

2023 年研修

福田俊介

2023/01/31

目次

1	はじめに	2
1.1	利用しているデータ	2
2	複数データをみる視点	3
2.1	複数データの具体的なイメージとしての地域の地価水準	3
3	複数データの中の多いものを把握するという視点	4
3.1	区間毎に集計する	4
3.2	具体的な数字を得る	5
3.3	平均値とそれ以外の視点	5
4	複数データに見られるデータの偏り	6
4.1	山形をしたヒストグラム	7
4.2	ヒストグラムと設定する区間	7
4.3	分布の形状に着目する確率密度曲線	8
5	分布を複数データの性質としてみる	10
5.1	最大と最小	10
5.2	データの集中、ばらつきの具合	10
5.3	確率密度曲線の縦軸の高さが表すもの	11
6	四分位を使ってデータのばらつき具合を考察	12
6.1	最大値と最小値	12
6.2	中央値	12
6.3	四分位数	12
6.4	箱ひげ図	12
7	複数のデータからみる価格の水準	15

1 はじめに

不動産鑑定士の目線での**取引事例**は、取引事例比較法を適用する場合に用いる資料であり、各々の取引事例は、地域の特性と個別性を持つ**個別の不動産の個別の価格**を示す資料です。そして、不動産鑑定士は、一つ一つの事例を念入りに見ることにはとても慣れていると思います。

しかしながら一方で、通常の鑑定評価の作業で、**複数の取引事例を分析**することはあまり行っていないのではないのでしょうか？複数の取引事例とは、幅広いデータの集まりだと「〇〇市の全ての住宅地域」の取引事例であったり、実務上で具体的にありそうな例であれば「3年以内で枚方市内の牧野駅を最寄り駅とする、土地面積が60㎡から130㎡までの住宅地域に掛かる取引全ての事例」などです。

例えば、REA ネットの事例閲覧ページで、自分の好きな検索条件を設定した結果の一覧をイメージしてください。不動産鑑定士は、通常の業務経験の積み重ねから、これら事例検索結果の一覧をざっと一瞥することで、町名や字名、価格の偏り、最寄り駅からの距離、前面道路の幅員や系統から対象不動産の存する近隣地域との優劣比較を行い、近隣地域の価格水準の見当を大まかにつけてしまいます。

このような複数の事例に対して鑑定士が行っている処理は、普段から鑑定士がこうした作業を繰り返し業務として行っていることによる経験に基づく専門性の高いものであると思っています。ですから一方で、このような作業自体を不動産鑑定士以外の人へ分かりやすく説明するということは、なかなか難しい部分だったりすると思います。

そこで、このテキストでは、一般的なデータ分析による**複数データの分析の仕方**を利用して、鑑定評価上の**複数の事例**の分析の仕方を一般の人（鑑定士以外の人）に説明する際のヒントを紹介しようと思います。もちろん、**複数データの分析**の知識自体が、鑑定士自身の複数事例に対する分析の精緻化にも寄与すると思います。

1.1 利用しているデータ

まず、この文書で用いている**複数のデータ**は、国土交通省が公表している不動産取引価格情報から、大阪府の土地のみの取引に掛かるデータ（2005年から2022年までの57,110件）をベースにしています。

<https://www.land.mlit.go.jp/webland/servlet/MainServlet>

各項目では、このデータから特定の条件で抜き出したデータを用いています。

2 複数データをみる視点

ここから複数データとして、大阪府の寝屋川市の住宅地に掛かる土地取引の取引単価（㎡単価）を扱って行きます。ここでまず、**複数のデータ**と具体的にどんな向き合い方が出来るのかを考えていきましょう。

不動産のことを考える時、鑑定士目線では（私がそうなのですが）、何か個別の不動産を価格を求めるための思考回路がすぐに働いてしまいます。すなわち、個別性や地域性の異なるものが混じった複数事例全体をみることの意味がいまいちピンとこないかもしれません。

2.1 複数データの具体的なイメージとしての地域の地価水準

そこで、ここでは、**複数や全体を具体的にイメージするために「寝屋川市の住宅地域の地価水準」**というものを考察することを考えてみてください。

もし、先生方が普段 REA 事例を利用して、或る市に掛かる事例を「住宅地域」という括りだけで検索すると、すごく沢山の事例が検索結果として表示されることが容易に想像できると思います。実際に寝屋川市の住宅地に掛かる事例の件数を数えると 1,236 件あります。当然、1 ページにおさまる件数ではありません。単純にページを切り替えて全体に目を通したとしても土地単価全データの傾向的なものを把握することは困難であることが予想できます。

では、そもそも、地域の特性等の類似性で検索して、データの数进行、見通しのよい状態になった検索画面では、私達は何を見ているのでしょうか？直感的に思うのは「どの辺りの数字が多いのか」ではないでしょうか。例えば、ざっとみて㎡単価 11 万円や 12 万円程度の事例が**多そう**だなあという感覚です。

一方で、ここでみるような目視で把握困難な大量のデータを目の前にすると、何が「多そう」なのかを把握しづらくなります。そこで、感覚的に「多そう」と感じている部分をもう少し精緻化する方法を検討しましょう。

3 複数データの中の多いものを把握するという視点

何が多くて何が少ないかは、単純に「数えて、その数えた数字で比較する」ことで簡単に把握することが出来ます。例えば、「単価 11 万円台の事例が何件、単価 12 万円台の事例が何件、、、」という具合に一つ一つの事例を単価何万円台なのかに分けて、その数を数えます。

先に、「簡単に把握することが出来ます。」と書きましたが、ここでの事例のデータ数は 1,236 件あり、**自分で一つずつ集計するのは簡単ではありません**。そのためには、この集計を行うシステムが必要です。しかし、この **集計さえ行ってしまうと、複数のデータを把握するのが簡単になる**という視点をまずは把握しましょう。

3.1 区間毎に集計する

さて、ここでは具体的な数字を見るために実際に、1 万円台毎の数をプログラムの数えてみます。

```
## .
##      (0,1e+04]      (1e+04,2e+04]      (2e+04,3e+04]      (3e+04,4e+04]
##              4              13              21              22
##      (4e+04,5e+04]      (5e+04,6e+04]      (6e+04,7e+04]      (7e+04,8e+04]
##              29              34              50              70
##      (8e+04,9e+04]      (9e+04,1e+05]      (1e+05,1.1e+05]      (1.1e+05,1.2e+05]
##              79              130              111              119
##      (1.2e+05,1.3e+05]      (1.3e+05,1.4e+05]      (1.4e+05,1.5e+05]      (1.5e+05,1.6e+05]
##              102              101              94              60
##      (1.6e+05,1.7e+05]      (1.7e+05,1.8e+05]      (1.8e+05,1.9e+05]      (1.9e+05,2e+05]
##              38              48              39              16
##      (2e+05,2.1e+05]      (2.1e+05,2.2e+05]      (2.2e+05,2.3e+05]      (2.3e+05,2.4e+05]
##              21              10              7              3
##      (2.4e+05,2.5e+05]      (2.5e+05,2.6e+05]      (2.6e+05,2.7e+05]      (2.7e+05,2.8e+05]
##              5              2              0              1
##      (2.8e+05,2.9e+05]      (2.9e+05,3e+05]      (3e+05,3.1e+05]      (3.1e+05,3.2e+05]
##              2              2              0              0
##      (3.2e+05,3.3e+05]      (3.3e+05,3.4e+05]      (3.4e+05,3.5e+05]      (3.5e+05,3.6e+05]
##              0              1              0              0
##      (3.6e+05,3.7e+05]      (3.7e+05,3.8e+05]      (3.8e+05,3.9e+05]      (3.9e+05,4e+05]
##              0              0              0              1
##      (4e+05,4.1e+05]      (4.1e+05,4.2e+05]      (4.2e+05,4.3e+05]      (4.3e+05,4.4e+05]
##              0              0              0              0
##      (4.4e+05,4.5e+05]      (4.5e+05,4.6e+05]      (4.6e+05,4.7e+05]      (4.7e+05,4.8e+05]
##              1              0              0              0
##      (4.8e+05,4.9e+05]      (4.9e+05,5e+05]
##              0              0
```

上記で示されている集計結果の表現は、ある区間毎に該当する件数が何件あるかを表現しています。

この出力書式は、数学やデータを表す表現の方法で割と一般的なもののなので、この機会に見慣れておきましょう。

まず、 $(1e+04, 2e+04]$ と書かれている部分は、10,000 より大きく、20,000 以下の区間ということを表しています。括弧が間違っていて印刷されていたり、数字が文字化けしているわけではありません。

まず、区間を表す場合、「〇〇より大きい」や「〇〇以下」という表現があるように、何かをきちんと分析する場合、そこで表示されている数字を「含む」、「含まない」ということを正確に表現する必要があります。そして、一般的（数学やデータ分析等の分野）には、区間の表現で値を含まないものを **开区間**、値を含むものを **闭区間**と呼んでいます。そして、具体的にこれを文字で表現する場合に、开区間を丸括弧である“(“で、闭区間を角括弧”[“で表現します。

次に、 $1e+04$ の部分は、桁数が大きい数字をコンパクトに表す表現として、一般的に使われる表現です。これは、“e”の後ろに、10 の何乗かという数字を書いて表現します。ですから $1e+04$ は、 1×10^4 つまり 10,000 を表しています。

最後に、これら区間を表す表記の下に書かれた数字がその区間にある事例数を表現しています。

3.2 具体的な数字を得る

このように、集計すると**具体的な数字**を得ることが出来ます。例えば、このデータの集計を見ると $(9e+04, 1.1e+05]$ 、すなわち、 m^2 単価が 9 万円より大きく 10 万円以下の区間の事例が最も多く 130 件あるという事が正確にわかります。そして、 $(1e+05, 1.1e+05]$ 、すなわち、 m^2 単価が 10 万円より大きく 11 万円以下の区間の事例は 111 件であり、 $(1.1e+05, 1.2e+05]$ 、すなわち、 m^2 単価が 11 万円より大きく 12 万円以下の区間の事例は 119 件であります。

この微妙な違いに意味があるかどうかの判断は別として、一覧をざっと見て多そうな数字を感覚的に把握するのと、実際に数えて具体的な数字で見比べるのとの違いが把握できると思います。

また、これらの集計を自分でどうやるのか？という点については、今の所、気にしなくて結構です。沢山のデータについては、ある区間ごとに集計して数えることで具体的な数字として得られる情報があるという事をまずは把握しましょう。

3.3 平均値とそれ以外の視点

ここで、大量のデータについて、簡単に数えたり計算できたりするのであれば、複数データを代表する数値として最も有名な**平均値**を調べて、それを一つの目安とすることも可能でしょう。

寝屋川市のデータの平均値は 120,784 円/ m^2 です。先程の最も多い区間と比べると、平均値の表す数字のある区間とは異なっていることがわかります。

この様に複数のデータを扱う場合、「平均値」はその複数のデータの性質を表す目安のうちの **一つ**であり、複数データを分析して得られる数字は他にも沢山あるのです。つまりは、複数データの平均値が常に「正しい」何かを表しているわけではなく、複数のデータを考察する際には、他にも色々な視点があるのです。

4 複数データに見られるデータの偏り

さて、ここまでは**複数データ**から、「多そう」な部分を見つけるために区間毎の集計を行い、「最も多い区間」を見つけました。

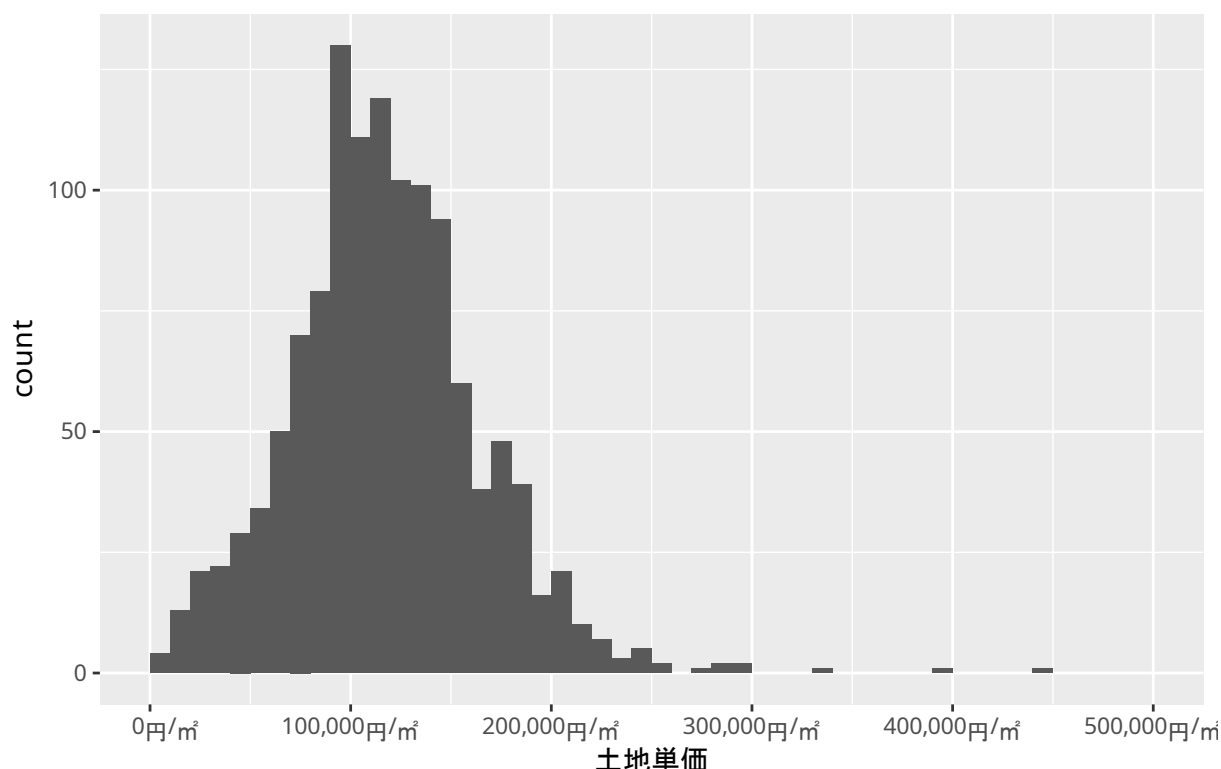
ここで、もともと考察しようと思っているのは「寝屋川市の住宅地域の価格水準」です。もちろん、「最も多い区間」イコール「当該地域の価格水準」と判断するのも一つの方法です。

しかし、今、私達は、具体的に数え上げた各区間の事例数が手元にあり、先に見たとおり、例えば、 m^2 単価 10 万円より大きくて、14 万円以下の 1 万円毎の各区間についても 100 件以上の事例があります。単純に 1 万円刻みの区間に決定するよりも、これら周辺の多そうな部分も含めることにも合理性が有りそうです。その一方で、 m^2 単価が 30 万円を超えるような事例は、ほとんどありません。

このように、複数のデータを考察して具体的に集計を行うと、もちろん「多そう」な部分の中から正確に「最も多い」部分を見つけることが出来るのですが、それ以外に、データ全体について、どの範囲にどれくらいの量の事例が**分布**しているのかを知ることができるので、多そうだと思える範囲と、逆にこれは無さそうと思える範囲などの **程度**を自分の物差しで判断できるようになります。

この程度を直感的に把握するには、先に見たどの区間に何件の事例があるという具体的な数字をみるだけではピンと来ません。実は、この分布からそれぞれの区間の多い少ないの程度を把握するにはグラフで見る方が直感的にわかりやすくなります。そこで、実際に上述の集計表を可視化したものが次のグラフになります。

土地単価のヒストグラム



このグラフは**ヒストグラム**と呼ばれるグラフです。このグラフの横軸は m^2 単価が書かれており、棒状のグラフの横幅がその**区間**を表し、棒の高さで区間にある事例の数を表しています。棒の高さや区間の具体的な数字

は、先の集計表のものです。

まずはじめに一般的にデータを分析する上で、**ヒストグラム**という単語を覚えます。そして、**ヒストグラム**と**棒グラフ**は異なるものであることを認識します。

普段耳にする「グラフ」と呼んでいるもの、例えば、棒グラフ、円グラフ、折れ線グラフ、帯グラフ等は、見た目の違いで、「〇〇グラフ」と認識していると思います。形的に言えば、「ヒストグラム」も「棒グラフ」も大体同じです。しかし、データ分析を行っている文脈で「ヒストグラム」というグラフが出てくると、先にやったように「あるデータを区間毎に集計して、その数量を表示している」グラフを指しています。つまり、形のことを言っているのではなくて、そのグラフが表している意味が付随しています。日常生活で使われる**グラフ**は、何かのデータを見やすくするために使われる事が多いですが、データ分析で使われる**グラフには意味がある**ということを認識しましょう。

4.1 山形をしたヒストグラム

さて、ヒストグラムの形を眺めると「山形」になっています。つまり、最も事例が多くあつまるところの区間があり、その一番多い区間から離れる程にその区間に存する事例数が減っていくという事を表しています。

なにかの物事を集計すると、大体この様に多い部分と少ない部分の**偏りがあるように分布**していて、可視化すると山形になっていることが多いのです。

我々はこのことを理由のあるなしは別としてなんとなく経験則的に知っているのですが、一番はじめに「寝屋川市の住宅地域の価格水準」を知る上で、「多そう」な価格から話を始めた事にそれほど違和感を感じなかったのではないのでしょうか？

4.2 ヒストグラムと設定する区間

さて、実は区間集計をとったり、それに基づいたヒストグラムというグラフを描くとき、**区間**をどの様にするかで、その得られる結果が異なります。先の分析では、単価 1 万円台毎に集計を取りましたが、これを単価 2 万円台毎の集計でやった場合の結果を以下に示します。

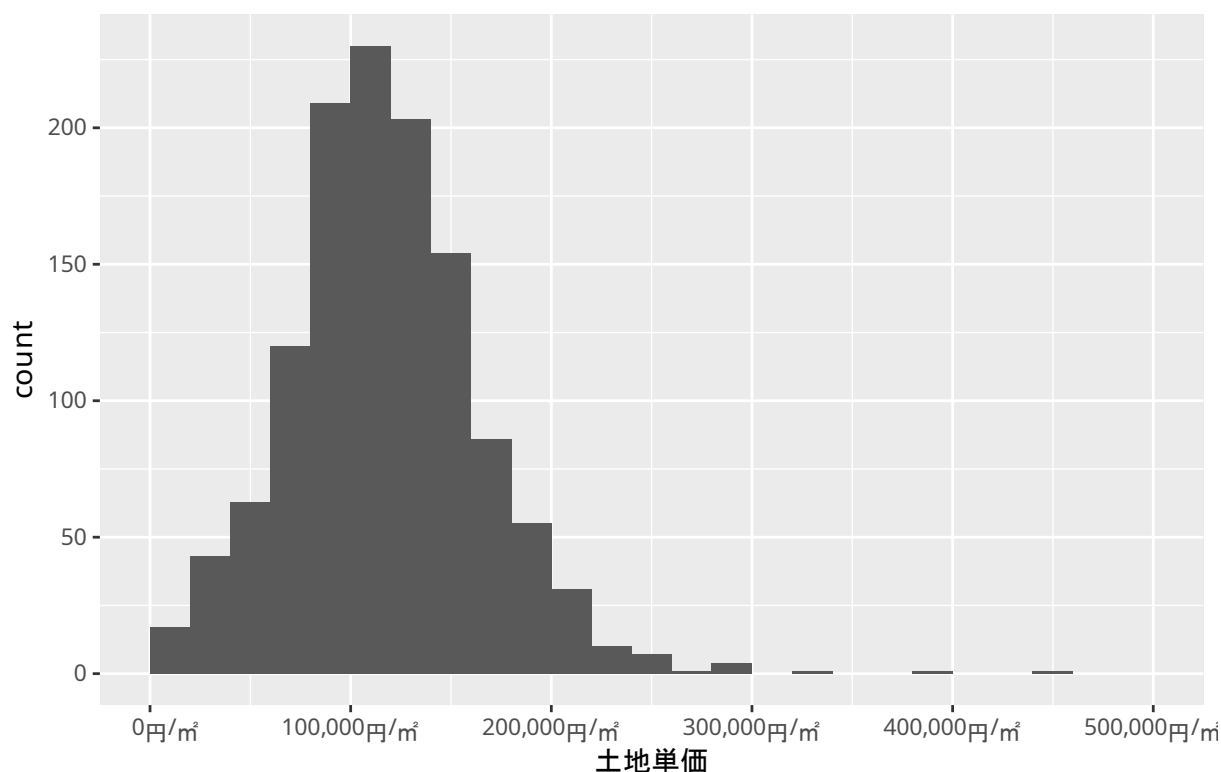
##	.			
##	(0,2e+04]	(2e+04,4e+04]	(4e+04,6e+04]	(6e+04,8e+04]
##	17	43	63	120
##	(8e+04,1e+05]	(1e+05,1.2e+05]	(1.2e+05,1.4e+05]	(1.4e+05,1.6e+05]
##	209	230	203	154
##	(1.6e+05,1.8e+05]	(1.8e+05,2e+05]	(2e+05,2.2e+05]	(2.2e+05,2.4e+05]
##	86	55	31	10
##	(2.4e+05,2.6e+05]	(2.6e+05,2.8e+05]	(2.8e+05,3e+05]	(3e+05,3.2e+05]
##	7	1	4	0
##	(3.2e+05,3.4e+05]	(3.4e+05,3.6e+05]	(3.6e+05,3.8e+05]	(3.8e+05,4e+05]
##	1	0	0	1
##	(4e+05,4.2e+05]	(4.2e+05,4.4e+05]	(4.4e+05,4.6e+05]	(4.6e+05,4.8e+05]
##	0	0	1	0
##	(4.8e+05,5e+05]			

##

0

これに基づいたヒストグラムは以下のとおりになります。

土地単価のヒストグラム



このように、区間のとり方によりその集計結果が変わるのは当たり前のことですが、例えば、1万円台毎の集計のときは、10万円/㎡以下の部分でピークがあったのに、2万円台の集計のときは10万円/㎡以上の部分にピークがあります。

このとき、「どっちが多いの?」と頭を悩ませることになるかもしれません。これは、「どちらかが正しい」というものではなく、はじめに書いたとおり「区間のとり方によりその集計結果は変わる」のです。

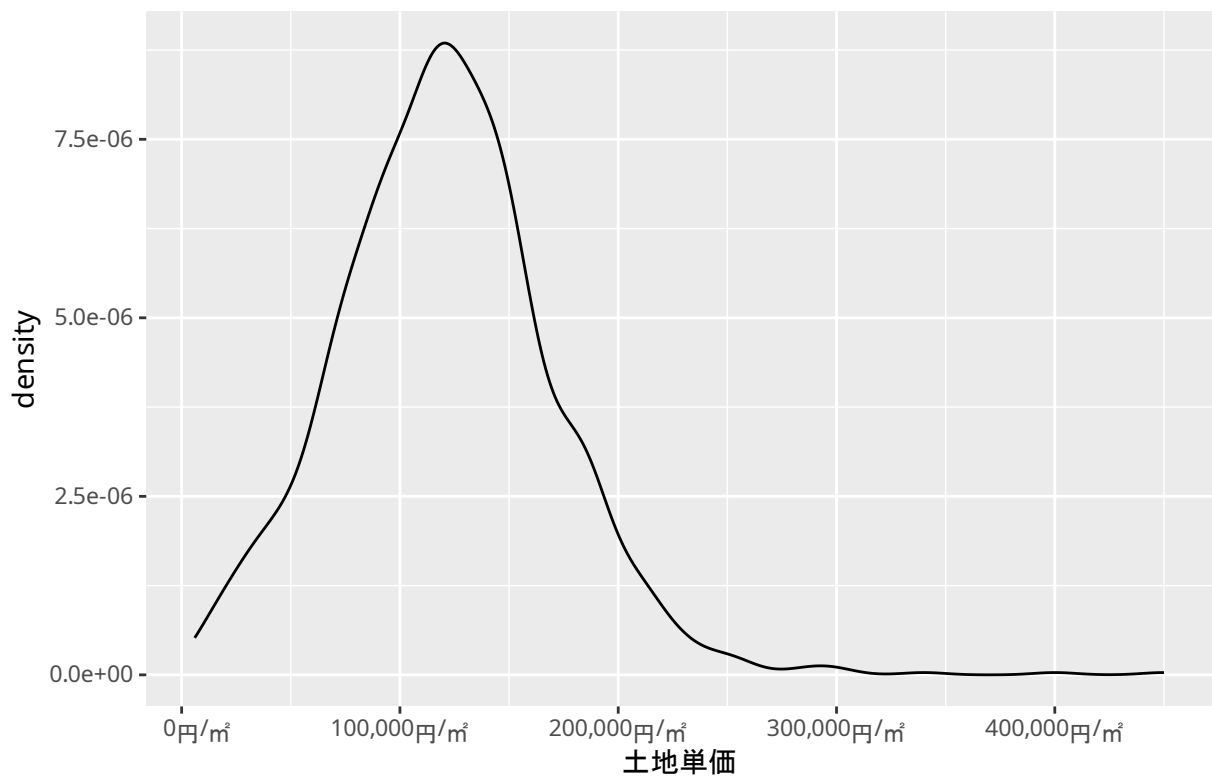
4.3 分布の形状に着目する確率密度曲線

さて、もともと、沢山の事例の中から「多そう」な部分を知るために、区間設定をおこない、それに該当する事例の数を数えて集計し、その集計結果をヒストグラムで描き分布の状態を知りました。

しかし、区間設定による集計では、その区間のとり方により、ヒストグラムの形状が変わってしまいました。

そこで、分布の様子を考察する時、いつでも同じデータなら同じ形状になるように可視化するグラフがあります。これは、区間ごとの集計を行うのではなく、全事例に対する、その値までの事例の個数の割合を面積として表示するグラフで**確率密度曲線**と呼ばれています。

土地単価の確率密度曲線



この**確率密度曲線**というグラフの名前もデータ分析でよく出てくる名前なので、まずは、まずは、名前を覚えてしまいましょう。はじめは、このグラフをどうやって描くのかという**書き方**は気にせず、**確率密度曲線**は、**データの分布の様子を表している**ということを把握していれば十分です。分布を考察する時にヒストグラムのように区間設定が必要なく、同じデータに対して同じ結果が得られます。

5 分布を複数データの性質としてみる

ここまでで、「寝屋川市の住宅地域の価格水準」を検討することを目的として複数の事例データを考察するために「多そう」な部分に着目することから初め、データが分布している様子を見れるようになりました。ここからは、この分布仕方がひとまとまりの複数データの性質を表すものだと考え、分布を考察する時どのようなところに着目できるのかを見ていきましょう。つまり、**分布の性質を紐解くことで、「寝屋川市の住宅地域の価格水準」に接近する**わけです。

5.1 最大と最小

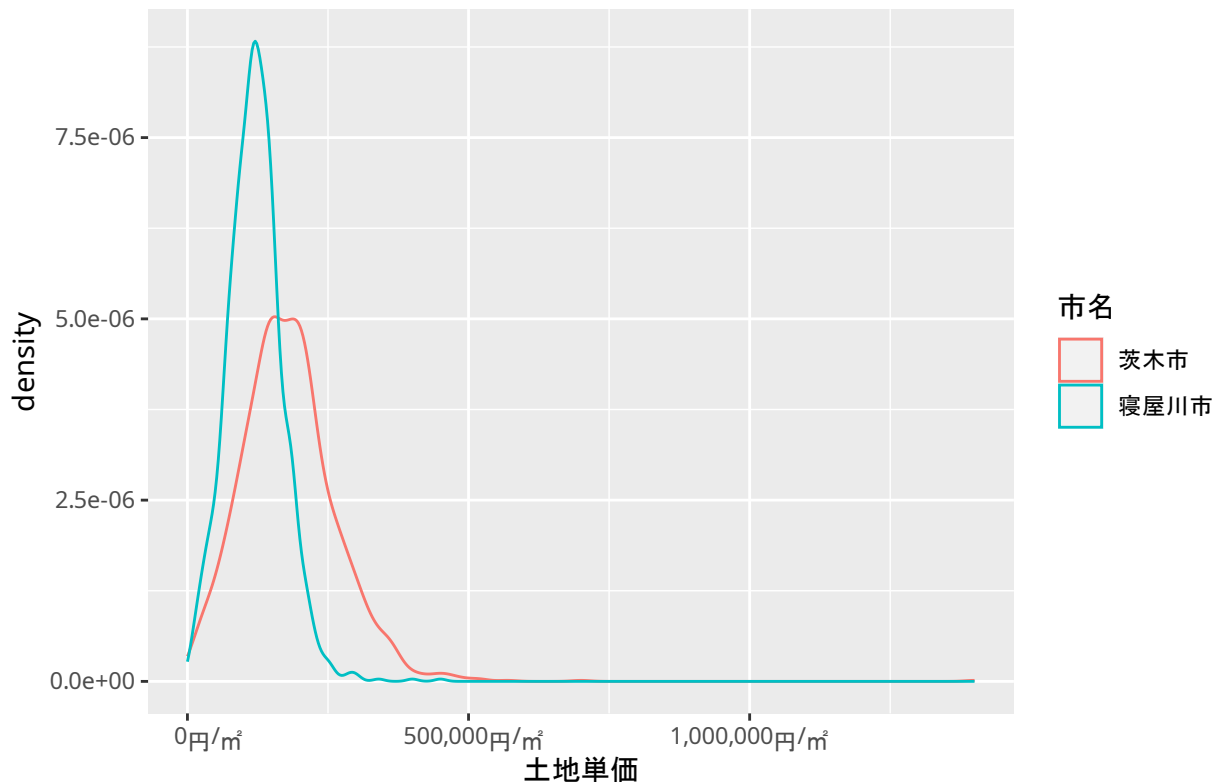
いままでは、直感的に **多そう**な部分に着目していましたが、今度は逆に **無さそう**な部分に着目してみましょう。つまり、複数データの中に含まれる最大値と最小値です。寝屋川市での土地単価の最大値は 450,000 円／㎡であり土地単価の最小値は 6,000 円／㎡です。

現実の取引での可能性は別として、得られている事例の中では最大値より大きな、また、最小値より小さな取引はなかったという事実が得られます。つまりは、この範囲の外側での取引は**無さそう**な領域として把握できます。このことを考えると最大値、最小値も**分布の性質**を表す一つの重要な要素と言えます。

5.2 データの集中、ばらつきの具合

今まで、寝屋川市の住宅地域だけを見ていましたが、一つのデータを見るだけでなく、茨木市の住宅地域のデータを**比較**してみましょう。

土地単価の確率密度曲線の比較



上のグラフは寝屋川市と茨木市の住宅地の単価の確率密度曲線です。

5.3 確率密度曲線の縦軸の高さが表すもの

複数の事例を確率密度曲線で比較する際に**注意が必要**なのは、縦軸が取引の絶対量を表しているのではないということです。つまり、このグラフでは、寝屋川市のほうが茨木市より**取引の量が多いわけではありません**。

このグラフでは、寝屋川市のグラフのピークは、横軸 12 万円/㎡前後の地点です。また、茨木市のピークは横軸で 20 万円/㎡強の地点です。そして、ピーク時の縦軸は何を表しているかということ、寝屋川市の全事例のうちの 12 万円台の事例の割合であり、つまり、そこに**集中している度合い**を表しています。

これに対して、茨木市の 20 万円/㎡強の地点のピークの高さが寝屋川市よりも低くなっています。これは、茨木市が寝屋川市ほど、価格の**偏りが集中していない**。すなわち、寝屋川市と**比較して価格がバラけている**ということを表しています。

あらためて、確率密度曲線を見ると、寝屋川市は比較的データが集中しているので山が**のっぼ**であり、茨木市は比較的データがバラけているので山が**なだらか**になっています。

このように、複数のデータがある場合、そのデータの**集中度合い**や**バラけ具合**が、そのデータの分布の特徴の一つになります。

6 四分位を使ってデータのばらつき具合を考察

複数データを考察する視点の一つとして「ばらつき具合」があることがわかってきました。ただ、先の項目では可視化されたグラフで、直感的に高い低いを把握しただけです。ここから、これをもう少し精緻に考察する方法を考えます。

6.1 最大値と最小値

先に見た最大値と最小値です。そして、先にも言ったとおり、これが複数データの端っこであり、バラけている限界を表しています。

6.2 中央値

中央値は文字通り、真ん中の値です (全データ数が偶数の場合は、真ん中 2 件の平均値)。具体的には、寝屋川市の事例は 1,236 件なので、事例を小さい方から順番に並べて 618 件目と 619 件目の平均値になります。

この真ん中の値が分かると、ここから最大値までの差と最小値までの差が異なる場合、データの集まり具合がどちら側に偏っているのかを知ることが出来ます。

6.3 四分位数

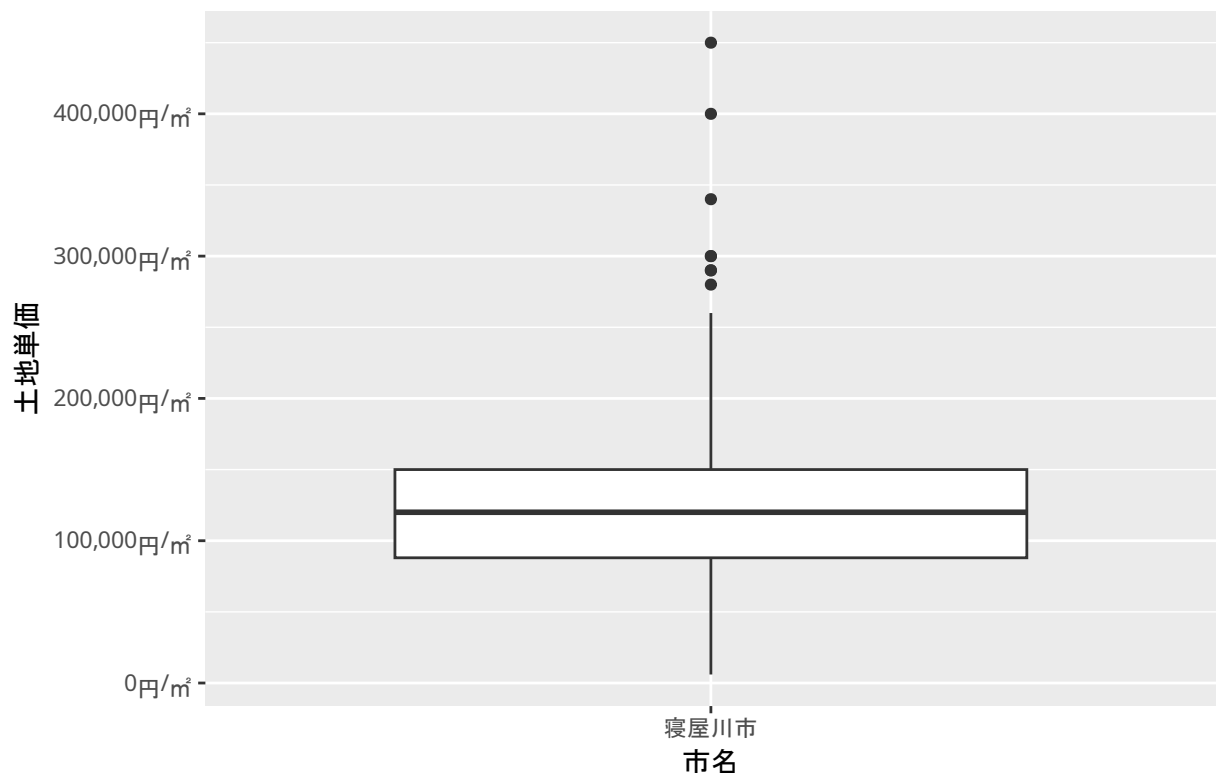
さて、最小、中央、最大の値を具体的に得ることで、分布の偏り具合を考察出来そうに思えてきました。しかし、もう少し細やかに把握するためにデータ分析の世界では **最小と真ん中の真ん中、すなわち 25% 位置と真ん中と最大の真ん中、すなわち 75% 位置**の数値も把握します。全体のデータの区間を 4 つに分割することから**四分位**と呼ばれています。第 1 四分位は、最小と中央の真ん中。第 3 四分位は、中央と最大の真ん中になります。(第 2 四分位は中央値、第 4 四分位は最大値に一致するので通常そう呼びません)

寝屋川市の四分位数	
最小値	6,000 円/m ²
第 1 四分位	88,000 円/m ²
中央位	120,000 円/m ²
第 3 四分位	150,000 円/m ²
最大値	450,000 円/m ²

6.4 箱ひげ図

前の項目では四分位数を具体的な数値でみました。この数値を視覚化して相対的な位置関係を直感的に把握できるようにしましょう。この四分位の位置関係を可視化したグラフを**箱ひげ図**と呼びます。以下に、寝屋川市の住宅地域の土地単価データの箱ひげ図を示します。

寝屋川市の土地単価の箱ひげ図



大きなコマのような形が表されています。コマ本体の長方形の部分の真ん中に太い横線が書かれていますが、この太い線が中央値の位置です。

次に、コマ本体の長方形の下側の辺の位置が第1四分位の位置です。逆に、コマ本体の長方形の上側の辺の位置が第3四分位の位置です。

更に、コマから突き出した上下にひげの様な線が突き出ています。まず、下側は線の終端が最小値の位置を表しています。次に、上側には、線が突き出し、その線の先に点がいくつか書かれています。この点の中で一番上にある点が最大値の位置を表しています。

この線は最大で第1四分位から第3四分位までの距離の1.5倍の位置まで引かれ、それを超えた部分のデータは点で表示されるようになっています。

6.4.1 箱ひげ図の見方

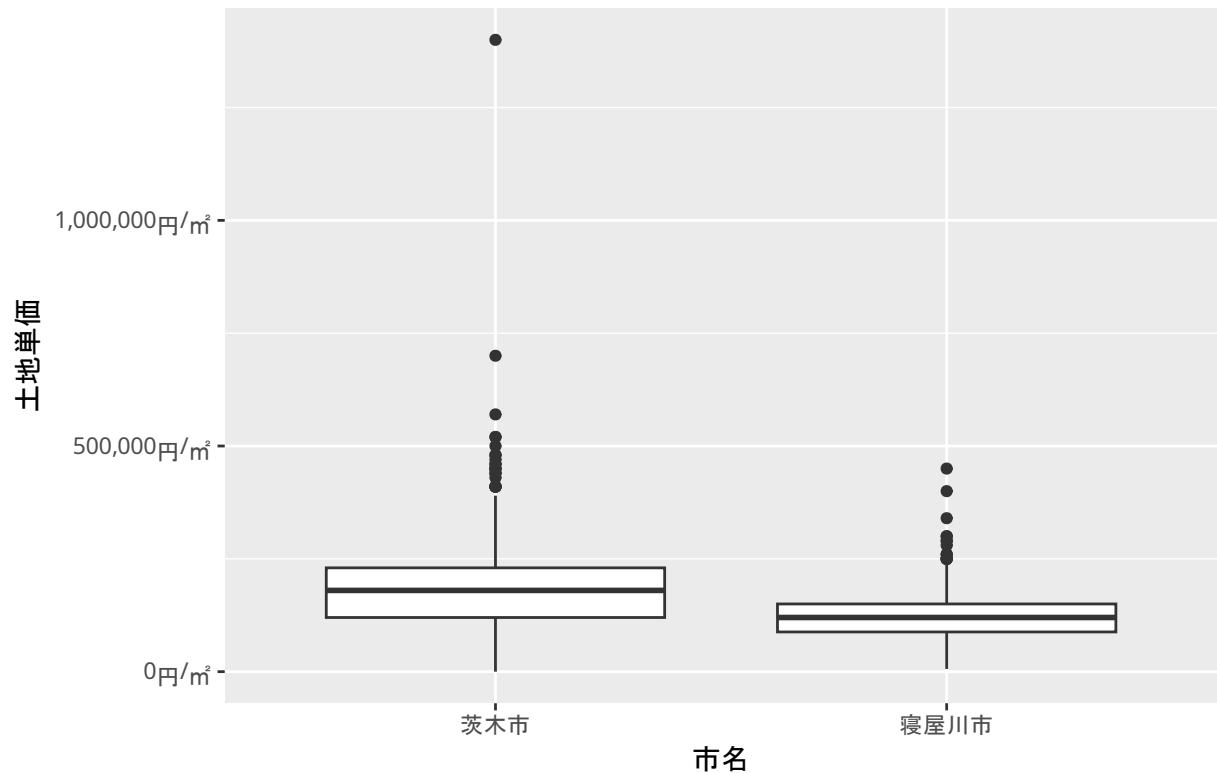
まず、箱ひげ図は基本的にコマ本体の長方形部分にデータ全体の50%のデータがあることを示しています。ですから、コマの長方形部分が上下に大きく太い場合、相対的にデータがバラけており、コマの長方形部分が上下に小さく薄い場合、相対的にデータが集中してまとまりがあると把握できます。

次に、ひげ及び点の部分は、データ内によく見られるデータから、離れた部分にあるデータで、その離れ具合を表しています。特に、点で表示される部分は、そのデータ群の中で飛び抜けて離れたデータを示しており、不動産の土地単価であれば、何らかの著しい事情があることが予測されるべきデータの存在を把握することが出来ます。

6.4.2 箱ひげ図で比較する

相対的な位置関係は比較することで明らかになるものがあります。茨木市と寝屋川市の箱ひげ図を並べてみます。あわせて、具体的な数値もそのあとに示します。

土地単価の確率密度曲線の比較



寝屋川市の四分位数

最小値	6,000 円/㎡
第 1 四分位	88,000 円/㎡
中央位	120,000 円/㎡
第 3 四分位	150,000 円/㎡
最大値	450,000 円/㎡

茨城市の四分位数

最小値	26 円/㎡
第 1 四分位	120,000 円/㎡
中央位	180,000 円/㎡
第 3 四分位	230,000 円/㎡
最大値	1,400,000 円/㎡

7 複数のデータからみる価格の水準

複数のデータの代表値である**平均**を価格水準の数値として採用することも一つの方法です。

しかし、ここで見てきたように、複数のデータを考察する際の視点として、複数のデータがどのような分布の状態にあるのかを把握し、それらの特徴を示す数値を用いて価格水準を説明することも一つの方法です。

具体的には、上記で求めた四分位数値を参考に、寝屋川市の住宅地域の取引では、88,000 円／㎡から 150,000 円／㎡の価格帯が主であるということに数値的な裏付けをもたらすことが出来ます。