百均の電卓で解ける統計入門

数学的な厳密さよりも直感的に理解することに重点を置いています・・・

散らばり？　－Ⅰ－

前回までに、データが数値的に広がって存在していて、その広がりの中から「全データを代表する数値」として**平均値**のお話をしてきました。

データたちは、平均値の周辺にどのように分布しているかを見るのが**ヒストグラム**でしたね。

しかし、平均値やヒストグラムでは、データがどのくらい広がっているのか、あるいは散らばっているのかはわからないのです。

で、そのデータたちが、どのくらい広がっているのか、ばらついているのかを考えていきましょう。

Aグループ：　4　5　6　7　8
Bグループ：　2　4　6　8　10

↓

Aグループ：　　4　5　6　7　8
Bグループ：　2　　4　　6　　8　　10

(1)平均値からの距離？

皆さん、第１回で「図の２つのグループを比べると、どちらがどのくらいバラツキが大きいと思いますか？並べ変えると・・・多くの方が『BはAより２倍ぐらいばらついているかなぁー』と感じるのではないでしょうか。」というお話をしたことを覚えていますか?

これを例題に考えていきましょう。

前号までに平均値の求め方はやりましたので、２つのグループはともに平均値が「６」であることはすぐに理解できると思います。

では、各データが平均値からどのくらい大きいか、あるいは小さいかを見てみましょう。

Aグループ	-2	-1	0	1	2
Bグループ	-4	-2	0	2	4

Aグループ：{(-2)+(-1)+(0)+(+1)+(+2)}÷5=0÷5=0
Bグループ：{(-4)+(-2)+(0)+(+2)+(+4)}÷5=0÷5=0

このように、平均値より大きいときはプラス、小さいときはマイナスで表した数値を統計学では『**偏差(deviation)**』と呼びます。私たちが知りたいのは、この５つの偏差を１つの数字で代表させたものですね。しかし、単純に算術平均(足して個数で割ったもの)を求めると・・・

実は、どんなデータに関しても、単純に偏差の算術平均をとると「０」になってしまうのです。ここでは、その証明は割愛しますが、算術平均を使う方法がうまくいかないことは、直感的にわかってもらえると思います。

では、どうしたらよいのでしょうか？

(2)符号をなくす？

こういった、不具合が起こる原因が符号であることに気付きましたか？平均値より大きいときはプラス、小さいときはマイナスで表した数値をそのまま使ったので打ち消し合いが起きてしまったのです。

そこで、マイナス側の偏差とプラス側の偏差を同等に扱うためには，絶対値にするかわりに二乗するのです。つまり，偏差の二乗和の平均をとるわけです。

$$\text{Aグループ二乗和の平均}=\frac{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}{5}=2$$

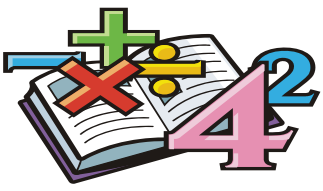
$$\text{Bグループ二乗和の平均}=\frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{5}=\frac{40}{5}$$

ここで出てきた数値量を『**分散(variance)**』と呼びます。

ただ・・・。これでは、ここの偏差より大きな数字となってしまい、「全データを代表する数値」とは言いにくいですね。

また、この例では単位がありませんが、mℓとか、分であったら、単位も二乗になってしまいます。ｃｍであったら、面積になってしまいますね。

そこで、第２回で述べた「**二乗平均**」を思い出してください。それぞれの差を二乗して足し、個数で割り、さらにルートする方法です。



$$\text{Aグループ二乗平均}=\sqrt{\frac{(-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2}{5}}=\sqrt{2}$$

$$\text{Bグループ二乗平均}=\sqrt{\frac{(-4)^2+(-2)^2+(0)^2+(2)^2+(4)^2}{5}}=\sqrt{\frac{40}{5}}$$

この統計量を「**標準偏差(Standard Deviation)**」と呼びます。この英語の頭文字を取って、「**S.D.**」と略されるのです。

この項　⇒