はじめに

統計を使えるように なるために

◎統計との出会い

この本を手にとっていただき、ありがとうございます。

私は現在、株式会社ミツカリというHRTechの会社を経営しながら日々自社やお客様の人事データと格闘しております。性格データを取得する適性検査ツールを開発し、人事部を中心とした企業のお客さまがどのように分析を行うかをサポートする仕事をしています。そう言うと、「理系ですか？」とか「大学時代は分析は学ばれていたんですか？」と聞かれることも多いです。しかしながら大学の学部は法学部、しかも卒業までほぼエクセルなど触ったこともない、ゴリゴリの体育会出身でした。

一番最初に入った外資系の証券会社には、債券を扱う部門採用で入社しました。そこでは新人は金利の計算を行うのがスタートなのですが、エクセルでどうやって階乗の計算をするのかもわからず、1.03*1.03*1.03*.....と数式を書いて、同期に笑われるを通り越して唖然とされたことを覚えています。

それでもやらないとクビになるという切迫感もあり、エクセルの数式を深夜まで作り（質より量！）、なんとか曖昧な理解ながらも計算できるようになっていきました。それでも、統計を理解して使っているというよりは、なんとか食らいついているというような状態でした。

その後、一念発起し留学したビジネススクール（いわゆるMBA）で、一番最初のセメスター（学期のようなもの）で受けた統計の講義の１つのスライドが、私の人生に大きな影響を与えました。それはカジノのディーラーが八百長をしているかどうかを統計的に分析してみようという題材でした。いつも勝率のよいディーラーがここ30日間は負けることが多くなり、しかも彼の父親が病気だとい

うことがわかる、といったストーリー仕立ての展開。その講義を受けるまでは、なんとなくわかったようなわからないような状態だった統計でしたが、カジノのディーラーの八百長確率を分析する！？　という題材が妙に面白く、講義にのめり込みました。留学先からラスベガスも近く、実際のカジノにも行ったことがある手触り感も影響していたと思います。

その後、統計への興味から、いくつかのより高度な統計ソフトを使ったデータ分析の講義も受講しました。統計ってめちゃくちゃ難解で、しかも現実にどう使っていいのかわからないというものだったのが、カジノという身近な題材で学ぶことで、いろいろなところに使えて、なによりすごく面白いものだということがわかりました。そして面白いから忘れない、さらには少しできるようになると嬉しくなってもっと学びたくなり、という好循環が生まれました。

今では実務でも使う機会が増え、人事のお客様が統計の知識をつけたいとおっしゃっていることが多く、それならばと講義を作り、あれよあれよというまに延べ1500名以上の方に講義を行うまでになり、ついには本まで出版することになりました。本当に、人生なにが起こるかわからないなと感じています。

◎統計学習で大切なこと

エクセルを開いたこともなければ、正規分布表にオバQの落書きをするような統計劣等生だった私が考える、統計学習で大切なポイントは3つあります。

①身近な事例で考えること
②実際に計算を行うこと
③公式の意味を考えること

1つ目の「身近な事例で考えること」について。

私はカジノの事例で、統計の具体的な利用法に興味が沸きました。とっつきやすい題材で、実際に使えるところをイメージできるようになる。これはすごく大切です。本書では、新たに配属された山田君が人事のデータを使って分析する、という内容にしています。人事の方はもちろん、応募者として面接を受けられた方も多いと思います。その中でどんなことができるかを感じてほしいです。

2つ目の「実際に計算を行うこと」について。

統計は、概念だけ学んでも実際に手を動かさないと身につきません。計算を何回も繰り返すことで、忘れにくくなっていきます。世の中にはデータをまとめた結果なども多く出回っていますが、自分で計算をしてみる癖をつけるのは、誰かにとって都合のよいデータに騙されないようになるためにも有効です。とにかく手を動かしてデータの整理をしてみること、そして計算をしてみることが大事です。本書では、エクセルのサンプルファイルをダウンロードしていただき、たくさんの演習をやっていただくことを前提としています。エクセルの画面のスクリーンショットもつけています。

3つ目の「公式の意味を考えること」について。

統計の学習で出てくる公式は、たとえ覚えていなくても、エクセルの関数を使えば計算してくれます。ただ、公式の中身を覚えることで「結局どういうことをしているのか？」を知ることができます。そこさえ覚えておけば、たとえやり方を忘れてしまっても、「何をしているのか」から逆算して「どういう計算をするべきか」を考えることができるようになります。公式は一般化されたものが多く、記号などが多くてとっつきにいくいものです。それでも、なぜそんな計算をしているのか？　を知ることで、理解が深まり記憶に残ります。

この3つを意識して、本書は作られています。具体的なイメージを持って勉強し、自分で計算をして、なぜこんな計算をしているのかを考えることで、明日から使える≒記憶に定着しやすい内容を目指しています。

◎ ビジネスの現場で役に立つ統計的思考

それでは、統計的な思考を身に付けることで、どんなことができるようになるのでしょうか？　大きなメリットは、何かを決める時の判断材料になるということです。例えば統計的思考がわかるようになると、なんとなく多いとか少ないとか、なんとなく成功しそうだとか失敗しそうだといった「なんとなく」を、「何%」という数値に変えることができるようになります。

例えば本書のような人事の場合、応募者さんがなんとなく営業として成功しそうだとか失敗しそうだといった要素を数値化したり、なんとなく会社を辞めそうだといった感覚を数値化したりすることができます。また営業の場合であれば、どのような特徴のあるお客様が買ってくれやすいかを数値化したり、マネージャーの場合であれば、自分のチームが売上目標を達成できる可能性は何%程度なのかを数値化したりできます。 それ以外にも、経理の場合なら必要な経費が平均いくらで、この水準以上までは5%以下の確率でしか行かないだろう、などと数値化したりできるのです。

もちろん、これらは過去のデータを参考にしているため、過去の延長線上でしかわからないという欠点があるのは事実です。ですが、こうした批判を聞くと、分析結果のすべてが正しい結果につながるわけではないものの、存在している過去のデータを参考にしないのはすごくもったいないのではないかと思います。

また、昨今ニュースで耳にしないことはない人工知能（AI）の基礎にも統計が関わっており、統計的思考の基礎を理解することは、そういったトレンドの理解にもつながります。

◎ この本の流れと到達点

本書は、山田君と部長という2人の登場人物を中心に展開されます。そして、部長の優しいけれど自分で考えさせられる指導によって、山田君が統計の面白さにはまっていきます。また、データを整理する記述統計、手元にあるデータから未知のことを推測する推測統計、そしてその結果をモデルにしていく実運用を理解するという流れで進んでいきます。

最終的に本書を読み終わった段階で、手元のデータから未知のことを予測する回帰モデルを作り、そのモデルの予測精度がどれくらいか評価できるようになるまでを目指します。

なるべく統計用語を使わずに、大枠をつかめ、そして学ぶきっかけをつかめることを意図して書いております。それでは、一緒に学んでいきましょう！

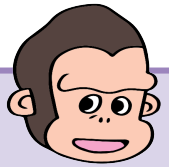
CONTENTS

第 3 章

相関

売上を上げているのは どんな要素を持った人？

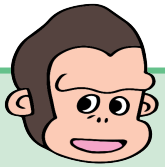
第 4 章



散布図

入社時面接の点数と売上の 関係を図にして理解しよう！

第 5 章



回帰分析

将来の売上予測にもっとも影響を与える要因を探せ！

サンプルファイルのダウンロードについて

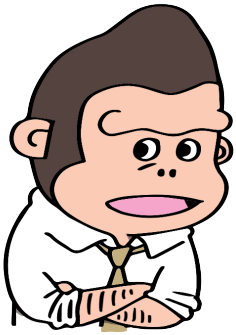
本書の解説で使用しているエクセルのサンプルファイルを、下記のWebページよりダウンロードできます。ブラウザーにURLを入力し、サンプルファイルのリンクをクリックしてください。ダウンロード後のファイルは圧縮されているため、解凍した上でご利用ください。

URL https://gihyo.jp/book/2021/978-4-297-12485-4/support

この本の登場人物

山田君

IT系企業に新設されたデータ分析チームに配属された、1年目の社員。不器用ではあるが、自分で考えてコツコツと挑戦することは得意。好きな言葉は「人間万事塞翁が馬」。



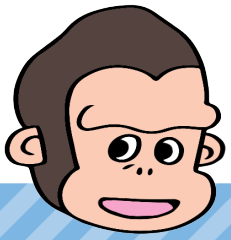
ゴリラ部長

山田君の上司。部下を育てることに情熱を持つ。昔はゴリゴリの体育会系だったが、伝わりやすい、そして自分でできるようになるコミュニケーションを心がけている。ラーメンやハンバーガーなど、B級グルメの大ファンで趣味は食べ歩き。

第 1 章

標準偏差

売上平均が同じでも
実態は違う
データの散らばりを
調べよう！



はじめに

新設されたデータ分析チームに配属された山田君。その辞令を見た時は驚きました。なぜなら大学時代は部活に4年間熱中。勉強はまじめに取り組まず、統計はもちろん、エクセルすらほとんど使ったことがなかったからです。それなのになぜデータ分析！？　新卒採用の最終面接を担当してくれたゴリラ部長の他は、自分を含め3名の小さなチーム。新設のチームだけど、データ分析とか統計って、ネットや新聞などでも見ない日はない言葉だから、期待されているチームなのではないかな。そんな期待と不安が入り混じる中、最初のMTGを迎えます。

contents

問題❶　社長が収益目標をぶらしたくないと言ったら、仕事を任せるべきは第一営業部、第二営業部どちら？

問題❷　社長が今年は第二営業部に仕事を任せるべきだ！　と言ったら、それはなぜ？

次につながる質問

部長はこう言いました「ところで、その意見って10人のデータを参考にしただけだけど、いったいどれくらい正しいのか？」

はじめまして

みなさん、今日からよろしく！　最近ネットニュースやテレビで「データ」や「統計」という言葉を見ない日はないよね。社内にある宝物のデータを分析して意味合いを見つけて、弊社の問題を解決するということが我々のチームのミッションになります。頑張りましょう。そして、今期もっとも取り組みたい問いがこちら。

そう言うと、ゴリラ部長はホワイトボードに次の文字を書いた。

「活躍する営業は採用時にどんな要素を持っているか？　を可視化する！」

「具体的なゴール：面接終了後にわかっているデータ（例：面接の点数、適性検査のスコア、起業経験の有無など）から、なるべく活躍する可能性が高い人を採用する意思決定に役立つモデルを作る」

弊社の競争力の源泉は、優秀な営業パーソンを多く採用し営業組織のレベルを上げること。もちろん、時代の変化によって求められるスキルは変わっていく。けれども役員陣と検討した結果、

活躍する営業の要素を定義して、採用の段階からそういった人を見抜く

ということが、分析チームが全社の売上に貢献できる期待値が一番高そうだという結論になりました。そのために、まずはみんなに分析に慣れてもらい、その後、ゴールを目指して実際の分析を進めていってもらいたい。それでは、新しく配属されたみなさん、自己紹介をお願いします。

（何人かの自己紹介のあと山田君の番が来る）

山田と申します！　大学時代はサッカーばかりの4年間を過ごしており、正直データ分析どころか統計すら挫折した経験があります。ただ、試合のデータから相手にどうやって勝つかを考えたり、コツコツとトレーニングを積むことで能力を改善してきた経験を生かして、一生懸命がんばりたいと思います。どうかよろしくお願いいたします！

（みんなから拍手）

よろしく！

データ分析で一番大切なこと

ところで山田君。早速だけど、エクセルの基本的な操作は学んできたかな？

はい。とはいっても内定者研修での事前課題+αでやってきた、合計や平均を計算するという程度ですが…。

それで十分。それでは、最初の仕事をお願いするね。ここに第一営業部と第二営業部、全体では50名ずついるチームだが、それぞれから抽出してきた10名の売上データがある。これらのデータがどうなっているか、気になる点を教えてほしい。

わかりました。

えーっと、わかったと言っているけど、具体的にどんなことをすればいいかイメージできてる？

えっ。

山田君。データ分析では、やるべきことを明確にしないで分析を開始すると、時間をすごく浪費することになる。

はい。

まだ始めたばかりなので難しいと思うけど、このあたりの心構えに関して少し話をします。まずデータ分析において私がもっとも大切だと思っていることを伝えたいんだけど、ちなみになんだと思う？

えっと。うーんと（こういう時の部長、まじゴリラっぽくて威圧感すごいな）。データ分析なので、数字のセンスとかですかね？

いや、分析にセンスなんていらない。大事なのは明確な目的。

も、く、て、き？

そう。データ分析をする時には、目的を明確にすることが重要なんだ。データ分析を開始してたくさんデータに触れていると、こんなこともわかる、あんなこともわかると面白くなり、無闇に時間が過ぎてしまうことがよくある。私も始めたての頃はこれによくハマって、翌朝まで分析して、面白かったけど結局なにも進んでいなかったということがよくあった。データ分析をする時は、何を目的にして、そのためにどんなデータが出ればいいのか、ポストイットに書いてPCに貼り付けておくとよいよ。先ほど山田君が簡単に「わかりました」と言ったのでちょっと注意が必要かなと思って聞いてみたけど、目的の整理や、分析でやるべきことが言語化できていなかったよね。

その通りです。

そんなにシュンとならなくて大丈夫。毎回データ分析を始める人には同じように接していて、いつも同じような反応になるから。正直な話、私はもっとひどかった。具体的に次のアクションがイメージできていない場合は、まず考えて、順序を言語化してみる。そして難しかったら聞くようにしてください。

ありがとうございます。それで今回の分析の目的は？

今回は、山田君にチームの作業に慣れてもらうことが第一。それから、第一営業部と第二営業部のどちらがいいチームなのかを数量的に明らかにできたら嬉しいな。あと、最初のうちは次の作業や方法がわからなくなったら30分程度自分で考え、検索などで調べても行き詰まったら早めに報告してください。

わかりました！

山田君は、メモにしっかりとこう書いた。

・データ分析をする時は目的をしっかり意識する（できればポストイットやメモに書いて、分析中はPCの画面隅に貼っておく）

・今回のデータ分析の目的

1） 今のチームの仕事に慣れること

2） 第一営業部と第二営業部のどちらがいいチームなのかを数量的に明らかにすること

（席に戻ると、部長からエクセルのデータが送られてきていた）

よし、それじゃあ最初に目的を考えよう。部長は仕事に慣れろと言っていたので、まずは「自分で分析→部長にチェックしてもらう→フィードバックをもらう→追加で分析する」という一通りのサイクルを回してみるのがよさそうかな。どうせ完璧にはできないだろうし、なるべく早くやってみて、部長の意見を早めにもらうことを心がけよう。あとは「いいチーム」の定義が難しいな…。ちょっとよくわからないから、まずはデータを見てみよう。

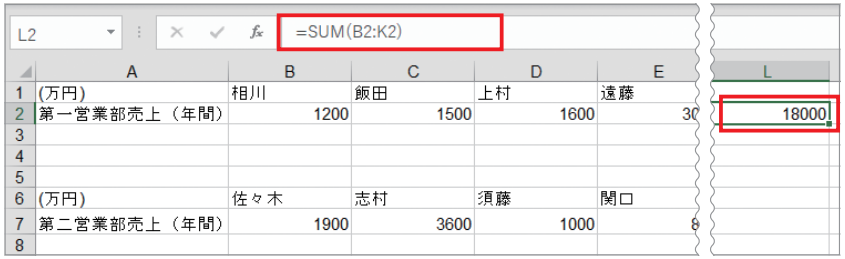
（そこにはそれぞれの営業の名前と売上だけが入っていた）

	A	B	C	D	K
1	(万円)	相川	飯田	上村	近藤
2	第一営業部売上（年間）	1200	1500		1800
3					
4					
5					
6	(万円)	佐々木	志村	須藤	戸口
7	第二営業部売上（年間）	1900	3600		3700
8					
9					

データとしてあるのは売上だけか…。ということはいったん

- **売上が高い営業さん＝活躍している**
- **売上が高いチーム＝いいチーム**

と定義して考えてみよう。まずはエクセルで、第一営業部の売上のデータを合計していこう。sum関数を使って…。



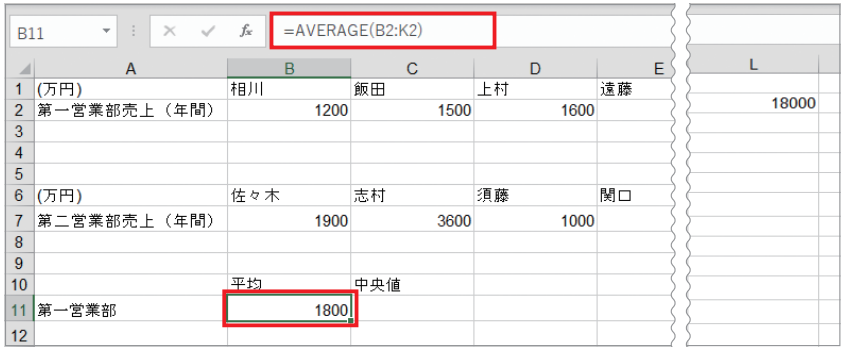
	A	B	C	D	E	L
1	(万円)	相川	飯田	上村	遠藤	
2	第一営業部売上（年間）	1200	1500	1600	30	18000
3						
4						
5						
6	(万円)	佐々木	志村	須藤	関口	
7	第二営業部売上（年間）	1900	3600	1000	8	
8						

エクセル関数

sum関数

複数のセルの合計を計算する際に、1つ1つのセルを足していってもよいのですが、sum（始まりのセル：終わりのセル）で選択すると一気にその範囲の合計値を出してくれます。またsum（）のかっこの中を「,」でつなげば、1つずつのセルを足していくことも可能です。

次は、売上の平均を出してみよう。このプロセスをまとめてできるのがaverage関数だよな。よし！　では平均の計算を関数を使ってやってみよう。

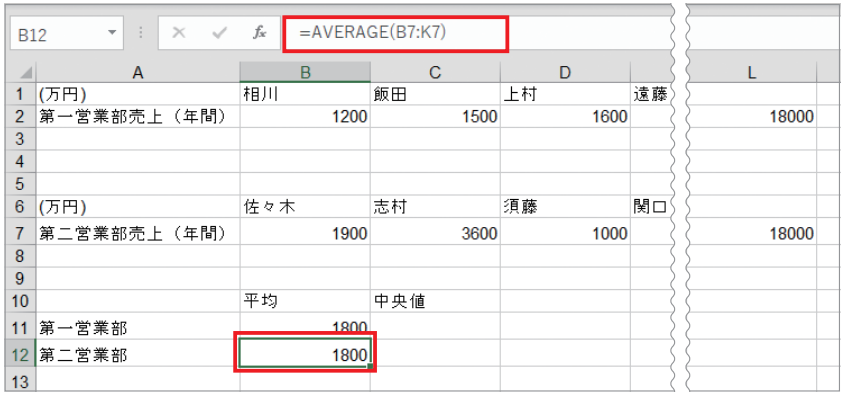


エクセル関数

average関数

average（始まりのセル：終わりのセル）で選択すると、一気にその範囲の平均値を出してくれます。またaverage（）のかっこの中を「,」でつなげば、1つずつ選択したセルの平均を計算することも可能です。

同様に、第二営業部の売上の合計と平均を計算してみよう。



なるほど。売上の合計や平均は第一営業部、第二営業部で同じになるんだな。

ってことは、どちらのチームも優劣がつけられない、ということになるのだろうか。けどなんかデータの中身を見ると志村さんとかが稼いでいるように見える特徴があるし。

（10分経過）

うーん。悩んでいても仕方がない。部長に報告しよう！

失礼します、部長。第一営業部と第二営業部の分析の件ですが、どちらも同じようにいいチームだと思いました。理由は、売上の平均が同じだからです。

それで？

それで？　えーと、うーんと。まず「いいチーム」の定義は難しいのですが、今回は売上のデータしかなかったので、いったん売上の高いチームを「いいチーム」と定義しました。

確かに、「いいチーム」の定義はそもそも何か？　を決めないと進まないね。売上が高いチームを「いいチーム」としたのは、よいと思います。で？

（部長の圧がすごいな）えーっと、とはいえ第一営業部と第二営業部では一番稼いでいる人が違ったりして。

ゴリラ部長の目がキラリと光る

ふむふむ。それってどういうこと？

そういえば、学校でデータを扱う時に平均はよく取ったのですが、なんかそれだけだと完璧じゃないというか。今回も第一営業部と第二営業部で平均は同じでどちらもいいチームなのかもしれないですが、個人として成績のいい人が多いのはどちらか？　とかがわからないというか。

いい視点だね。分析とか統計と言うと、まずは平均が出てくる。そして平均はとても大事な統計指標です。ただ、それだけじゃ山田君が今感じたように足りなくて、例えば5人全員が50点のテストと、5人が0点、30点、60点、70点、90点のテストとでは、意味が違うよね。今度はエクセルで中央値、偏差、分散、標準偏差を調べてみて報告してもらえる？

は、えっと…確認のためもう1回言っていただいていいですか？

（こいつ絶対わからなかったな）もちろんいいよ。第一、第二営業部の中央値、偏差、分散、標準偏差を調べてほしい。

統計とは？

ここからは、ゴリラ部長から少しお話させていただきます。

統計の定義を調べると、「集団の性質や傾向を数量的に明らかにすること」が統計であるという記述を見つけることができます。例えば手持ちの買い物のデータ（集団）があったとします。その平均が100円だったことがわかると、データの傾向が数量的に明らかになってきますよね。このように、単なる数値の集まりである集団（＝データ）の持っている傾向や性質を数値によって理解できる形にしていくことが、統計の定義になります。

平均とは？

次に、平均とはなんでしょうか？　私の対面での講義で同じ質問をすると、答えられる人は10人に3人くらいの割合です。それでも

1, 2, 3, 4, 5

の平均は何か？　と聞くと、ほぼ100％みなさんお答えいただけます。

平均は3

ですよね。このように統計は定義も重要ですが、具体的な数値で計算できることが大事です。ちなみに平均の簡単な定義は

データの合計をデータの数で割ったもの

と言えると思います。上記の例ではデータの合計は1＋2＋3＋4＋5で15、それをデータの数5で割った3が平均になります。

中央値とは？

次に、中央値の定義です。対面の講義では、中央値の定義をお聞きして答えられるケースは平均よりも減り、20%程度の方がお答えいただけます。ただ

1, 2, 3, 4, 5

の中央値は？　と聞くと、みなさんお答えいただけるんですよね。なんとなく3だと思った方！　正解です。ちなみに中央値の定義は

データを大きい順ないし小さい順に並べた時に真ん中にくるもの

になります。この場合、データは小さい1から大きい5の順に並んでいるため真ん中の3が中央値になります。
それでは、

1, 2, 3, 4, 5, 6

の中央値はなんでしょうか？　あれ？　真ん中の数がなくなりましたね。データの数が奇数の場合は真ん中の数が存在しますが、データの数が偶数の場合は真ん中の数は存在しません。偶数の場合は、データを大きい順か小さい順に並べた時の真ん中の2つの値の平均をとるのが、中央値を出す際のルールです。この例の場合は、3と4の平均の3.5が中央値となります。

平均と中央値は何が違うの？

もしかしたら「あれ？」と思った方はいらっしゃらないでしょうか？
実は1, 2, 3, 4, 5のデータの平均と中央値は、どちらも3。1, 2, 3, 4, 5, 6のデータの平均と中央値はどちらも3.5なのです。平均と中央値が同じ結果になるなら、なぜ統計指標として平均と中央値が存在するのでしょうか？　不思議ですよね。

実は、1, 2, 3, 4, 5, 6のように差が均等（等差）など、特定の条件を満たす時にデータの平均と中央値が同じになるだけで、多くの場合は異なる値になります。
平均と中央値の差がとりわけ大きくなるのが、飛び抜けた値がある時です。ちょっと大げさな例ですが、

1, 2, 3, 4, 100000

というデータがあったとしましょう。この時平均は

100010/5＝20002

中央値は

3

になります。
このように飛び抜けた値がある時は、平均と中央値の乖離が大きくなります（こうした飛び抜けた値を外れ値と言ったりします）。
上記の特性を理解した上で、分析の目的に合わせて平均と中央値を使い分けてくださいね。

偏差とは？

次に、偏差です。偏差は「偏った差」と書きますが、「差なので引き算をする」と考えると覚えやすいです。ではどうやって引き算をするかと言うと、それぞれのデータの値からデータ全体の平均の値を引いたものが偏差になります。つまり、平均からどれくらい偏っているかを表したものが偏差ということです。

例えば

1, 2, 3, 4, 5

のデータの例をとると、平均は3なので偏差はそれぞれ

1－3＝－2
2－3＝－1
3－3＝0
4－3＝1
5－3＝2

を計算して

－2, －1, 0, 1, 2

となります。
平均も中央値も（このあと出てくる分散も標準偏差も）、データの集まり（データセット）に対して1つしか出てきません。データセットに1000億個のデータがあっても平均は1つだけ。偏差は、データの数だけ出てくるという特徴があります。

● 分散とは？

分散とは、散らばりを表す統計指標です。「散らばっている」と言うと、その基準は？　と思う方もいらっしゃると思います。分散の基準は、平均からの散らばりです。分散を出すには、

偏差をそれぞれ2乗して、その平均をとる

という計算をします。

先ほどの1, 2, 3, 4, 5の例で言うと

偏差は　　－2, －1, 0, 1, 2
その2乗は　4, 1, 0, 1, 4
その平均は（4＋1＋0＋1＋4）/5

ということで

分散は2

となります。

● 標準偏差とは？

標準偏差も、データの散らばりを表す統計指標です。計算の仕方は、分散の平方根をとります。平方根とは何かと言うと、2乗するとXになる数を2乗する前の数に戻すことを言います。
1, 2, 3, 4, 5の例ですと、分散は2でしたから、2乗すると2になる数は

$$\sqrt{2} = 1.41431$$

となり、これが標準偏差となります。

売上の平均、中央値、分散、標準偏差を計算しよう

よーし！　まずは中央値を計算していくぞ。大きい数から数えて中央値を見つけることもできるけど、ちょっと面倒だな。エクセルに関数はあるのかな？

そうか

median

という関数で中央値を計算できるのか。

エクセル関数

median関数

median（始まりのセル：終わりのセル）で選択することで、中央値を出してくれます。median（）のかっこの中を「,」でつなげば、1つずつ選択したセルの中央値を計算することも可能です。

C11 =MEDIAN(B2:K2)

	A	B	C	D	K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3						
4						
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800				
13						

答えは1750。

一応確認もしてみよう。一番売上が低いのが相川さんと毛塚さんの1200。その次が飯田さんの1500。次が上村さんの1600。5番目が北村さんの1700ときて6番目が近藤さんの1800。データセットが今回は10個で偶数なので、真ん中2つの平均をとって（1700＋1800）/2で1750。合ってるな。同じようにmedianで計算すると、第二営業部の中央値は1700か。平均と中央値は近いけど、一致はしていないんだな。

C12 =MEDIAN(B7:K7)

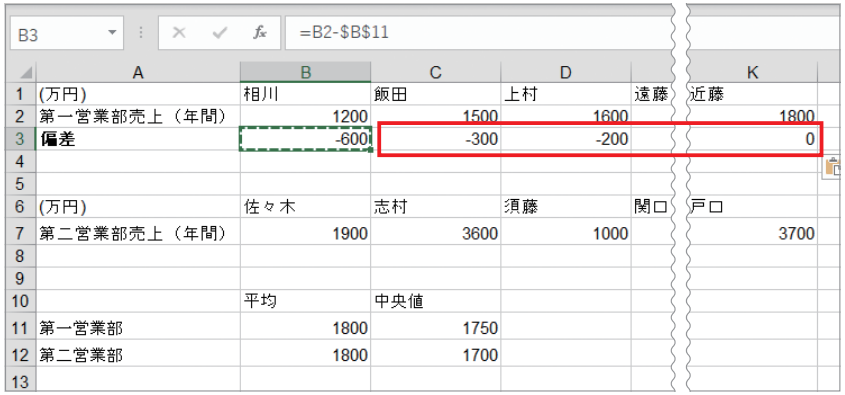
	A	B	C	D	K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3						
4						
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800	1700			
13						

次に、偏差の計算をしてみよう。偏差は「それぞれのデータから平均を引いたもの」だから、相川さんの偏差は1200−1800で、−600かな。平均はコピー＆ペーストした際に参照セルが動いてしまうので、絶対参照（$をつける）にすると便利だな。やり方は簡単で、WindowsでもMacでもF4キーを押すとアルファベットと数字の前に$マークがついて、その値が固定される。

B3 =B2-B11

	A	B	C	D	K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3	偏差	-600				
4						
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800	1700			
13						

よーし。これで偏差が出たぞ。これをこのままコピー＆ペーストして、最後の近藤さんまで計算しよう。



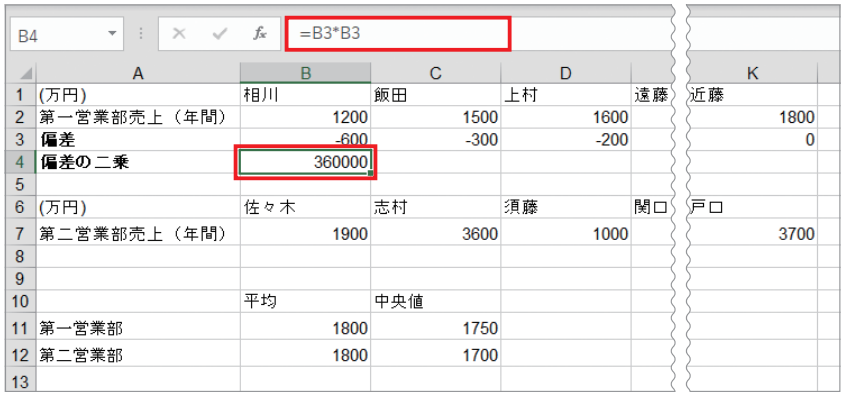
	A	B	C	D		K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3	**偏差**	-600	-300	-200		0
4						
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800	1700			
13						

分散は「偏差の2乗の平均をとったもの」だったな。
えーっとエクセルで2乗するのは、確か2つやり方があって

① 単純に同じセルを2回掛ける
② ^を使って2乗する

があるんだよな。
①のように同じセルの値を掛けるのでもよいし、

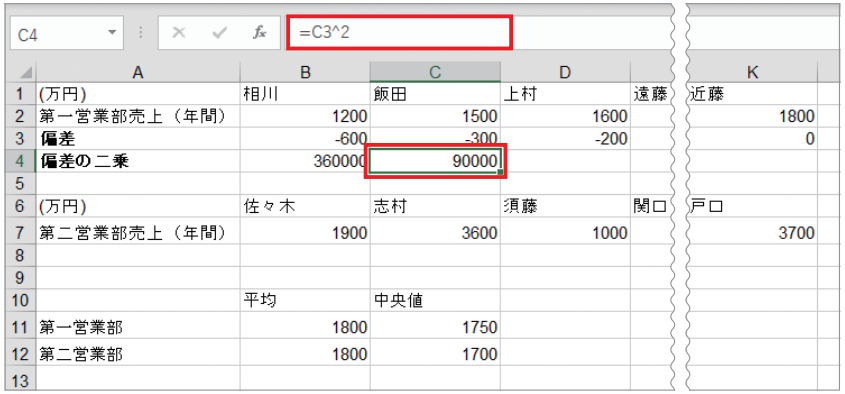
①のやり方



	A	B	C	D		K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3	**偏差**	-600	-300	-200		0
4	**偏差の二乗**	360000				
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800	1700			
13						

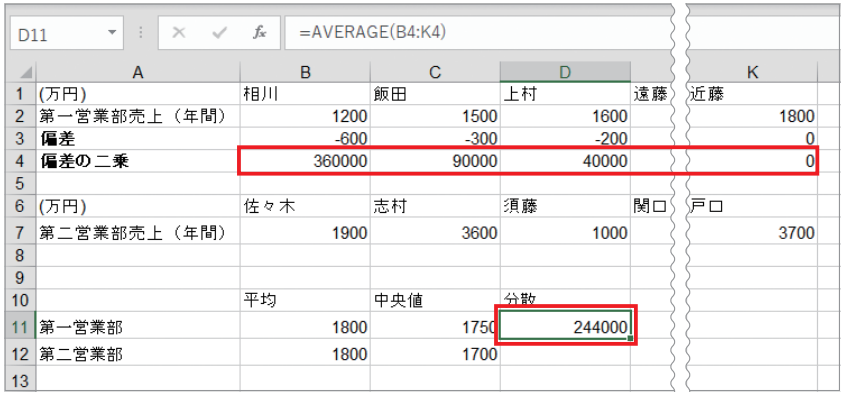
②のように^を使って2乗してもよい。

②のやり方



	A	B	C	D		K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3	**偏差**	-600	-300	-200		0
4	**偏差の二乗**	360000	90000			
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値			
11	第一営業部	1800	1750			
12	第二営業部	1800	1700			
13						

全員の「偏差の2乗」を平均すると、分散がでる。



	A	B	C	D		K
1	(万円)	相川	飯田	上村	遠藤	近藤
2	第一営業部売上（年間）	1200	1500	1600		1800
3	**偏差**	-600	-300	-200		0
4	**偏差の二乗**	360000	90000	40000		0
5						
6	(万円)	佐々木	志村	須藤	関口	戸口
7	第二営業部売上（年間）	1900	3600	1000		3700
8						
9						
10		平均	中央値	分散		
11	第一営業部	1800	1750	244000		
12	第二営業部	1800	1700			
13						

最後に、標準偏差は「分散の平方根」をとるんだった。平方根をとるっていうことは、2乗したらその数になる数に戻すということ。例えば4の平方根は2乗したら4になる数ということで、±2がその数になる。標準偏差はばらつきの大きさを表すので、正の値になるはずだ。

エクセルでの計算のやり方はこれも2つあって

① sqrt関数を使う
② ^(1/2)　2分の1乗する

のどちらかで計算できる。

（部長注：ちなみに細かい話ですが、②のやり方を選んだ場合、エクセルで「^1/2」と書くとエクセルは1乗してから2で割る、つまり2で割ったのと同じ結果になってしまいます。1/2に括弧（）をつけるのを忘れないようにしてください）

平方根をとるというのは、別の言い方をすると1/2乗することに等しいんだな。

①のやり方

E11　=SQRT(D11)

	A	B	C	D	E	
1	(万円)	相川	飯田	上村	遠藤	大森
2	第一営業部売上（年間）	1200	1500	1600	3000	
3	**偏差**	-600	-300	-200	1200	
4	**偏差の二乗**	360000	90000	40000	1440000	
5						
6	(万円)	佐々木	志村	須藤	関口	相田
7	第二営業部売上（年間）	1900	3600	1000	800	
8						
9						
10		平均	中央値	分散	標準偏差	
11	第一営業部	1800	1750	244000	493.96	
12	第二営業部	1800	1700			
13						

②のやり方

E11　=D11^(1/2)

	A	B	C	D	E	
1	(万円)	相川	飯田	上村	遠藤	大森
2	第一営業部売上（年間）	1200	1500	1600	3000	
3	**偏差**	-600	-300	-200	1200	
4	**偏差の二乗**	360000	90000	40000	1440000	
5						
6	(万円)	佐々木	志村	須藤	関口	相田
7	第二営業部売上（年間）	1900	3600	1000	800	
8						
9						
10		平均	中央値	分散	標準偏差	
11	第一営業部	1800	1750	244000	493.96	
12	第二営業部	1800	1700			
13						

よし計算できたぞ！　第一営業部の分散は244000、標準偏差は493.96になった。

9					
10		平均	中央値	分散	標準偏差
11	第一営業部	1800	1750	244000	493.96
12	第二営業部	1800	1700		
13					

第二営業部も同様に計算完了。部長に報告しよう！

	A	B	C	D	E	
1	(万円)	相川	飯田	上村	遠藤	大森
2	第一営業部売上（年間）	1200	1500	1600	3000	
3	**偏差**	-600	-300	-200	1200	
4	**偏差の二乗**	360000	90000	40000	1440000	
5						
6	(万円)	佐々木	志村	須藤	関口	相田
7	第二営業部売上（年間）	1900	3600	1000	800	
8	**偏差**	100	1800	-800	-1000	
9	**偏差の二乗**	10000	3240000	640000	1000000	
10		平均	中央値	分散	標準偏差	
11	第一営業部	1800	1750	244000	493.96	
12	第二営業部	1800	1700	1040000	1019.80	
13						

分散と標準偏差をより深く理解しよう

部長、計算できました。このエクセルシートを見てください。第一営業部は平均1800、中央値1750、分散244000、標準偏差493.96。第二営業部は平均1800、中央値1700、分散1040000、標準偏差1019.80となりました。平均は同じで、中央値は50異なり、分散と標準偏差にはそれぞれ約80万と約526の差があり、いずれも第二営業部の値が大きかったです。

なるほど。ここまではできるようになったんだね。すばらしい！それで？

それで？　と言いますと？

山田君、大事なことを忘れてないかな？

あ！　そうだ。目的…。

そう、目的。データ分析って、分析を始めるといろいろ発見があって面白いことも多い。そうなると、目的を忘れがちになるんだよ。どちらがいいチームと言えるかな？

なるほどー（ブツブツ言いながら考える山田君）。なんかこう一概には言えないというか。売上の平均は同じで、合計も変わらないけど散らばりが異なる…。

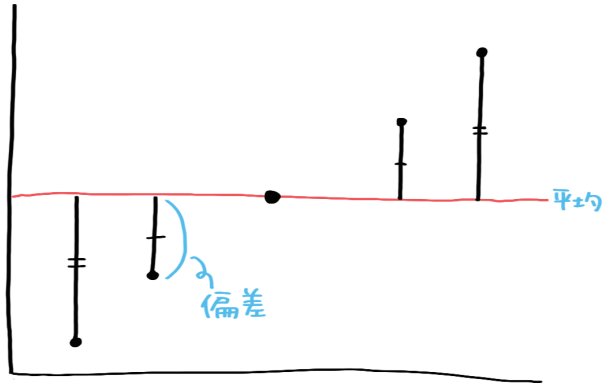
山田君、「散らばり」って言っているけど、そもそも分散ってなんなの？

うっ。そもそもですか！？

それでは、分散の公式って結局何をしているのかということを、図で考えて説明してみてくれるかな？　図に描いて公式の意味を理解しておくと、忘れない知識として定着するからね。例えば1, 2, 3, 4, 5 っていうデータセットがあるとしたらどうする？

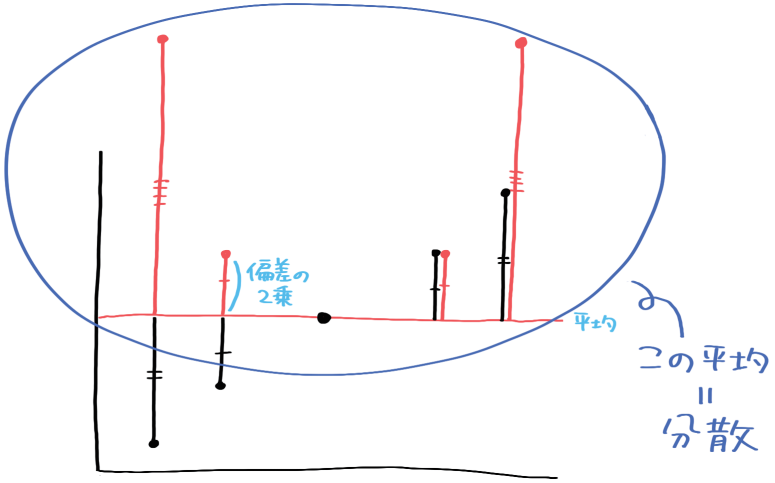
えっと、まず平均を計算すると3になって、そこから偏差を計算しますね。

そうだね。1, 2, 3, 4, 5のデータセットの偏差を図で書くと、こんなイメージになるのがわかるかな？



なるほど。それぞれの黒い点が1, 2, 3, 4, 5のデータで、赤い線が平均。そこからの差分が偏差っていう意味ですね。

その通り。そして分散はそれぞれの偏差を2乗して正の値に直してその平均をとっているので、こんなイメージになるよね。



なるほど、そうか。データ全体の平均からの距離が離れていればいるほど、分散は大きくなる。偏差のままだとマイナスがあったりするので、分散は距離として計算するため、2乗して正の値にしているんですね！　確かに散らばりが大きいデータって、それぞれのデータごとに値が変わるからイメージしづらいけど、平均を基準にどのくらい散らばっているか？　を考えているんですね。

そう、その通り！

ところで部長。これって、単純にマイナスをとった絶対値ではダメなんでしょうか？

いい質問だね。すごく簡単に言うと、「処理が面倒だから」という回答になるかな。絶対値は場合分けという形で整理して計算しないといけないので、なにも考えずに2乗してしまえばよいこちらの処理の方が簡単なケースが多いから、というのが答えになるね。

ありがとうございます。

山田君、実は標準偏差も分散と同じくデータの散らばりを表す統計指標なんだけど、そもそもなぜ存在するんだと思う？

（ご、ゴリラ部長のこの「そもそも攻撃」つらいな）えーっと、ちょっと待ってください。

散らばりを表す統計指標は、2つも必要ないんじゃないかな？

標準偏差でやっている平方根って、2乗した数を元に戻すという処理だから。分散を計算する時に2乗したので元に戻している的な感じですかね…。

分散の2乗のプロセスから整理して考えられているのは素晴らしいけど、じゃあなぜそもそも元に戻さないといけないのかな？

うっ。うーんと。

少し質問を変えると、平均、中央値、分散、標準偏差の4つの数値の中で、分散だけが仲間外れなんです。その理由はなんでしょうか？

	A	B	C	D	E	
1	(万円)	相川	飯田	上村	遠藤	大森
2	第一営業部売上（年間）	1200	1500	1600	3000	
3	偏差	-600	-300	-200	1200	
4	偏差の二乗	360000	90000	40000	1440000	
5						
6	(万円)	佐々木	志村	須藤	関口	相田
7	第二営業部売上（年間）	1900	3600	1000	800	
8	偏差	100	1800	-800	-1000	
9	偏差の二乗	10000	3240000	640000	1000000	
10		平均	中央値	分散	標準偏差	
11	第一営業部	1800	1750	244000	493.96	
12	第二営業部	1800	1700	1040000	1019.80	
13						

もっと言うと、売上と偏差を含めても、分散だけが仲間外れなんだけど。

分散だけ2乗している値ってことですかね…。何が仲間外れの理由かって言われるとちゃんと説明できないな。

うんうん。いい観点だね。言い方を変えると「単位が違う」んだよね。平均、中央値、標準偏差の単位はここでは「万円」なんだけど、分散は単位が「万円万円」という単位なんだ。だから、例えば平均の売上に対して散らばりがどれくらいなのかの比較がすごくイメージしづらいんだ。散らばりの数値だけを見るのであれば分散でよいけど、例えば平均1800万円に対して2500万円の売上がどれくらいの散らばりと言えるのかを考える際の基準としては、分散は単位が異なって比較できないから使えない。そのために平方根をとって2乗してしまった単位を元に戻した標準偏差が必要になる。

なるほど！　だから散らばりを示すという意味では同じ指標が2つもあるんですね。

その通り。ではここまで話したところで、再度結果を見てみよう。

	A	B	C	D	E	
1	(万円)	相川	飯田	上村	遠藤	大森
2	第一営業部売上（年間）	1200	1500	1600	3000	
3	**偏差**	-600	-300	-200	1200	
4	**偏差の二乗**	360000	90000	40000	1440000	
5						
6	(万円)	佐々木	志村	須藤	関口	相田
7	第二営業部売上（年間）	1900	3600	1000	800	
8	**偏差**	100	1800	-800	-1000	
9	**偏差の二乗**	10000	3240000	640000	1000000	
10		平均	中央値	分散	標準偏差	
11	第一営業部	1800	1750	244000	493.96	
12	第二営業部	1800	1700	1040000	1019.80	
13						

練習問題
分散と標準偏差を意思決定に生かす

今回の分析は、「どちらがいいチームか？」を調べることが目的だったね。実は、どちらがいいチームか悪いチームかは、状況によって変わるんだ。そこで、次の2つの問題に対して答えを教えてほしいです。あっ、次の会議の時間になってしまったから、あとは山田君の方で考えて、答えはメールでもらえるかな。

問題❶　**社長が収益目標をぶらしたくないと言ったら、仕事を任せるべきは第一営業部、第二営業部どちら？**

問題❷　**社長が今年は第二営業部に仕事を任せるべきだ！　と言ったら、それはなぜ？**

はい！　わかりました。

（自分の机に戻った山田君）

この問題は部長の視点で書かれているんだな。こういう問いに対して、部長はデータを元に社長に提案しているのか…。まず問題❶収益をぶらしたくない時。儲かればいいってもんじゃなくて、特に上場したりすると予測との乖離が大事になるからとにかくブレがない方がいいのか。平均から予測を作るとすると、そこからブレないのは分散や標準偏差の小さな第一営業部ということかな。なのでこのデータからわかることは、売上ベースのブレは第一営業部の方が少ないので、第一営業部に任せるということでよさそうだな。

次に 問題② 。このデータから見てわかることは、第二営業部の方が標準偏差が大きいってことか。平均からブレてもよいっていうケースは…。うーんと、儲かっている時とかなのかな。儲かっていればとりあえずゼロでもいいし、大きく儲けを狙うことも考えてもよい。よし、これでいったんメールしてみよう。送信！

（3分後。早速、ゴリラ部長からの返信が届いた。）

お疲れ様山田君。

問題①はその通りだと思います。問題②は半分正解。もう1つの答えとしては、儲かっているというケースもあるけど、逆に儲かってなさすぎて一か八か勝負する時にばらつきが大きい（大きく儲けた経験がある）第二営業部にかけてみようというパターンもあるかな。

今回の問題をお願いする中で伝えたかったのは、統計とか分析って、ほとんどが平均を計算して終わりっていうことが多いんだよね。けど平均が同じでも、そのデータの中の散らばりを見ていくことが大事。同じ平均1800万円でも、極端な話、売上0万から5000万円までいるチームと1750万と1850万の相対的に狭い範囲に収まっているチームとでは、全然意味が違う。統計というのは、その平均ありきで、どれくらいの範囲まで、どれくらいの確率で数値が行きうるのか。それを考えていくものです。別の言い方をすれば、95%の確率で0万～5000万になるチームと1750万から1850万になるチームがあれば、同じ平均1800万でもデータが表す現実はまったく異なることになる。その点を理解してもらえたら嬉しいです。

で、次に考えてほしいのは、これってそもそも10名のデータを見ただけだけど、第一営業部って50名いるんだよね。だから、本当にその分析って正しいの？　ということです。一部の傾向だけで、全体の傾向を見て取れるのか？どう思うかな？

確かにそうなんだよな。一部のデータだけ見ても、全体の傾向を掴めてるとは言いづらいけど、参考にならなくもない。さて、どうしたらよいのだろうか？

column

記述統計、推測統計、ベイズ統計

ここでいったん、統計にはどんな種類があるのか整理してみましょう。統計を学ぶ上で全体像を理解していただく、そして本書の内容が、データを整理し推測するためにそれを生かすという流れで書かれている理由を理解していただくのに役に立つからです。

統計には、大きく分けて記述統計、推測統計、ベイズ統計の3種類があります。

記述統計

現在ある数値データを整理して、その意味合いを知ることを目的としたもの。例えばこの地域で穫れた穀物がどれくらいの量だったか、去年と比べて多いのか少ないのか、といった目的でデータの取得を始めるものです。記述統計は、得られたデータからその特徴を抜き出すテクニックのこと（今回計算したように、すでにある第一営業部と第二営業部の売上から平均値などを分析する）。17世紀頃の欧州で、死亡率などの数字を使って分析を行っていたところに起源があるものです。世界史を受講していた人は、資料集の中に羊皮紙にデータがびっしり書かれていたという記憶がある人もいらっしゃるはず。持っている標本データの傾向をわかりやすく把握するための統計と言えるでしょう。

推測統計

統計学と確率分布を使って「調査しきれないほどの大きな対象」や「将来に関すること」の推測を行うものです。調査しきれない理由は、費用です。ちなみに、日本の国勢調査にかけられている予算は720億円。20世紀になって確立された方法論で、「部分から全体を推測する」というのが特徴です。持っているデータ（この後説明していきますが、標本と言います）から全体の傾向（母集団）を推測するものになります。

ベイズ統計

データが不十分でもある事態が発生する確率を分析することが可能で、データの母集団が変化することを前提に行う考え方になります。Googleやマイクロソフトの検索エンジンやアマゾンのリコメンドエンジンなど、いわゆるGAFAM企業がこの統計手法を使っていると言われています。こちらの内容は、基本的には本書では取り扱いません。

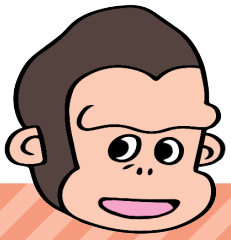
第1章まとめ

- ☑ 平均は統計の基本中の基本であり、とてもよく使われかつ非常に大切な統計指標。ただし平均だけを見ているとデータの全体の散らばりを把握できない。
- ☑ 中央値と平均は、外れ値がある時に結果が大きく異なる。
- ☑ 散らばりを表すためには、分散や標準偏差が使われる。
- ☑ 分散と標準偏差の一番の違いは単位。元のデータとの比較でどれくらい散らばっているかの指標には標準偏差を使用する。
- ☑ 平均と分散や標準偏差のセットで見ていくことで、データの実態がよりわかってくる。

第 2 章

正規分布

過去のデータから
将来の売上を
予測しよう！



前章までのあらすじ

部長から第一営業部と第二営業部の売上を分析するよう依頼された山田君。そこで平均を計算してみると、どちらも平均1800万円と出てきました。しかし第一営業部は多くの人が売上1500万円近くだったのに対し、第二営業部は700万円の人から3700万円の人まで、売上が散らばっていました。散らばりの指標である分散と標準偏差を学び、意思決定に生かすやり方も学んだ山田君。そこに、部長からの質問「これってそもそも10名のデータを見ただけだけど、第一営業部って50名いるんだよね。本当にそれって正しいの？　一部の傾向だけで全体の傾向は見て取れるのか？　どう思うかな？」を受けて、山田君は考え始めます。

contents

少ないデータから多数のデータを推測する

入社してすぐにデータ分析のチームに配属された山田君。手元にあるデータに対して、平均や標準偏差などを計算することで整理、可視化するところまでは学んだが、そこで部長から飛んできた質問

「これってそもそも10名のデータを見ただけだけど、第一営業部って50名いるんだよね。本当にその分析って正しいの？　一部の傾向だけで全体の傾向を見て取れるのかな？」

に、頭を悩ませていた。

確かに50名のデータは10名のデータとは違う。けど、10名の傾向を取れば50名の傾向を推し量ることはできるような気がする。できる、いや、まったくできないわけじゃないという言い方が正しいかな。本でも、統計には記述統計と推測統計があると書いてあって、なんかこの辺が関係しているのかな。混乱するなあ。

“トントン”

はい！

お疲れ様。なんか頭から湯気が出てるよ。

あ、部長（こんなに図体大きいのに気配がまったくなかった）。お疲れ様です。

何か悩んでる？

はい、先ほどいただいた宿題に関して、考えが堂々巡りしてます。50名のデータを10名で予想できるのか？　について考えていたのですが、まったくできないわけでもないような気がして。けどやっぱり10名だと全部の傾向を正確に見るのは難しいような。

なるほど。できないような、できるような、というその感覚を、もう少し言葉にして説明できる？

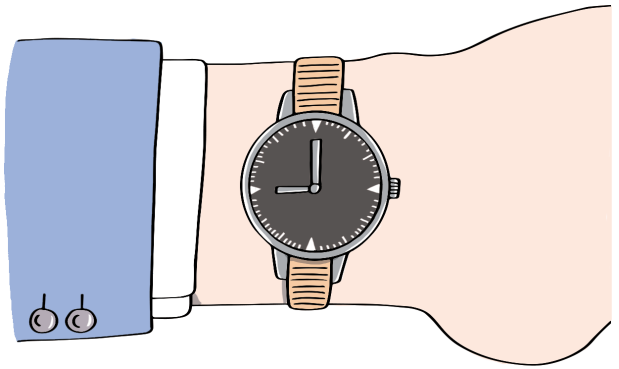
そ、そうですね…。うーん。（沈黙）

そうだ、山田君。そんな時はね。

は、はい。

ラーメン食べに行こう！

ラーメンですか？



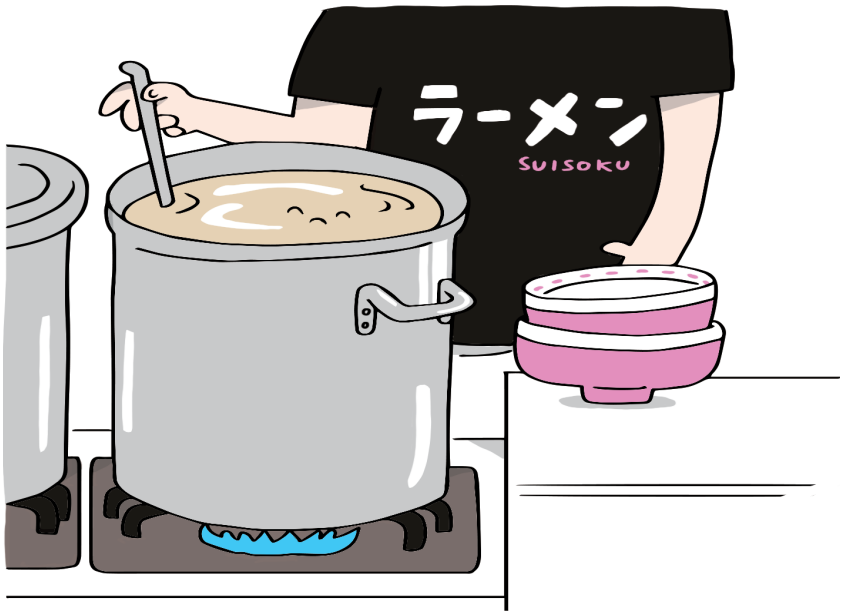
あー、もう21時か！　ラーメン大好きなので、ぜひ行きたいです（仕事終わらないな）。

ゴリラ部長は、山田君を連れてオフィス裏の小ぎれいなラーメン屋さんへ向かった。店内は古いながらも清潔に保たれていて、豚骨と魚介が混ざったいい香りがする。仕事終わりのビジネスパーソンと学生で店内は混んでいたが、ちょうどカウンターに2席空きがあり座ることができた。カウンターからは、厨房でラーメンを作る親方らしき人とサポート役の人の動きがよく見えた。

ここのラーメン、とっても好きなんだよね。スープが上品なのにしっかりとした味があって。麺も好みの中太たまごちぢれ麺。

そうなんですね。私も学生時代からラーメン大好きでした。

山田君、ここから見える風景は普段通っているラーメン屋さんに近いかな？　寸胴や丼やカウンター。寸胴のスープは、親方のこだわりで仕込みから48時間も煮込んでいるらしい。



はい。大学時代に私が行っていたラーメン屋さんも、こんな雰囲気でした。

親方の動きを見てみよう。よーく見ていてね。

親方は大きな寸胴の中から煮込まれたスープを取り出し、温められていた丼に移し替えた。丼の中は空で、ただスープを移し替えただけである。

山田君、ここで1つ質問。いま、丼の中は空だったね。そして丼は適度に温められていた。寸胴と丼は温度や湿度などの条件が同じで、丼が空だったと仮定した時、寸胴のスープとたった今注がれた丼のスープの味は同じだと思う？

えっ。スープの味ですか？

そう。寸胴のスープと丼のスープの味。

ほとんど同じじゃないですかね？　ただ移し替えただけだし。

ゴリラ部長の目がキラッと光る。

なんでいま、「ほとんど」と言ったのかな？　親方としては困るよね、スープの味が変わったら。

えーっと。そうですね、けど「厳密に同じ」かと言うと、ちょっとくらい違いがあるんじゃないかなと。微妙な濃度とか。

いいポイントだね。確かに、寸胴のスープと丼のスープの間に大きな差はなさそう。けど、まったく同じとは言い切れない。じゃあ、2つのスープの味をなるべく同じに近づけるためには、どうしたらよいと思う？

えーっと。すっごく単純な答えですが、

「よく混ぜる」

とかですかね。

そうなんだよ！！　その通り！！　あと、ほとんど同じ意味だけど、均等に全体から掬うということも大事だよね。底の方に溜まった豚骨の骨とかばっかりにならないようにとか。

そうですね。きちんと混ぜて均等に掬えば、だいたい同じになりますよね。親方もきちんと混ぜてましたね。ん、待てよ…。

どうした？

なるほど部長、そういうことだったんですね！　そうか、50名の部員を10名で推測できるかどうか。わかったような気がします。

よかったよかった。データの可視化という観点では、ヒストグラムというものを調べてみるとよいかもしれないね。

山田君はラーメンをすごい勢いで食べ終わると、オフィスに戻って分析を開始しました。

推測統計とは？

ここで、ゴリラ部長から「推測統計とは何か？」ということについて簡単に説明させていただきます。P.25で説明したように、統計の定義は

集団（データ）の傾向や性質を数量的に明らかにする

ということでした。この表現を使って推測統計の説明を行うと、手元にある一部のデータに確率分布という「型」や「パターン」を当てはめて、データ全体の性質を数量的に明らかにするということになります。もう少し噛み砕いて言うと、

今持っている少ないデータを使って、より多くのデータや未来のことを、あるパターンに当てはめて予測できるようにすること

という風に表現できると思います。

標本と母集団

データ分析を行う時に、常に頭に入れておいてほしい図がこちらになります。