データのばらつきと統計的取り扱い

広島大学 AI・データイノベーション教育研究センター
稲垣知宏

目標

データにはばらつきがあることを理解し、ばらつきのあるデータを統計的に扱えるようになる

この授業で紹介すること

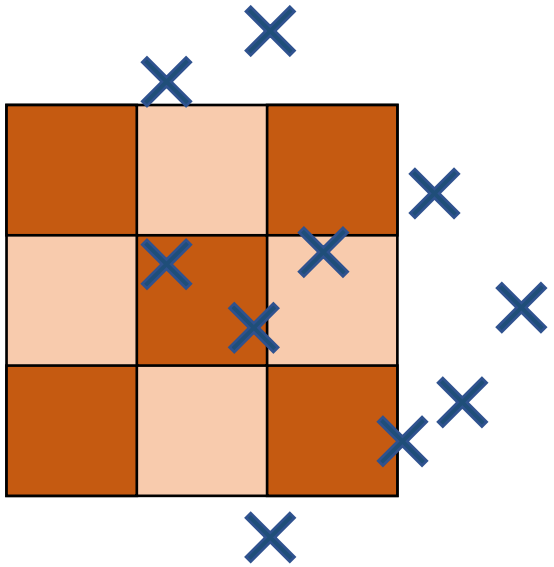
- ばらつきのあるデータを代表する値
- データのばらつきの評価

キーワード

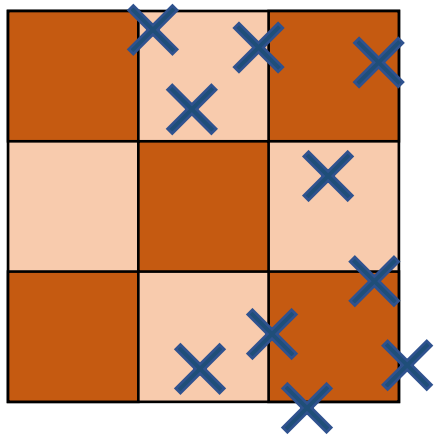
中央値、最頻値、平均値、確率分布、ヒストグラム

こんなことはありませんか？

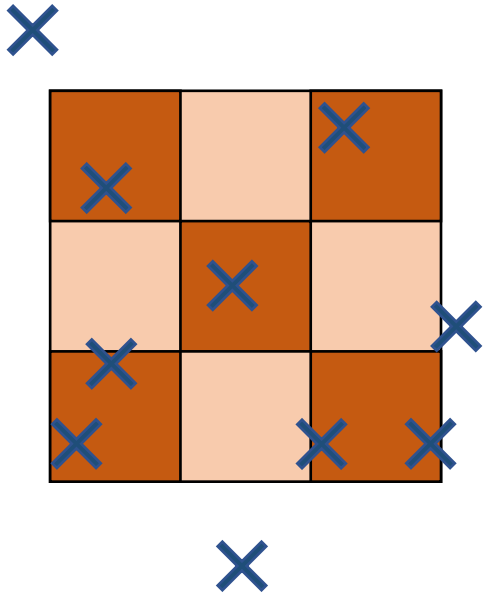
以下は、Aさん、Bさん、Cさんが、的の中央に向かってボールを10回投げた結果です。一番上手なのは誰だと思いますか？



Aさんの結果



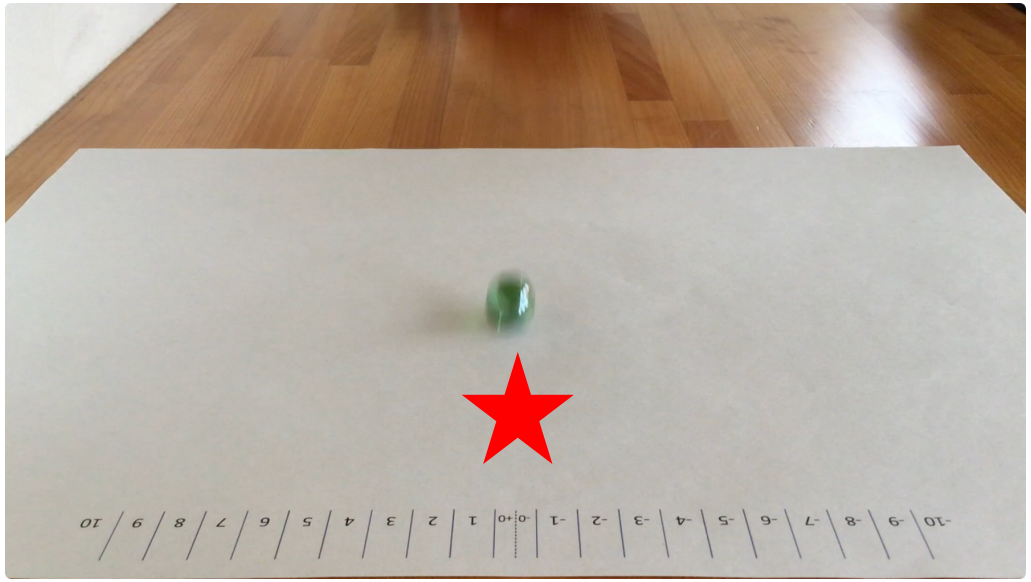
Bさんの結果



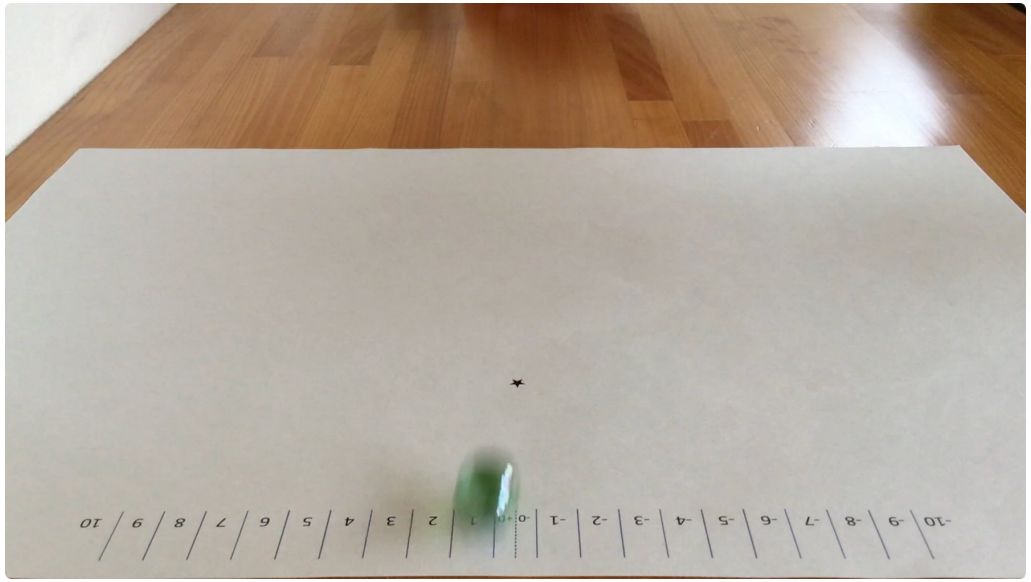
Cさんの結果

ビー玉を転がす

ばらつきのあるデータ例として、★印に向けてビー玉をゆっくり転がしてもらい、ビー玉が転がってきたビンの数字を記録します。



10 9 …　　… -10



10 9 …　　… -10

ビー玉を転がす

AさんとBさんに転がしてもらった結果です。これを分析します。

Aさんの結果

試行回	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ビンの値	-1	-3	2	-5	-1	1	-2	-4	2	0	4	1	5	-1	-4	-1	0	-2	4	1

Bさんの結果

試行回	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ビンの値	2	-5	-1	4	5	1	-4	2	-3	-1	-5	0	4	1	-3	-3	-1	1	3	-1

結果の分析

それぞれのビンに入った回数を数えます。

Aさんの結果

試行回	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ビンの値	-1	-3	2	-5	-1	1	-2	-4	2	0	4	1	5	-1	-4	-1	0	-2	4	1

回数

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0

結果の分析

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0

1. ビンの値の順にデータを並べたとき、真ん中に来るデータのビンの値を中央値（メディアン）と呼びます。20個のデータがある場合、下から10番目の値-1と11番目の値0の間になりますから、-0.5が中央値です。

2. 回数が最も多いビンの値が最頻値（モード）と呼ばれます。上の例では、回数が最も多いビンは-1で、これが最頻値です。

結果の分析

3. 次に、各ビンに入った回数をビー玉を転がした回数の20で割って、確率を計算します。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0
確率	0	1/20	1/10	1/20	1/10	1/5	1/10	3/20	1/10	0	1/10	1/20	0

ビンの値に確率をかけて足しあげると、平均値が求まります。

$$-5\times\frac{1}{20}-4\times\frac{1}{10}-3\times\frac{1}{20}-2\times\frac{1}{10}-1\times\frac{1}{5}+1\times\frac{3}{20}+2\times\frac{1}{10}+4\times\frac{1}{10}+5\times\frac{1}{20}$$
$$=-0.2$$

例題

Bさんの結果について中央値、最頻値、平均値を求めなさい。

Bさんの結果

試行回	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ビンの値	2	-5	-1	4	5	1	-4	2	-3	-1	-5	0	4	1	-3	-3	-1	1	3	-1

解説

試行回	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
ビンの値	2	-5	-1	4	5	1	-4	2	-3	-1	-5	0	4	1	-3	-3	-1	1	3	-1

まず、それぞれのビンに入った回数を数えます。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0

解説

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0

中央値：下から10番目の値-1と11番目の値0の間の-0.5です。

最頻値：回数が最も多い-1です。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0

平均値：ビンの値に確率をかけて足しあげると、-0.2になります。

解説

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0

中央値：下から10番目の値-1と11番目の値0の間の-0.5です。

最頻値：回数が最も多い-1です。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0

平均値：ビンの値に確率をかけて足しあげると、-0.2になります。

解説

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0

中央値：下から10番目の値-1と11番目の値0の間の-0.5です。

最頻値：回数が最も多い-1です。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0

平均値：ビンの値に確率をかけて足しあげると、-0.2になります。

解説

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0

中央値：下から10番目の値-1と11番目の値0の間の-0.5です。

最頻値：回数が最も多い-1です。

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0

平均値：ビンの値に確率をかけて足しあげると、-0.2になります。

上手だったのは？

AさんとBさんの結果は、データを代表する値として、中央値、最頻値、平均値で見ると、差がありませんでした。

AさんとBさんの結果のばらつきはどうなっているでしょうか？

Aさんの結果

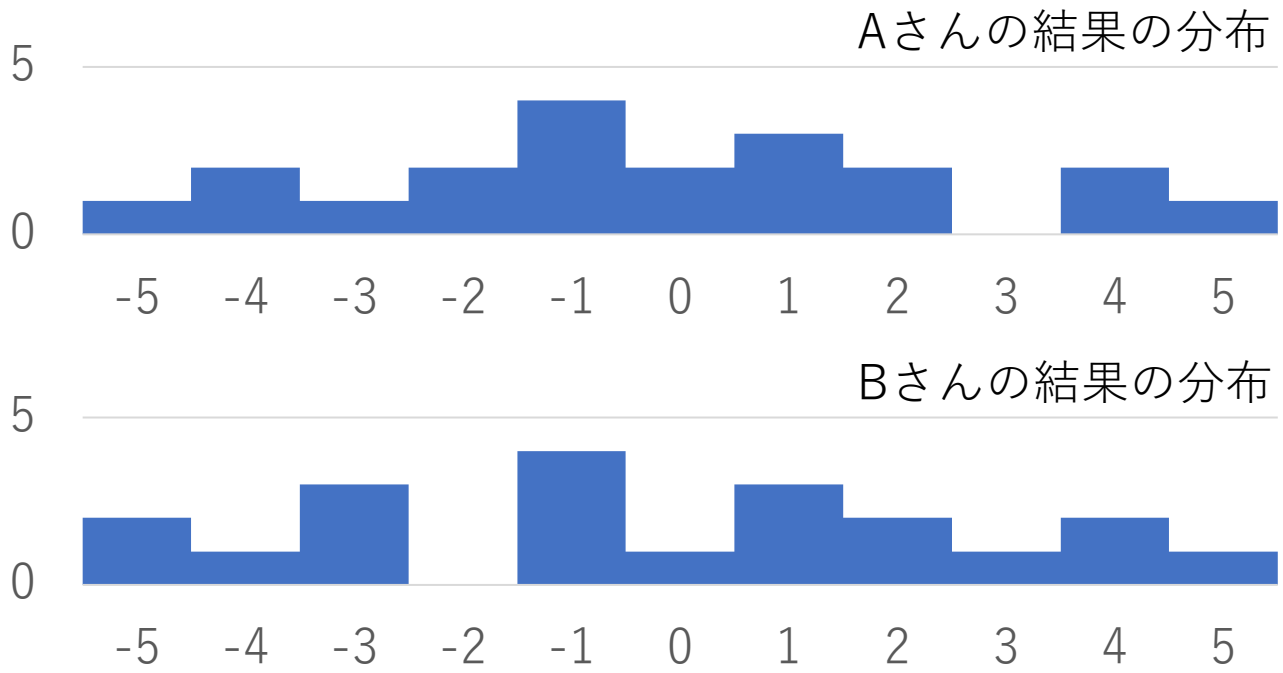
ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0
確率	0	1/20	1/10	1/20	1/10	1/5	1/10	3/20	1/10	0	1/10	1/20	0

Bさんの結果

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0

データのばらつき

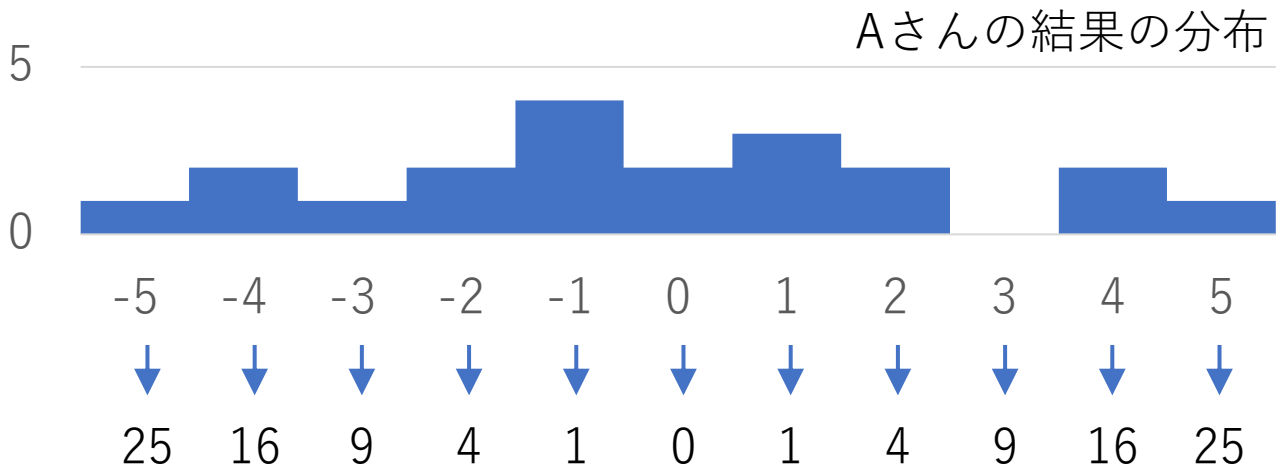
ビンに入った回数の分布（度数分布）をヒストグラムで表すと、AさんとBさんで結果のばらつきに違いがありそうです。



AさんとBさんの結果について、それぞれのビンに入った回数の分布のヒストグラム

ばらつきを評価する

ばらつきを評価するために、結果が中央にある0のビンからどれだけ離れていたか、ビンの値を２乗して考えます。



ビンの値の２乗に確率を掛けて足しあげます。

ばらつきの評価（１）

ビンの値の２乗に確率を掛けて足しあげると、Aさんの場合、

値の２乗	36	25	16	9	4	1	0	1	4	9	16	25	36
ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0
確率	0	1/20	1/10	1/20	1/10	1/5	1/10	3/20	1/10	0	1/10	1/20	0

$$25 \times \frac{1}{20} + 16 \times \frac{1}{10} + 9 \times \frac{1}{20} + 4 \times \frac{1}{10} + 1 \times \frac{1}{5} + 1 \times \frac{3}{20} + 4 \times \frac{1}{10} + 16 \times \frac{1}{10} + 25 \times \frac{1}{20}$$
$$= 7.3$$

ばらつきの評価（２）

0のビンからではなくて、平均値の-0.2からどれだけ離れているかの２乗で考えます。表から同様に計算すると7.26になります。

値の２乗	33.64	23.04	14.44	7.84	3.24	0.64	0.04	1.44	4.84	10.24	17.64	27.04	38.44
平均値との差	-5.8	-4.8	-3.8	-2.8	-1.8	-0.8	0.2	1.2	2.2	3.2	4.2	5.2	6.2
ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	1	2	1	2	4	2	3	2	0	2	1	0
確率	0	1/20	1/10	1/20	1/10	1/5	1/10	3/20	1/10	0	1/10	1/20	0

任意の位置からどれだけ離れているかでばらつきを評価できます。このばらつきが最も小さくなるのは、平均値です。

問題

1. Bさんの結果のばらつきを評価して、AさんとBさんのどちらが上手くビー玉を転がしたのか議論しなさい。
2. 平均値(-0.2)からどれだけ離れているかの2乗でばらつきを評価すると、0のビンからどれだけ離れているかの2乗で評価したばらつきの値より小さくなることを確認しなさい。

Bさんの結果

ビンの値	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6
回数	0	2	1	3	0	4	1	3	2	1	2	1	0
確率	0	1/10	1/20	3/20	0	1/5	1/20	3/20	1/10	1/20	1/10	1/20	0