■目次

1 仕事と数値

1.1　あなたの仕事は何ですか？

美容師なら　お客様の美容を整えること

イラストレーターなら　イラストを描くこと

料理人なら　料理を作ること

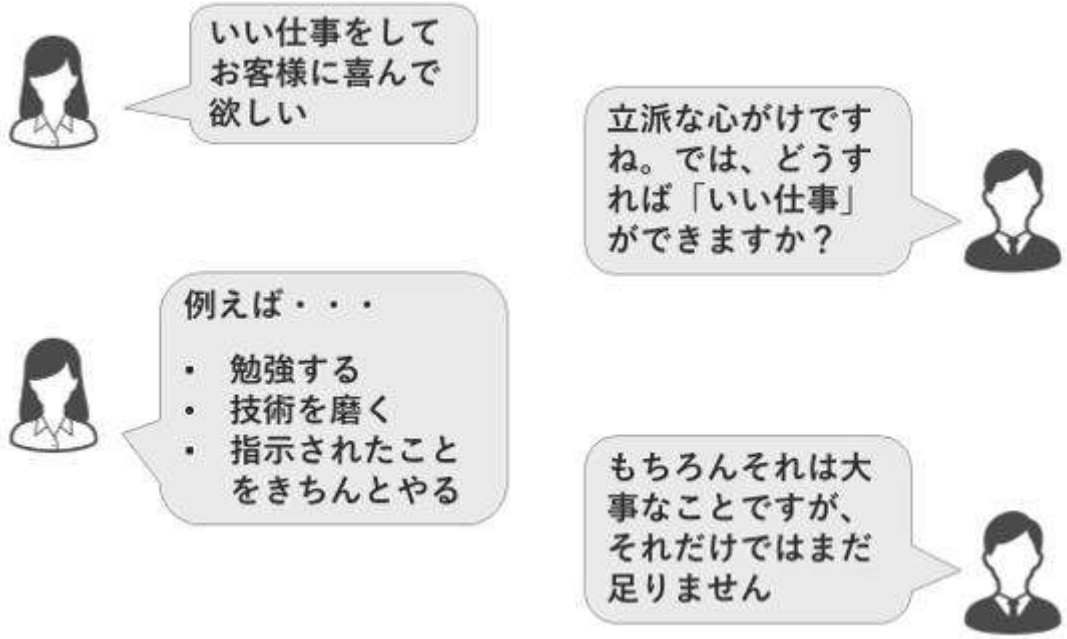
……それだけ、だと思ってはいませんか？

あなたは今、美容/美術/料理/医療などさまざまな専門分野の知識・技術を学んでいて、将来その技術を活かした職業につく方が多いことでしょう。

- 美容師なら、お客様の美容を整えること。
- イラストレーターなら、イラストを描くこと
- 料理人なら、料理を作ること

それが将来自分のすべき仕事である、と考えている方が多いはずです。もちろんそれらの仕事が一番大事な「柱」であることは間違いありません。しかし、それだけでは不十分です。「いい仕事」をしたいと思うのであれば、他にも考えなければならないことがあります。

1.2　「いい仕事」はどうすればできますか？

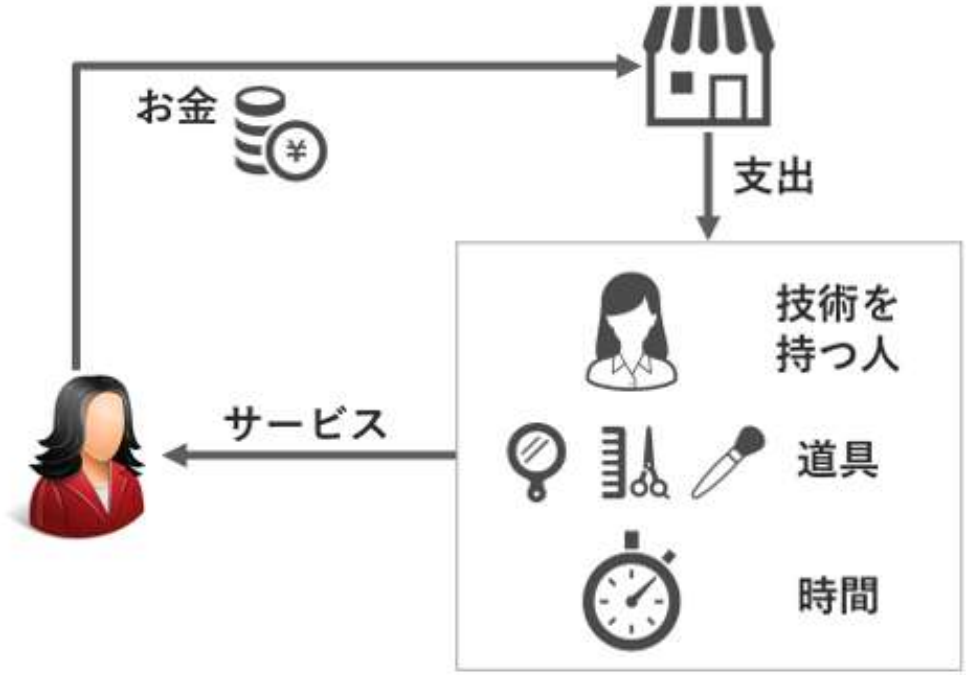


「いい仕事」をしてお客様に喜んでもらいたいというのは立派な心がけですし、そのために

- 勉強する
- 技術を磨く
- 先輩/上司に指示されたことをきちんとやる

のは大事なことです。しかしそれだけでは「いい仕事」はできません。特にあなたが 5 年 10 年と経験を積んで店長を任されたり、独立して自分の店を構えたりするようになると、「技術」だけでは足りないのです。

1.3　「いい仕事」をするには「お金」が必要

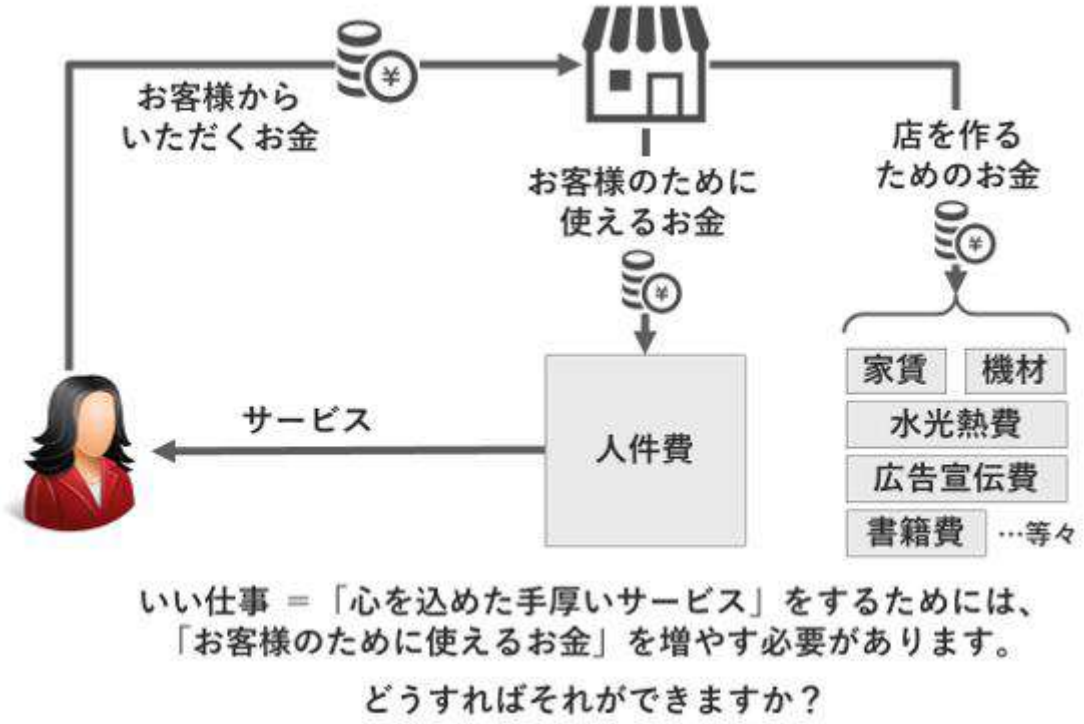


いい仕事をするためには「お金」が必要です。お金は当然、お客様からいただくものです。たとえば美容室の場合、技術を持つ人を雇い、道具を買いそろえ、時間を使ってお客様に美容のサービスをしています。人・道具・時間を確保するためには十分なお金が必要です。これは美容室に限らず、どんな仕事でも本質は同じです。

「お金がなくても精一杯やります」という心がけは必要ですが、結局のところ心がけではどうにもならない部分が多いのです。

（注：ここでいう「サービス」は仕事をするという意味であり、無料奉仕という意味ではありません）

1.4　お客様のためにお金を使うには？



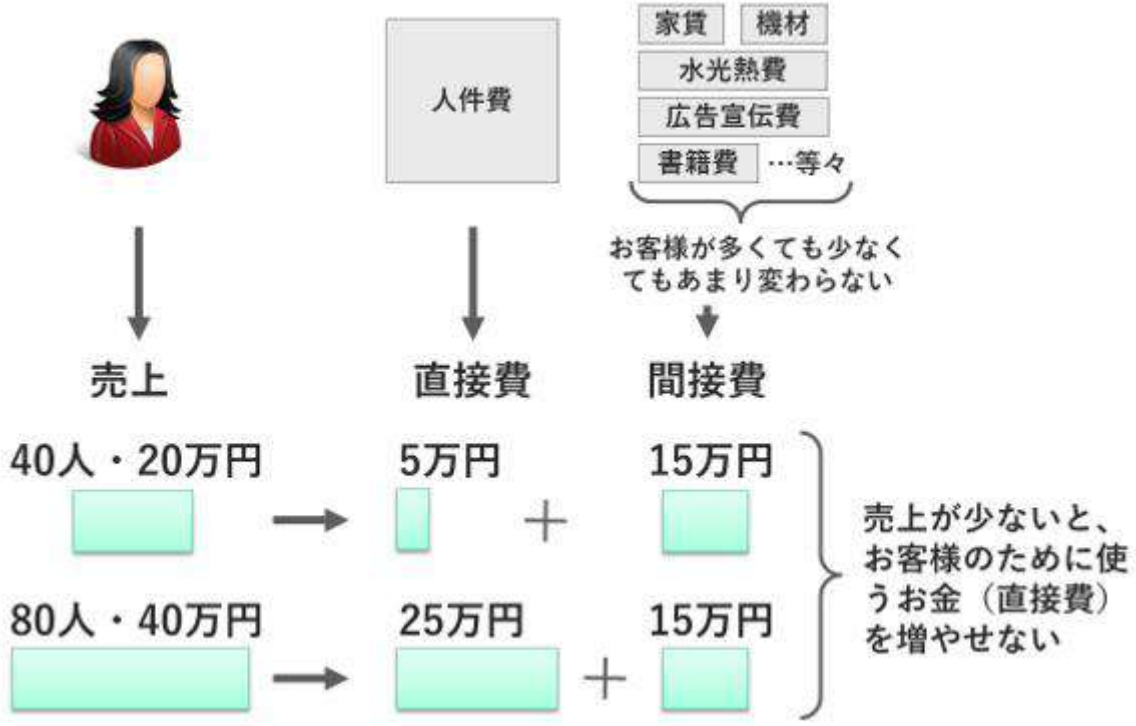
お客様からいただく「お金」は実は大まかに 2 種類の目的に使われます。

1 つは家賃/機材/水光熱費/広告宣伝費等、「店」を成り立たせるために使われるお金です。
もう一つは、お客様のために使われるお金です。多くの場合、「人件費」がこれに当たります。製造業や飲食業の場合は「材料費」もここに含まれます。

「いい仕事をする」ためには「心を込めた手厚いサービスをする」必要があり、そのためには「お客様のために使えるお金」を増やす必要があります。

どうすればそれができるでしょうか？

1.5　売り上げを増やせば直接費を増やせる



「お客様のために使えるお金」と「店を作るためのお金」をここではそれぞれ直接費、間接費と短く呼んでおきます(注)。たとえば家賃はお客様が 0 人でも 100 人でも変わらないように、間接費はお客様が多くても少なくてもあまり変わらないため、総額で 15 万円固定でかかると仮定します。客単価を 5,000 円として、お客様が 40 人・売上 20 万円のばあいと、80 人・40 万円の場合でお客様一人当たりに使える直接費を計算すると、

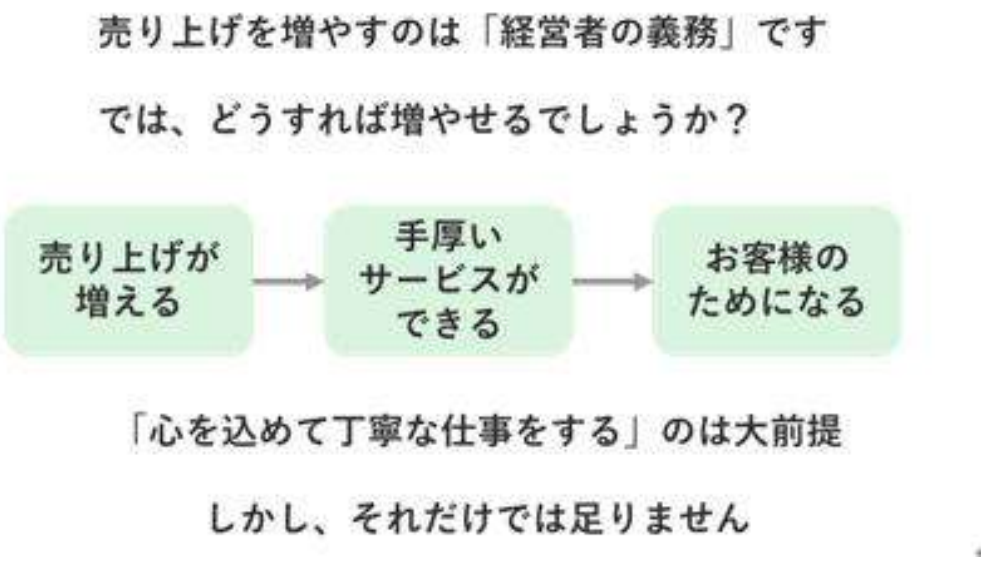
40 人・20 万円の場合：　直接費 5 万円÷40＝1 人当たり 1,250 円
80 人・40 万円の場合：　直接費 25 万円÷80＝1 人当たり 3,125 円

となり、一人当たりで 2.5 倍の差が出てしまいます。「お客様のために使えるお金＝直接費」が多ければその分丁寧な「いい仕事」ができます。

いい仕事をするには「技術を磨く」だけでなく、「売り上げを増やす」ことも非常に重要なのです

(注) 会計学上は、人件費がすべて直接費に分類されるわけではなく、水光熱費等がすべて間接費になるわけでもありませんが、大まかに言って人件費は直接費に、家賃や水光熱費等は間接費になるなる割合が大きいためこのような書き方をしています。

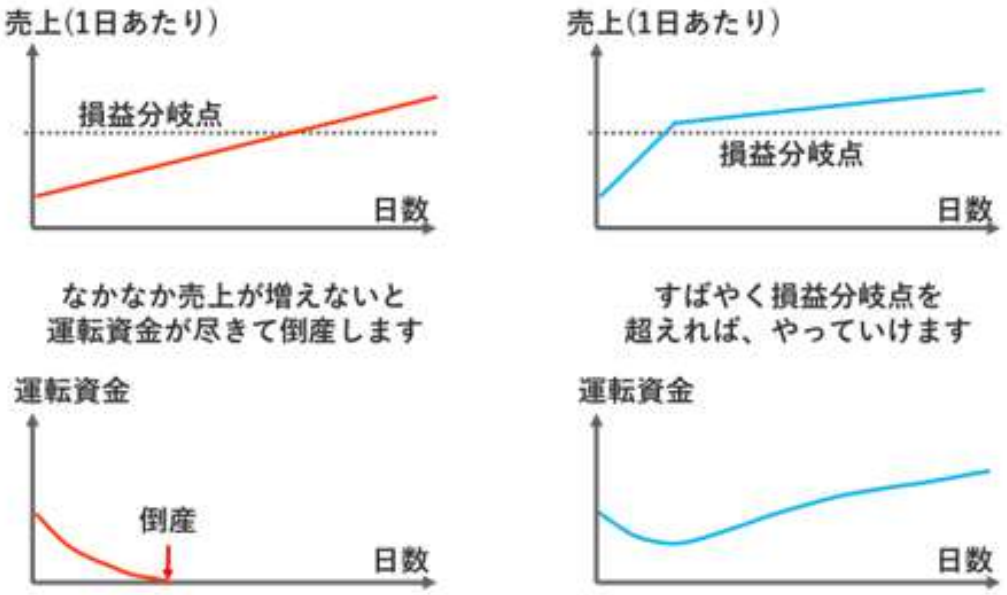
1.5.1 どうすれば売上を増やせるか？



売り上げが増えれば、お客様の支払額が変わらなくても手厚いサービスができますので、お客様のためになります。したがって、売上を増やすのは「経営者の義務」と言えます。では、どうすれば増やせるでしょうか？

「心を込めた丁寧な仕事をして、お客様に満足してもらえば、また来てもらえて売上が増える」という面は確かにあります。しかしそれだけでは「お客様が増える」のに長い長い時間がかかります。集客ペースが遅く、損益分岐点を超えるのに時間がかかると運転資金が尽きて倒産してしまいます。

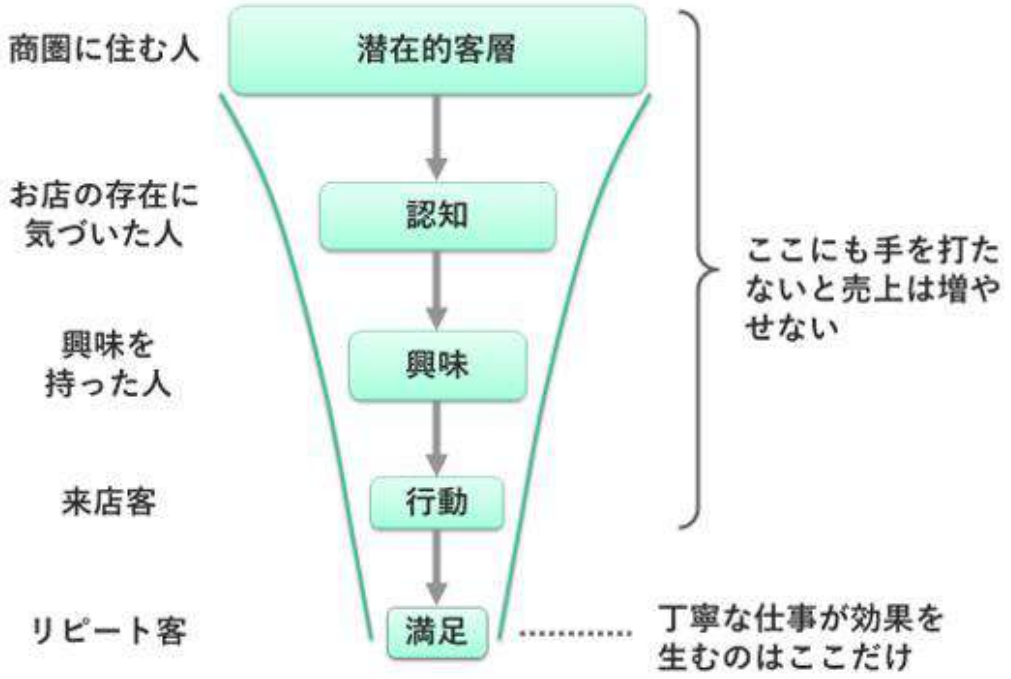
1.6　集客ペースが遅いと運転資金が尽きる



損益分岐点：売上がこの額以上になれば利益が出る、という金額のこと

運転資金：お店の定常的な支払に必要な資金。これがゼロになると倒産

1.7　丁寧な仕事が効果を生むのは最後だけ



この図は「マーケティング・ファネル」と呼ばれるもので、「潜在顧客がリピート客になるまでのステップ」をモデル化したものです。

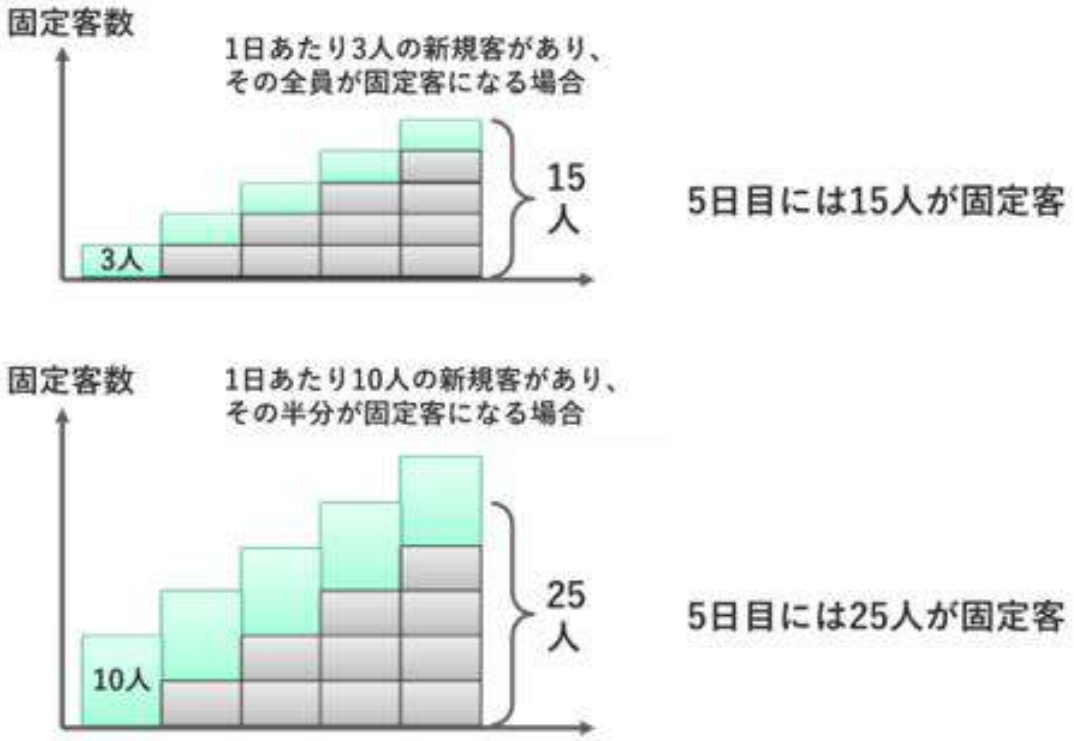
最上部の「潜在的客層」は商圏に住む人全員です。たとえば日常の買い物をするコンビニエンスストアの商圏はおおむね半径 500 メートル以内、スーパーならば 1～数 km、大規模ショッピングモールなら 10km 以上になります。美容室は日常的に来店しないためコンビニよりは広いですが、これといった特色の無い店の場合はコンビニと同程度です。ただし特色のある店ならば 10km 単位の遠方からの来店も期待できるため、店の性格によって大差があります。

しかし、いくら商圏に住んでいても、存在が知られていない店には客は来ません。通勤の路上に店がある、店の広告がある、チラシがポスティングされていた、ネットで検索したら見つかった、など、なんらかの理由で「気がついた人（認知）」のうち、「興味を持った人」が来店という「行動」を起こし、「満足」すると再来店（リピート）を期待できます。

「潜在的客層」から「認知」→「興味」→「行動」→「満足」までのそれぞれの過程でどんどん人数が減っていくため、漏斗（英語でファネル）の形に似ていることからマーケティング・ファネルと言います。**新規客を獲得して固定客（継続的にリピートしてくれるお客様）へと転換していく**ためには、潜在的客層に認知させる、認知した者に興味を持たせる、さらに行動させ満足を得る、と

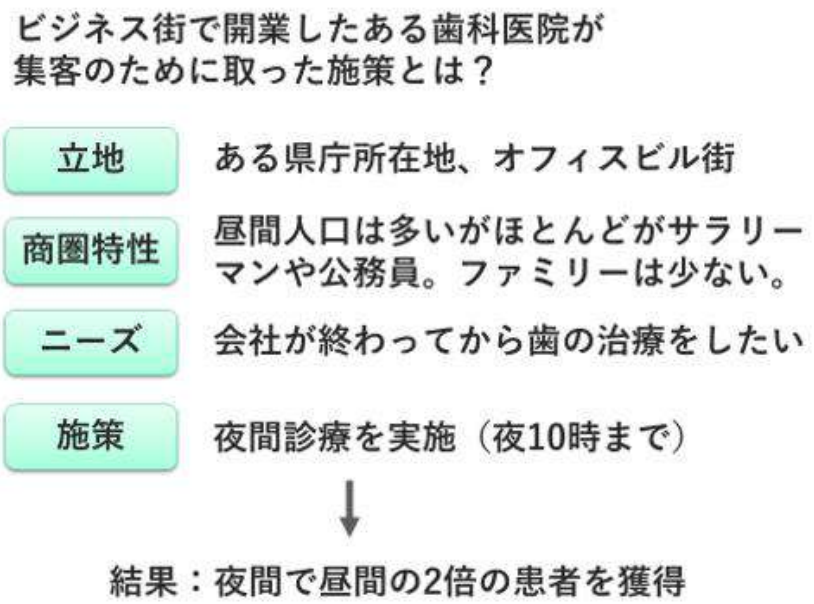
いう各段階でどんな手を打つのかをきちんと考えていく必要があります。
このファネルの中で「心を込めた丁寧な仕事」が効果を生むのは最後の「満足」の段階だけです。1日の新規来店客が3人しかいなければ、どんなに「いい仕事」をしたとしても固定客は1日最大3人しか増えません。しかし、「認知」「興味」の段階で適切な手を打って1日に10人の新規客を獲得できれば、固定客になるのがそのうち半分しかなかったとしても1日に5人ずつ増えることになります。

1.8　新顧客が増えないと固定客も増えない



「丁寧な仕事が効果を生むのは最後だけ」であり、「行動」までの段階はそれ以外の方法で集客を図らなければいけない、ということを肝に銘じてください。

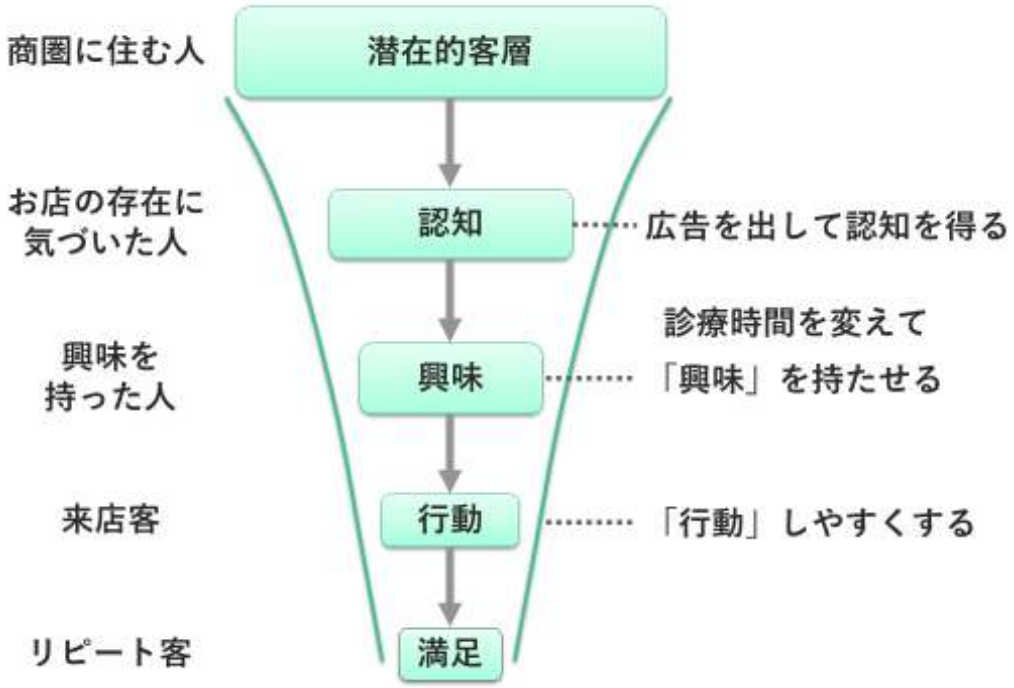
1.9　事例：歯科医院の集客戦略



ここで、「認知」「興味」段階に力を入れて集客を果たした歯科医院の事例を紹介します。ある県庁所在地のオフィスビル街で開業したある歯科医院は、「昼間の人口は多いがほとんどサラリーマンや公務員であり、昼は忙しいので仕事が終わってから歯の治療をしたい」というニーズがあることに目を付けて、当時は珍しかった夜 10 時までの夜間診療を実施したところ、夜間で昼間の 2 倍の患者を獲得して一気に経営を軌道に乗せた例があります。

このように、「新規客の来店を得る」までの過程では「仕事の技術」以外の部分で集客施策を考えたほうがよいケースが少なくないのです。

1.10 何がどこで効果を生むかを考える



もう一度、マーケティング・ファネルに照らし合わせて考えると、夜間診療というのは「お、便利だな」という「興味」を持たせるための施策であるとともに、興味を持った人が「行動」しやすくもなるという一石二鳥の効果があります。

しかし「興味」を持たせる材料があったとしてもその前に「認知」が必要です。医院そのものの前を通る人しか気づかないようでは「認知」する人数が少なすぎます。たとえば、「認知」を増やすためには「主要な通りに広告を出す」といった方法が可能です。その場合、「認知」と同時に「興味」を持たせるために、「夜間診療実施」を大々的にアピールした広告が適切でしょう。つまり下記 A タイプのほうが B タイプよりも「興味」を持たせる効果が高いと思われます。広告を出す場合はこのように「強調するキーワード」の選択で効果が大きく違ってきます。

Aタイプ

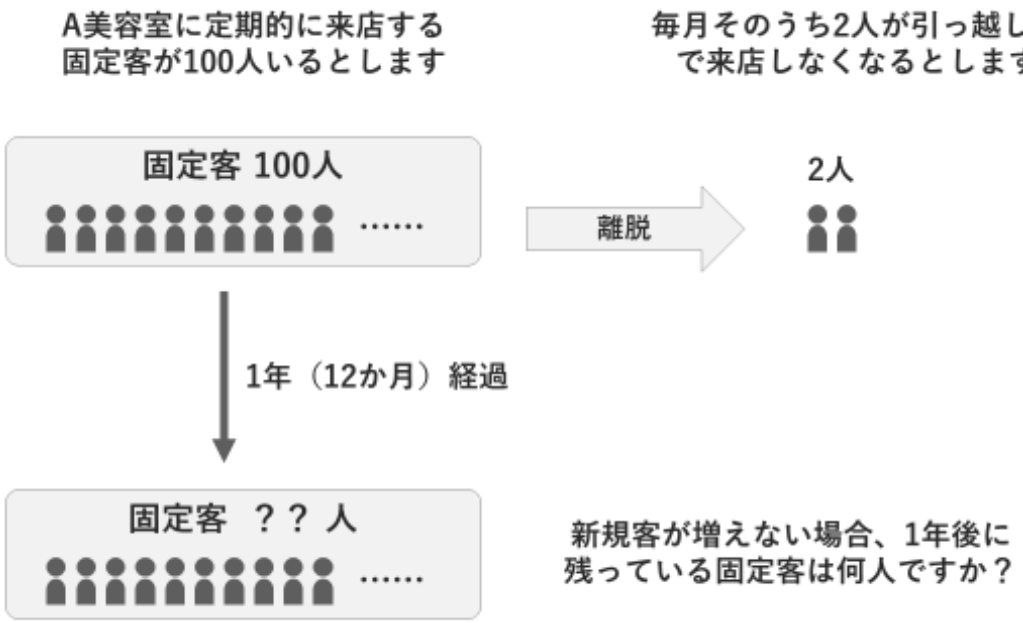
○○歯科医院
夜間診療（午後10時まで）
XXビル2F

Bタイプ

○○歯科医院
歯周病/小児歯科/インプラント
XXビル2F

なお、歯科医院向きの方法ではありませんが、もし飲食店のような業種であれば、行動を促すために期間限定のスペシャルメニューを作る、クーポンを配布するといった方法もよく使われます。

1.11 数字で考えてみましょう



マーケティング・ファネルを踏まえて集客施策を考えるためには、「数字で考える」習慣が欠かせません。

練習問題として 1 つ考えてみましょう。
あるとき、A 美容室に定期的に来店する固定客が 100 人いたとします。しかし毎月そのうち 2 人が遠方への引っ越し等で来店しなくなるとします。新規客が増えない場合、1 年（12 ヶ月）経過後に残っている固定客は何人でしょうか？

解答：月に 2 人ずつ×12 ヶ月間減るわけですから、答えは当然

100 － 2×12 ＝ 100 － 24 ＝ 76 人

となります。どんな業界でもこのように固定客は必ず減っていくものですので、常に新規客を獲得していく努力をしなければなりません。それには、

当店の固定客は現在、何人いる
通常のペースだと月に何人ぐらい減っていく

といった数字を把握する意識・習慣が必要です。

1.12 数字を把握すれば異変に気付ける

A美容室では毎月末にその月の来店客を新規と既存に分けて人数を把握しています。なお、同じ人が月に2度以上来店した場合も1人と数えています。

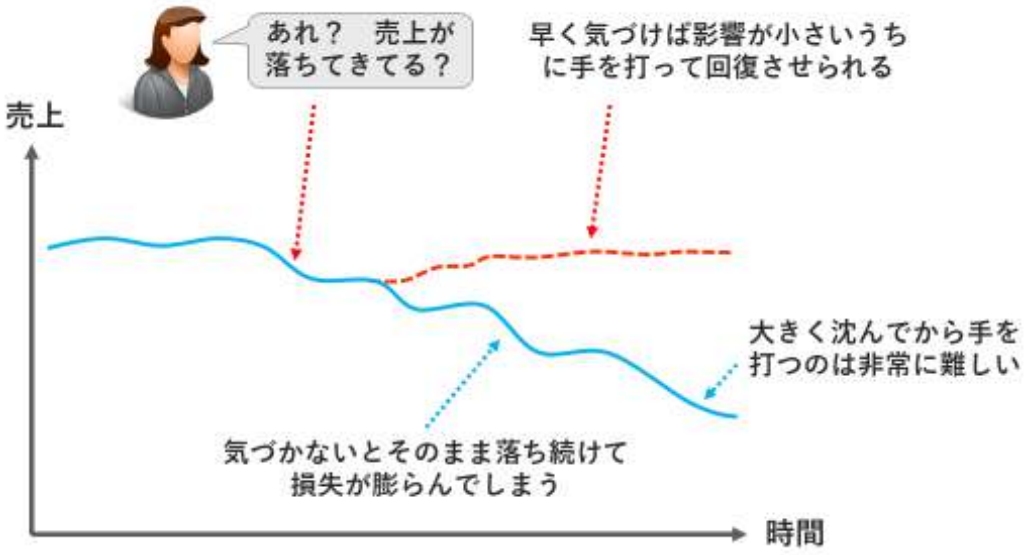
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
既存	88	80	99	95	89	87	80	82	81	85	82	80
新規	2	3	2	10	3	20	3	2	3	4	2	2

1. この数字から気になるところを探してください。「気になる」というのは、「あれ？　どうしてこうなったの？」と思うような、多すぎるか少なすぎる数字です

2. そのような異変が起きた理由としてはどんなものがありえますか？　ありうる理由を思いつく限り挙げてください

数字を把握するトレーニングとして今度はこの問題を考えてみましょう。
A 美容室では毎月末にその月の来店客を新規と既存に分けて人数を把握しています。なお、同じ人が月に 2 度以上来店した場合も 1 人と数えています。上図中の指示に従って、1. 気になるところを探し、2. その異変が起きた理由を思いつく限り挙げてみてください。

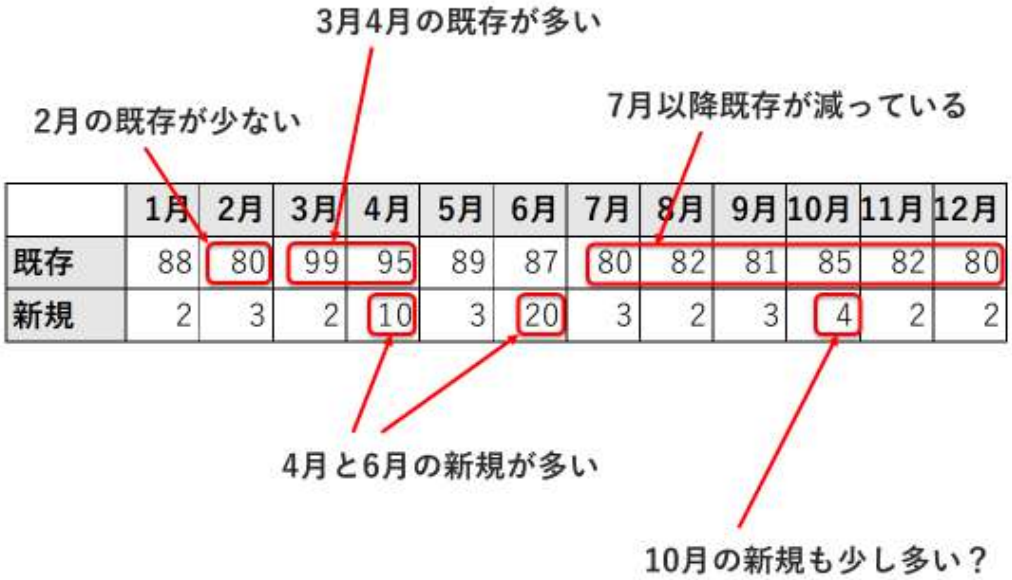
1.13 「問題」に早く気づかなければならない



1.13.1 では、どうすれば「問題に早く気づく」ことができるのか？

ふだんから数字に注意していると、「売上が落ちてきた」といった「問題」にいち早く気づいて手を打つことができます。早く気づけば影響が小さいうちに手を打って回復させられますが、気づかないとそのまま落ち続けて損失が膨らんでしまいます。大きく沈んでから手を打つのは非常に難しいものです。

1.14 気になる数字



	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
既存	88	80	99	95	89	87	80	82	81	85	82	80
新規	2	3	2	10	3	20	3	2	3	4	2	2

答え合わせです。「気になる数字」としては、上図のような各点が挙げられます。皆様が気づいた部分が他にあれば書き足しておいてください。

では、このような数字になった理由は何なのでしょうか？

「2 月の既存が少ない」理由の候補

- もともと 2 月は 1 月に比べて短いので、客も少なかった
- 新年や成人式のため 1 月の来店客が多かった

「3 月 4 月の既存が多い」理由の候補

- 3 月 4 月は卒業式・入学式や人事異動といった儀式が多いから

「4 月の新規が多い」理由の候補

- 進学や転勤で引っ越してきた人が多いから

などなど、このようにちょっとした数字の変化に注目して「なぜだろう？」と疑問を持ち、「理由はこれではないか？」という「仮説」を考える習慣が極めて重要です。

たとえば 2 月と 7 月の既存客は同じ数になっていますが、7 月は 2 月よりも 3 日長いのに同数というのは不自然です。1 日短い 6 月に比べて 7 人減っているのも不自然です。これがたとえば「学生が夏休みに入って帰省していなくなるので減ったのだろう」といった理由であれば問題はありませ

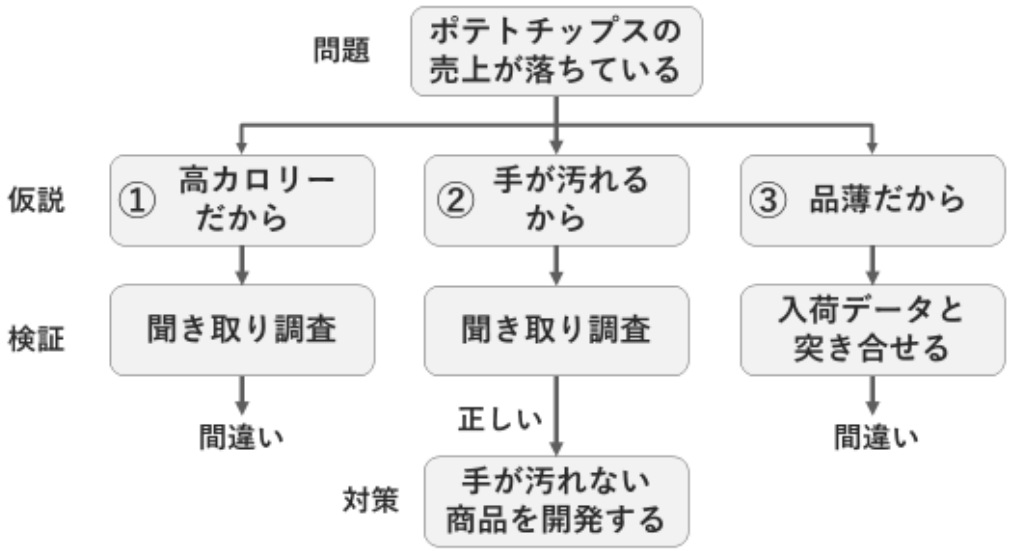
んが、何か店に不満があって離れていった場合は何か手を打たなければならないでしょう。

6 月の新規客が不自然に多いのも気になるポイントです。飲食店などでよくあるケースですが、TV などに取り上げられて一時的に客が増えるとそれまでの常連客が来づらくなり、その後かえって売上が落ち込むことがあります。たとえば 6 月は近隣の別な美容室が一時休業していてその客が流れてきたために新規が多く、しかしその結果既存客が予約を取りにくくなってさらに別な店に流れたために 7 月以降は落ち込んだ、といった可能性はないでしょうか？

もちろん、数字だけでは真相はわかりません。数字はあくまでも考えるきっかけです。そのきっかけを手がかりに真相を探り、問題があれば手を打つ必要があります。

1.15 「仮説」を考える習慣を持とう

- 何か困ったこと（問題）が起きたら、まず原因の「仮説」を考えます。
- 仮説の大半は間違いですので、「検証」を行います。
- 検証の結果、正しかった仮説を元に「対策」を考えます。



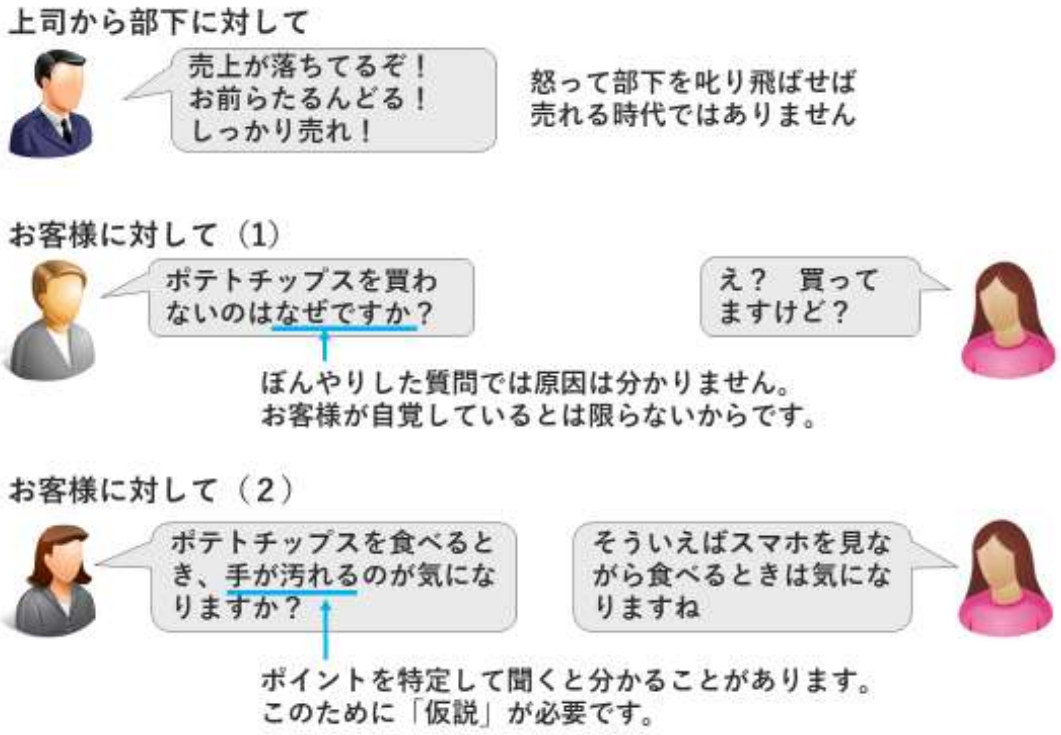
「仮説」を考える習慣は非常に大事です。何か困ったこと（問題）が起きたら、まず原因の「仮説」を考えるようにしてください。

「仮説」は間違っていてもかまいません。というよりたいていの場合間違っているので、「検証」を行います。上図ではある食品スーパーでポテトチップスの売上が落ちているという問題から「仮説」を立て、「検証」して「対策」を考えた例です。

①高カロリーだから敬遠されている、②手が汚れるから敬遠されている、③そもそも品薄で売り切れただけ、という 3 つの仮説を立てた場合、①と②についてはお客様への聞き取り調査をすれば、③については入荷数量のデータと突き合せれば検証できます。検証の結果②が正しいと分かれば、手を汚さずに食べられる種類の商品を開発してそれをアピールすれば売上は回復する可能性があります。

実際、近年スマートフォンの普及にともなって手が汚れるポテトチップスが敬遠されているという事実があるため、一部のメーカーは商品のパッケージを工夫して「手を汚さず食べられるチップス」を開発しています。商品そのものではなく、パッケージ等を変えることで売上が変わることはよくあるのです。

1.16 なぜ仮説が必要なのか？



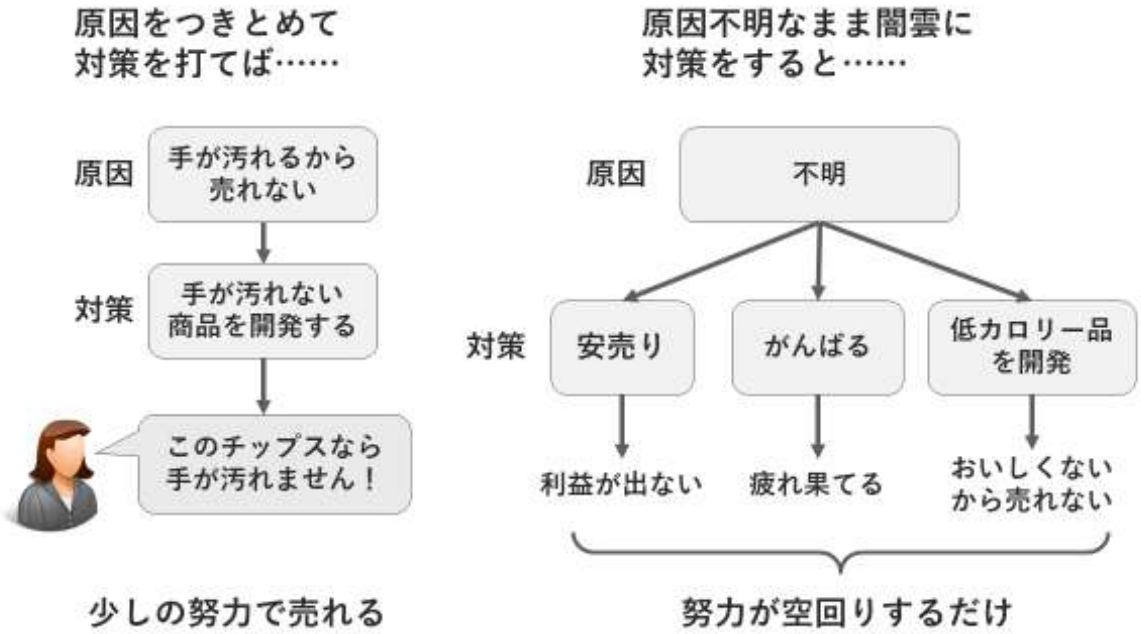
「仮説」を立てないとどんなことが起きるのでしょうか？
「売上が落ちている」という問題が起きたとき、1 番目の上司のように「お前らしっかり売れ！」と怒るだけというのはダメなパターンです。原因をさぐらずにただ「頑張る」だけで売れる時代ではありません。

かといって次の「お客様に対して（1)」のようにぼんやりした質問をしても原因はわかりません。人は自分の行動の理由を自覚していない場合も多いのです。

「お客様に対して（2)」のようにポイントを特定して聞くと、「そういえば……」とお客様自身もそこで気がつくことがあります。しかしこのような質問をするためにはその前に「仮説」が必要なのです。

なお、「なぜですか？」のようにキーワードを特定しない質問をオープン・クエスチョンといい、「手が汚れるのが気になりますか？」のように特定して聞く質問をクローズド・クエスチョンと言います。クローズド・クエスチョンの長所は相手が自覚していない答えがわかることで、短所はこちらが仮説を持っていない答えは得られず、誘導尋問になりかねないことです。オープン・クエスチョンはその逆の特徴があります。どちらも一長一短があるので、問題の原因を探るために聞き取り調査を行う場合は、オープンとクローズドの質問を適材適所で使い分ける必要があります。

1.17 原因不明だと努力が空回りしてしまう



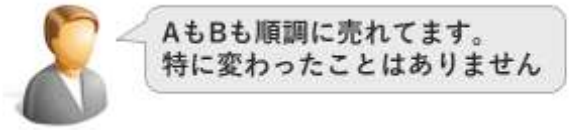
何か問題が起きたときには、必ず「原因を突き止めて手を打つ」ことを心がけてください。「手が汚れるから売れない」という原因がわかれば、その一点の対策をするだけで、少しの努力で売れるようになります。これに対して原因不明なままで対策をすると

- 安売り　→　売れても利益が出ない
- がんばる　→　疲れ果てるだけで売れない
- 低カロリー品を開発する　→　おいしくないから売れない

のように努力が空回りしてしまいます。

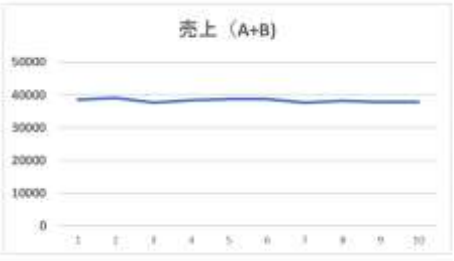
「手に職」系の仕事をする人は、「自分は真面目に丁寧に仕事をするだけ。いい仕事をしていればお客様は評価してくれる」のように考えがちですが、いくら「真面目に丁寧に」仕事をしてもそもそもお客様の喜ぶ方向でなければ「いい仕事」にはならないので注意してください。

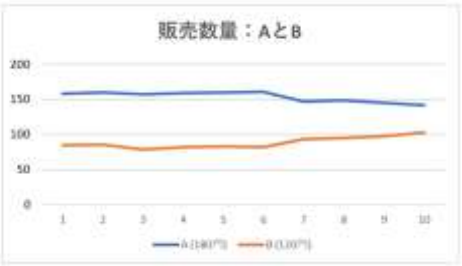
1.18 「数値」を把握していれば早く気づく



というAとB（いずれもポテトチップス）の販売個数と売上総額を見てみると？

	1週	2週	3週	4週	5週	6週	7週	8週	9週	10週
A（180円）	158	160	157	159	160	161	147	149	145	142
B（120円）	85	86	79	82	83	82	93	95	98	103
売上（A+B）	38640	39120	37740	38460	38760	38820	37620	38220	37860	37920

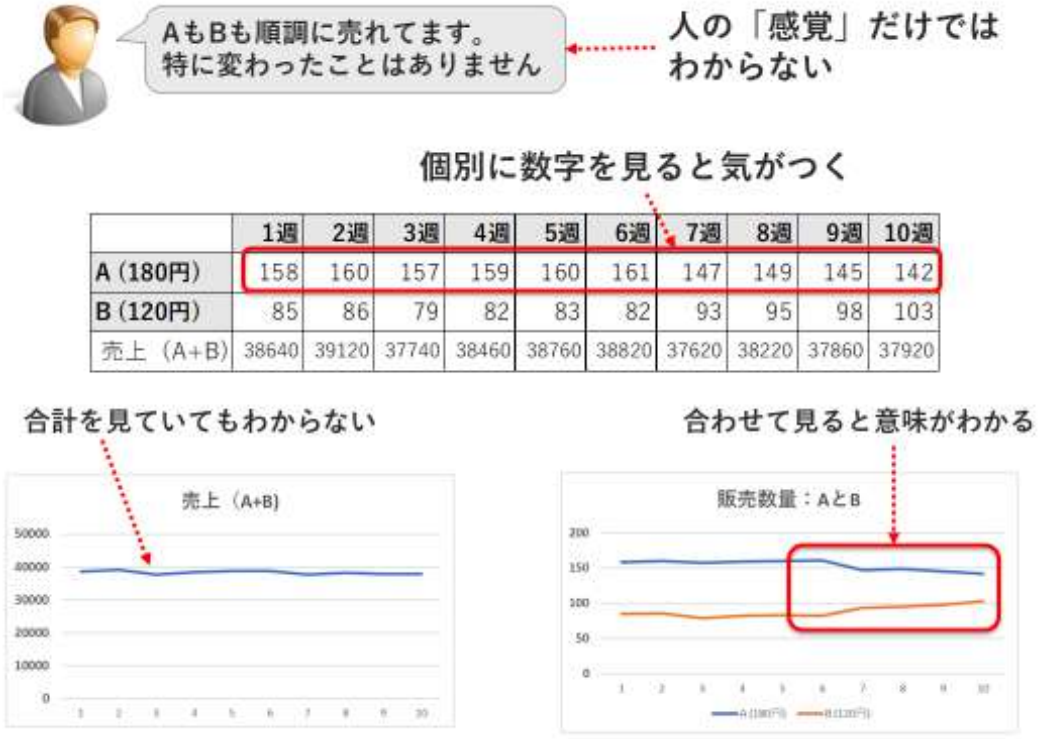




「問題」が起きたら原因の仮説を考えて検証し対策をするわけですが、そのためには「問題」に早く気づかなければなりません。そこで役に立つのが「数値」です。それも、いろいろな角度で数値を見る必要があります。

あるスーパーマーケットでポテトチップスをＡとＢの２種類だけ販売していたとします（実際には何十種類もあるのが普通ですが、話を単純化するため２種類にします）。担当者は「ＡもＢも順調に売れています」と言っていますが、本当でしょうか。

1.19 「感覚」だけではわからない



	1週	2週	3週	4週	5週	6週	7週	8週	9週	10週
A（180円）	158	160	157	159	160	161	147	149	145	142
B（120円）	85	86	79	82	83	82	93	95	98	103
売上（A+B）	38640	39120	37740	38460	38760	38820	37620	38220	37860	37920

確かにAとBの合計額では特に目立った変化はありません。しかし、個別の数字では動きがあります。それも、Aだけあるいは Bだけだとわずかな変化ですが、2つを合わせて見ると「逆方向に動いている・・・ひょっとして低価格品への乗り替えが起きているのか？」という仮説が立てられそうです。

もし、「低価格品への乗り換えが起きつつある」のが正しければ、もう少し別の価格帯、たとえば150円や100円の商品も仕入れてみるべきかもしれません。逆に、一般的に高価格帯のほうが利幅は大きいので、売上が落ちているA商品のほうのテコ入れを図るべきかもしれません。

こうした細かい変化は数字を見ていかないと気がつきません。人間の感覚は意外に当てにならないのです。

1.20 問題を見つけるためには数字を見よ！

店頭でお客様や商品をよく見て接客・販売することも大事ですが
数字を見たほうが気がつきやすい問題もあります

現場で気づいた
ことを数字で確認

どちらも
大事！

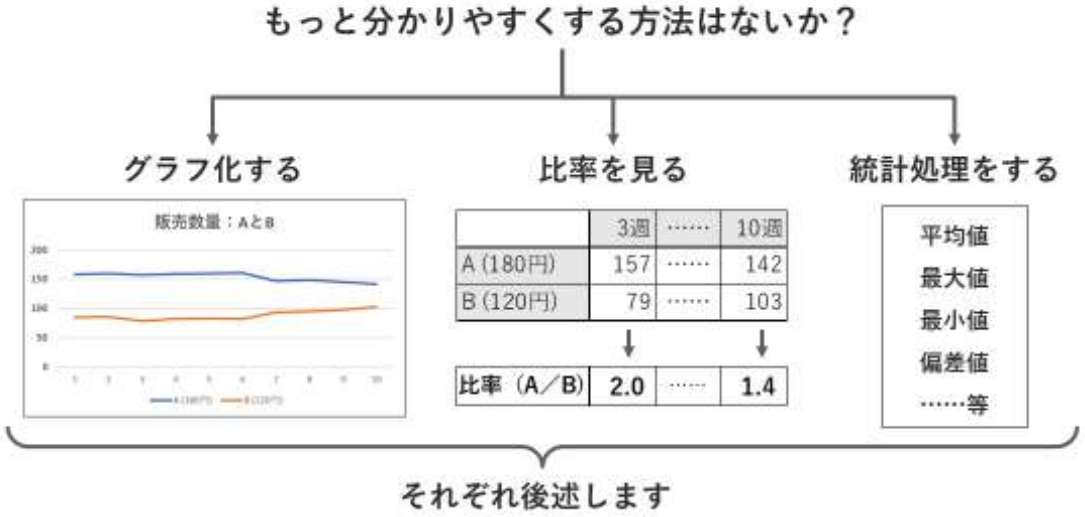
数字で気づいた
ことを現場で確認

	1週	2週	3週	4週	5週	6週	7週	8週	9週	10週
A（180円）	158	160	157	159	160	161	147	149	145	142
B（120円）	85	86	79	82	83	82	93	95	98	103
売上（A+B）	38640	39120	37740	38460	38760	38820	37620	38220	37860	37920

もちろん、数字だけではわからないこともあります。お客様の表情や会話は数字の上には現れませんが、それもマーケティングを考える上では非常に重要な情報です。店頭の現場で気づいたことを数字で確認すること、逆に数字で気づいたことを現場で確認することのどちらも大事です。

1.21 とはいえ、数字を見るのは面倒だ

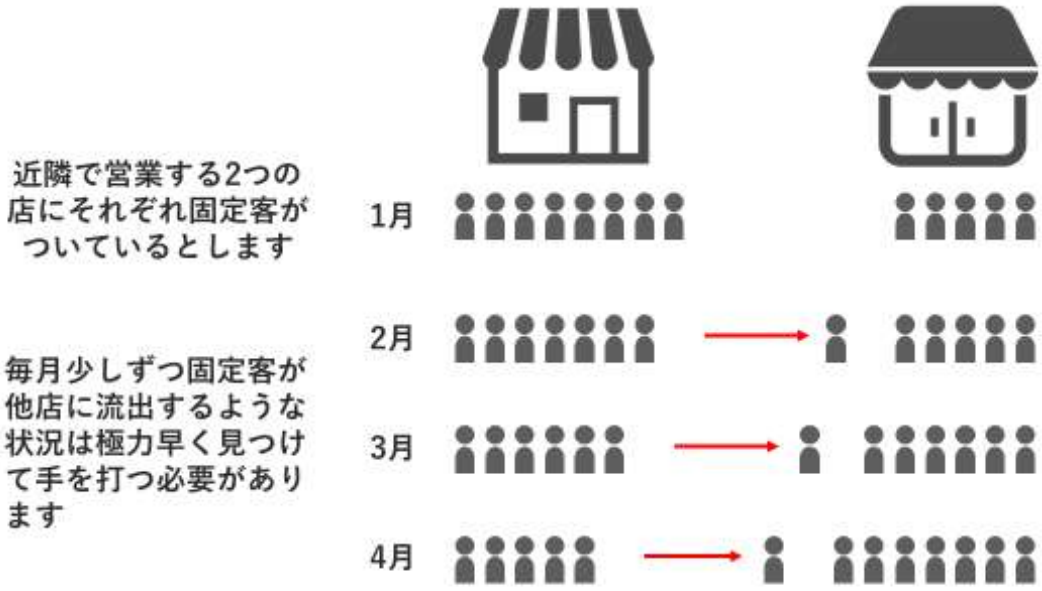
	1週	2週	3週	4週	5週	6週	7週	8週	9週	10週
A（180円）	158	160	157	159	160	161	147	149	145	142
B（120円）	85	86	79	82	83	82	93	95	98	103
売上（A+B）	38640	39120	37740	38460	38760	38820	37620	38220	37860	37920



とはいえ、数字を見るのは面倒なものです。数字だけが羅列された表をぼんやり眺めていても何もわかりませんし、集中して一つ一つ見ようとしてもすぐに目が疲れてきます。何かもっとわかりやすくする方法はないのでしょうか？

そこでよく使われるのが、グラフ化する、比率を見る、統計処理をする、などの方法です。それぞれ本講座の 2 章以降で扱います。

1.22 仕事と数値：固定客の減少



「店に魅力がなくなっている」のを放置すれば、流出はさらに続くでしょう

仕事をする上で数値に気を使わなければいけない場面をいくつか取り上げます。

まずは「固定客の減少」を察知することです。
美容室や飲食店は「いつも利用してくれる固定客」で経営が成り立つ場合が多く、固定客の維持に気を使わなければなりません。近隣で営業する 2 つの店があって、それぞれ固定客がついているとします。なんらかの理由で片方に魅力がなくなると、少しずつ他方に固定客が流出するため、できるだけ早く手を打たなければなりません。

たとえば「店に魅力がなくなった」の原因が「内装が古く、貧乏くさくなった」というものだとすると、「手を打つ」ためには「内装のリフォーム」が必要でかなりのお金がかかります。しかし固定客が減って売上も落ちているときは、お金のかかる投資にはなかなか踏み切れません。そうして対策を先延ばしにしたために、さらに売上が落ちてどうしようもなくなるケースも多いものです。売上があまり落ちていないうちに気づけば思い切った投資もしやすく、傷が浅いうちに回復させられます。

そこで、自店に固定客がどの程度いるのかを常に把握し、減少していないかどうかをチェックする必要があります。そのためによく使われる手法の 1 つに Recency（リーセンシー）分析があります。

1.23 仕事と数値：Recency 分析

（Recency=最近、来ているかどうかを見る分析）

①最終来店日を記録しておく

会員ID	最終来店日	経過日数
1	3月30日	93
2	6月14日	17
3	5月14日	48
4	6月30日	1
5	5月25日	37
6	6月20日	11
7	3月24日	99
8	6月20日	11
9	4月29日	63
10	4月7日	85
11	5月20日	42
12	4月24日	68
13	6月2日	29
14	6月7日	24
15	6月5日	26
16	4月10日	82
17	5月14日	48
18	6月4日	27
19	5月25日	37
20	4月11日	81

②経過日数でソート

会員ID	最終来店日	経過日数
4	6月30日	1
6	6月20日	11
8	6月20日	11
2	6月14日	17
14	6月7日	24
15	6月5日	26
18	6月4日	27
13	6月2日	29
5	5月25日	37
19	5月25日	37
11	5月20日	42
3	5月14日	48
17	5月14日	48
9	4月29日	63
12	4月24日	68
20	4月11日	81
16	4月10日	82
10	4月7日	85
1	3月30日	93
7	3月24日	99

③一定の日数以内の来店者数をカウント（期間集計）

30日以内　8

60日以内　13

Recency とは、「少し前の時間、最近」を表す英語で、Recency 分析は顧客が最近来ているか（購入しているか）どうかを見る分析です。

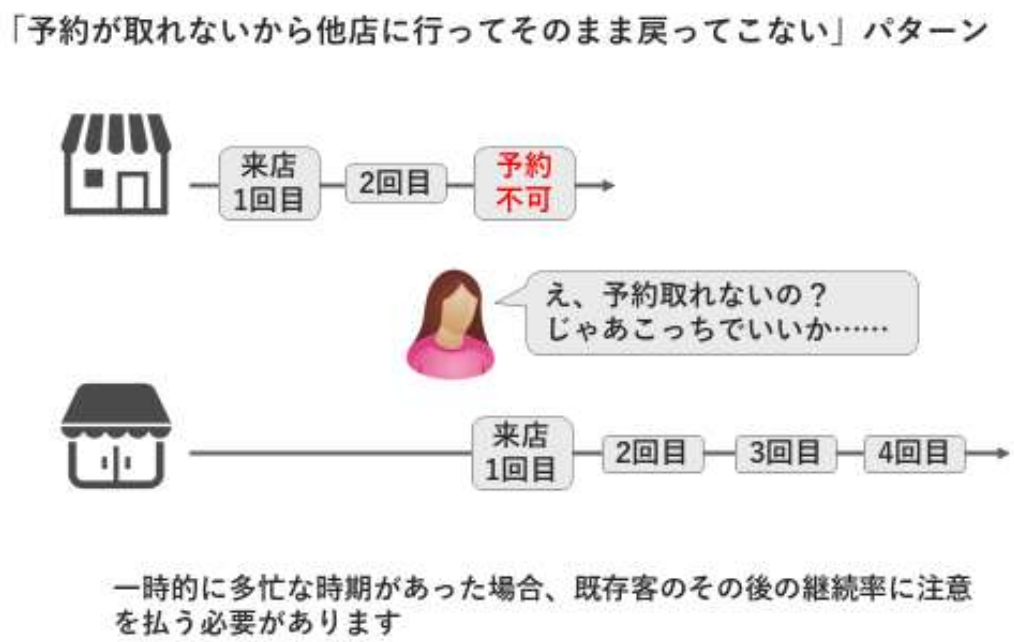
Recency 分析をするには、まず①顧客の最終来店日を記録しておき、来店後の経過日数がわかるようにします。さらに②経過日数が小さい順にソート（並べ替え）したり、③一定の日数以内の来店者数をカウントします。美容室であれば固定客は 2 ヶ月に一度程度は来ることが多いので、この数字が小さくなったときは固定客が減少していると考えられます。

もちろん、集計期間は業種・業態によって違い、たとえば食品スーパーならばより短い期間で区切る必要があります。

小売業では Recency の他に、Frequency（フリークエンシー、来店頻度、頻繁に来ているかどうかを見る）、Monetary（マネタリー、購入金額の大きさを見る）分析を同時に行う場合もあり、総称して RFM 分析と言われています。

最終来店日を記録するのは手間がかかるため、ポイントカードを発行して自動的に記録されるシステムを作るのが一般的で、経過日数によるソートや期間集計には表計算ソフトを使ったり、専用のRFM 分析ツールを使います。

1.24 仕事と数値：多忙の反動

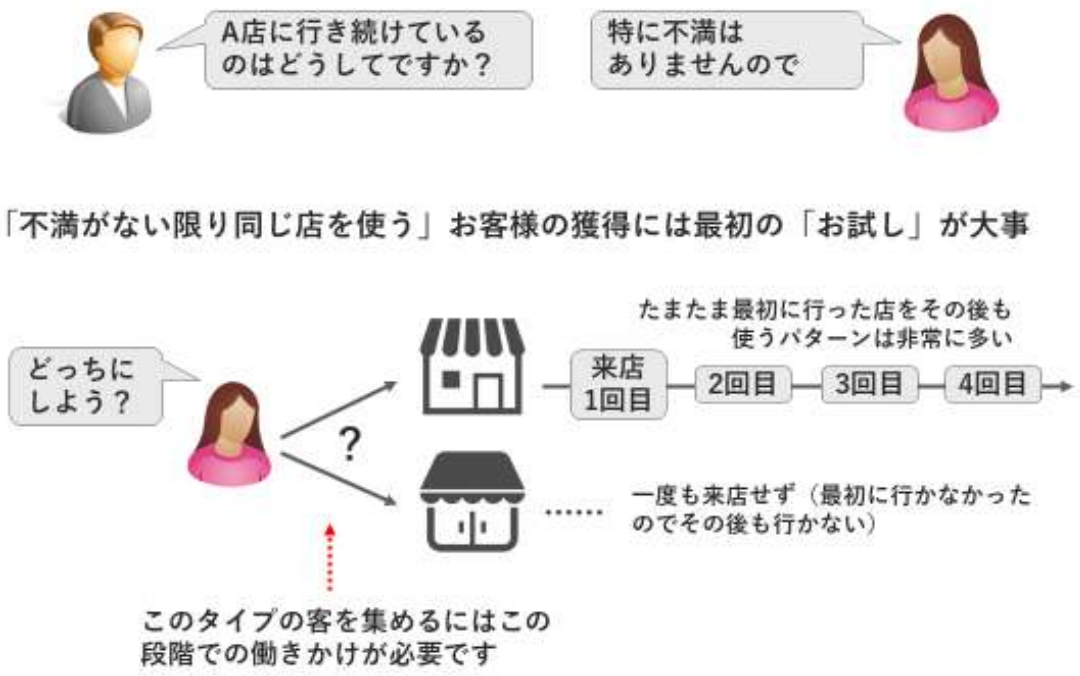


お店そのものに特にこだわりのない客の場合、何もなければそのまま固定客だったものが、何かのきっかけで他店を利用するとそのまま戻ってこない、というケースがしばしばあります。メディアに取り上げられるなど、一時的に多忙な時期（＝繁忙期）があるとこの種の「きっかけ」になりやすいため、その後の状況に注意が必要です。たとえば、

- 繁忙期に予約を受けられなかった客のリストを作っておき、繁忙期を過ぎたらそのリストに特典付お知らせメールを出す
- 常連客が事前に予約を入れやすいように、空き状況をネットで参照・予約可能なシステムを導入する

などの対策が考えられます。

1.25 仕事と数値：初回来店勧誘



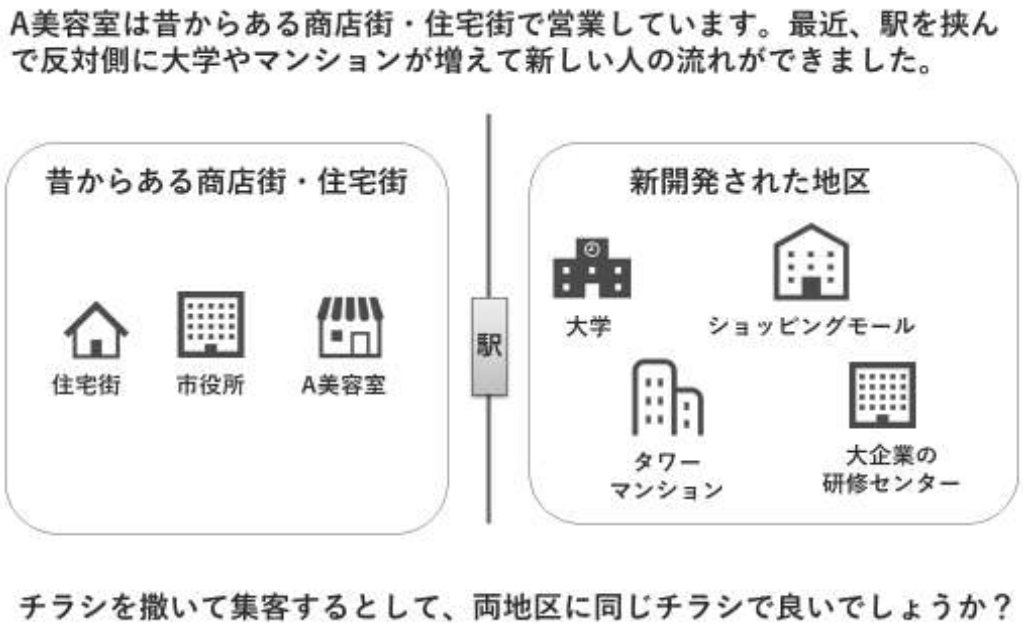
「不満がない限り同じ店を使い続ける」お客様は少なくありません。このタイプの客を集めるには初回来店の獲得が大事で、それを逃がして他店に行かれてからひっくり返すのは容易ではありません。

卒業・進学・就職・人事異動等で引っ越しが増える 3 月～4 月は特に新規客の獲得チャンスであり、次いで 9～10 月も転勤による引っ越しが多いため、この時期に街頭やポスティング、新聞折り込み等の手段でチラシ配布などの広告をするのは効果があります。

ただしこれもベースの「数字」をある程度見込んで考えなければなりません。たとえば近隣に大学の多い地域であれば 4 月に学生の大幅な異動がありますが、学生はほとんど新聞を取らないため新聞折り込みチラシでは効果が望めないのに対して、ワンルームマンションやアパートは学生の比率が高いと考えられるため、ポスティングは効果が期待できます。その場合はチラシの文面自体も学生向けに構成すべきです。一方、地方都市の旧市街地や郡部など、年齢構成が高く人口移動の少ない地域では別な方法を取らなければなりません。

居住者あるいは通勤・通学者の属性は地域によって大差があります。営業地域の特徴を、正確でなくても構わないので、ある程度数字で把握する努力をしておきましょう。

1.26 仕事と数値：エリアマーケティング



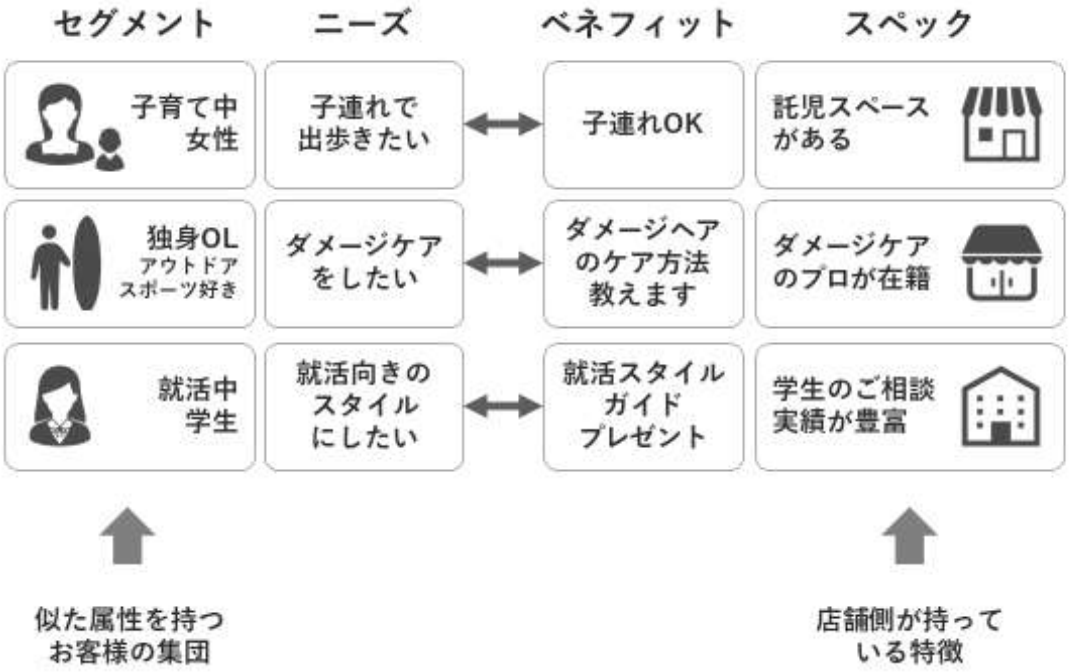
地域の特徴に合わせて営業方針を考えることを一般に「エリアマーケティング」と言います。線路や川、幹線道路などを 1 本またぐとまったく違う性格の地域になるケースも少なくありません。

上図の A 美容室は昔からある商店街・住宅街で営業していますが、最近駅を挟んで反対側の地区が新開発されて、大学やショッピングモール、タワーマンション、研修センターなどが増えてきました。A 美容室は新開発地区からも客を集められそうな立地ですが、その場合、両地区に同じチラシを使うべきでしょうか？

この場合、たとえば次のような「仮説」を立ててチラシを作り、実際に配布して反応を見て修正していくと良いでしょう。

仮説	方針
タワーマンションには 30～40 代の高所得者夫婦が多い	上品な装いを中心にデザインする
大学の新入生が入学後に初めて美容室に来る時期は 4 月下旬に集中する	4 月中旬から GW 期間に学生向けのチラシを配布する
ショッピングモールの客層には小さい子連れの母親が多い	子どもが遊べるスペースを用意しておく

1.27 ニーズとベネフィットのマッチング



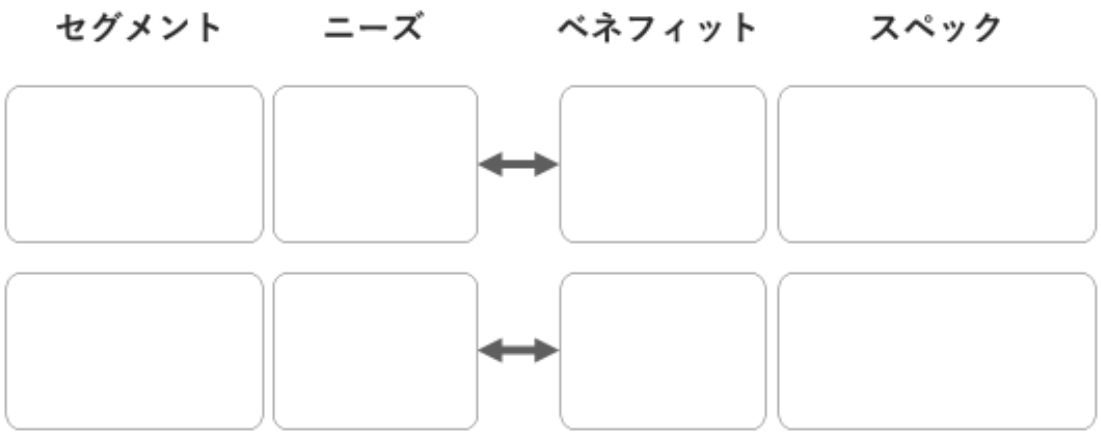
数値を考えながら仕事をするために役立つ考え方に、「ニーズとベネフィットのマッチング」というものがあります。これは細かく言うと上図のように「セグメント、ニーズ、ベネフィット、スペック」の 4 つの視点で考えます。

「セグメント」というのは似た属性を持つお客様の集団で、同じセグメントの客は似たニーズを持つ、と考えます。たとえば小さな子供がいる子育て中の女性は、「子連れで出歩きたい」というニーズがある場合が多いですが、独身 OL や学生には当然そのニーズがありません。一方、「スペック」というのは店舗側が持っている特徴のことで、それを「ニーズ」にマッチする形で表した言葉が「ベネフィット」です。

一般に、広告等で集客を図る場合は、1 つの店が強くアピールできるベネフィットは 1 つだけです。したがって、自店が強みを持てるベネフィットは何なのかをよく考えて、それがどんなセグメントの客を惹きつけるのかを考えて集客手段を考えなければなりません。そうしてたとえば「当店のスペックなら、ダメージヘアを気にする女性の興味を引けるのではないか」と仮説を立てると、そこで初めて「そういう客層が商圏内のどこにどれぐらい住んでいるか」といった「数値」を考えたり調べたりできるようになります。つまり、セグメントからスペックまでのつながりを考えないと、「そもそも何の数値を調べたら良いのか」がわからないのです。

1.28 身近な業界で考えてみよう

飲食店（例：立ち食い蕎麦店とハンバーガーショップ、焼き肉屋とピザレストラン）、美容室（例：自分がよく行く店と行きにくい店）、英会話教室、アパレルショップなど、身近な業界で具体的なお店をイメージし、それぞれがどんなスペック（強み）を持っていてどのようなセグメントを対象に集客しているのかを考えてみましょう



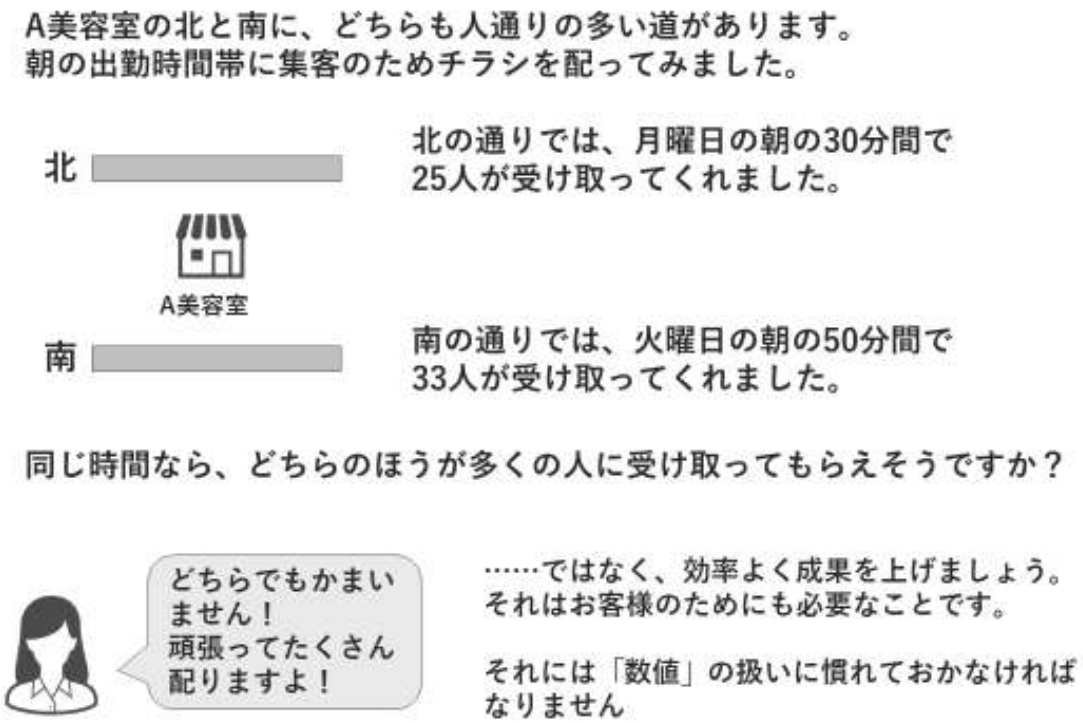
練習のために、自分が利用したりする身近な業界でいくつかのお店を例にとってセグメントからスペックまでどのようになっているかを考えてみましょう。

たとえば同じ喫茶店チェーン業界でもドトールとスターバックスではハッキリと客層が違います。英会話教室には英語を使って仕事を有利にしようというチャレンジングな雰囲気を持つところもあればコンプレックス解消をテーマに集客するところもあります。ファーストフードであれば立ち食い蕎麦店と牛丼やハンバーガーではやはり客層が違います。美容室やラーメン店など、個人営業が中心の業界は特に個々の店による差が出やすいものです。

外装、内装、価格設定やメニュー構成などの具体的な店づくりのすべてをこの「セグメント～スペック」をもとに組み立てるものなので、自分が知っている店を例に考えてみると良いでしょう。

同じ業界でも行きやすい店と行きにくい店がある場合は、何がその違いを生んでいるのかを考えてみるとよいでしょう。

1.29 数値の扱いに慣れておこう



「数値を手掛かりに考える」ためには、あたりまえですが数値の扱いに慣れておく必要があります。
たとえば上図のような状況ではどちらのほうが良さそうでしょうか？　一人の人間が 2 か所に同時にいることはできないので、「チラシ配り」という同じ努力をするなら、より効率の良い場所を選ぶべきです。

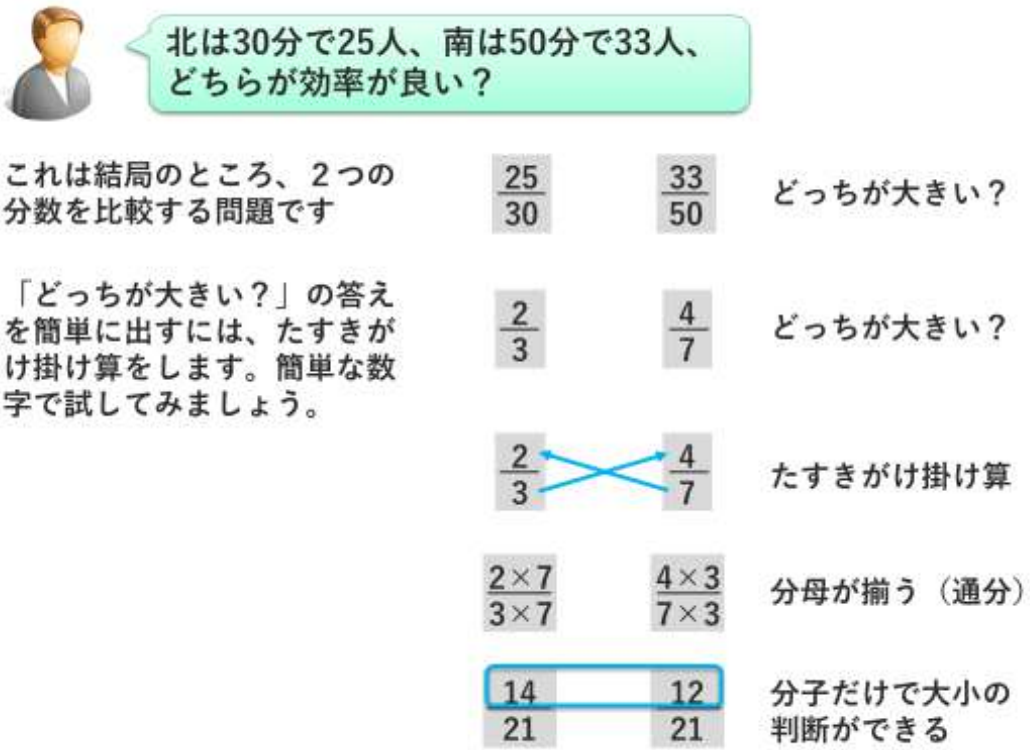
「どちらでもかまわないので頑張ります！」という考え方はよくありません。というのは、「チラシ配り」のような集客のための作業は「客を集めるための手段」であって、お客様に直接役に立つものではないからです。「集客作業」はできるだけ短時間で済むようにして、それで浮いた時間をお客様のために使うべきです。

そこで、北と南で「1 分当たり」何人にチラシを配れたかを計算してみましょう。そうすると、

北：25 人÷30 分 ＝ 0.83 人/分　　南：33 人÷50 分 ＝ 0.66 人/分

となるので、同じ時間を使うなら北の方が効率が良さそうです。こういった計算は電卓でしてもかまわないので、「面倒くさがらずに数値を考える」習慣は持っておいてください。

1.30 分数の比較



「北は 30 分で 25 人、南は 50 分で 33 人、どちらが効率が良い？」のような判断をしなければいけない場合はよくあります。これは結局のところ２つの分数のどちらが大きいかを比較する問題です。電卓をつかってもかまいませんが、ざっくりと大小比較をするだけなら簡単な計算で済むので、できれば暗算で出せるようにしておきましょう。

たとえば　「２／３と４／７、どちらが大きい？」　という問題の場合、分母を同じ数字に揃えるため、それぞれの分子に他方の分母の数字を掛けて「２×７，４×３」を計算します（小学校で習う「通分」です)。すると分母が揃うので分子だけで大小の判断ができます。33×30 と 25×50 なら、0 は無視して

「33×3 はだいたい 100、25×5 は 100 より多い、だから 25×5 のほうが大きい」

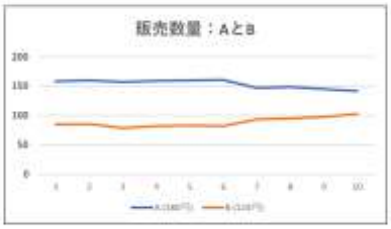
という程度の大小だけ判断できればいいので、正確な数字を出す必要はありません。暗算で大まかに大小を把握できることをめざしましょう。ですが、不安な時は遠慮なく電卓を使ってください。

1.31 数値の並びはグラフにする

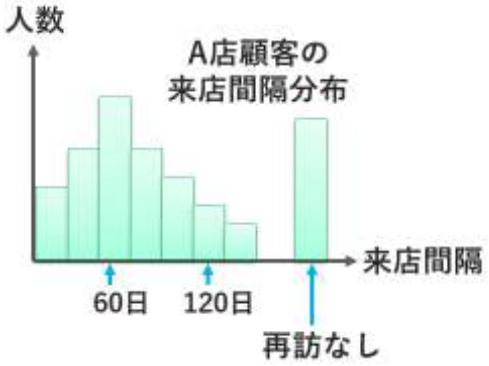
表１

	1週	2週	3週	4週	5週	6週	7週	8週	9週	10週
A (180円)	158	160	157	159	160	161	147	149	145	142
B (120円)	85	86	79	82	83	82	93	95	98	103
売上 (A+B)	38640	39120	37740	38460	38760	38820	37620	38220	37860	37920

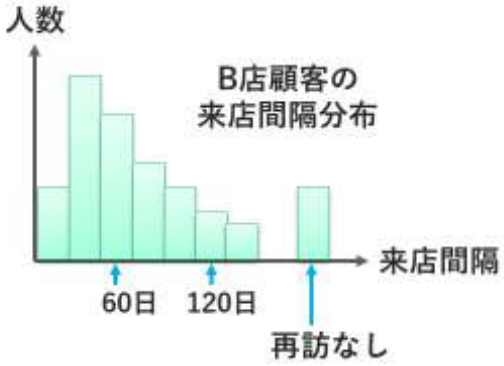
グラフ１



グラフ２



グラフ３

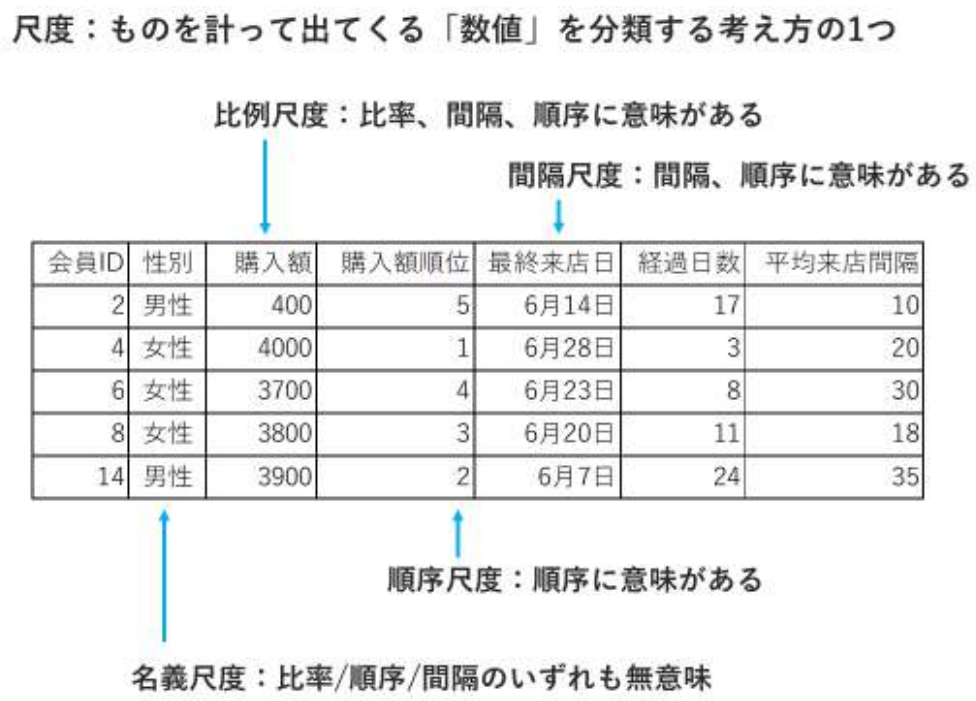


たくさんの数値が並んでいるときに役に立つのがグラフです。表１は２つの商品の売上個数の推移、それをグラフにしたものがグラフ１です。数字ではわかりにくい、ちょっとした変化がグラフにするとわかるようになります。

グラフ２とグラフ３は別な店（美容室）の顧客を来店間隔別にグラフ化したものです。比べてみると、A 店は 60 日周期で来る客がもっとも多いのに対して B 店はそれより短い 40 日周期が多いこと、A 店は一度来たあと再訪しない客が B 店よりも多いことがわかります。

もし A 店の立地や客層が似ているのにこのような違いがあるのならば、A 店の接客や技術、プロモーションに何か問題があるのかもしれません。グラフを作ることによって数値を読まなくてもこうした違いが目に見えるようになります。現在は表計算ソフトウェアで簡単にグラフを作れるので、主要なグラフの作り方・使い方は知っておきましょう。本講座では後の章で扱います。

1.32 ものを「計る」尺度のいろいろ



会員ID	性別	購入額	購入額順位	最終来店日	経過日数	平均来店間隔
2	男性	400	5	6月14日	17	10
4	女性	4000	1	6月28日	3	20
6	女性	3700	4	6月23日	8	30
8	女性	3800	3	6月20日	11	18
14	男性	3900	2	6月7日	24	35

数値を考えて仕事を進めようとすると、数値と言ってもさまざまな種類があることに気づきます。ここで「尺度」という考え方を知っておきましょう。

上図はある店の顧客に関するデータの一部ですが、このリストの中に「比例尺度、間隔尺度、順序尺度、名義尺度」という 4 種類の「尺度」が出てきます。

1.32.1 比例(比率)尺度

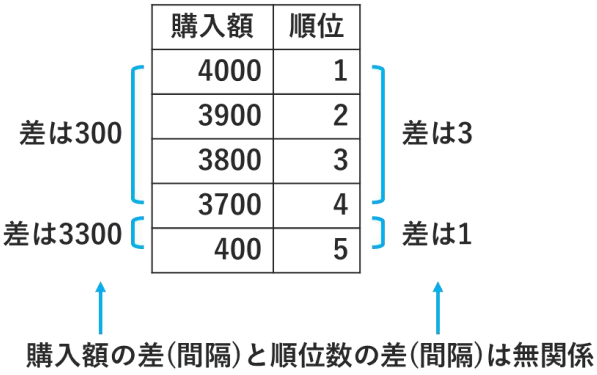
「購入額」についてはたとえば「会員 ID=4 番の女性の購入額 4000 円は、2 番の女性の購入額 400 円の 10 倍である」のように、比率、倍率に意味があります。このような種類の数値が出てくる項目を「比例尺度を持つ項目」と言います。比率尺度と呼ぶ場合もあります。

1.32.2 間隔尺度

「最終来店日」という日付について、たとえば「6 月 1 日は 5 月 1 日の何倍か？」のように比率を考えるのは無意味ですが、「6 月 1 日は 5 月 1 日の何日後か？」のように間隔を考えるのは意味があります。このように「間隔に意味がある」種類の数値が出てくる項目を「間隔尺度を持つ項目」と言います。

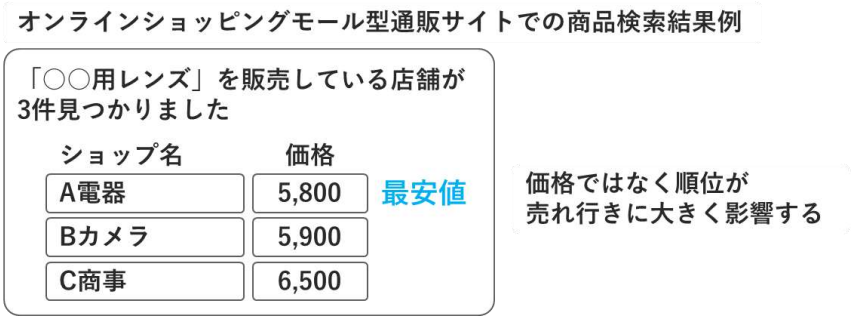
1.32.3　順序尺度

「購入額順位」の項目はたとえば「順序３位と４位」の数字について「４／３」のように比率を計算するのも「４－３」のように間隔を計算するのも意味がなく、単に「４は３よりも後(下位)」という「順序(順位)」だけが意味を持ちます。実際、購入額を見ると 1〜4 位はすべて 3700〜4000 円という狭い範囲に収まっていて、5 位だけが 400 円と極端に少ない数値です。このような場合、順位の数字で差をとっても購入額の差(間隔)とは何の関係もありません。このように順位(順位)だけが意味を持つ尺度を順序尺度と言います。



順序尺度はたとえば陸上競技で「上位入賞者から金・銀・銅メダルを与える」のような場面では重要です。たとえば購入額なら 4000 円と 3900 円で 1 位 2 位と順位をつける意味はほとんどありませんが、100 メートル走なら 9.89 秒と 9.90 秒のようにほとんど同タイムでも順位をつけなければならず、それが金と銀という大きな名誉の差になります。

ビジネスの場面では、オンラインショッピングモール型通販サイトで複数の店舗から横断的に商品をできるような機能があると、「価格」ではなく「順位」が売れ行きに大きく影響する例があります。



順位の影響は商品の種類によって違い、1 円でも安いものが圧倒的に多く売れる場合もあれば、逆に最安値が敬遠される場合、中間が売れる場合もあります。

1.32.4　名義尺度

「性別」の項目は比率/順序/間隔のいずれも無意味です。このような項目を「名義尺度を持つ項目」と言います。性別、住所、職業などは多くの場合名義尺度になります。

1.32.5　尺度の判別が難しい/紛らわしい例

場合によっては尺度の判別が難しいときもあります。

【経過日数や平均来店間隔】
「6 月 20 日」のような日付で書かれた「最終来店日」は間隔尺度ですが、それを「（最終来店後の）経過日数」として「2 日、4 日、10 日」のような日数で表したときは、比率の計算が意味を持ってくる可能性があります。特に、たとえばそのデータを長期間取って「平均来店間隔」を産出した場合は比率に意味が出てきます。美容室や飲食店のような業界では 1 回来店あたりの客単価が大きく変わらないと考えられるため、平均来店間隔が 2 倍に伸びると売上が 1／2 に下がる可能性が高いためです。

【会員番号、ID】
基本的には名義尺度です。しかし、会員登録された順に 1 から増やしていく方法で番号をつけている場合、番号が「会員登録の時期」に連動するため、間隔や順序にある程度意味があるケースがあります。

【学校種別】
小学校・中学校・高校のような学校種別は通常は名義尺度ですが、在籍する子供の年齢が影響する情報については、順序が意味を持つ場合があります。

1.32.6　比例尺度についてはすべての計算が可能

４種類の尺度について、大小、差、比率のそれぞれを計算することに意味があるかどうかをまとめると下記のようになります。比例尺度についてはすべての計算が可能です。

	大小	差	比率	
名義尺度	×	×	×	性別、住所、電話番号など
順序尺度	○	×	×	金・銀・銅メダル、人気ランキングなど
間隔尺度	○	○	×	温度、日付など
比例尺度	○	○	○	金額、重量、速度、長さなど

1.33 用語集

直接費	商品を作ったりお客様にサービスをするために直接使われる費用。材料費、加工やサービスのための人件費などが該当し、売上に応じて上下する傾向がある
間接費	売上が上下してもあまり変わらない傾向のある費用。家賃、水光熱費、機材費などが該当する
損益分岐点	売上がこの額以上になれば利益が出る、という金額のこと
マーケティング・ファネル	潜在顧客がリピート客になるまでのステップをモデル化したもの。認知・興味・行動・満足の段階に分かれる。集客にはそれぞれに合ったマーケティング施策が必要
認知	店や商品の存在に気がつくこと。通勤途上に店がある、チラシがポスティングされていた、SNS で目に留まったなどの段階を表す
興味	「なにこれ？　ちょっといいかも？」のような印象を持つこと。広告を出すときや商品を設計するときは興味を引くように考えなければならない
行動	興味を持った見込み客が実際に行動すること。電話やネットでの予約や注文、来店などの具体的行動がとりやすいようにしなければならない
仮説	売上が落ちた、集客がうまく行かない、などの困ったことが起きたときに、その原因を考えることを「仮説を立てる」などという
オープン・クエスチョン	「○○を買わないのはなぜですか？」のように、選択肢を限定しない質問。店側では想像もつかない答が得られることがある
クローズド・クエスチョン	「手が汚れるのが気になりますか？」のように、選択肢を限定して聞く質問。客が自覚していない答が得られることがある
RFM 分析	顧客が最近来ているかどうか（Recenccy)、どの程度の頻度で来ているか(Frequency)、いくら購入しているか(Monetary)の３点を分析する方法
セグメント	「主婦」「大学生」など、顧客を特定の共通性で分類した集団。その共通性に注目してマーケティングの施策を考えるので、セグメントを意識することが重要
比例尺度	「Ａさんは Ｂさんの２倍購入している」のように、比率に意味がある数字
間隔尺度	「Ａさんは３日、１０日、１７日、２４日と、１週間おきに来店している」のような場合、３と１０の比率（３／１０）には意味がなく、間隔（３－１０）には意味がある。間隔に意味がある数字を間隔尺度という
順序尺度	「当店の売上ランク１位はＡ商品、２位がＢ商品、３位がＣ商品です」のように、順位に意味があり、比率や間隔には意味がない数字
名義尺度	「性別」の項目は比率/順序/間隔のいずれも意味がない。住所、職業、趣味などは多くの場合名義尺度となる

1.34 確認問題

【問 1】「間接費」の説明として正しいものはどれですか？

【選択肢】

（A）売上額の変化に比例してかかる傾向がある費用。例えば飲食店で 100 人分の料理を出すためには 100 人分の食材を仕入れる必要があり、仕入れ費用は売上にほぼ比例する。

（B）売上額が変化してもあまり変わらない傾向のある費用。例えば家賃は客が多くても少なくても変わらない。

（C）来店を促すために店のことをよく知ってもらえるように広告を出したり、来店者への景品をつけるための費用

【問 2】「損益分岐点」の説明として正しいものはどれですか？

【選択肢】

（A）売上額がこの額以上になれば利益が出る、という金額のこと

（B）損失を減らして利益を増やすための重要な経営判断ポイントのこと

（C）売上額が増加に転じたタイミングのこと

【問 3】「RFM 分析」の R,F,M の組み合わせとして正しいものを選びなさい

【選択肢】

（A）R＝直近の来店日、F＝来店頻度、M＝購入金額

（B）R＝購入金額、F＝来店頻度、M＝直近の来店日

（C）R＝来店頻度、F＝購入金額、M＝直近の来店日

【問 4】「顧客セグメント」の説明として正しいものを選びなさい

【選択肢】

（A）例えばスーパーでは野菜、肉、魚、調味料など、販売する商品（食品）の種類ごとに棚を分けていることが多い。このような商品分類のことをいう。

（B）ママチャリは主婦が近所に買い物に行くために使う、ライターは煙草に火をつける

ために使うなど、商品ごとに想定する主な用途がある。その「主な用途」のことをいう。

（C）大学生は就活情報に興味がある、主婦は家事を便利にする生活雑貨に興味があるなど、消費者の属性ごとに主に興味をもつ分野が違ってくる。このようになんらかの共通の属性を持つ消費者の集団のことをいう。

【問 5】次の 4 種類の数字がどの尺度に該当するか、正しい選択肢を選びなさい

【選択肢】

	（A）	（B）	（C）	（D）
顧客ごとの購入金額	順序尺度	比例尺度	比例尺度	順序尺度
ある顧客の来店日付	間隔尺度	名義尺度	間隔尺度	間隔尺度
商品別の売上順位	比例尺度	順序尺度	順序尺度	名義尺度
顧客の性別や職業、趣味など	名義尺度	間隔尺度	名義尺度	比例尺度

■確認問題解答

【問 1 解答】

（B）

（A は直接費、 B は広告宣伝費や販促費）

【問 2 解答】

（B）

【問 3 解答】

（A）

【問 4 解答】

（C）が顧客セグメント。

（（A）は商品カテゴリー、（B）は使用シーンと呼ばれることが多い）

【問 5 解答】

（C）

2 仕事と比較

2.1　比べてみると、違いが分かる

東京のＺ町には有名な古書店街があり、専門的な古書を求める客が全国から集まります。下の図は古書店街マップで、Bは古書店、S,L,Fは大手コンビニエンスストアのチェーン店、それ以外は他の業種の店です。古書店にはそれぞれ専門があって、店によって品揃えがまったく違います。古書店とコンビニの立地を比べてみましょう。

問1）古書店とコンビニの立地はどのように違いますか？

問2）その違いはなぜ生まれたのでしょうか？

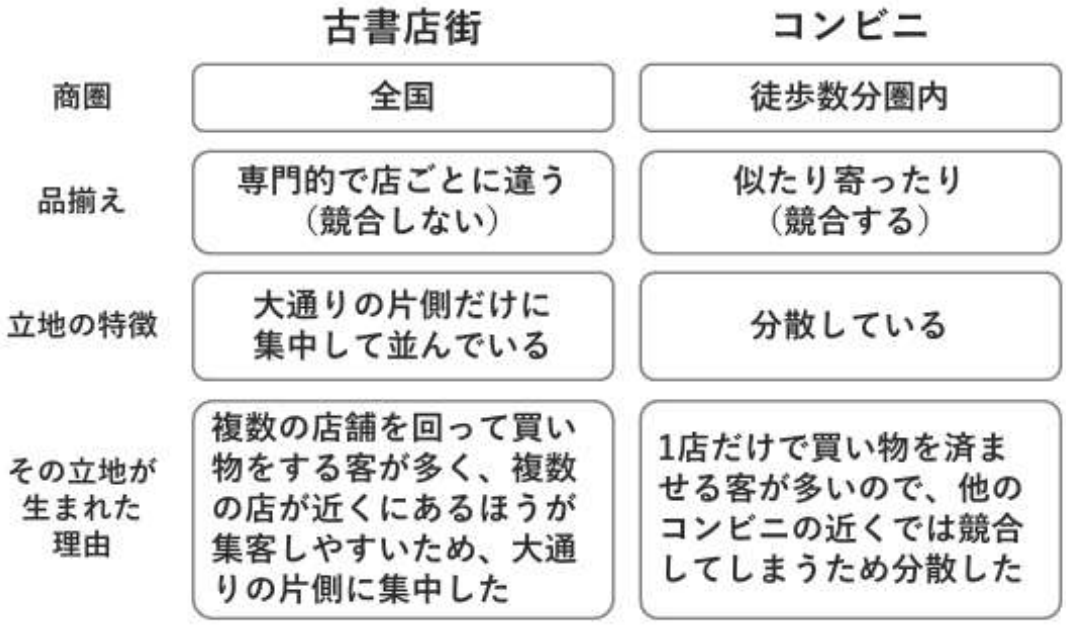
東京のＺ町には有名な古書店街があり、専門的な古書を求める客が全国から集まります。上の図はそんな古書店街の地図で、B は古書店、S,L,F は大手コンビニエンスストアのチェーン店、それ以外は他の業種の店です。古書店にはそれぞれ専門があって、店によって品揃えがまったく違います。古書店とコンビニの立地を比べてみましょう。

問 1）古書店とコンビニの立地はどのように違いますか？

問 2）その違いはなぜ生まれたのでしょうか？

コンビニと古書店では客の行動が違います。専門的な古書を買いに来る客は、古書店を何軒も回って買い物をするのが普通ですが、コンビニでそのような行動を取る客はほとんどいません。それを考えると、このように立地パターンが違う理由がわかります。

2.2　コンビニと古書店街の立地の違い



立地の特徴を一言で言うと、コンビニは分散しているのに対して古書店街は大通りの片側だけに集中して並んでいます。

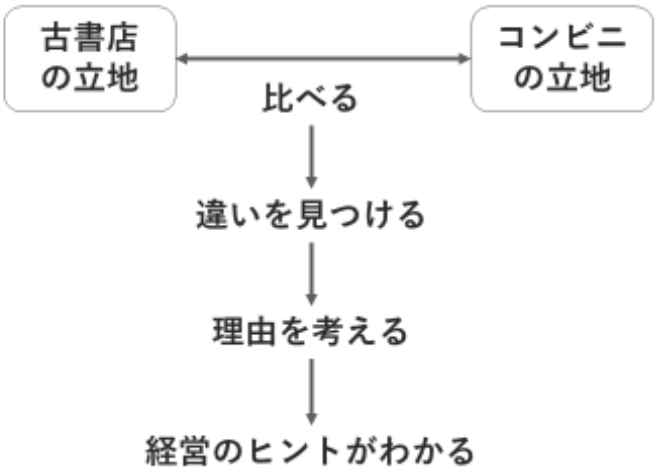
古書店街では複数の店舗を回って買い物をする客が多いため、複数の店が近くにあるほうが客にとって便利です。しかも、大通りはなかなか信号が変わらず横断するのに手間がかかるため、片側に集中するようになったと考えられます。もしこれが狭く横断しやすい通りであれば通りの両側に並んだことでしょう。一方、コンビニの客はほとんどが 1 店だけで買い物を済ませるので、他のコンビニの近くに出店すると競合してしまうため分散していると考えられます。

どんな業種であれ、独立開業して店を出すような場合、どこに出店するかは集客をするための最初の重要な選択になります。一度出店した店は簡単には移動できません。コンビニ的な店なのか専門古書店的な店なのかに応じて、立地の考え方は大きく変わってきます。専門古書のように「複数の店舗を回って買い物をする」タイプの商品を「買い回り品」、コンビニで売っている生活雑貨のように「手近な店で買う」タイプの商品を「最寄り品」と言います。

ちなみに、古い本を売る書店でも専門性のない「古本屋」と呼ばれる業態の場合はコンビニに近い面があります。

2.3　違うものには理由がある

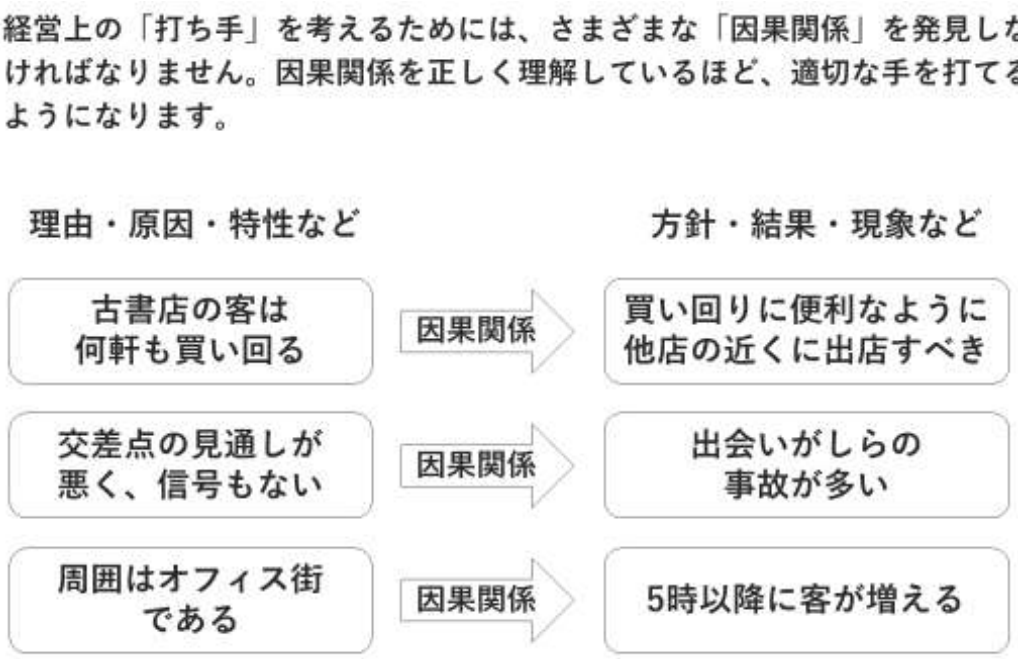
「違う」ものにはたいてい何か理由があります。
理由が分かれば、それを手がかりにして、店を経営する上での「打ち手」のヒントが得られます。



集客や仕事の進め方を改善しようとするとき、「比べてみる」ことは非常に大事です。比べてみると違いが分かり、「なぜだろう？」と考えるきっかけになります。きちんと考えたり調べてみたりすると違う理由が分かり、理由が分かれば店を経営する上での「打ち手」のヒントが得られます。

「比べる対象」は同じ業種でなくても役に立ちます。コンビニと古書店のようにまったく違う業種を比べることでヒントが得られる場合も珍しくありません。

2.4　因果関係を発見しよう



経営上の「打ち手」を考えるためには、さまざまな「因果関係」を発見しなければなりません。因果関係を正しく理解しているほど、適切な手を打てるようになります。

上図は因果関係の例ですが、因果関係の「原因」の側は他にも「理由」や「特性」などの名前で呼ばれることがあり、「結果」の側は「方針」や「現象」などの名前で呼ばれることがあります。呼び方は違っていても広い意味では因果関係であるというケースは非常に多いものです。上図はいずれも単純でわかりやすい例ですが、実際には何と何が因果関係になっているのかがわかりにくい例も非常に多いので、注意深く探っていかなければなりません。

2.5　数値で見ると違いがわかる(1)

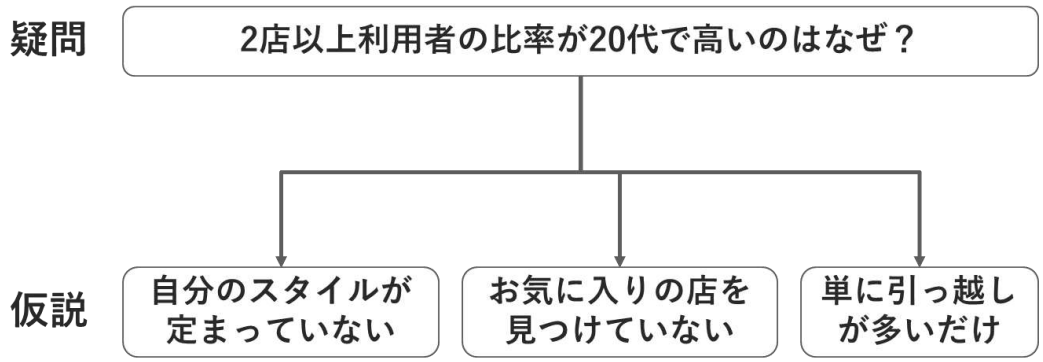
美容室利用状況に関する調査で、下記のような結果が得られたとします。
20代と50代の数値を見るとどんな違いがわかりますか？

年代	1年間の利用回数	1年間に2店以上を利用した人の割合(%)
20代	5.28	36.2
30代	4.95	28.2
40代	6.67	15.6
50代	7.69	13.4

出典：全国理美容製造者協会(2000年)

「比べて」違いを見つけようとするとき、数値を比べると違いがわかる場合が非常によくあります。
上図は美容室の利用状況に関する調査結果です。20 代と 50 代の数値を見るとどんな違いがわかりますか？　そしてなぜその違いが生まれているのでしょうか？

2.6　違う理由について、仮説を考える



数値を見て直接わかることは大まかに下記の 3 つです。

(1) 2 店以上利用者の比率が 20 代で高い
(2)50 代のほうが 1 年間の利用回数が多い

これを経営上の意味に言い換えると

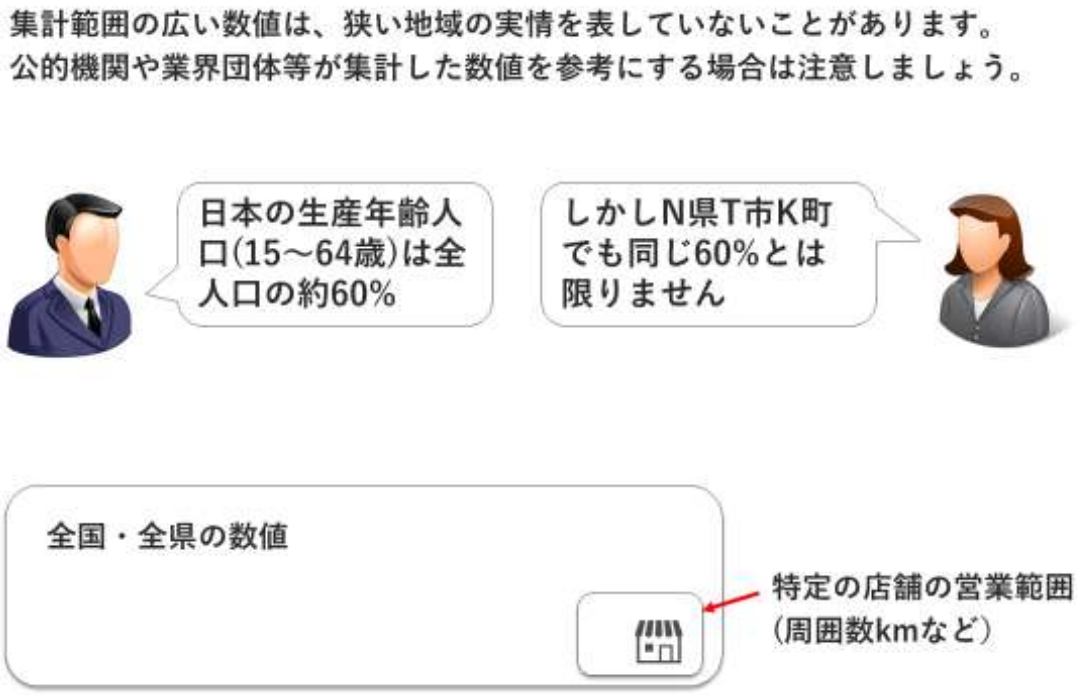
20 代の客は新規獲得しやすいが固定客化は難しい
50 代の客は新規獲得しにくいが一度満足すれば長く使ってくれる可能性が高い

と言えそうです。もし「20 代が 2 店以上利用している比率が高いのはなぜか？」という理由がわかったら、そこに手を打つことで「20 代の新規獲得・固定客化」の双方を改善できるかもしれません。

そこでその理由の仮説を考えると、(1)自分のスタイルが定まっていないのであちこちの店を試している、(2)しかしいまだにお気に入りの店を見つけていない、(3)単に 50 代よりも引っ越しが多いだけ　といったものが考えられます。もし(1)が正しいとすると、「納得の行くスタイルを提案」することができれば、(2)お気に入りの店として固定客化できる可能性があります。一方、(3)が正しい場合は固定客化は難しいものの、引っ越しの時期に合わせた広告宣伝等が有効でしょう。

「数値」というのは小さな違いでも見つけやすく、「違いを見つけて理由を考える」ための最初のきっかけにしやすいものです。経営のさまざまな部分を数値化して比べてみることを心がけましょう。

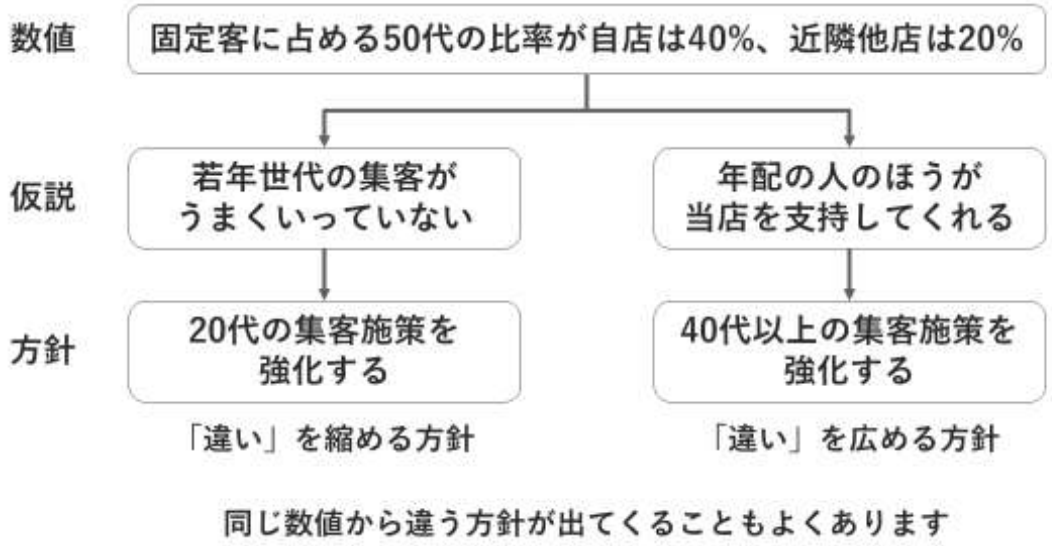
2.7　数値の集計範囲を確認せよ



「数値を見ると違いが分かる」場合がよくあるのは事実ですが、公的機関や業界団体が「集計」した数値は集計範囲が広いため、自店の営業地域の実情と合っていない場合があることには注意してください。「周囲数 km の商圏で営業している店」にとって重要なのはその範囲の実情であって、それは全国や全県のデータとはかけ離れていることがあります。

2.8　「違い」を縮めるか広げるか？

「違い」を縮めるか広げるかに正解はありません。同じ店でも1年経てば答えが違う場合もよくあります。各自の経営環境の中で自店の強みが生きる方針をその都度自分自身で考えてください。



同じ事実（数値）を元に、まったく逆の方針が出てくることは珍しくありません。「違い」を短所と考えて縮める方針を立てることも、長所と考えて広げる方針を立てることもありえます。店主やスタッフの個性と客の相性が大きく影響する業界・業態では「正解は人の数だけある」と言えます。どの方針であれば自分の強みを発揮できるのか？　を自分自身で考えましょう。

【事例】
アメリカの子供向け家具・おもちゃ小売店 Children's Supermart は 1957 年に不調であった家具から撤退し、好調であった子供向けおもちゃに絞り込んだ。同社はその後世界最大のおもちゃ小売チェーン「トイザらス」となった。→得意分野を見つけて「違いを広げた」方針の例。

ガラス製品を製造している A 社の香川工場と浜松工場にて、同じ製品の不良品率が 10 倍近く違っていたため、相互に技術者を派遣して改善に取り組んだ結果、不良品の大幅削減に成功した。→「違いを縮めた」方針の例。

2.9　数値で見ると違いがわかる(2)

年代ではなく満足度別に美容室の利用状況を調べると、下記のような結果が得られたとします。経営上気をつけるべきポイントは何でしょうか？

満足度	月1回以上利用する人の割合(%)	1年間に2店以上を利用した人の割合(%)
非常に満足	26.2	9.8
満足	19.0	19.4
やや満足	11.2	28.8
満足していない	11.5	47.5

出典：全国理美容製造者協会(2000年)

今度は「１年間に２店以上を利用した人の割合」という同じ項目を、年代別ではなく満足度別に見てみましょう。満足度別に見るとさらにハッキリと数値に差が出てきます。

この数値から読み取れることの１つは、当たり前ですが「不満を感じた客は他店に流れる」ということです。経営上は最低限「満足」のレベルをキープすることを目標にしなければならないでしょう。そのためには、満足度を測る何らかの仕組みを持っておかなければなりません。

もうひとつは、「満足度を上げると来店頻度を上げられる可能性がある」ということです。今まで２ヶ月に１回だったお客様が毎月来てくれるようになれば、そのお客様からの売上は２倍か、少なくとも 1.5 倍にはなるでしょう。こちらの観点でもやはり「満足度」は非常に重要な指標であると言えそうです。

2.10 直接比較しにくいときは割合をとる

数値に大きな差があって比較しにくい場合は割合をとってみましょう

A、B両店の固定客の年齢階層分布

年齢階層	20代	30代	40代	50代	合計
A店	23	63	98	87	271
B店	82	221	210	156	669

大きな差のある数値は比較しにくい

年齢階層	20代	30代	40代	50代
A店	8%	23%	36%	32%
B店	12%	33%	31%	23%

「割合」にすると比較しやすい

グラフを作ると傾向がわかる

数値を比べてみようとしても、両者に大きな差があるときは比較しにくいことがあります。上図はA・B両店の固定客数を年齢階層別に調べたものですが、人数で表示した上の表で合計の数字を見ると271対669と2倍以上の差があるため、直接比較しにくい状態です。

そこで下の表のように「割合」にすると似た数字が揃うため、比較しやすくなります。グラフを作るとさらに傾向が分かりやすくなり、B店のほうが顧客の年齢層が若くなっていることが一目で分かります。

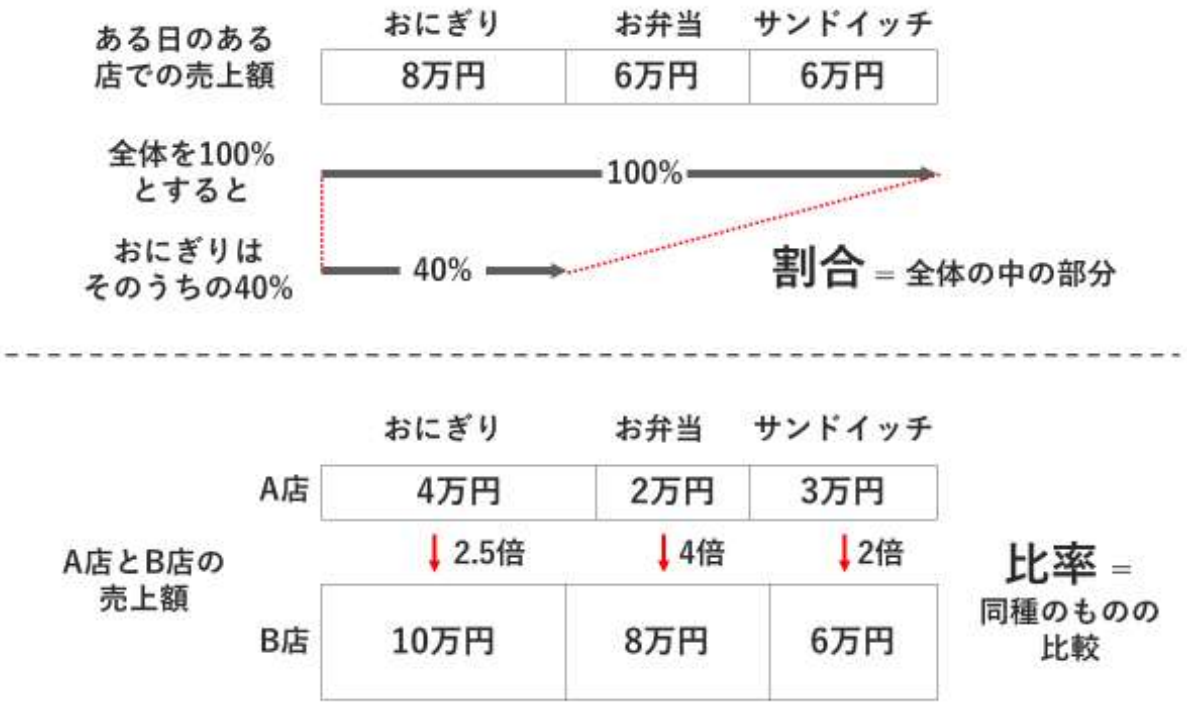
また、割合ではなく比率を出す方法も役に立ちます。

年齢階層	20代	30代	40代	50代	合計
A店	23	63	98	87	271
B店	82	221	210	156	669
B/A比率	3.6	3.5	2.1	1.8	2.5

20～30代の比率が高い

B店の固定客数がA店の何倍あるか、年齢階層別に比率を計算すると、「合計」よりも20～30代で比率の数値が高くなっているため、B店のほうが若い客が多いことがわかります。

2.11 割合と比率



「比率」と「割合」の基本的な意味は同じですが、実用的には少し違う場面で使われます。
「割合」は、全体の中の部分を表す場合によく使われます。

【割合を表す表現の例】

- 当校の学生の 8 割は県内で就職します。
- このビールのアルコール度数は 5.5%です。
- 当美容室のお客様の 3 割がカットのみのご利用です。

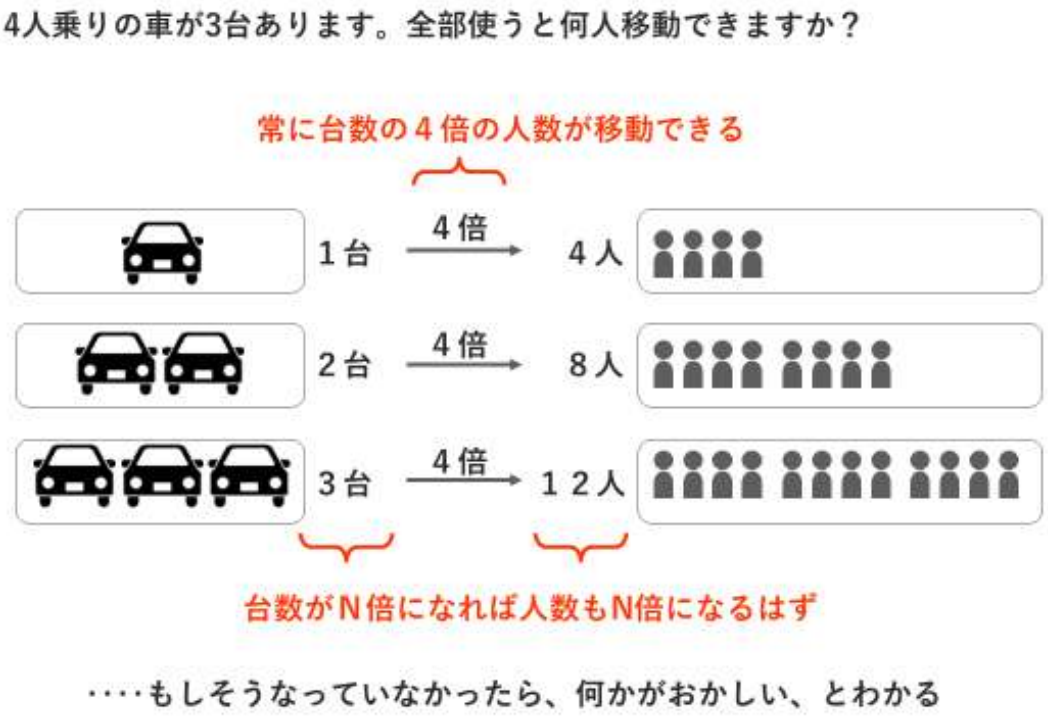
「比率」は同種のものを比較する場合によく使われます。

【比率を表す表現の例】

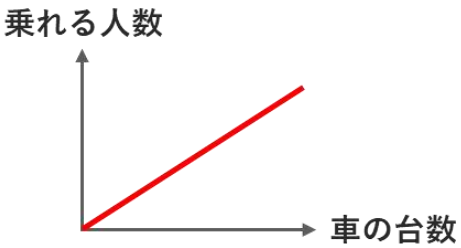
- A 校の野球部員の数は B 校の野球部員の数の約 2 倍でした。
- 減塩醤油の塩分量は一般の醤油の 1 ／ 2 です。
- その日の売上は前日の 3 倍でした。

ただしこの違いは厳密なものではなく、ある程度の目安です。いずれにしても「数値を直接比較しにくいときに、比較しやすくする手段」と言えます。

2.12 比例関係とは？



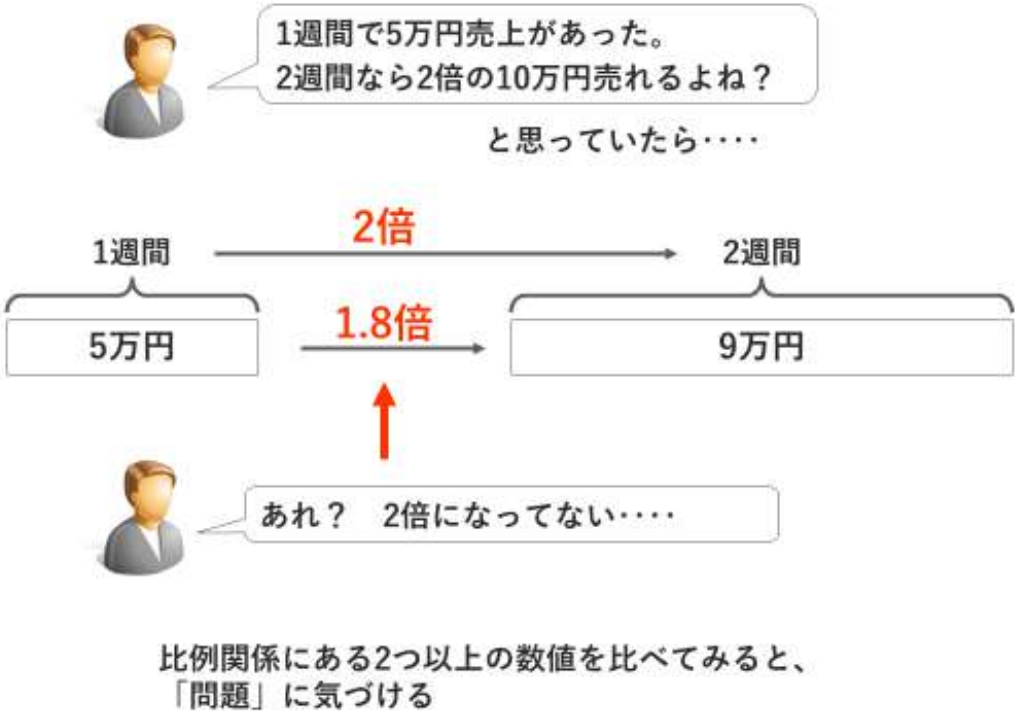
「比率」は比例関係のある数値に使うのが基本です。
比例関係というのは上図のように、一方が N 倍になれば他方も N 倍になるような関係です。
グラフを書くと原点を通る直線になります。



もちろん、「車の台数」というときの「車」は同じ種類のものでなければなりません。たとえば小型乗用車とバスを同じ 1 台と数えることはできません。

本来、同じ種類のものが複数ある場合は比例関係になるはずなので、もしそうなっていなかったら、何かがおかしい、とわかります。

2.13 比率が役に立つとき



たとえば、1 週間で 5 万円の売り上げがあったので、2 週間では 2 倍の 10 万円売れると思っていたら 9 万円にしかならなかったとします。そんなときに「おかしいな？」と考えてよく調べてみると、表の看板に木の枝が伸びて見えにくくなっていることに気がついたりします。

モノやサービスの売れ行きはほんのちょっとしたさまざまな理由で落ちていくものです。
落ちかけたときに素早く原因をつきとめて手を打てば、影響が大きくなる前に食い止められます。
そのためには「問題」に素早く気がつかなければなりません。そこで役に立つのが「数値」です。
といっても、数値が違いすぎると比べにくいのに対して、比率は一定の数値に近くなるので比べやすいのです。

年齢階層	20代	30代	40代	50代	合計
A店	23	63	98	87	271
B店	82	221	210	156	669
B/A比率	3.6	3.5	2.1	1.8	2.5

数値が違いすぎると、
比べにくい

比率は一定の数値に近く
なるので比べやすい

そこで、比例関係にあるはずのものの比率を比べて異常が起きていないかを確かめることが、「問題」を見つけるための基本テクニックの 1 つになっています。

2.14 異常値を発見すれば原因がわかる

ある店のある月の1週目と2週目の売上に大きな差が出たので調べてみると

	1週目	2週目	2週目/1週目比率	
客数	227	218	0.96	客数があまり変わらないのに売上が落ちている
売上	287,082	259,321	0.90	
商品A	135,717	129,013	0.95	主力商品は落ちていない
商品B	126,905	120,840	0.95	
商品C	24,460	9,468	0.39	← オプション商品が異常に落ちている

実際の事業の中での比例関係にはさまざまなものがあります。

【ファーストフード店】ハンバーガーを買う客はドリンクも買うことが多いので、バーガーとドリンクの売上に比例関係がある。
【携帯電話】携帯電話のついでにモバイルバッテリーを買っていくので比例関係がある

これらのさまざまな比例関係のうちの一部に異常値が出る場合があります。
上図はある店のある月の１週目と２週目の売上に大きな差が出た例です。
調べてみると、客数があまり変わらないのに売り上げが落ちています。客数が変わらないということは、集客の問題ではなく店内の問題の可能性が高いでしょう。
そこで商品別の売上を見ると、主力商品であるＡとＢは客数に見合った数字なのにオプション商品のＣが異常に落ちています。そこでＣに関してのみ調べてみると・・・

【ケース 1】商品Ｃは、ＡやＢを買おうとする客に対して「Ｃもあると便利ですよ」とお勧めして売る商品である。ところが２週目の初日にたまたま不機嫌な客がいて「余計なモノを売りつけるな！」と怒って帰ってしまわれたので、お勧めするのが怖くなってしまった・・・

【ケース 2】展示入れ替え中に商品Ｃの宣伝パネルを一時的に撤去していて、そのまま忘れていた。主力商品ではないので気に留めていなかった。

といったことが分かったりします。このようなちょっとしたことでも売上は落ちていきます。それを早く察知して手を打つことが大事です。

2.15 間隔尺度の比に意味はない

気温は間隔尺度ですので、比率（割合）には意味はありません。
「差」には意味があります。

		月曜	火曜	水曜	木曜	金曜
2	かき氷	52	23	2	2	2
3	ソフトクリーム	31	48	4	8	6
4	アイスコーヒー	40	37	5	28	20
5	ホットコーヒー	5	7	47	22	26
6	気温	33	30	22	25	22
7	気温の比率（対月曜）	100%	91%	67%	76%	67%
8	前日との温度差		-3	-8	3	-3

← 比率（割合）は無意味
← 差には意味がある

比率（割合）を計算すると数値を比べやすくなることは多いのですが、数値だからといって常にその方法が使えるわけではありません。たとえば「気温」について比率を計算しても意味はありません。

上図はある店の商品の曜日別の販売個数とその日の最高気温を表にしたものです。気温は月曜日が最高で火曜日に少し下がり、水曜日に大きく下がっています。

かき氷、ソフトクリーム、アイスコーヒーのような商品は暑いときによく売れますが、だからといって気温の比率を計算して「月曜日の 33℃に対して火曜日は 30℃だから 91%だ！」などと計算してもその比率は売れ行きとは無関係です。(参考までに：一般的には気温が 25℃を超えるとアイスクリームの売れ行きが良くなり、30℃を超えるとかき氷のほうがよく売れると言われています)

しかし、「温度差」には意味があります。水曜日は前日よりも 8 度下がっています。このように温度差が大きいと人間は「寒い」と感じやすくなるため、温かいものが売れます。しかし金曜日の気温は同じ 22 度ですが前日とあまり差が無いため水曜日に比べるとアイスコーヒーもよく売れています。ある商品がなぜ売れたのか、売れなかったのかを探る際、上記のように「温度差」との関係を見るとわかることがあります。

このように、「比を計算しても意味が無いが、間隔には意味がある」種類の量を間隔尺度と言います。気温は間隔尺度の代表的な数値です。

2.16 用語集

因果関係	「雨が降ったので道路が濡れた」、「居眠り運転によって交通事故を起こした」のように、何らかの「原因」によって「結果」が生じる関係を言う。「原因」が分かればそれに応じて手を打つことにより問題を改善できる。
予想と行動	「雨が降りそうだったので傘を持ってきました」というのは「予想と行動」である。これは「雨が降る（原因）と身体が濡れる（結果）ことを予想したのでそれを防ぐために傘を持ってきた（行動）」わけで、これも「因果関係」の応用形。このように「因果関係」を元にした考え方であっても、「原因/結果」とは違う用語のほうがピッタリくる場合もよくある。他に「予想」の代わりに「理由」や「特性」、「行動」の代わりに「方針」や「現象」といった名前が使われることがある。
最寄り品	消費者が買う店をあまり選ばず手近な店で買って済ませるタイプの商品。洗剤、ドリンク、タバコ等、多くの店で売っていて品質も変わりない種類の商品であり、コンビニやスーパーで販売される商品の大半が該当する。
買い回り品	消費者がベストな価格や品質を指向していくつもの商店を探し回って買い物をするタイプの商品。高価な耐久消費財、趣味嗜好性の高い商品などが該当する。最寄り品に比べて商圏が広く、専門店で販売される。
商圏	店舗が集客できる地域のこと。コンビニでは半径 500m、スーパーでは 1～数 km と言われている。買い回り品を扱う店は一般に商圏が広い。
新規客	店に初めて来店し買い物をする客のことを言う。新規客の来店を促すためには広告宣伝が必要な場合が多い。
固定客	継続的に自店を利用してくれる客のことを言う。固定客が経ると経営は成り立たないため、満足度の維持向上を図らなければならない。
経営指標	経営状態を表す様々な数値のことを言う。たとえば来店客数、客単価、売上額、利益率、来店頻度など。事業の種類によって重要な指標は異なる。
異常値	経営指標が通常とは大きく異なる値を示すことを言う。たとえば通常は 1 日の来店客が 100 人程度の店にある日 200 人来店したなら異常値と言える。異常値は何らかの問題や商機の存在を示唆することが多い。
比率	経営上のある指標と別な指標を割り算して得る数値。たとえば「客単価」は売上額を来店客数で割ったもの。たとえば牛丼店で一日当たりの来店客数は日によって大きく変わっても、ほとんどの客は牛丼を一杯食べて帰るので客単価は変わらないように、「比率」はあまり変わらない傾向がある。逆に、比率が大きく変わったときは「異常値」であり、何らかの経営上の問題や商機の存在を示唆していることがある。
比例関係	二つの数値の一方が N 倍になれば他方も N 倍になるような関係のことを言う。 比率を取って意味があるのは何らかの比例関係にある数値どうしである。

2.17 確認問題

【問 1】「買い回り品」「最寄り品」の説明として正しいものを 1 つずつ選びなさい

【選択肢】

（A）同じ、または似たような商品を複数の店で売っているので、近所の店で買う傾向が強い商品。この種の商品を主に扱う店は他店の近くに出店すると競合しやすい

（B）野菜、肉、鮮魚、乳製品等、温度管理の必要な食品類のこと

（C）専門性が高い商品であり、個々の店ごとにそれぞれの得意分野の品揃えをする傾向があるため他店と競合しにくい

（D）日常的に使用する商品のうち、衣類、化粧品、文房具など比較的安価で身につけたり持ち歩いたりするもの

【問 2】「因果関係」と言えるものを選びなさい

【選択肢】

（A）雨の日に車を運転すると交通事故を起こしやすい

（B）アイスクリームを販売している A 社では 2020 年に前年比 50%増の広告費を使った。同年の売り上げは前年比 60%増だった。

（C）冬になると火事が多くなる

■確認問題解答

【問 1 解答】

買い回り品：（C）、最寄り品：（A）

【問 2 解答】

（A）

「雨が降る」→「視界が悪くなる。路面が滑る」→「交通事故」　という流れで因果関係がある。

（B）は文中に明記されている情報だけでは因果関係があるとは言えない。「広告を増やしたから売り上げが増えた」という因果関係があると考えがちだが、逆に「売り上げが増え

たから広告費を増やせた」という逆の因果関係もありうるし、まったく関係がない可能性もある。

（C）は因果関係とは言えない。一見すると「冬になる」→「気温が低いので暖房のために火を使う。湿度が低いので火が燃え広がりやすい」→「火事が多くなる」　という、雨と交通事故の因果関係と似た流れのように見えるが、実際は「冬になったことが原因で気温と湿度が下がる」わけではなく、「気温と湿度が下がる季節を冬と呼んでいる」というのが正しい。「視界が悪く路面が滑る状態を雨と呼ぶ」という文は明らかに間違っていることからも、A と C は論理構造が全く違うことがわかる。

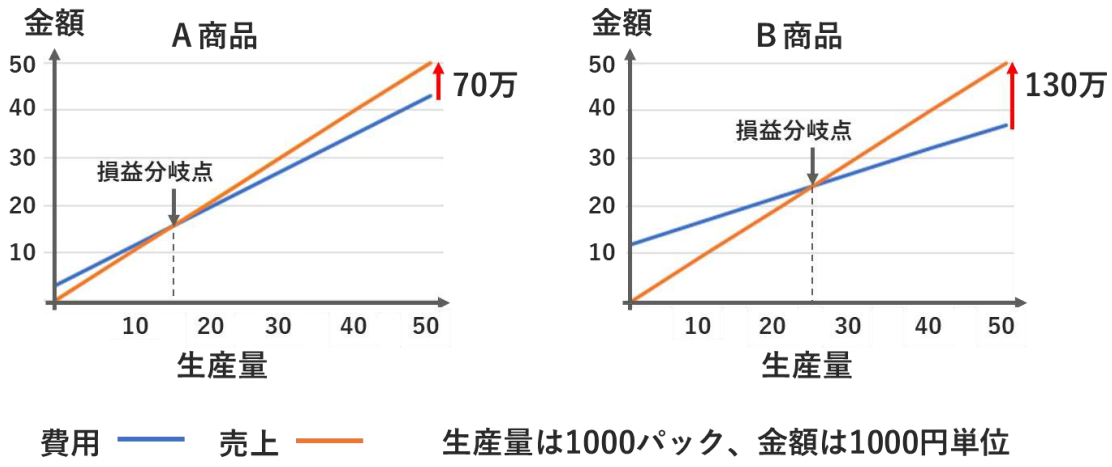
3 仕事と変化

3.1　どちらが儲かりますか？

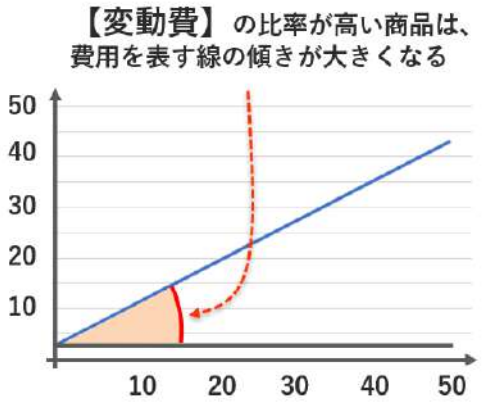
ある製麺所ではそばやうどんのような麺類を作って販売しています。現在の販売数はA・B商品とも月に3万パックで、どちらも1パックにつき10円の利益が出ます。どちらか片方を月5万パックまで増産可能で、増産分をすべて同価格で販売できるとしたら、どちらを増産する方が儲かりますか？

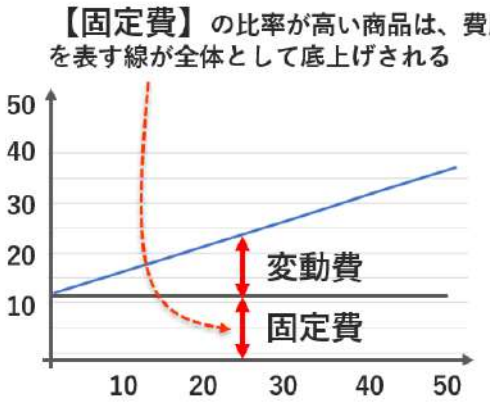
2019年1月の収支計算表		A		B	
販売価格(1パック)		100		100	
変動費	材料費	40	変動費計 80	30	変動費計 50
	労務費	30		10	
	販売費	10		10	
固定費	工場間接費	10	固定費計 10	40	固定費計 40
費用計(1パック当たり)		90		90	
営業利益(1パック当たり)		10		10	

A商品、B商品のどちらも1パックにつき同じ10円ずつ儲かるのであれば、どちらを増産してもかまわない……とはなりません。「1パックにつき10円儲かる」というのは生産量が月3万パックの場合の話であり、生産量が変わると差が出てきます。実際に生産量が0~5万までのA、B商品の費用と売上の関係をグラフにすると下記のようになります。



どちらも単価100円なので、生産量に対する売上のグラフは変わりませんが、費用のグラフは大きく違います。その結果、利益（売上－費用）を計算すると、月産3万パックの場合はどちらも30万円で変わりませんが、5万パックになると、A商品の70万円に対して、B商品は130万円と大きな差がつきます。AとBでは固定費と変動費の構造が違うことに注意しなければなりません。





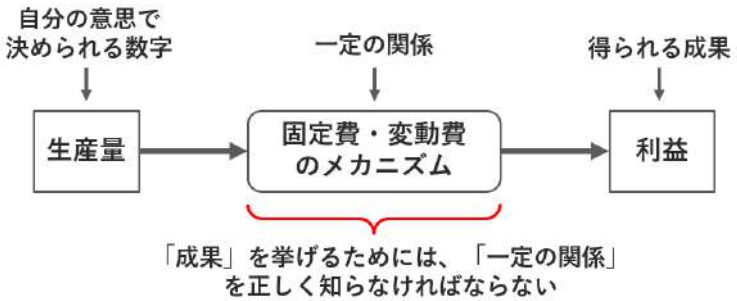
「**固定費**」というのは売上に関係なくかかる費用で、家賃などが代表的です。固定費の比率が高い商品について、生産量と費用のグラフを描くと、費用の線が全体として底上げされたグラフになります。

「**変動費**」というのは売上が増えるとそれに応じて増える費用で、原材料費などが代表的です。変動費の比率が高い商品は、費用の線の傾きが大きくなります。その構造が違うため、A 商品と B 商品には下記のような差が生まれます。

A 商品は固定費が低く変動費が高いので、
　損益分岐点が低い（生産量が少なくても利益が出る）
　しかし、分岐点以上に生産量が増えてもあまり儲からない

B 商品は固定費が高く変動費が低いので
　損益分岐点が高い（たくさん生産しないと利益が出ない）
　しかし、分岐点以上に生産量が増えれば急速に利益が増える

ここで、「自分の意思で決められる数字（生産量）」と「それによって得られる成果（利益）」の間に「一定の関係（固定費・変動費）」があることに注意しましょう。このような場合、高い「成果」を挙げるためには「一定の関係」がどのようなものなのかを正しく知らなければなりません。



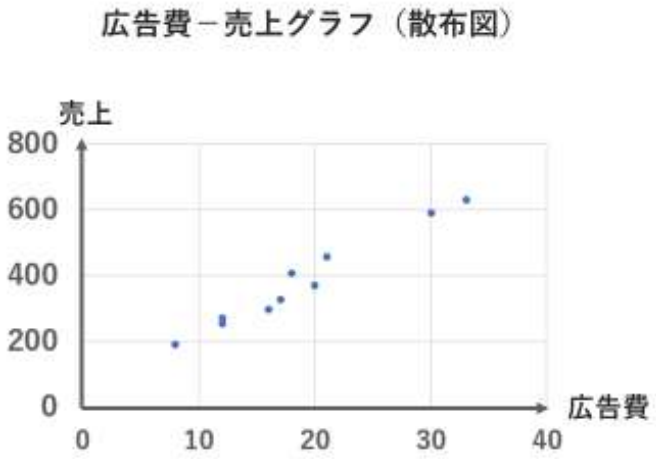
「一定の関係」には固定費・変動費の他にも様々な種類があるため、自分の仕事のどこにどのような「関係」があるかを常に意識して考えていく必要があります。

3.2　変数どうしの関係を数式で表す

ある会社の1号店から10号店までの店舗で使っている広告費と、その店舗での売上額の表とグラフです。グラフを見ると、広告費と売上額の間に、ある「関係」がありそうです。どんな関係でしょうか？

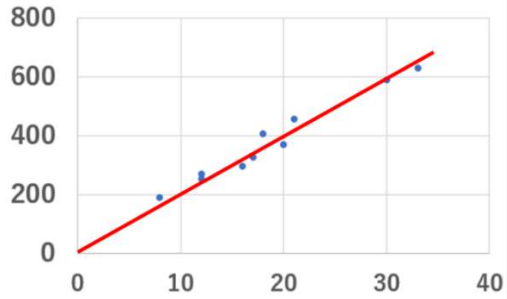
	広告費	売上
1号店	10	255
2号店	12	327
3号店	11	297
4号店	5	190
5号店	24	590
6号店	18	458
7号店	17	371
8号店	33	631
9号店	8	270
10号店	16	409

単位：万円



ある会社には 1 号店から 10 号店までの店舗があり、それぞれ個別に地域のポスティングや折り込みチラシ等の広告費を支出しています。各店の広告費と売上の関係を散布図というグラフにすると、1 本の直線の上に乗りそうです。

実際、グラフの上に直線を 1 本引いてみるとこのように、売上＝広告費×20 という式であらわされる、原点を通る一本のきれいな直線状になりました。この式は何を意味しているのでしょうか？



売上 ＝ 広告費×20

候補 1：広告を出すと、それにかかった費用の 20 倍ぐらいの売上が上がる

式だけを見ると上記候補 1 のようにも見えますが、実際にはほぼあり得ないと言って良いでしょう。
もしこれが正しいなら、例えば今まで広告費に 10 万円かけていた店が 1000 万円かければ売上も

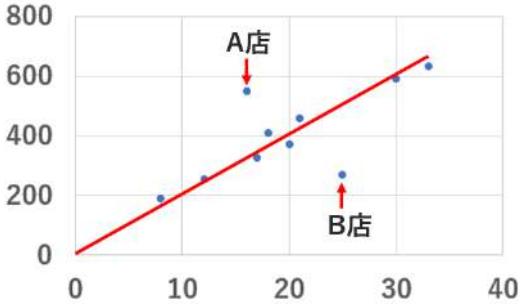
100 倍になる、ということですが、これは率直に言って非現実的です。

候補 2：どの店も、売上額の 20 分の 1（5％）ぐらいの費用で広告予算を出している

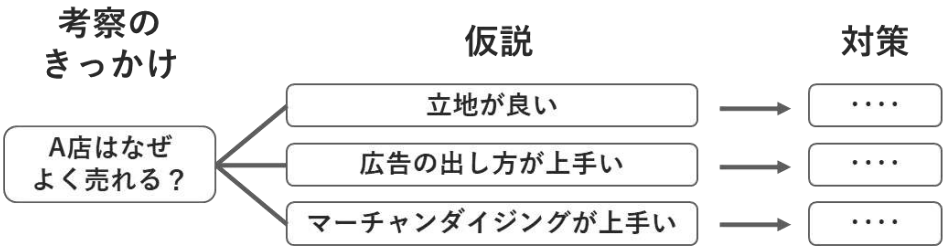
現実にはこの候補 2 が正しい可能性が高いでしょう。実際、このように「売上額の○○%」という基準で広告費を使っている会社は数多くあります。

さて、それではこのような数式がわかったとして、それが何の役に立つのでしょうか？

実際にリアルな会社のデータでこのような「広告費－売上」グラフを作ってみると、上記例のようにきれいな直線にはならない場合が多く、「外れ値」がしばしば出てきます。たとえば下の図では「A店」は掛けた広告費以上によく売れている例、「B 店」はその逆です。



こうした「外れ値」からは、思いがけない改善のヒントが得られることがあります。A 店がなぜよく売れるのか？　と考えた時に、理由がおおよそ 3 つありえたとしましょう。



「立地が良い」のであれば他店では真似できませんが、「広告の出し方が上手い」のであれば、A 店の方法を学ぶことで他店の売上も改善できる可能性があります。あるいは、単に広告にとどまらず、商品の選択や店内の演出などの「マーチャンダイジング」の良さが効いているのかもしれません。その場合は広告の方法よりもマーチャンダイジングを学ぶべきでしょう。自立して事業を営むためには、このように何らかの「考察のきっかけ」から「仮説」を立てて「対策」をとっていくことが重要です。

その出発点となる「考察のきっかけ」は、「外れ値」からわかることが多いのですが、外れが外れであると気がつくためには「普通」がハッキリしていなければなりません。つまり、

普通の店は　広告費＝売上の 5%　になる

ということがわかっているからこそ、A 店や B 店はそれに当てはまらない「外れ値」である、と気がつきます。このように、「普通はこうなるよね」を数式で表せる場合は非常に多く、それができるとグラフを描いて「外れ」を発見しやすいので、「変数間の関係を数式で表す」ことはとても重要です。

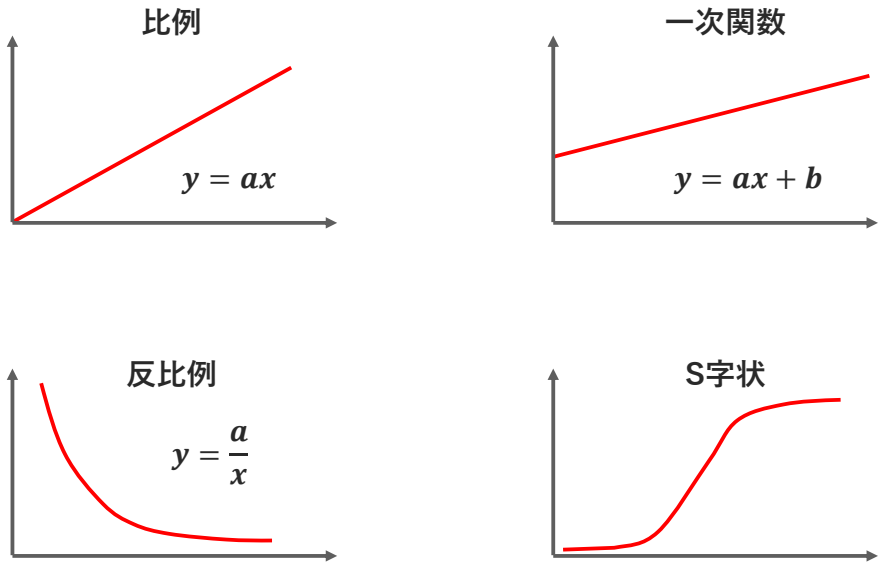
「変数」とは、**「変えられる／変わる」**数値のことを言います。「広告費」はいくら出すかを店長／社長が自分で「変えられる」数値です。「売上」はその結果「変わる」数値です。

仕事の中で出てくる様々な「数値」はすべて「変数」であると言えます。それらの変数の中には、関係を数式で表せるものがあります。可能な限り、変数間の関係を数式で表していくことが重要です。

（注：なお、実際には「数値」ではない名義尺度のような値を持つ「変数」もあります）

3.3　さまざまな関係

2変数間の関係にもさまざまな種類があります。



2 つの変数の関係にはさまざまな種類があります。もちろん現実にはきれいな直線や曲線を描けることは少ないですが、基本的な類型は知っておきましょう。

【比例】

原点を通る直線のグラフで表せる関係です。ある商品を 1 つ作るのに 10 分かかる→1 時間あれば 6 個作れる、など、商品の生産に投じる時間と生産量の関係などによくあります。

【一次関数】

比例と同じく直線ですが、原点を通らないものです。実店舗では広告費をゼロにしても売り上げはゼロにはならない場合が多く、原点を通らない直線のほうが実態に近くなります。

【反比例】

商品ごとに売上額が多い順に並べると、一部の商品がよく売れて残りは急激に落ちていき、大半はゼロに近いという、反比例のグラフに似た形になりがちです。

【S 字状】

広告を出しても最初はほとんど売れず、ある程度回数を重ねると売れ始め、限界に近づくと頭打ちになる場合が多いため、広告投入量と売上の関係は S 字状の曲線になりがちです。

3.4　ローデータと単純集計

「変数」の値を得るためによく使うのが「集計」です。
下記はローデータを単純集計した例です。

ローデータ
（raw＝生）

売上#	ビール	ワイン	チーズ	パン	ソーセージ
1	○	×	×	×	○
2	○	○	×	○	○
3	×	○	○	×	×
4	×	○	○	○	×
5	○	×	×	×	○
6	×	○	○	○	○
7	○	×	×	×	×
8	○	×	×	○	○
9	○	○	○	○	○
10	×	○	×	×	×

集計

単純集計

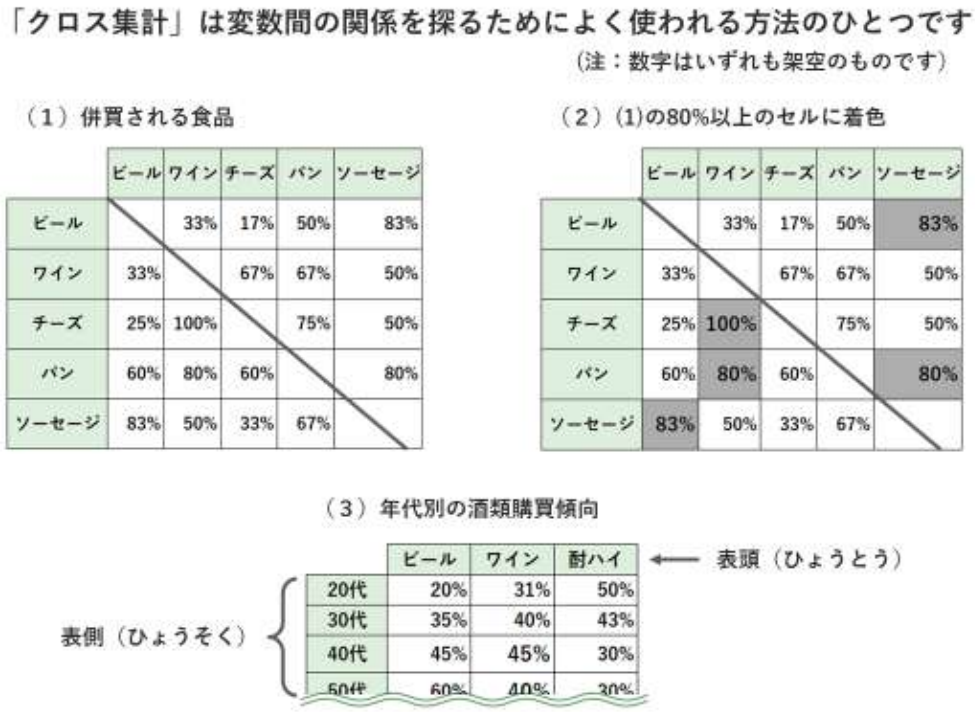
集計	60%	60%	40%	50%	60%

「変数」の中には「集計」しないとわからない数字もあります。たとえば上図はビールやワインなどの購買データですが、10 件の売上のうちで 60%がビールを、60%がワインを、40%がチーズを購入している……といった情報は「集計」しなければわかりません。

「ローデータ」のローは raw（=生）の意味で、1 件ずつの明細データのことです。たとえば売上＃の欄に 1～10 までの数字がありますが、これは 10 人の個客の購入記録を表していて、1 番の客はビールとソーセージを買い、3 番の客はワインとチーズを買った、などの情報がわかります。

「単純集計」は、そのローデータをビール、ワイン、チーズ等の商品カテゴリごとに集計したものです。売れ行きが良いか悪いかを判断するには、ローデータではなくこのような集計データが必要です。

3.5　クロス集計



	ビール	ワイン	チーズ	パン	ソーセージ
ビール		33%	17%	50%	83%
ワイン	33%		67%	67%	50%
チーズ	25%	100%		75%	50%
パン	60%	80%	60%		80%
ソーセージ	83%	50%	33%	67%	

	ビール	ワイン	チーズ	パン	ソーセージ
ビール		33%	17%	50%	83%
ワイン	33%		67%	67%	50%
チーズ	25%	100%		75%	50%
パン	60%	80%	60%		80%
ソーセージ	83%	50%	33%	67%	

	ビール	ワイン	酎ハイ
20代	20%	31%	50%
30代	35%	40%	43%
40代	45%	45%	30%
50代	60%	40%	30%

集計方法には単純集計の他にクロス集計という方法もあります。

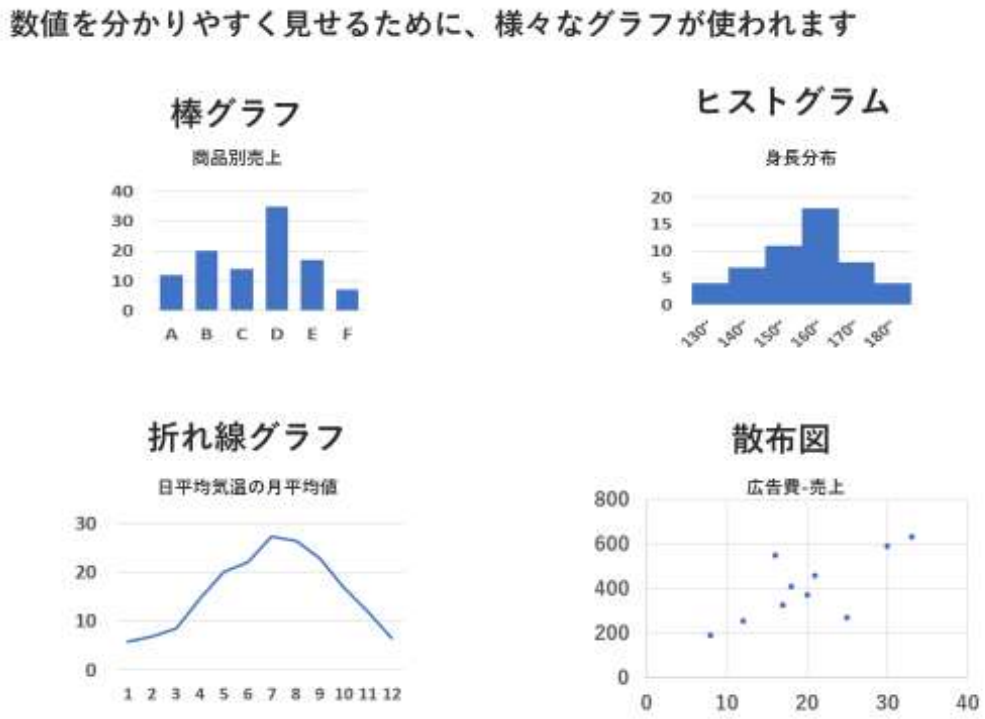
（1）は併買される（同時に買われる）商品を知るためのクロス集計の例で、縦軸横軸ともに同じ商品カテゴリーが並んでいます。このような集計により、「ビールを買った客の 17%がチーズを、83%がソーセージを同時に買っている」といったことがわかります。これがわかると「ビールの売り場でソーセージのお勧め品を展示しておくとよく売れるかもしれない」といったアイデアが得られます。このように「併買される商品の傾向を知る」ことは小売業のマーチャンダイジングにおける重要な課題です。

（2）は（1）と同じ集計表ですが、80%を超える数字のセルを着色して目立たせたものです。クロス集計は高い数字や低い数字に意味がある場合が多いので、単に数字を出すだけでなく、このような工夫で注目点を目立たせるとわかりやすくなります。

実用的に最もよく使うのは、(3）のように縦と横で別な分類軸を使って行うクロス集計です。年代別、性別、居住地別などで商品の好みの傾向が分かれることは非常に多いため、この方法がよく使われます。

なお、クロス集計の横軸のことを「表頭（ひょうとう)」、縦軸のことを「表側（ひょうそく)」と呼びます。

3.6　グラフによる表現



数値を分かりやすく見せるためには様々なグラフが使われます。

【棒グラフ】

商品別売上など、連続性のないデータを表すときによく使われます。たとえばA、B、C がそれぞれリンゴ、牛乳、ウィスキーのような商品だとすると、それらの売上には通常連続性がないため、棒グラフの「棒」どうしはスペースを空けて分離して書きます。順序を変えて表示するのも自由にできるのが普通です。

【ヒストグラム】

1 クラスの生徒の身長分布、日本人の年齢構成など、連続性のあるデータを人為的にいくつかの「階級」に区切って集計表示するために使われます。たとえば身長を 150cm~、160cm~のように 10cm 単位の階級に区切って集計することはよくありますが、このような区切りは集計の都合上設定しているだけで、本来何の意味もありません。そこで、本来は連続しているデータである、ということを示すために棒グラフの「棒」どうしの間にスペースを空けずに書いたものがヒストグラムです。棒グラフと違って、ヒストグラムの階級は順序を変えて表示することはできません。

【折れ線グラフ】

連続性があり、長期間継続しているデータを表すためによく使われます。たとえば「気温」は前月

と当月で極端に違うことはなく、連続性があるため折れ線グラフで表現するのに向いています。

【散布図】

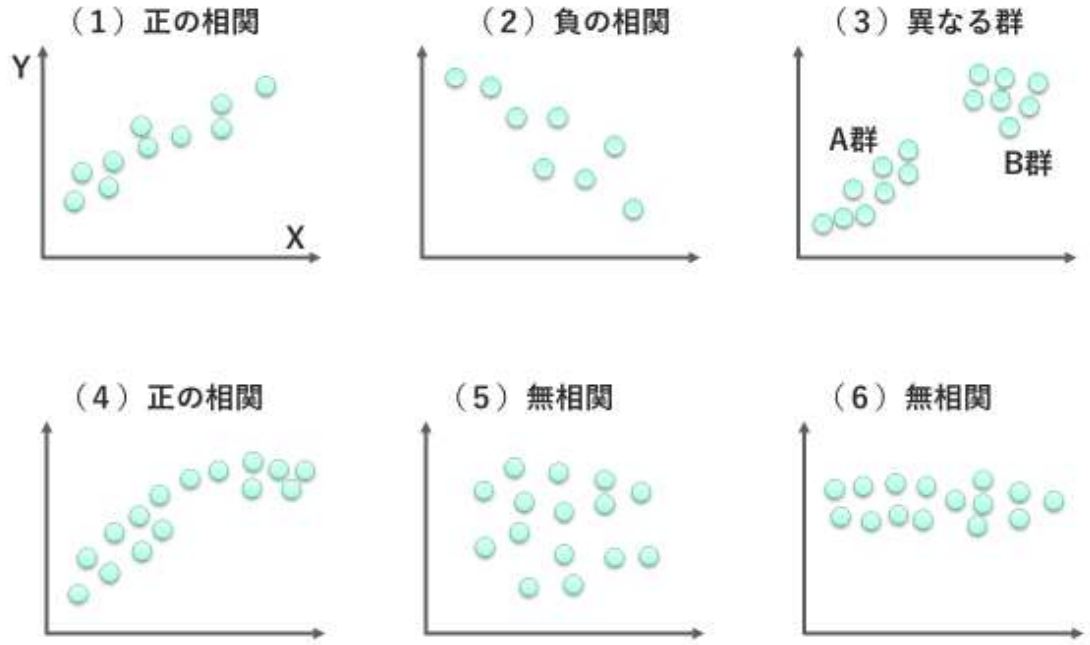
関連性があるかないかわからない複数のデータをタテ軸と横軸にとって作る、「バラバラに散らばった点」のようなグラフが散布図です。データに関連があるかないかを検証するためによく使われます。

例１：「数学が得意な人は理科も強いだろう」という仮説を検証するため、数学と理科の得点を縦・横の軸にして散布図を描く

例２：「高所得者はワインを好むだろう」という仮説を検証するため、「年収」と「年間のワイン消費量」を縦・横の軸にして散布図を描く

散布図を描いたときに、なんらかの直線が引けそうな分布をしていれば、関連がある可能性が高いと言えます。

3.7　散布図からわかる関係



散布図のパターンで関係の有無を判断する例です。

（1）は右上がりの直線状に分布しています。これは横軸Xが増えるほど縦軸Yも増える関係であり、このような関係を「XとYの間に正の相関がある」と言います。「年齢が高いほど身長が高くなる」、「年収が高いほど外食日数が増える」などがこの例です。

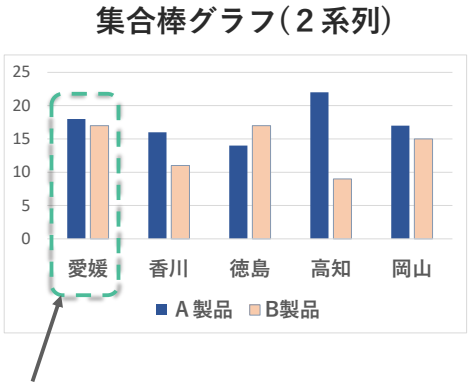
（2）は（1）とは逆に右下がりの直線状に分布しています。このような関係を負の相関と言います。

（3）ではA群とB群の間にギャップが見えます。このような場合、おそらくA群とB群は異なる性質を持った集団です。たとえば日本人と外国人に同じアンケートをとった場合など、このような形になる場合があります。（3）の例ではA・B両群の間にギャップがあるため分かりやすいですが、調査の種類によってはギャップがはっきりしないことがあります。A群だけを調べれば「正の相関」がありそうですが、AとBのギャップがはっきりせず両者を分離できないと、それを発見するのが難しくなります。

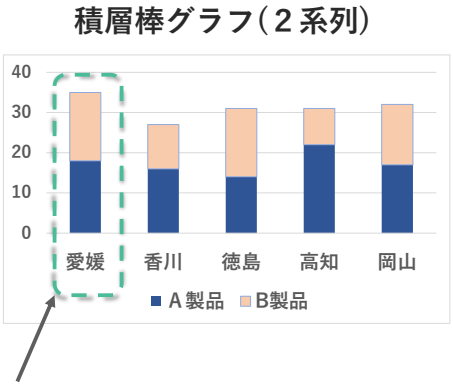
（4）の左半分はXとYに正の相関が見えますが、右半分はXが増えてもYは変わらなくなっています。たとえば携帯電話は一人1台以上持つことはあまりないため、全員が持つようになるとそれ以上は売れません。普及が頂点に達して需要が飽和したような商品でよくあるパターンです。

XとYに関係がない場合は、（5）のように線ではなく「面」として広がった形になるか、（6）のように傾きのない直線状になります。「関係がない」ことを「XとYの間には相関がない」または「無相関である」と言います。

3.8　さまざまな棒グラフ



どっちが多い？　と比べやすい

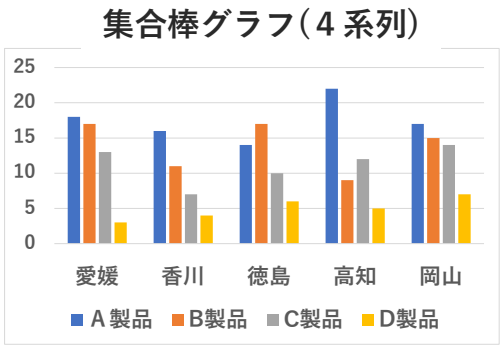


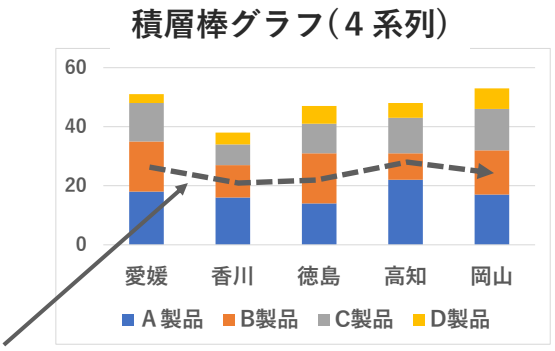
まとめて一番多いのはどこ？　を見つけやすい

棒グラフのちょっとしたバリエーションとして、「集合棒グラフ」と「積層棒グラフ」というタイプがあります。たとえばＡとＢの２製品のデータがあるとき、上図左のようにそれを横並びにして描くのが集合棒グラフ、縦に積み上げて書くのが積層棒グラフです。いずれもデータは同じで、ＡとＢの県別の売上を表したグラフです。

集合棒グラフは「愛媛県ではＡとＢどっちの売上が多い？」という比較をしやすいのに対して、同じデータを積層棒グラフにするとそれができません。一方、積層棒グラフは「ＡとＢを合わせた売上が一番多い県はどこ？」という質問にはすぐ答えられますが、集合棒グラフではそれはわかりません。

また、「１つの系列の推移を追いかける」のは積層棒グラフのほうが簡単です。

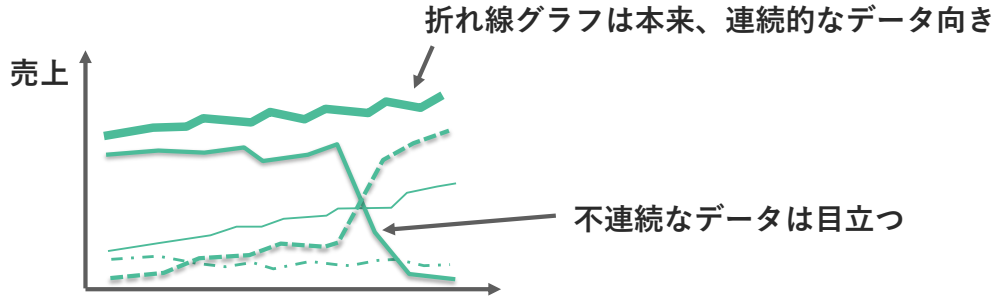




１つの系列の推移を追いやすい

データの系列数が多いときはこれが問題になることがあるため、どんなデータをどのように見せたいかをよく考えてグラフの形式を選びましょう。

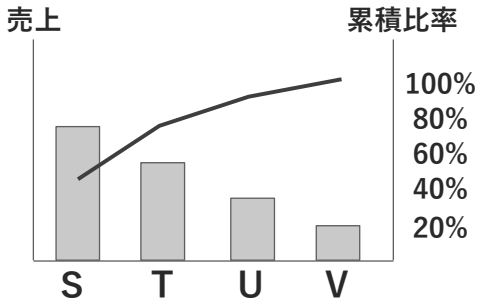
3.9　さまざまな折れ線グラフ



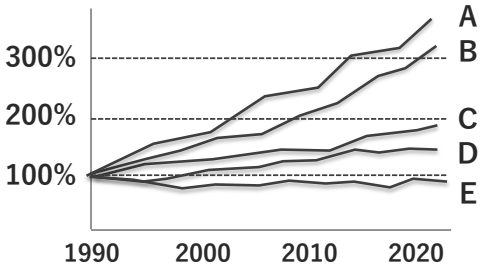
折れ線グラフは最もよく使われる種類のグラフで、時間の経過にしたがって連続的に変化するようなデータを表すのに向いています。「売上高」は通常急激に変化することはないためこの条件に当てはまります。逆に、折れ線グラフの中に不連続な（急激な変化を示す）データがあると目立ちます。上図は５つのデータ系列を折れ線グラフにしたものですが、ちょうど同じタイミングで急上昇・急降下している２本の線が目を引きます。もし同じデータを集合棒グラフで書いていたらこれほどには目立ちません。

その他、棒グラフとの組み合わせで「パレート図」という形式が経営分析をするためによく使われます。たとえば下記の例は商品 S、T、U、V の売上を大きな順から棒グラフにした上で、それを順に足していった「累積比率」を折れ線グラフで書いたものです。多くの商品を取り扱う事業であっても実際の売上額としては数品目だけで売上の大半を占めているケースが多く、それを重点的に管理するためにこのパレート図が使われます。

パレート図

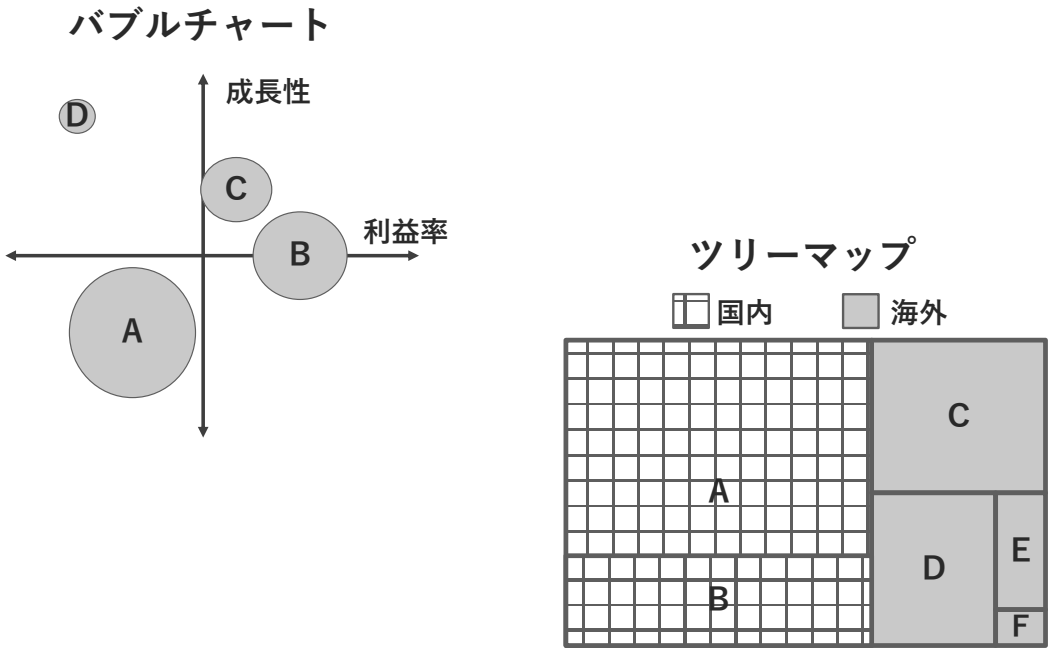


ファンチャート



「ファンチャート」は基準年（上図では 1990 年）の数値を 100%として、それ以後の変化を基準年に対して何%かという比率で表す折れ線グラフです。複数の会社や事業の売上推移をこの形式で表して成長性を比較する用途でよく使われます。「ファン」とは扇のことで、基準年を中心に扇を広げたような形になることからファンチャートと呼ばれます。

3.10 バブルチャート、ツリーマップ



「バブルチャート」は散布図の一種です。通常の散布図では「点」を使うところをバブルチャートでは点の代わりに大小の円（バブル）を描き、その面積で第３の数値を表します。上図の例ではある会社が持っている４つの事業 A～D についてタテヨコの軸で成長性と利益を、円の大きさで売上高を表しています。このバブルチャートを見ると次のようなことがわかります。

A 事業：売上額は最大だが縮小傾向。しかも儲かっていないので会社のお荷物になりかねない。
B 事業：成長してはいないが安定的に利益を稼ぎ出している
C 事業：少ないながら利益を挙げており今後も成長が見込める
D 事業：大きな成長が期待できるため思い切った投資を考える分野

バブルチャートは３つの数値を１度に把握して判断したい場合によく使われます。

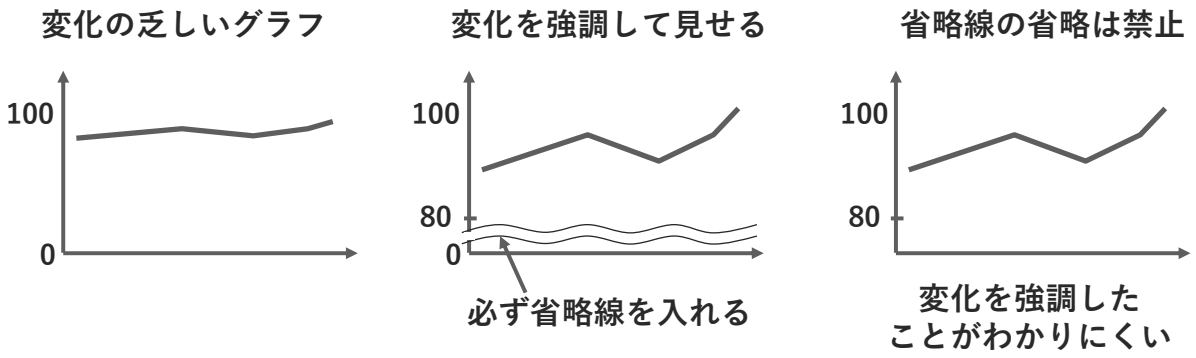
「ツリーマップ」は数値を面積で表し、長方形を分割する形で「構成比」を階層的に視覚化できるグラフです。上図例ではある会社の事業が大まかに「国内」と「海外」に分かれ、「国内」がＡとＢ、「海外」が CDEF に分かれていることをそれぞれの大きさとともに示しています。通常、「構成比」の表現には円グラフや積層棒グラフを使いますが、いずれも「大きな差のある数値」を表すのには向いていません。ツリーマップは面積で数値を表すので、Ａと Ｆのように大きな差のある数値でも直観的に把握しやすい形で表現できます。

3.11 やってはいけない禁止・注意事項

グラフを使う際に知っておくべき基本的な禁止事項・注意事項がいくつかあります。

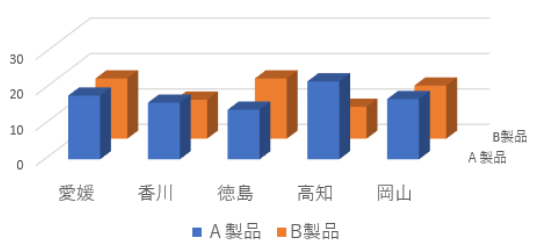
数値の変動が乏しい時、変化をわかりやすくするために軸の一部を省略して「変化のある部分を拡大して見せる」場合がありますが、その場合は「省略したことがハッキリ分かるように省略線を明示」しなければなりません。

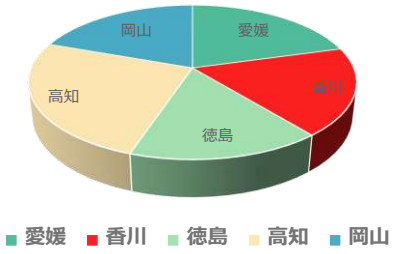
下記例の左側がその「変化の乏しいグラフ」であり、数値が 100 弱でごくわずかしか変わっていません。その変化を強調したのが中央の図で、軸の一部を省略して 80～100 の間だけを拡大して見せています。このように軸の一部を省略した場合は、そのことを示す二重波線を必ず挿入して省略を明示する必要があります。右の図のように「省略線を省略」すると、下の「80」という数字を見落されて「あ、こんなに変動大きいんだ」と誤解されやすくなります。TV 番組ではこの「省略線の省略」をしている例がよくありますが、詐欺的な表現として厳しく批判される種類の手法ですので決して真似しないでください。



3D グラフは数値を正確に把握しづらいため、使わないことを基本にします。特に円グラフを 3D 化すると形がゆがんで非常に誤解を与えやすいので基本的に禁止です。

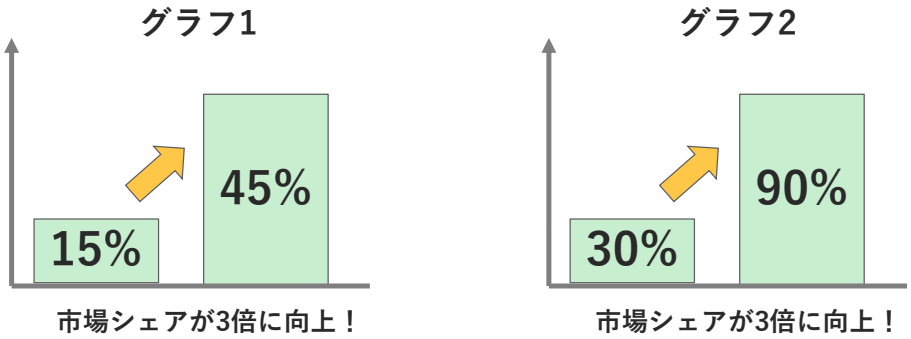
3.11.1 3D グラフは正確な数値を把握しづらいので使わないようにする





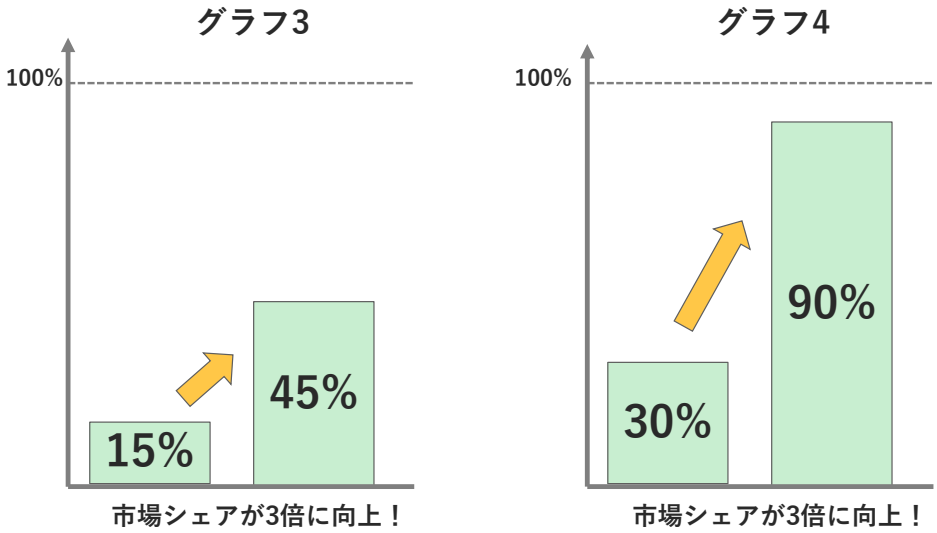
グラフの内容によっては「100%」が決まっているものがあります。その場合は必ず「100%」がどのレベルなのかを明示するようにします。下記のサンプルはそれをしていない例です。グラフ 1，2 はいずれも「市場シェアが 3 倍に向上」したと言っていますが、その結果達成できた数字は 45%対 90%と大幅に違うのに、「100%」を明示していないのでどちらも似たものに見えてしまい、その差がわかりません。

3.11.2 「100%」を示していないグラフを書いてはいけない



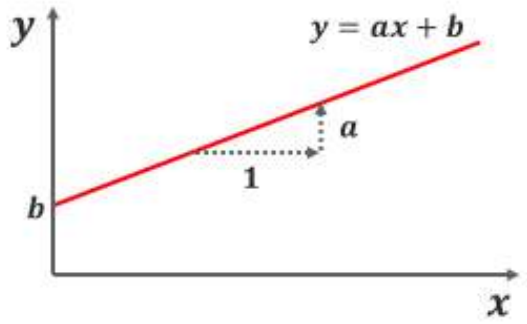
それを改善すると下記のようになります。「市場シェア」のような数値は「100%」が上限でありそれを超えることはありません。したがって「90%」はそれ以上の向上余地がほとんどないのに対して「45%」ならまだまだ伸びしろがありそうです。100%のラインを示すとこの重要な違いがハッキリと伝わります。100%があるときは必ずそれを明示するようにしてください。

3.11.3 「100%」が決まっているときは必ず明示する



3.12 一次関数

「一次関数」は、$y = ax + b$ という数式で表せる関係です。
仕事の中では非常に多くの場面で出てきます。



x が 1 増えると y が a 増える。
x =0のときの y の値は b 。

代表的な例：　製造個数と売上原価の関係、
水光熱費等の従量制サービスの費用、

「さまざまな関係」の項でも出てきましたが、仕事の中で出てくる数値の関係の中には「製造個数と売上原価」「水光熱費等の従量制サービスの費用」など、「一次関数」で表せるものが非常によくあります。

一次関数は数式にすると $\boldsymbol{y=ax+b}$ という形で、グラフを描くと直線になります。

【製造個数と売上原価】

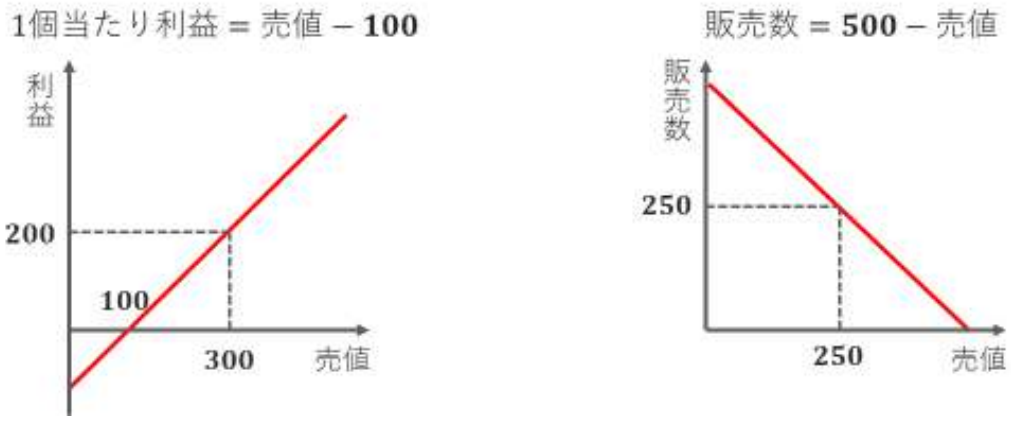
うどんやそばのような商品を製造していて、「売上原価」にたとえば小麦粉やそば粉のような材料費と工場の家賃が含まれるとしましょう。通常、材料費は製造個数に比例して増えます。これが「$\boldsymbol{x}$ が 1 増えると $\boldsymbol{y}$ が $\boldsymbol{a}$ 増える」の部分、$\boldsymbol{ax}$ に該当します。一方、工場の家賃は製造個数に関係なく一定なのが普通です。これが「$\boldsymbol{x}$ =0 のときの $\boldsymbol{y}$ の値は $\boldsymbol{b}$」の部分、$+\boldsymbol{b}$に該当します。

【水光熱費等の従量制サービスの費用】

水光熱費等は、まったく使わなくてもかかる基本料金＋使った分に比例してかかる従量制料金という構成になっていることが多く、これも $\boldsymbol{y=ax+b}$ になります。

3.13 一つの変数が複数の数字に影響する

ある商品1個あたりの原価が100円だとします。
売値を高くすれば1個あたりの利益は増えますが売れる個数は減ります。
売値を安くすれば1個あたりの利益は減りますが売れる個数は増えます。
もし1個 x 円で売ると1日当たり 500-x 個売れるとしたら、利益を最大にするためにはいくらで売れば良いでしょうか？



一次関数 $y=ax+b$ で表せる関係の場合、x が増えたとき y は減るか増えるかどちらか片方に決まります。

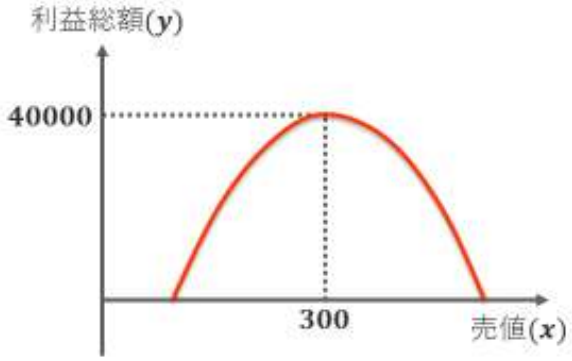
y= ax+b　　x が増えれば y は増える
y= -ax+b　　x が増えれば y は減る

しかし、仕事で現れる数字の中にはこうならないものもあります。
たとえば「値段」と「利益」の関係はその 1 つで、値段を上げると利益は増えますが、上げすぎると売れなくなってかえって利益が減る傾向にあります。このような関係は一次関数では表せません。

3.14 二次関数

1個当たり利益×販売数　で利益の総額になります。

利益総額 = 1個当たり利益 × 販売数
= (売値 − 100) × (500 − 売値)
= −売値^2 + 600 × 売値 + 40000
= −(売値 − 300)² + 40000



このような場合、売値と利益総額の関係は左のような曲線（放物線）を描きます。

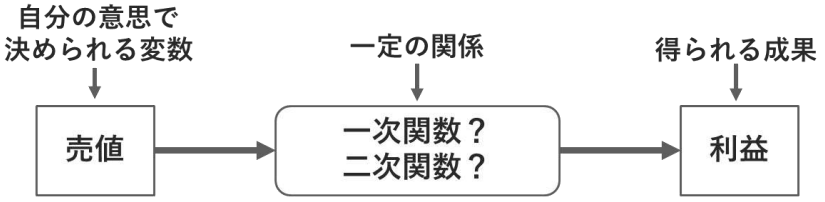
x^2 を含む式を二次関数式といい、そのグラフは放物線になります。

$$y = -x^2 + 600x + 40000$$

1 個あたりの利益に販売数を掛けると利益総額になります。1 個あたり利益が 売値-100、販売数が 500-売値になる場合、売値と利益総額のグラフを描くと上図のような曲線になり、売値=300 円のときに利益総額が最大になります。

この場合、式は $y = -x^2+600x+40000$ のようになります。x2 の項を含む式で表せる関数を二次関数と言い、そのグラフは放物線になります。

この例は「二次関数」を説明するための極端な設定であって、実際には売値と利益がこのような単純な放物線の関係になることはほとんどありません。しかし、経営をするためには「自分の意思で決められる変数」を変えたときに「どんな成果が得られるか」を予測する必要があります。変数と成果の関係は一次関数や二次関数である程度近似できる場合があり、この考え方を知っておくと予測・判断をするために役立ちます。



「一定の関係」を数式で表せると、成果の予測に役立つ

3.15 用語集

変動費	売上が増えるとそれに応じて増える費用で、原材料費などが代表的。
固定費	売上に関係なくかかる費用で、家賃、水光熱費、機材費などが代表的。人件費も固定費に近い性質がある。
変数	経営に影響するさまざまな数値に名前をつけたもの。販売数、売上額、来店客数などはいずれも変数である。
一次関数	2 つの変数 x と y について $y=ax+b$ という関係が成り立つとき、y は x の一次関数であるという。製造業では生産量と費用が一次関数の関係になる場合が多い。
二次関数	2 つの変数 x と y について $y=ax^2+bx+c$ という関係が成り立つとき、y は x の二次関数であるという。販売価格と販売数の間に「値上げすると販売数が減少し、値下げすると増える」というような明確な比例関係がある場合、販売価格と利益の関係は二次関数になることが多い。
クロス集計	データを集計する際、ある項目と他の項目の関連性を把握できるように集計を行うことを言う。たとえば「顧客の年齢層」を表頭、「商品カテゴリー」を表側においてクロス集計をすると、どの年齢層の顧客がどんなカテゴリーの商品を買っているのか/いないのかを把握できる
表頭	集計表の上端の見出し項目のこと。
表側	集計表の左端の見出し項目のこと。
棒グラフ	数値を棒の長さで表したグラフであり、商品別売り上げなどの連続性のないデータを表すためによく使われる。
ヒストグラム	1 クラスの生徒の身長分布、日本人の年齢構成など、連続性のあるデータを人為的にいくつかの「階級」に区切って集計表示するために使われる、特殊な形式の棒グラフ。
折れ線グラフ	数値を点の高さで表し、点と点の間を線でつないだグラフであり、連続性があって長期間継続しているデータを表すためによく使われる。
散布図	点が散らばったような形で書くグラフ。縦軸と横軸のデータに関連があるかないかを検証するためによく使われる
データ系列	一つの項目に関するデータを蓄積したものをデータ系列という。たとえば「A 商品の月別売り上げ」は一つのデータ系列であり、「B 商品の月別売り上げ」はまた別なデータ系列である。
集合棒グラフ	複数のデータ系列を横並びの棒で表現した棒グラフ
積層棒グラフ	複数のデータ系列を縦に積み上げた棒で表現した棒グラフ

ファンチャート	基準年の数値を 100%として、それ以後の変化を基準年に対する比率で表現した折れ線グラフ
省略線	数値の変化が乏しいグラフを描くときに変化を分かりやすくするために軸の一部を省略する際、省略したことを示すために挿入する波線のこと
3D グラフ	グラフを立体的な形状で表現したグラフ。誤った印象を与えやすい。

3.16 確認問題

【問 1】次の選択肢の中から、変動費を説明しているもの、固定費を説明しているものを 1 つずつ選びなさい

【選択肢】

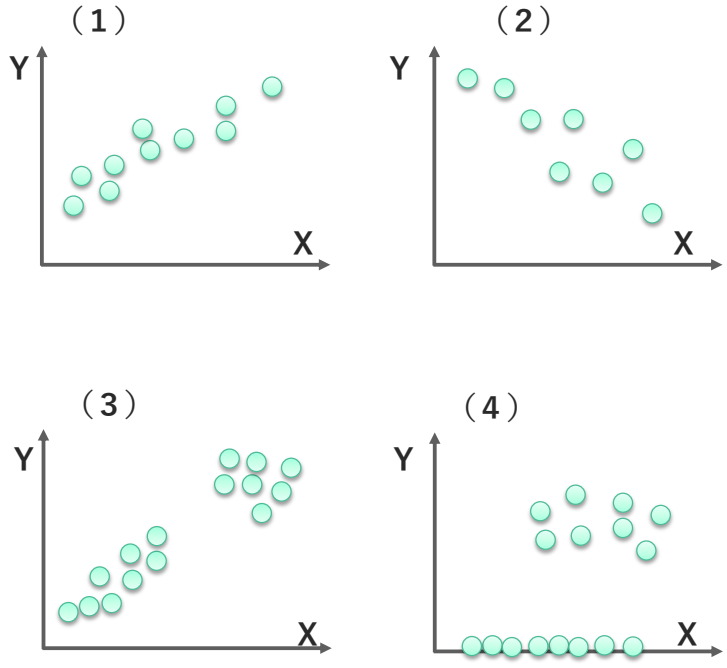
（A）商品の販売価格に対する利益の割合

（B）商品の生産コストのうち、原材料費、労務費など「商品の生産量に比例してかかる」傾向のある費用

（C）売上の大小に関係なくかかる費用であり、家賃が代表的

（D）会社の売上が少ないうちは赤字だが、売上が増えてある額を超えると利益（黒字）が出るようになる、その赤字と黒字の境目となる売上額のことを言う

【問 2】（1）～（4）のグラフがそれぞれどのデータのものなのか、適切と考えられるものを A～D の中から選びなさい



【選択肢】

（A）X 軸は気温、Y 軸はタバコの販売数

（B）子供向けおもちゃ店への来店客の年齢（X 軸）と身長（Y 軸）

（C）X 軸は 1 日当たりの来店客数、Y 軸はその日の売上額（一店舗についてのみ）

（D）X 軸は気温、Y 軸はあつあつの中華まんの販売数

■確認問題解答

【問1解答】　**変動費：B、固定費：C**

なお、Aは利益率、Dは損益分岐点

【問2解答】

（1）－（C）　通常、来店客数と売上額には正の相関がある。

（2）－（D）　あつあつの中華まんは寒い日のほうが売れる傾向がある（気温と販売数に負の相関がある）

（3）－（B）　子供向けおもちゃ店は親が子連れで来店する場合が多いので、ある年齢以下の子供世代と親世代に分かれたグラフとなる

（4）－（A）　未成年者はタバコを買わない。成年に達しても買わない人はまったく買わないため購入額はゼロに張り付く。買う人の購入額だけが散らばる

4 仕事と集合Ⅰ

4.1　分類してください

あるお店のお客様リストをいくつかのグループに分類してください。
分類する基準にはどのようなものが考えられますか？

氏名	性別	住所	年齢
田中一郎	男	東区境田...	22
鈴木さとみ	女	東区山形...	25
林田亜里沙	女	西区横河...	37
佐藤浩二	男	南区磯部...	26
榊原安昌	男	東区片田...	41
伊達みどり	女	西区上町...	52
野沢公子	女	東区岩瀬...	32

上の表はあるお店のお客様リストの一部ですが、これを分類するとしたらどのような基準が考えられるでしょうか？

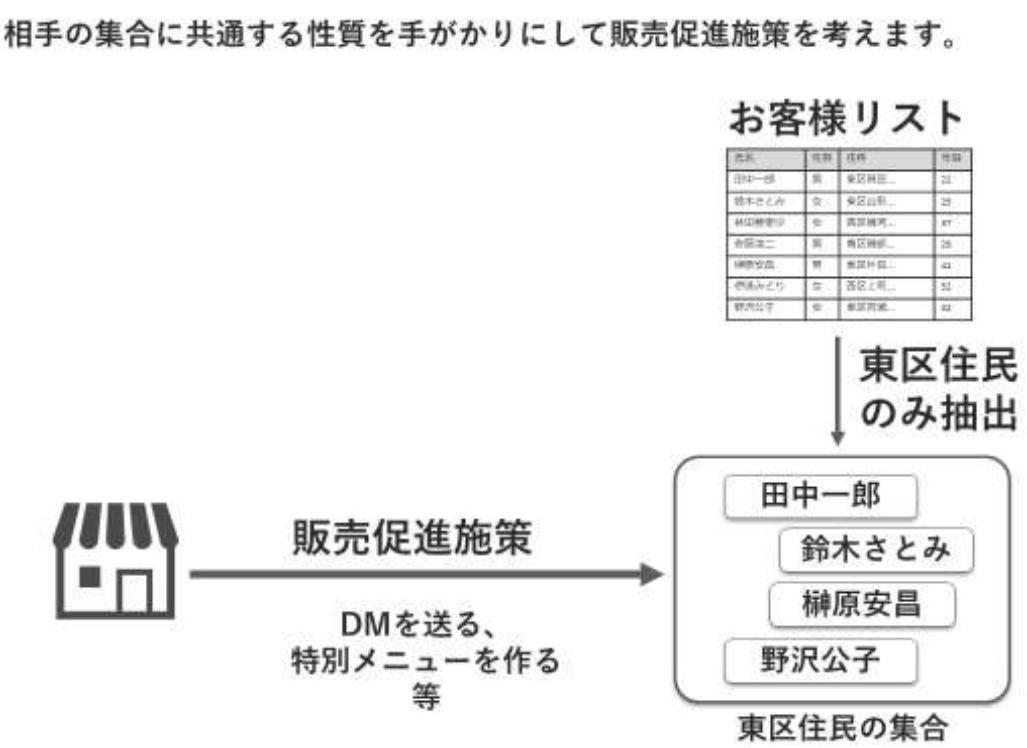
【性別】
女性向けファッション商品の企画をしても男性客の興味を引くことはできませんし、サッカーや野球などのスポーツ用品に興味を持つのは男性が中心です。性別で商品やサービスの好みが違うケースは非常によくあるため、性別での分類はよく使われる基準です。

【住所】
住所からもさまざまなことがわかります。店に近いか遠いかは来店頻度に影響します。幹線道路沿いの住人は騒音や大気汚染対策グッズに興味があるかもしれません。

【年齢】
年齢層によっても当然さまざまなニーズが変わってきます。年齢層によって好まれる服装の傾向は違いますし、20~30 代と 40-50 代では子供の教育に関するニーズも違うでしょう。

このように、「お客様リスト」は一つであってもそれを分類する基準は千差万別なのが普通です。

4.2　それはどのような「集合」ですか？

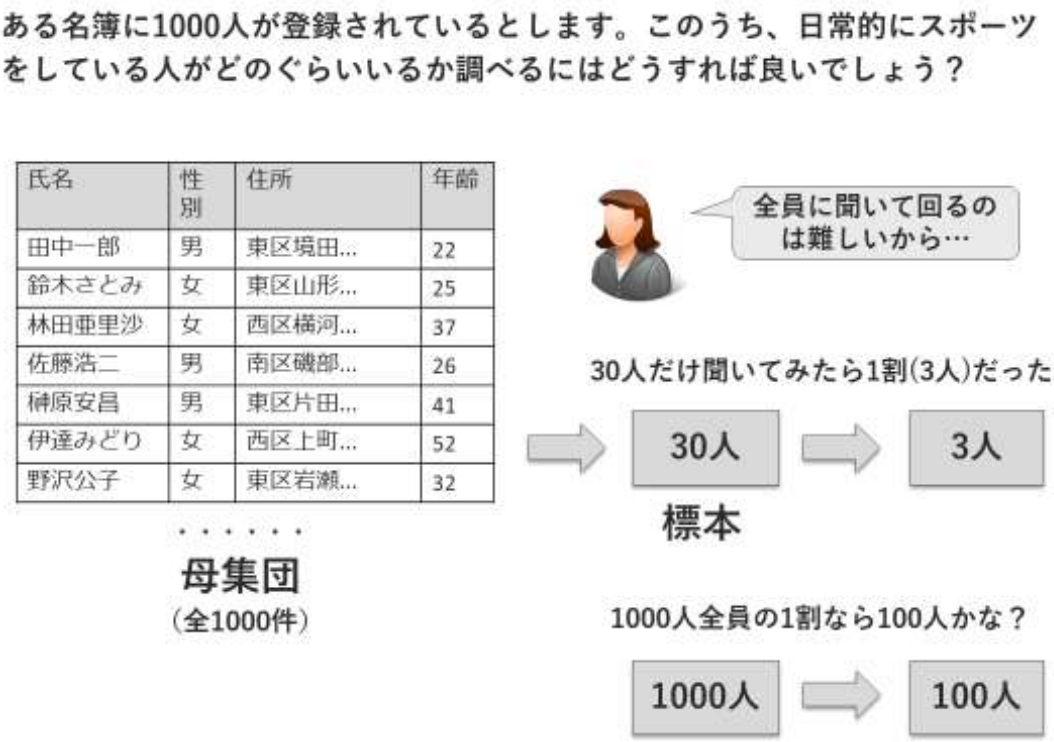


販売促進施策を考えるには、相手の「集合」に共通する性質を手がかりにします。
「集合」というのはもともとは数学の用語ですが、本書では「何らかの共通な性質を持つ要素を集めたもの」のという意味で使います。たとえばお客様リストからは

女性客の集合
東区住民の集合
30 代のお客様の集合

など、様々な「集合」を抜き出すことができます。「全員が興味を持つ商品／サービス」というのはあり得ませんので、「女性客」「音楽をやっている人」「映画好きな人」など、対象とする顧客のイメージを描いてその興味を引くように販売促進の企画を立てるのが基本です。

4.3　集合の全数を調査することは難しい



ある集合に共通する性質が分かれば経営判断の重要な手がかりになりますが、それが簡単ではない場合もあります。たとえばある店で「日常的にスポーツをしている人」向けの栄養食品を安く仕入れられる機会があって、ただし最低発注数が 20 個だったとしましょう。20 個をすべて売ろうとすると、興味を持ちそうな人がその 2 倍の 40 人は必要だとします。

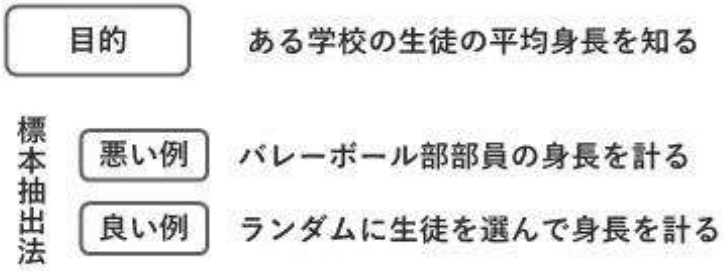
その店の顧客リストに 1000 人の人がいたとして、そのうち「日常的にスポーツをしている人」が 40 人いるかどうかを、どうすれば調べられるでしょうか？

1000 人全員に聞いてみるのは不可能ですので、そのうちの 30 人にだけ聞いてみたらそのうちの 1 割、3 人が該当したとしましょう。1000 人全員でもこの比率が成り立つのなら、全体では 100 人になります。40 人よりはかなり多そうなので、20 個発注しても大丈夫でしょう。

このような場合、「顧客リスト全員」にあたる 1000 人を「母集団（ぼしゅうだん）」、実際に聞いてみた 30 人を「標本（ひょうほん）」と言います。母集団の数が多い場合、全数を調査するのは難しいことが多いので、一部を取り出した「標本」で全体の傾向を推定する方法をとります。

4.4　母集団と標本の代表性

母集団の性質をよく表す（代表性のある）標本を選ぶ必要あり



標本を選ぶときは母集団の性質をよく表す標本になるように気をつけなければなりません。
たとえば、ある学校の生徒の平均身長を知りたい場合、「バレーボール部部員」を標本として選んだのでは正しい結果は得られません。バレーボール部には通常、高身長の生徒が多く集まっているからです。ランダムな基準で選べばそのような偏りはないと考えられるため、適切な標本と言えます。

このように、「標本が母集団の性質をよく表している」ことを「代表性がある」と言います。代表性がある標本を選ぶのは意外に難しいので、慎重に考えなければなりません。

たとえば、あるコンビニエンスストアに車で訪れる客の人数を調べたいとしましょう。その店に1日じゅう張り付いて車で来店する客を数えれば良さそうにも思えますが、おそらく平日と休日では傾向が違いますし、年末年始やゴールデンウィークのような時期もあれば、花火大会やコンサートなどの大きなイベントでも変わってきます。周辺の事故や工事で車の流れが変わる場合もあります。

4.5　代表値：最大・最小・平均・中央・最頻

最大値	32	集合に含まれる要素の最大の値
最小値	5	集合に含まれる要素の最小の値
平均値	16	集合に含まれる要素の平均の値
中央値	13	要素の中で中央に現れる値
最頻値	10	要素の中で最も頻繁に現れる値

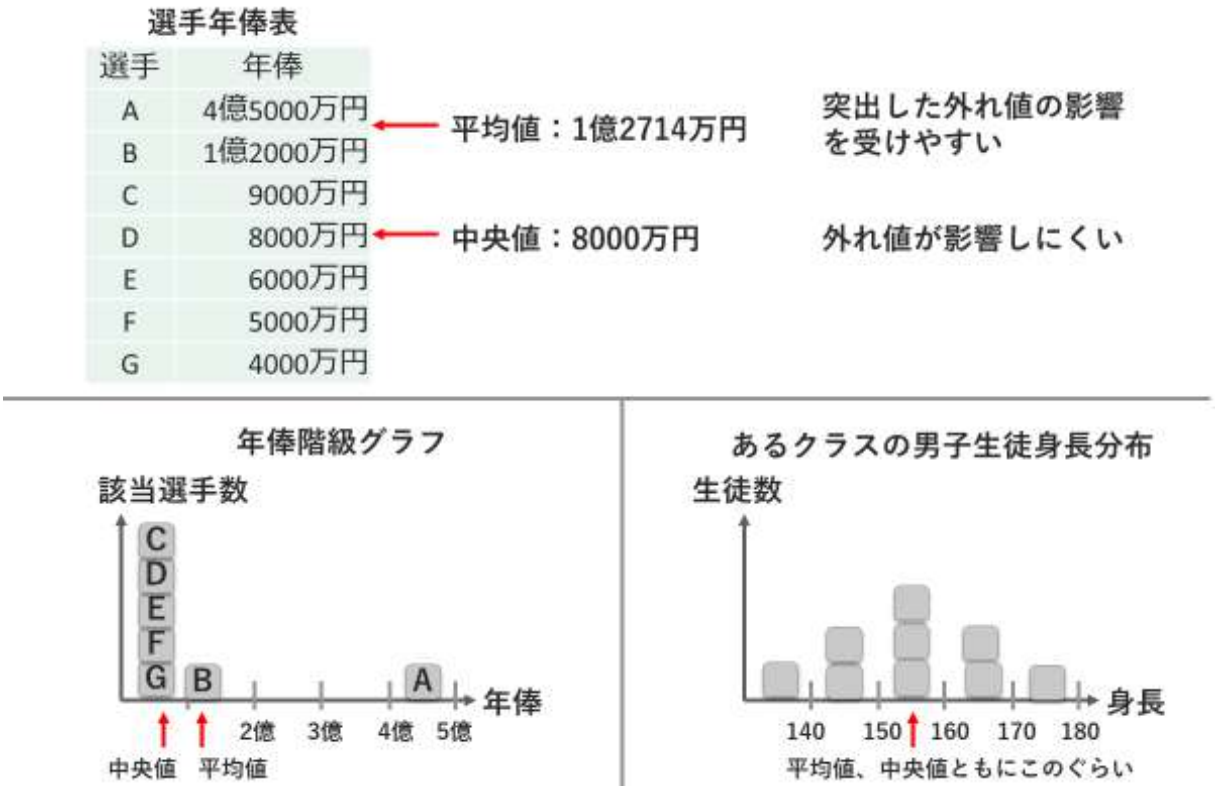
要素が数値であるような集合の性質を表す「代表値」が何種類かあるので覚えておきましょう。(前ページの「代表性」とは別な概念です)。

「最大値（さいだいち)」「最小値（さいしょうち)」は集合に含まれる要素の中のそれぞれ最大と最小の値、「平均値（へいきんち)」はすべての要素の数値を足して要素数で割った値です。平均は「平均点」などの用語でも使われているので有名です。

「中央値（ちゅうおうち)」は、要素を大きさ順に並べたときに中央に出てくる値です。平均値と紛らわしいので区別しましょう。

「最頻値（さいひんち)」は、最も頻繁に現れる値です。上図の例では「10」が 3 回現れているので最頻値になります。

4.6　平均値と中央値の違い



平均値と中央値には重要な違いがあります。上図左上の「選手年俸表」はあるプロスポーツ・チームの選手の年俸表だとします。A 選手が突出して大きく、他は高くても 1 億円を超える程度です。このようなデータの平均値を取ると 1 億 2714 万円、中央値では 8000 万円です。

では、「このチームの普通の選手の年俸、どれぐらい？」と聞かれたとき、どちらの数字のほうが現実に近いでしょうか？　明らかに中央値の「8000 万円」に近い金額をもらっている選手のほうが多いので、中央値のほうがこのチームの選手年俸の実態を代表していると言えます。

上図左下の「年俸階級グラフ」は、1 億、2 億・・・という年俸階級別に、該当する選手数をグラフにしたものです。A が明らかに突出していますが、「平均値」はこのように「突出した外れ値」の影響を受けやすい欠点があります。
右下の「あるクラスの男子生徒身長分布」は、身長の数字で同様のグラフを作ったものです。このように「突出した外れ値」が少なく、中央に山ができて左右になだらかに減っていくタイプのデータでは平均値と中央値に大きな差は出ません。

データの偏りが大きい場合、「平均値」は「集合の代表値」としては使えないので注意してください。

4.7　最頻値と中央値の違い

生データ

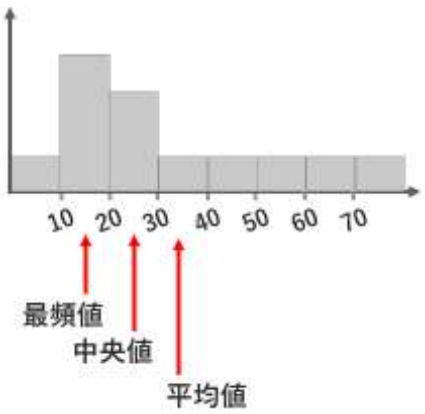
番号	数値
1	9
2	13
3	15
4	18
5	19
6	22
7	24
8	27
9	32
10	35
11	45
12	50
13	60
14	62
15	71

階級表

データ区間	頻度
10	1
20	4
30	3
40	1
50	1
60	1
70	1
80	1

平均値	33.2
中央値	24
最頻値	11~20

ヒストグラム



上図左側の「生データ」を見ると数値が 9~71 まで分布しています。これを 0~10、11~20、21~30 のように 10 ずつの「階級」に区切ってデータの出現個数を数えたのがその右の「階級表」です。たとえば階級表の 2 行目、データ区間 20 の頻度が 4 というのは、生データで数値 11~20 までの値をとるデータが 4 個あったという意味です。その下の平均値、中央値は生データに対して平均値と中央値を求めたものです。

一方、「最頻値」はデータの階級単位で考えます。この例では 11~20 の区間が最頻値だったことになります。

右の「ヒストグラム」は階級表をグラフにしたもので、データ区間ごとに何個のデータが出現したかが一目でわかります。前ページの「年俸階級グラフ」などのグラフもヒストグラムの一種でした。

ヒストグラム上に平均値、中央値、最頻値を描くと上図のようにやはりそれぞれ違う数値になります。「平均値は必ずしも最も普通のデータではない」、というのは前ページで記載したとおりですが、では最頻値と中央値の違いは何でしょうか？

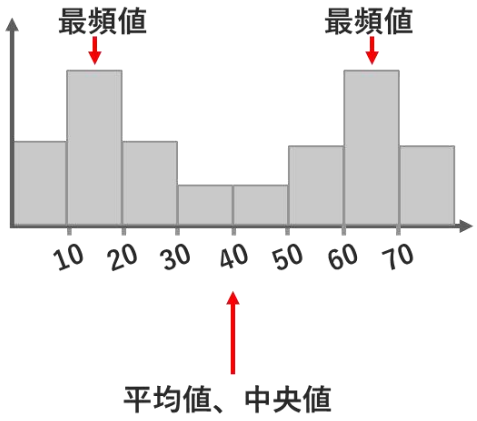
実は、販売促進の企画を立てる場合など、「中央値」よりも「最頻値」を使って考える場合がよくあります。たとえば「10 代の人向け商品」「20 代の人向け商品」のように年齢層で区切って企画を考えるとしましょう。上図の「数値」が年齢だとすると、10 代（データ区間 20 に該当）は 4 人、20

代（データ区間 30 に該当）は 3 人います。つまり「10 代向け商品」のほうが、買ってくれそうな客が一人多いわけです。

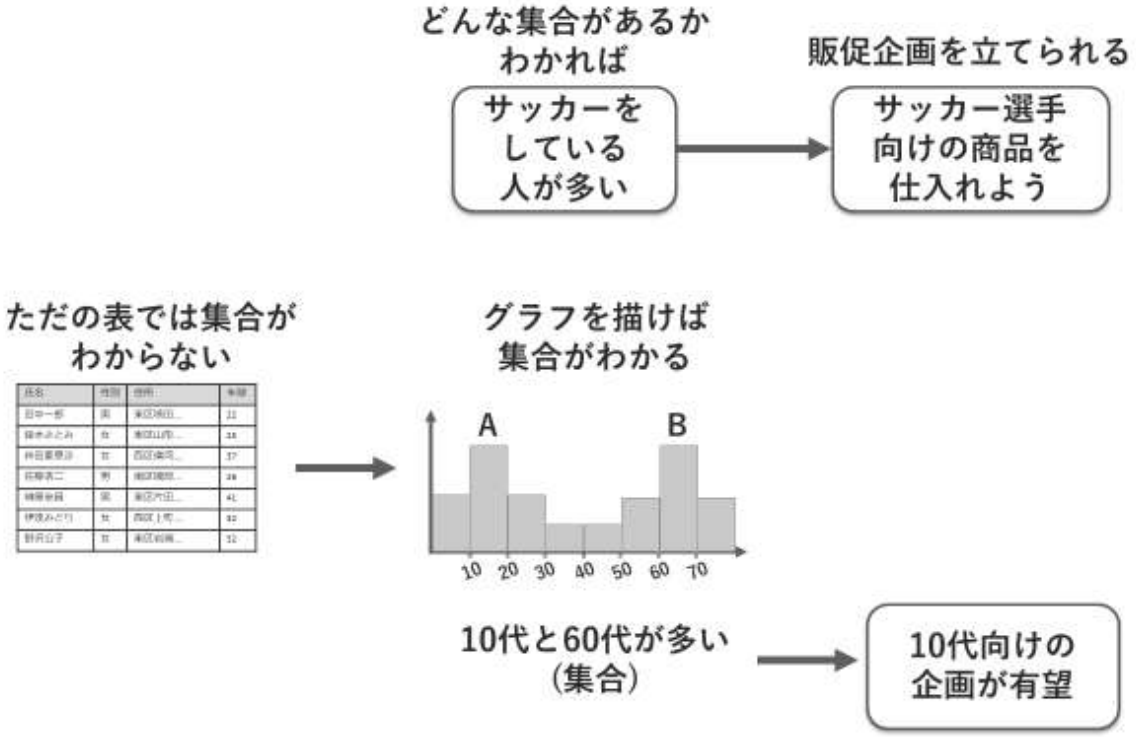
このように、データを階級で区切って考える場合は「最も数が多い階級」が重要になるので、最頻値という考え方も知っておきましょう。

最頻値は平均値や中央値とも大きく違う値になることがあります。たとえばデータが下図のように二つの山があるタイプのものの場合、平均値、中央値はいずれも真ん中付近になりますが、最頻値は右(61～70)と左(11～20)の二か所に出てきます。このような場合、平均値や中央値で企画を立ててもろくに売れません。祖父母が孫を連れて来店するような店の客の年齢階級を調べるとこんな数字になることでしょう。つまり商品企画は最頻値を知って考えることが重要で、しかも最頻値は一つだけとは限らない、そんな違いがあります。

フタコブラクダ型データ



4.8　グラフを描けば集合が見つかる

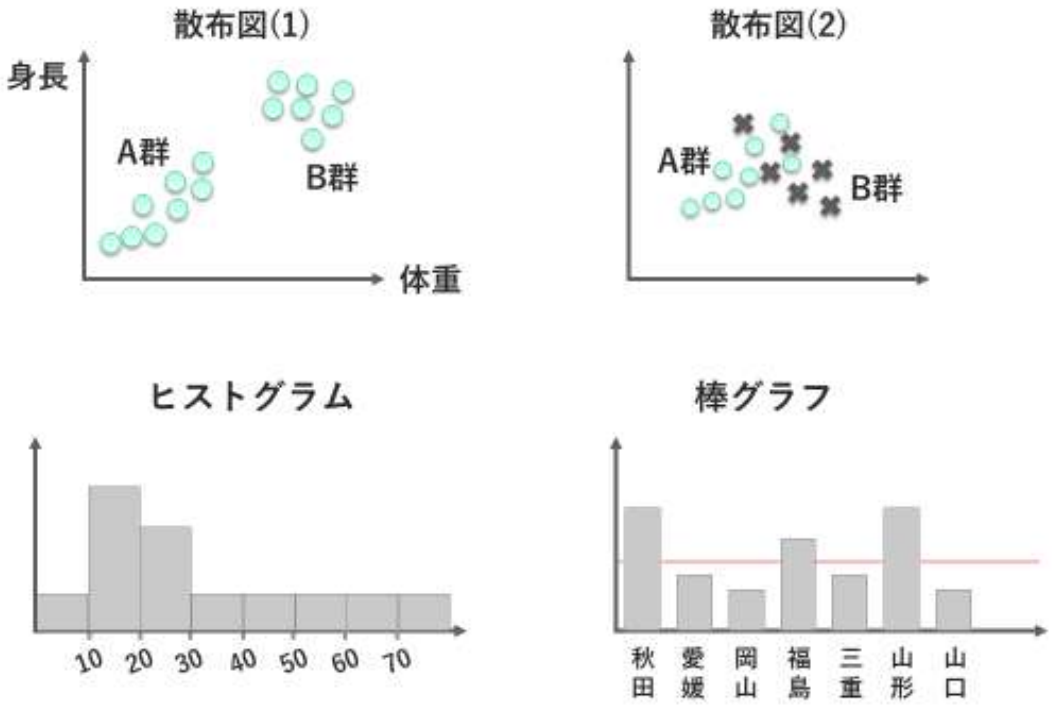


既に触れたように、販売促進の企画（販促企画）は特定の「集合」に対して立てるものです。たとえば自社の顧客の中にサッカーをしている人が多ければ、「サッカー選手向けの商品を仕入れる」など、サッカーに関わる企画が有望です。

それにはまず「集合」を見つける必要がありますが、顧客リストという「ただの表」を見ていてもなかなかわからない「集合」が、グラフを描くと分かることがあります。たとえば上図中央のヒストグラムを見ると、数字が年齢だとして、10 代と 60 代に高い山があります。であればその集合、つまり 10 代や 60 代向けの企画が有望だろう、と分かります。

顧客リストはたとえば「A さんは 18 歳、B さんは 22 歳、・・・」のような細かい数字をまとめた表ですが、「集合」というのは細かい数字を捨てて「10 代の顧客は○○人」のように大ざっぱな範囲や傾向を見る考え方です。そのためには「グラフ」のほうが扱いやすいのです。

4.9　集合を見つけるためのグラフ(1)



「グラフ」にも非常に多くの種類がありますが、「集合を見つけるためにグラフを描く」ためによく使われるグラフの種類や注意点を挙げておきます。

たとえば公園に遊びに来ている親子連れの身長と体重を量ると散布図(1)のようなグラフになるでしょう。A群は成長途中の子供でB群は親です。子供はある程度育つと「親と一緒に公園に遊びに来る」ことは少なくなるため、A・B群の間にはギャップがあります。散布図ではこのように「ギャップ」という形で集合の区切りが見えることがあります。

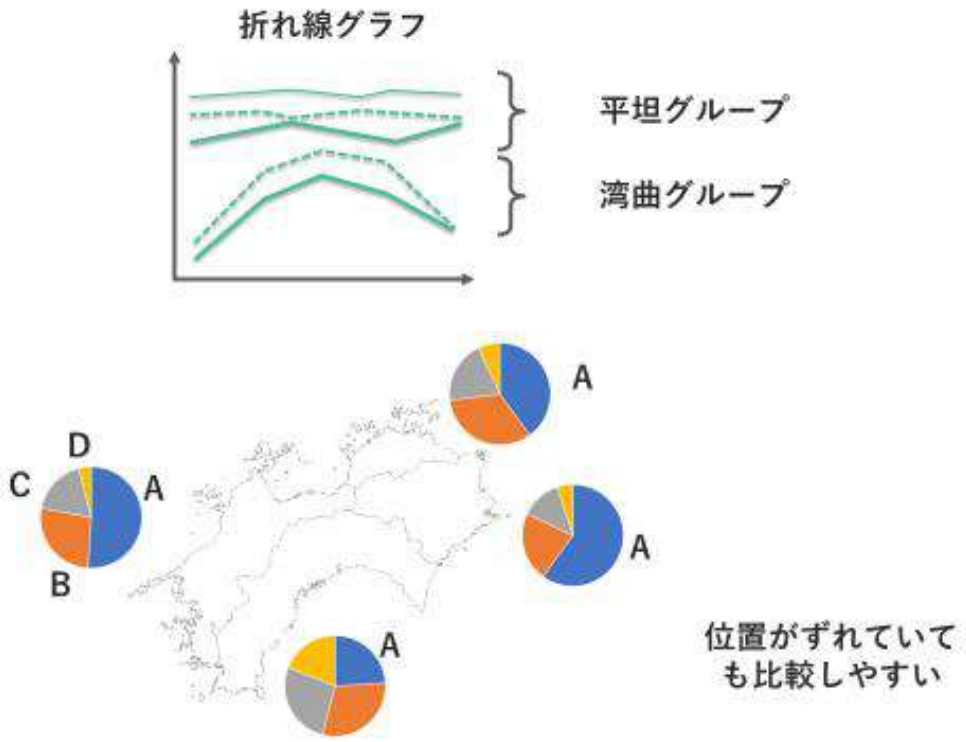
一方、散布図(2)はそのギャップが見えにくい例です。A群（○印）は右上がり、B群（×印）は右下がりになっていますが、AとBが重なっていてギャップがないので、もしすべてを○印で描いたらその区別はわかりません。このようなデータの場合、散布図以外の方法でA群・B群という集合は見つけておく必要があります。

左下のヒストグラムは、年齢のような連続的なデータを「階級」に区切ってその出現数（度数）を表すグラフです。「階級」自体が集合であり、度数が多いまたは少ない階級を見つけるために使います。

棒グラフはヒストグラムに似ていますが、データの並び順に注意が必要です。上図右下の棒グラフ

はいくつかの都道府県についての数値ですが、これを見てもすぐには「東北地方の県は数字が大きい」ことには気がつかないでしょう。これは県を 50 音順に並べているからで、東北・関東・東海・・・などの地方別に並べればすぐにわかります。ヒストグラムの場合、「階級」は本来連続的なデータなので並び順は自動的に決まるのに対して、棒グラフでは並び順を自由に変えられる場合があり、その結果、集合がわかりにくくなるケースがあるので気をつけましょう。

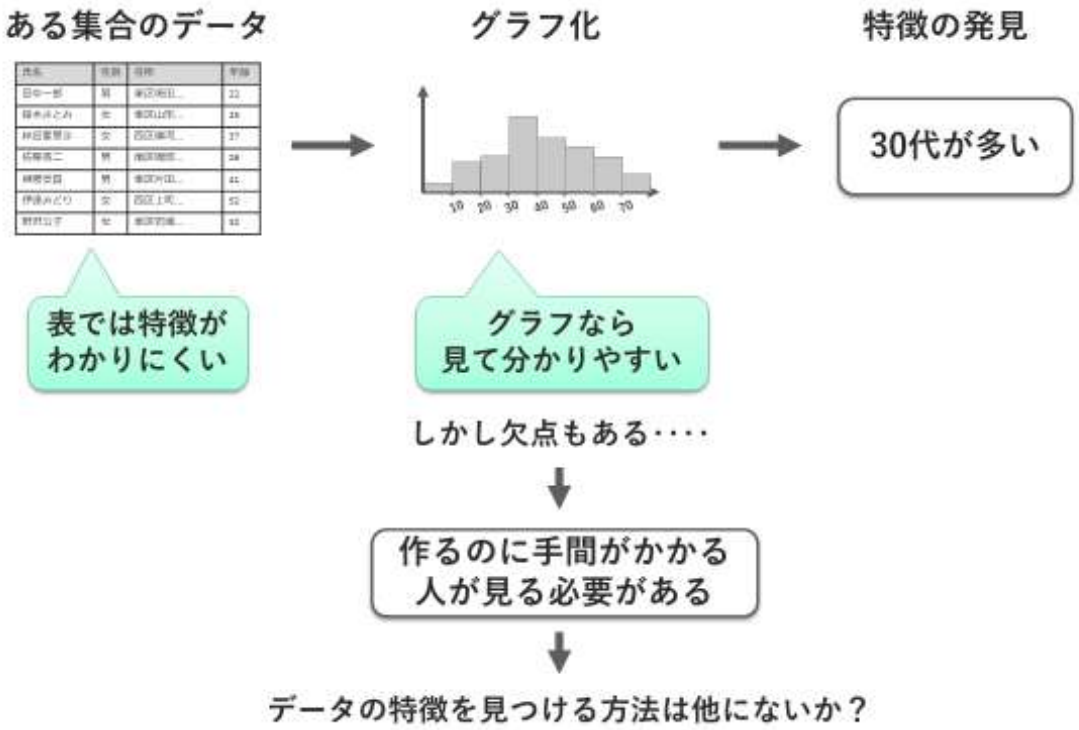
4.10 集合を見つけるためのグラフ(2)



「折れ線グラフ」で集合を探す場合もあります。上図の例では複数の折れ線グラフを載せていますが、おおまかに「平坦なグループ」と「湾曲しているグループ」という集合に分けられそうです。（ちなみに、平坦グループは熱帯地方、湾曲グループは温帯地方の年間の平均気温のグラフです）

円グラフには、位置がずれていても比較しやすいという特徴があります。たとえば上図が四国各県の産業別生産額（比率は架空のもの）だとしましょう。「A分野の比率が一番大きい県はどこ？」と聞かれた場合、円グラフならすぐにわかりますが、もしこれが棒グラフだとわかりにくくなります。棒グラフは縦か横にまっすぐ揃えて配置しないと比較しにくいのに対して、「角度」で大きさを表す円グラフは位置がずれていても比較しやすいからです。このため、円グラフは地図にひもづけてグラフを描くためによく使われます。

4.11 「グラフ」は便利だが万能ではない

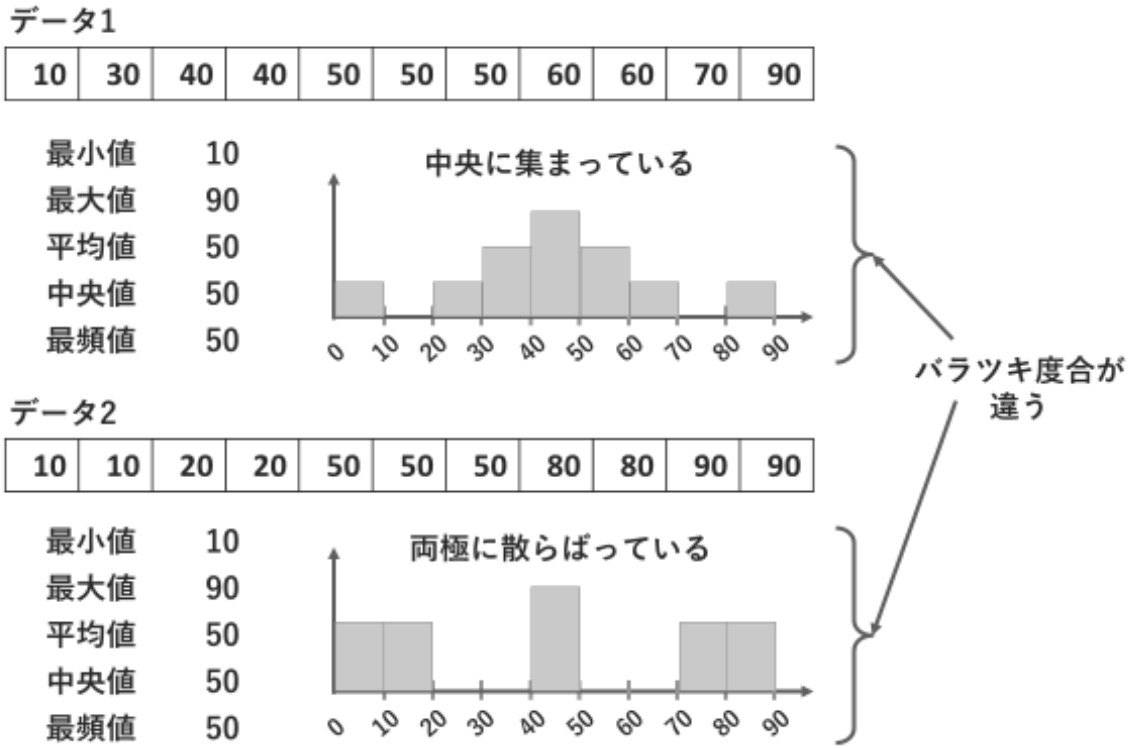


ある集合のデータが表になっていたとしても、数字の並んだ表を読んでいくのは難しいものです。そこでグラフをつくるとパッと見て分かりやすいので「30 代が多い」といった特徴の発見が楽になります。

とはいえ、グラフは便利ですがもちろん万能ではなく、欠点もあります。それは、「作るのに手間がかかる」ことと、「人が見る必要がある」ということです。1 枚だけなら「パッと見てわかりやすい」グラフでも、それがたとえば 100 枚もあったらいいかげんうんざりすることでしょう。たとえば 1 軒の雑貨店に約 10000 種類の商品があったとして、その商品別の売上推移をグラフにすると 10000 枚のグラフになります。特徴を見つけようと思っても、とても全部見ている時間はありません。

そこで、「人が見なくてもデータの特徴を見つけられる方法」があると便利ですが、そんな方法は何かないでしょうか？

4.12 データの「バラツキ」を考えよう

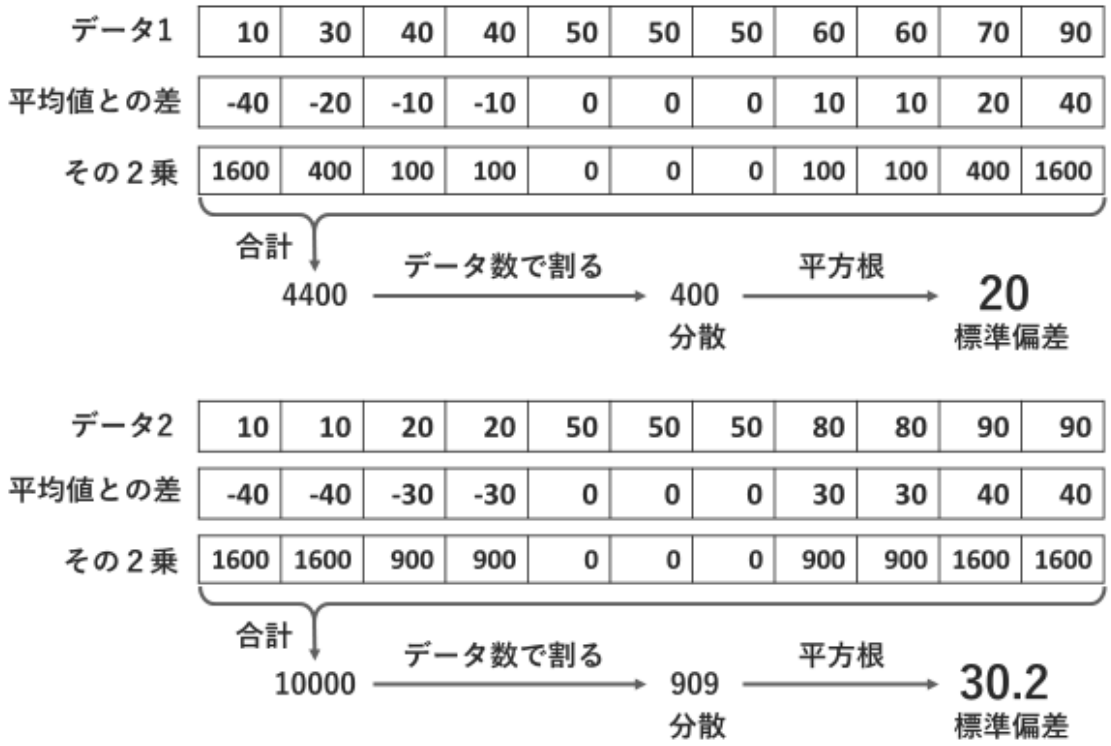


ここで、上図中のデータ 1 とデータ 2 を比べてみましょう。どちらもデータ個数は 11 個で値は 10～90 の範囲です。ヒストグラムを作ってみるとデータ 1 は中央に集まっているのに対してデータ 2 は両極に散らばっている（バラついている）という大きな違いがあります。しかし最小値、最大値、平均値、中央値、最頻値を計算してみると完全に一致してしまうため、これらの代表値ではその違いはわかりません。

仮にこの数字が年齢であれば、データ 1 の客層を持つ店なら 50 歳前後を主な客と考えれば済みそうですが、データ 2 の客層を持つ店なら低年齢層と高齢者層も別に考えなければうまく行かないでしょう。

データ 1 と 2 の違いは「集まっているか、散らばっているか」です。実はこのようなバラツキ度合を計算するための特別な方法があります。

4.13 平均値との差を考える



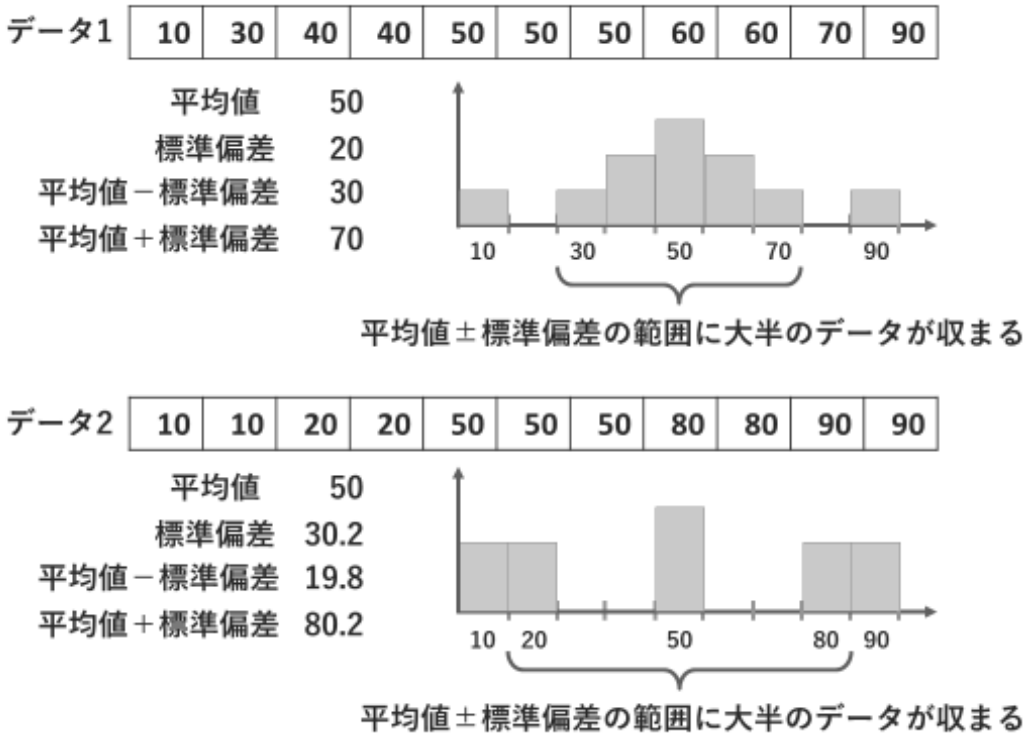
しばらく無味乾燥な計算が続きますが、いったんは下記で説明する計算過程を地道にたどってみてください。

「データ１」が元になるデータで、個数は 11、平均値は 50 です。次に、11 個の各データと平均値との差を求めます。10,30,40・・・と平均値の差は-40、-20、-10・・・になるわけです。
次に、その結果（平均値との差）を２乗します。-40 の 2 乗は 1600、-20 の 2 乗は 400、以下同様に 11 個のデータすべてについて平均値との差の 2 乗を計算して、それをすべて合計すると 4400 になります。この「4400」をデータ個数 11 で割ると 400 になり、これを「分散」と言います。さらにその平方根（2 乗すると 400 になる数）を計算すると 20 になり、これを「標準偏差」と言います。

「データ２」について同様の計算をすると、分散は約 909、標準偏差は 30.2 になります。

「標準偏差」と「分散」という２つの専門用語が出てきましたが、ここでは「標準偏差」に注目してください。データ１と２の計算結果を比べてみると、中央に集まっているデータ１よりも、両極に散らばっているデータ２の標準偏差のほうが大きいですね。これが重要なポイントで、**「標準偏差はデータのバラツキの大きさを表している」**のです。

4.14 標準偏差が意味するもの



あらためて、ヒストグラムを使って標準偏差の意味を図示したのが上図です。

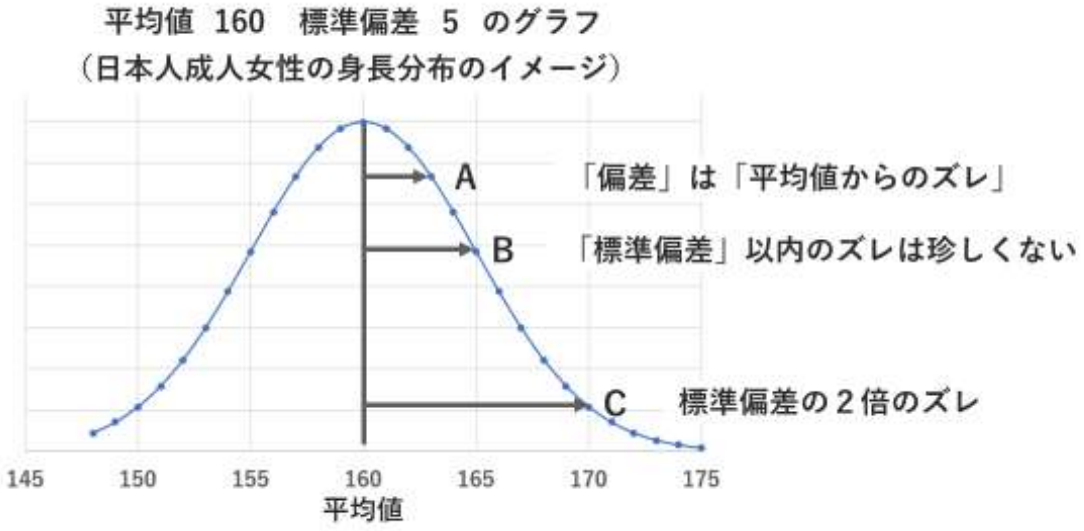
データ１と２の平均値は同じですが、標準偏差は違います。しかしどちらにしても、

平均値±標準偏差の範囲に大半のデータが収まっている

ことに注意してください。前ページの手順で計算した「標準偏差」にはそんな性質があります。そのため、標準偏差が大きいほど「データのバラツキが大きい」ことを意味します。

たとえばサッカー選手しかいない集合に対してならばサッカー選手向けの商品企画だけで済みますが、野球、ラグビー、バレーボールなど他のスポーツの選手も混じっているなら、それぞれに向けた商品企画をしなければなりません。「バラツキが大きい」というのはそういうことです。「標準偏差」は、グラフを見なくても「計算」だけでバラツキの大きさを判断できる指標です。

4.15 「標準」と「偏差」の意味



数式では標準偏差は σ （シグマ）という記号で表記します

「標準偏差」という用語は難しそうに見えますが、

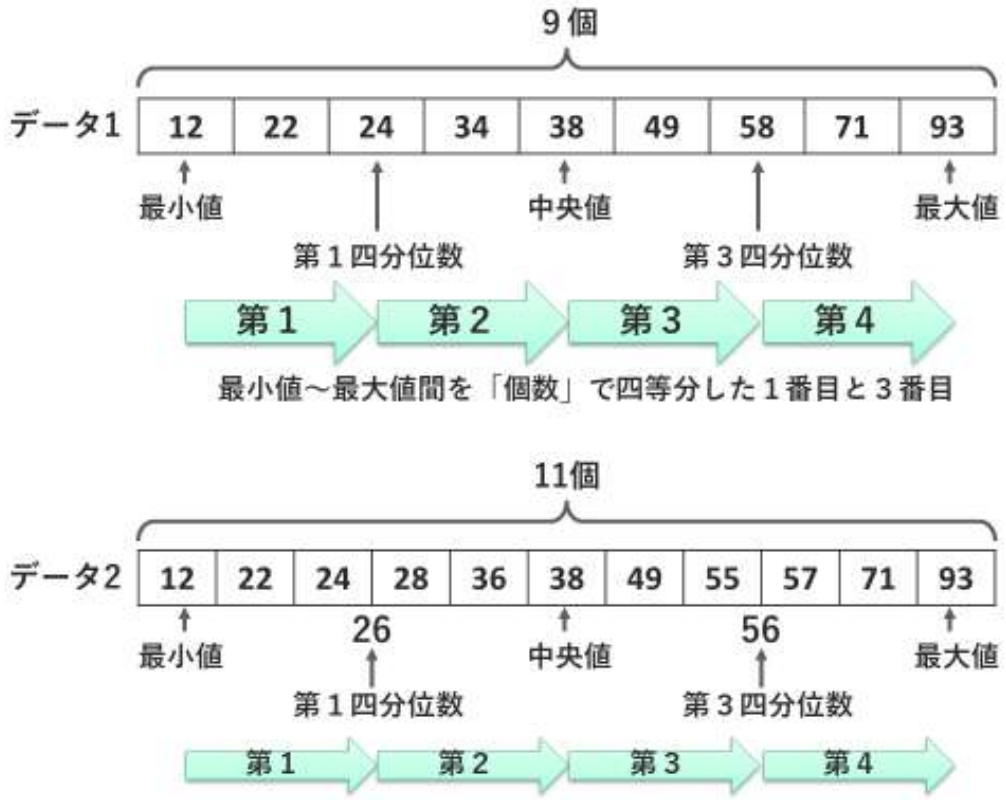
「偏差」は「平均値からのズレ」
「標準」は「普通に良くある範囲（珍しくない）」

のような意味と考えてください。上図は、「平均値 160、標準偏差 5 の正規分布」と呼ばれるデータをグラフにしたものです。日本人の成人女性の身長分布をイメージしてもらうと良いでしょう。

A 点は 163cm ぐらいの人で、この場合「平均値からのズレ＝偏差」は 3cm です。
B 点は平均値(160)+標準偏差(5)=165cm です。このぐらいの身長ならそれほど珍しくありません。これを越えるとだんだん「背が高いですね」というイメージが出てきて、C 点(170cm、つまり標準偏差の 2 倍のズレ)になるとかなり少なくなるのがわかります。

「標準偏差」は今後もよく使う非常に重要な概念なので覚えておいてください。数式の中で標準偏差を表す場合は σ という記号を使い、「シグマ」と呼びます。

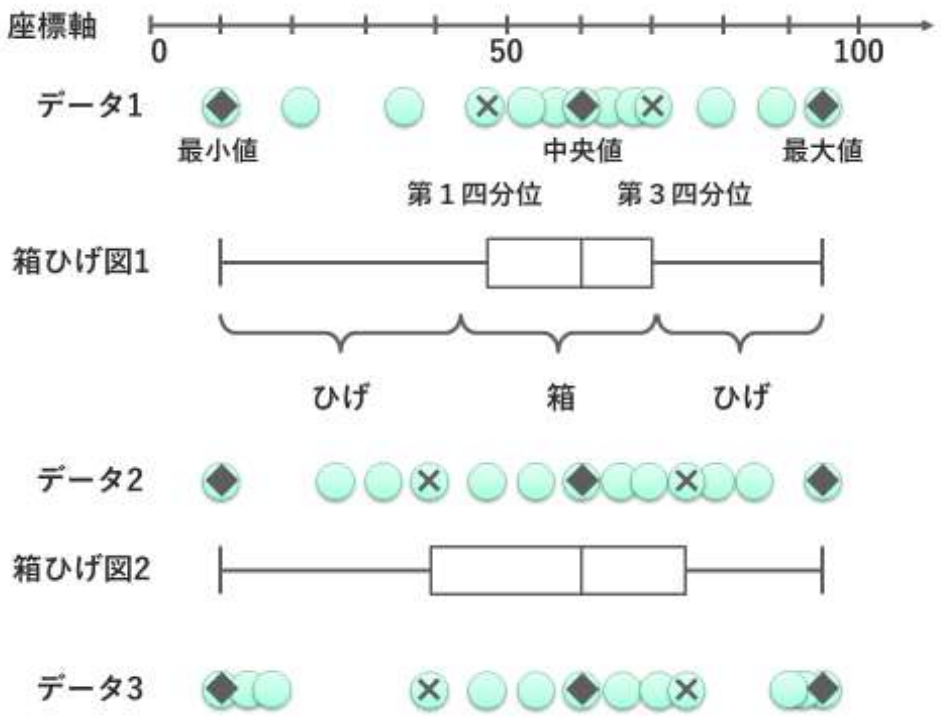
4.16 四分位数の考え方



データのバラツキを計算で示せる指標として、標準偏差の他に四分位数（しぶんいすう）というものも使われます。これは最小値～最大値間をデータの個数で 4 分割した各点の値のことです。上図のデータ 1 の場合、 4 分割した小さい方から 1 番目の「24」が第 1 四分位数、 3 番目の「58」が第 3 四分位数です。一般に「四分位数」というとこの第 1 と第 3 の 2 つを指します。「第 2 四分位数」は中央値、「第 4 四分位数」は最大値のことですので、これらについてはそのまま中央値、最大値と呼ばれます。また、最小値は「第 0 四分位数」に該当します。

11 個のデータで構成されているデータ 2 のほうは、第 1 四分位の位置が 24 と 28 の中間なので両者の平均をとって 26、第 3 四分位は 55 と 57 の平均をとって 56 になります。

4.17 箱ひげ図



四分位数を用いて「箱ひげ図」というグラフを書く方法があります。
上図の最上段にあるのは、0 から 100 までの目盛りを刻んだ単なる座標軸です。
その下の「データ 1」は、その座標軸を使ってあるデータをプロットしたものです。最小値・中央値・最大値に◆印を、第 1・第 4 四分位数に×印を付してあります。
このデータを「箱ひげ図」法でグラフ化したのが「箱ひげ図 1」です。この図で次のようなことがわかります。

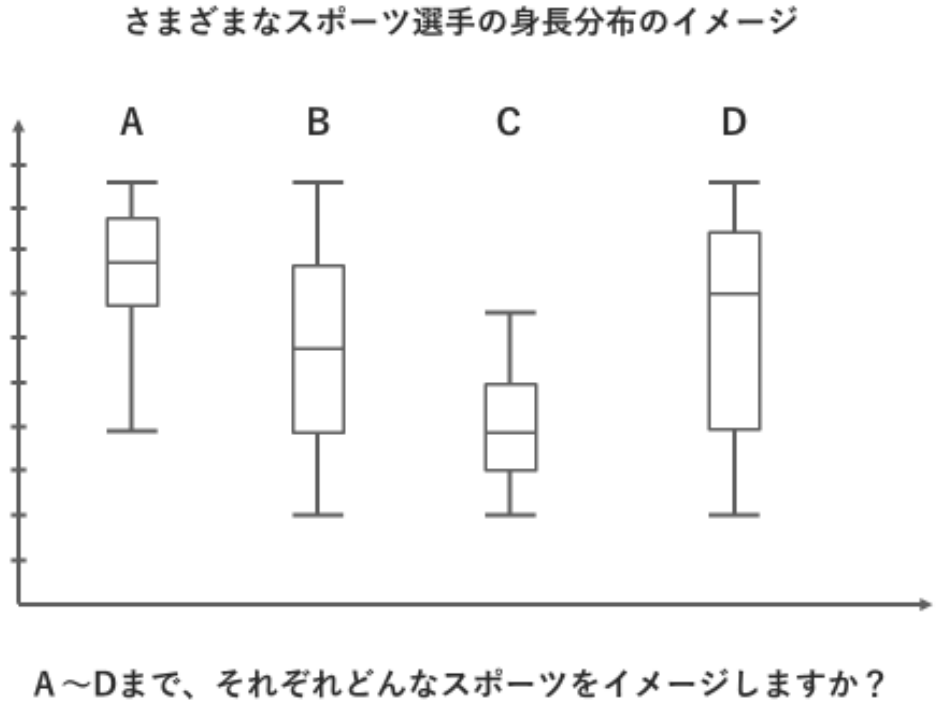
- おそらく中央値の周囲にデータが集中していて、最小値・最大値付近は少ない
- 最小値よりも最大値側に寄ったデータが多い

一方、その下の「データ 2」を基にした箱ひげ図 2 を見ると「箱」の範囲が箱ひげ図 1 よりも大きいので、最小・最大・中央値は同じでもデータ 1 よりもバラつきが大きいだろうということがわかります。

ただし箱ひげ図には欠点もあります。「データ 3」を使って箱ひげ図を描くと実は箱ひげ図 2 とまったく同じになりますが、データ 2 とデータ 3 を見比べるとずいぶん分布が違います。データ 3 では最小値・最大値付近にもデータがまとまっていることに注意してください。おそらくデータ 3 は最小値付近、中央値付近、最大値付近の 3 つのグループに分けて考えるべきです。箱ひげ図（および

その下になる四分位法）ではこのような違いが見えなくなってしまいます。
実は、箱ひげ図は中央値付近にだけ集中するタイプのデータを表現するのに向いた手法であり、データ 3 のように「複数の集中点があるデータ」のバラツキを表すのには不向きです。

4.18 箱ひげ図によるバラツキの表現



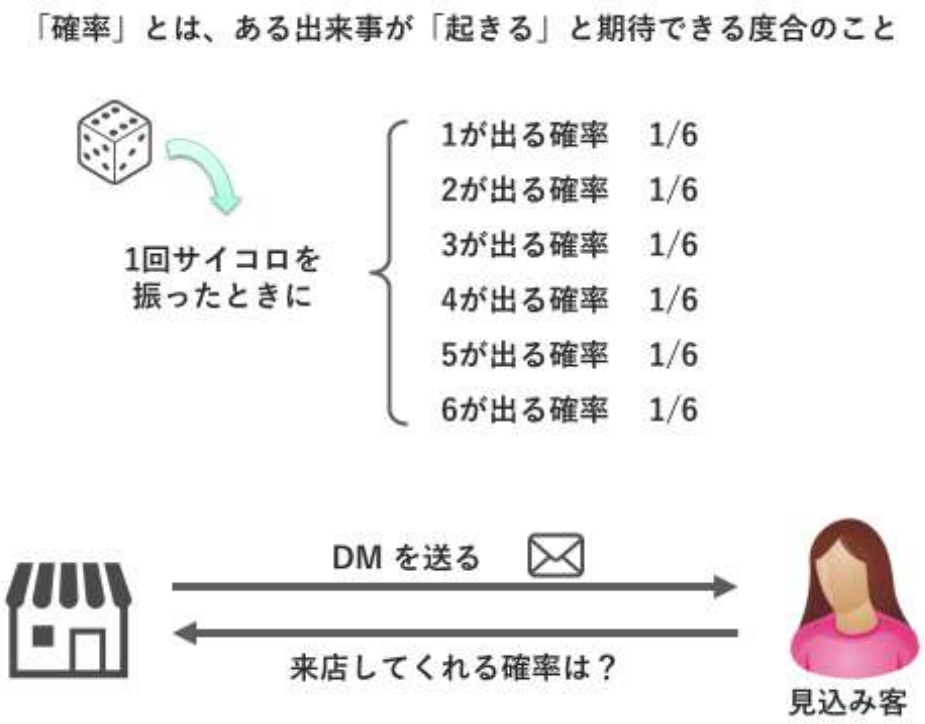
上図は、さまざまなスポーツの選手の身長分布を箱ひげ図で表したイメージです。

A は高身長に集中していますが低身長の選手も少しいる、これはバレーボールやバスケットボールなどに多いパターンです。

B は低身長から高身長までまんべんなくいるスポーツで、卓球や柔道が当てはまりそうです。
低身長側に寄った C は競馬や競艇など、体重制限のある分野でこの傾向があります。

D は B に似て低身長から高身長まで範囲が広いですが、中央値が高いので B に比べると高身長側が多くなっています。サッカーやアメリカンフットボールはこの傾向があります。ただし、アメリカンフットボールの場合はポジションによって身長の傾向がまったく異なり、「複数の集中点があるデータ」になるため、本来は箱ひげ図を使うのには向いていません。

4.19 確率とは？



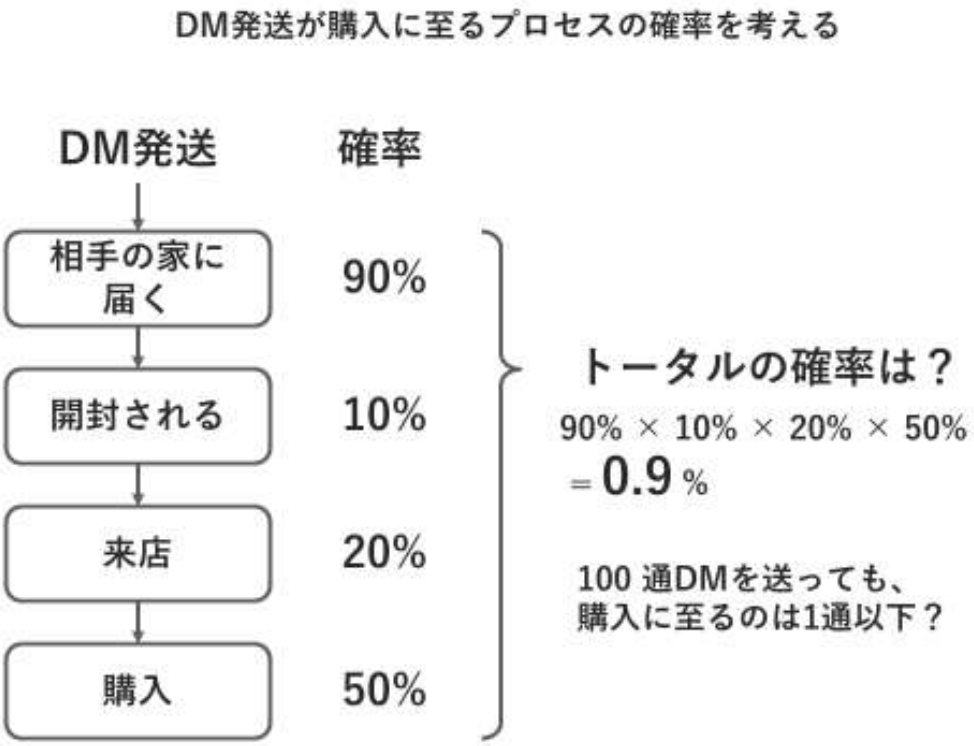
ある出来事が「起きる」と期待できる度合いのことを「確率」と言います。たとえば天気予報では「明日の降水確率は 80%」のように使われます。「降水確率 80%」と言えば「おそらく雨が降る」という意味ですし、「降水確率 0%」なら「絶対に降らない」という意味です。このように日常生活でも「確率」という言葉は使われます。

確率の中には分かりやすいものもあります。たとえば正しく作られたサイコロを 1 回振ったときには、1～6 までどの目が出る確率も 1／6 です。

仕事をする上でもこの「確率」が重要な場面があります。たとえば見込み客に DM（ダイレクトメール）を送ったとして、その人が来店してくれる確率はどのぐらいでしょうか？　もし DM を 1 通送るのに 100 円の費用がかかり、受け取った客の 10 人に 1 人が来店する（来店確率＝1/10）なら、来店客 1 人を集めるために 1000 円の費用がかかる計算なので、1000 円以上の利益が出る買い物をしてもらわないと採算が合いません。

DM に限らず、販売促進施策による来店確率は非常に変わりやすいので、十分考えて販促施策の手を打つ必要があります。

4.20 仕事の成否を左右する確率を考えよう



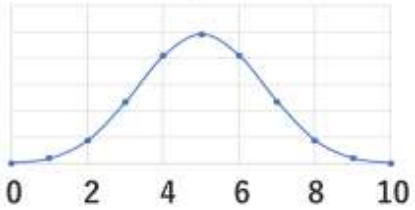
1 つの販促施策が成果を上げるまでには複数の段階があり、それぞれに確率があります。たとえば DM を発送してそれが相手の家に届く確率、開封されて読まれる確率、興味を持たれて来店する確率、店頭で購入される確率がそれぞれ 90%、10%、20%、50%だとすると、それを全部掛けたトータルの確率は 0.9%となり、100 通 DM を送っても購入に至るのは 1 通以下ということになります。

この数字は具体的な内容によって大きく変わります。たとえば「DM が相手の家に届く確率」の 90%は、最近会員登録された自社の顧客リストなら高めに出ますし、古いリストや他社から購入したものなら下がります。「開封される」から後の確率には、たとえば DM のデザイン、顧客ロイヤルティ、企画内容が大きく影響します。「DM を打っても売上につながらない」という場合、どのポイントで確率が低くなっているのかを突き止めて、ピンポイントでそこに手を打つことが求められます。

4.21 確率の基本：二項分布

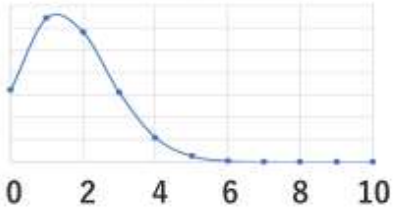
コインを10回投げたとき、
表がN回出る確率は？

表の回数	確率
0	0.0010
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.0010



サイコロを10回振ったとき、
6の目がN回出る確率は？

6が出る回数	確率
0	0.1615
1	0.3230
2	0.2907
3	0.1550
4	0.0543
5	0.0130
6	0.0022
7	0.0002
8	0.0000
9	0.0000
10	0.0000



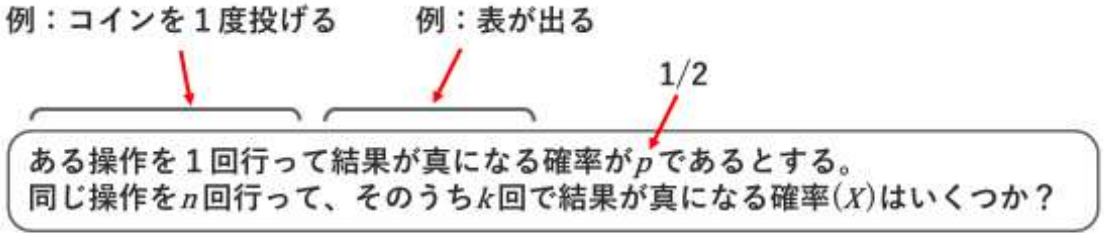
仕事上のプロセス、特に集客のようなものが「うまくいく確率」を見積もるのは非常に難しいものですが、単純なプロセスであれば確率は計算で正確に求められるものがあります。

たとえば 10 円玉のようなコインを 10 回投げたとき、表が出る回数は 0 回から 10 回まで考えられますがその確率は上図左側、同様にサイコロを 10 回振ったときに 6 の目が N 回出る確率は上図右側です。

コインを 1 回振って表が出る確率は 1/2 ですので、10 回振ったら 5 回が一番多くなりそうです。実際その通りなのですが、4 回・6 回も意外に多く出ることがわかります。

一方、サイコロを 1 回振って 6 の目が出る確率は 1/6 ですから、10 回振るとその１０倍で 10/6＝1.66 回になりそうです。この数字は 1 よりも 2 に近いので、直感的には「2 回出る確率が高い」ような気がしますが、実際には 1 回のほうが少し高確率で出ます。人間の直観が実際の確率と大きくズレるケースは珍しくありません。直観だけで確率を考えていると、経営上の判断を間違えがちです。少なくともこのような単純なプロセスの確率については、計算方法を知っておきましょう。今回例示した「表 or 裏」「6 or その他」のように「結果が 2 種類しかない操作を複数回繰り返すと一方の結果が何回出るか」を示す分布は二項分布と呼ばれていて、その確率は簡単な式で計算できます。

4.22 二項分布の計算式



この答えは次の式で求められます。

$$X = \binom{n}{k} P^k (1-P)^{n-k}$$

$$= \frac{n!}{k!\,(n-k)!} P^k (1-P)^{n-k}$$

（Excel では BINOM.DIST 関数で計算できます）

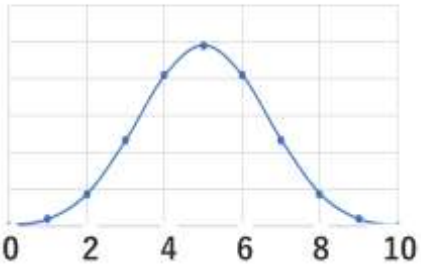
二項分布は「結果が二種類しかない操作を独立して複数回行う」場合を想定しています。「結果が二種類」のところでいう「結果」は実際には Yes/No、表/裏、真/偽 などの名前で呼ばれることが多いですが、「6 の目が出るか否か」のような想定もその結果は「真/偽」と見なせるので適用できます。

たとえば「サイコロを 1000 回振って 5 か 6 が 100 回出る確率」ならば、n=1000, p=1/3, k=100 とすれば計算できます。この式を手で計算するのは大変ですが、現在は表計算ソフト等でやれるので活用しましょう。例えば Excel なら BINOM.DIST という名前の関数でこの 2 項分布を計算する機能があります。

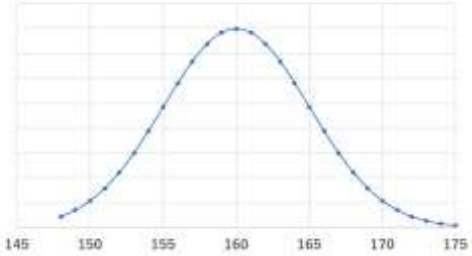
4.23 二項分布と正規分布は似ている

二項分布のグラフは、以前出てきた正規分布のグラフによく似ています。

二項分布のグラフ



正規分布のグラフ



実は二項分布を「正規分布」に近似する方法が実務的によく使われます。

正規分布のグラフを表す式

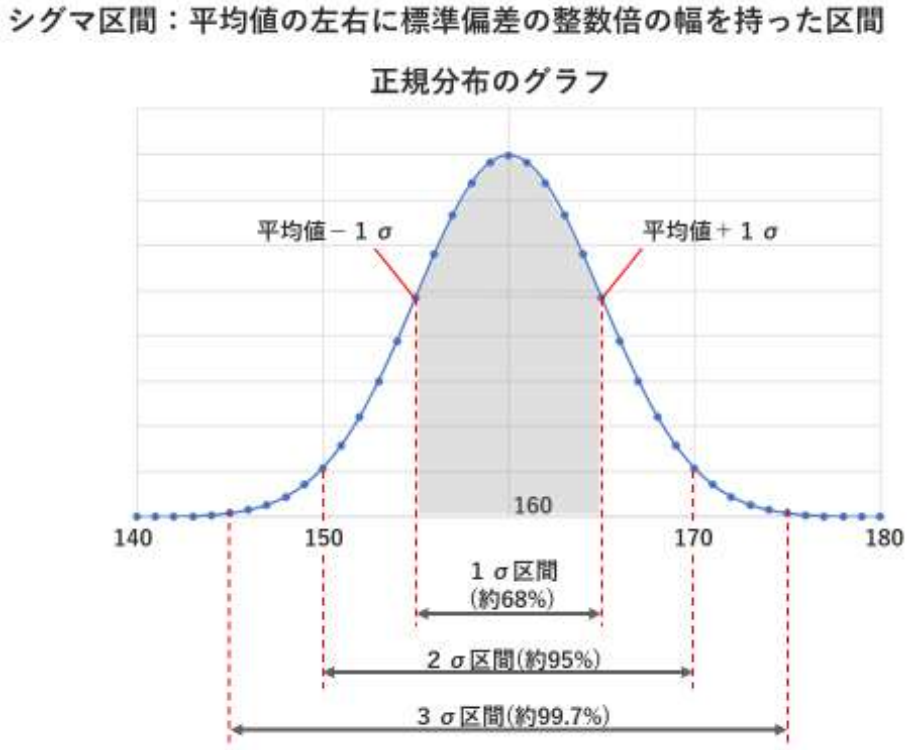
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

μ は平均値
σ は標準偏差

標準偏差の説明(p.16, 「標準」と「偏差」の意味)の中で「正規分布」という用語が出てきたのを覚えていますか？　正規分布というのは平均値と標準偏差で決まる関数で、数式では上図最下部の式で表され、山形のグラフになります。二項分布とは違うものなのですが、並べてみてみるとよく似ているのがわかります。

実は、二項分布は n（操作を試行する回数）が多くなるほど正規分布に近づいていく性質があります。
さらに、正規分布のほうが二項分布よりも実務的に扱いやすい場合が多いので、大まかな傾向が分かればいい場合は二項分布ではなく正規分布を使って確率の計算を初めとする判断を行うことが多いのです。

4.24 正規分布とシグマ区間



「コインを複数回投げて表が出る数」は、投げる回数が多くなるほど正規分布に近い結果が得られます。他にも、（人間を含む）動物の身長や体重、雨粒の大きさなど、自然界のランダムな現象は正規分布を示すことが多いので、ある分野について「どのぐらいの範囲のデータがありうるか」という見当をつけるには正規分布を知っておくことが役に立ちます。

たとえば、「日本人の成人女性の平均身長は 160cm、標準偏差は 5cm」だったとしましょう（この数字は仮定です）。その場合、身長 170cm の女性は何人に一人ぐらいいると考えればよいのでしょうか？　少ないのは間違いありませんが、具体的には 20 人に一人なのか、それとも 100 人に一人あるいはもっと少ないのでしょうか？　その答えによって、経営上の判断が変わってくることがあります。たとえば「170cm 以上の人が 50 人に 1 人よりも少ないなら、そのサイズの商品を作っても売れる見込みがないから作らない。それより多いなら作る」のような基準で商品企画をしている会社にはこの答えが必要です。そこで役に立つのが、

平均値を中心とする１σ区間に約 68%、２σ区間に約 95%、３σ区間に約 99.7%

のデータが集まるという、正規分布の特徴です。170cm はちょうど２σ区間の境界ですので、150~170cm の範囲に 95%の人が収まっているはずです。ということは残り 5%つまり 20 人に一人はそこからはみ出します。これは低いほう（150cm 以下）と高いほう（170cm 以上）の両方を含む

ので、高いほうだけであれば半分の「40 人に一人」程度はいるでしょう。したがって「日本人の成人女性で 170cm 以上の人は 50 人に一人以上いそうだ」ということが推定できるため、その商品を作ろう、という方向で経営判断ができます。

このように「平均値と標準偏差がわかれば、どの範囲にどれぐらいのデータがありそうかを簡単に推定できる」のが正規分布のメリットの一つです。1σ、2σ、3σと「標準偏差の整数倍」単位で境界を考えることが多いので、その範囲の概数とともにこれらの用語を覚えておきましょう。

名前	定義	含まれるデータ数
1σ区間	平均値±1σの範囲	約 68%
2σ区間	平均値±2σの範囲	約 95%
3σ区間	平均値±3σの範囲	約 99.7%

4.25 用語集

集合	何らかの共通な性質を持つ要素を集めたもの。
標本	なんらかの集合の性質を知るための調査をする際、具体的な調査対象とするために集合の一部を抜き出したもの。
母集団	標本抽出の対象となった「集合」全体のことをいう。たとえば「全校生徒のうち 10 人にアンケート調査を行いました」という場合は「全校生徒」が母集団、「10 人」が標本である。
サンプルサイズ	一つの標本の中に含まれる要素数を言う。「10 人にアンケート調査を行いました」という場合、サンプルサイズは 10。
標本数	標本調査を何度も行う場合の数を言う。「3 つの高校でそれぞれ 10 人ずつにアンケート調査を行いました」という場合、標本数が 3 になる。
代表性	標本が母集団の性質をよく表しているかどうか。
代表値	標本の特徴を表すようないくつかの数値のこと。最大値、最小値、平均値、中央値、最頻値などがある。
最大値、最小値	一つの標本に属す要素のうち、値が最大のものと最小のもの。
平均値	一つの標本に属す要素の数値をすべて足して、要素数で割ったもの。
突出した外れ値	一つの標本の中に、他の要素とは極端に違う数値がある場合それを突出した外れ値という。突出した外れ値があると、平均値が実態と合わない数値になりやすい。
中央値	一つの標本に属す要素を一定の基準で並べたとき、中央に位置する数値。
最頻値	一つの標本に属す要素を一定の基準で分類したとき、最も多数の要素が属す分類をいう。たとえば一日の来店客を年齢層別に分類したところ 10 代の客が最も多かったとすると、その日の来店客の年齢別最頻値は 10 代となる。
偏差	標本の中の一つの要素が、平均値からどれだけずれているかを示す値
標準偏差	一つの標本の中の要素がどれだけバラツキがあるかを表す指標。「平均値±標準偏差」の範囲の偏差は珍しくないと判断できる。σ と書いて「シグマ」とも呼ぶ。
二項分布	コイントスのよううに「1 回の試行で結果が 2 種類しかない操作」を複数回繰り返したとき、一方の結果が何回出るかを示す分布のこと。簡単な式で計算できる。
正規分布	動物の身長や体重、雨粒の大きさなど、自然界のランダムな現象によく見られるデータのバラツキ方。

シグマ区間	正規分布の中央から標準偏差の 1 倍、2 倍、3 倍の範囲をそれぞれ 1σ、2σ、3σ区間と呼び、1σ区間に約 68%、2σ区間に約 95%、3σ区間に約 99.7%のデータが含まれる
フタコブラクダ型データ	正規分布と違って、平均値や中央値付近の要素が少なく、その上下に外れたところに 2 つの山があるタイプの分布をするデータのこと。平均値や中央値ではその集合の実態と合わないため販促企画を立てるときなどに注意が必要。

4.26 確認問題

【問 1】ある店舗での顧客別の購入金額の表から、最大値・最小値・平均値・中央値・最頻値・標準偏差を求めなさい。小数点以下は四捨五入し、最頻値は 1000 円単位の階級で「○千円台」のように解答すること。

顧客番号	購入額
1	700
2	1300
3	1200
4	16900
5	2500
6	5100
7	2800
8	4200
9	3400
10	2400
11	6800

【問 2】広告 DM（ダイレクトメール）を 1000 人の顧客に郵送するとして、以下のような条件だった場合、そのうち商品の購入に至る顧客は何人いると考えられるか計算しなさい

【条件】

（A）発送した DM が顧客の家に届く確率は 80%とする（名簿が古い、住所を間違えている、郵便事故などの理由で届かないものが 20%あると想定）

（B）届いた DM を顧客が開封する確率を 25%とする

（C）開封した DM を読んだ顧客が来店する確率を 20%とする

（D）来店した顧客が購入する確率を 40%とする

■確認問題解答

【問１解答】

最大値：16900

最小値：700

平均値：4300

中央値：2800

最頻値：2 千円台

標準偏差：4339

【問２解答】

16 人

1 人が購入する確率は 0.8×0.25×0.2×0.4=0.016

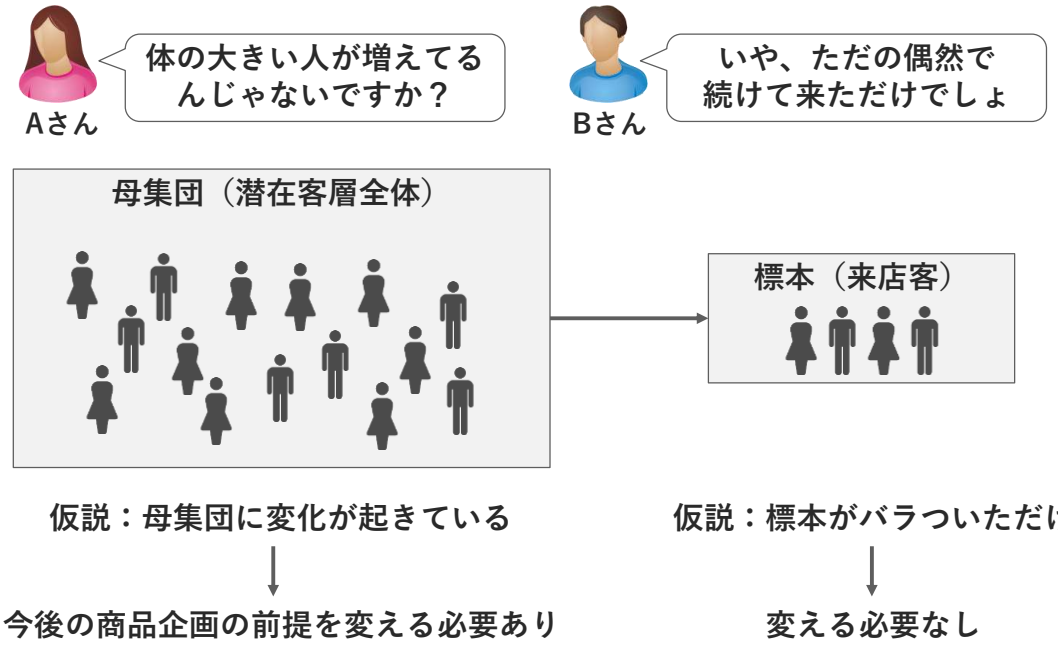
同じＤＭを 1000 人に発送するので 1000×0.016=16

したがって 16 人が答えとなる

5 仕事と集合 Ⅱ

5.1　「珍しい出来事」をどう解釈する？

ある洋服店である週に、ふだんはなかなか売れない商品(例：XXLサイズのシャツ)が連続して売れました。これは何を意味するのでしょうか？



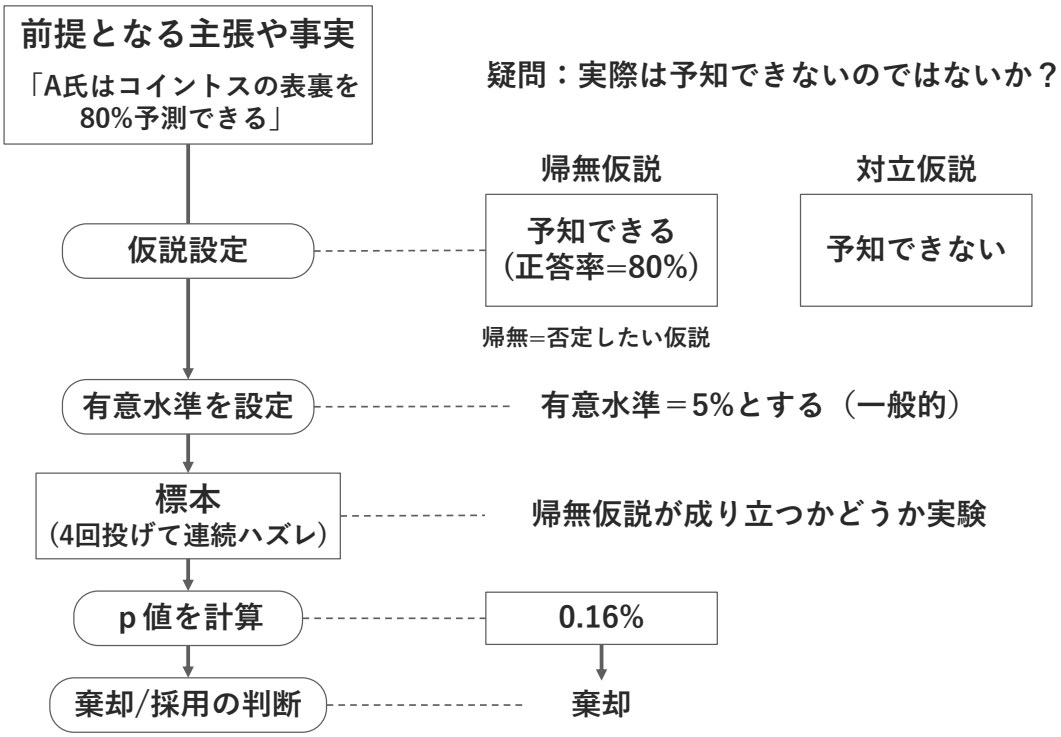
あるときある洋服店で、ふだんはなかなか売れない XXL サイズのシャツが連続して売れたとします。これに対して A さんは「体の大きい人が増えているんじゃないか」と考えたのに対して B さんは「ただの偶然でしょ」と意見が分かれました。

お店への来店客というのは、「母集団」に対する「標本」と考えることができます。母集団は「そのお店に来る可能性がある客層全体」です。通常はその店の近くに住んでいる、通勤している、近くに遊びに来ることがある人が該当します。一方、標本はその週に実際にその店に来た客です。

A さんの意見が正しいなら、母集団に変化が起きているということですから、今後の商品企画の前提を変えて「大きなサイズを増やす」などの手を打つ必要があります。B さんの意見が正しいなら、特に何も変える必要はないでしょう。

こんなときに、ただの「カン」ではなくデータを分析してどちらの意見がより「確からしい」かを判断することはできるでしょうか？　このような目的に対して使えるのが「仮説検定」という統計的なデータ処理の方法です。「検定」とは「正しいかどうかを確かめる」ことをいうので、「仮説検定」とは「仮説が正しいかどうかを確かめる」一般的な方法です。

5.2　仮説検定の基本



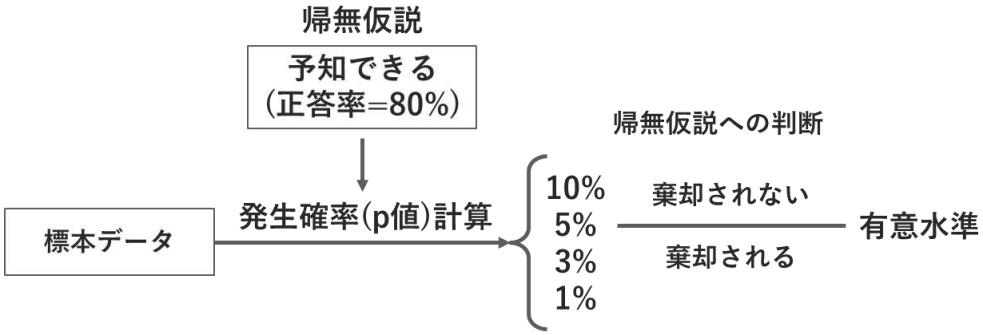
仮説検定の基本を知るため、わかりやすい極端な例から考えましょう。ある人（A 氏）が「私はコイントスの表裏を 80%予測できる」と予知能力があることを主張していたとします。当然、誰もが「出来るわけないじゃないか」と疑問を持つはずですので、ここで仮説検定を行います。

仮説検定を行う対象は何らかの「前提となる主張や事実」です。たとえば「私はコイントスの表裏を 80%予測できる」というのは A 氏の主張です。他には「文部科学省によると、男子の小学校卒業時の平均身長は XXcm、標準偏差は YYcm である」というようなものなら「過去に計測された事実」です。過去の事実は現在までの間に変わることがあるので、いったんは「事実」と確かめられたものでも、時間が経つなどして状況が変われば仮説検定の対象になります。

まずはその「主張や事実」に対して仮説を設定します。このとき設定する仮説には「帰無（きむ）仮説」と「対立（たいりつ）仮説」の 2 種類があります。「帰無」というのは「おそらく間違っている」という意味あいでとらえましょう。つまり「無に帰る」、ざっくばらんに言えば「無い無い、それ絶対ありえない」というような種類の仮説が帰無仮説です。したがって「コイントスを 80%予知できる」というような荒唐無稽な仮説が該当します。それに対する「おそらく正しい、普通の意見」を「対立仮説」と言います。今回は「予知なんかできるわけがない」という、いかにも普通の常識的な意見が対立仮説です。この段階ではどちらもただの「仮説」であって、正しいとも間違っているとも言えません。

今回は極端な例を出しているので「帰無と対立」の違いがわかりやすいですが、現実に仮説検定を行う場面ではこの差がはっきりしないときもあります。たとえば「前提となる主張」が「地球は温暖化している」というものだったらどうでしょうか。これに対しては人によって「している」「していない」と意見が分かれます。基本的に、帰無仮説には「自分が否定したい仮説」を選びます。自分が「地球は温暖化している」という意見を持っているなら「温暖化していない」を帰無仮説に選び、逆の意見なら逆にします。したがって意見が分かれる問題の場合は人によって帰無と対立の選択が逆になります。

次に「有意水準」を設定します。有意水準の位置づけは下記の図で説明します。



帰無仮説を立てた後、「標本データ」の「発生確率（p 値)」を計算します。p 値は「帰無仮説が正しいと仮定したとき、標本データの事象が発生する確率」のことです。たとえば「正答率=80%」という帰無仮説のもとで 1 回コイントスを行い、予測がハズレた場合の p 値は 20%（80%正答できるはずなのにハズレたので、残り 20%に該当する事態が起きた、と考える)。2 回続けてハズレた場合は 20%の 2 乗ですから p 値=4%です。これに対して、同じ 1 回ハズレ、2 回ハズレでも、帰無仮説が「正答率=60%」というものだった場合の p 値はそれぞれ 40%、16%です。つまり標本データが同じでも、帰無仮説が異なれば発生確率（p 値）の計算結果は異なります。

標本が同じでも帰無仮説が違えばp値は異なる

標本データ	帰無仮説 正答率=80%	帰無仮説 正答率=60%
1回ハズレ	p値=20%	p値=40%
2回連続ハズレ	p値=4%	p値=16%

ある帰無仮説をもとに具体的な標本データの p 値を計算したあと、それが「有意水準」より高いか低いかを判断します。もし有意水準より低ければ、

帰無仮説が正しいという仮定の下では、非常に低い確率でしか発生しない
標本が得られてしまったので、おそらく帰無仮説は正しくない

というロジックで帰無仮説を「棄却」します。「棄却」とは捨て去ることですので、帰無仮説は「無かったこと」になります。

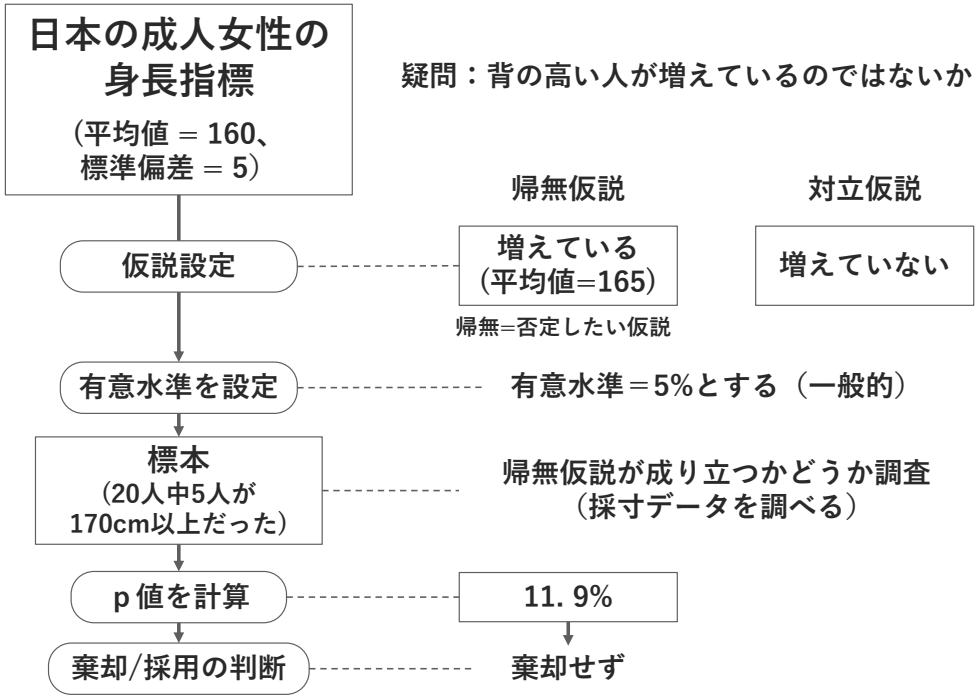
このように「帰無仮説を捨てるかどうか」を左右する重要な数字が「有意水準」です。一般的には5%に設定されますが、人命が関わるような重大な問題については1%がよく使われます。有意水準は勝手に決めてよいものではなく、問題の性質に応じて適切に選ばなければなりません。

> 【注意！】p 値が有意水準より高ければ帰無仮説は棄却されませんが、これは棄却するには証拠が不十分というだけで、「帰無仮説が正しい」という意味ではありません。たとえば p 値=10%つまり 10 回に 1 回というのは偶然でも十分起こりうる数字です。「俺はコイントスで必ず表を出せる」と宣言した人がいたとして、その人が実際に 3 回連続して表を出したとしても「必ず表を出せる」と信じる人はあまりいないでしょう。3 回連続偶然に表が出る確率は 1/8 あるので、この程度では「必ず出せる」能力の証明にはなりません。4 回連続でもまだ不十分で、多くの人は 5 回連続つまり 1/32 ぐらいになってようやく「ひょっとして本当かも？」と信じ始めます。

さて、有意水準を設定した上で標本を採るのは、「帰無仮説が成り立つかどうか」を実験することです。4 回連続でハズレた場合の p 値は 0.16%となり、有意水準 5%より小さいので帰無仮説は棄却され、対立仮説が採用されます。

以上が「仮説検定」の基本ですが、次はもう少し複雑で現実的なケースを考えましょう。

5.3　ある洋服店での仮説検定



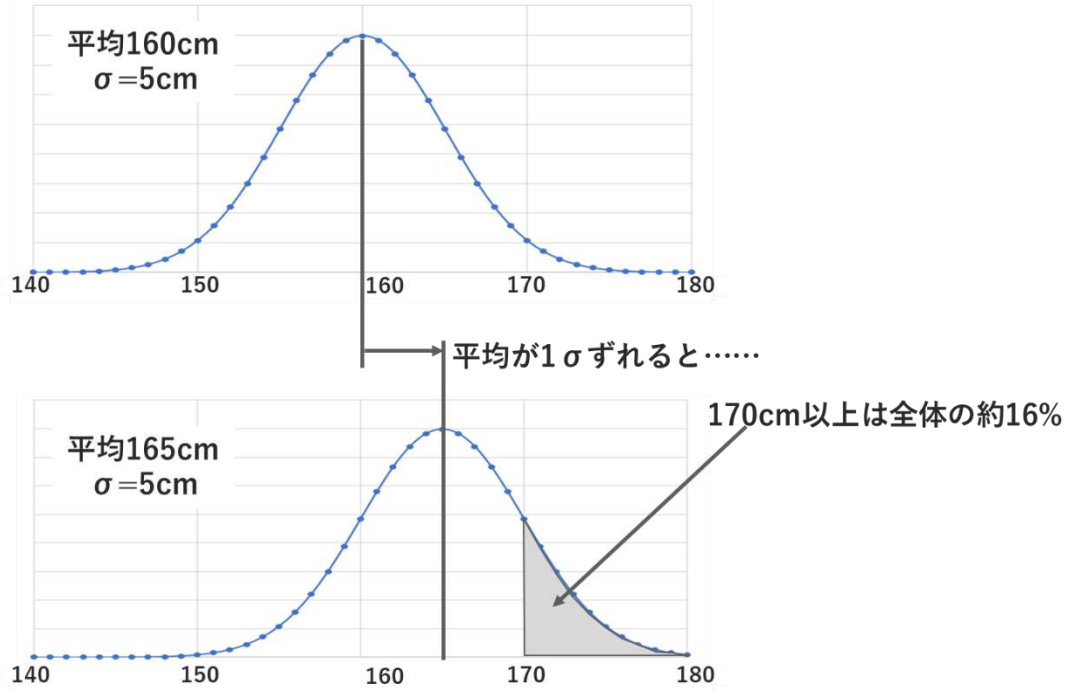
ある洋服店で短期間に大きなサイズが売れたために、「背の高い人が増えているのではないか？」という仮説を立てた場合の仮説検定の流れです。

まず、「前提となる主張や事実」としては、「日本の成人女性の身長指標」として一般的な数字を使います。ここでは、話を単純化するために平均値=160cm、標準偏差=5cm としておきます。一方、疑問は「背の高い人が増えているのではないか？」というものでしたが、成人の身長はそう簡単に伸び縮みするものではないので、ここでは「いやそれはただの偶然で、実際は増えていないだろう」と反論する立場で考えてみましょう。そこで、「増えていない」を対立仮説、「増えている」を帰無仮説とします。

「増えている」という帰無仮説の枠内で「平均値=165」と書かれているのに注目しましょう。増えているとしてそれは何 cm なのかというこの数字を決めないことには p 値の計算ができないので、決めておく必要があります。ここでは+5cm つまり 1σ分増やしていますが、σ単位でなければならないわけではなく、p 値の計算ができさえすれば何でもかまいません。

有意水準は一般的な 5%に設定しましょう。標本として過去 1 週間の採寸データを調べたところ、20 人中 5 人が 170cm 以上だったとしましょう。帰無仮説（平均値=165cm、標準偏差は変わらず 5cm）が正しいとしたら、標本がこの分布を示す確率（p 値）はいくらになるでしょうか？

平均値が 160cm から 165cm へと 1σずれると、身長 170cm 以上は全体の約 16%になります。これは、1σ区間=160~170 が 68%で、170cm 以上はその残り 32%のうち右側（高身長側）半分だからです。

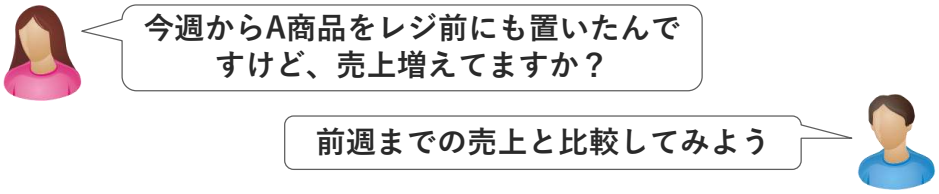


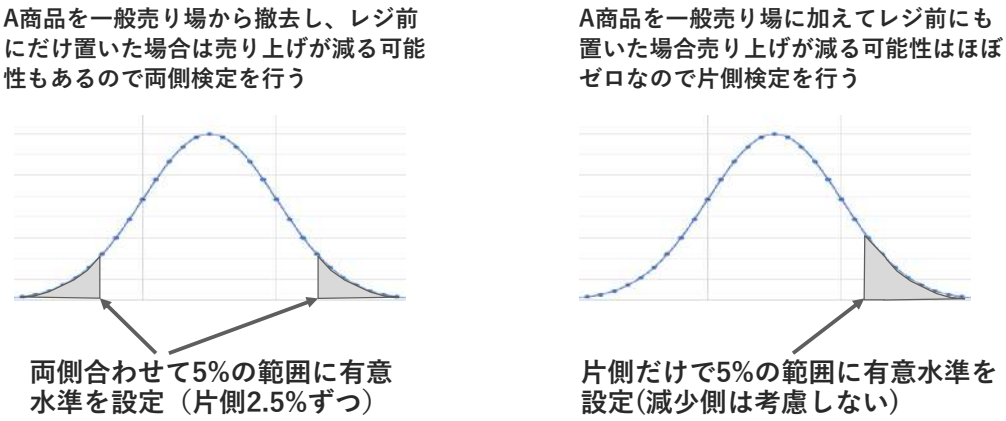
それがわかれば、20 人のうち 5 人が 170cm 以上である確率は二項分布の式で計算できます。「1 人の客」を 1 回の試行と考え、1 回あたり 16%の確率で真になる試行を 20 回繰り返して 5 回真になると考えればよいわけです。これを計算すると p 値は 11.9%で、優位水準 5%よりも大きいため、帰無仮説「平均身長が 165cm に増えている」は棄却できません。

ちなみに帰無仮説と対立仮説を逆にして、「平均身長は 160cm のまま、増えていない」を帰無仮説にして p 値を計算すると 0.01%となり、優位水準より極めて小さいため棄却されます。したがって、この検定によれば、「20 人中 5 人が 170cm 以上」であった場合、その店の客層の平均身長は具体的な数値は不明ながら 160 よりは上である可能性が高いと考えられます。

なお、実際には「標本」に偏りがある状態で仮説検定を行うと非現実的な答えが出がちなので注意が必要です。たとえば標本が「ある店のある 1 週間の来店客」のようなデータだと、たまたまその週に近所のホテルに泊まっていたバレーボールの選手団が買いに来ていた、といったバイアスを排除できません。

5.4　片側検定と両側検定





スーパーマーケットなどの業態では、「レジ前商品」というジャンルがあります。店内を回って目的の商品を探し終えた後、会計のためレジ前に並んでいるときに目に留まると衝動買いされやすい商品をレジ前に置きます。通常、レジ前商品はレジ前「にも」置くので、それで売上が下がることはほとんど考えられません。このような場合は減少側を考慮に入れる必要がないので、「売上が上がったかどうか？」と、片側だけを判定します。

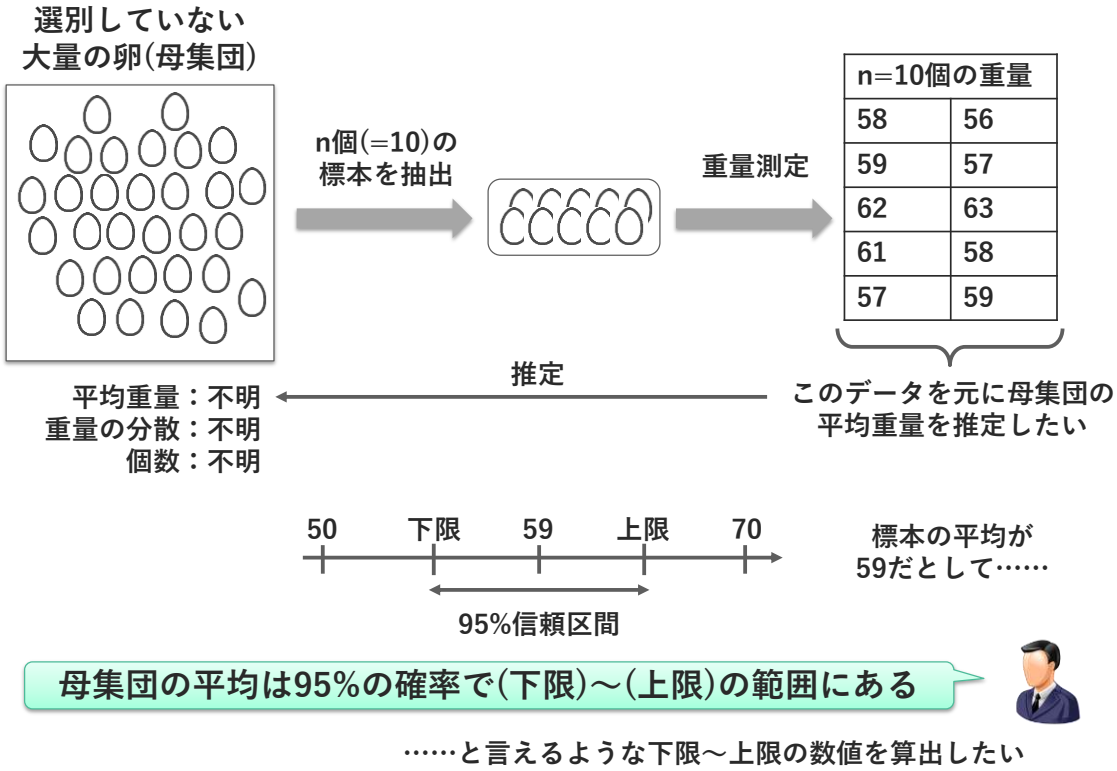
一方、商品を一般売り場から撤去してレジ前に置く場合はかえって売上が下がることもあり得ます。そのような場合は両側を判定しなければなりません。

【両側検定】
正規分布の下位に外れた場合、上位に外れた場合の両方に有意水準を設定する。

【片側検定】
正規分布の下位・上位のどちらか片方にだけ有意水準を設定する。

有意水準は通常 5%や 1%で設定します。よく使われる「5%」というのは両側検定の場合で平均値 $\pm 2\sigma$ を外れた範囲です。片側検定の場合、同じ 5%でも片側検定だと片側だけで 5%の範囲を有意と判定するため、両側検定よりも甘い有意条件を設定していることになるため注意してください。

5.5　標本のデータを元に母平均を推定したい



鶏卵はよく使われる食材で、スーパーなどでは重量別にM, Ｌなどのサイズで分類されて売られています。1 個あたりの重さによって値段が違うので、重さは重要な情報です。通常は重量別に選別して出荷しますが、仮に選別していない大量の卵があるとしましょう。これを母集団とします。

選別していないので重量は正規分布に従うと考えられますが、個数も平均値も分散も不明です。この母集団の平均重量を知りたい、しかし大量なのですべての卵を個別に測定したり個数を数えるのは手間がかかりすぎるためやりたくないとします。このような場合、標本調査によって母集団の平均値を一定の精度で推定することは可能でしょうか？

直観的には、標本の平均値は母集団の平均値に近いと考えられるため、標本の平均をそのまま母集団の平均値とみなしても良さそうにも思えます。しかし、標本抽出にはどうしてもある程度の偏りが出るため、単純に「標本平均＝母集団の平均」と考えるのは危険です。

たとえば 10 個の卵を標本として抽出したところ、その平均値が 59 グラムだったとすると、

A：母集団の平均は 58～60 グラムの間にある
B：母集団の平均は 57～61 グラムの間にある

の 2 つの主張のどちらもある程度正しそうですが、ではそれぞれどの程度正しいのでしょうか？
B のほうが A よりも範囲が広いので、B のほうが正しい確率は高そうです。しかし、範囲を広げたほうが当たるからといって、たとえば

C：母集団の平均は 50〜70 グラムの間にある

のようにむやみに範囲を広げても意味がありません。一般的な鶏卵の重量は 60 グラム前後なので、母集団の平均が 50 グラム以下や 70 グラム以上になることはあり得ないため、C のような主張をすれば必ず当たりますが実際には何も言っていないのと同じで何の役にも立ちません。

したがって、目標にしたいのは

D：母集団の平均は 95%の確率で（下限)〜(上限)の範囲にある

と言えるような下限、上限の数値を算出することです。「95%の確率で」のように「確からしさ」を明確にすることが重要で、A や B の主張についても

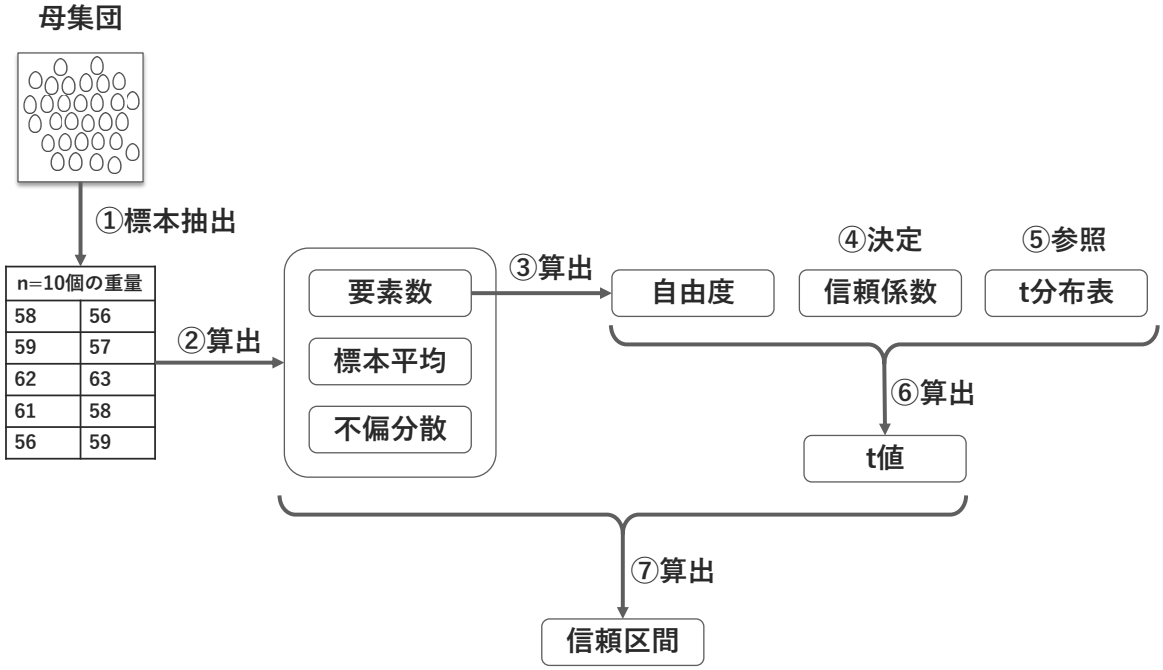
A：母集団の平均は 70%の確率で 58〜60 グラムの間にある
B：母集団の平均は 90%の確率で 57〜61 グラムの間にある

のように「確からしさ」が明示されていれば、必要な「確からしさ」のレベルに応じてどちらかの主張を採用して次の行動を決定できます。

以上のような「確からしさ」のことを**信頼係数**と言い、統計学では通常の問題については 95%、人の命に関わったり大きな損失を招くような重大な問題については 99%を使うのが一般的です。また、信頼係数を明示して示された下限〜上限のことを「**信頼区間**」と言います。たとえば「平均値の 95%信頼区間は 57〜61 グラムである」と言えば、95%の確率で 57〜61 グラムの範囲に平均値があることを示します。

では、次項で標本データを元に母集団の平均値の信頼区間を算出するまでの大まかな流れを説明します。

5.6　標本抽出から信頼区間算出までの流れ



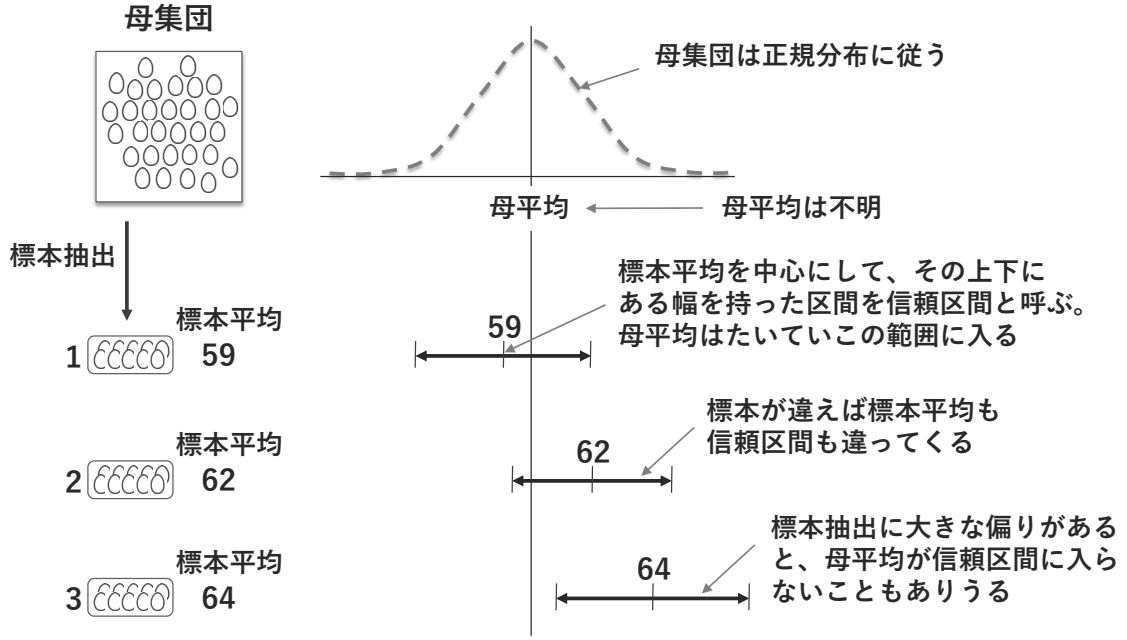
ここでは、信頼区間算出までのおおまかな流れを示します。全体像を示すために、まだ説明していない用語/概念も使用しているので、分からない部分にはあまりこだわらず、後続のページを読んでからもう一度このページに戻って読み直してみてください。

まず母集団から①標本を抽出します。その標本を調べて②要素数、標本平均、不偏分散を算出します。要素数というのは 1 つの標本に含まれるデータ数のことです。不偏分散については後述します。

次に要素数を元に③自由度を算出します。といってもこれは単に要素数から 1 を引くだけです。
④信頼係数は任意に決定します。一般的には 95%や 99%を使用します。
⑤t 分布表は t 値と呼ばれる統計量をあらかじめ計算した表のことで、既存の表を参照します。
⑥自由度、信頼係数、t 分布表を元に t 値を算出します。
⑦t 値と要素数、標本平均、不偏分散を元に信頼区間を算出します。

大まかに以上のような流れで母集団の平均値の信頼区間を算出することができます。
詳細は次項以降に記述します。

5.7　母平均・標本平均と信頼区間



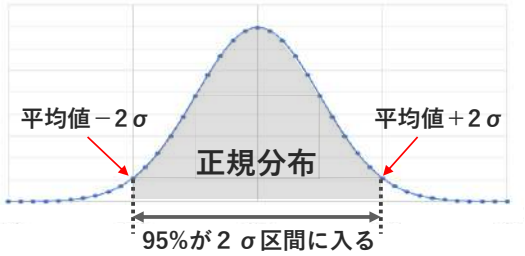
$$\text{信頼区間} = \text{標本平均} \pm \text{t値} \times \sqrt{\frac{\text{不偏分散}}{n}}$$

母平均と標本平均、信頼区間の関係を詳しく見てみましょう。正規分布に従う母集団があり、その平均（母平均）は分からないとします。そこで標本を抽出して平均を調べれば、母平均はその標本平均を中心にして上下にある幅を持った区間内にある確率が高いと考えられます。この区間をその確率と合わせて「信頼区間」と呼びます。

当然、標本を何度も取るとその都度標本平均は違ってきます。上図では標本を 3 回とったら標本平均が 59，62，64 と変わった例を示しています。1 回目と 2 回目では標本平均が違い信頼区間も違っていますが、いずれにしても母平均は信頼区間の範囲内にあります。3 回目は標本抽出に大きな偏りがあって母平均が信頼区間に入らなかった例です。このように標本調査による母平均の推定が外れる場合もありますが、それは標本抽出にたまたま偏りが出た場合であり、そのようなケースは少ないのでたいていは信頼できる、という考え方をします。

そこで問題は、「たいていは信頼できる」という「上下にある幅をもった信頼区間」を具体的にどのように決めるかです。そのために、標本平均に加えて t 値(t 分布をもとに決まる値)、不偏分散、n という数値を使って上図末尾のような式で計算します。n は標本の要素数で、t 値と不偏分散については次項で説明します。

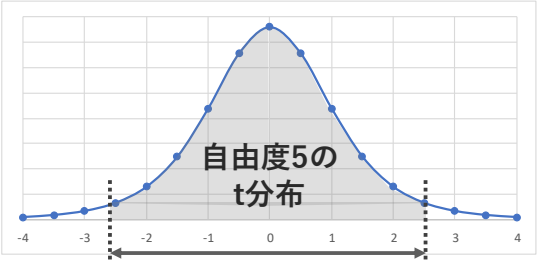
5.8　正規分布の標準偏差とｔ分布のｔ値



正規分布するデータは、95%が
平均値±2σの範囲に入る

式の構造は似ているが
細部に差がある

$$\text{標本平均} \pm t\text{値} \times \sqrt{\frac{\text{不偏分散}}{n}}$$



95%が入る区間のt値は
−2.57〜+2.57

自由度が変わるとこの数値も変わる

t分布の形は「自由度」によって変わる。

自由度 = 要素数-1

自由度が小さいt分布は正規分布よりも裾野が高く、95%区間のt値は2よりも大きい

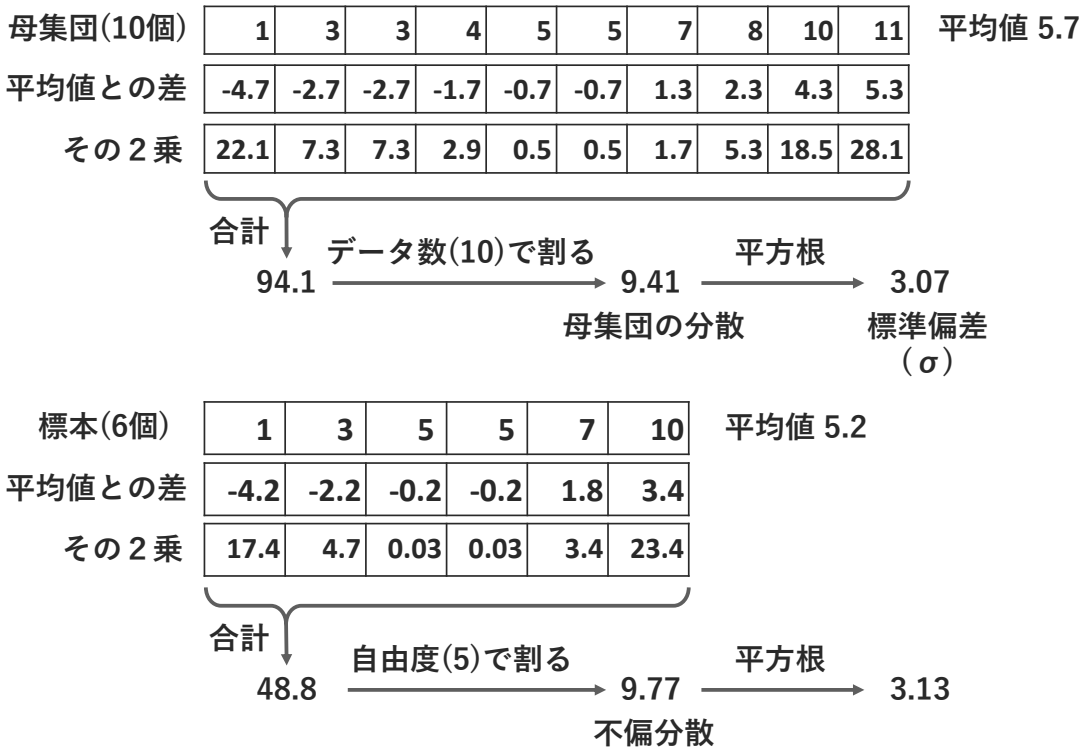
既に 4 章で触れた内容ですが、正規分布するデータは 95%が　**平均値±2σ**　の範囲に入ります。一方、t 分布を使って信頼区間を計算する式は

$$\text{標本平均} \pm t\text{値} \times \sqrt{\frac{\text{不偏分散}}{n}}$$

です。±2σ の 2 が t 値に変わり、σ が　不偏分散／n　の平方根に変わった形で、式の構造はよく似ています。ここで t 値を求めるために使うのが t 分布で、その形は要素数－１で計算される「自由度」によって決まり、自由度が小さいほど（つまり要素数が少ないほど）裾野が高く、自由度が大きいほど（要素数が多いほど）正規分布に近づきます。上図に掲載した自由度 5 の例では 95%区間の t 値は-2.57〜+2.57 です。正規分布の 95%区間は 2σ でしたが、t 分布は正規分布より裾野が高いため 95%区間の t 値は一般に 2 より大きな数値であり、自由度が大きくなるほど 2 に近づきます。

なお、自由度が 30 以上の t 分布はほぼ正規分布と一致しますが、もともと t 分布は「調査の手間を減らすため、少ない標本で母集団の平均値を推定したい」というときに使うので自由度をむやみに増やすことは意味がなく、かといって少なすぎると信頼区間が広がりすぎて推定の精度が悪くなるため、実際に t 分布を使うのは自由度(要素数-1)が 5〜20 程度の場合がほとんどです。

5.9　母集団の分散と不偏分散



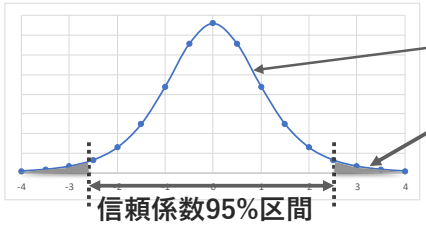
「不偏分散」または「標本不偏分散」は、標本の分散をもとにして母集団の分散(母分散)を推定した量のことです。上図にその算出の例を示しました。

母集団として 10 個の数値があるとします（標本調査が必要な場合、母集団ははるかに大きな数値になるのが普通ですがここでは例示のため小さな数にしてあります)。その平均値 5.7 との差をすべての要素について算出し、それを 2 乗して合計した数 94.1 をデータ数 10 で割って得た 9.41 が母分散であり、これの平方根を取ると標準偏差 σ になります。

一方、その母集団から 6 個の標本を取り、以下同じように標本の平均値を計算し、平均値との差を出し 2 乗して合計すると 48.8 という数値が出ます。これを自由度 5 で割ったものが不偏分散（標本不偏分散）9.77 です。標本の要素数は 6 ですが「自由度」はそこから 1 を引いて 5 になることに注意してください。こうして出した不偏分散 9.77 は母分散 9.41 にかなり近い数値です。もし割る数を自由度 5 ではなく要素数 6 にすると 8.14 になり、母分散との差が大きくなります。

理論的な説明は省略しますが、標本をもとにして母集団の分散を推定する場合は要素数から 1 を引いた「自由度」で割ることを覚えておいてください。これによって得られる、元の母集団の分散に近い値のことを不偏分散（標本不偏分散）と呼びます。

5.10 t 分布表による t 値の決定



自由度5のt分布のグラフ

信頼区間を外れるのは
両側で5%、片側で2.5%(=0.025)

t分布表

	片側確率（有意水準）				
自由度	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	**2.571**	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

t分布表で片側確率0.025、
自由度5の交点を探して

t値=2.571

を得る

自由度 5 の t 分布において信頼係数 95%に該当する t 値を求める場合を考えましょう。標本平均の上下両側に信頼区間を設定する場合、自由度 5 の t 分布のグラフで信頼係数 95 の区間を外れる範囲(有意水準)は両側を合わせて 5%ですので片側では 2.5%(=0.025)になります。

t 分布表は「自由度」と「片側確率(有意水準)」の組み合わせによる t 値の変化をまとめた表で、この表から片側確率 0.025、自由度 5 の交点を探すと t 値=2.571 が得られます。

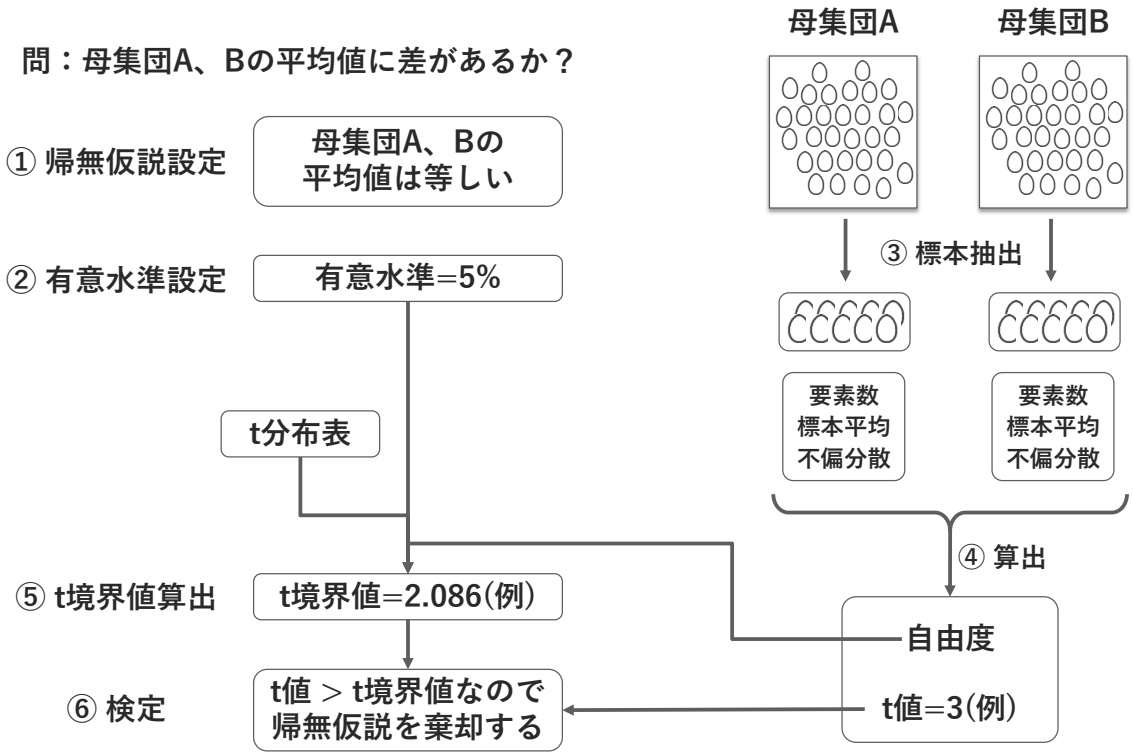
「片側検定と両側検定」の項でも触れたように、信頼区間は標本平均の上下両側に設定する場合もあれば、片側だけ考えればよい場合もあります。片側だけでよい場合は有意水準を半分にしなくてよいので、信頼係数 95%の場合は片側確率 5%(=0.05)の列で t 値を探します。

こうして t 値がわかれば、あとは標本平均、不偏分散、要素数(n)と合わせて

$$\text{標本平均} \pm \text{t値} \times \sqrt{\frac{\text{不偏分散}}{n}}$$

を計算すれば母平均の信頼区間を算出できます。

5.11 対応のないデータの t 検定



t 分布の考え方を使って、２つの母集団からの標本によって母集団の比較をすることができます。

正規分布に従う独立した母集団 A、B があるとします。この両者の平均値がわからず、分散も違う可能性があるときに、標本調査によってこの両者の平均値に差があるかどうかを判断したいとしましょう。具体例としては「２つの養鶏場で生産された卵の平均重量に差があるか？」「２つの学校の生徒の平均身長に差があるか？」などを調べる場面をイメージしてください。どちらの場合も、母平均・母分散の両方とも一致しない可能性があります。

既に説明した「t 分布を使って、標本のデータから母集団の平均値を推定する」方法を使えばそれぞれの母平均の信頼区間を算出することはできます。しかし仮に２つの母平均が実際には完全に一致していた場合でも、標本から算出した信頼区間はほとんどの場合一致せず、ある程度ズレてしまいます。では、どの程度のズレなら「母平均に差がある／ない」と判断できるのでしょうか？

このような判断をしたい場合に使うのが、やはり t 分布を使った「対応のないデータの t 検定」という方法です。具体的な手順を大まかに示したのが上図です。

まず、①帰無仮説を設定します。仮に「母集団 A、B の平均値は等しい」としておきましょう。
次に②有意水準を設定します。仮説検定では一般的に 5%や 1%を使います。帰無仮説が正しいと仮

定したときに、標本で得られた値の組み合わせが発生する確率が有意水準以下しかなければ、帰無仮説を棄却する、つまり「母集団 A、B の平均値には差がある」という対立仮説を採択する基準がこの有意水準です。

次いで母集団 A、B のそれぞれから③標本を抽出します。標本の要素数は一致していなくてもかまいません。標本を調べるとそれぞれの要素数、標本平均、不偏分散がわかります。そして 2 つの標本から得たそれらのデータを元に④自由度と t 値を算出します。t 値は次のような式で算出できます。

$$t\text{値} = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\overline{x_1}, \overline{x_2}$ **標本の平均値**

s_1^2, s_2^2 **標本の不偏分散**

n_1, n_2 **標本の要素数**

自由度は標本が 1 つの場合は単純に要素数-1 で計算できましたが、標本が 2 つある場合は複雑な計算が必要になり、本書の範囲を超えるため計算式は省略します。実務的には t 値、自由度のどちらも表計算ソフトの分析ツール等で簡単に計算できるため、計算式を覚えておく必要はありません。

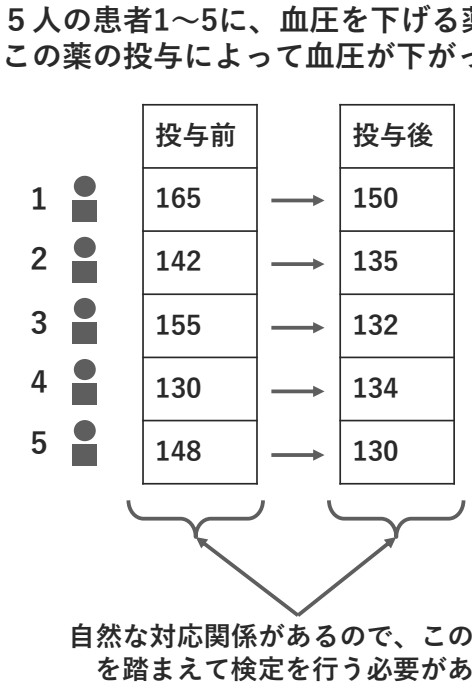
そうして算出した自由度と有意条件、t 分布表を用いて⑤ t 境界値を算出します。t 境界値とは、有意水準の境界に該当する t 値のことです。

最後に t 境界値と t 値を比較して⑥検定を行います。上図例の場合、t 境界値=2.086 に対して t 値=3 となり、t 値＞t 境界値であるため帰無仮説は棄却されます。つまり「母集団 A、B の平均値には差がある」と言えます。

以上のように、2 標本が属す母集団の平均値を比較する方法を「対応のない 2 標本 t 検定」と言い、その中でも今回のように「母集団の分散が等しいと仮定できない」場合に採用する手順を「Welch の t 検定」と言います。

「対応のない」とは、２つの標本の間に対応関係がないという意味です。今回冒頭で具体例として挙げた「２つの養鶏場で生産された卵」「２つの学校の生徒の平均身長」はいずれも独立した母集団であり、対応関係はないと言えます。対応関係がある場合の t 検定については次項で説明します。

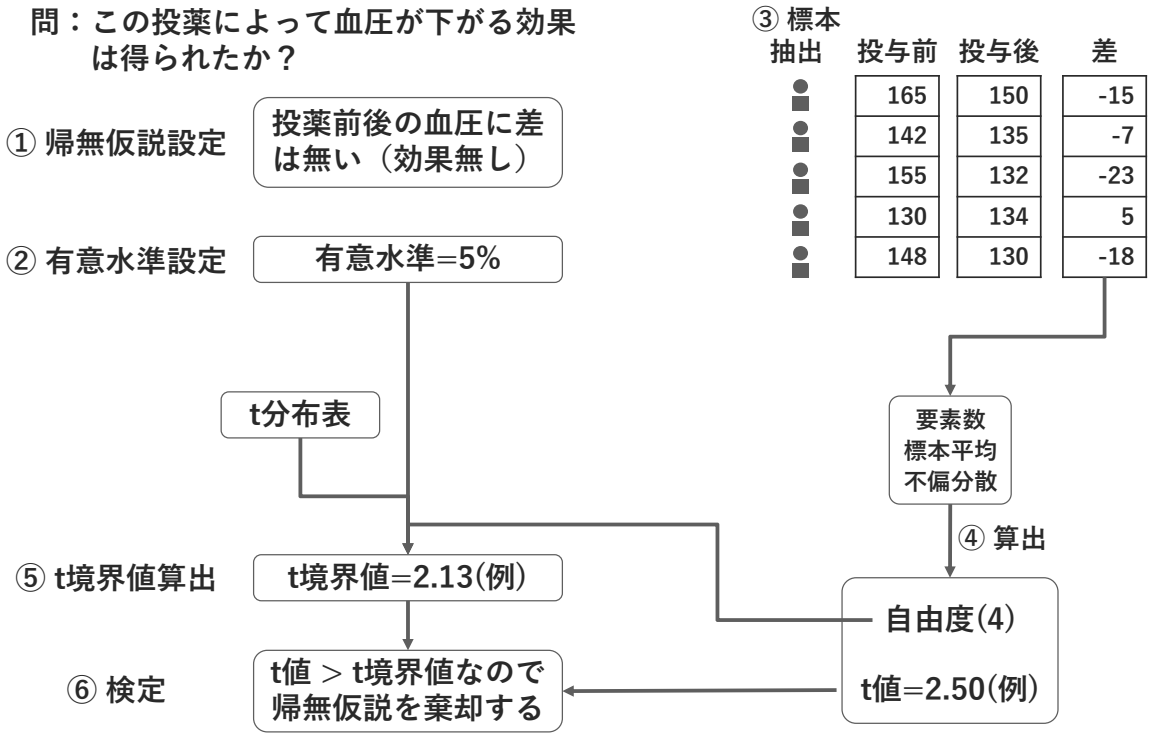
5.12 対応のあるデータとは



たとえば医薬品の効果を判断する場合には、薬剤の投与前と後で指標がどのように変わったかを調べます。この場合、同じ人の投与前と投与後のデータには一対の自然な対応関係があります。このようなデータを、「対応のあるデータ」と言い、検定はその対応関係を踏まえて行わなければなりません。

上図で例示した医薬品の他にも、同じ人に 2 種類の食品を食べ比べて採点してもらうケース、ダイエット法を試す前と後の体重変化を比べるケース、あるスプリントシューズを履いたときと他の製品とで 100 メートル走のタイムの変化を比べるケースなど、さまざまな場面で「対応のあるデータの検定」が使われます。このための検定でも t 分布の考え方を利用します。

5.13 対応のあるデータの t 検定

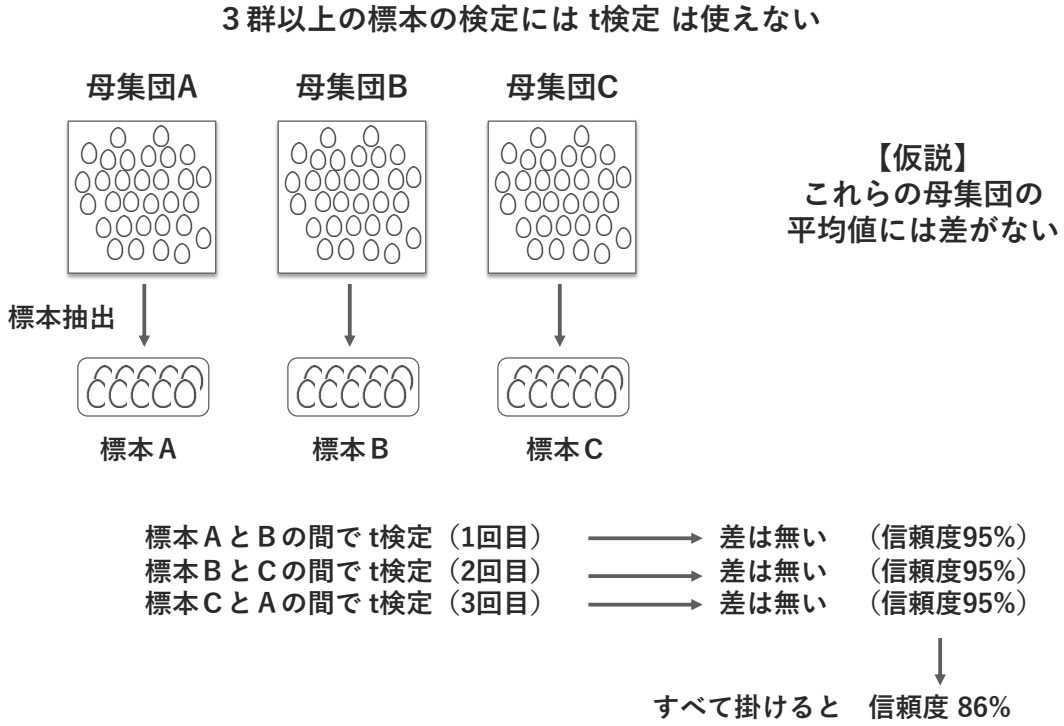


対応のあるデータの t 検定も、基本的な手順は対応の無いデータの t 検定と同じで、まず①帰無仮説と②有意水準を設定します。

次いで③標本を抽出して、例えば「投与前と投与後の血圧の差」のように差分を出してその要素数・標本平均・不偏分散を算出します。「対応のないデータ」の場合は 2 つの標本から複雑な計算でこれらの数値を出しましたが、対応のあるデータはいったん差分をとることで標本が 1 つになるため、この計算が単純になります。

後は対応のないデータの t 検定と同じで、④自由度、t 値を算出し、⑤t 境界値を算出して、⑥検定を行います。上図例の場合、t 境界値=2.13（血圧が下がったかどうかだけを検定するので片側検定の境界値を使用）に対して t 値=2.50 と t 値＞t 境界値のため帰無仮説は棄却されます。つまり「投薬前後の血圧には差がある（効果があった)」と言えます。

5.14 ３群以上の標本を検定するには？



たとえば食品会社が新商品を開発して、Ａ県、Ｂ県、Ｃ県という３つの地域で発売しようとしたとします。しかし、味の好みには地域差があるため、３県の住人が同じように満足してくれるとは限りません。実際、日本全国で同じ名前で販売されている食品でも、関西版と関東版では味付けを変えているようなケースがあります。

そこで、３県の住人に試食してもらって満足度のデータを集めたとします。県が違っても満足度が同じなら同じ味付けでかまいませんが、もし県によって差があるなら味付けを変えなければならないかもしれません。

そのためにはまず「３つの母集団（３県）から集めた３群の標本の平均値に差があるのかどうか？」を判断する必要がありますが、どうすればそれができるのでしょうか？

これは一見すると「対応のないデータの平均値に差があるかどうかを調べる」問題ですから、ｔ検定で可能なように見えます。つまり、標本ＡとＢの間でｔ検定を行い、以下同様に標本ＢとＣ、標本ＣとＡの間で同じように、合計３回のｔ検定を行って、すべて「差が無い」という結果であれば「３群の標本の平均値に差は無い」と判断できるだろう……というわけです。しかし、実はこの方法では信頼性の高い結果は得られません。

なぜ「t 検定を 3 回行う」方法ではダメなのかを理解するために、今までも「信頼係数」や「信頼区間」という用語の中に出てきた「信頼」という言葉の意味を改めて考えましょう。たとえば、

「標本 A と B の平均値には差が無い」という仮説の信頼度は 95%である

という場合、言い換えれば

「標本 A と B の平均値には差が無い」という仮説は 5%の確率で間違っている

のです。それでも、人の命が関わらないような分野なら厳密な判断は必要ないので、「間違いの確率が 5%」というのは十分低い確率であり実用的に問題ないため、有意水準 5%がよく使われます。

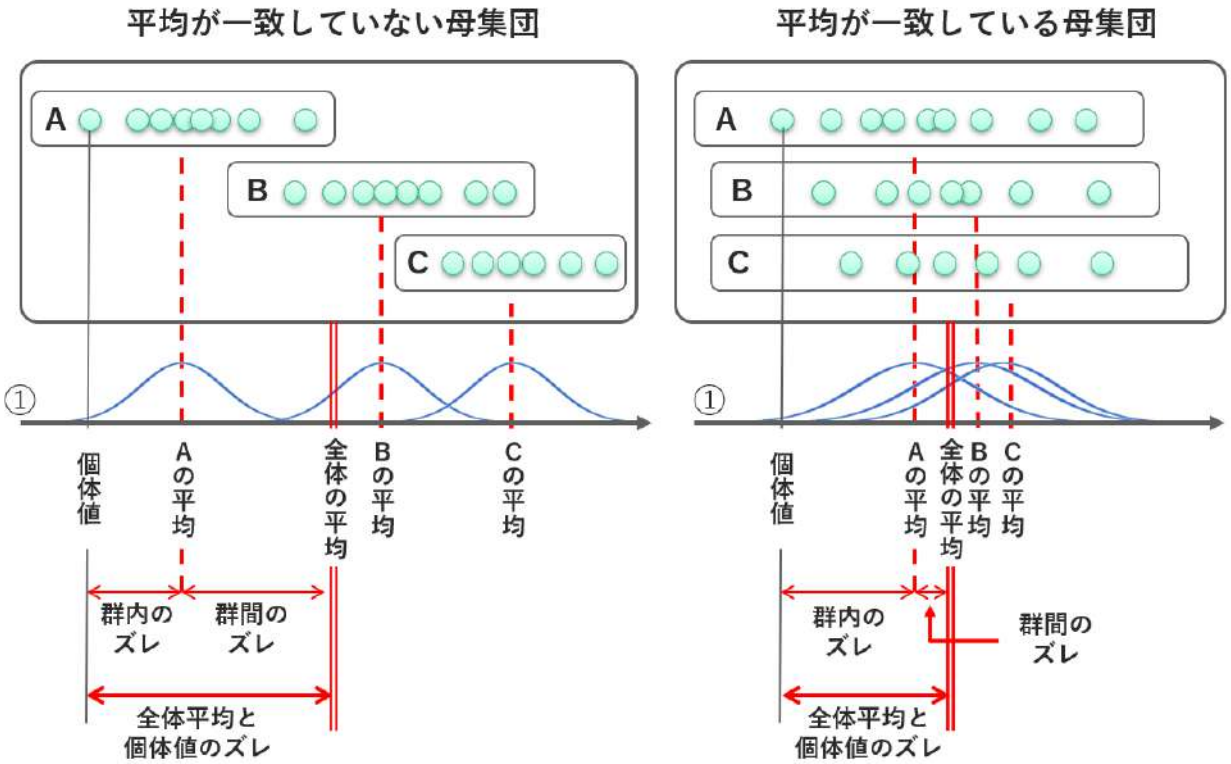
しかし、「5%の確率で間違っている検定を 3 回重ねて行った場合」は話が違ってきます。3 回行って一度も間違えない確率を計算すると

$$95\% \times 95\% \times 95\% \fallingdotseq 86\%$$

となり、「14%の確率で間違う」ことになります。これでは「十分低い確率」とは言えません。しかも、実際にこの種の調査を行う場面を考えると、「3 県」で済むとは限らず、4 県 5 県 6 県……と標本数が際限なく増える可能性があり、そのうちの 2 県の組み合わせの数はそれ以上に増えます。増えるとその分だけ「95%の確率」を重ねて掛け算することになり、「一度も間違えない確率」はどんどん下がってしまいます。

したがって、3 群以上の標本について「平均値が一致しているかどうか」を判断する用途には t 検定は使えません。それでは、どんな方法を使えば良いのでしょうか？

5.15 群内のズレと群間のズレに注目する



３群以上の標本について平均値が一致しているかどうかを判断するための基本的な考え方は、「群内のズレと群間のズレに注目する」というものです。

上図左側は「平均値が一致していない母集団」、右側は「平均値が一致している母集団」から取った標本のイメージです。平均値が一致していない母集団から取った標本のデータを①に示す数直線上にプロットするとＡＢＣのデータがそれぞれ独立した山を描くのに対して、平均値が一致している母集団から取った標本では山の重なりが大きいことがわかります。

ここで、「群間のズレ」と「群内のズレ」に注目しましょう。それぞれ、

群間のズレ＝標本の平均値　－　全体の平均値
群内のズレ＝個体値　－　標本の平均値

で計算されます。「全体の平均値」は、ＡＢＣの３つの標本をひとまとめにして平均値を出したものであり、「標本の平均値」はＡＢＣそれぞれの標本ごとに平均値を出したものです。「群間のズレ」は、標本Ａ群、Ｂ群、Ｃ群をそれぞれ「ひとまとまり」としてとらえて、「群どうしがどれぐらい離れているか」を表す数値です。一方、「群内のズレ」は、Ａ群ならＡ群内の一つ一つの個体値がＡ群

の平均値からどれだけ離れているかを表します。

それを踏まえて上図の左側と右側を比べると、おおむね

群間のズレ＞群内のズレ　ならば　平均値が一致してない
群間のズレ＜群内のズレ　ならば　平均値が一致している

と言えることがわかります。母集団の平均値が一致していないならば、「標本の平均値」と「全体の平均値」との差が大きくなるため群間のズレが大きくなるのに対して、母集団の平均値が一致しているときは「標本の平均値」も「全体の平均値」の近くにまとまるため群間のズレが小さくなることから、こうした傾向が言えます。

しかし、たとえば上図左側のＢ群だけを見ると「Ｂ群の平均値」は全体の平均値に近いため、「群間のズレ＜群内のズレ」　となり、「平均値が一致している」という答になってしまいます。つまり、３つ以上の標本群があるとき、「群間のズレ」と「群内のズレ」の大小関係は一定にはなりません。したがって、「3 つ以上の標本群に対して、群間のズレを群内のズレを一度にまとめて総合的に判断する方法」が必要です。

それを行うのが「分散分析」という方法です。

5.16 分散分析の手順（１）

平方和計算表

	A群	B群	C群	全体	
	6,8,8,9,9,11,12	8,10,10,12,12,14,15	9,10,12,14,15,15,17		
サンプルサイズ	7	7	7	21	
標本平均	9.00	11.57	13.14	11.24	
標準偏差	1.85	2.26	2.70	2.86	
標本分散	3.43	5.10	7.27	8.18	← 全体の分散
平方和	24.00	35.71	50.86	171.81	← 全体の平方和
				110.57	← 群内の平方和
（群内平均－全体平均）の２乗	5.01	0.11	3.63		
×サンプルサイズ	35.06	0.78	25.40		
				61.24	← 群間の平方和

分散分析表

	平方和	自由度	平均平方	F
群間	61.24	2	30.62	4.98
群内	110.57	18	6.14	
全体	171.81	20		

分散分析の具体的な手順は以下のとおりです。
上図はA、B、Cという３つの標本群に対して分散分析を行ったものです。たとえば「A群」という箱の中にある「6, 8, 8, 9, 9……」という数値が実際のA群標本内の個々のデータです。標本群の数が４つ以上に増えても同じ手順で分析できます。

まずはそれらの標本の「サンプルサイズ」を出します。上図例では３群ともサンプルサイズ＝７となっていますが、群ごとにサンプルサイズが変わってもかまいません。「全体」の列にはＡＢＣのすべてを合算した全体の数値を書きます。この例では 7+7+7=21 が全体のサンプルサイズです。

「標本平均」には群ごとの平均値を記入します。ここでも「全体」の欄ではＡＢＣの全体をひとまとめにした平均値を書きます（以下同様）。

「標準偏差」には群ごとの標準偏差を記入します。

「標本分散」は標準偏差の２乗です。

「平方和」は　標本分散×サンプルサイズ　で計算します。

「全体の平方和」と「群内の平方和」については、

全体の平方和　＝　全体の分散　×　全体のサンプルサイズ
群内の平方和　＝　標本の平方和すべての合計

であることに注意してください。「群内の平方和」は厳密には群内のすべての個体値に関して「群内のズレ（個体値－標本平均)」を２乗して足したものです。

さらに、「(群内平均－全体平均）の２乗」を各群について計算し、それに各群のサンプルサイズを掛けて合計して「群間の平方和」を出します。これが「群間のズレ」を２乗して足したものです。なお、

全体の平方和　＝　群内の平方和 ＋ 群間の平方和

となっているのは偶然ではなく必ずこれが成り立つので、もし違っていたらどこかで計算ミスがあります。

ここまで計算した上で、次は分散分析表を作ります。
「平方和」の列は群間、群内、全体の平方和をそれぞれ転記します。
「自由度」の列は以下のとおり計算します。

群間の自由度　＝　標本群の数 － 1
群内の自由度　＝　標本群ごとの「サンプルサイズ－１」を計算してすべて合算
全体の自由度　＝　標本群すべてのサンプルサイズの合計　－　1

群間と群内の「平均平方」は以下のとおり計算します。

平均平方　＝　平方和　÷　自由度

最後にＦ値を求めます。

Ｆ　＝　群間の平均平方　÷　群内の平均平方

「群間のズレ」が「群内のズレ」より大きいほど「平均値が一致していない」ことを意味するので、Ｆ値が大きければ大きいほど「標本群の平均値には差がある」という判断になりますが、では、上図の例にある　Ｆ＝4.98　は何を意味するのでしょうか？　それを判断するためには、Ｆ分布表が必要になります。

5.17 分散分析の手順（２）

A群	B群	C群
6,8,8,9,9,11,12	8,10,10,12,12,14,15	9,10,12,14,15,15,17

帰無仮説　A、B、C群の母集団の平均値に差はない

分散分析の結果　群間の自由度：2　群内の自由度：18　F値：4.98

有意水準5%のF分布表

		群間の自由度				
		1	2	3	4	5
群内の自由度	10	4.965	4.103	3.708	3.478	3.326
	15	4.543	3.682	3.287	3.056	2.901
	20	4.351	3.493	3.098	2.866	2.711
	25	4.242	3.385	2.991	2.759	2.603
	30	4.171	3.316	2.922	2.69	2.534

4.98 ＞ 3.493　のため帰無仮説は棄却される（＝平均値には差がある）

算出されたF値　　**群間自由度=2、群内自由度=20に該当するF値**

標本A群、B群、C群のデータは前ページまでと同じです。
分散分析を行う場合は

すべての標本群の母集団の平均値に差はない

という帰無仮説を設定します。これが棄却された場合は、いずれか１つ以上の母集団の平均値が他と異なっていることを意味します。

その判断をするために、分散分析の結果から「群間の自由度」、「群内の自由度」、「F値」を使います。
F値がある水準より大きければ帰無仮説は棄却されるのですが、その「ある水準」は群間および群内の自由度の数値に応じて変わります。その水準を計算したのがF分布表です。

この例の場合、群間の自由度：２　と　群内の自由度：１８　に最も近い値を5%有意水準のF分布表から探すと3.493です。実データから算出されたF値4.98はこれよりも大きいため、帰無仮説は棄却されます。つまり、「95%以上の確率で、標本A、B、C群の母集団のどれか一つ以上の平均値が他と異なる」と言えます。

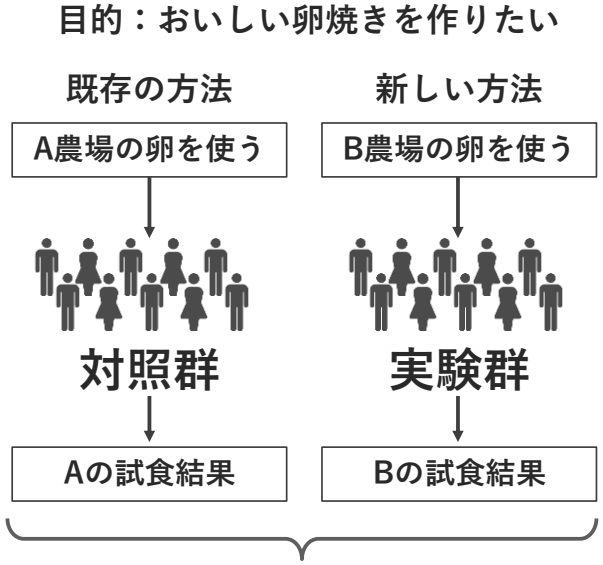
なお、有意水準が変われば F 分布表も変わるため、有意水準 1%で検定を行う場合はその水準に合った F 分布表を使わなければなりません。有意水準 1%の場合、群間の自由度：２　と　群内の自由度：１８に対応する数値は 5.85 となり、4.98 はこれより小さいため帰無仮説は棄却されません。つまり、

「標本Ａ、Ｂ、Ｃ群の母集団のどれか一つ以上の平均値が他と異なる確率は
95%から 99%の間である」

と言えます。

5.18 対照群と実験群

ある目的を達成するための既存の方法があり、それに対して新しい方法を考えた場合は、比較対照する実験を行って今後の方針を決める



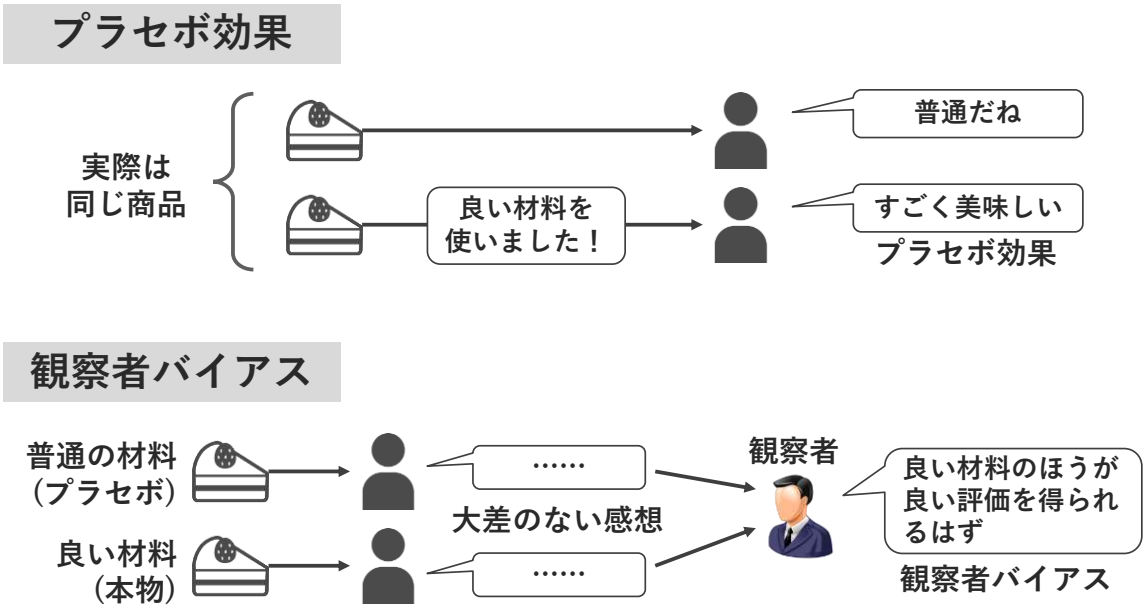
たとえばレストランのシェフが「もっとおいしい卵焼きを作りたい」と考えて、「卵を変えてみてはどうだろうか？」と思いついたとします。そこでシェフは今まで使っていたA農場の卵と、新しいB農場の卵で同じ卵焼きを作り、両方を食べ比べてもらって評価が良い方を使うことにしました。

このように、既存の方法に対して新しい方法を考えたので両方を比較して決めたい、という場合がよくあります。このとき、新しい方法を試す集団を「実験群」、既存の方法で行う集団を「対照群」と言います。「実験」とは「実際に試してみる」ということで、「今までやったことがない新しい方法を実際に試してみる」のが実験群です。

それに対して、「今までやってきた方法」で行う集団を「対照群」と言います。これは、実験群の結果と比較対照してみる群」という意味です。

レストランの新メニューに限らず、新しい方法を採用するかどうかを決めるために実験をすることはよくありますが、そうした実験を行い評価する際に注意すべきことがいくつかあります。その注意事項を知るためにまずは「対照群」「実験群」という用語を覚えておいてください。

5.19 プラセボ効果と観察者バイアス



実際は同じ商品であったとしても、片方を「このケーキはいい材料を使ってるんですよ！」とアピールすると、そちらを高く評価する人が増えてしまいます。これをプラセボ効果と言い、実験をするとき慎重に防がなければならない要注意点のひとつです。「プラセボ」というのは医薬品分野で「偽薬（効果のないニセの薬）」を意味する言葉で、まったく効果のないはずの薬、たとえばただの小麦粉や砂糖水を薬と偽って与えても実際に病状が改善してしまう場合がある現象を指します。「病は気から」とも言うように、人間の身体状態は心理状態の影響を大きく受けるため、「良い薬をもらった」と思い込んで心理状態が好転することによって身体の状態まで改善するものと考えられています。

医薬品の試験のような場面でプラセボ効果が起きると、実際には効かない薬でも「治療効果がある」という結果が出てしまうことがあります。そのままその薬が市販されると、「効かない薬」が多くの患者に処方されてしまい、有効な治療ができない事態を招くため、それは絶対に防がなければなりません。

プラセボ効果は「思い込み」による心理的な効果なので、思い込みを防げればプラセボ効果は起きません。具体的には、対照実験をするとき対象者にはプラセボなのか本物なのかわからないようにして行います。たとえば「今までと同じ材料のケーキ」、「良い材料を使ったケーキ」の両方を用意して「どちらを食べているのかわからない」ようにして食べてもらい、それでも「良い材料」のほうが良い評価が得られれば、プラセボ効果ではなく、実際に差があると判断できます。なお、医薬品の試験などでは、薬を処方される患者と処方する医師の両方ともプラセボなのか本物なのかわか

らないようにする「二重盲検法」という方法が使用されます。

一方、プラセボ効果に似ていますが少し違う「観察者バイアス」という現象にも注意が必要です。これは、実験結果を評価する「観察者」が思い込みをしてしまう問題です。たとえば、ケーキの新製品を試す際、プラセボ効果が起きないように、試食をしてもらう対象者にはプラセボなのか本物なのかわからないようにしてケーキを食べてもらったとして、感想が次のようなものだったらどう判断すべきでしょうか？

旧製品（プラセボ）の感想　：とても美味しかったです
新製品（本物）の感想　　：すごく美味しかったです

「とても」と「すごく」には差があるのでしょうか？　普通に考えれば「大差のない感想」です。しかし、「良い材料を使った本物のほうが良い評価のはず」だと観察者が思い込んでいると、「本物のほうがより好評だった」と解釈してしまいがちです。結果が定量的に出るもの（１００点、８０点のように数字で出るもの）については観察者バイアスは起きにくいですが、「すごく美味しい」のように定性的に出るものは起きやすいので注意が必要です。

プラセボ効果と同様、「思い込み」がなければ観察者バイアスも起きないので、これを防ぐためには「観察者にもプラセボか本物かがわからないように」しなければなりません。つまり、「実験結果を分析評価する人」に対しても、プラセボの結果なのか本物の結果なのかを知らせずに分析してもらう必要があります。

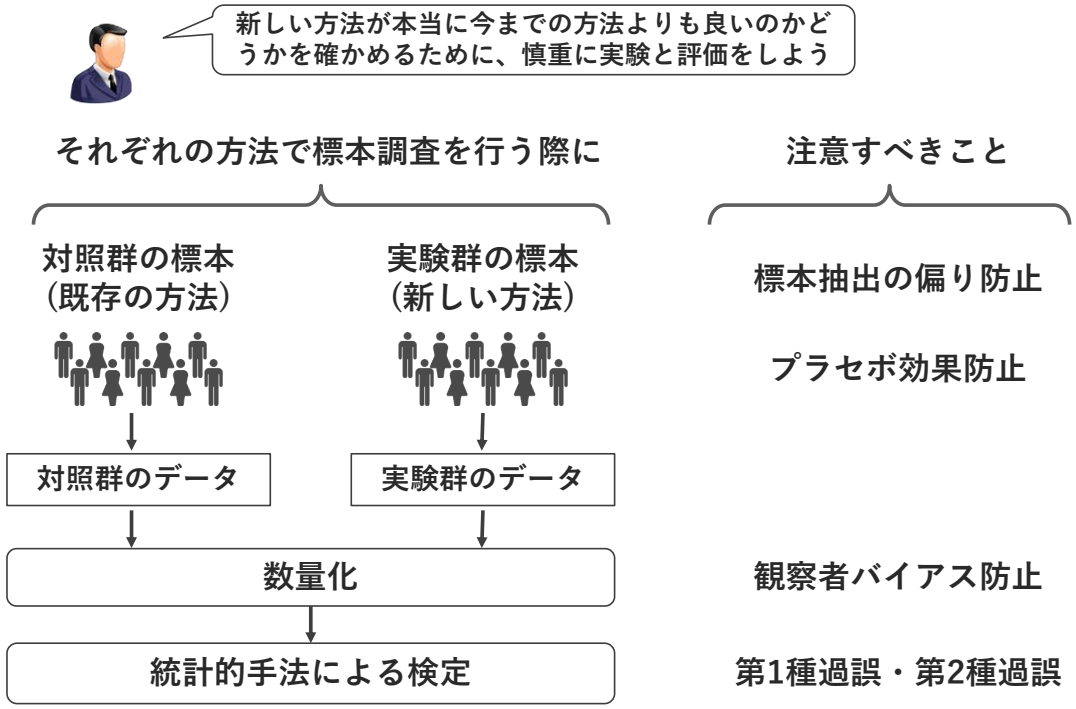
「新しい方法を考える」のは「今までよりも良くしたい」からです。そのために大きな手間暇をかけて新しい方法を探すわけですから、それがうまくいくかどうか試す実験をするときはどうしても「上手くいって欲しい」という願いがあります。それは人の心理として当然なのですが、しかしそれがプラセボ効果や観察者バイアスを招いてしまうのも事実なので、そのようなバイアスが起きないようにする慎重な配慮が求められます。

繰り返しますが、

新しい方法を考えて試そうとする者は、それが有効であって欲しいと願うもので、
その「願い」があるからこそ判断を間違えやすい

というのは重要なポイントです。その間違いを防ぐためにこそ、本書のような統計的な知識が必要です。

5.20 対照実験の流れと注意事項

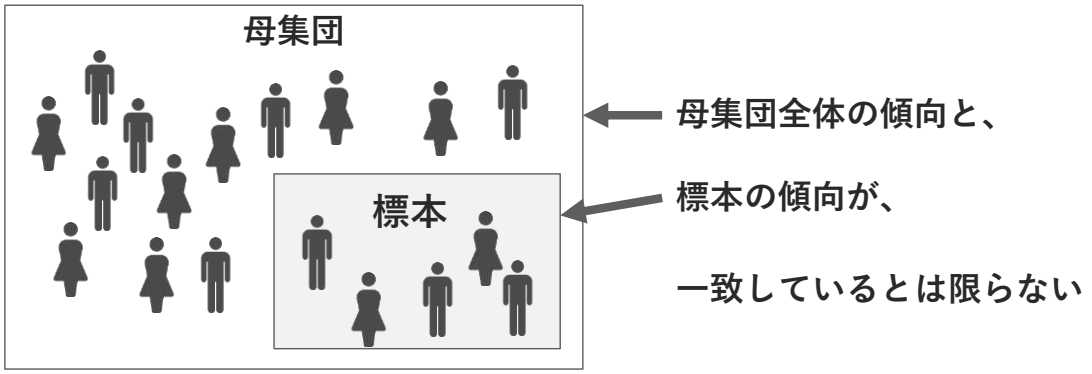


対照実験の流れと注意事項をまとめておきましょう。
まず、標本抽出に偏りがあってはいけないので、両群の条件を徹底的に揃えるように注意して「対照群（既存の方法を使う群)」と「実験群（新しい方法を使う群)」の標本を抽出します。
たとえば、新しい薬の効果を調べるのに「対照群は男性のみ、実験群は女性のみ」のような標本では偏りが出てしまいます。あるいは食品の評価をするような場合、食味に対する感性は出身地や年齢による差も大きいため、それらの条件も偏らないようにします。

次に、それらの実験参加者にプラセボ効果が起こらないようにする必要があります。一般に「新製品」「新製法」など「新」がつくものには良い印象が生まれがちですので、実験参加者は自分が対照群なのか実験群なのかを知ることができないようにしておきます。たとえば医薬品ならば患者と医師の双方とも、使用している薬が本物なのか偽薬なのかがわからないようにします。

対照群と実験群のデータが出た後、それを「数量化」するときには観察者バイアスが起きがちです。「数量化」とは、「すごく美味しかった」のような定性的な感想を「満足度８０点」のような数値に換算することで、統計処理を行うためには必要ですが、このときに観察者バイアスが起きないようにしなければいけません。データがもともと数値で出る実験であれば観察者バイアスは起きません。
そして最後に残るのがこれから説明する「第１種過誤・第２種過誤」です。

5.21 標本が母集団を正しく表すとは限らない

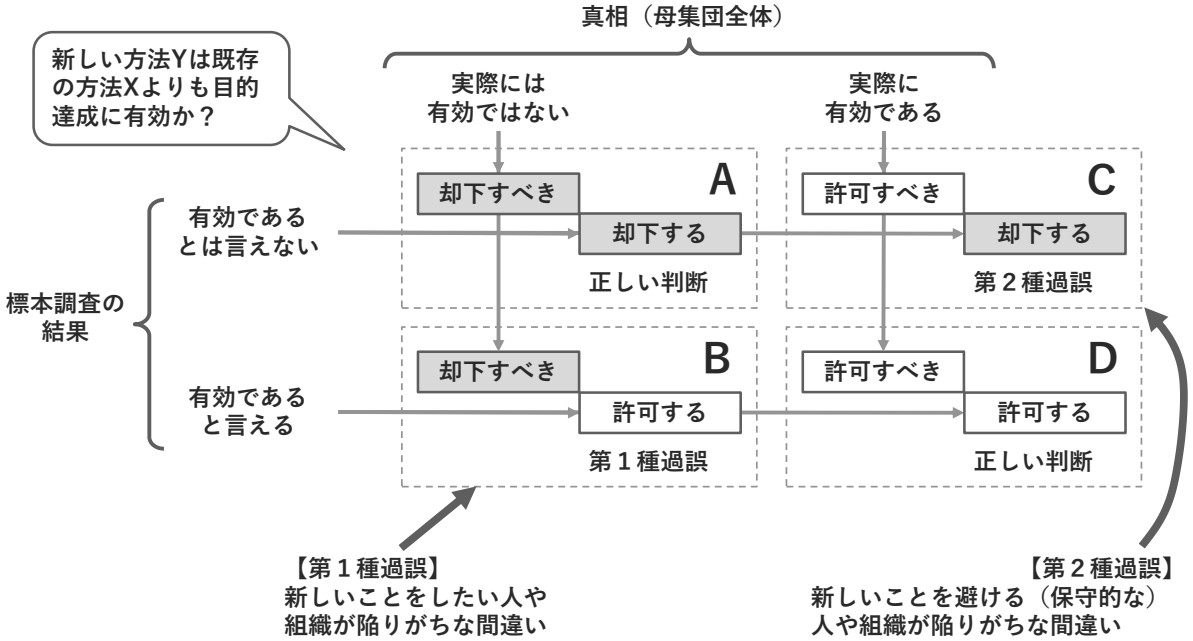


そもそも標本調査とは母集団全体の一部についてのみ調査を行うもので、母集団全体の傾向と標本の傾向が一致しているとは限りません。したがって、もしその両者が大きくズレていた場合、標本のデータを元にして行った判断は母集団全体に対しては「誤り」になります。たとえば新しいお菓子を開発する場合、標本として選んだ人々にたまたま「極端な甘いもの好きが多かった」場合、「標本調査では人気が高かった」としても、一般発売すると「甘すぎて売れない」といったことが起こり得ます。これが「標本を元にした判断が実際の母集団には合っていなかった」という「誤り」です。

もちろん、そのような誤りが起きないように統計的な処理（検定）をするのですが、どんなに正確に統計処理をしても防ぎきれない誤りもあります。

5.22 第１種過誤・第２種過誤

標本調査による比較対照実験のデータに統計的な検定処理をしても、正しい結果を出せるとは限らない。標本調査が誤った結果を出すパターンが2種類あり、それぞれ第1種過誤（誤り）/第2種過誤（誤り）と呼ばれる



新メニュー開発、新薬開発のように「新しい方法が既存の方法よりも有効かどうか確かめたい」という場面では、対照群と実験群の標本を比較対照して有効であるかどうかを検定することになります。

図中の「真相」とは「母集団全体に対して新しい方法が有効かどうか」の答で、これは「実際には有効ではない／実際に有効である」の２つに１つです。「真相」が「実際には有効ではない」のであれば新しい方法 X の採用は却下すべきですし、「実際に有効である」のならば許可すべきです。

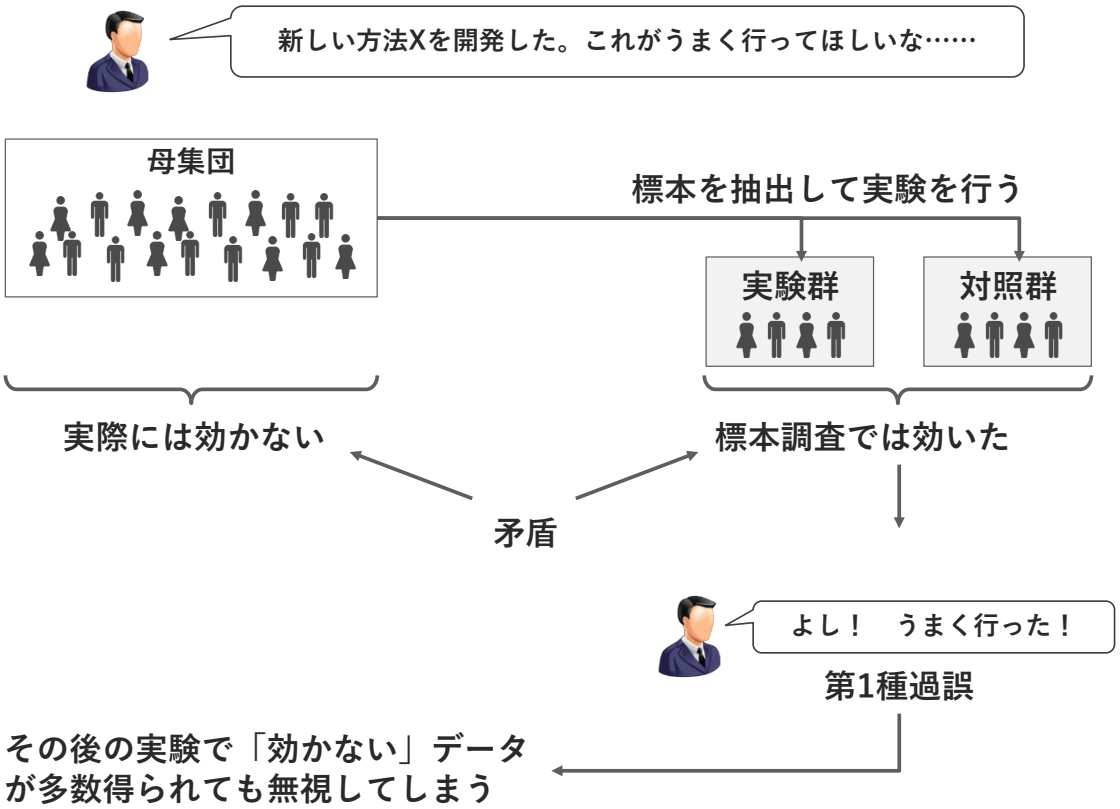
しかし現実には「母集団全体」への調査は不可能なので標本調査で「真相」を推定しなければなりません。標本調査の結果は「有効であるとは言えない／有効であると言える」の２つに１つです。新しい方法 X を採用するかどうかの判断はこの結果に応じて行うため、「有効であると言えない」ならば却下されますし、「有効であると言える」ならば許可されます。それが「真相」と一致していればいいのですが、現実には一致しないことがあります。つまりそれが「誤り」です。

「真相」と「標本調査の結果」はいずれも 2 種類なので、その組み合わせは 2x2 で 4 種類あり、そのうち 2 種類が「誤り」です。
A は「却下すべき」ときに「却下した」わけですから、正しい判断です。

B は「却下すべき」ときに「許可してしまった」わけで、誤りです。これを第 1 種の過誤と言います。
C は「許可すべき」ときに「却下してしまった」わけで、誤りです。これを第 2 種の過誤と言います。
D は「許可すべき」ときに「許可した」わけですから、正しい判断です。

「過誤（かご）」は「誤り」という意味で、「第 1 種過誤・第 2 種過誤」は標本調査の結果を基に統計的な検定を行う際に必ずつきまとう誤りです。これをゼロにすることは不可能であり、ある程度は必ず発生すると考えなければなりません。

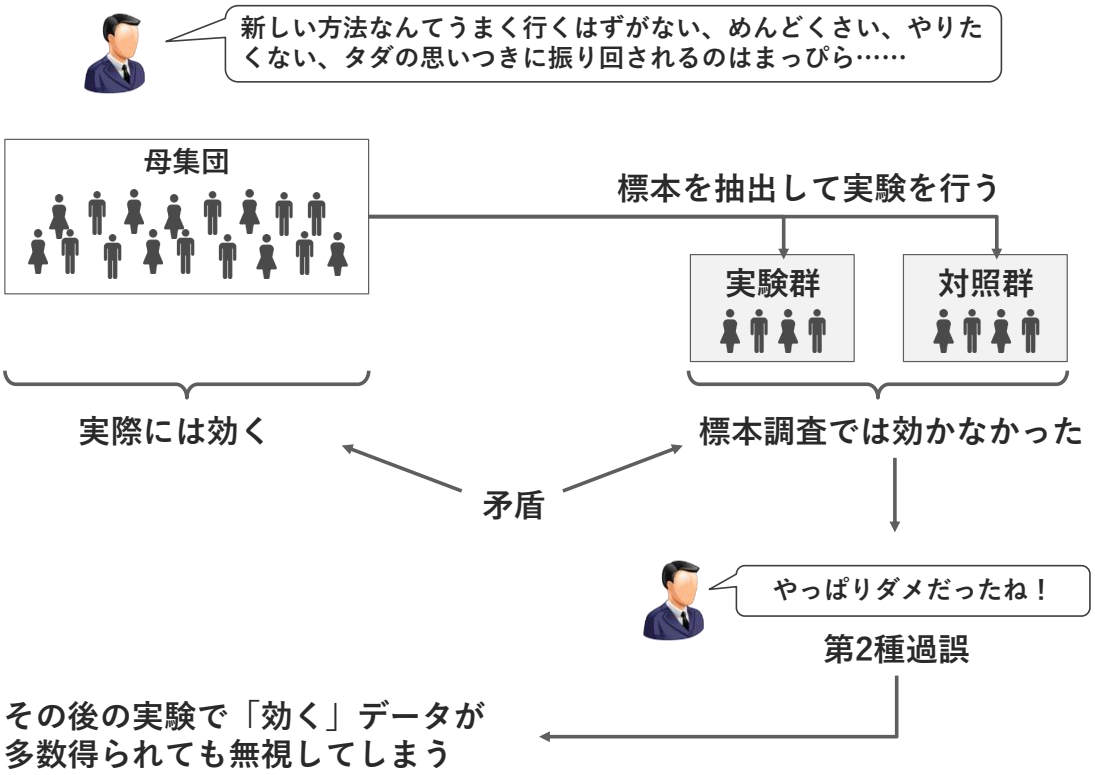
第 1 種過誤に特に注意しなければいけないのは、「新しい方法」への期待が高い状況です。



新しい方法 X が母集団に実際には効かないのに標本調査では「効く」という結果が出てしまう事態は確率的に必ず発生するものです。それを「よし、うまく行った！」と考えること自体は問題ありませんが、その思い込みが強すぎるとその後の実験で「効かない」データが多数得られても無視してしまいがちです。いったんは「期待通りにうまく行った」データが得られたとしても、それが第 1 種の過誤である可能性を常に忘れず、その後のデータも注意深く検証していく姿勢が必要です。

逆に、第 2 種過誤に特に注意しなければならないのは「新しい方法」が期待されていない状況です。
たとえば、偉い人が「新しいこと」をむやみやたらに提案するものの、タダの思いつきなので結局

うまく行かない、ということが繰り返されているような職場では、「どうせうまく行くはずがないのにめんどくさいなあ」という空気が蔓延していることがあります。



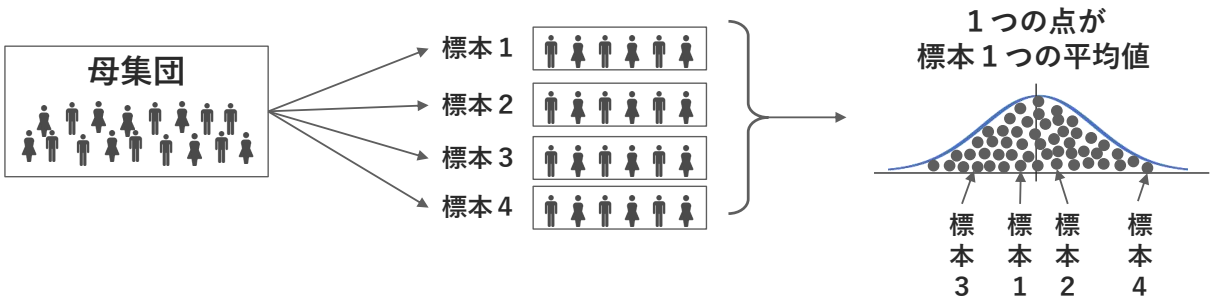
その場合は「効かなかった」というデータのほうが「期待通り」なので容易に受け入れられてしまい、それが第２種の過誤であってもその後のデータを無視してしまいがちです。

いずれの誤りもゼロにはできないので、標本調査で得られる結果にはこれらの過誤がありえます。最初の標本調査で良い結果（悪い結果）が出たとしても決めつけず、できる限りその後も調査をして過誤がないかどうかを確認していく必要があります。

この問題を理解するために、そもそもなぜ正しく統計処理をしても過誤が起きるのか、その理由を知っておきましょう。

5.23 なぜ過誤が起きるのか？

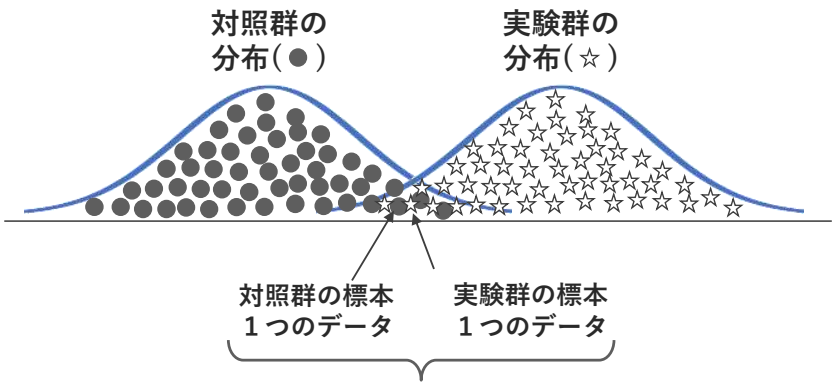
同じ母集団から何度も標本抽出をして平均値を調べると、バラツキが出る



同じ母集団から何度も標本調査をする場面を考えてください。たとえば「全校生徒からランダムに10人選んで身長を調べる」という調査を10回、20回、30回……と繰り返した場合、標本の各回毎に平均値はバラつき、上図の右端のように正規分布をします。この正規分布の中央は母集団の平均値と一致します。したがって、標本1や2のように母集団の平均に近い値が出る場合が多いものの、標本3や4のように大きく外れた値になることもあります（その確率は小さいですが）。

次に、対照実験を行う場合を考えます。新薬開発のように、新しい方法で効果があるかどうかを調べる対照実験を何度も繰り返したとすると、もし「効果がある」場合は対照群と実験群の値は下記のようにはっきりと違う山に分かれて分布します。当然、それぞれの山の中央付近のデータが出るケースが一番多いものの、下記のように「対照群と実験群が偶然、近い値になる」確率もゼロではありません。その場合、データを素直に読めば「新しい方法には効果はない」、つまり、実際には有効なのに「有効であるとは言えない」わけです。これが第2種の過誤です。

新しい方法で効果がある場合の分布例

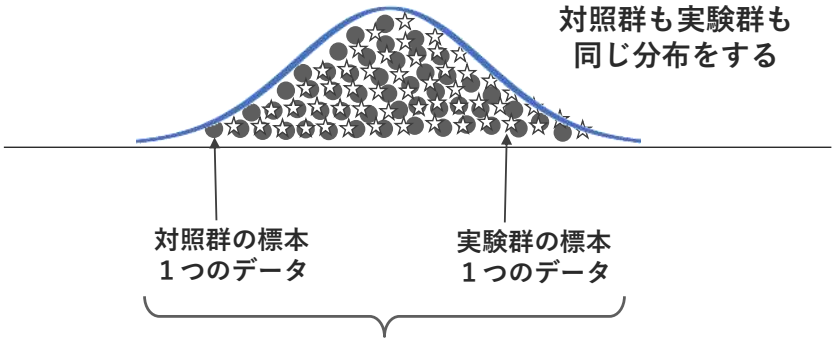


実際には効果があるのに、偶然近い値が出てしまうと「効果がある」とは言えない（第2種の過誤）

一方、新しい方法が実際には効果がなかった場合、対照群も実験群も同じ分布をするので山が重なります。しかしこれも「対照群と実験群が偶然、大きく離れた値になる」場合があります。そのデ

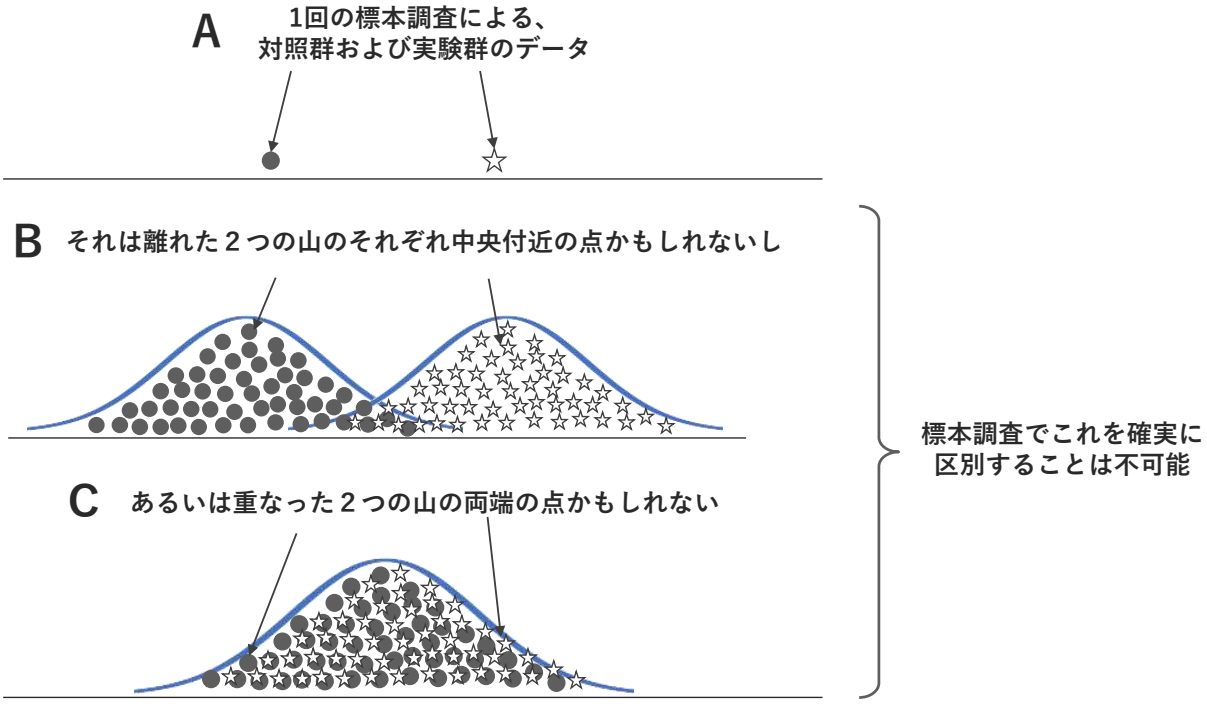
ータを素直に読めば「新しい方法には効果がある」、つまり、実際には有効ではないのに「有効であると言えてしまう」わけです。これが第１種の過誤です。

新しい方法では効果がない場合の分布例



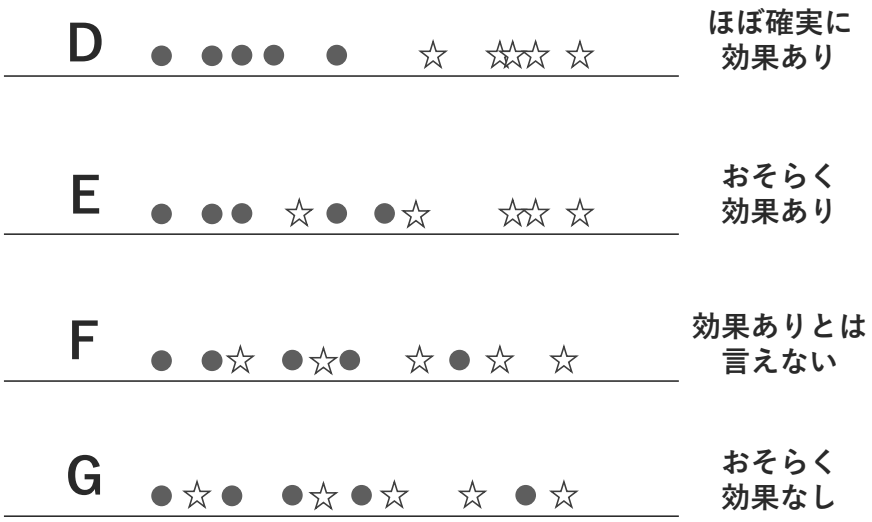
実際には効果がないのに、偶然離れた値が出てしまうと「効果がある」ように見える（第1種の過誤）

以上の例のように山形の分布（正規分布）のグラフとともに見ていると、どうして過誤が発生するのかわかりにくいかもしれませんが、実際に「１回の標本調査」で得られるデータは下図のＡのような１点だけのものです。それがＢ（新しい方法には効果がある）を意味しているのか、Ｃ（新しい方法には効果はない）を意味しているのかを確実に区別することはできません。１回の調査では対照群と実験群が１点ずつしかわからないので「山が見えない」ためです。



では、どうすれば真相がＢなのかＣなのかを区別できるのでしょうか？　単純な発想としては、標本調査の数を増やせば良いでしょう。たとえば対照実験を５回やって下記のような結果が出たら、Ｄはほぼ確実に効果があると言えます。Ｅはおそらく効果あり、Ｆは微妙、Ｇは効果なしの可能性が高いでしょう。

5回の標本調査による、
対照群および実験群のデータ



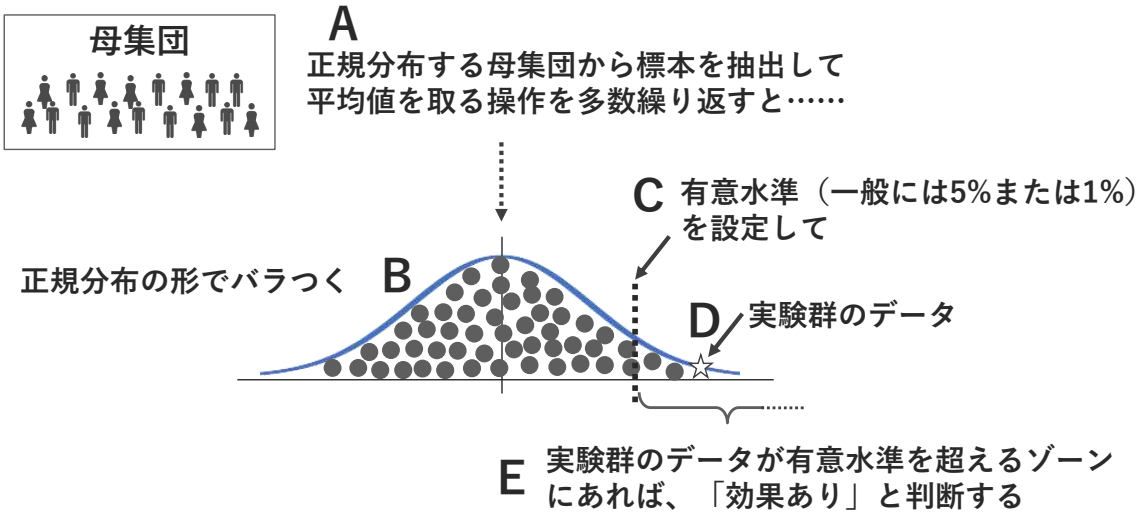
しかし、このような方法には２つの問題があります。

１．調査を増やすほど、時間・費用その他の労力がかかる
２．「ほぼ」「おそらく」のような判断を人間の感覚で行うと、観察者バイアスがかかる

１については、「調査」をするほど単純に時間とお金がかかるだけでなく、たとえば医薬品の試験（治験と呼ばれます）を行う際は、一部の患者にはプラセボが投与されます。プラセボは「まったく効き目のない偽薬」ですから、その患者は「病気を治す機会を失う」わけです。これは患者にとっての不利益になるため、できるだけ少ない方が良く、そのためにもできるだけ少ない数の標本で判断できることが望ましいのです。（プラセボではない方も「効果があるかどうかわからない薬」なのでやはり標本数をできるだけ少なくしたいのは同じです）

しかし、「標本数が少ない」と当然、分布の山は見えにくくなりますし、それを人の感覚で判断していると観察者バイアスがかかってしまいます。そこでその判断は統計的な検定により行う必要があります。

5.24 対照実験の効果判定の基本的なイメージ



対照実験で実験群に効果があるかどうかを判定する際の基本的な考え方のイメージです。

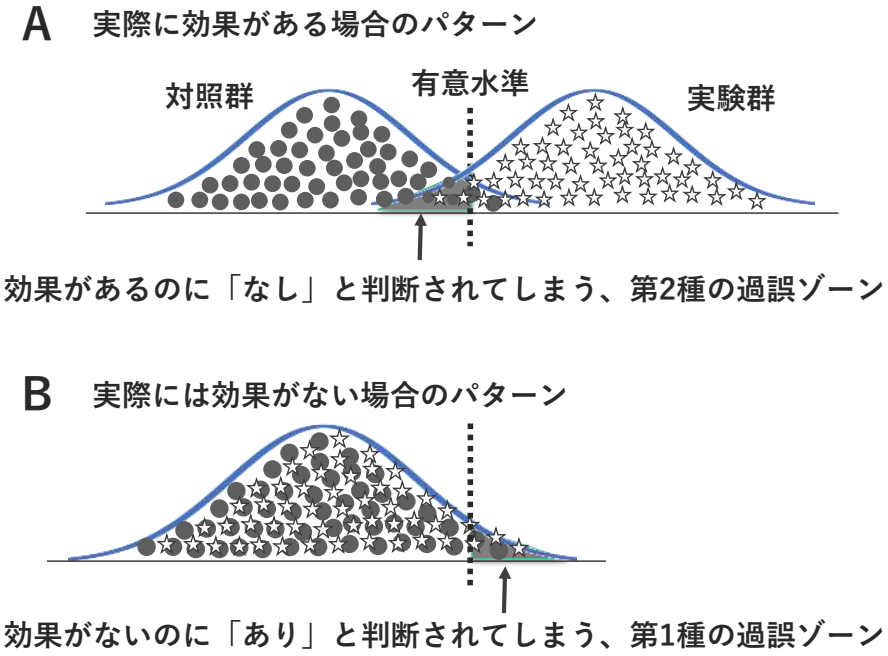
まず、正規分布をするなんらかの母集団から(A)標本を多数回抽出して平均値を取ったとすると、その値は同じく(B)正規分布の形でバラつきます。B で示した山形のグラフがそのバラツキ方のイメージで、多数の黒丸は標本 1 つのデータです。対照実験では「既存の方法」と「新しい方法」を対照するわけですが、B は「既存の方法」をとった対照群のバラツキだと思ってください。

次に、本章の仮説検定の項で説明した(C)有意水準を設定します。一般には 5%または 1%が使われます。そのうえで(D)実験群のデータ(☆)をプロットして、もし(E)有意水準を超えるゾーンにあれば、「効果あり」と判断します。「有意水準を超えた」ということは、「偶然ではなかなか起きない数字が出た」ということなので、「偶然ではないだろう」、つまり「新しい方法には効果がある」という判断をします。

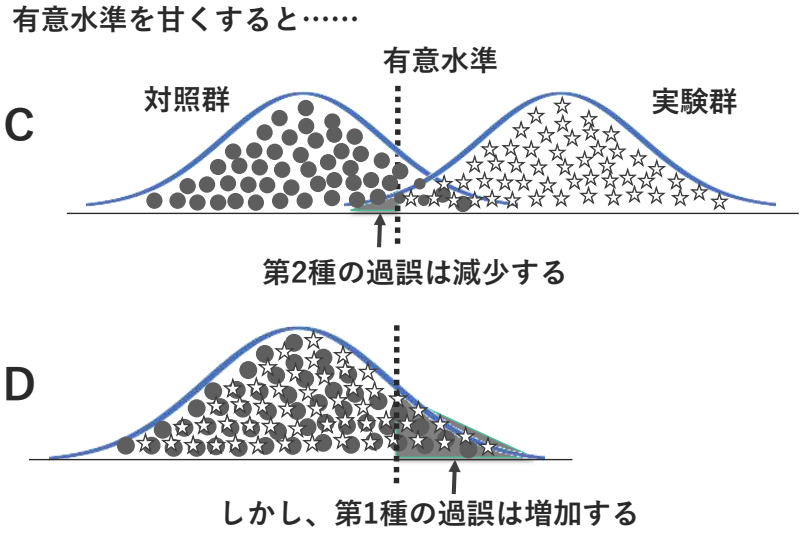
> 【参考】「有意水準 5%」というのは、「偶然では 20 回に 1 回しか起らない」という意味ですが、20 回に 1 回というのはたとえばサイコロを 2 個振って 6・6 が出る確率よりも大きく、医薬品のような人間の生死に関わる分野では頼りないため、医薬品では一般に有意水準 1%が使われます。

以上を念頭に置いて、次に「有意水準と第 1 種・第 2 種の過誤」との関係を考えましょう。

下記 A は「実際に効果がある場合のパターン」で、対照群と実験群が 2 つの山に分かれています。しかし、「実験群」の山で、有意水準より左のデータがたまたま出た場合は「効果があるのに、なしと判断」されてしまいます。これが第 2 種の過誤ゾーンです。それに対して B は「実際には効果がない場合のパターン」です。「実験群」の山で有意水準より右のデータがたまたま出た場合は、「効果がないのに、ありと判断」されてしまいます。これが第 1 種の過誤ゾーンです。



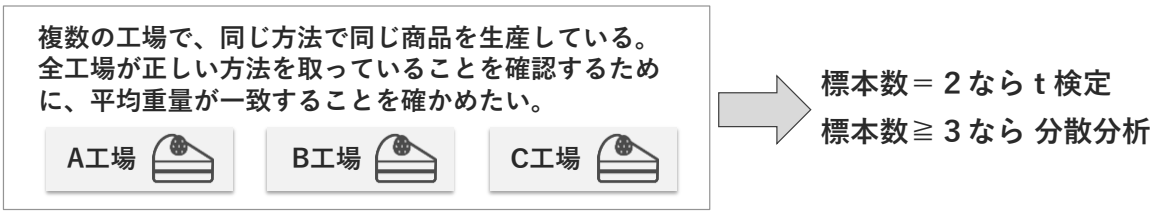
有意水準を挟んで左が第 2 種の過誤、右が第 1 種の過誤となるため、その両者を同時に小さくすることはできません。たとえば第 2 種の過誤を減らすために有意水準を甘くすると、第 1 種の過誤が増えてしまいます。



5.25 複数の標本を比較分析するパターンの例

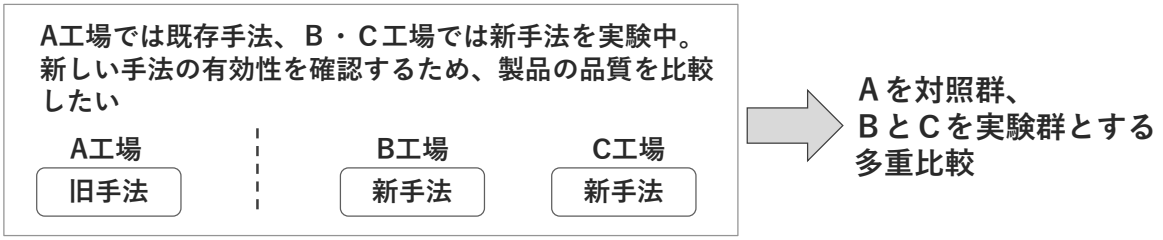
複数の標本を比較分析したいケースは非常に多いのですが、そのために使われる手法も非常に多様です。ここでは代表的なパターンをいくつか挙げておきます。

「複数の工場で、同じ方法で同じ商品を生産している。正しく生産が行われていれば商品の平均重量が一致するはずなのでそれを確かめたい」という場合は、既に説明したｔ検定（標本数＝２のとき）や、分散分析（標本数≧３のとき）を使用します。この場合はどの工場も対等であり、どの工場から取った標本も同じように扱われます（対照群や実験群という概念はありません）。



「Ａ校とＢ校で生徒の学力に差があるか確認するため、複数の教科の成績分布を比較したい」という場合はどうでしょうか。このような場合は教科ごとの成績を比較する「多重比較」を行います。多重比較の考え方については後述します。

A、 B、Ｃの３工場があり、Ａ工場ではこれまでと同じ長年実績のある手法で生産しているのに対し、Ｂ とＣでは品質向上が期待できる新手法での生産を実験中であるとします。実際に新手法で品質が向上したのかどうか判断するには、Ａ工場の標本を対照群、B、Ｃ工場の標本を実験群とする多重比較を行います。新手法を採用する工場の数がさらに増えた場合も基本的に同じです。多重比較の考え方については後述します。

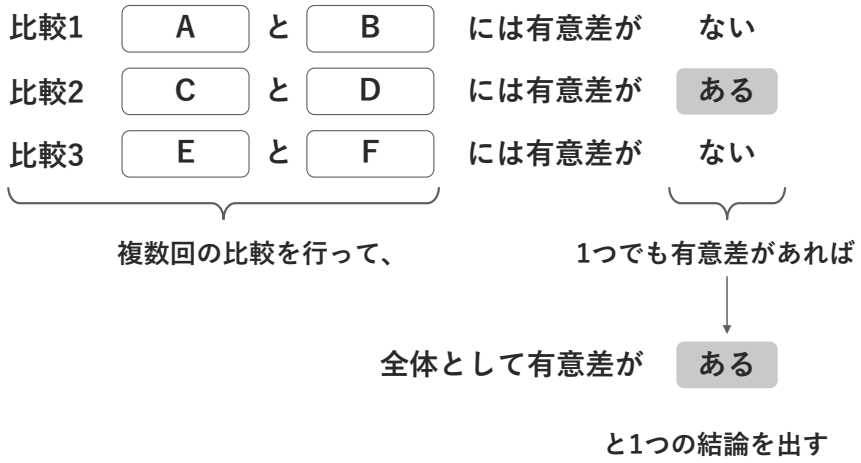


商品のパッケージデザインを決める場合など、複数の案をつくってそれを購入しそうな人々（見込み客）に見せて意見を聞き、最も印象の良い案を採用したいというケースがあります。このような

場合によく使われるのが「総当たりの一対比較によるランクづけ」です。これも多重比較の一種で、具体的な方法については後述します。

「多重比較」とは

「多重比較」というのは、複数回比較して 1 つの結論を出す操作のことを言います。



「複数回の比較で 1 つでも有意差があれば」という部分がポイントで、比較を単純に繰り返すと過誤が発生しやすくなるため特別な注意が必要です。

5.26 比較を単純に繰り返してはいけない

フレッシュジュース専門店が新品種のリンゴで作ったリンゴジュースの新商品のテストをしようとしている。既存品と新商品を飲み比べてそれぞれを10点満点で評価してもらうテストを違う店で10人ずつ、計3回行った。

			既存品		新商品			
	場所	人数	平均	分散	平均	分散	t値	有意差
1回目	A店	10人	7.3	2.61	7.5	2.45	0.32	なし
2回目	B店	10人	6.9	0.49	7.7	0.81	2.23	あり
3回目	C店	10人	7.0	1.00	7.4	2.44	0.80	なし

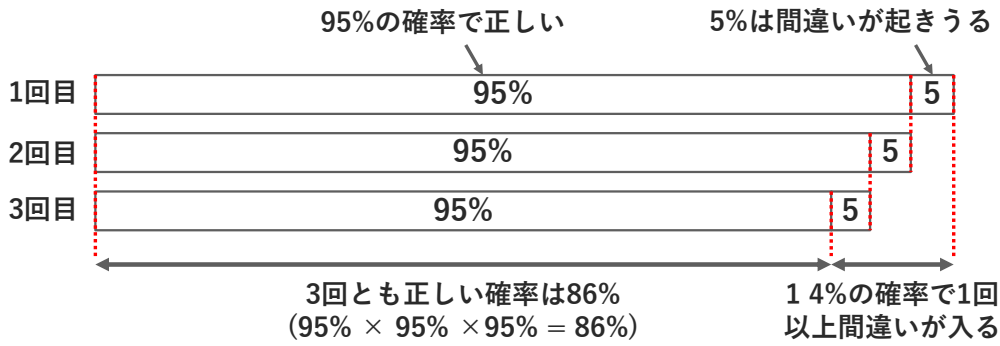
有意水準5%片側検定のt値：1.833
（対応のあるt検定、自由度9）

結論　B店で有意差があったから、新商品のほうが美味しい！

‥‥という結論を出してよいか？

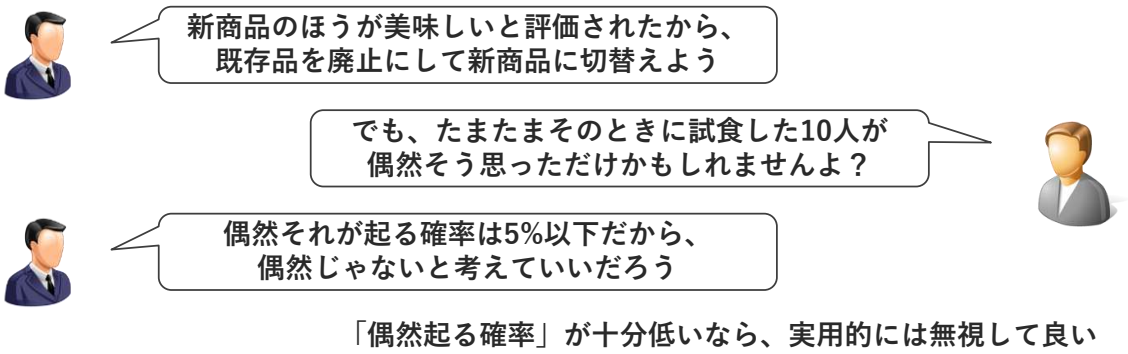
比較の単純な繰り返しで過誤が発生する具体例を見ましょう。フレッシュジュース専門店がリンゴジュースの新商品を既存品と飲み比べてもらうテストを 3 つの店舗で 10 人ずつ行いました。これは「同じ人に両方を飲み比べて点数をつけてもらう」ので、「対応のある t 検定」に該当します。細かなデータは省略しますが、2 回目の B 店の結果を見ると新商品と既存品の評価差が他店よりも大きく、t 値も 2.23 と有意水準 5%（「美味しいか？」だけを評価するので片側検定）の 1.833 以上なので B 店については「有意差あり」、つまり「新商品のほうが既存品より美味しいと評価された」と言えます。

では、そこで「B 店で有意差があったから、新商品のほうが美味しい！」と結論を出してよいのでしょうか？　これが「比較の単純な繰り返し」という落とし穴で、有意水準 5%つまり 95%の確率で正しい比較を 3 回繰り返すと、「3 回とも正しい」確率は 86%まで下がってしまいます。



問題は、「複数回の比較で 1 つでも有意差があれば、全体として有意差があると判断する」というロジックです。これは結局のところ、「**複数回の標本調査で 1 回でも過誤が起きれば（真相とは違う結果が出れば）、他のすべての標本が正しい答を出していても、全体として間違った結論を出してしまう**」ことを意味します。

既に触れたように、標本調査ではどうしても「過誤」が起きます。「有意水準」という考え方は、「過誤が起きるとしてもそれが十分小さいなら、正しいとみなしてよい」というものです。具体的には新商品を開発・発売するときにこんな判断をすることがよくあります。

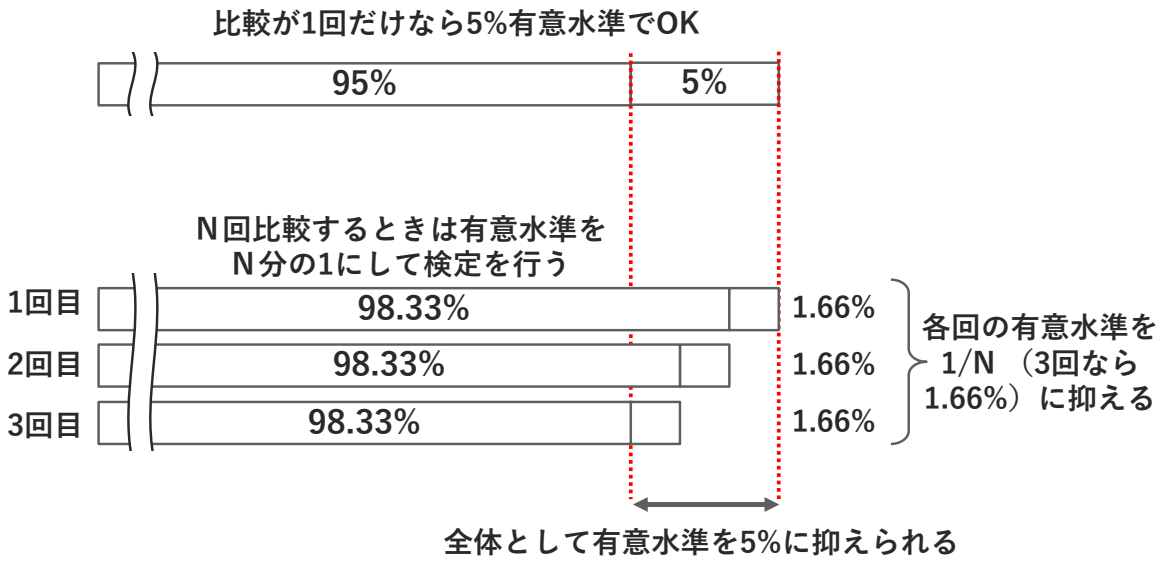


標本調査の結果が正しいとは限りませんが、「偶然かもしれない、かもしれない」とばかり言っていたのでは新しいことは始められません。偶然の可能性が十分低いなら、実用的には無視してよいのです。その「十分低い」の基準が有意水準であり、よく使われるのが 5%（医薬品のような分野では 1%）です。

確かに、5%以下なら 20 回に 1 回以下なのである程度当てにして良さそうですが、前述のフレッシュジュース専門店のように「3 つの店舗でテストする」という場合、どこか 1 店でその「偶然」が起きる確率は 14%に上がってしまいます。もしこれが 4 店舗 5 店舗と増えるとさらにその確率が上がっていく（有意水準が悪化する）ため、比較を単純に繰り返してはいけません。

そこで、複数の比較を適切に行うための方法を一般に「多重比較」と言います。多重比較の具体的な方法は非常に多様です。本書では最も簡単なボンフェローニ法から紹介します。

5.27 有意水準を調整するボンフェローニ法

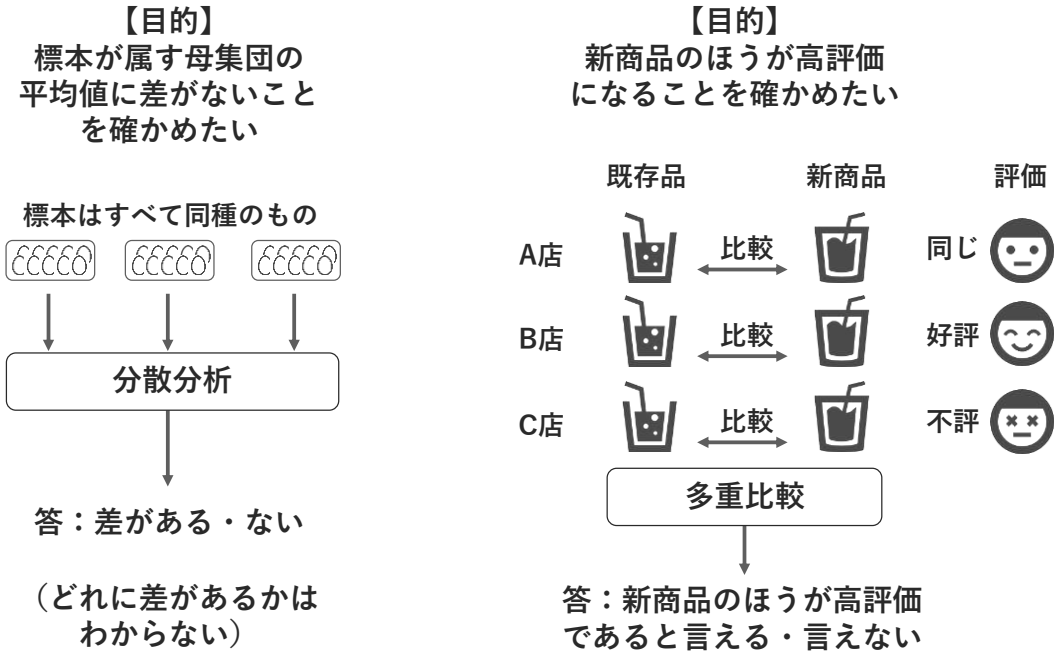


前述のフレッシュジュース店のようなケースで有意水準の悪化を防ぐための簡便な方法が、比較の回数に応じて有意水準を厳しくするボンフェローニ法です。比較が 1 回だけであれば 5%有意水準で検定するところで、3 回の比較をする場合は 5/3=1.66%まで下げることにより、全体としての有意水準を 5%に抑えることができます。

このボンフェローニ法は有意水準を 1/N にするだけなので簡単な方法ではありますが、第 2 種の過誤、つまり「実際には有意差があるのに、ないと判断してしまう間違い」が起りやすくなる欠点があります。具体的には「新品種のリンゴで作った試作品のほうが実際には美味しいと評価されているのに、差が無いと判断してしまい、新商品を出せなくなる」という事態が起きやすいのです。N が大きくなるほどこの欠点も大きくなることに注意してください。

ところで、この「検定を単純に繰り返してはいけない」という話題は本章の「分散分析」の項でも出ていました。その際は、「3 つの標本の平均値が一致するかどうかを確かめるために t 検定を 3 回繰り返してはいけない。その代わりに分散分析を使う」という趣旨でした。そこで、分散分析と多重比較の違いを確認しておきましょう。

5.28 分散分析と多重比較



分散分析が適しているのは「標本が属す母集団の平均値には差がない可能性が高く、それを確かめたい」という場面です。分散分析の対象になる N 個の標本はいずれも同種のものなので、たとえば「重量が一致するはずである」、という仮説が有力であり、それに対して YES または NO の答えを出すのが分散分析です。また、N 個の標本をまとめて一気に分析するため、「一致していない」という答が出たとしてもどれが有意に大きいのか小さいのかはわかりません。

一方、フレッシュジュース店の例は「新商品のほうが高評価になる」つまり「差がある」ことを確かめたい場面です。既存品と新商品はそもそもが異種のものなので差がある可能性が高く、それを 1 店ごとに比較したうえでその結果をすべて合わせて結論を出すのが多重比較です。1 店ごとに比較をするので、どの店では好評だった/不評だった、などの評価の差もわかります。

多重比較にはこの他にも多様な方法がありますが、いずれにしても「差がある可能性が高い」場面で使うものであり、「1 組ごとの比較を複数回行う」という点で分散分析とは違います。

ちなみに、フレッシュジュース店の例では分散分析を行う意味はありません。分散分析で分かるのは「差があるかないか」だけです。ジュース店としては「新商品が好評なのか不評なのか？」を知りたいので、「差がある」という答が出ても「好評・不評」を区別してくれないと役に立たないからです。

5.29 複数の項目による多重比較

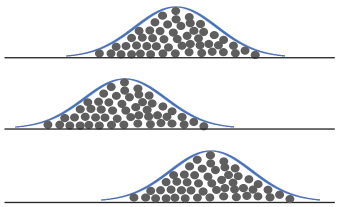
【目的】A校とB校で学力に差があるかどうかを知りたい

調査項目は３つ、標本は１つ

A校生徒の番号と点数の表

	1	2	3	4	5	6
国語	78	67	54	82	44	56
数学	89	77	62	61	50	52
英語	54	82	43	77	71	91

教科別成績分布グラフ
（A校の標本分）

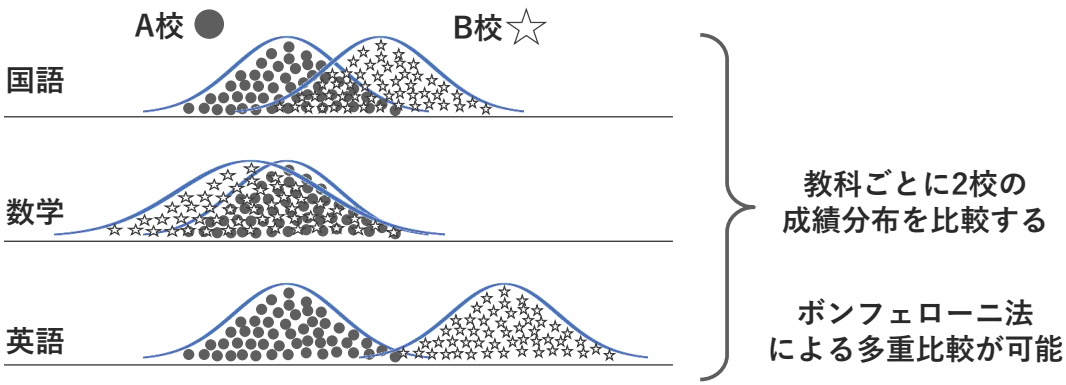


３項目のデータを分けて扱う

２つの学校、A 校と B 校で学力に差があるかどうかを知りたいとします。上図中の表は A 校生徒の一部を標本として抽出し、国語・数学・英語の３教科の成績を表にまとめたものです。この場合、調査項目は３つ、標本は１つであり、そのデータは右側の「教科別の成績分布グラフ」のように３項目のデータを分けて扱う必要があります。たとえば血圧が 120 で身長が 165 だったとしてその２つの数字を比較しても意味がないように、別な教科の成績は相互比較をしても意味がないからです。

そこで、２校について学力を比較する場合は各校それぞれから標本を取り、教科ごとに成績分布を比較する必要があります。

２校分のデータをグラフ化したイメージ

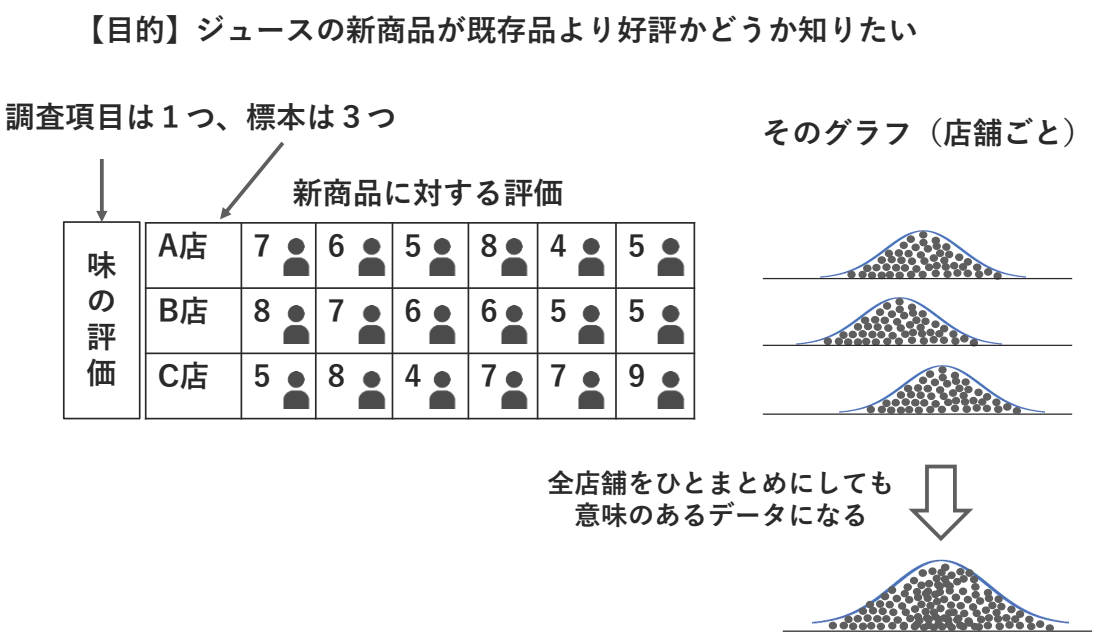


その場合、複数回比較を行って「A 校と B 校の学力には差がある/ない」という１つの結論を出す

ため、多重比較になります。３教科程度の項目数であればボンフェローニ法で有意水準を調整して比較することができます。

一方、既に触れたフレッシュジュース店の新商品の調査の例ではＡ・Ｂ・Ｃの３店で標本調査を行い、「店舗毎に比較する」「やはり多重比較になるのでボンフェローニ法で調整する」というものでした。しかし実はこちらの場合は「全ての店舗のデータをひとまとめにして扱う」ことも可能です。その場合、比較が１度で済むため多重比較になりません。

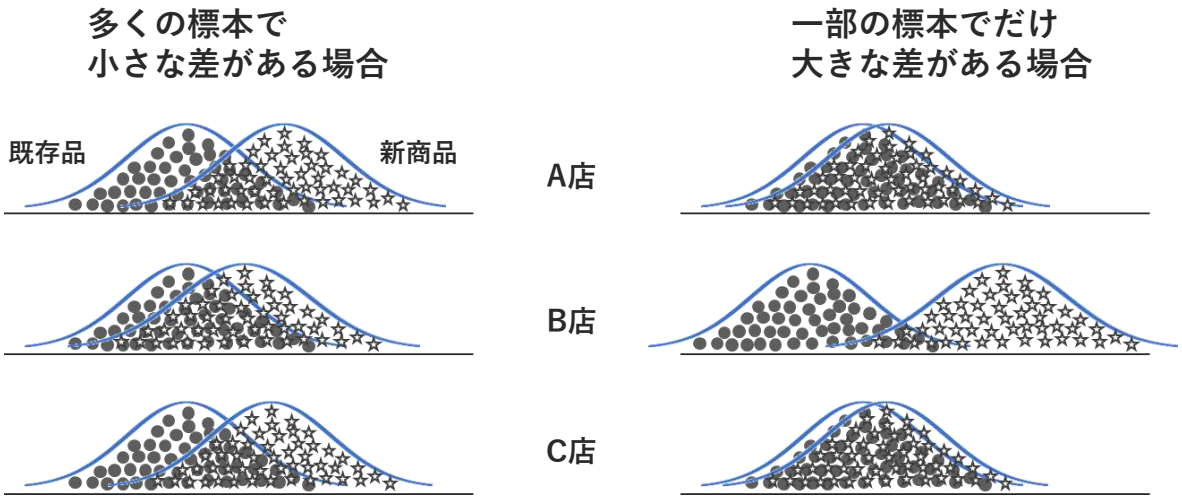
5.30 複数の標本データをひとまとめにする



学力調査のデータは「調査項目は３つ、標本は１つ」でしたが、ジュース店の新商品調査のデータは「調査項目は１つ、標本は３つ」です。この場合、すべてのデータが「味の評価」という同質のものになるため、異なる標本のデータをまとめて（一括して）扱うことが可能です。そこでたとえばABC3店舗の標本を１つにまとめて１回の比較で済ませることもできます。その場合は多重比較ではないためボンフェローニの調整は必要ありません。ボンフェローニ法は比較の回数が多くなるほど第２種の過誤が起こりやすくなる欠点がありましたが、ひとまとめにすればその欠点を回避できます。

一方で、多重比較をしたほうが良い面もあります。たとえば３店舗のうち１店舗でだけ特別変わった評価が出ているというようなデータを他店とまとめてしまうと分からなくなります。どちらのほうが適切かは一概に決まりませんので、個々の状況に応じて判断していくしかありません。

5.31 多重比較と一括比較の違い



多重比較：検出できない

一括比較：検出できる

多重比較：検出できる

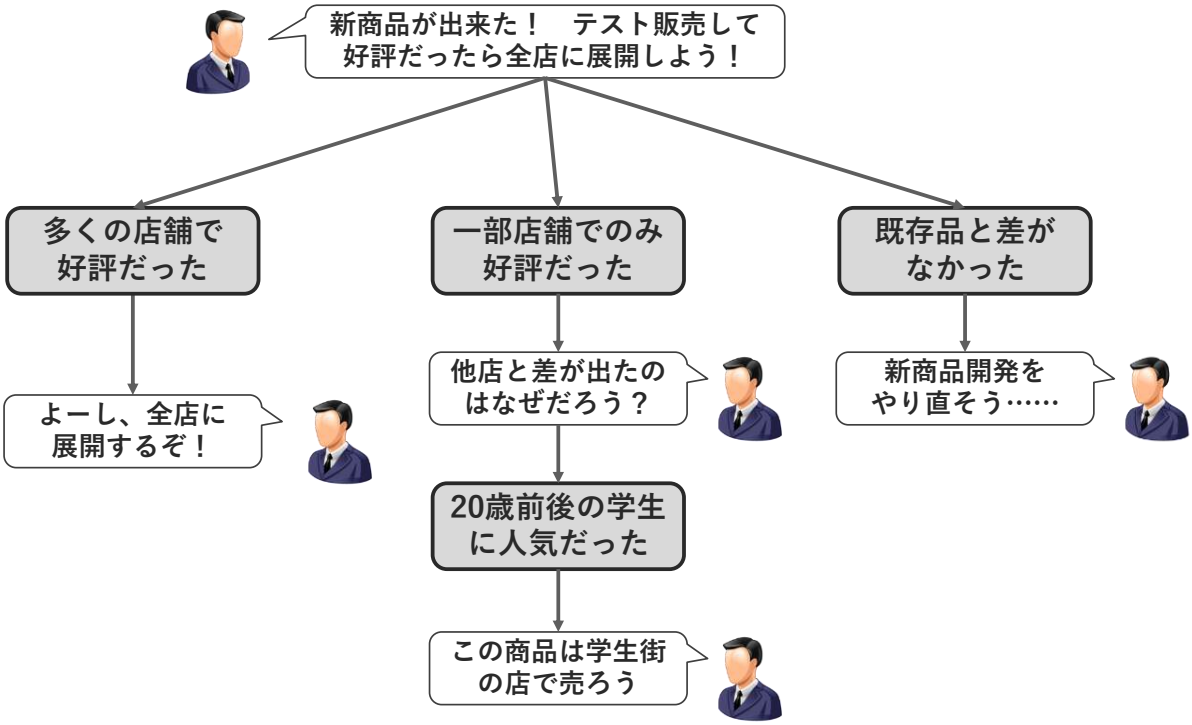
一括比較：検出できない

フレッシュジュース店の新商品を既存品と比較する調査をしたところ、「多くの標本で小さな差があった」としましょう。グラフにすると上図の左側のようなパターンで、ほとんどの標本で既存品と新商品の山が「少しだけズレている」タイプです。商品自体が実際に改善された場合はこのようなデータが出やすいですが、ボンフェローニ法による多重比較ではこの差を検出できず第 2 種の過誤になりやすい欠点があります。ボンフェローニ法では 1 回の比較での有意水準を厳しくするため、「小さな差」を検出できないからです。逆に、全標本のデータを一括して比較すれば検出できます。

次に、「一部の標本でだけ大きな差があった」ときを考えましょう。グラフにすると上図の左側のようなパターンで、ほとんどの標本では既存品と新商品の山がほとんど一致しているのに、一部の標本でだけ大きくズレているタイプです。このような場合、多重比較なら差を検出できますが、一括比較では検出できません。このパターンが出たときは B 店の条件だけ他店と何か違っていた可能性があります。たとえば「B 店にだけ、新商品を好むタイプの客層がいる」、「B 店の気温や湿度などの環境が新商品に合っていた」などです。

このように比較の方法によって向き不向きがあるため、それを理解して使い分けていく必要があります。

5.32 統計処理の目的は次の行動を決めること



あらためて確認しますが、統計処理の目的は「差がある/ない」という答えを出すことではなく、「次の行動を決めること」です。たとえば新商品が出来たので一部の店舗でテスト販売してみた結果、「多くの店舗で好評だった」のであればそのまま全店に展開できるでしょうし、「一部店舗でのみ好評だった」のであれば「なぜその差が出たのか？」を考える/調べるきっかけが得られます。その結果たとえば「好評な店舗では 20 歳前後の学生客が多く、そこで人気だった」ということが分かれば「この商品が学生街の店で売ればいい」とわかります。あるいは「まったく差が出なかった」のであれば「新商品開発をやり直そう」ということになるでしょう。

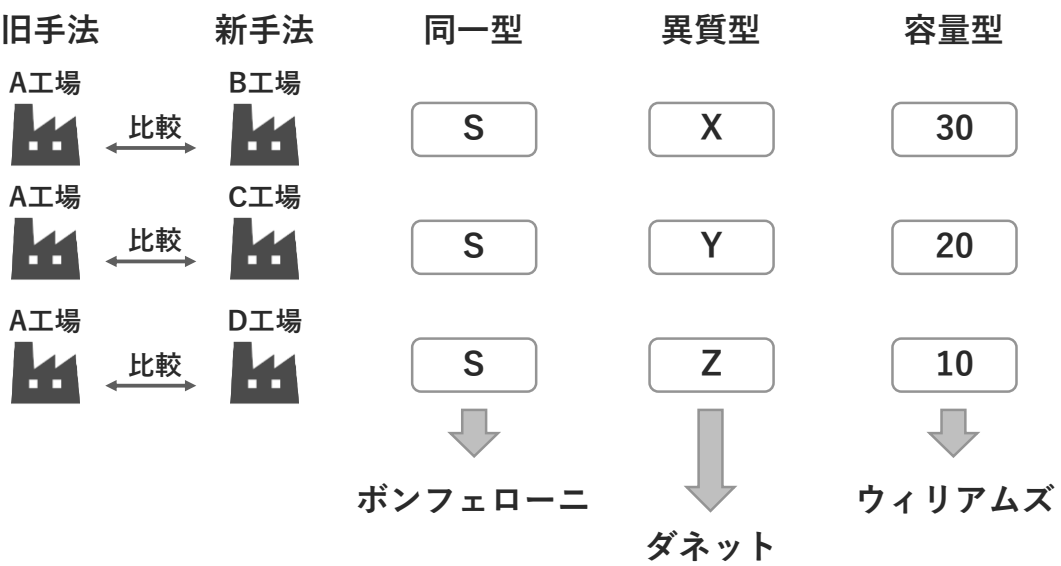
「全店に展開する」「なぜ差が出たのか調べる」「学生街の店で売る」「新商品開発をやり直す」……これらはいずれも「行動」です。これらの「行動」をして初めて経営を改善することができます。統計処理の目的はこうした「行動」を決めるための判断材料を得ることです。

前ページで書いたように、多重比較は「多くの店舗で少しだけ好評」というパターンを検出しにくく、一括比較は「一部店舗でのみ好評」というパターンを検出しにくい傾向があります。これらの特徴を理解して統計手法を使い分けることにより、「次の行動」を的確に決めることができます。

5.33 複数の対照実験の比較

【目的】A工場では既存手法、Ｂ・Ｃ・D工場では新手法を実験中。
新しい手法の有効性を確認するため、製品の品質を比較したい。

A を対照群、Ｂ、Ｃ、Dを実験群とする多重比較を行う



次は、「A 工場では既存手法、他の工場では新手法での生産を実験しており、新手法が有効かどうかを確認したい」という場面を考えます。「工場」に限らず、「農家が一部の畑を使って新しい栽培法を試す」「学習塾が一部の教室で新しい教材を使用する」「医療分野で新薬の試験をする」などさまざまな場面でも同様の状況があるため応用できます。「既存手法」とは今まで長年使われてきた実績のある手法なのに対して「新手法」はその実績がないため、実験の結果ハッキリと効果が見いだせない限り新手法は採用されないと考えてください。

基本的には A を対照群、他の工場（B、C、D）を実験群とする多重比較を行います。つまり比較の片方は必ず既存手法を使っている A 工場であり、他方が B、C、D と変わります。ただし、「新手法」といってもいくつかのバリエーションがあり、それぞれ適した方法が違います。ここでは、「同一型」「異質型」「用量型」という 3 種類を説明します。

「同一型」とはすべての実験群で同じ方法（S）を使用するもので、「新手法」が既にある程度確立して全面的に展開する前の最終テストのような局面でよく使われます。　多重比較の方法としては既に説明したボンフェローニの方法がよく使われます。他に、テューキーの方法、シェッフェの方法などがありますが本書では省略します。

「異質型」とはそれぞれの実験群で違う方法（X、Y、Z）を使用するもので、「新手法」の開発初期

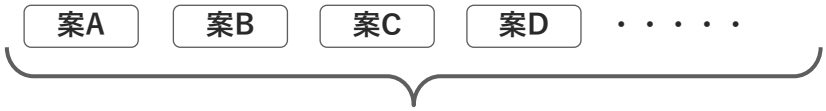
の段階、いくつかの候補の中から有望なものを探そうとするときによく使われます。たとえば料理のレシピで使用する食材をこれまでとは違うものに変更しようとするなら、品種 X、Y、Z など有望そうな候補をいくつか選んですべて試してみることでしょう。その場合、どれがどのぐらい良かったのか/悪かったのかを知る必要があります。この用途に向いている多重比較の方法として「ダネットの方法」があります。

「用量型」とは、それぞれの実験群で使う方法自体は同質で、ただし「用量」が異なるものです。料理のレシピに例えるなら、これまでは使っていなかった新しい調味料を加える際、その量を 30g、20g、10g と変えてみてどれが最適かを試すような場面です。この方法がよく使われるのは医薬品です。医薬品は容量が少ないと効果がなく、ある量以上で効果が出始め、さらに増やすとそれにつれて効き目も増すものの無限に増えるわけではなく、あるところで打ち止めになることが多いものです。つまり「使う量（用量）の大小に連れて効果が変わる」ことが仮定できる分野です。この用途に向いている多重比較の方法としてウィリアムズの方法があります。

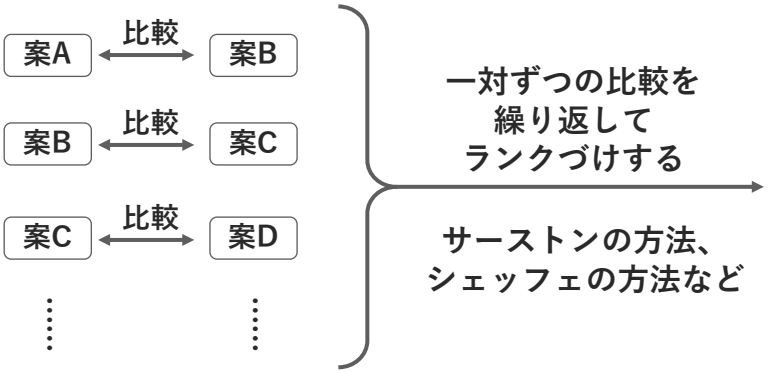
ダネットやウィリアムズの方法の具体的な手順は煩雑なため本書では省略します。多重比較には他にもさまざまな方法があり、それぞれ向き不向きがあるため、必要に応じてより専門的な書籍等を参照し学んでください。

5.34 一対比較法

【目的】新製品のパッケージデザイン案を数種類考えた。
その中で最も印象の良い案を採用したい。



多数の案すべてを1度に見て「良いものを選ぶ」のは難しいので



順位	案	スコア
1	C	92
2	B	78
3	D	70
4	A	56
5	E	24

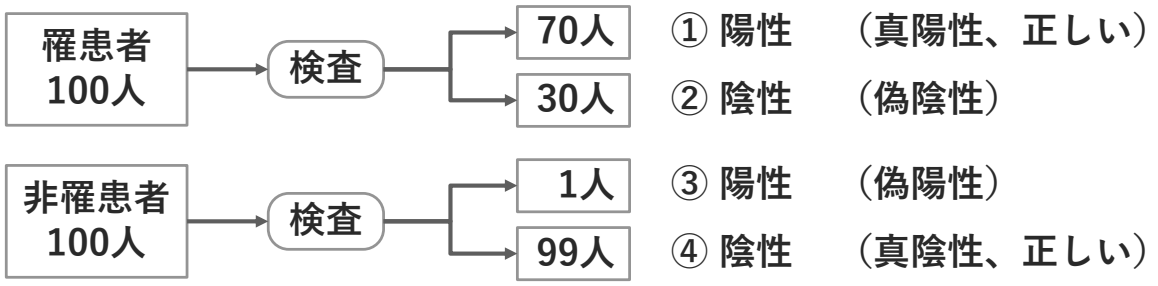
商品のパッケージデザインを考えるような場合、コンセプトの異なる案をいくつか作ってその中から最も好感度の高いものを選びたいですが、案の数が多い場合はそれらすべてを 1 度に見て「一番良いものを選ぶ」のは難しいことが知られています。たとえば 5 案を一列に並べて「一番良いものを選んでください」と指示すると、両端の案は目に止まりやすく、中間に並んだものは目立たない傾向があります。一列に並べるのではなく、カードに印刷して一枚ずつ見てもらったとしても同じで、「並び順」が評価に大きく影響しがちです。

そんなときに使われるのが「一対比較法」で、「A と B ではどちらがいいですか？」「B と C では？」「C と D では？」のように「一対ずつ見て良いほうを選ぶ」という比較を総当たりで行い、そのデータをまとめて「一番評価が高いのは C 案」のようにランクづけを行う方法です。

比較をする際に「単に良い方を選んで○×をつける」方式と、「どちらがどのぐらい良いか」という採点をする方式の 2 種類があり、それぞれサーストンの一対比較法、シェッフェの一対比較法と呼ばれています。

5.35 偽陽性と偽陰性

ある病気を診断する検査が感度70%、特異度99%だとするとその意味は……



2020 年には新型コロナウィルスの影響により「PCR 検査」が社会的に大きな話題になりました。ここでその種の検査にともなう「偽陽性・偽陰性」という問題について知っておきましょう（既に出てきた「第 1 種過誤・第 2 種過誤」と本質的に同じ問題です)。

ある病気を診断する検査があり、その精度は「感度 70%、特異度 99%」だとしましょう。「感度」とは、「実際に病気にかかっている人（罹患者）が検査を受けたときに陽性（病気にかかっているという意味）の結果が出る確率」です。たとえば罹患者 100 人が感度 70%の検査を受けると、①そのうち 70 人は正しく陽性と診断されます（真陽性）が、②残る 30 人は陰性（病気にかかっていないという意味）の結果が出ます。これを「偽陰性」と言い、正しくない結果です。

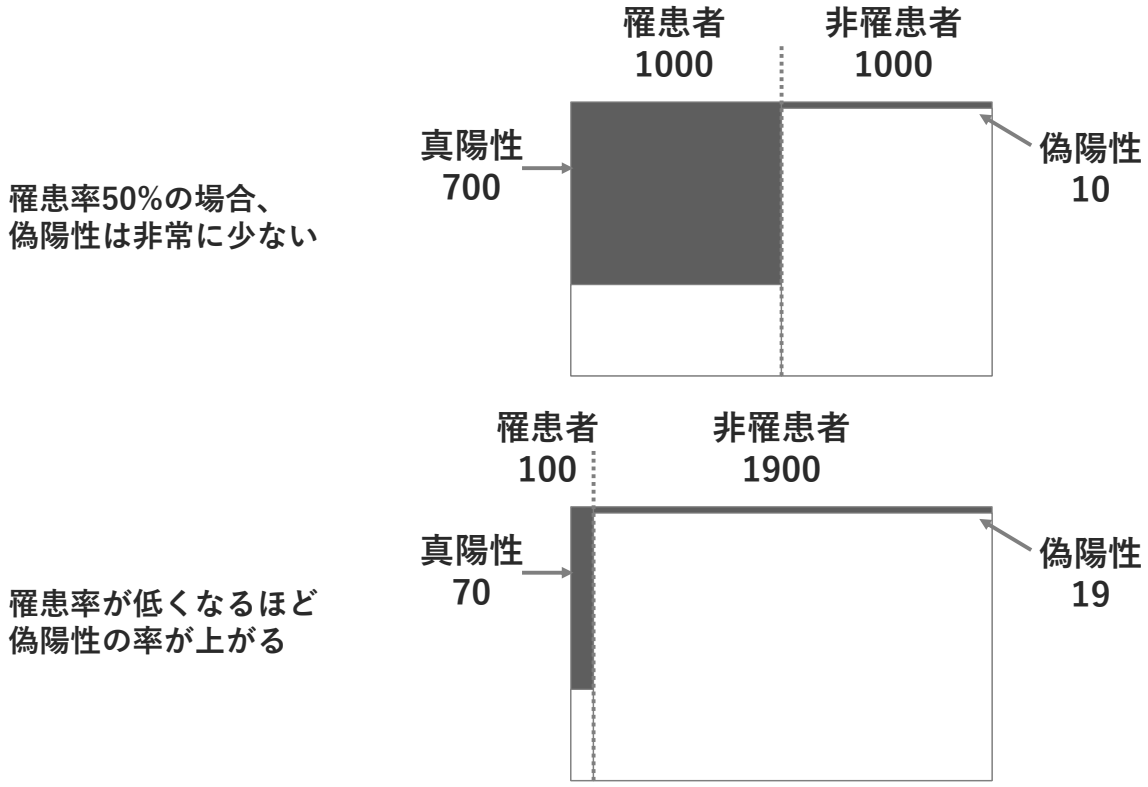
一方、「特異度」とは、「病気にかかっていない人（非罹患者）が検査を受けたときに陰性の結果が出る確率」です。たとえば非罹患者 100 人が特異度 99%の検査を受けると、③そのうち 1 人が陽性、④99 人が陰性と診断されるということを意味します。③は正しくない結果であり、「偽陽性」と呼ばれています。

では、1 万人のうち 100 人がその病気にかかっているとして、1 万人全員が検査を受けたとき、陽性と診断されるのはそのうち何人でしょうか？　「実際に病気にかかっている人」が 100 人なのですから、その 100 人が陽性と診断されると思いがちですがそうはなりません。

罹患者 100 人のうち実際に陽性と診断されるのは 70 人です(A、真陽性)。一方、非罹患者が 9900 人いてそのうちの 1%、つまり 99 人が陽性と診断されてしまいます（C、偽陽性)。合計で 169 人が陽性と診断されますが、そのうち本当の罹患者は 70 人しかいません。つまりこの場合、陽性となった人のうち本当の罹患者は半分以下しかいないため、「陽性」という結果が出てもあまり当てになりません。さらに 30%は偽陰性が出るため、罹患者 100 人のうち 30 人は見過ごされます(B、偽陰性)。

このイメージをつかむためには下記の図を見てください。罹患率 50%（罹患者・非罹患者が同数ずついる）の集団 2000 人に対して感度 70%、特異度 99%の検査を行うと、真陽性 700 人に対して偽陽性は 10 人と非常に少なくなるため、「陽性」は非常に高い確率で「実際に罹患している」ことを意味します。ところが、罹患率 5%の集団では真陽性 70 人に対して偽陽性が 19 人と、偽陽性の率が大幅に上がります。

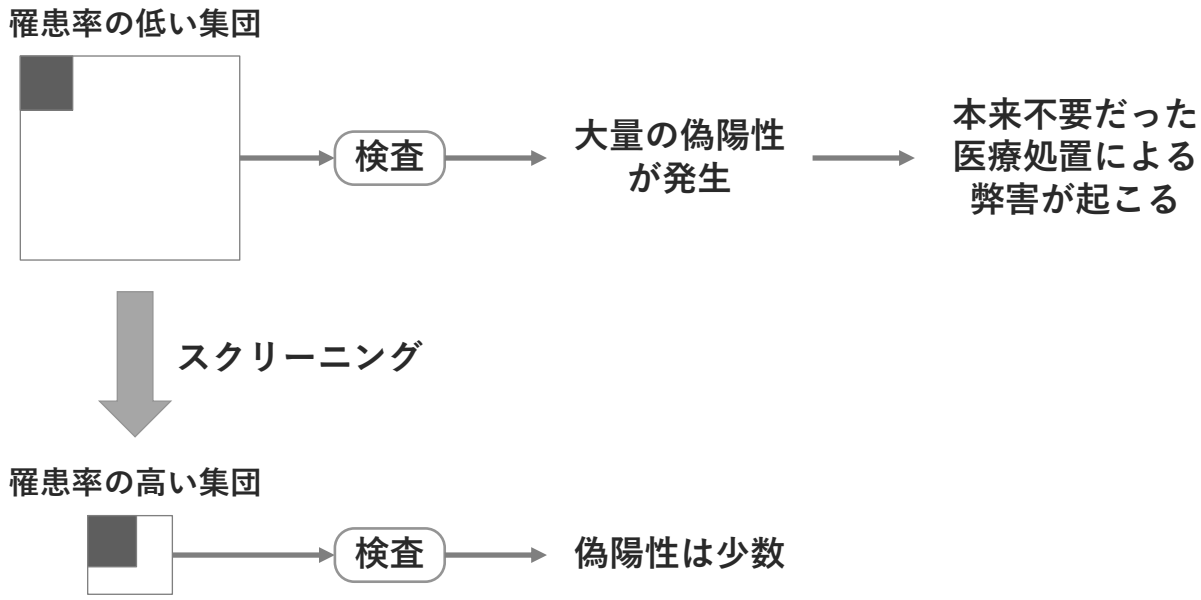
感度70%、特異度99%の検査を行う想定での陽性・偽陽性の比率



つまり、「罹患率の低い集団に対して検査をして陽性が出てもあまり当てにならない」のです。これは非常に多くの医学検査につきまとう問題で、偽陽性・偽陰性をゼロにすることはほぼ不可能です。

なお、これは既に出てきた「第 1 種過誤・第 2 種過誤」と本質的には同じ問題で、「偽陽性」は第 1 種の過誤、「偽陰性」は第 2 種の過誤に該当します。

5.36 偽陽性を減らすには？



罹患率の低い集団に対して検査を行うと、大量の偽陽性が発生し、本来不要だった医療処置による弊害が起きてしまいます。たとえば「胃がんの疑いがある」と診断されたら、レントゲン・内視鏡・CT などによる検査を追加で受けなければなりません。お金も時間もかかりますし、身体にも負担がかかる場合があります。偽陽性でそのような負担を負う人が増えるのは防ぐべきです。

偽陽性を減らすためには検査対象を「罹患率の高い集団」に絞ります。たとえば「肺がん」は若い人にはほとんど起きないので 40 歳以下の人に対して肺がん検診を行う必要はありません。年齢に限らずさまざまな条件で「その病気に罹患している可能性が低い人々」を検査対象から除外することを「スクリーニング」と言います。スクリーニングにより罹患率の高い集団に絞り込んでから検査をすると偽陽性を減らせるため、「本来不要だった医療処置による弊害」を防ぐことができます。

偽陽性・偽陰性は医学検査での用語ですが、本質的には第 1 種の過誤・第 2 種の過誤ですので、統計的な判断を行うあらゆる分野で起きる問題だと考えてください。一般の仕事でもたとえば「A という条件が満たされた時は B の処置を行う」のように、「条件→処置」という業務マニュアルを組んでいる場合がありますが、「条件」を「検査陽性」と読み替えれば医学検査と同じ構造です。そこでたとえば「A という条件」の範囲を「念のために……」と安易に広げてしまうと、「本来必要なかった余計な仕事を増やす」結果を招きます。これは偽陽性の問題とまったく同じです。このような落とし穴にはまる事態を防ぐためにも、統計の基礎をしっかりと身につけておいてください。

5.37 用語集

帰無仮説	仮説検定を行う際に設定する、おそらく棄却される（否定される）であろう仮説。
対立仮説	帰無仮説が棄却されたときに採択（採用）される仮説。
棄却・採択	厳密には少し違うが、「棄却」は否定、「採択」は採用や肯定と考えてよい。仮説検定では、否定・採用の代わりにこの用語を使う。
有意水準	仮説の棄却と採択を分ける境界となる基準値。通常 5%または 1%を使うことが多い。
p 値	ある現象が「どの程度の確率で起こりうるか」を表す数値。p 値を計算し有意水準と比較して仮説の棄却・採択を決定する。
片側検定・両側検定	有意水準を正規分布の片側のみで設定するのが片側検定、両側で設定するのが両側検定。
信頼区間	標本をもとにして母集団の平均値（母平均）を推定するとき、母平均は「おそらくこの上限と下限の間に存在するだろう」と思われる上限と下限の範囲をいう。実際は「おそらく」ではなく確率を指定して「母平均の 95%信頼区間は A～B である」のように使う。
信頼係数	なんらかの推定を行う場合、その推定がどの程度確からしいかを表す数値をいう。信頼区間を推定する場合の「母平均の 95%信頼区間は A～B である」という用法の「95%」が信頼係数である。
t 検定	2 つの母集団から抽出した標本を元に、その標本が属す母集団の平均に差があるかどうかを判断する検定。
t 分布表	t 検定で用いる「t 値」をあらかじめ計算した表。
不偏分散	標本の分散をもとにして母集団の分散を推定した量
分散分析	それぞれ異なる母集団から抽出した 3 群以上の標本をもとに、母集団の平均値が一致しているかどうかを分析する方法
対照群と実験群	ある目的を達成するための新しい方法が有効かどうかを判断する際、既存の方法を使う対照群と新しい方法を使う実験群を作り、他の条件を同じにして比較する
プラセボ効果	偽薬でも「これは良く効く薬です」と言われると実際に薬効が出てしまう現象
観察者バイアス	対照実験の結果を自分が期待する方向に解釈してしまうバイアス。定性的な解答を数量化する際に起こりやすい。
第 1 種過誤・第 2 種過誤	実際には有意差がないのに「有意差あり」と判断してしまうのが第 1 種過誤、実際には有意差があるのに「有意差なし」と判断してしまうのが第 2 種過誤。

多重比較	複数回の比較を行って 1 つの結論を出す方法。有意水準の調整などに注意が必要。細かな条件応じて多様な方法がある。
一対比較法	複数のデザインを比較してランクづけをする場合に、一対ずつの比較を繰り返してその結果を集計することにより行う方法。
偽陽性と偽陰性	医学検査では第 1 種過誤・第 2 種過誤をこの名前で呼ぶ

5.38 確認問題

【問 1】サイコロを 4 回続けて振ったところ、すべて 6 の目が出た。偶然このような現象が起こる確率を求めなさい。

【問 2】日本人男性の平均身長が 170cm、標準偏差が 6 であるとする。ある店への来店客のうち男性を 20 人無作為に選んで身長を測ったところ、182cm 以上の人が 4 人いたとする。帰無仮説を「この店へ来る客の身長の分布は日本人男性全体と一致する」として、「20 人のうち 4 人が 182cm 以上」という現象の p 値を計算せよ

【問 3】第 1 種過誤・第 2 種過誤がそれぞれどれに該当するか、選択肢の中から選びなさい

【選択肢】

（A）ケーキ店で「とても甘い新品種のイチゴを使いました」と言ってショートケーキを販売したところ、「本当だ、以前より甘いですね！」と好評だったが、実際には誤って従来と同じイチゴが使われていた。

（B）九州の食品会社が主に東京の店で販売する新商品を開発したが、地元では非常に好評だったものの東京ではあまり人気が出なかった。

（C）新薬の効き目を既存薬と比べる対照実験を行ったところ、新薬の方が効果が高いとは言えないことがわかった。しかしその後の調査により、新薬の効き目は実際には既存薬よりも高いことが分かった。

（D）新薬の効き目を既存薬と比べる対照実験を行ったところ、新薬の方が効果が高いという結果が出た。しかしその後の調査により、新薬の効き目は実際には既存薬と変わらないことが分かった。

■確認問題解答

【問 1 解答】

1296 分の 1

1 回振って 6 の目が出る確率は 1/6、それが 4 回続く確率は $1/(6\times6\times6\times6)=1/6^4=1/1296$ となる。

【問２解答】

p 値=0.0013 （0.13%）

182cm は平均＋2σに該当し、これ「以上」となる人口の比率は 2.5%。そこで、20 人中 4 人が 182cm 以上となる確率は、1 回につき 2.5%成功する試行を 20 回繰り返して 4 回成功する確率として 2 項分布により計算できる。これを計算すると

Excel での計算例：BINOM.DIST(4, 20, 0.025, FALSE) = 0.001262 = 0.13%

したがって p 値=0.13%となるため、有意水準 5%でも 1%でも帰無仮説は棄却される。

【問３解答】

第 1 種過誤：D，第 2 種過誤：C

なお、A はプラセボ効果、B は標本抽出の偏り。

6 仕事と集合Ⅲ

6.1　A 店でよく売れているのはなぜだろう？

コンビニのように支店をいくつか持つチェーン店のうち1店で、Xという商品がよく売れているそうです。なぜA店でだけ売れているのでしょう？

各店舗の基礎データとX商品の売上

	A店	B店	C店	D店		相関係数
X商品売上	100	80	82	75		
来店客数	4200	4300	4200	4100		0.19
客単価	670	660	680	620		0.56
年齢：平均値	40	42	41	39		-0.04
年齢：標準偏差	12	14	14	11		-0.08
年齢：中央値	39	40	42	37		0.11
駐車場の広さ	3	1	2	0		0.91
商圏人口	3300	3200	2900	3400		0.06

基礎データを見てもA店でXが売れる理由はよくわからない

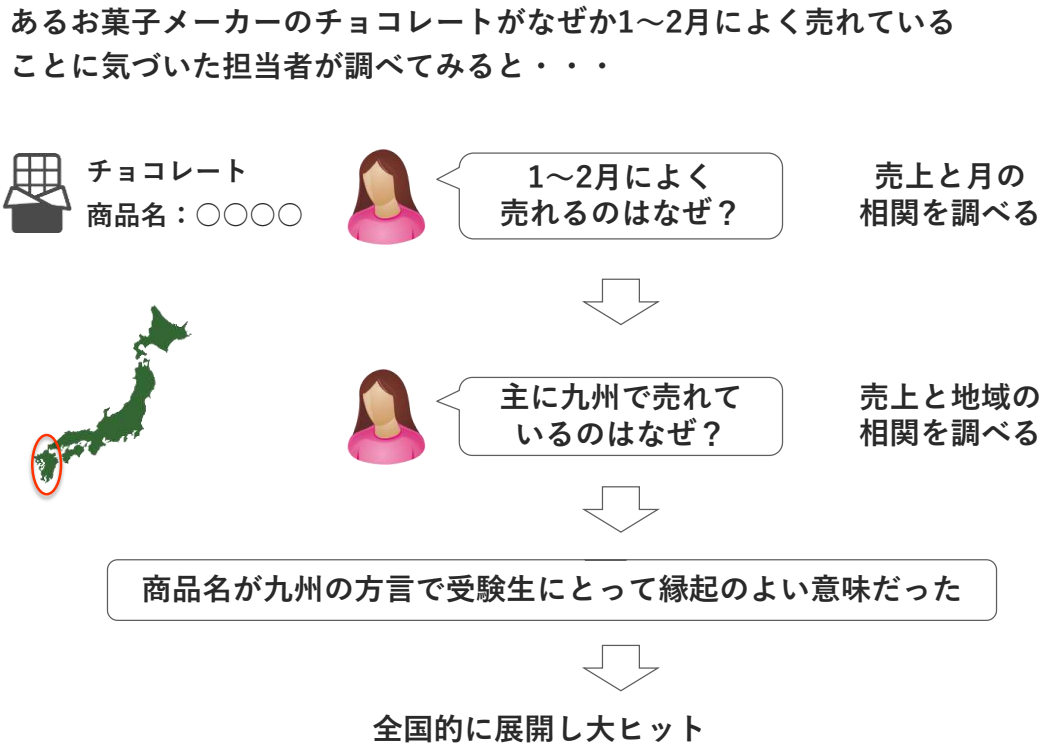
しかし、相関係数を取ってみると・・・

コンビニのように支店をいくつか持つチェーン店のうち 1 店で、X という商品がよく売れているそうです。なぜ A 店でだけ売れているのでしょう？

来店客数や商圏人口といった基礎データを見ても、A 店でだけ売れる理由はよくわかりません。「特定の商品が一部の店でだけよく売れる」というケースはよくありますが、食品や雑貨、衣料品などを売っている普通の小売店には何千種類もの商品があり、よほど目立つものでなければ、特定の商品が売れた理由を調べることはないでしょう。そもそも、上図の表のように売上が A 店の 100 に対して他店は 80 前後という程度では、「A 店でよく売れている」ことに気づいてさえいないことも多いものです。

しかし、もしその理由が分かれば適切な手を打てて B～D 店でも売れるようになるかもしれません。実はそんな「なぜ売れている？」という疑問への答えを探るためによく使われるのが相関分析という手法です。上図右側に「相関係数」という列がありますが、これは「X 商品売上」と「来店客数、客単価、年齢・・」などの基礎データにどのぐらい「関連性があるか」を計算した数値です。相関係数について詳しくは後述しますが、見ると「駐車場」の項目で 0.91 という最も大きな数字が出ています。となると X 商品は駐車場が多い店で売れる＝車で来店する客が買っているのかもしれません。だとしたら、他にも駐車場の多い店を中心に販促に力を入れることで売上を増やせる可能性があります。このような手がかりを見つけるのに役立つのが相関分析です。

6.2　異変に気づけばヒットにつながる

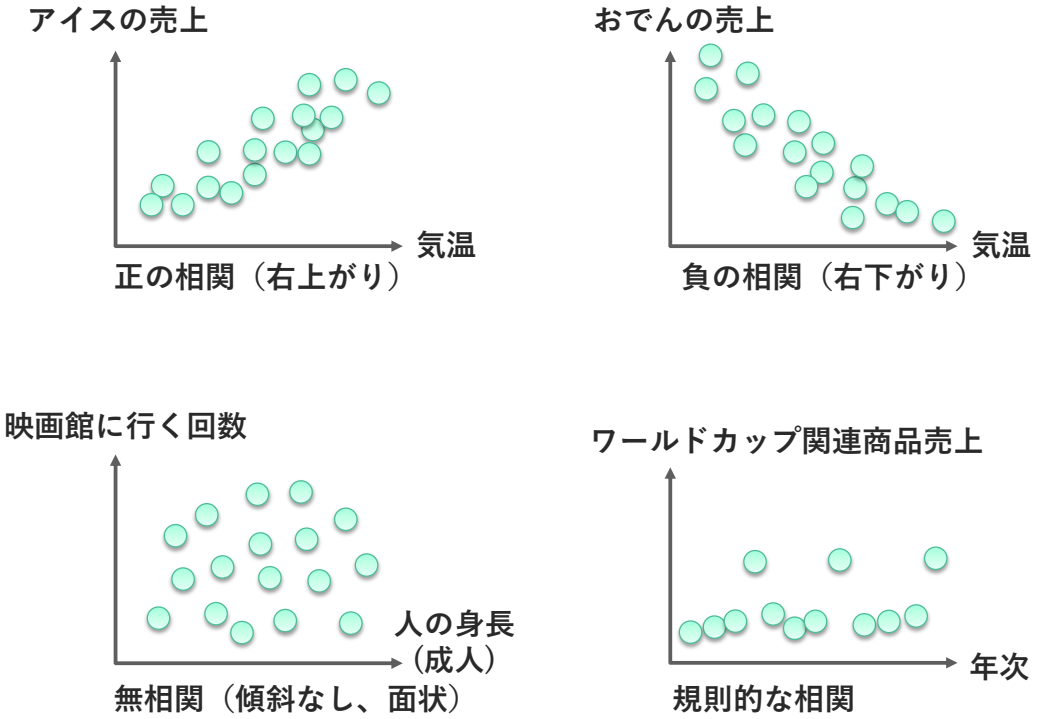


あるお菓子メーカーのチョコレートが特別な施策もしていないのになぜか 1～2 月によく売れているという現象がありました。担当者が不思議に思ってさらに調べてみるとその売上のほとんどが九州であることがわかりました。さらに調べてみると、商品名が九州の方言で受験生にとって縁起の良い意味になっていたために、受験シーズンに縁起物として大量に買われていたことが分かりました。

その後、そのメーカーはそれをヒントに商品パッケージを変え広告キャンペーンも変えて全国的に展開した結果大ヒット商品となりました。

しかしその大ヒットのきっかけは九州でのちょっとした異変でした。商品別の売上と月の相関を調べる、地域との相関を調べる、などの方法で、そんな「ちょっとした異変」に気づくことができます。「相関係数」は簡単な計算で出せますから、生の数字だけでは見逃してしまいそうなちょっとした異変を見つけるために活用できます。

6.3　相関分析



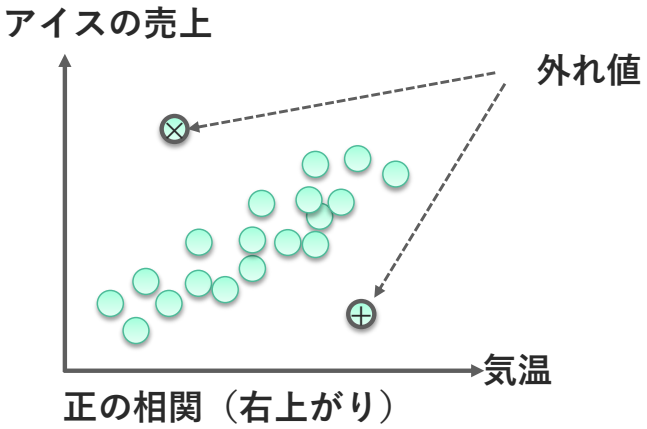
「相関」とは「相互に関連がある」という文を略した用語です。たとえば「夏の暑い日にはアイスクリームがよく売れる」ことから、「アイスの売上と気温の間には関連（相関）がある」と言えます。実際、ある店の一か月間の記録をもとに、一日の平均気温を横軸に、その日のアイスの売上を縦軸にとって散布図を描くと上図左上のようになりました。直線に近い右肩上がりの傾向がはっきりとみて取れます。このように、ある変数（気温）が増えると他方の変数（アイスの売上）も増える、というような関係を「正の相関」と言います。逆に寒い時期のほうがよく売れる「おでん」について同じことをすると上図右上のように直線に近い右下がりの傾向が見えます。この場合は「負の相関」があると言えます。

一方、人の身長（成人）とその人が映画館に行く回数にはほとんど関係がない（無相関である）と考えられます。このような場合は上図左下のように散布図では面上に広がるなど、傾斜のない形になります。

相関の中には右下のように「右上がりでも右下がりでもなく、面状にも広がらず、規則的に変化する」タイプのものもあります。たとえばサッカーのワールドカップ関連商品は４年に一度の開催年によく売れるのでこのような形になります。しかし、右上がりまたは右下がりの「直線的な関係」は最も単純でかつ応用する機会も多いので、単に「相関」というとそれらの関係を表している場合がよくあります。本書でもその用法を使うため、これ以後単に「相関」といえば直線的な関係のことであると考えてください。

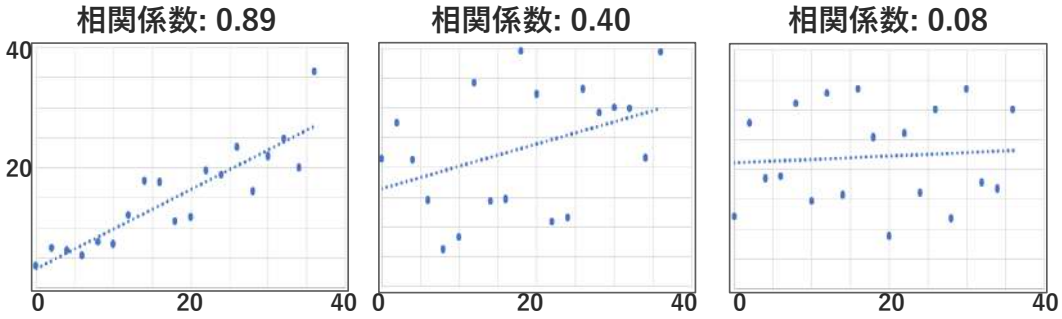
散布図を描くとこのような「複数の変数の関係」を視覚的に確認できるため、「あ、AとBは正の相関があるのか。じゃあ、Bが増えそうなときにはA商品を多めに仕入れておくべきかもしれない」といった予測・判断に役立ちます。

散布図を描くことには「外れ値」を見つけやすくなるというメリットもあります。
「外れ値」とは「普通とは違う異常なデータ」のことで、下記の図であれば×と＋の記号つきの点が「外れ値」です。



「外れ値」がなぜ起きたのかを調べると経営改善のヒントが得られることも多いので、散布図を描いてみて外れ値があった場合はその理由を探ってみましょう。

6.4　相関係数・回帰式



回帰直線
$y = 0.66x + 3.1$

回帰直線
$y = 0.37x + 16$

回帰直線
$y = 0.06x + 21$

相関係数は-1～1の値を取る

-1または1に近いほど相関が強い。0は無相関

ばらばらの点の動きに最も近い一次関数が回帰直線

「2 つの変数の間に関連がある」といっても、ではどの程度の強さの関連なのか、という「程度」の差はあります。その相関の強さを数値化したものが「相関係数」です。

上図では 3 つの散布図の例を示していますが、左側ははっきりとした右上がりの傾向が見えるのに対して中央は「右上がりと言われればそう見える」程度、右側は「上がっても下がってもいないように見える」グラフです。これらの散布図について「どの程度相関が強いかを示す相関係数」をそれぞれ計算すると 0.89, 0.40, 0.08 になります。相関係数は±1 の範囲の値をとり、1 に近いほど強い正の相関、-1 に近いほど強い負の相関、0 に近いほど無相関であることを示します。0.89 は+1 に近いのではっきりとした右上がりが見えるのに対して、弱い相関だと散布図でも右上がりの形が見えにくいものです。実際、中央(0.40)と右(0.08)の散布図を比べて見ても、中央のほうが相関が強いとはわかりにくいですね。しかし数字であれば 0.40 と 0.08 はハッキリとした違いなので、散布図ではわかりにくい「弱い相関」も発見することができます。

一方、点線で引いた「回帰直線」は、散布図の点の分布にできるだけ近くなるように 1 次関数の線を引いたものです。左側の散布図の中の回帰直線は「点」の動きと大筋合っていると言えるでしょう。相関の強いデータに関してはこのように回帰直線を引くことで「明日は気温 25 度の予想だからこのぐらい売れるだろう」といった予想の道具として使えます。相関が弱くても回帰直線を引くこと自体は可能ですが、中央や右のグラフで分かるとおり、その場合は回帰直線が現実と合わないケースが多発します。

6.5　相関係数の求め方

i	X	Y
1	10	8
2	20	15
3	30	24
4	40	27
5	50	39
6	60	67
7	70	97
8	80	62
相関係数		0.89

ピアソンの積率相関係数 r の計算式

XとYの共分散

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

Xの標準偏差　　**Yの標準偏差**

表計算ソフトでは CORREL 関数で計算可能

$$\boldsymbol{r = CORREL(X}\text{の配列}\boldsymbol{,Y}\text{の配列)}$$

相関係数の計算方法にもいくつかの種類がありますが、最もよく使われるのはピアソンの積率相関係数と呼ばれる方法です。たとえば上図左の表の X と Y のデータの相関係数は、右側の枠内の式で計算できます。式の意味としては「X と Y の共分散」を「X の標準偏差と Y の標準偏差の積」で割ったものです。

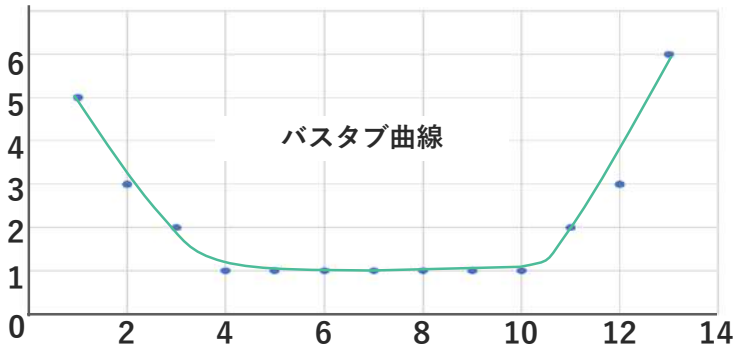
ただし実務的には表計算ソフトの CORREL 関数（具体的な名前はソフトによって違う可能性があります）で行うことが多く、この式を覚えておく必要はありません。CORREL 関数の引数に、X と Y のデータ配列を渡すとピアソンの積率相関係数が得られます。上図のデータの場合、X が増えるにつれて Y も増える傾向にあるため、相関係数は 0.89 という 1 に近い数字が得られ、強い正の相関があることが分かります。なお、相関係数の値は一般的には次のように解釈されます。

0.7 ～ 1.0 かなり強い正の相関がある
0.4 ～ 0.7 正の相関がある
0.2 ～ 0.4 弱い正の相関がある
-0.2 ～ 0.2 ほとんど相関がない
-0.4 ～ -0.2 弱い負の相関がある
-0.7 ～ -0.4 負の相関がある
-1.0 ～ -0.7 かなり強い負の相関がある

6.6　相関係数では相関がわからないケース

X	Y
1	5
2	3
3	2
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	2
12	3
13	6

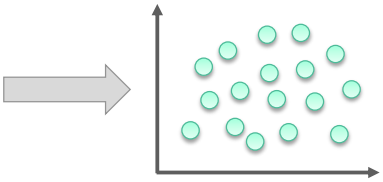
散布図が
直線状に
ならない



相関係数を計算すると0に近い

相関係数＝0.08（ほぼ無相関）

しかし本当に無相関なら
散布図はこんな形のはず



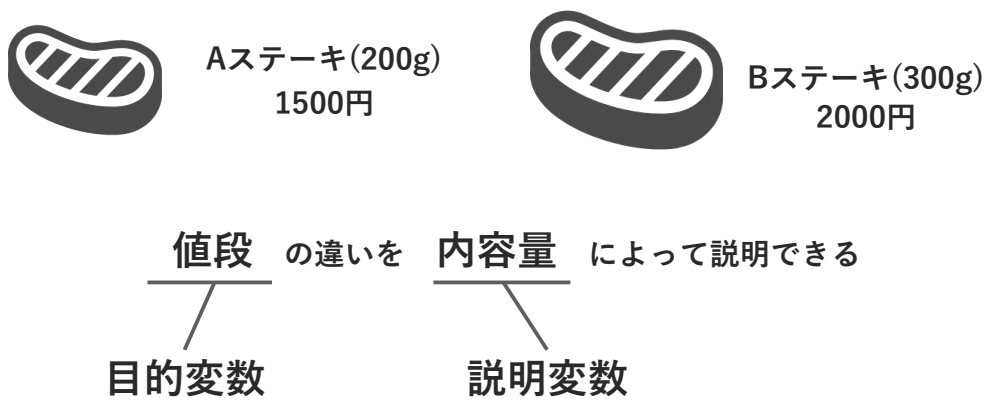
「相関係数」は複数のデータに関連があるかどうかを手軽に調べられる便利な方法ですが、弱点もあります。上図左側の表にある X と Y のデータについて散布図を描くと直線上にはならず、右上の図のように両側が上がって中央が凹んだ形になります。この形はバスタブ（浴槽）を横に見た形に似ているので「バスタブ曲線」と呼ばれています。実例としては、機械類の故障率と使用期間がこの種の曲線を描くことが知られています。機械類は出荷直後に初期不良が見つかり、寿命が近くなると部材の劣化による故障が増えるので、使用期間の初めと終わりに故障が出やすいためです。

ところが、この X、Y に対して一般的なピアソンの積率相関係数を計算すると 0.08 とゼロに近い値となり、「ほぼ無相関」という結果になります。しかし本当に無相関ならば散布図は右下のように平面状に散らばった形になるはずで、バスタブ曲線にはなりません。

「故障率」と「使用期間」には明らかに相関がありますが、その関係は直線状でないため一般的な相関係数の計算式では相関を検出できません。ピアソンの積率相関係数は一次関数で表せる直線状の相関を検出するもので、バスタブ曲線のように二次関数的な関係性は分かりません。

バスタブ曲線のようなデータに対して関係性を調べる場合、使用期間の前半と後半に分けて計算したり、あるいはピアソンとは別の計算方法を使う必要があります（ただし本書ではその具体的な方法については省略します）。

6.7　説明変数と目的変数



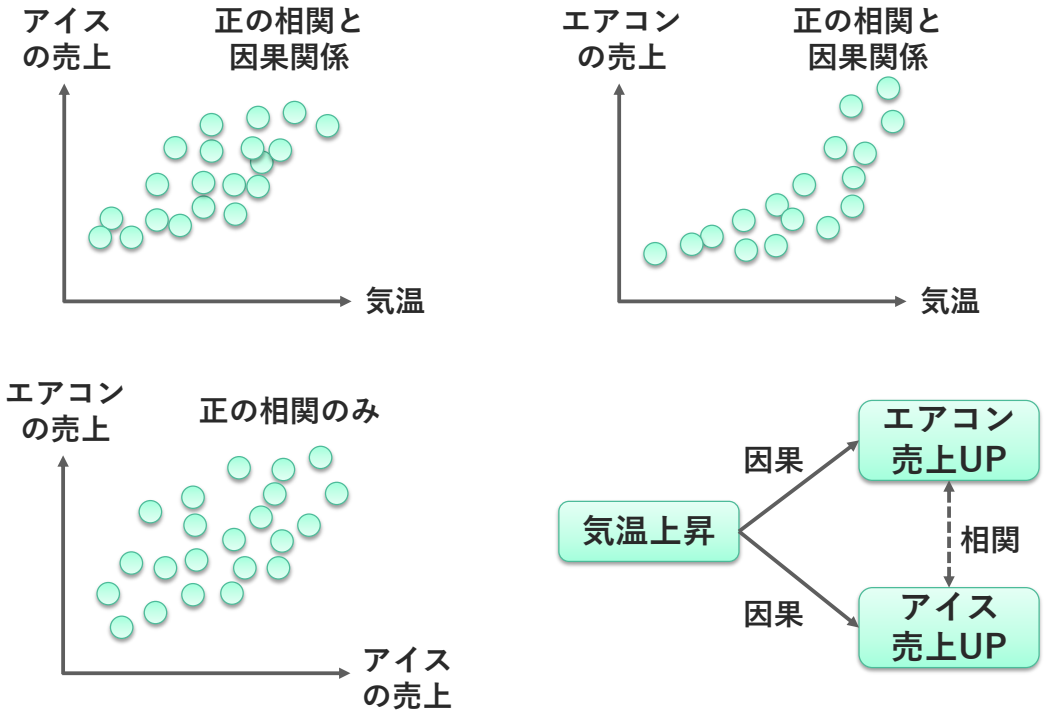
因果関係が存在する場合が多いが、必ずそうなるわけではない

相関分析の話をする前に、「説明変数と目的変数」という考え方を知っておきましょう。

同じ店で出しているAステーキよりもBステーキのほうが値段が高いのはなぜ？　と問われたら、単純に「Bステーキのほうが大きいから」というのが自然な答えでしょう。これは値段の違いを内容量によって説明しているもので、ここでいう「値段」を目的変数、「内容量」を説明変数と言います。

本章の最初に出てきた「なぜA店でX商品が売れるの？」という疑問の例で言えば、「X商品の売上」が目的変数、「駐車場の広さ」が説明変数です。目的変数と説明変数の間には因果関係が存在する場合が多いですが、必ずそうなるわけではありません。

6.8　相関関係と因果関係

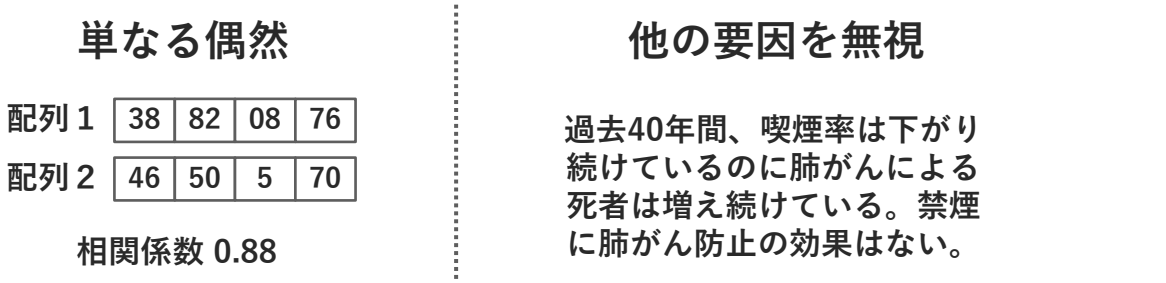


相関関係と因果関係の違いはよく誤解されています。あるお店のアイスクリームの売上と、その日の平均気温データを数十日分記録してその組み合わせを散布図にすると図中左上のようになったとします。全体として右上がりの「正の相関」を示しています。「暑くなると冷たいデザートが欲しくなる」のは自然な感覚ですからおそらく因果関係もあるでしょう。

「気温とアイス」が典型的な「因果関係のある正の相関」であるのに対して、電機店では「気温が高いとエアコンが売れる」ことも知られています。これも「因果関係のある正の相関」なのは間違いありません。

では、「エアコンの売上」と「アイスの売上」で散布図を描いたらどうなるでしょうか？　実はこれもおおむね右肩上がりの正の相関を描きます。しかしこの３種類のうち、「気温とアイス」「気温とエアコン」には因果関係がありますが「アイスとエアコン」には因果関係がないことに注意してください。共通の原因（気温上昇）を持つ２つの変数（エアコンとアイスの売上）の間には相関がありますが、だからといってがんばってアイスクリームを多く売ってもエアコンの売上は増えません。当たり前すぎる話ですが問題が複雑になるとこのレベルの間違ったロジックで「がんばれ！」と無駄な努力をしているケースも少なくないので、因果と相関の違いは知っておく必要があります。

6.9　相関・因果関係を誤認するパターン



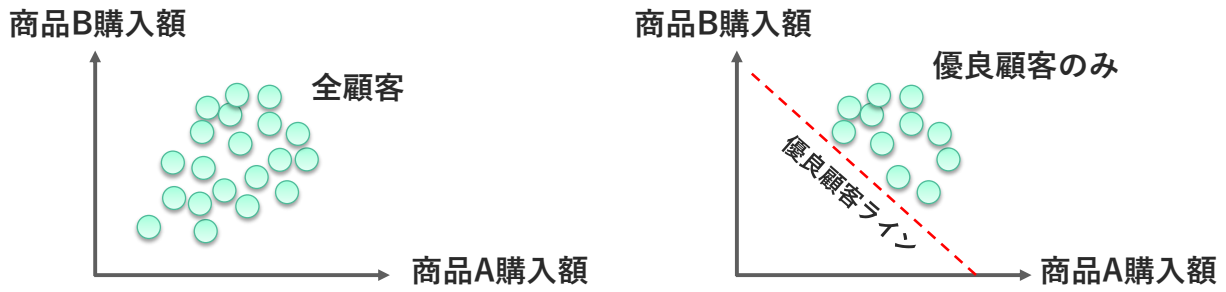
相関関係や因果関係を間違って認識するパターンでもっとも単純なのが「単なる偶然」です。上図左上の配列 1 と配列 2 について相関係数を計算すると 0.88 と非常に高い相関を示しますが、実はこれはある日のある道路上で 1 分間上り方面と下り方面へ向かう車のナンバー下 2 桁をとっただけで、何の関係もありません。単なる偶然でも相関係数が高くなりうる場合があるため、標本調査は何度も行って検証する必要があります。

「喫煙率は下がり続けているのに肺がんの死者は増え続けている」というのは医療不信の文脈で「喫煙→肺がん」の因果関係を否定するためによく言われているものです。実際には、「喫煙習慣→肺がん発病」の間には何十年単位の長い時間がかかるため、現在増えている肺がんの多くは何十年も昔に喫煙していた人に発症しています。また、医療が進歩して「がん以外の病気ではなかなか死ななくなった」ため、結果として肺がんによる死者が増えているように見えています。これらの要因を無視して単純な相関関係だけを見ても適切な判断はできません。

「因果関係を逆にとらえてしまう」間違いもよくあります。たとえば犯罪が多い地域では警察官が多く必要になるため、地域別に「人口に対する警察官の数」と「犯罪発生率」を比べると正の相関があります。もちろん、「犯罪が多い→警察官を増やす」という方向の因果関係が正しいのですが、これを「警察官が多いから犯罪が起きるのだ！」というような誤解をしてしまうケースがあります。

この他、「合流点での選別」にも注意しましょう。
下図左側はある店の全顧客に関して商品 A の購入額と商品 B の購入額で散布図を描いたものです。

6.10 合流点での選別



これを見るとAとBの購入額には正の相関があります。あるときその店ではAとBの購入額の合計が一定ラインを越えたら優良顧客として特別なサービスをすることにしました。そこで優良顧客のみを見るとAとBの購入額には負の相関があるように見えてしまい、まったく逆の判断をしてしまうかもしれません。

そうすると何が困るのでしょうか？　もし「AとBの購入額に正の相関がある」ならそれは「Aを買う客はBも買いたがる可能性が高い」ことを意味するので、「AとB、セットでお得キャンペーン」のような企画が有望です。一方、「AとBの購入額に負の相関がある」ならそれは「AとBの一方を買った人は他方を必要としない傾向がある」ことを意味するので、「今月の来店者にはAかBのどちらか1つをお値引きします」というような販促施策が有望です。相関が正か負かによって施策の打ち方が変わってくるので、その読みを間違えると適切な手を打つことができません。

このようなミスは2つ（以上）のグループを合わせてある条件で選別したときに起こりやすく、「合流点での選別」と言われています。「商品Aの購入」と「商品Bの購入」がそれぞれグループに当たり、「合計購入額が高い客を優良とする」が「合流点での選別」にあたります。

こうしたミスを防ぐために重要な方法の1つが、相関係数だけに頼らず、散布図を描いてみることです。既に書いた通り、「ピアソンの積率相関係数は一次関数で表せる直線状の相関を検出する」ものですので、一次関数にならない相関は検出できません。したがって、散布図を見たときに「大まかに直線状になっている」場合に限って使用するようにすれば、この種のミスはある程度防げます。(相関係数の計算方法にはその他の方法もありますが、本書では省略します)。

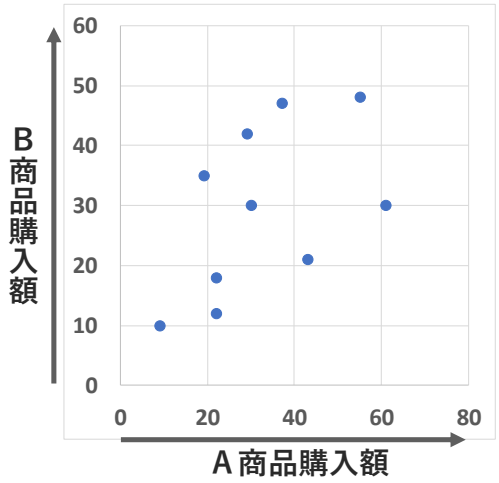
6.11 その相関係数に意味はあるか？

顧客別商品別半期売上

顧客	A 商品	B 商品
1	22	12
2	30	30
3	22	18
4	55	48
5	29	42
6	19	35
7	43	21
8	37	47
9	61	30
10	9	10

単位：千円

あるお店の顧客のうち10人のＡ商品購入額と
Ｂ商品購入額を一定期間集計した表と散布図



相関係数	0.52

← ＡとＢの購入には相関があると言えるか？

ここまで見てきたとおり、相関係数を計算して高い数値が出たとしても実際にその数値に意味があるとは限らないので、「その数値に意味があるかどうか？」は別途確かめる必要があります。散布図を描いてみるのはそのために役立つ方法の 1 つですが、場合によってはそれでも分からないケースがあります。

例として、あるお店の顧客から 10 人をランダムに抽出して、一定期間（例えば半年）のＡ商品購入額とＢ商品購入額を集計した場合を考えましょう。たとえば、ビールとワインが主力商品である酒屋さんで、Ａ商品＝ビール、Ｂ商品＝ワインだとすると、顧客の中には大まかに 3 種類の人がいそうです。

ビール派：ビールを中心に買い、ワインはあまり買わない
ワイン派：ワインを中心に買い、ビールはあまり買わない
両刀派：ビールもワインも同じぐらいに買う

両刀派が多い場合はビールとワインの売上には正の相関が出ます。その場合、ビールもワインも区別せずに「お酒の購入額が一定額を超えたら割引」のような販促策が使えそうです。

逆に両刀派が少ない場合はビールとワインの売上には負の相関が出ます。その場合はビール派向け

とワイン派向けで別な販促策を考えた方がよいかもしれません。

さて、それでは実際に 10 人の顧客を抽出して集計した前ページ記載のデータはどう解釈できるでしょうか？　相関係数は 0.52 と、強くはないが弱くもない正の相関がある、と出ています。これを素直に解釈すると「両刀派が一番多い」ことになります。

相関係数だけでは当てになりませんので散布図を描いてみると、大まかに右肩上がりの直線上になっていると言えそうです。これならピアソンの積率相関係数を使えるので、0.52 を素直に解釈しても良さそうに見えます。

しかし実はこの結論は間違いで、このデータについては相関係数 0.52 が出ていても「ビールとワインの売上に正の相関がある（＝両刀派が一番多い)」とは言えません。

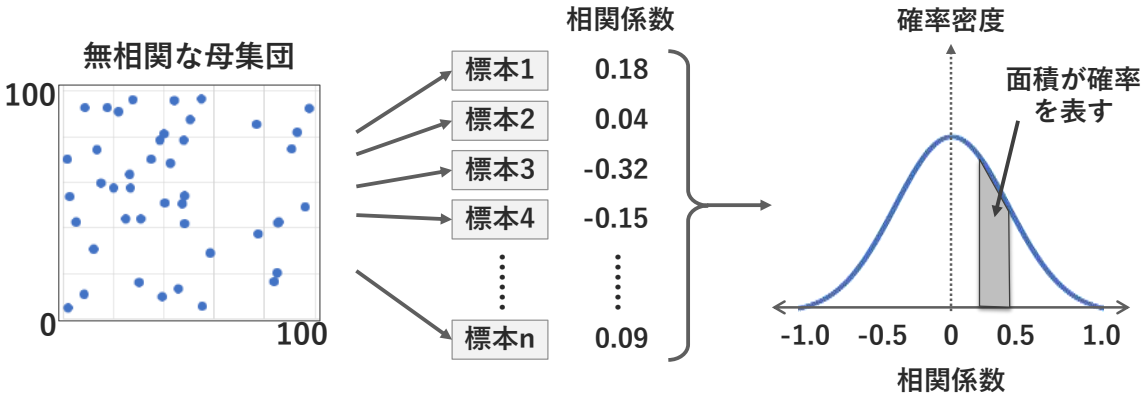
問題は「標本の数が少ない」ことで、10 件程度の標本数ではタダの偶然でも 0.52 という相関係数が出ることがあるため、通常使われる「相関係数が 0.4 〜 0.7 であれば正の相関がある」という判断基準が使えません。つまり「この標本数では、相関係数 0.52 には意味がない」のです。

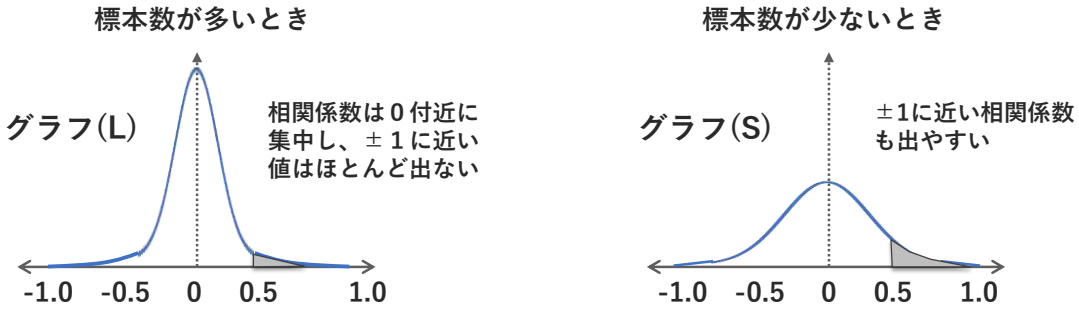
そこで次は、

- 標本数が少ないと具体的にどのような影響が出るのか？
- 標本数 10 件の場合、相関係数がどの程度あれば意味がある（相関があると判断できる）のか？

を考える必要があります。

6.12 標本の相関係数の分布を考える





まずは標本数の大小によってどのような影響が出るかを考えましょう。
まったく相関のない母集団から一定個数の標本を取って相関係数を調べる、という調査を何度も繰り返したとします。

母集団が完全に無相関なので、標本を取っても相関係数は 0 になる……とは限りません。サイコロを 6 回振ったら 1～6 の数字が 1 回ずつ出るとは限らないように、標本抽出にはある程度の偏りが出るのが普通です。

とはいえ、偏りが出る確率は小さいので、標本調査を何度も何度もやったときに得られる相関係数は 0 に近いほど多く、±1 に近いほど少なくなります。そこで、標本調査を無数に繰り返して出た相関係数の数をグラフ化すると、上図右上のような正規分布に似たグラフになります。このグラフには以下のような性質があります。

① 横軸は相関係数を表し、-1.0～1.0 までの範囲となる
② グラフの山部分の面積は正確に 1 になる
③ 横軸上で一定の範囲を区切ると、その範囲の山の面積が、その範囲の相関係数が出る確率を表す

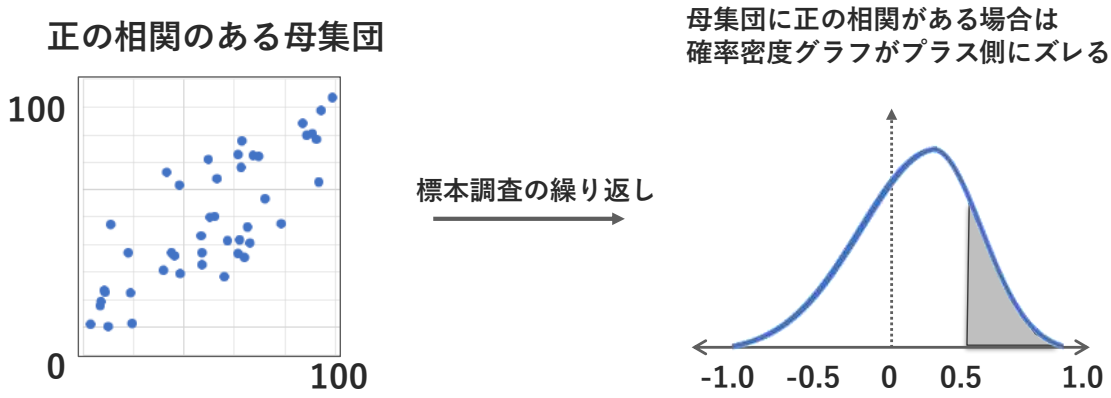
上記③のような性質から、このグラフは確率密度グラフと呼ばれます。

さらに、標本数の大小によって下記のような影響が出ます。

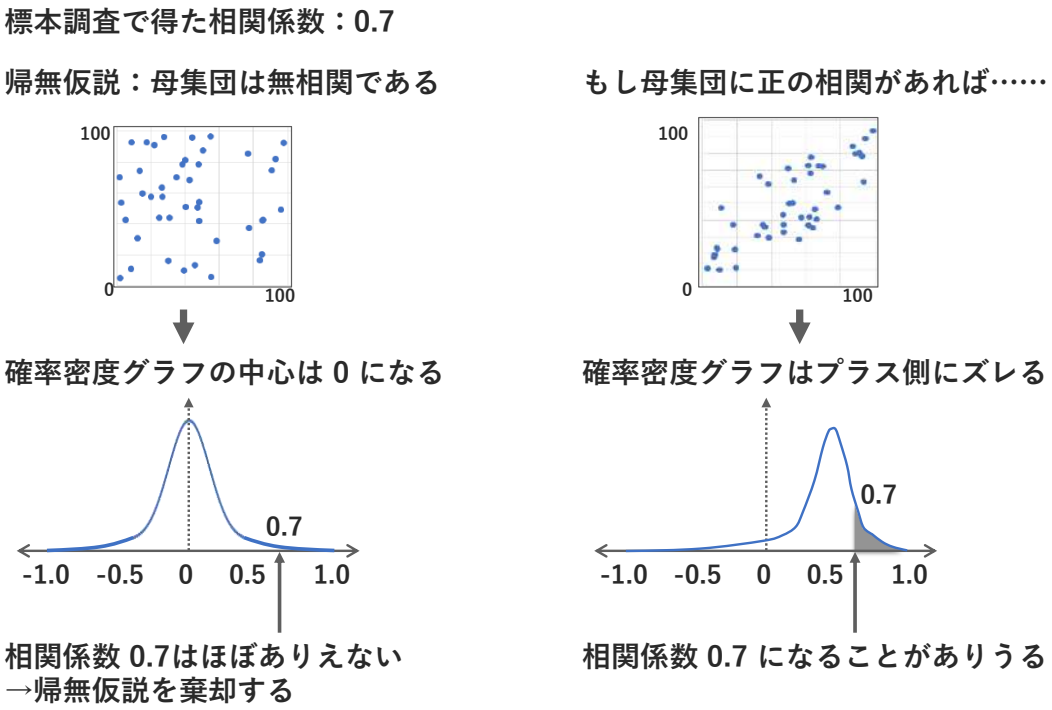
④　標本数が多いとき（グラフＬ）は相関係数は０付近に集中し、±１に近い値はほとんど出ない
⑤　標本数が少ないとき（グラフＳ）は±1 に近い相関係数も出やすい

グラフＬとグラフＳの 0.5 以上の面積を比較すると、標本数が少ないグラフＳのほうが 0.5 以上の面積が大きくなっています。つまりそれが⑤の意味で、標本数が少ないときは「タダの偶然で」高い相関係数が出ることが十分ありうるのです。今回の酒屋さんの例ではＡ商品とＢ商品の相関係数として 0.52 という数字が出ていましたが、標本数が 10 件しかないためこれは「偶然で起こりうる」数値であり、「ＡとＢに相関がある」ことを意味しません。

なお、「母集団が無相関ではない」場合、たとえばＡ商品とＢ商品の売り上げに正の相関がある（両刀派が多い）ときは、確率密度グラフの中心が０ではなくプラス側にズレます。つまり、標本調査で１に近い値が出やすくなります。



6.13 「無相関」の帰無仮説を立てて検定



現在検討している酒屋さんの例のように、「標本調査で得た相関係数に意味があるかどうかを調べる」ことを無相関検定と言います。大まかな考え方は下記のとおりです。

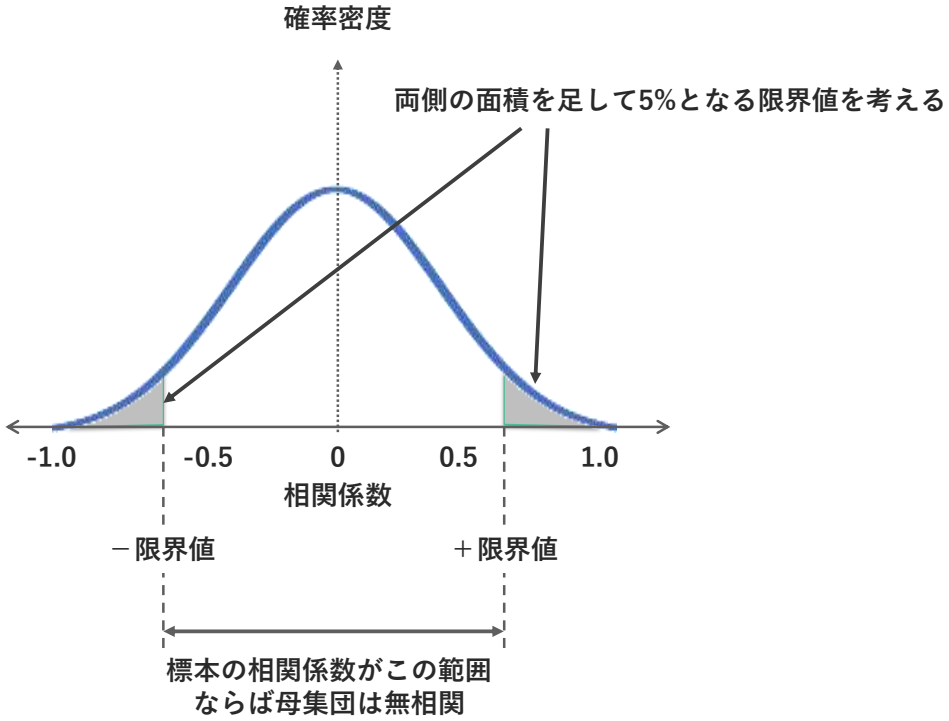
標本調査で得た相関係数がある程度高い数値、たとえば 0.7 だったとします。これが実際に「母集団には正の相関がある」と言えるかどうかを確かめるために、まず「母集団は無相関である」という帰無仮説を立てます。

無相関ですから標本調査で相関係数を調べたときの確率密度グラフの中心は 0 になります。それを前提として相関係数 0.7 が出る確率を考えると非常に低い確率であり、ほぼありえません。帰無仮説が正しければ相関係数 0.7 が出る確率は非常に低い→しかしその数値が出た→おそらく帰無仮説は正しくない（棄却される）→つまり母集団には相関があると考えられる、というわけです。

もし母集団に正の相関があれば確率密度グラフがプラス側にズレるため、相関係数が 0.7 となる確率はその分高くなります。当然こちらのほうが「ありうる」と考えられるので、「正の相関がある」と考えるのが妥当ですね。

それでは、もし相関係数が 0.7 ではなく 0.5 だったらどうでしょうか？　0.4 だったら？　0.3 だったら？　「ありうる」と「ありえない」の境界をどこに引けばよいのでしょうか？

6.14 相関係数が有意となる限界値



「有意水準」という考え方が以前出てきましたが、無相関検定でも同じ考え方を使います。
確率密度グラフの両側を足して面積が全体の 5%となる値を考え、実際に得られた相関係数がそれよりも大きければ（マイナス側なら小さければ)、発生確率が非常に小さいと考えて帰無仮説を棄却します。帰無仮説は「母集団は無相関である」というものですので、それを棄却するということは「母集団には相関がある」ということであり、「相関係数には意味がある（有意である)」ことになります。

有意水準としては通常は 5%、厳しく判断する場合は 1%を使います。

有意水準 5%（または 1%）となる相関係数の数値を「限界値」と呼びますが、限界値が具体的にどんな値になるかは標本数によって変わります。無相関検定を行う場合は、標本数別に限界値を計算した「限界値表」を参照します。

6.15 限界値表を参照して無相関検定

標本数別に計算した、相関係数の有意水準の限界値

標本数	5%有意水準	1%有意水準
3	0.997	1
4	0.95	0.99
5	0.878	0.959
10	0.632	0.765
15	0.514	0.641
20	0.444	0.561
50	0.279	0.361
100	0.197	0.256

標本の相関係数がこの限界値を超える場合に、帰無仮説が棄却される
（つまり、母集団は無相関ではないと考えられる）

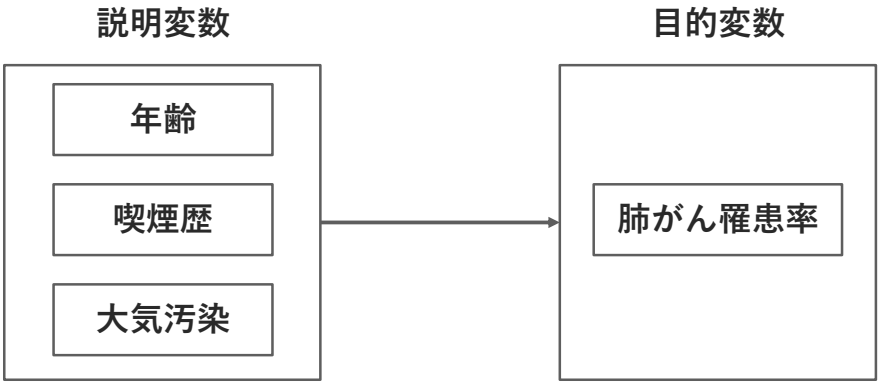
上図が実際の限界値表（の一部）です。この表から標本数ごとに 5%有意水準と 1%有意水準の限界値を調べて判断します。本書では例として一部しか載せていませんので、実務上はより詳細な表が掲載されている一般の統計学の書籍を参照してください。

今回検討している酒屋さんのケースでは標本数 10 件でしたので、5%有意水準の値を見ると 0.632 です。実際に得られた相関係数は 0.52 で、

相関係数 0.52　＜　限界値 0.632

ですから有意水準に達していません。したがって帰無仮説「母集団は無相関である」は棄却されず、「相関係数 0.52 には意味がない」という結論になります。

6.16 説明変数は複数ありうる

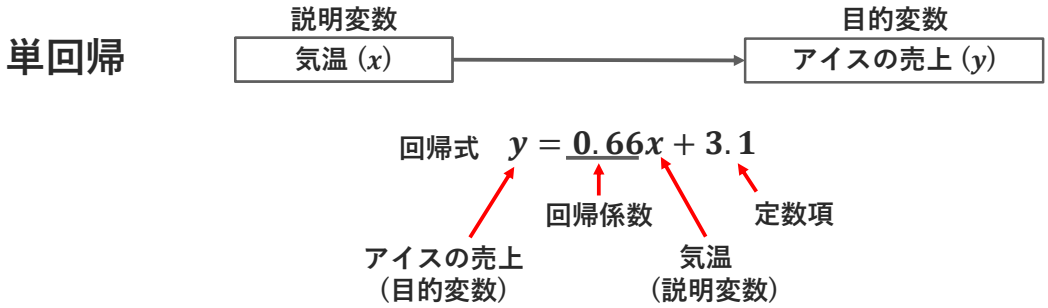


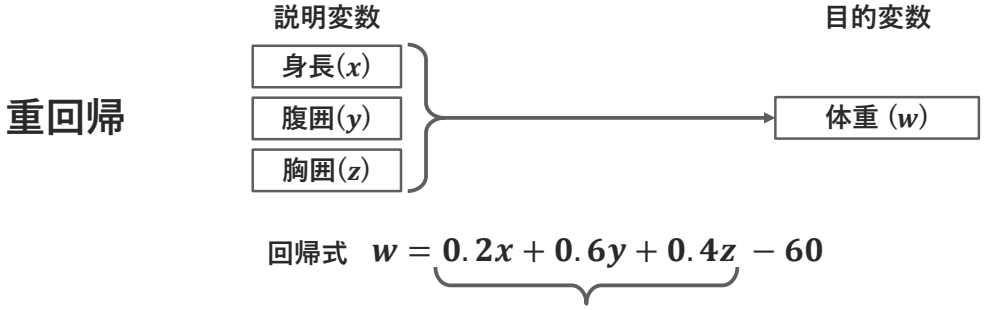
目的変数 1 つに対して説明変数が複数存在することはよくあります。たとえば「肺がん」という病気には、年齢、喫煙歴、大気汚染などの要素が関係していると考えられています。

なお、A（目的変数）の変化を B（説明変数）で説明できるとしても、B と A の間に因果関係があるとは限りません。前ページでも書いたように、相関関係と因果関係は混同しやすく、単なる相関関係である場合もよくあります。

それでも、目的変数と説明変数の関係を明らかにすることには意味があります。たとえばある人が「肺がんにかかっているかどうか」は手間とお金のかかる検査をしなければわかりませんが、「若い時期に長い喫煙歴がある」かどうかは簡単なアンケートをとるだけでわかります。該当する集団は「肺がん罹患率の高い集合」になるので、それに対して本格的な検査をすれば効率良く肺がんを早期発見し治療することが可能で、結果として多くの人を救うことができます。

6.17 単回帰分析と重回帰分析





相関のある変数どうしの関係を数値で表した「回帰式」を導くと、さまざまな予測や意思決定に使えます。回帰式のうち、説明変数が 1 個しかないものを単回帰式、複数あるものを重回帰式といい、それらを導く分析のことを単回帰分析、重回帰分析などと言います。

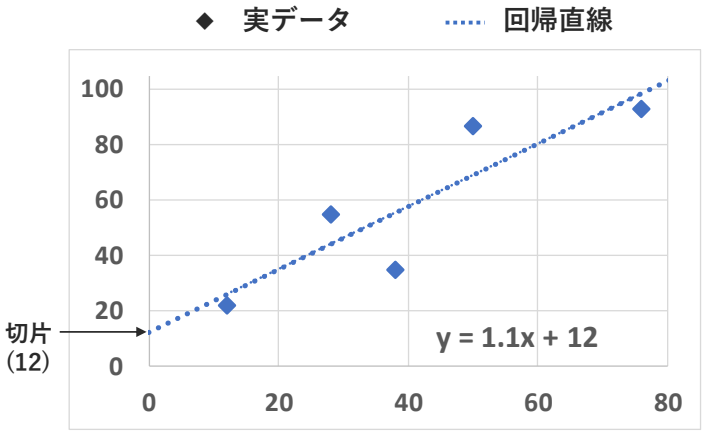
上図では体重を予測する重回帰式の例を示していますが、身長と体重に相関があるのは当然として、腹囲や胸囲も体重との相関があるため、回帰式中の説明変数の項が 3 つになっています。なお、説明変数の項に掛ける定数（アイスの売り上げの式で言えば 0.66 の部分）のことを回帰係数と言います。

世の中のさまざまなものごとのうち、「1 つの説明変数だけで結果を見通せる」というものは多くはありません。肺がんには年齢・喫煙歴・大気汚染が影響し、店の売上には商圏人口・駐車場・店舗面積が影響し、などのように複数の説明変数が目的変数に影響するのが普通です。重回帰分析はその複雑な関係を数式にして扱いやすくする分析方法です。

6.18 回帰式の求め方

あるメーカーの支店別売上

支店	商品A	商品B
高松	12	28
松山	38	35
秋田	28	55
山形	50	87
岡山	76	93



回帰式　商品B売上 ＝ 回帰係数 × 商品A売上 ＋ 切片

相関係数 × 商品Bの標準偏差 / 商品Aの標準偏差

商品Bの平均 －（回帰係数 × 商品Aの平均）

実際に回帰式を計算してみましょう。回帰分析にもさまざまな方法がありますが、本書ではその中でも最も単純な「線形単回帰分析」という方法を紹介します。「線形」は説明変数と目的変数の関係が直線的である（だから回帰「直線」を引く）ことを、「単回帰」は説明変数が 1 つだけであることを意味します。

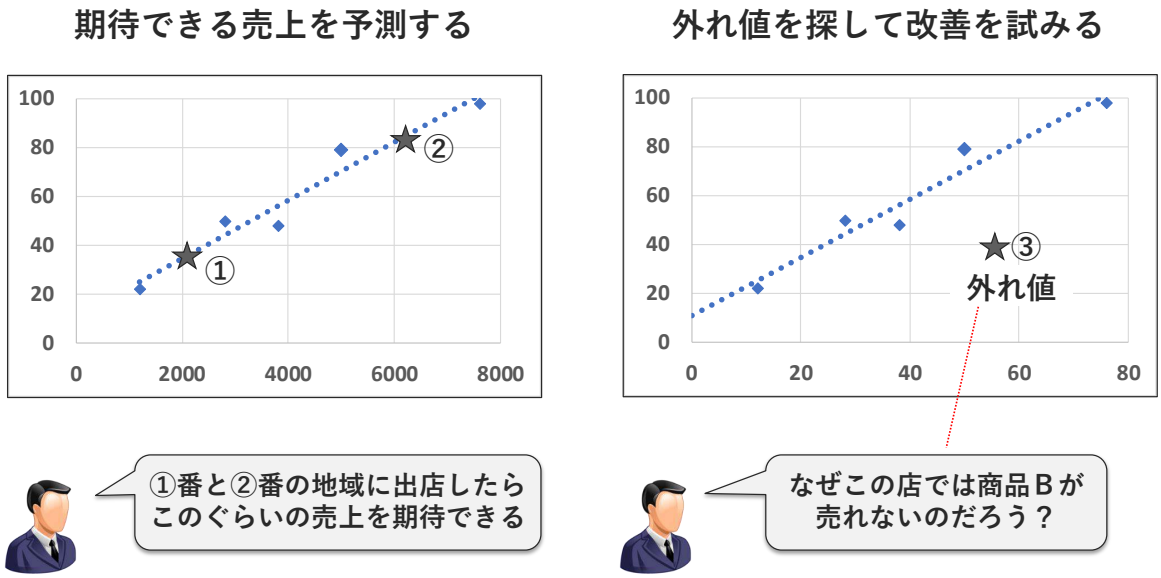
上図例はあるメーカーの商品 A と商品 B の営業支店別売上データを元に散布図を描き、さらに回帰直線を引いたものです。グラフの中の

$$y = 1.1x + 12$$

という式が回帰式で、商品 A の売上(=x)を元に商品 B の売上(=y)を予測します。このグラフを見る限り、おおむね実際のデータに近い回帰直線を引けていると言ってよいでしょう。1.1 が回帰係数、12 は切片（せっぺん）といって x=0 の場合の y の値を示します。この例では切片=12 ですので、回帰直線と縦軸の交点（横軸=0 の点）が 12 になります。

回帰係数は　相関係数 ×（商品 B の標準偏差÷商品 A の標準偏差）で求められ、
切片は　商品 B の平均 －（回帰係数×商品 A の平均）で求められます。

6.19 回帰式の役立て方



２つ以上の変数の関係を回帰式で表す「回帰分析」の用途で代表的なものの１つが「期待できる売上を予測する」ことです。たとえばコンビニエンスストアのように食品・日用品を取り扱う店への来店客のほとんどは近くに住居や職場がある人ですので、商圏人口（来店できる距離内に住んでいるか職場がある人の数）と売上の間に高い相関があるため回帰分析が有用です。

上図の左側はいくつかの店舗の商圏人口と一日当たりの売上額について回帰直線を引いたものです。ここでもし商圏人口 2000 人の地域①と 6000 人の地域②に出店したら、それぞれどのぐらいの売上を期待できるかが回帰直線で分かります。どれぐらいの売上が見込めるかは新しい店を出すか出さないかを決める極めて重要な情報なので、回帰分析はそれを判断する根拠を与えてくれます。（もちろん、実際には商圏人口だけで売上が決まるものではないため、あくまでも手がかりの１つになるだけです）

回帰分析の代表的な用途にはもう１つ、「外れ値を探して改善を試みる」というものもあります。上図の右側は横軸に商品 A、縦軸に商品 B の売上をとって回帰直線を引いたものですが、③で示した店舗だけ、回帰直線から大きく外れていることがわかります。もし何かその店の運営に問題があるなら、詳しく調べて手を打てば改善できるでしょう。そうした「調べるきっかけ」を得るためにはまず「この店だけ大きく外れている」という「外れ値」を探す必要があります。回帰式は「標準」的な値を表しているため、それとの差が大きい場合は「外れ値」であると判断できます。

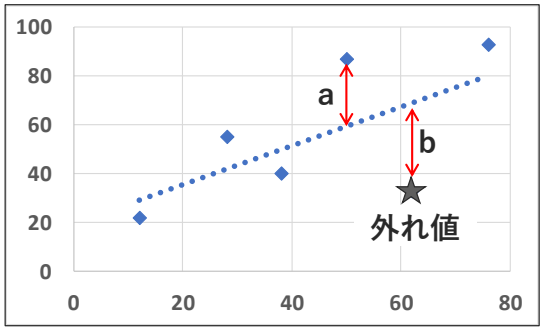
6.20 外れ値は除外しておく

支店	商品A	商品B
高松	12	22
松山	38	40
秋田	28	55
山形	50	87
兵庫	62	33
岡山	76	93

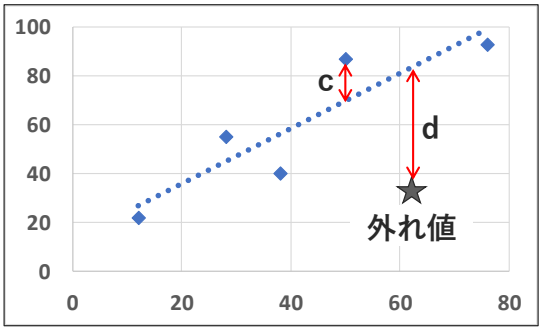
データの中に明らかな外れ値が含まれる場合は、
それを取り除いて回帰分析を行うことが望ましい

← **外れ値**

外れ値を含むデータによる回帰直線



外れ値を除くデータによる回帰直線



回帰分析を行う際の注意事項の 1 つに、「明らかな外れ値は除外しておくことが望ましい」というものがあります。上図は商品 A と商品 B の売上データで、「兵庫」の 62-33 というのが外れ値です。

外れ値を含むデータで回帰直線を引くと、山形のデータ（図中 a の部分）は兵庫のデータ（図中 b、外れ値）と同程度に回帰直線から外れているように見えます。つまり「外れ値」が目立ちませんので、このままだと「b が外れ値かな？　よくわからないな」となって改善の手を打つのが遅れてしまいかねません。

一方、外れ値である兵庫のデータをあらかじめ除外して回帰直線を引くと、山形のデータ（図中 c の部分）と回帰直線の差は小さくなり、逆に兵庫のデータ（図中 d、外れ値）は大きく外れる形になりました。これなら「まずい、兵庫がどうなってるのか調べよう」と判断できるでしょう。

全体のデータ数に対する「外れ値」の件数が大きい時はこのように「回帰分析をしても外れ値がわかりにくい」という事態が起りがちですので、明らかな外れ値は回帰分析の前に除外しておくことが望ましいのです。

6.21 回帰式の当てはまりの良さを表す R2 値

データセットA

相関係数=0.87

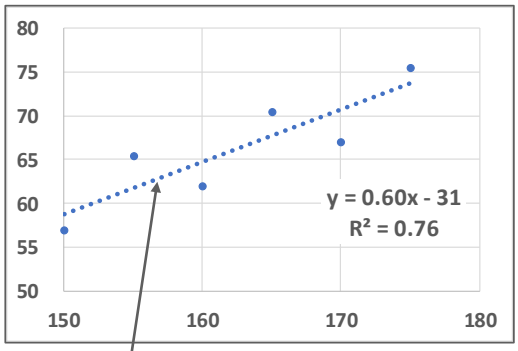
身長	体重
150	57
155	66
160	62
165	71
170	67
175	76

データセットB

相関係数=0.72

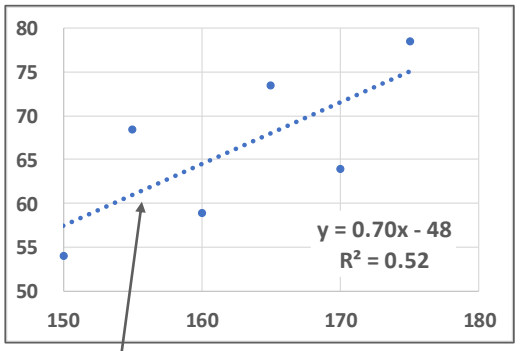
身長	体重
150	53
155	70
160	58
165	75
170	63
175	80

グラフA



回帰直線が良く当てはまる

グラフB



うまく当てはまらない

説明変数と目的変数の関係を回帰式で表せればさまざまな予測に使えるので便利です。しかし、計算で回帰式を出してみてもそれが実際のデータに良く当てはまるとは限りません。

上図は二組の架空のデータ（データセットA、B）について回帰式を算出してグラフに回帰直線を引いた例です。イメージしやすいように身長と体重の関係としてあります。

相関係数を計算するとデータセットAは0.87、Bは0.72と出ます。いずれも0.7以上ですから、これは「かなり強い正の相関がある」ことを示しています。強い相関があるデータについては回帰分析により予測できる可能性が高いので回帰式を算出すると、それぞれ

データセットAの回帰式　$y = 0.60x - 31$
データセットBの回帰式　$y = 0.70x - 48$

となります。実データの散布図と，この式によって表される回帰直線をグラフ化したものがグラフA、Bですが、明らかに印象が違います。グラフAのほうは実データと回帰直線の差が小さく、グラフAではそれが大きいことが分かります。つまり、データセットBの回帰直線は実データにうまく当てはまりません。これでは、回帰直線を予測に用いることはできません。相関係数が高いから

といって、回帰式が当てになるとは限らないのです。

このような場合に「回帰式（回帰直線）が当てになるかならないか」を判断することはできるでしょうか？　グラフを描いて「なんとなく差が大きい」と見た目で判断するのは人によってあるいは気分によって判断がバラツキやすく、良い方法とは言えません。そこで役に立つのが、決定係数あるいは R2 値[1]と呼ばれる数値を計算する方法です。

> ＜決定係数とは＞
> 目的変数の変化を説明変数がどの程度うまく説明できるか、つまり「回帰式がどの程度当てになるか」を表す数値で、1 に近いほど「当てになる」ことを意味する。

前ページのグラフＡ、Ｂの中にこのような式がありました。

グラフ A：　R2 = 0.76
グラフＢ：　R2 = 0.52

この R2 が決定係数です。計算方法が数種類あり、最も単純な計算方法では R2 は相関係数の二乗に一致します。計算方法によってそれぞれ違う数値を出すため一概には言えませんが、おおまかに以下のような基準値がよく目安として使われます。

決定係数の値	実データに当てはまるか？
0.9 以上	非常に良く当てはまる
0.7〜0.9	良く当てはまる
0.5〜0.7	あまり当てはまらない
0.5 未満	当てはまりは悪い

グラフ A の決定係数は 0.76 と「良く当てはまる」水準だったのに対して、グラフ B の決定係数は 0.52 でしたので、「あまり当てはまらない」の下限に近い数値ですから、やはりグラフ B の回帰式（回帰直線）は当てにできないことが分かります。

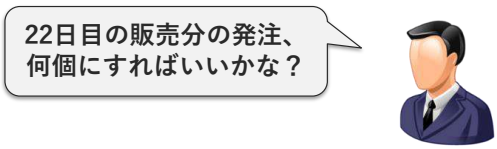
---

1 上付き文字を使える場合は R2 を R2 と表記します

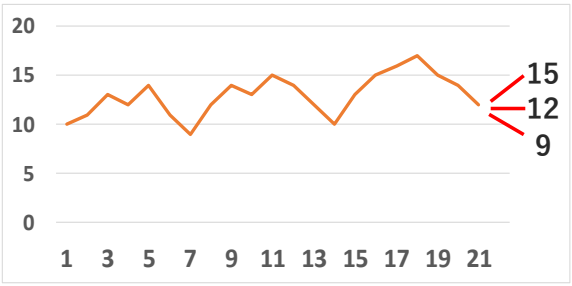
6.22 時系列に沿って変わるデータの分析

ある商品の日ごとの販売数量データ

日付	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
販売数	10	11	13	12	14	11	9	12	14	13	15	14	12	10	13	15	16	17	15	14	12	



日付/販売数グラフ



販売回復を期待
21日と同数
過去数日の推移を延長

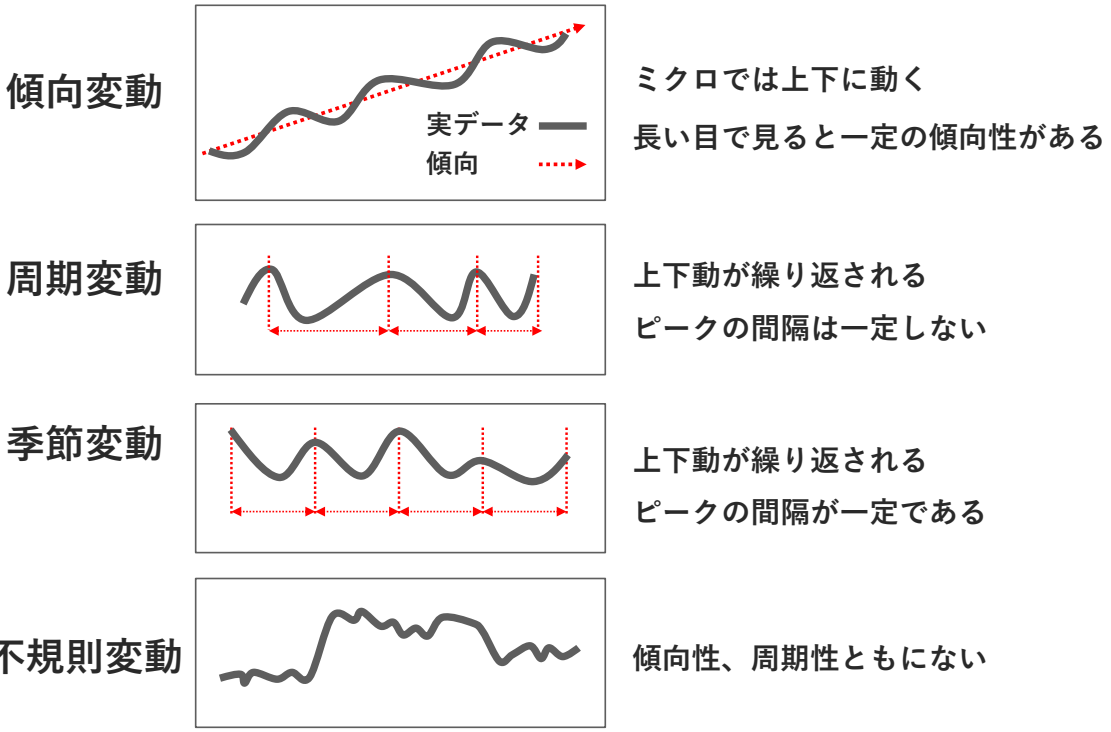
コンビニやスーパーのような小売店で、ある商品が 21 日目までは上図の表のように売れていて、明日が 22 日目だとします。その商品はお弁当のように賞味起源の短いもので、その日の終わりに翌日分を発注することになっています。発注した品は翌日朝に入荷します。入荷したら当日のうちに売り切らなければならないため発注数を慎重に考えなければなりません。発注が少なすぎると販売機会損失、多すぎると廃棄でどちらにしても損になってしまいます。21 日めには 12 個売れましたが、明日の分はいくつ発注すればよいでしょうか？

ほんの数日前には 17 個売れてるので、販売回復を期待して 15 個
21 日めと同じぐらい売れると考えて　12 個
過去数日は下がり基調なのでその線を延長すると　9 個

など、いずれもそれぞれ一理ありそうです。ここで注目したいのは、この 3 つの考え方はいずれも「過去の販売数を根拠に未来を予測している」ということです。特に 3 番目の「9 個」という予測は、過去数日の動きの傾向を延長しています。このように「時間軸に沿った動き」を考える方法を時系列分析と言います。

過去数日の傾向線を延長
（時系列分析の一種）

6.23 時系列変動の種類



「時間軸に沿った動き」にもいくつかの種類があり、主なものが傾向、周期、季節、不規則の 4 つです。

「傾向変動」は短期では上下しても長い目で見ると「上がる」「下がる」のどちらか一方へ動くという一定の傾向性があるものを言います。たとえば日本の人口は近年減り始めていますが、それ以前は 150 年ほど増加傾向を続けていました。成長期にある会社の売上額も、一時的に減ることはあっても長い目で見ると増加します。

「周期変動」はある範囲で上下動を繰り返すがその周期は一定ではない、というタイプの動きです。たとえば道路上のどこか一点で通過する車の台数を 1 分単位で数えると、渋滞している時はほとんど止まってゼロになる一方で、数十台単位で通過していく時もあることでしょう。しかしたとえ渋滞が解消されても片側 2 車線の道路で毎分 1000 台の車が通ることはありえません。1 つの道路を走れる車の数には限界があるため、上限と 0 の間で増えたり減ったりする動きをします。上限を超えて増えることはないし、0 を切ってマイナスになることもないため「傾向変動」にはなりません。また、変動のピークから次のピークまでの間隔が一定でなくても「周期変動」と呼びます。「周期」は、「上限と下限の間を行ったり来たりする」という意味と考えてください。

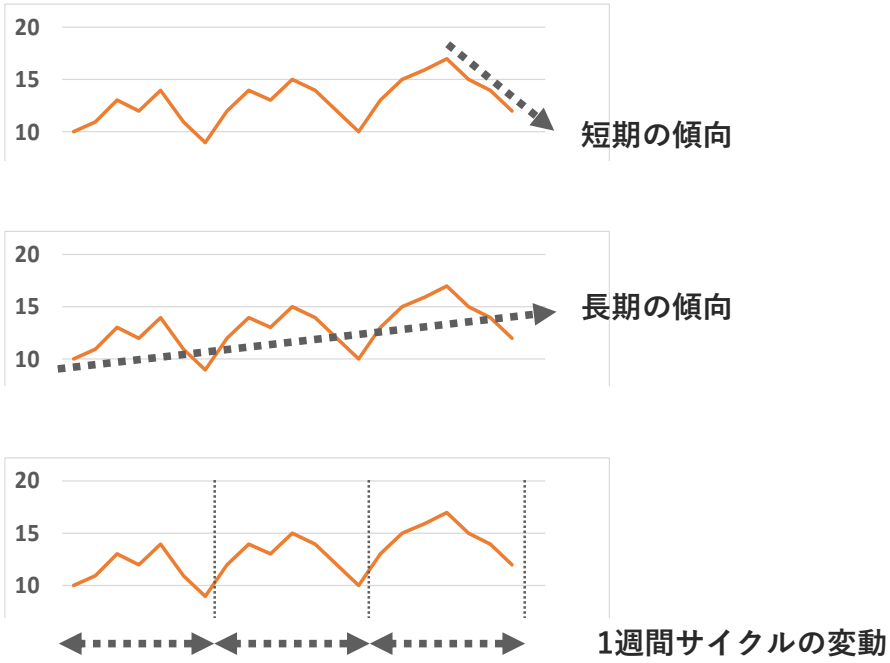
しかし、周期変動の中には一年という一定の期間で春夏秋冬が繰り返されるものもあります。この

ようにピークとピークの間隔が一定であるような種類の周期変動を「季節変動」と呼びます。「季節」変動といっても一年単位のものとは限らず、1ヶ月単位・1週間単位・1日単位のようなものも「季節変動」と呼ぶので注意してください。道路の交通量は、深夜早朝は少なく日中は増えるという1日単位で変動することが多いので、その意味では季節変動とも言えます。

「不規則変動」はそれ以外のもの、傾向・周期・季節に当てはまらない変動をするものを言います。たとえば大きな台風が来るとその後しばらくの間は台風で壊れた家屋の修理需要が一気に上がり、高止まりして少しずつ下がっていきます。このような動きは傾向とも周期とも言えないため不規則変動として分類されます。台風の場合は「毎年秋に多い」という季節性の面もありますが、地震ではそれもないため完全に不規則です。また、2019年に「タピオカ」ブームが起きたように、特定分野の人気が一時的に高まる「ブーム」現象は一般に予想がつかないもので、不規則変動を起こす場合が多いです。

実際には多くの場合、これらのうち複数の種類の変動が複合します。たとえばある駅の1日あたりの乗降客数を考えると、平日と休日で差があるため1週間単位の変化、つまり「季節変動」をするのが普通であると同時に、もしその駅周辺で住宅開発が進んでいて周辺の人口が増加しつつある時期ならば長期的に増加していく「傾向変動」も重なりますし、台風や地震のような自然災害があれば「不規則変動」も複合してきます。したがって、時系列のデータを扱う場合は最低限このような種類があることを踏まえて考えなければなりません。

6.24 データだけでは判断できないときもある



あらためて以前出てきた「21 日間の販売数量データ」を見ると、少なくとも 3 つの可能性がありそうです。

1 つめは過去数日の動きを「短期の傾向変動」と考えてそれを延長する方法です。この考え方だと販売数がやや少なくなると見るのが妥当です。

2 つめは「長期の傾向変動としてはわずかに増加している」と考える見方です。この場合、同程度か微増を期待しても良さそうです。

3 つめは「1 週間単位のサイクルがある」と考える見方です。平日によく売れる商品ならば土日は減って月曜日に増えますから、1 週間単位の「季節変動」をします。この見方が正しいなら、22 日目はまた増える日に当たるため、微増よりもハッキリ増えることもありえます。

以上 3 つの見方のどれが現実に合っているのかはデータだけではわかりません。しかしたとえば「ある商品」の主な購入者が会社員や学生であって平日に多く来店する人々なのであれば、「1 週間サイクル」と考えるのが一番妥当です。その仮説が正しいかどうかは「販売数」のデータだけでは分かりません。統計学は未来を予測するために使える強力なツールではありますが、同じデータでも複数の見方が成り立ちそれぞれが違う答を出す場合があります。それを適切に選ぶためには、データに表れていない部分の理解がいる場合もあることに注意が必要です。

6.25 用語集

相関関係	ある変数と別な変数の変化に一定の法則がある場合、その２つの変数には相関関係があると言う。因果関係がある場合は必ず相関関係もあるが、逆は成り立たない。
相関係数	ある変数と別な変数の間に相関関係があるかないかを示す数値。＋1 から－１の範囲で、関連がない場合は 0、ある場合は±1 に近い値を取る
回帰式	２つの変数の間に相関関係がある場合、 y= ax + b のような一次関数の形で近似できることがあり、その式を回帰式という。
回帰係数	回帰式 y = ax + b における a を回帰係数という
バスタブ曲線	機械製品の故障率は使用開始直後に高く、その後低くなり、製品寿命の前にまた高くなる傾向があり、それをグラフ化した形をバスタブ曲線という。この現象の場合、「故障率」と「使用期間」の間に相関関係があるが、曲線ではないため、一般的な相関係数の計算式では相関関係を検出できない
説明変数と目的変数	ある変数の変化を別な変数の変化によって説明できる場合に、前者を目的変数、後者を説明変数と言う。回帰式 y=ax+b では y が目的変数、x が説明変数。
共通の原因	相関関係を因果関係と誤認するパターンの１つ。共通の原因を持つ複数の変数の間で相関係数を計算すると「相関あり」という答えが出る。「共通の原因」に気がついていないと、この相関を因果関係だと誤認しやすい。
合流点での選別	相関関係を誤認するパターンの１つ。「A 商品または B 商品の購入額合計１万円以上」のように「複数の変数にまたがった条件」で標本を選別することを言う。その選別後に A と B の相関分析をすると、誤った相関係数が算出されやすい。
無相関分析	通常、相関係数の絶対値が 0.4～0.7 であれば弱い相関があるとされるが、標本数が少ない場合にはその相関係数が意味を持たない場合がある。相関係数に意味があるかどうかを分析することを無相関分析と言う。
単回帰分析と重回帰分析	相関のある変数どうしの関係を回帰式で表すことを回帰分析と言い、説明変数が１個しかないものを単回帰分析、複数あるものを重回帰分析と言う。
R2 値（決定係数）	回帰式が実際のデータによく当てはまるかどうかを表す値。強い相関のあるデータで回帰式を作っても実際のデータにはよく当てはまらない場合があるため、回帰式を実務に利用する前に R2 値を算出して確認しなければならない。
時系列分析	データの時系列に沿った変動に一定の法則がある場合にその法則を明らかにする分析のことを言う

傾向変動	長い目で見たときに一定の傾向性がある時系列変動のこと。国の人口や経済規模、会社の売上など
周期変動	ある範囲で上下動を繰り返す変動。ただし「周期」は一定でなくてもよい。
季節変動	ピークとピークの時間間隔が一定であるような変動。「季節」といってもその間隔は1年とは限らず、1ヶ月、1週、1日など長くても短くてもよい。
不規則変動	傾向・周期・季節のいずれにも当てはまらない変動。

6.26 確認問題

【問 1】ピアソンの積率相関係数に関する説明の中で、間違っているものはどれか？

【選択肢】

（A）２つの変数の動きに何らかの関係があるかどうかを判断する指標である

（B）-1 から+1 の範囲の値を取り、+1 に近い場合は「正の相関がある」、-1 に近い場合は「負の相関がある」という。0 に近い場合は「無相関」である

（C）2 変数の関係が直線状（一次関数）になる場合に限らず、曲線状の関係がある場合も相関を検出できる

（D）計算のみで相関係数を算出できるため、無数の変数の中から関係がありそうなものを自動的に見つけ出すためにも使える

【問 2】ある店の顧客別の購入明細の一部です。「チーズとワイン」「ワインとソーセージ」「ソーセージとチーズ」の 3 種類の組み合わせの相関係数を計算しなさい。

顧客番号	チーズ	ワイン	ソーセージ
1	500	2800	600
2	200	0	200
3	700	1500	0
4	800	1900	1200
5	1000	4000	3000
6	0	700	2400

■確認問題解答

【問 1 解答】

（C）が間違い。

ピアソンの積率相関係数は直線状の関係しか検出できない。

【問 2 解答】

チーズとワイン　　：0.79　（強い相関）

ワインとソーセージ　　：0.48　（弱い相関）

ソーセージとチーズ　　：0.15　（無相関）