2023 年研修

福田俊介

2023/02/14

目次

1 **はじめに** 3
1.1 利用しているデータ 3

2 **複数データをみる視点** 4
2.1 複数データの具体的イメージとしての地域の地価水準 4

3 **複数データの中の多いものを把握するという視点** 5
3.1 区間毎に集計する 5
3.2 具体的な数字を得る 6
3.3 平均値とそれ以外の視点 6

4 **複数データに見られるデータの偏り** 7
4.1 山形をしたヒストグラム 8
4.2 ヒストグラムと設定する区間 8
4.3 分布の形状に着目する確率密度曲線 9

5 **分布を複数データの性質としてみる** 11
5.1 最大と最小 11
5.2 データの集中、ばらつきの具合 11
5.3 確率密度曲線の縦軸の高さが表すもの 12
5.4 ヒストグラムで見る 13

6 **四分位を使ってデータのばらつき具合を考察** 14
6.1 最大値と最小値 14
6.2 中央値 14
6.3 四分位数 14
6.4 各区間の間隔が集中ばらつき具合を具体的に表す 14
6.5 箱ひげ図 15

7 **ヒストグラムから地価を推定する** 19

7.1 分布と確率の話 19

8 **正規分布** 22

8.1 正規分布と平均 23

9 **分布のばらつき具合の物差しである標準偏差** 26

10 **標準偏差と正規分布の確率** 29

10.1 標準偏差と正規分布と確率 31

11 **複数のデータからみる価格の水準の説明** 33

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**ということはあまり行っていないのではないでしょうか？複数の取引事例とは、幅広いデータの集まりだと「○○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、データの数を絞り、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**だなぁという感覚です。

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、**自分で一つづ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えば、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的に数えてみます。

```
## .
##         (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、「○○より大きい」や「○○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 **開区間**、値を含むものを **閉区間**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、開区間を丸括弧である”(“で、閉区間を角括弧”[“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、㎡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、㎡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、㎡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/㎡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

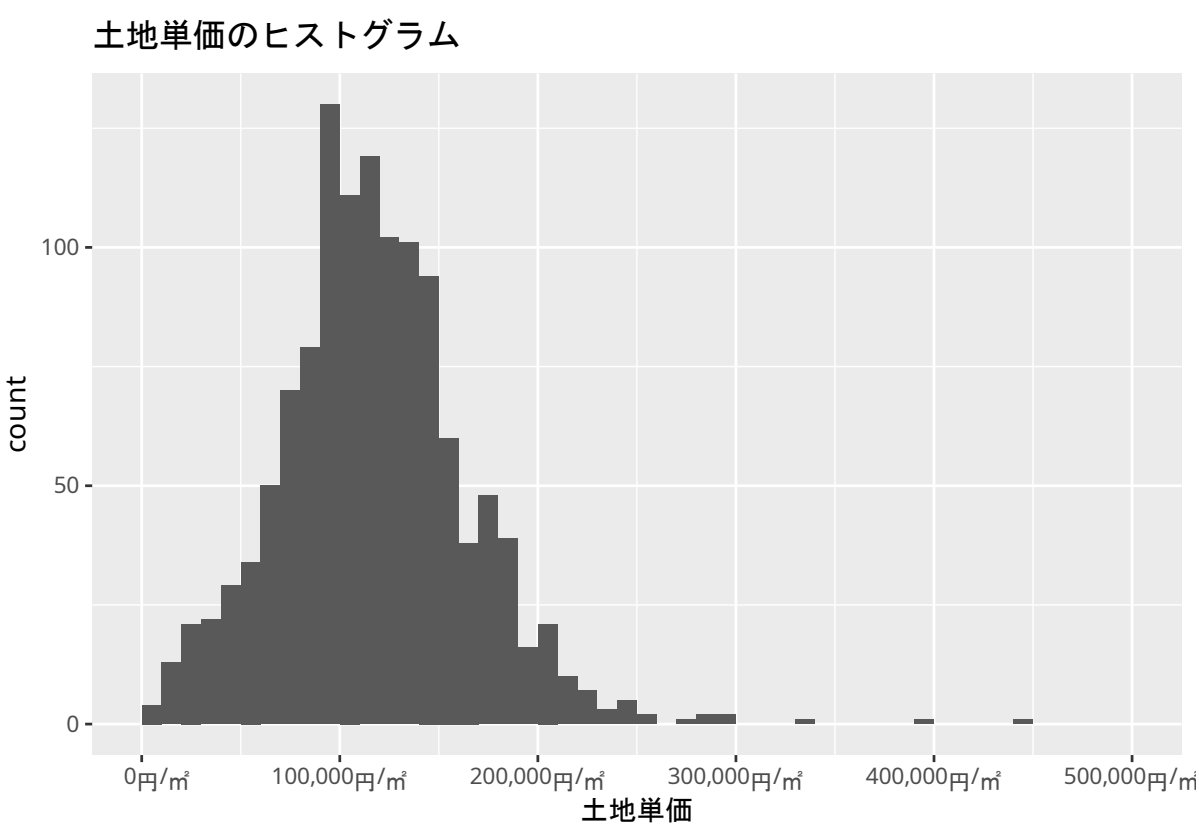
さて、ここまでは**複数データ**から、「多そう」な部分を見つけるために区間毎の集計を行い、「最も多い区間」を見つけました。

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、㎡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、㎡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。



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㎡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。日常生活で使わえる**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ような或る区間があり、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、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？

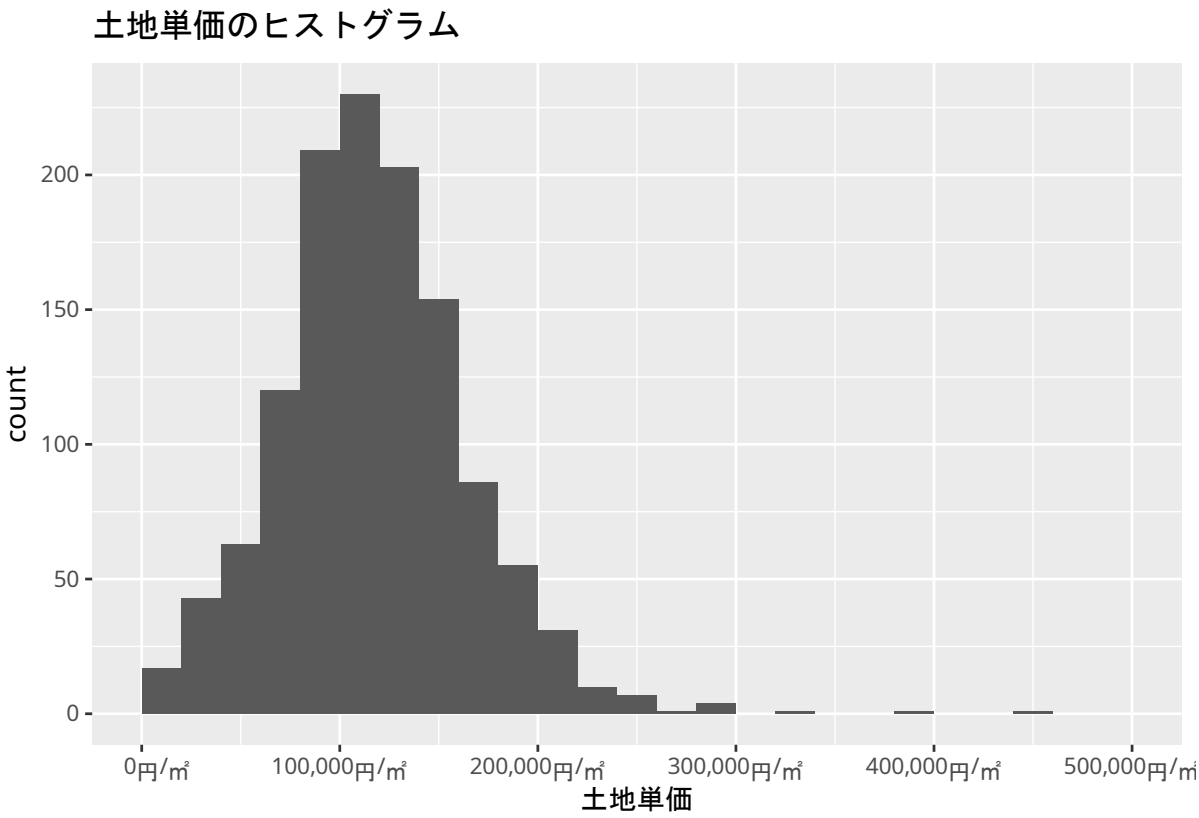
4.2 ヒストグラムと設定する区間

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**をどの様に取るかで、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。



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。

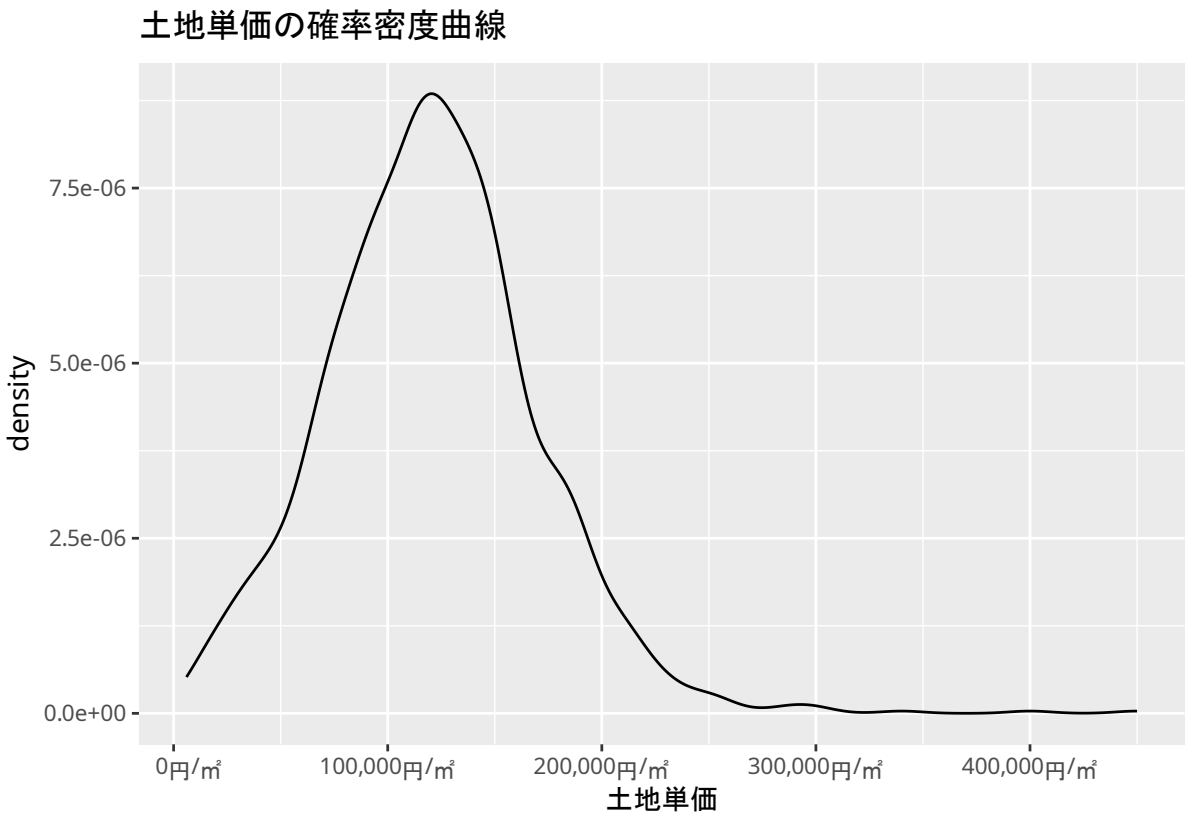
このとき、「どっちが多いの？」と頭を悩ませることになるかもしれません。これは、**「どちらかが正しい」というものでは無く**、はじめに書いたとおり**「区間のとり方によりその集計結果は変わる」**のです。

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。



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かのかという**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。

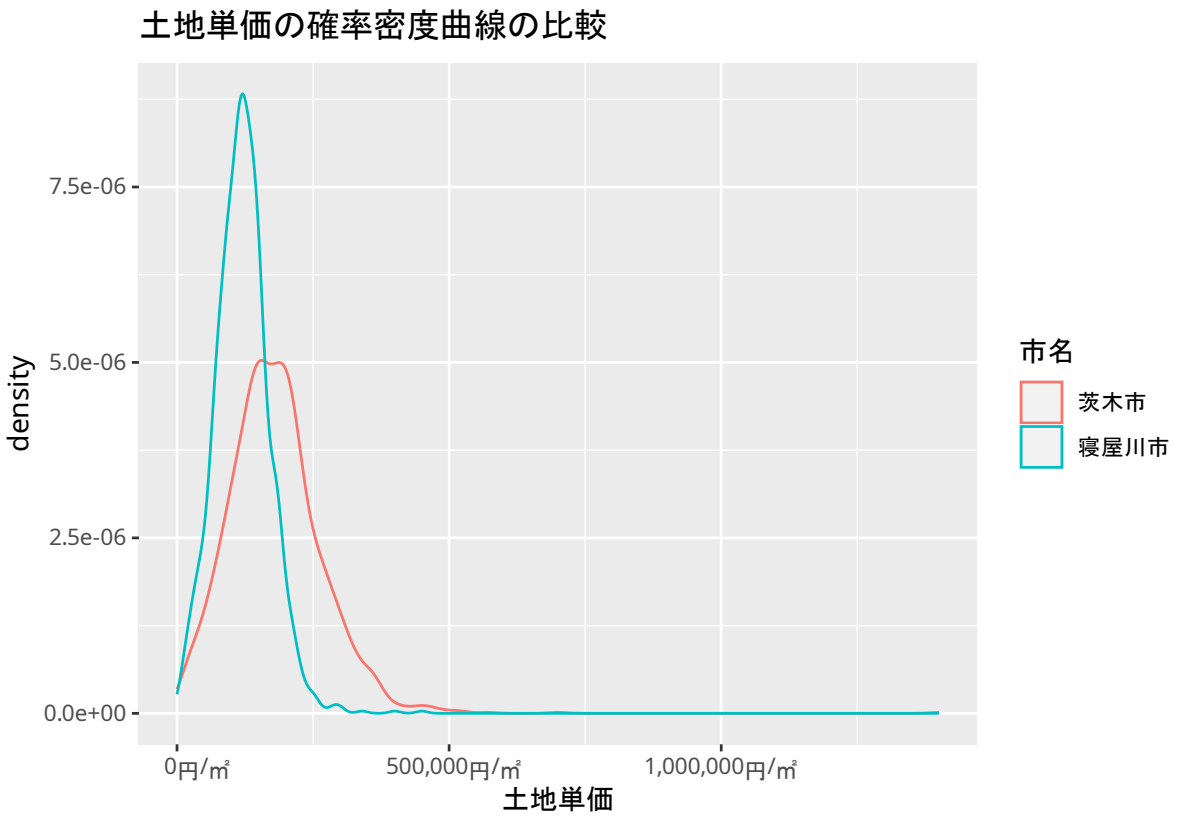
5.1　最大と最小

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。



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

複数の事柄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。

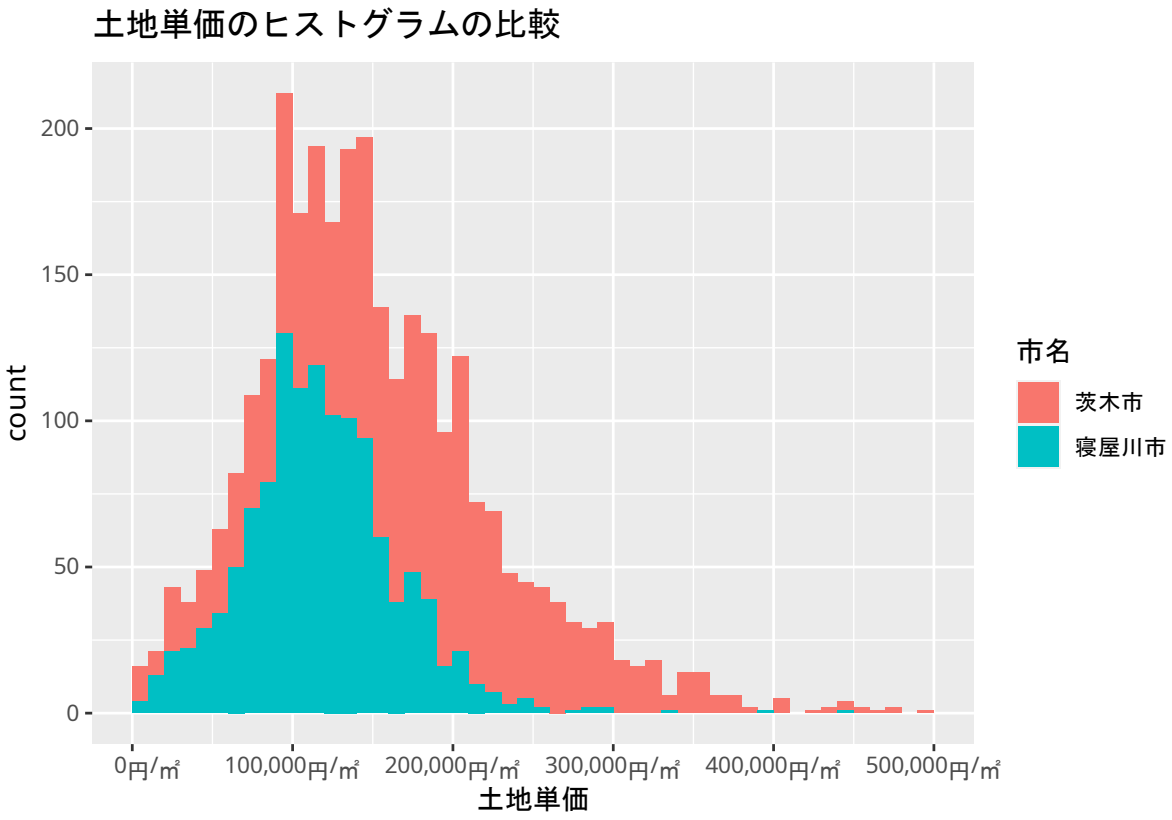
これに対して、茨木市の 20 万円／㎡強の地点のピークの高さが寝屋川市よりも低くなっています。これは、茨木市が寝屋川市ほど、価格の**偏りが集中していない**。すなわち、寝屋川市と**比較して価格がバラけている**ということを表しています。

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**のっぽ**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。

5.4　ヒストグラムで見る

では、ヒストグラムでも見てみましょう。こちらだと、取引の量の違い（どちらが多いのか少ないのか）がひと目でわかります。また、あらためて比較すれば茨木市の事例の分布の仕方は、寝屋川市に比べて広範囲に渡っている、すなわち、バラけていることが確認できます。



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具体的な数字で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っち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のように呼びません）

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／㎡

6.4　各区間の間隔が集中ばらつき具合を具体的に表す

四分位というのは、ある一定の順番の値です。この四分位各値の間隔に着目することで、その分布の偏り具合が数値的に把握できます。

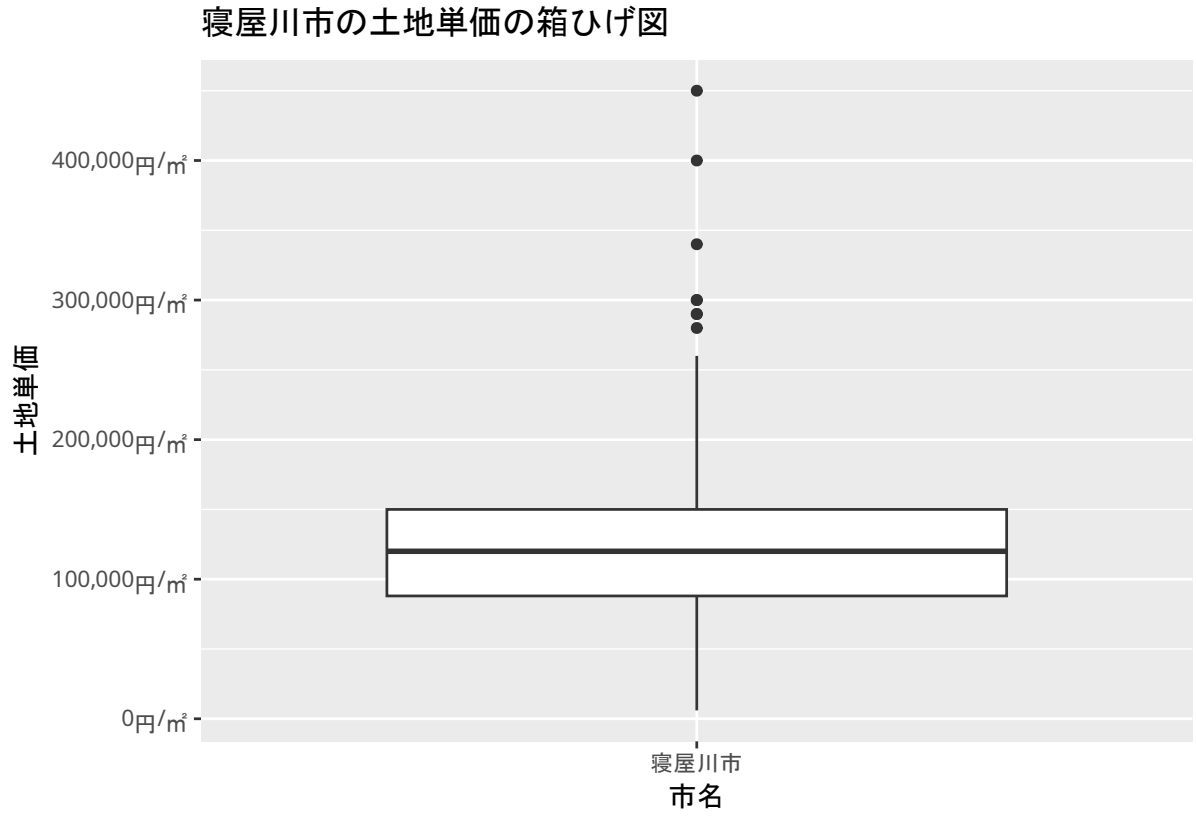
まず、第１四分位から中央値、中央値から第３四分位の間隔はともに３万円程度です。この間の事例は、全体の５０%、約６００件の事例あり、この６００件の事例が３万円＋３万円の６万円の間に密集しています。

一方で、第３四分位から最大値の間は３０万円もあります。この間の事例は全体の２５%、約３００件の事例ですが、先の６万円の間隔よりもずっと広い３０万円の間にまばらに存在していることになります。

四分位数を使うと、どれくらい集中していたり、疎らだったり、あるいは、中央値からどちらの方に分布の偏りがみられるかを具体的な数字で見ることが出来るようになります。

6.5 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。



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、その先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。そして、先の部分は、第１四分位から第３四分位までの距離の 1.5 倍の位置を示しています。

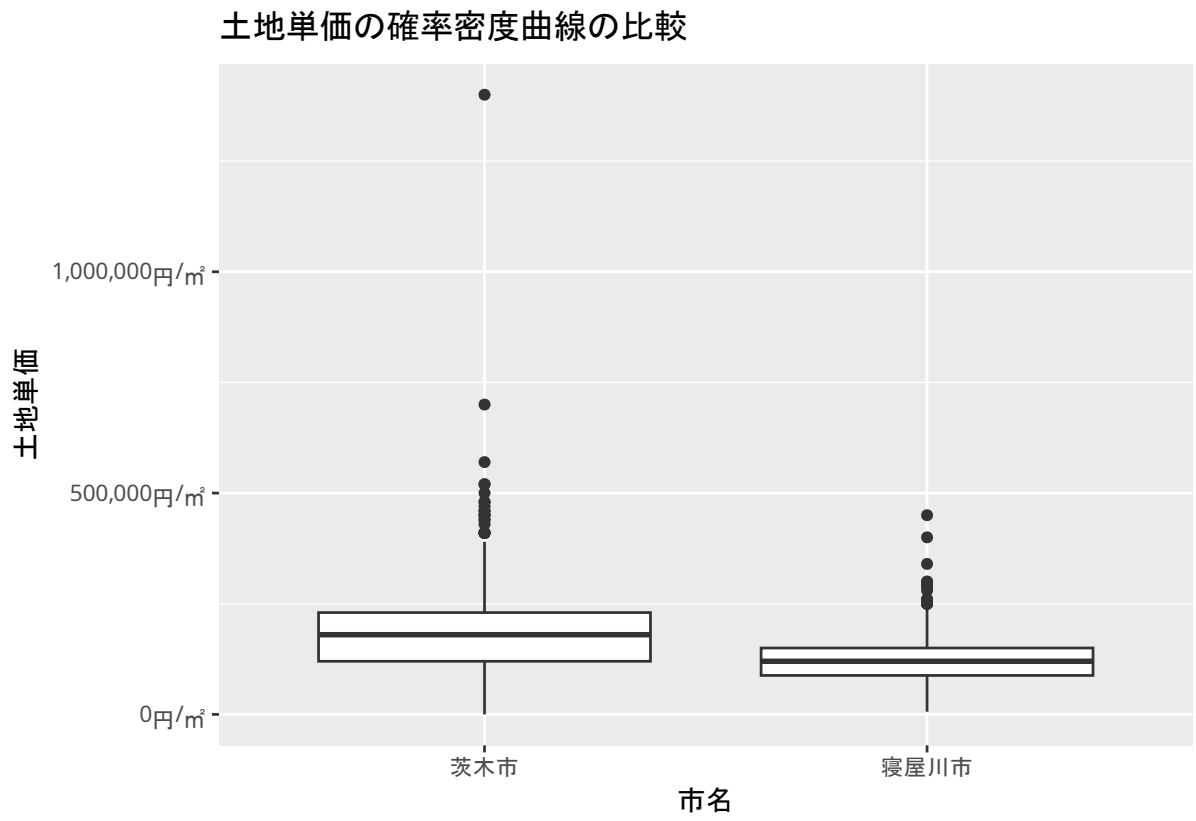
6.5.1　箱ひげ図の見方

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。

6.5.2　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てみます。あわせて、具体的な数値もそのあとに示します。



	茨城市の四分位数	寝屋川市の四分位数
最小値	26 円／㎡	6,000 円／㎡
第 1 四分位	120,000 円／㎡	88,000 円／㎡
中央位	180,000 円／㎡	120,000 円／㎡

	茨城市の四分位数	寝屋川市の四分位数
第 3 四分位	230,000 円／㎡	150,000 円／㎡
最大値	1,400,000 円／㎡	450,000 円／㎡

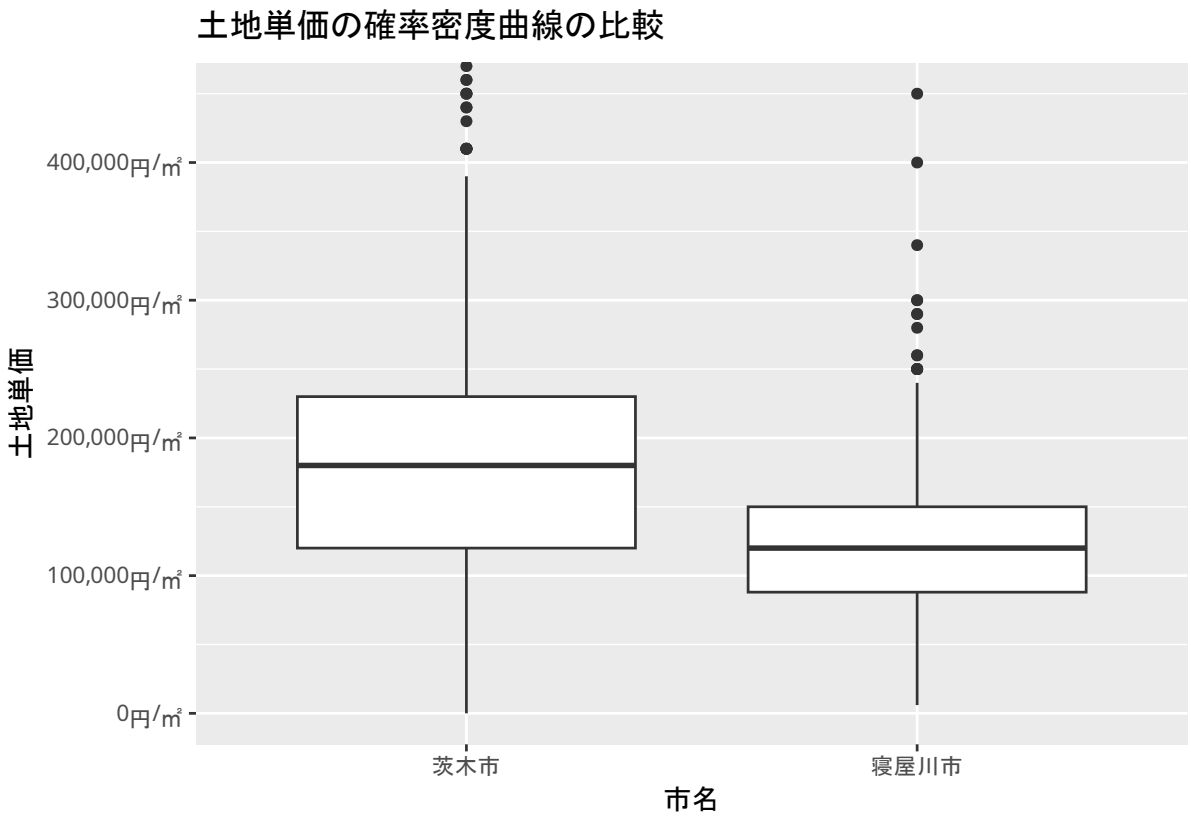
6.5.3 着目したい部分にズームイン

データを視覚化した場合、沢山データのある部分から凄く離れた部分にデータがあると上記の箱ひげ図のように、全体を見えるようにするために、相対的に長方形部分が小さくなって、でーたの詳細がわからなくなるときがあります。

そんな時は、「**グラフが見えにくい**から離れたデータを除外する」ということは、データを分析することになりません。なぜなら、「離れたデータが存在する」という事自体が、**複数データ**を分析していることになるからです。また具体的で分かりやすい、不都合な点としては、「データを外す」と、中央値や平均値全てが変わります。つまり、みやすさのために意味もなく代表値が変わってしまうわけです。

但し、あまりにもレベルの違うデータが存在するのは、それなりの理由があります。価格形成要因の分析目線で言えば、単価レベルであれば価格形成要因の異なるデータが混在していて、それらを分類して分析すべきであれば、逆に、結果として離れた位置にある**データを外さなければなりません**。このように、データを外すはずさないは、グラフが見にくいからではなくて、意味のある分析を行うための分類で決める必要があります。

一方で、たしかに上記のようにせっかく可視化したのに見にくい場合、データを除外して見やすくするのではなくて、グラフの表示する位置を変更して見やすくする必要があります。

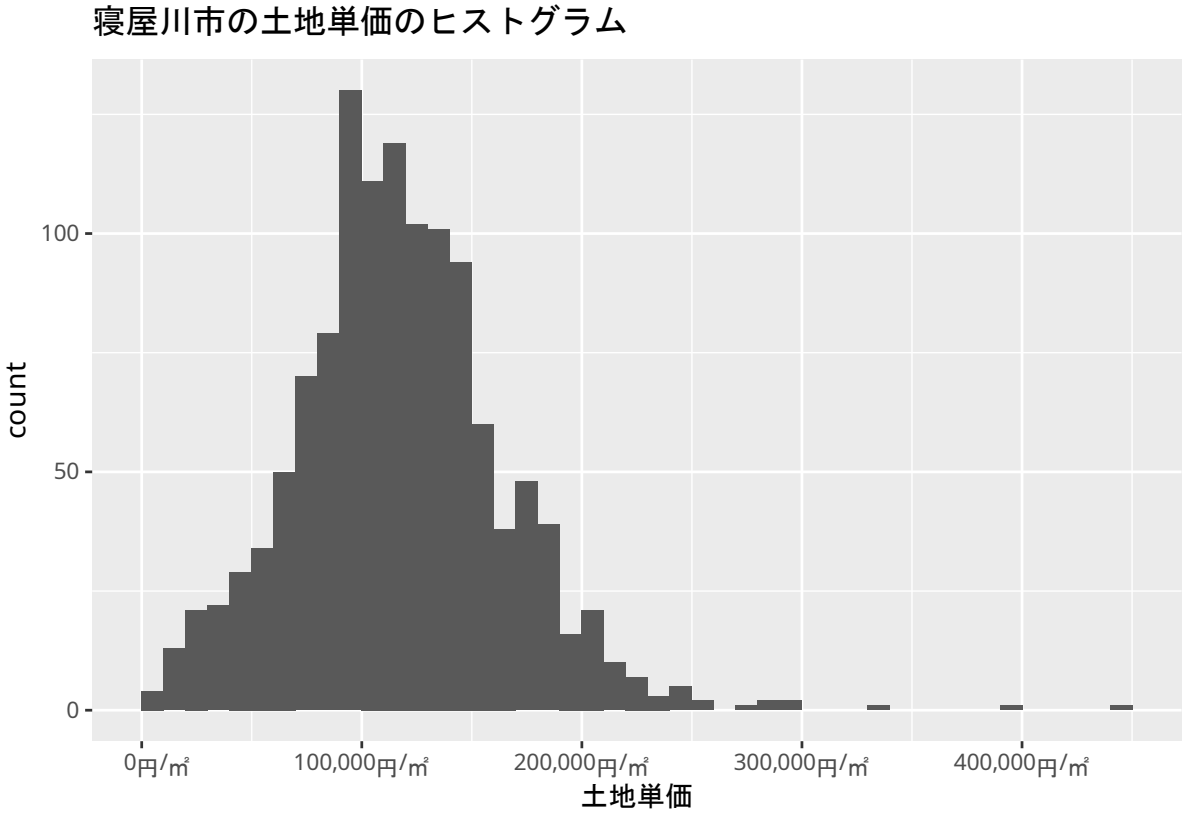


こうすることで、茨木市と寝屋川市の中央値の高低や、50% データ範囲の分布の厚みや水準の違いがはっきりと見えるようになります。

7　ヒストグラムから地価を推定する

7.1　分布と確率の話

今まで、みてきた複数事例のデータ分析の流れを振り返ってみます。まず、データがどのへんに沢山あるかをみるために、区間を区切ってそこにある事例の数を数えて、それをグラフにしました。すなわち、ヒストグラムです。



さて、このグラフは実際に取引された事例のデータに基づいた分布を表しているに過ぎません。しかし、ここで、あなたが事例作成担当者であり、その担当の地域が寝屋川市だとして、次に事例作成依頼が来るとしたら、その事例の単価はだいたい幾らぐらいでしょうか？

7.1.1　次の事例はいくらくらい？

もちろん、次に割り当てられる事例なので、何の用途が来るか、規模がどれくらいか、まったくわかりません。鑑定士の先生に聞けば、「用途も規模も地域も特性もわからないのに分かるわけない」と言われそうですが、5,000,000 円／㎡の事例とかあると思うかと聞けば、「うーん、もしかしたらあるかもしれんけど、普通無いわなぁ」と言うんじゃないでしょうか？上のヒストグラムの分布を見ている限り、現実的に考えて、1,000,000 円／㎡の事例でさえ、実際にあったとして、自分で事例の入力をしていたら「アンケートの人、桁間違えてる

んちゃうか?」、もしくは、「単価計算の桁間違えたかな?」と思うのではないでしょうか？

7.1.2　次の事例の単価を当てたら１０億円貰えます

さて、そんな怪しげなキャンペーンが万が一あったとして、「はい、５，４，３，２，１,、答えてください」と言われたら？

鑑定士の先生も「わからんけど、、100,000 円／㎡くらい?」と答えたくなるんじゃないでしょうか？そこには、価格形成要因がわかる、わからないの要素は何もいりません。**ヒストグラムで一番多いところが確率的に高いだろう！**と常識的に素直に推測したからです。

7.1.3　ヒストグラムで次の一手を予測するのはアリ？

そもそも、この推測は正しいのでしょうか？

物事の事象に偏りのある分布がなされている時、それらが**ランダム**に抽出されるのならば、その中の多いものほど抽出されやすいのは当たり前といえます。

ここで、振り返ってほしいのは、上述のヒストグラムは、アンケートから帰ってきて事例化されたデータの分布です。ですので、「次の事例」正確にい言えば、この事例化されたデータの分布とは関係がありません。

では、「次の事例」というのは、何の分布を元に予測すれば、10 億円の賞金に近づくのでしょうか？答えは「寝屋川市の全ての取引単価の分布」、若しくは、「寝屋川市の全ての土地の土地単価の分布」です。これらは、現実には知ることの出来ないデータですが、実質的に存在するデータであり、この分布をもとに考えれば、「次の事例」の単価水準にもっとも近づくことが出来ます。

7.1.4　母集団と標本

これらは言われてみれば、当たり前の話です。実際に事例化されたデータの集まりは**標本（サンプル）**といわれるもので、その背後にある事例化されていない土地の価格の全データが**母集団**といわれるものです。まずは、標本と母集団というものがあって、それらは別々のものであるということをしっかりと意識しておきましょう。そして、今扱っているデータが、標本なのか母集団なのかをきっちりと認識するようにしましょう。

では、**なぜ、母集団と標本の違いを何故意識しないといけない**のでしょうか？

母集団と標本の違いは、母集団が全部のデータで、標本は母集団のデータの一部であり、母集団より少ないデータだということです。そして、一番はじめに考えたとおり、母集団の分布と標本の**分布が一致しているのかどうなのかわからない**からです。もし、一致しているのならば標本の分布の一番多いこところを選ぶことで、母集団から得られる次の一手を予測できます。しかし、一致していないのならば、標本の分布を使っても意味がありません。

常識的に考えると「標本の数が増えれば」、標本の分布が母集団の分布に近づきます。経験的に知っている「サイコロの目」の話です。

母集団の分布と標本の分布が一致しているかどうかは、標本の数が少ないと、一致しているかどうか自体がわからず、標本の数が増えるほど一致している確率が高まっていきます。つまり、ざっくり言えば、**標本が沢山あるとき**には、標本の分布が母集団の分布に一致している可能性が高いので、標本の分布で「次の事例」の価格水準を推定することで賞金 10 億円に近づけるのです。

7.1.5　統計学からの視点

大雑把な感覚としての話では、この「標本が沢山ある時なら、確率が高い」ということが **大体正しい**という認識で十分です。

但し、統計学的な話になると、どういう分布の種類の時で、どれくらいのサンプルがあれば、どれくらいの確率で**平均**が一致しているかを **計算で理論的に**求めることが出来ます。

ここでは、その計算等は紹介しません。ただ、ここまで見てきたように、物事に偏りがあり、その偏りに規則性があって、それらの抽出を確率で考えれば、色々なことが確率として計算できるということを認識すれば十分です。

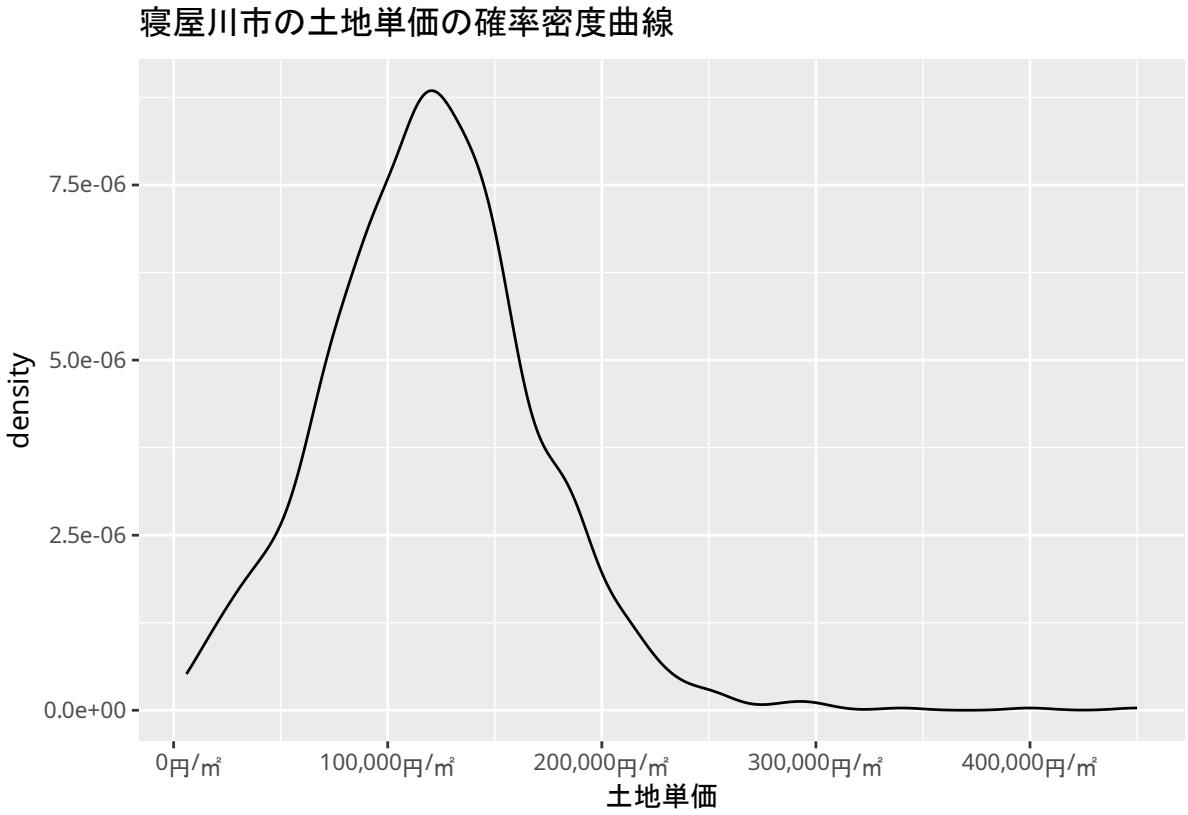
データの多い少ないを感覚的に判断した、**なんとなくの経験則ではない**のです。

7.1.6　次の事例の地価予測と地価の予測

もともとのテーマは、複数事例のデータの見方であり、沢山の複数事例を地価水準として把握するという話です。この前の項目での話題では、「次の事例」というテーマでしたが、これは、この地域からランダムに土地を指定した時の価格の様なものです。その事自体に鑑定評価上、何か意味があるのか無いのかは別として、これも複数事例から得られる不動産の価格の一視点だと言えます。

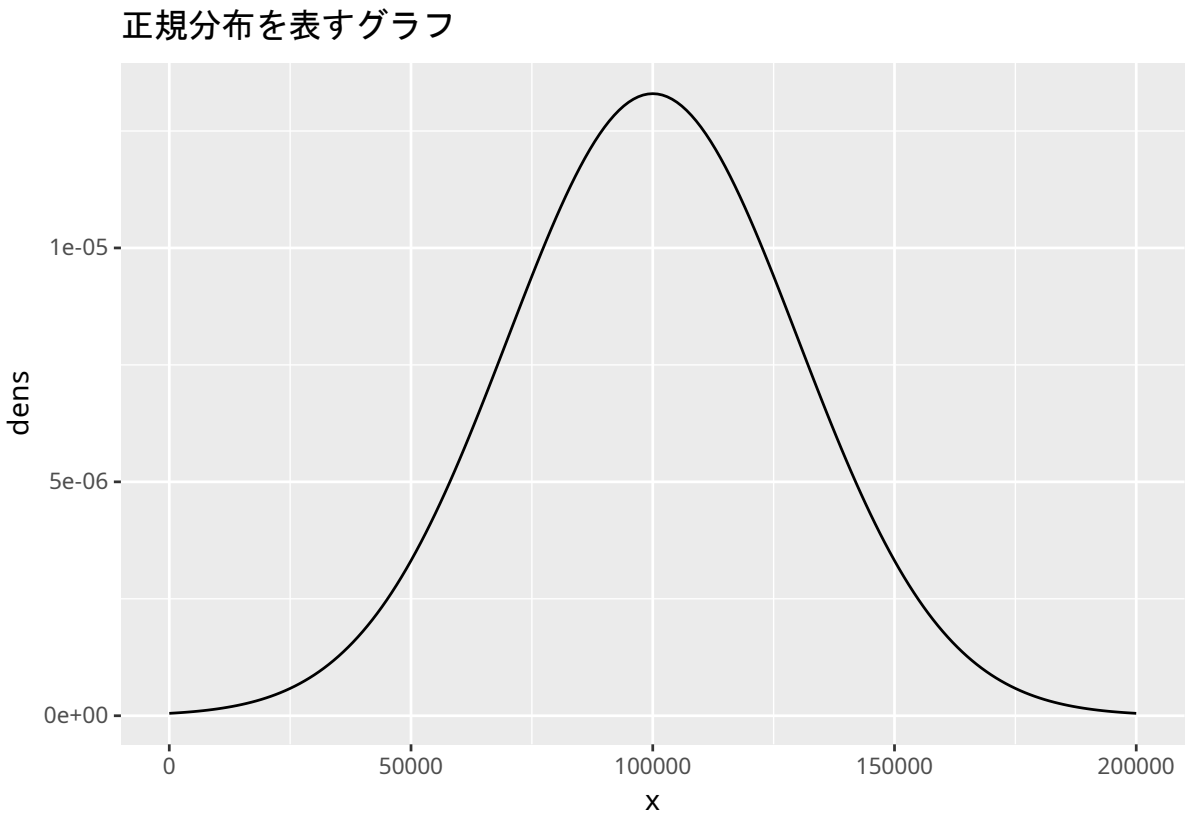
8　正規分布

先に、ヒストグラムの他に確率密度曲線というものを見ました。実際に可視化したのは寝屋川市の事例である「標本」を可視化したものです。あらためてみてみましょう。



これは、現実の標本の分布です。現実の分布なので、偏っていたり凸凹していました。

一方、これに対して、ある規則に従った綺麗な分布を可視化したものが次のグラフで、正規分布と呼ばれています。



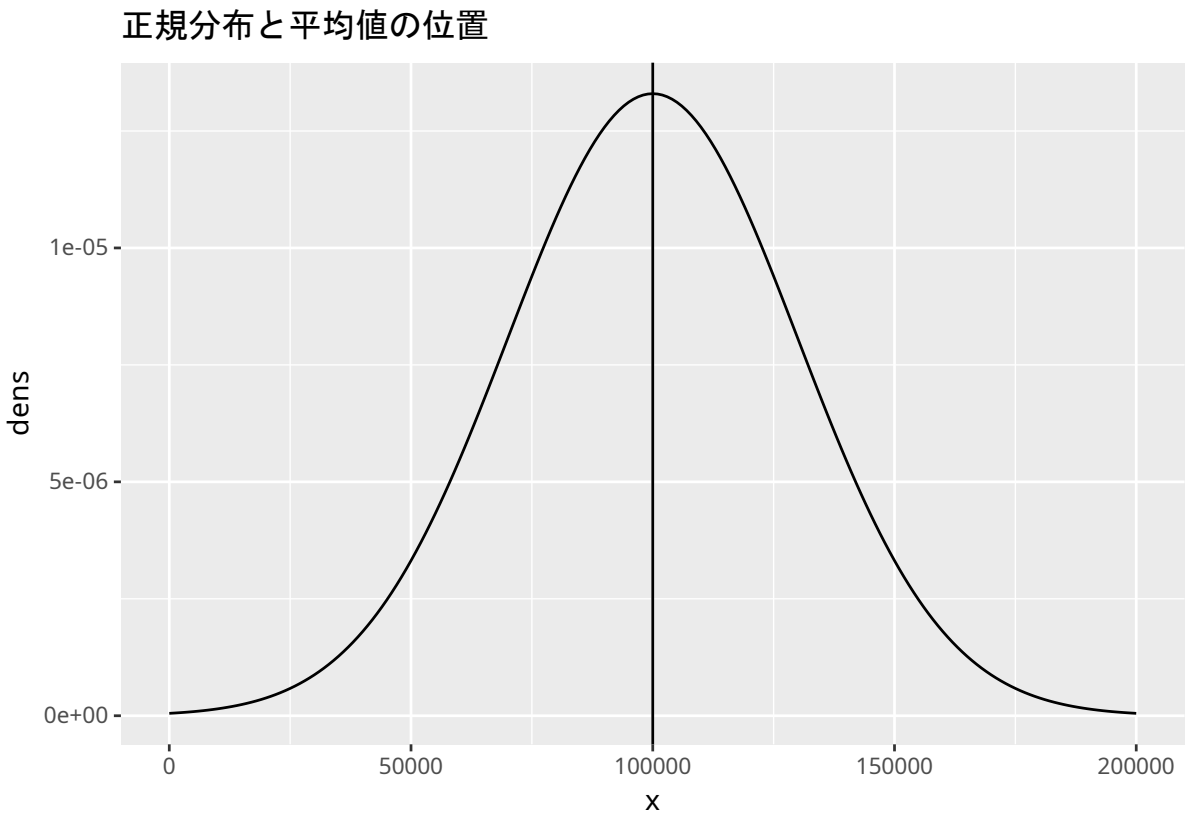
このグラフは、左右対称で綺麗な凸型になっています。よく見ると、横軸の 100000 の目盛りのところが凸型のピークです。つまり、今までみてきた分布のグラフの見方から考えるに、100000 が最もありそうな値で、それから増えても減っても、同じ様な具合にその頻度が減っていく値の集団を表現していると考えることが出来ます。

これが、どこかの地域の取引事例の確率密度曲線を描いたものだとすると、100,000 円／㎡の事例が最もたくさんあり、そこから高かったり安かったりする事例が綺麗に減っていってることを表していると読めます。

「**正規分布**」という言葉は、統計やデータ分析を齧ると、どこかで出てくるので聞いたことのある言葉だと思います。形も見たことがあるのではないでしょうか？しかし、これが何を表しているのかあまりわからなかったとしても、今回ここまで複数データの分布の話をしてきた今では、物事の**分布の話**であるということがなんとなく把握できたのでは無いでしょうか？

8.1　正規分布と平均

さて、なぜ、正規分布が話題としてよくでてくるのでしょうか？最も分かりやすいのが**平均**です。みなさん、複数のデータがあると**すぐに平均**をとります。なぜでしょう？それは、事象の分布が正規分布の時、平均値は分布のピークになっているからです。つまり、分布が正規分布の場合には、一番多いところが平均なわけです。そして、物事の分布は正規分布であることが多いからです。

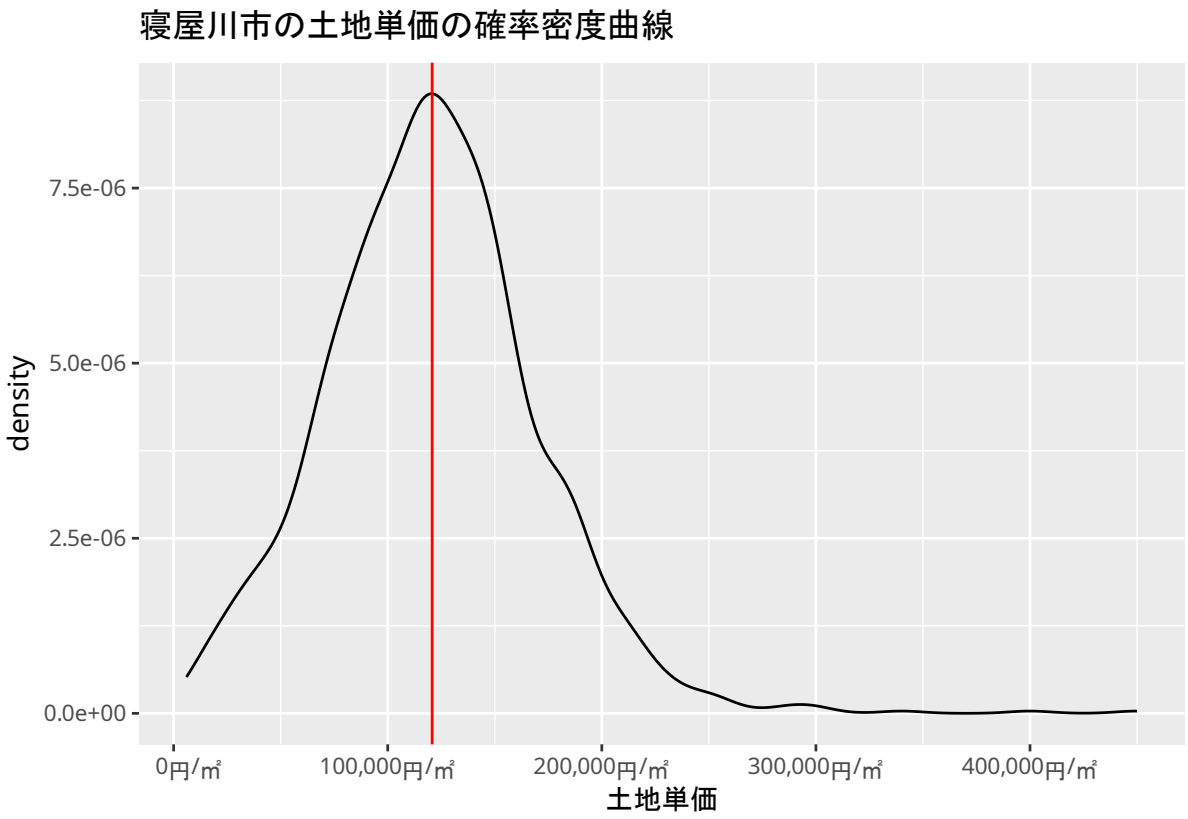


更には、サンプルがある程度沢山あって、正規分布に分布している場合、母集団も同じ様な分布である確率が高まります。実際には、母集団が正規分布している場合、標本の平均（何回も標本を取る）は正規分布になって、その平均が母集団の平均になります。

すなわち、ここまでの知識を総動員すると、何らかのサンプルを得て、そのサンプルの数がある程度沢山ある時、だいたい物事の母集団は正規分布の可能性が高いので、次に抽出される事象（次の事例）は、サンプルの平均値の確率がだいたい高く、あながち間違ってない値ということになります。だから、みんな**平均をすぐにとる**のです。

その上で、さらにここまでの知識を付け加えれば、母集団が正規分布でない場合や、標本の数が母集団に対して精度が出ないほど少ない場合には、**当てにならない事もアリ得る**ということがわかったはずです。

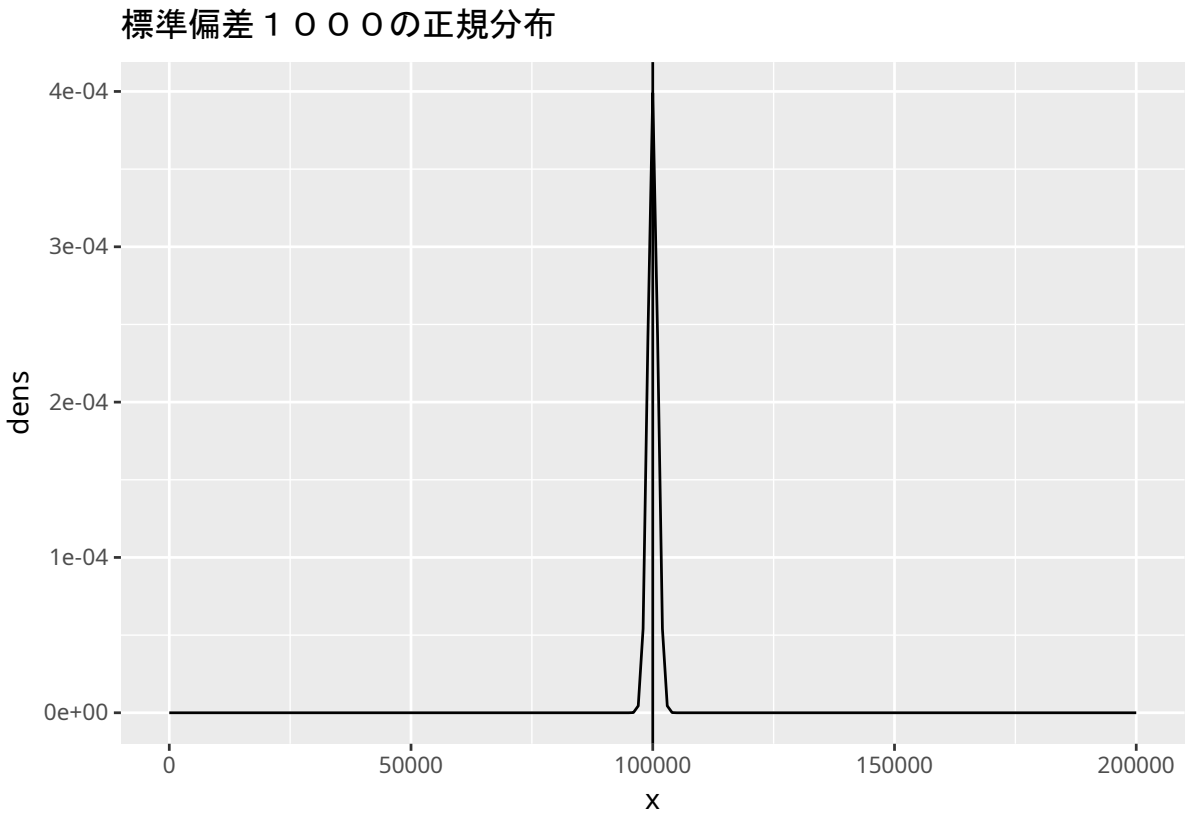
ちなみに、寝屋川市の取引に関する分布と平均の関係は次のようになります。


寝屋川市の土地単価の確率密度曲線
7.5e-06
5.0e-06
2.5e-06
0.0e+00
density
0円/㎡
100,000円/㎡
200,000円/㎡
300,000円/㎡
400,000円/㎡
土地単価

9　分布のばらつき具合の物差しである標準偏差

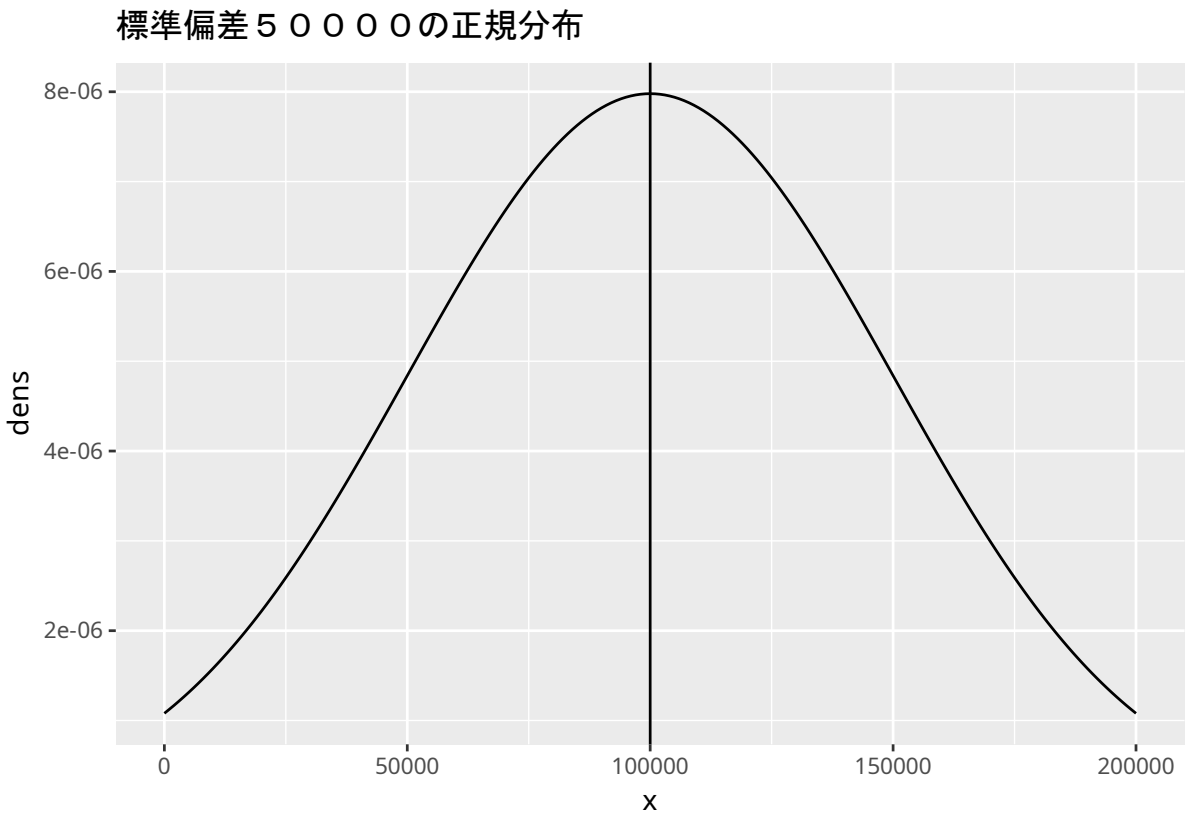
ここまで、**最も多い部分**に着目していましたが、複数のデータの性質である、分布の**ばらつき具合**にも着目してみましょう。次に、同じ平均値 100000 を持つ正規分布である次の 2 つのグラフを見てください。

まずは、平均の辺りに集中した分布です。



この分布の標準偏差は 1000 という値です。

一方で、次のグラフは平均から離れた部分にも分布のある緩慢な分布です。

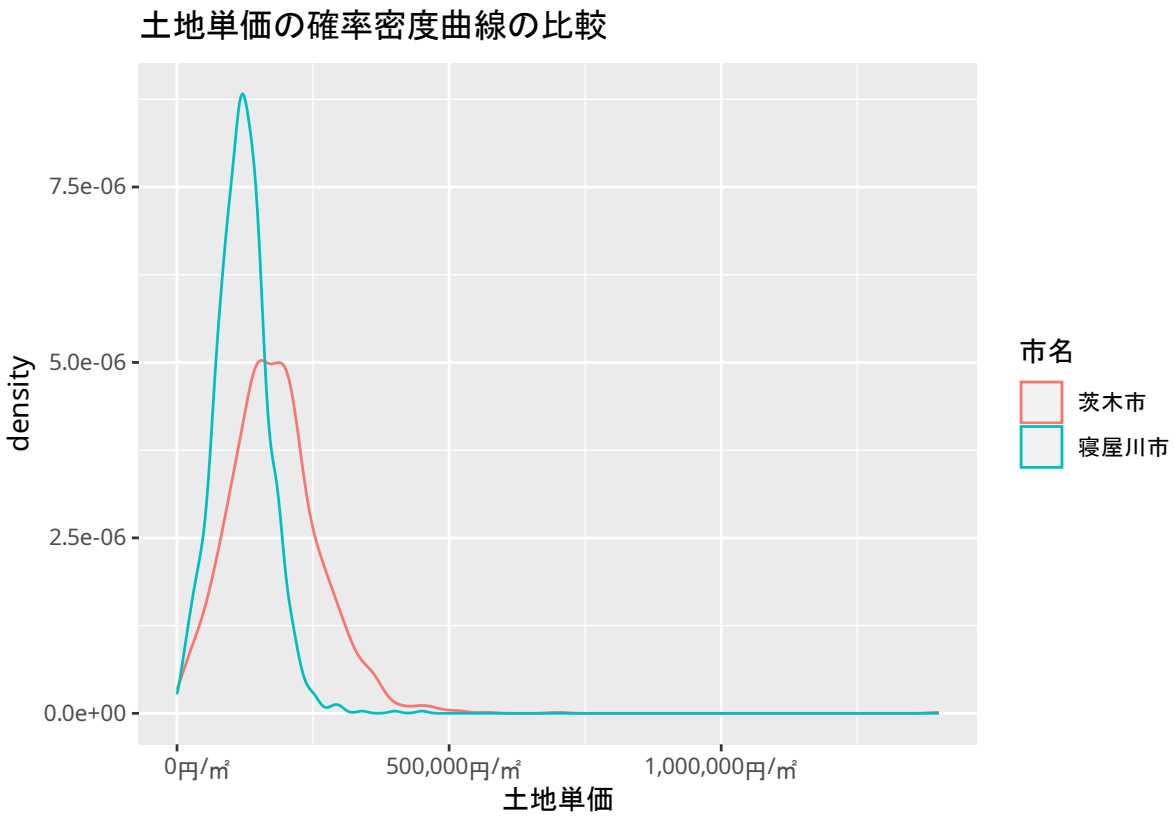


この分布の標準偏差は 50000 という値で、先に分布の 1000 という標準偏差に比べて凄く大きい値です。

さて、ここで**標準偏差**とは何かというと、このデータ集団の個別の各データとこのデータ集団の平均との距離の平均の平方根です。つまり、計算式が表しているのは、平均と各データの距離の大きさであり、この標準偏差という数値が小さいと、はじめの分布のように、平均の近くにデータが集中して分布しており、数値が大きいと、平均から離れた場所にもデータがある疎らな分布になります。なので、あるデータの複数データについて標準偏差を求めることで、その分布の**ばらけ具合が数値化**出来るということになります。

9.0.1 茨木市と寝屋川市の標準偏差

茨木市と寝屋川市の土地単価の分布の様子は、先に見たとおり、次のようになっていました。



この茨木市と寝屋川市の土地単価の分布について、標準偏差を求めると次の様は数値になっています。

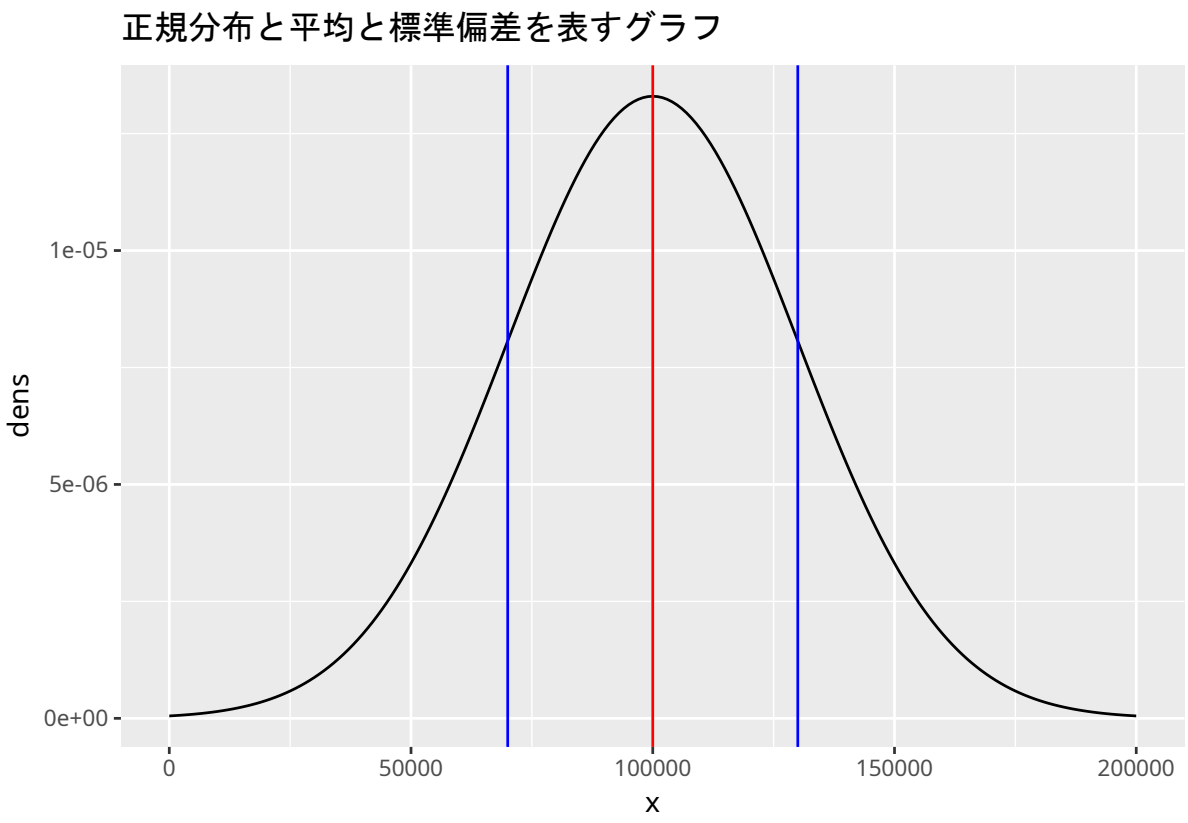
	茨城市	寝屋川市
平均	180,000 円／㎡	180,000 円／㎡
標準偏差	88,176 円／㎡	49,315 円／㎡

先に話したとおり、グラフから見ても茨木市のほうがデータがバラけていますが、標準偏差という具体的な数値を使うと、「寝屋川市と茨木市を**比較**して、茨木市は約 40,000 円／㎡程、標準偏差でバラけている」と具体的な数字で比較の説明をすることが出来るようになります。

10　標準偏差と正規分布の確率

先に、確率密度曲線は、グラフの下側の面積自体がその時点までの分布の割合を表しているという話をしました。

ある正規分布のグラフをもう一度見てみます。ここのデータの分布は平均値が 100,000、標準偏差が 30,000 の正規分布になるデータです。そして、このグラフに分布の平均値 100,000 の位置に赤い線を、平均値から標準偏差の値だけ離れた位置となる 70,000 と 130,000 の位置に青い線を引きます。

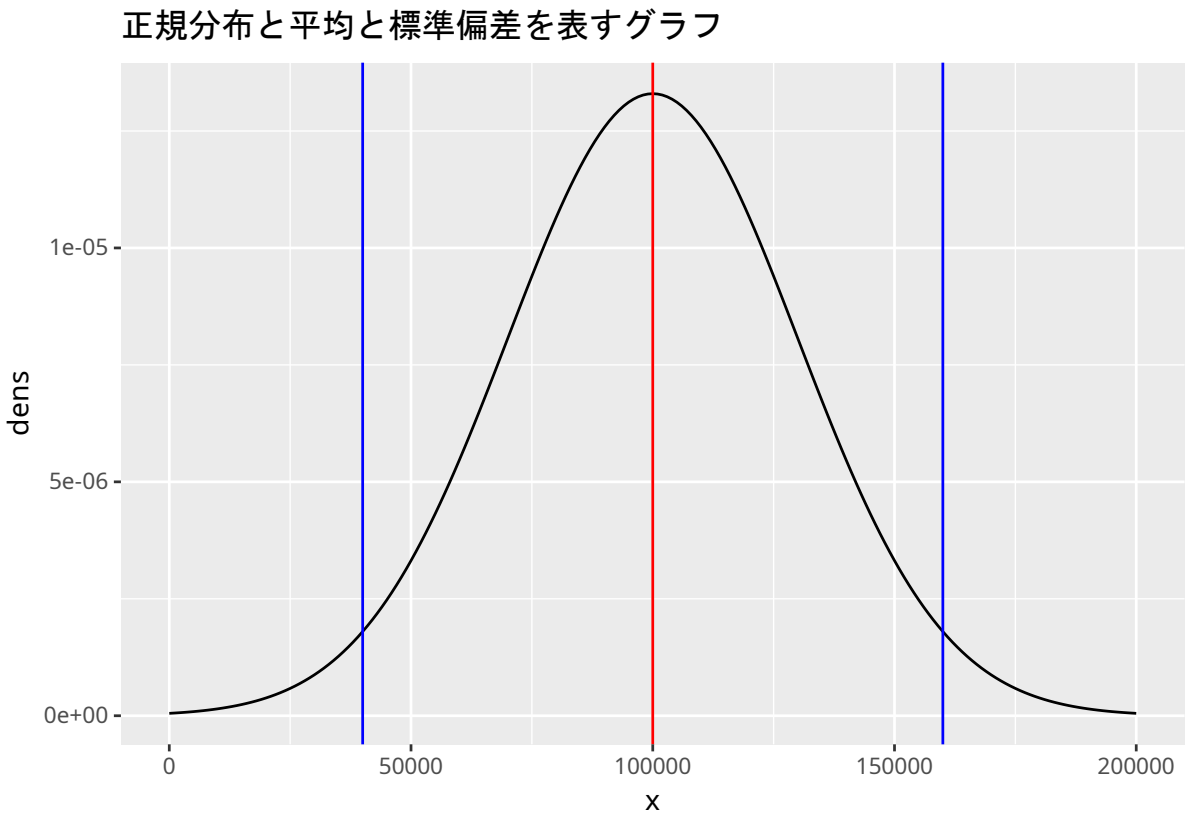


この時、実はこの線で区切られた部分の面積割合は決まっています。分布曲線の下側の面積全体に対して、70,000 の位置に引かれた青い線の左側の部分の面積の割合は必ず約 16% と決まっているのです。逆側の 130,000 よりも右側の部分も同様の約 16% です。なので、その内側の中央部分は約 68% です。

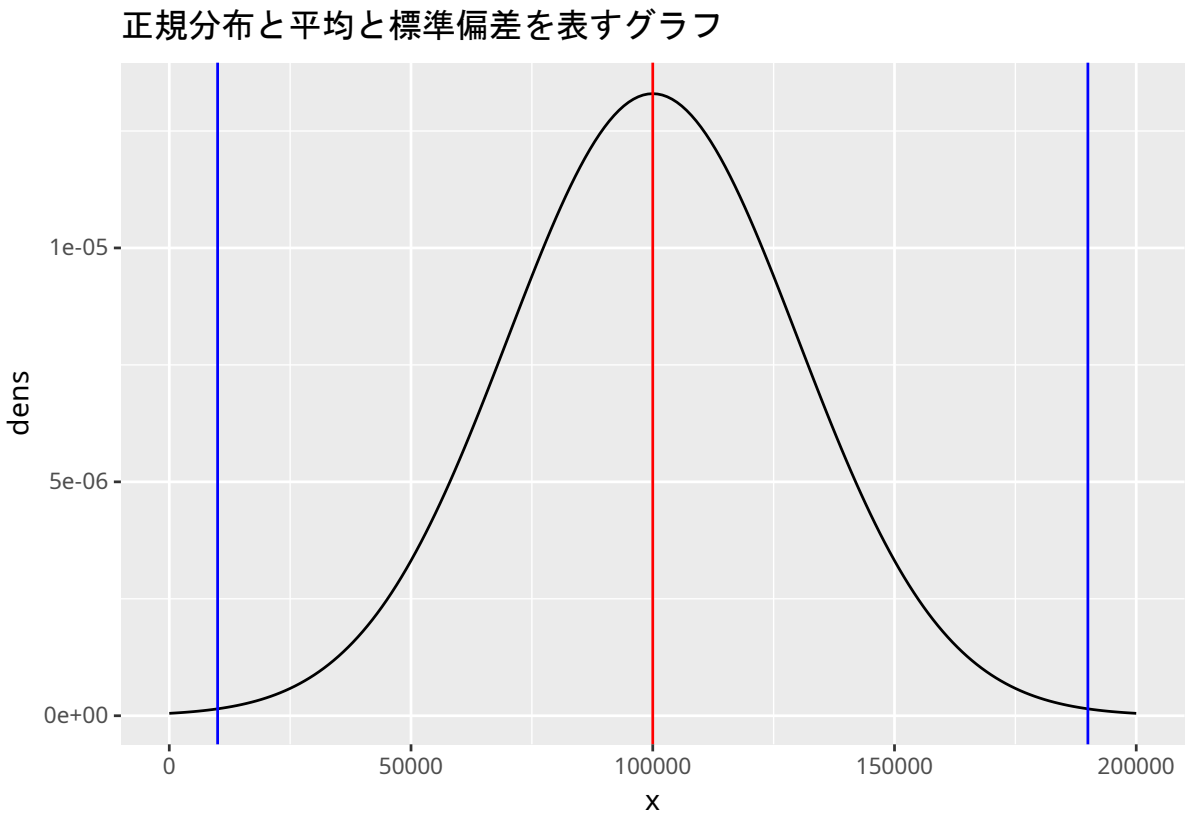
この割合自体は、平均や標準偏差の値によって変動しないのです。分布が**正規分布**である時の性質であり、計算により求まるのです。平均から標準偏差だけ離れた位置の内側に分布するデータの割合が約７割であるということですが、しかしこの**約７割**というのは何かを語る時には微妙は数字です。

そこで、よく用いられるのが、平均から標準偏差の２倍、若しくは、３倍離れた位置です。

まずは、２倍の位置、40,000 と 160,000 の位置に線を引いてみましょう。



この平均から標準偏差の 2 倍までの距離の内側の部分の全体面積に対する割合は、**約 95%** になります。
さらに、 3 倍の位置、10,000 と 190,000 の位置に線を引いてみましょう。



この平均から標準偏差の３倍までの距離の内側の部分の全体面積に対する割合は、**約 99.7%** になります。

10.1　標準偏差と正規分布と確率

はじめに話したとおり、ここで表示されている分布のグラフは確率密度関数による分布曲線で、グラフの下の部分の面積は分布の**確率**を表現しています。例えば、何が約 99.7% なのかといえば、寝屋川市の母集団が上記の正規分布であるとしたら、

先の１０億円の賞金のキャンペーンで、「10,000 から 190,000」と答える事が出来るのなら、99.7% の確率で１０億円が貰えるということです。応募要領として、範囲は 120,000 までといわれたなら少し確率は下がりますが、「40,000 から 160,000」と答え、約 95% の確率で１０億円を貰いましょう。

10.1.1　標準偏差の大きさ

ここで、平均から標準偏差までの距離の内側にある確率が計算できて決まっていることから、その範囲のことを**標準偏差の何倍**と表現することが多く、標準偏差を σ シグマという記号を使って表され、2σ や 3σ と表現されたりします。

そして、「１０億円チャレンジ」の例えで考えて、もし、いくつかある分布の中から分布自体を選ぶことが出来るなら、標準偏差が小さい分布を選べば良いことになります。なぜなら、それら分布の標準偏差の具体的な数字が大きくなればなるほど、同じ割合でも広い範囲から選ぶ必要があり、逆に、標準偏差が小さいと、選ぶ範囲が少なくなるからです。選べる範囲が固定ならば、標準偏差が小さい分布を選んだほうが１０億円のチャ

ンスが近づくのです。

11　複数のデータからみる価格の水準の説明

はじめに話したとおり、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様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があることもわかりました。

平均と分布の関係やデータには標本と母集団というものがあり、そこには何らかの関係性があることも把握しました。

さらに、分布というものの中には正規分布というものがあり、正規分布では、標準偏差等を物差しにして確率を利用できることも把握しました。

この講義では、そういうものが**ある**ということを紹介しただけで、実際に何故そうなのかや、具体的な計算での証明はされていませんが、まずは、経験や勘で決めなくても、統計的な具体的な数値や計算を物差しにすることが**出来る世界がある**という事を認識しておくこと自体に意味があると思います。

この文書では、複数の事例データを考察する具体的な目的として、寝屋川市の地価水準ということをイメージしていました。鑑定評価での市場分析として「よく見られる取引の価格帯は、〇〇円／㎡から〇〇円／㎡程度」なんていう言い回しをよく使うと思います。

もちろん、専門的な業務を繰り返して行っている経験からよく知っている地域ならば、この幅を瞬時に決めることも出来るかもしれません。しかし、この文章を読んだ先生達は、鑑定の目線と全く違ったデータの考察という観点、すなわち、四分位数値や、標準偏差を用いて具体的な数値を記述してみたり、または、寝屋川市と茨木市のように代替関係にある地域間の相対的位置関係を分布そのものグラフ、それを数値的に表現した平均や標準偏差、最大、最小、四分位等を **道具**として利用し、説明する事の可能性を感じたのではないかと思います。