロットすると次のようになります。
- このようなグラフは「散布図」とよばれ、x軸（横軸）で1つ目のデータを、y軸（縦軸）で2つ目のデータを表します。
- ここで示したデータのように、横軸の値（参加匹数）が増加すると縦軸の値（競技時間）も増加するという関係がある場合には「正の相関関係がある」といいます。逆に、横軸の値が増加すると縦軸の値は減少するという関係がある場合には「負の相関関係がある」といいます。



- このデータから相関係数を求めると-0.69となり、強い負の相関があることが分かります。しかしこのグラフを見ると、1つだけ大きく外れた点があるようです。「黄組」の毛糸玉ころがしでかかった時間です。実は、毛糸玉を転がしている途中で毛糸がほどけて黄組の猫たちに絡まってしまい、なかなかゴールできないというハプニングがあったのです。

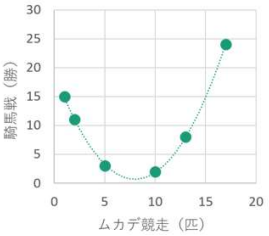


- 「ムカデ競争でころんだ猫の数」と「騎馬戦ではちまきを取った数（勝数）」をまとめたものです。このデータをプロットしてみます。

学年	ムカデ競走（匹）	騎馬戦（勝）
1年生	1	15
2年生	2	11
3年生	5	3
4年生	10	2
5年生	13	8
6年生	17	24

- 100匹の中からランダムに選ばれた1匹が「劇をやってみたい」と答える確率P(A)は次のように計算できます $\frac{10+10}{100}=\frac{20}{100}=\frac{1}{5}$
- もし、100匹の中からランダムに選ばれた1匹がオスであったときに、その1匹が「劇をやってみたい」と答える確率P(A)は次のように計算できます $\frac{10}{60}=\frac{1}{6}$
- このように、何か条件がある場合の確率のことを「条件付き確率」といいます。条件付き確率は次の式から求められます。 $P(B|A)=\frac{P(A\cap B)}{P(A)}$
- このとき、P(B|A)は「Aという条件のもとでBが起こる確率」を、P(A∩B)は「AかつBが起こる確率」を、P(A)は「Aが起こる確率」を表します。
- 100匹の中からランダムに選ばれた1匹がオスであったとき（Aという事象とします）に、選ばれた1匹が「劇をやってみたい」と答える（Bという事象とします）確率をP(B|A)とします。ランダムに選ばれた1匹がオスである確率P(A)は $P(A)=\frac{60}{100}=\frac{3}{5}$

- 相関係数は2つのデータによる直線的な相関関係の強さを表すものであり、線形ではない場合には相関関係の強弱は正しく表すことができません。したがって、いきなり相関係数を計算するのではなく、まずはデータをプロットした散布図を確認するようにしましょう。



- ランダムに選ばれた1匹が「オス」かつ「劇をやってみたい」と答える確率は、 $P(A)=\frac{10}{100}=\frac{1}{10}$
- であることから、求める確率は次のように計算できます。 $P(A)=\frac{\frac{1}{10}}{\frac{3}{5}}=\frac{1}{6}$
- 同様にして、100匹の中からランダムに選ばれた1匹が「ダンスをやってみたい」と答えるとき（Aという事象とします）、選ばれた1匹がメスである（Bという事象とします）確率P(B|A)を求めると、 $P(A)=\frac{\frac{15}{100}}{\frac{15+15}{100}}=\frac{\frac{15}{100}}{\frac{30}{100}}=\frac{1}{2}$

条件付確率

- 次の表は1年生100匹が来年の学習発表会でやってみたいと思ったことを集計したものです。

	オス	メス
劇	10	10
ダンス	15	15
和太鼓	25	2
ミュージカル	5	10
群読	5	3
合計	60	40