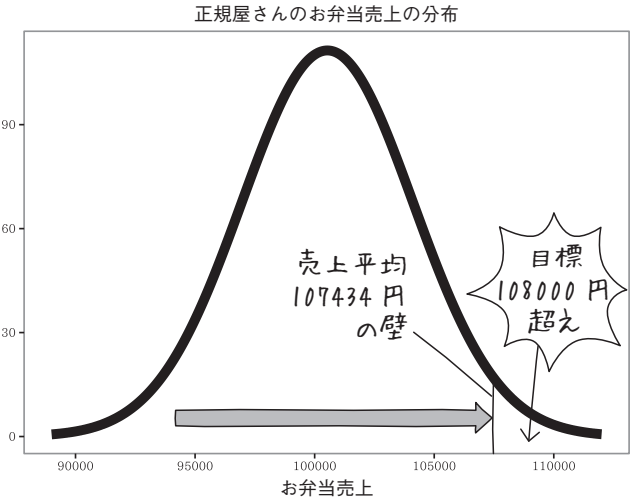
1 2 3 4 5 6

データ分析の王道？

平均値と標準偏差

翌週、正規屋をふたたび訪れた文太は、事務室で店長と乱子を前に、「今日はデータ分析の王道についてお話しします」と胸をはりました。

乱子は今日も、洗いたてらしい白い割烹着に白の三角巾という服装です。もっとも乱子は経理担当で、調理をすることはないらしいので、割烹着には1点のシミもなく、眩しいくらいに真っ白です。膝上まである割烹着の下からは、白く素足が伸びていますが、その先は例によって下駄でした。



の半分、つまり2・5％です。おおまかな金額で108000円を超える額です。

正規屋さんの売上が常時このレベルに達するようになるのは、いまの段階では確率的にはとても低くて、ハードルとしてはかなり高いのですけど、目標として頑張りたいと思うのです。

「よくは分からないけど、うちの売上高が大きく伸びるように努力してくれるというのを、正規分布にキンジさせて、おっしゃっているわけですね」

乱子が大きく微笑みました。白いズキンからこぼれているポニーテールが、窓から差し込む日差しに反射して眩しく見えました。

「ところで、さっきから動かしているこのソフトはなんなの？」

文太の正面のノーパソに乱子が顔をぐっと近づけてきました。文太の顔のすぐ前で乱子のポニーテールが揺れています。シャンプーの香りが文太の鼻をくすぐります。

「え、え、え。あ、これはRってソフトです。どんな分析もグラフ作成もこなしてくれるすごい高性能のソフトウェアなんです」

「ふーん。こういうの入れるから、うちのシステム契約料高くなるのね」

「いや、このソフトは、じつはフリーです。つまり、無料です」

「は？　無料？　だとすると、あんたの会社からの請求額、高すぎるんじゃないの？」

「す、すいません、社長に伝えておきます！」

「このレポートにあるグラフなんかも、このソフトで作ったわけ？」

「はい、そうです」

「ちょっと作ってみせて？」

「え？　あ、はい。たとえば、あの箱ヒゲ図は、こんな命令を実行しました」

```
> boxplot(data)
```

「……なに……これ？」

「Rというソフトでは、ユーザーが命令をプログラミング言語で書くんです。するとRがこれを解

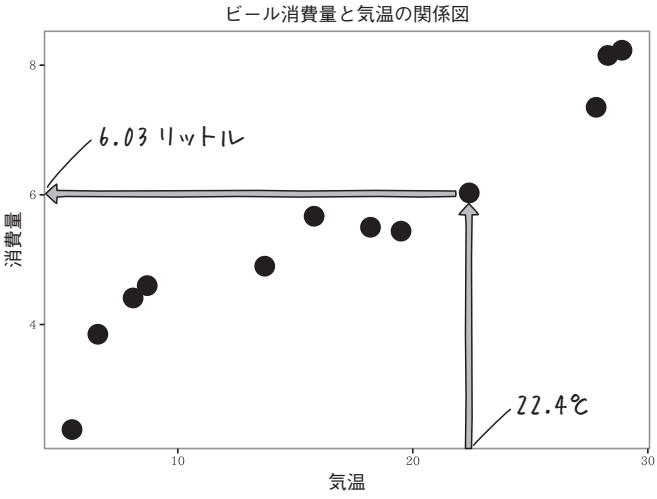
文太はキーボードを操作すると、数字の一覧のようなものを表示させて、乱子のほうにディスプレイを向けました。

月	気温	消費量
1	5.5	2.38
2	6.6	3.85
3	8.1	4.41
4	15.8	5.67
5	19.5	5.44
6	22.4	6.03
7	28.3	8.15
8	28.9	8.23
9	27.8	7.35
10	18.2	5.50
11	13.7	4.90
12	8.7	4.60

「データ分析では、データをこんなふうに行と列に記録します。左端の列は月です。その隣、つまり左から2列目の気温というのは、その月の平均気温のことです。その隣の消費量というのは各家庭でのビールの平均消費量です。たとえば一番上の行は、平均気温が5・5℃であった1月に、平均的な家庭で2・38リットルのビールが消費されたということです。

データ分析では、こういう列に記録された数値を『変数』とか『変量』といいます。この場合、『月』と『気温』と『消費量』という3つの変数があります」

「へんすう？　変わる数ってこと」

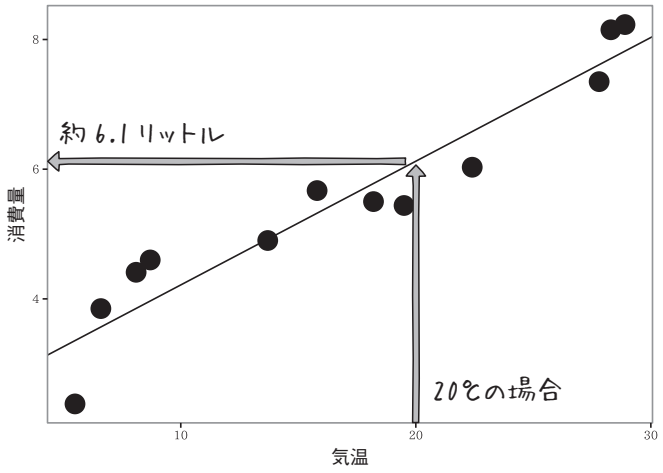


確かに、もうデータは集められているんですから、これが変わったりすることはないです。でも、来年、もう一度、気温と消費量の関係を調べても、今年と正確に一致するとは思えないですよね。年によって数字は変わってきます。それで、データ分析では「変数」といういい方をするのです。

このデータでは、「気温」と「ビール消費量」に注目してみます。この表からは、気温が高くなると消費量も伸びるという関係があることに気がつくと思います。この関係は、グラフにしてみると、もっとはっきりとします。描いてみますね。

このグラフを散布図といいます。グラフの描かれた正方形の真下にある目盛りは「気温」を表していて、これがX軸です。正方形の左の目盛りは「ビール消費量」を表しています。こっちはY軸でしたよね。

たとえばX軸の右のほうで22・4℃の位置から、



「ええ、相関係数は、2つのデータに直線以外の関係がある場合には使えないのです。

でも実用面ではバカにしたもんでもないんです。たとえばビールのデータでは平均気温が20℃の場合の記録はなかったです。でも気温と消費量にほぼ直線の関係があることが分かりましたから、グラフのX軸の20に対応するY軸の数値を見れば、消費量が予想できるようになります」

「グラフを目で追って判断するとか、いまどき原始的な気もするけど、まあ、確かに役には立ちそうね」

「……いや、これから、もっとデータサイエンティストらしい説明が始まるんです。それが最初にふれた回帰分析なんです！」

「怪奇隕石？」

「そう、つっこんでくるだろうと予想していましたよ……」

回帰分析

「それで、繰り返しになりますが、気温とビールの消費量関係は、次の式で表されます」

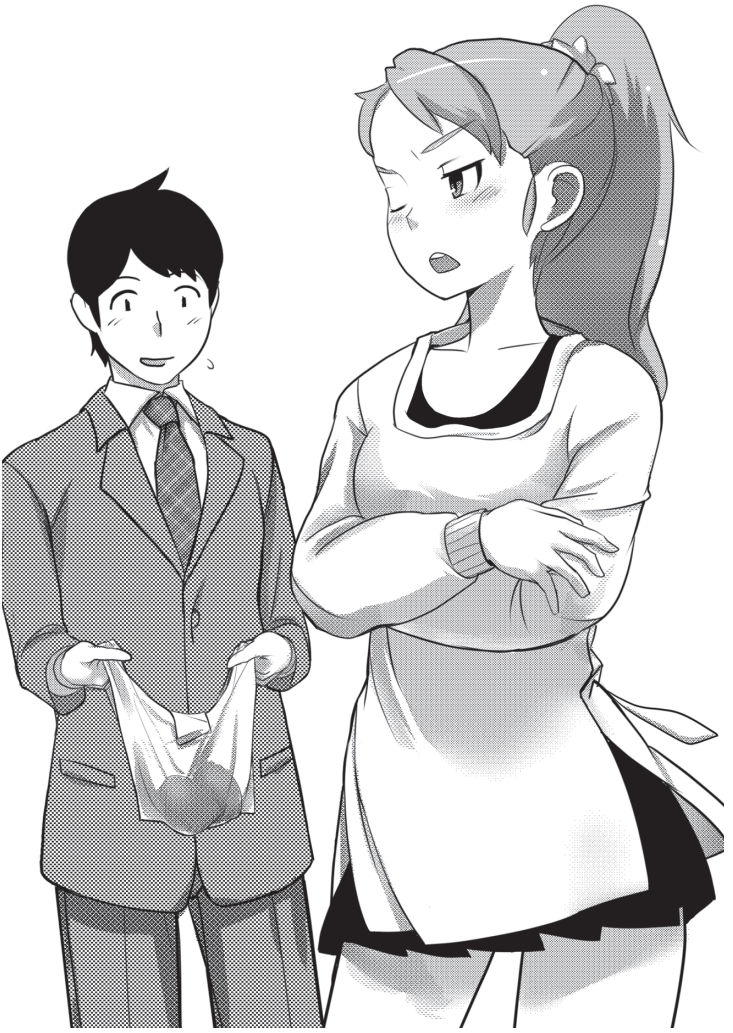
$$売上 = a \times 気温 + b$$

「やっぱり、ピンとこないんですけど」

「そうですよね。コンピューターにビールのデータを突っ込んで実際に計算させると、もっと具体的な式に変わります」

$$売上 = 0.19 \times 気温 + 2.3$$

「aが0・19に、bが2・3に変わったのは分かるけど、だいたい、この2つはなんなの？」

「まずですね、こうした式のことを回帰式といいます。それから、aは係数、bは切片といいます。係数aのほうは、この式の場合、気温が1℃上昇するとビールの平均消費量が0・19リットル増えることを意味します。つまり、温度による消費量の増加分を表します。それと切片のbですが、こちらは気温が0℃の場合の消費量になります」

「ちょっと、待った。0℃ってどういうこと？　そんな気温でも2・3リットルも消費するお調子者がいるわけ？」