ビジネスパーソン向け
統計データ利活用1dayセミナー

総務省統計局統計 データ利活用センター

総務省統計局

マニュアル内容

第1章　統計学でできること……P3
1)統計学の使い方（記述、探索、検証、予測）
2)ビジネスにおける公的統計の活用事例
3)データのまとめ方と比べ方（平均・割合・t検定とχ二乗検定）
4)数値データ同士の関係性の見方（単回帰分析）
5)複数の説明変数を同時に使う方法（重回帰分析）

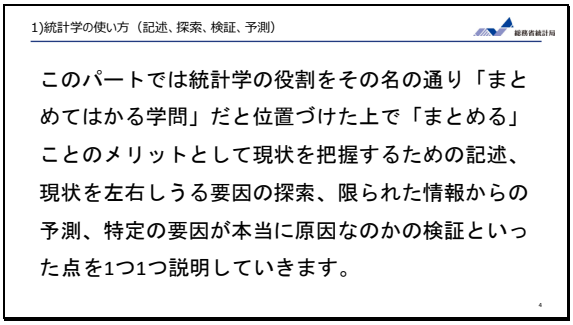
第2章　統計データに触れてみよう……P150
1)e-Statからデータのダウンロード
2)エクセル上でのデータ加工
3)デモ：公的統計のデータ分析

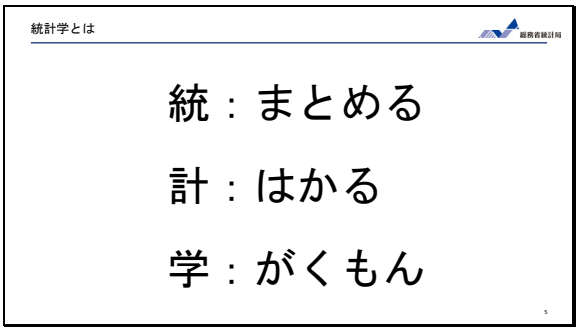
第3章　統計データを組み合わせよう……P197
1)サンプルデータと公的統計の結合(vlookup関数)
2)デモ：重回帰分析を用いた店舗ごとの売上関連要因の探索
3)jSTAT MAPを用いた有望エリアの探索
4)まとめとビジネスへの活かし方

総務省統計局

第1章　統計学でできること

このパートではまず、統計学を使う一般的な意義を説明するとともに、サンプルデータを触りながら基礎的な分析手法を学んでいきます。

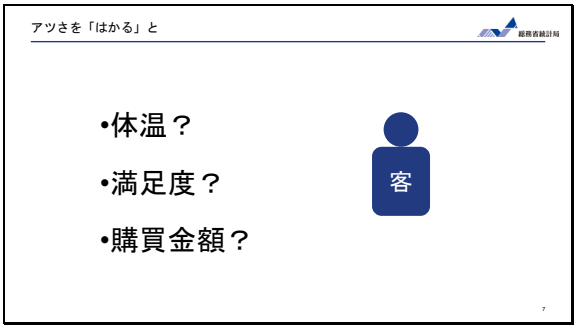




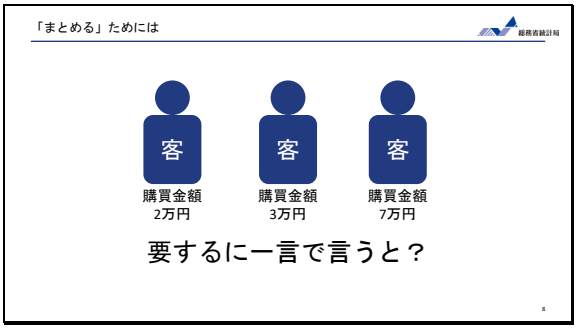
統計学の統とはまとめる。
統計学の計ははかる。
まとめてはかる学問が統計学です。



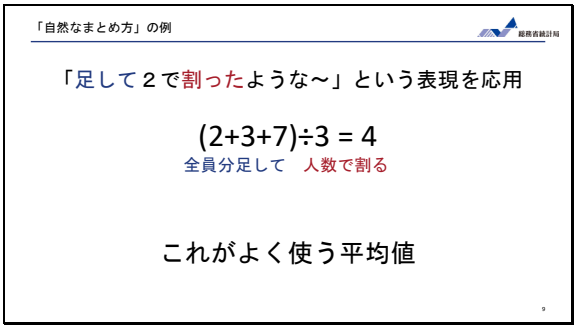
まとめて、はかる前の状態だと。
一般的に「あつい」という言葉があったとしても何を意味しているか曖昧です。



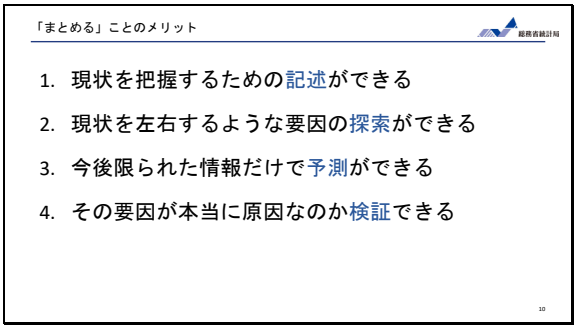
例えば、あついといっても体温であったり、お客さんが盛り上がって満足していたということかもしれません。
もしくは、購買額が多かったという意味で「あつい」という言葉を利用したかもしれません。
具体的に何の指標で数量化できるかということがはかるということになります。



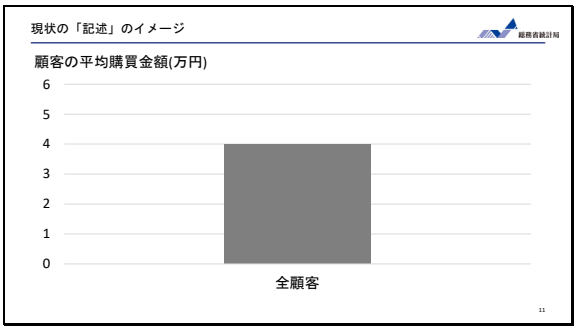
例えば3人顧客がいた場合、それぞれの購入額が異なると全て表現するのが大変となります。そういった際に統計学の表現が必要になります。



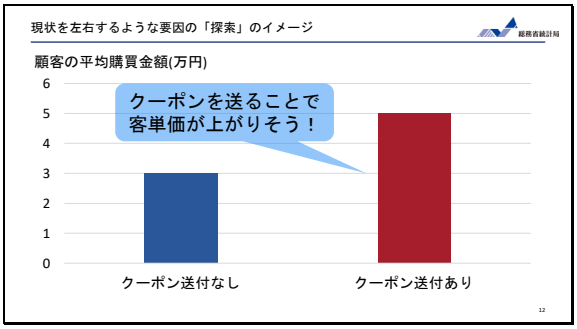
一例として、すでに私たちは、平均値という表現を一般的に利用しています。
全員分の金額を足して、人数で割ると平均値が求められるというのはよく利用されていることと思います。



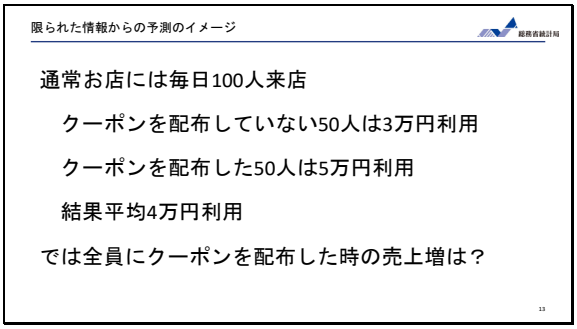
平均値以外にもいろいろなまとめ方がありますが、そのメリットとしては
1.個別の数字を全て含める必要がなく、一言で記述できます。
2.人によって異なることがある場合、現状を左右する要因を探索できます。
3.例えば一人平均いくらの購買力があるかわかれば、来客数を掛け合わせることで売上予測ができます。
4.自分が考えた要因が原因となっているか検証ができます。



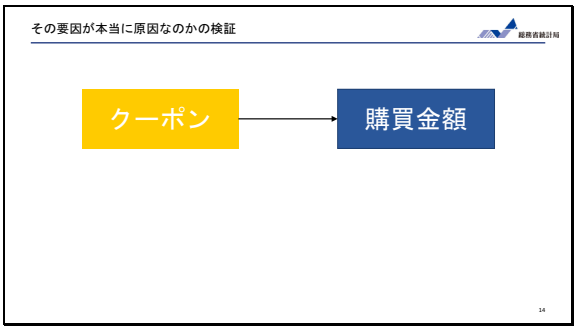
現状の記述例として、全体の顧客の平均購買金額をグラフで把握します。
顧客の平均購買金額が4万円とわかります。



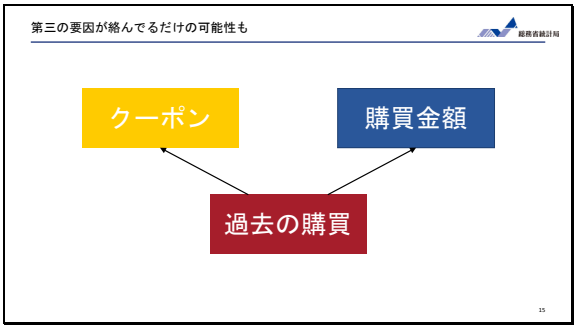
例えば、現状を左右する要因を探索しようとすると自社の顧客の状況を見て、クーポンを送付した顧客とクーポンを送付しなかったに分けて比較した場合、そこに差があるとクーポンを送ることで客単価を上げることがイメージできます。



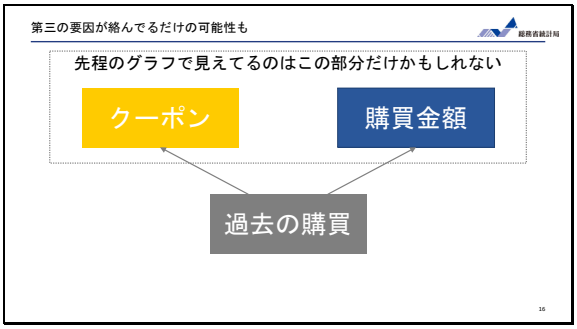
さらに予測で使用する場合、クーポン配布者、未配布者でそれぞれ利用金額が異なるが平均で一人当たり4万円の売り上げだと一日で400万円の売上が見込まれます。
この場合、もしクーポン未配布者へクーポンを配布するといくらの売上増が予測されるでしょうか。
未配布者が50人いるので、全員がクーポン配布者と同額利用すると仮定すると一人2万円×50人で100万円の売上増が予測されます。



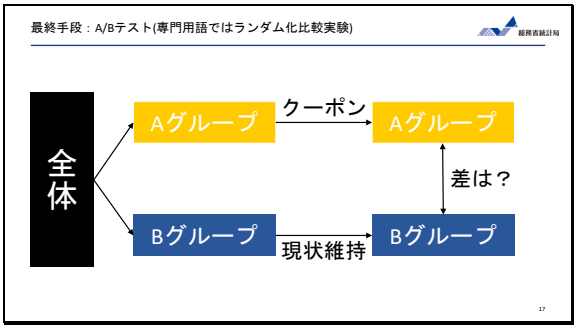
ただし、予測をする際に注意する必要があるのが、本当にクーポンを送ることが原因で、売上増という結果になっているかという関係を実際には、検証をする必要があります。



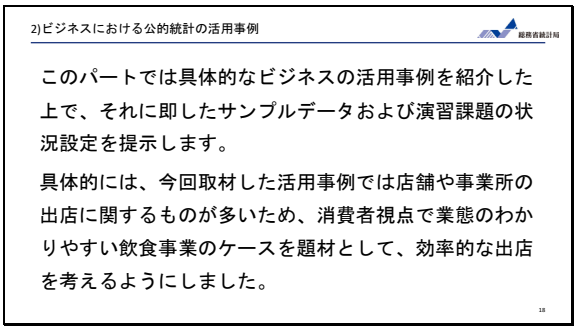
例えば、クーポンから購買金額を上げられるという因果関係はなく、実は第三の要因が関係している可能性があります。
一例として過去の購買が多い顧客にクーポンを配布している場合、そもそも購買金額が多い顧客であるためそのことが購買金額に影響している可能性があります。



もし過去の購買金額を見ていなかった場合、クーポンと購買金額は一見関係しているように見えますが、今後、クーポンを顧客に配布しつづけても実際の購買金額増につながらない可能性があります。



結局今あるデータだけで分析をするのは限界があります。
そういった因果関係を検証するためには、ランダム化比較試験（A/Bテスト）といわれる手法を利用します。
全体の顧客をランダムにAグループ、Bグループに振分けます。（例えば半分ずつや件数が多い場合は一部の顧客をピックアップ）
そしてAグループにはクーポンを配布し、それぞれのグループの購買金額を比較します。
もしAグループの方が購買金額が高い場合、クーポンが原因で購買金額が上がったと言えます。





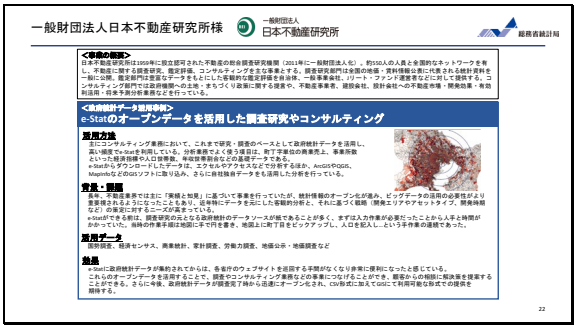
フランチャイズ研究会様の事例となります。



株式会社ゼンリンマーケティングソリューションズ様の事例となります。



株式会社パスコ様の事例となります。



一般財団法人日本不動産研究所様の事例となります。

本研修の状況設定

総務省統計局

- あなたは全国に「ちょっと高級な」焼肉店を展開する会社の幹部
- 新たな店舗を出店したい
- デモデータから「どのような店舗を」「どのあたりに出店するのが良さそうか」を考えましょう

23

今回の研修の設定です。焼肉店の幹部として、新店舗出店に向けた分析を行います。デモ・演習用のサンプルデータは本設定に基づいたデータとなります。

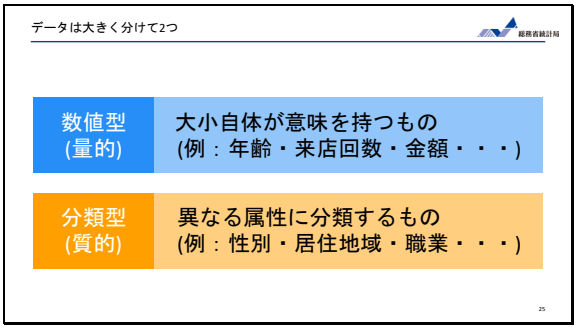
3)データのまとめ方と比べ方（平均・割合・t検定とχ二乗検定）

総務省統計局

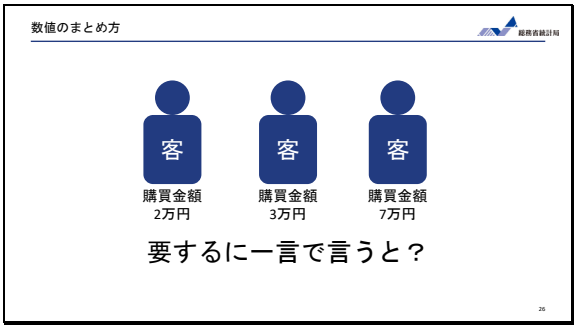
まずは「記述」の観点から、データには量的なものと質的なものがあることを説明した上で、量的なものは平均値、質的なものは割合の形で集計を行うことを説明します。

その上で、何かのグループで分けて平均値や割合を集計することで「探索」ができることを示し、その差がたまたまの差と言えるようなのかどうか、それぞれ対応する統計的仮説検定のエクセルでの実行方法を示します。

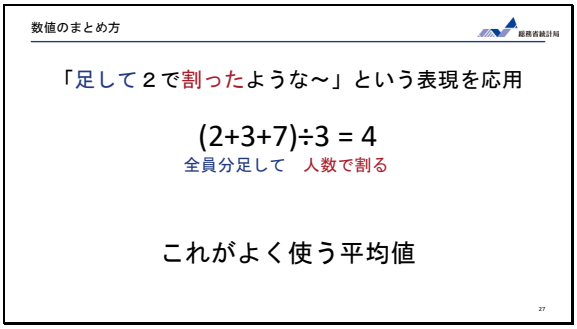
24



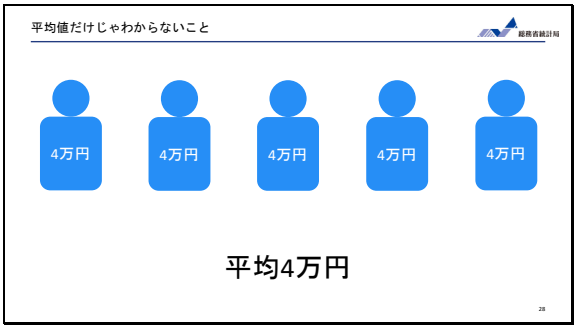
データのまとめ方は大きく2つになります。
年齢・来店回数・金額といった量的なもので表現されるデータは数値型となります。
性別・居住地域といった質的なものは分類型となります。
ちなみに郵便番号は、数字で表現されますが数字の大小に意味はなく、地域に割り当てられている番号なので分類型となります。



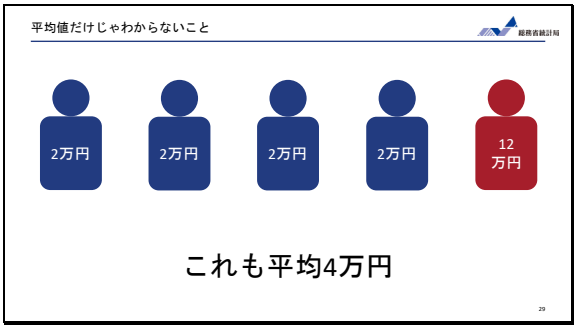
数値のまとめ方で、それぞれ異なる購買金額という数値型についてまとめると。



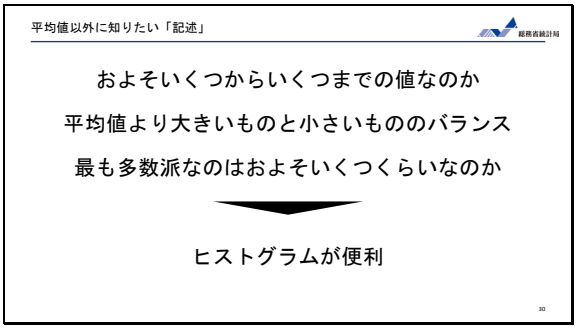
すでに触れたように平均値という考え方を利用するのが一般的となります。
ただし平均値だけで全てまとめられるかというと少し難しいことがあります。



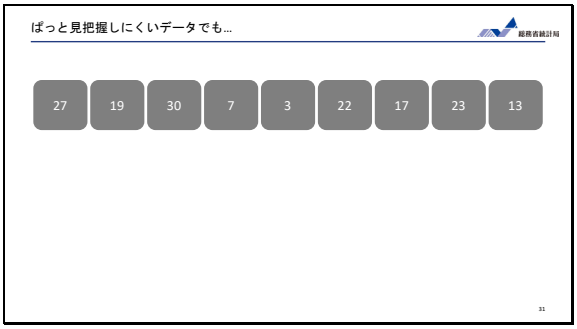
もし仮に全員の購買金額が4万円だった場合平均値は4万円となります。



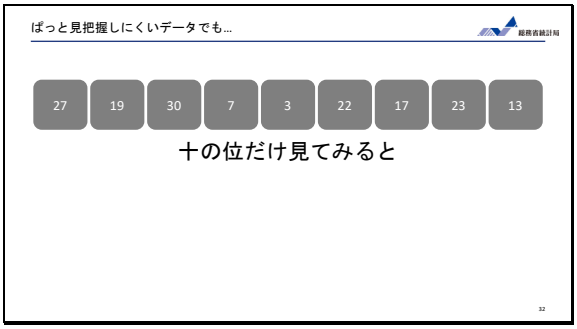
また、仮に2万円が4人、12万円が1人いた場合も平均値は4万円となります。
先ほどの全員4万円のケースと、平均値だけ見ると違いが判らなくなります。



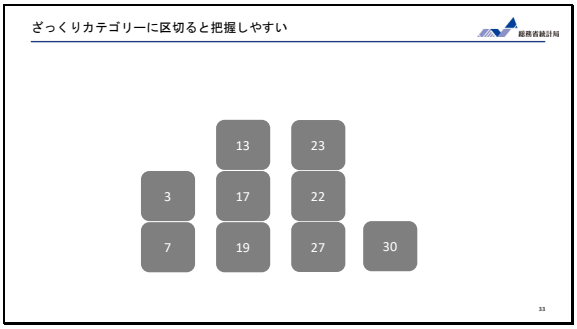
そういった場合どうまとめればよいかというと、ヒストグラムという表を用います。



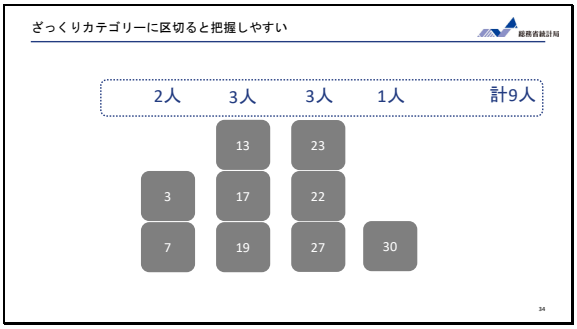
9名分のデータが仮にあるとします。中の数字は購買金額です。そのまままとめようとすると大変ですので



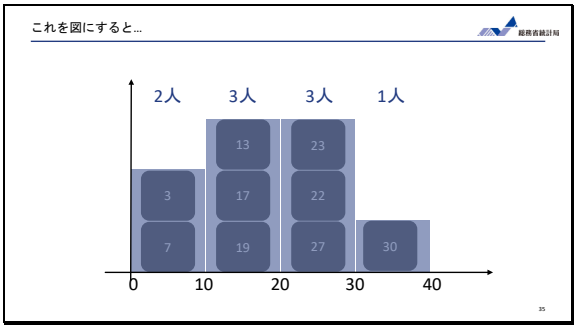
まずは10の位に注目します。



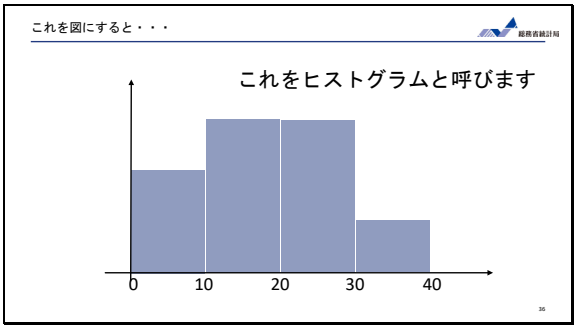
10万円未満、10万円台、20万円台、30万円台のグループに分けることができます。



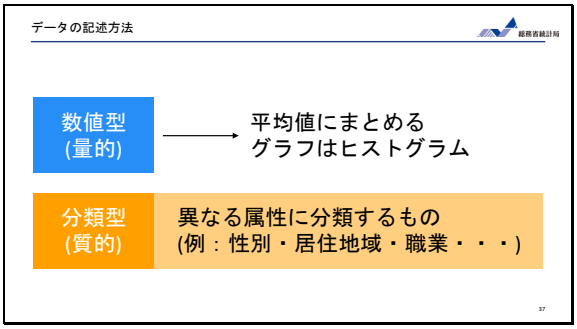
それぞれの人数を集計します。2人、3人、3人、1人合計9人となります。



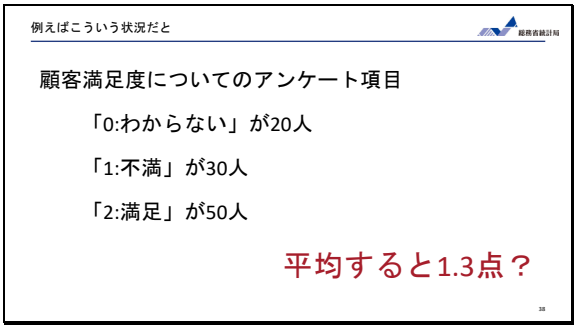
そしてまとめた情報をグラフで表現します。
0-10、10-20、20-30、30-40にそれぞれの人数を含めます。



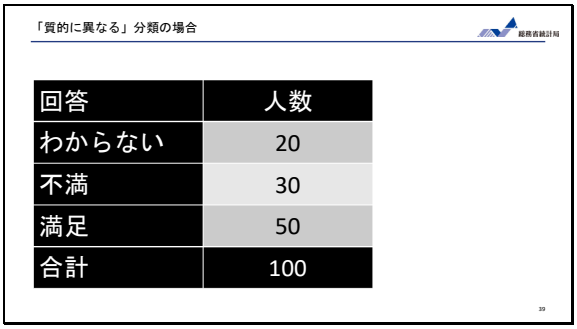
それぞれのカテゴリに何人含まれているかをまとめたものをヒストグラムといいます。
平均値でのまとめに比べてそれぞれのカテゴリにどのぐらいの人数が含まれるのかをグラフィカルに確認することができます。



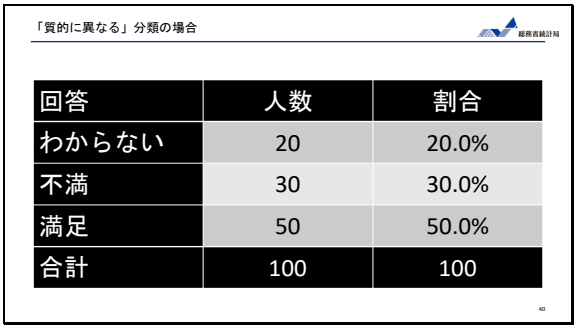
分類型についてもまとめることができます。



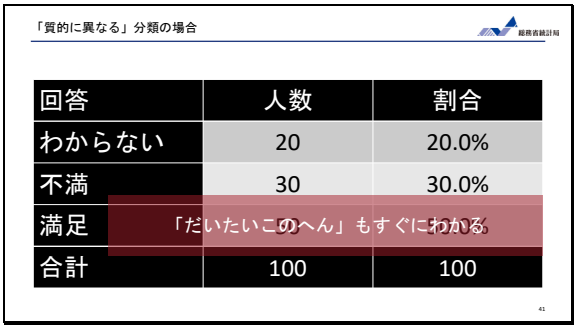
顧客満足度についてアンケートを実施し、結果をまとめる際にアンケート項目の0.1.2をもちいると平均1.3になります。
ただし、この数字には意味はなくそれぞれの回答者の情報が全くわかりません。



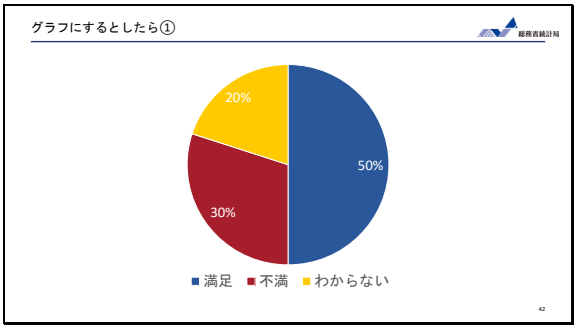
こういった場合は、それぞれのカテゴリの人数を集計します。



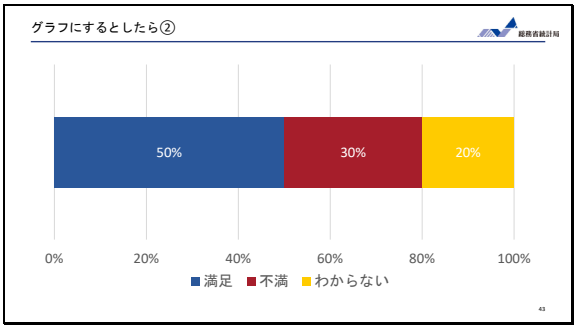
もしくは、全体に占める割合を算出します。



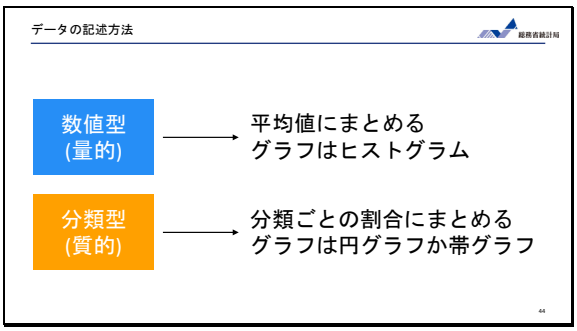
割合を出すことで、人数が増えた際にも結局どこが最も多くの人数が含まれるのか一目で確認することができます。



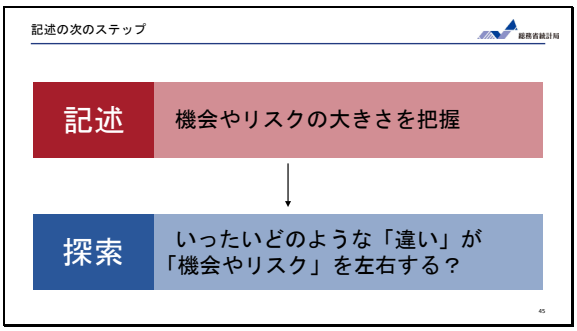
グラフとしてまとめることもできます。
一例として円グラフだとこのようになります。
この場合のメリットはどこで過半数になっているかなどが直感的に理解できます。



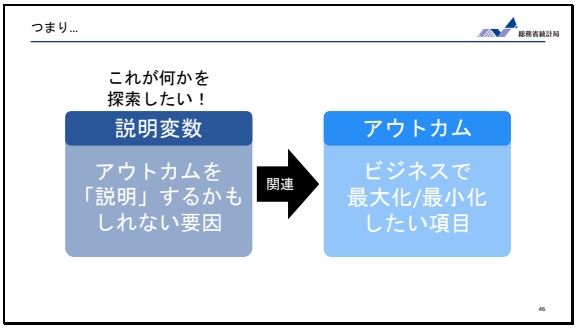
もう一つの例としてグラフとして帯グラフがあります。円グラフに比べると見にくくなりますが、グラフに占める面積で割合を理解することできます。



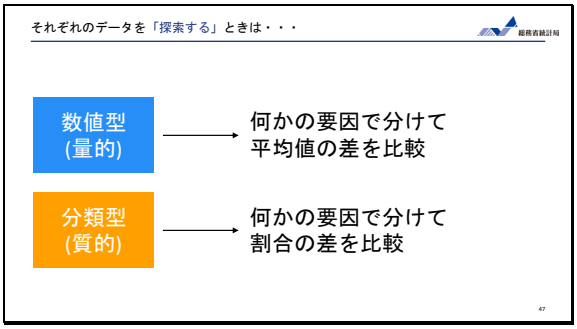
数値型については、平均値にまとめ、ヒストグラムを利用します。
分類型については、割合にまとめ、円グラフか帯グラフを利用します。



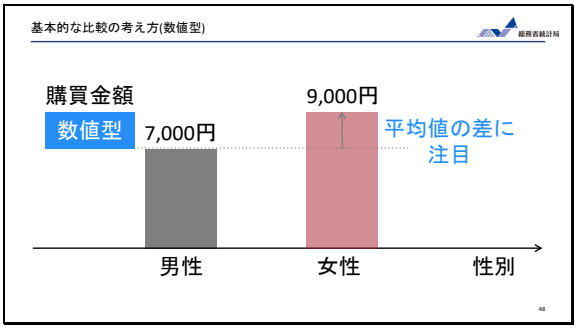
記述の次にやるべきことは、探索となります。探索では、どのような「違い」が「機会やリスク」を左右する原因となるかを明らかにします。
そしてその原因を取り除くことで現状を改善します。



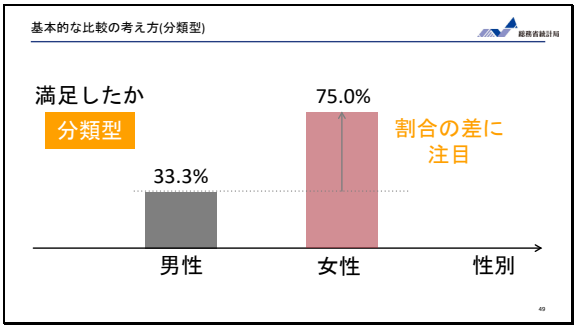
探索を行う際にポイントとなるのは以下2点です。
1）ビジネスで最大化/最小化したい項目をアウトカム
2）原因の候補となるものを説明変数
アウトカムの大きさや小ささを説明するかもしれない要因なので「説明変数」と呼びます。



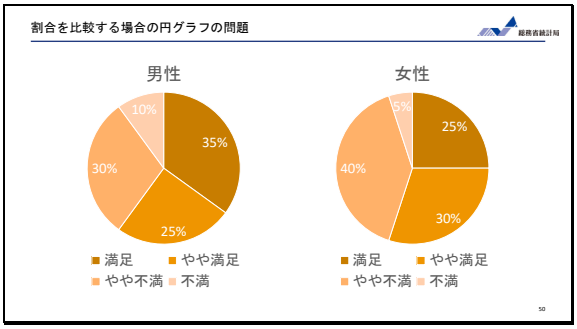
数値型は、何かの要因で分けてそれぞれの平均値の差を比較します。
分類型は、何かの要因で分けて割合の差を比較します。



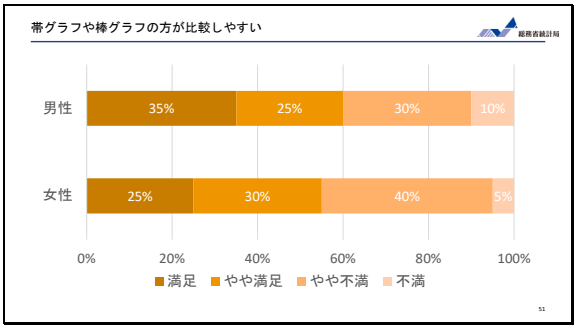
例えば、男女の購買金額それぞれの平均値をグラフ化しその差に着目します。
女性の購買金額平均が高いことがわかるので女性に着目しビジネスを展開した方が良いのがわかります。
つまり性別という要因が探索できたことになります。



同様に満足度で比較すると、同様に女性の満足度が高いのがわかります。
この場合、男性の満足度を上げる活動をするか、満足度が高い女性に集中するかという選択が考えられます。



ちなみにグループ分けをして比較する際に円グラフを用いると見づらいのでお勧めではありません。



その場合帯グラフを用いることで見やすく、値の比較も容易となります。

ピポットグラフを用いてグラフを作成します。

ピポットグラフで平均値を集計するには(この後デモします)

- データ全体を選択してピボットグラフを挿入
- グループ分けしたい項目を「列」に
- 平均値を求めたい項目を「Σ値」に
- 値フィールドの設定を「平均値」に

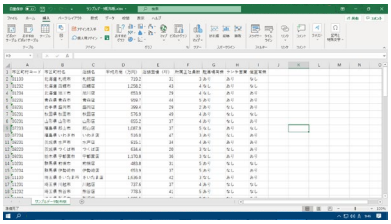
52

デモ①

デモ①-1

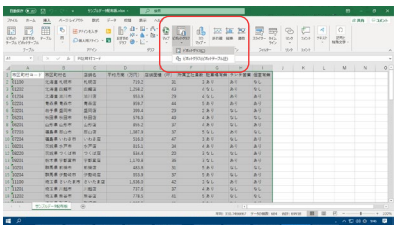
53

デモ①



54

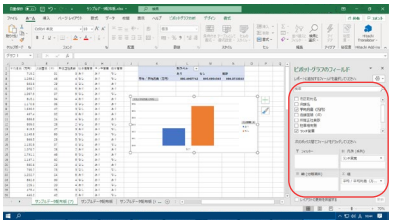
デモ①



②データ全体を選択して、挿入＞ピボットグラフを選択します。

55

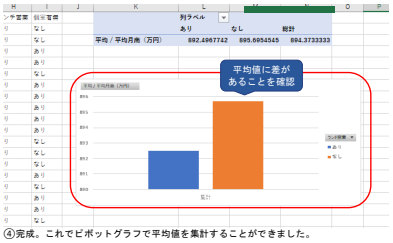
デモ①



③割合を求めたい項目を「Σ値」に、「凡例」に比較したい項目を入れて、値フィールドの設定を「平均値」に設定します。

56

デモ①



④完成。これでピボットグラフで平均値を集計することができました。

57

ピポットグラフで割合を集計します。

ピポットグラフで割合を集計するには(この後デモします)

- データ全体を選択してピポットグラフを挿入
- グループ分けしたい項目を「軸（分類項目）」に
- 割合を求めたい項目を「Σ値」に
- 「凡例」比較したい項目を
- 値フィールドの設定で計算の種類を「行集計に対する比率」に

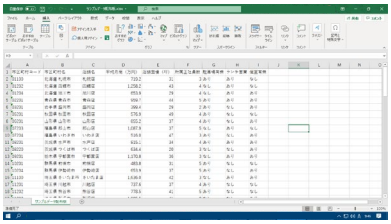
58

デモ①-2

デモ①-2

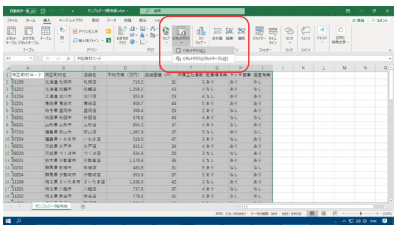
59

デモ①-2



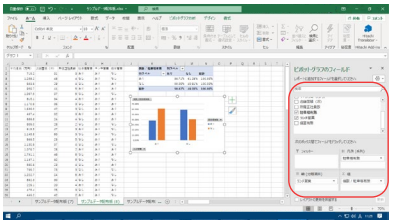
60

デモ①-2



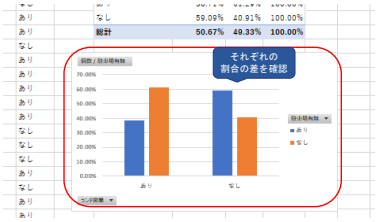
②データ全体を選択して、挿入＞ピボットグラフを選択します。

デモ①-2



③割合を求めたい項目を「Σ値」に、「凡例」に比較したい項目を入れて、値フィールドの設定で計算の種類を「行集計に対する比率」に設定します。

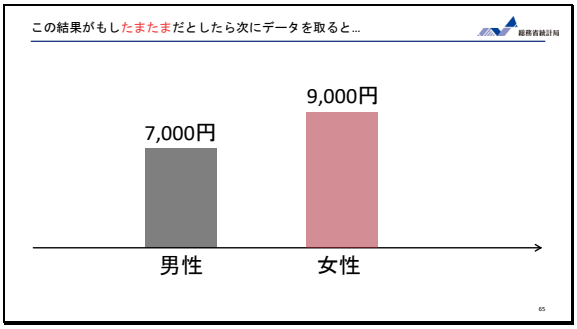
デモ①-2



④完成。これでピボットグラフで割合を集計することができました。

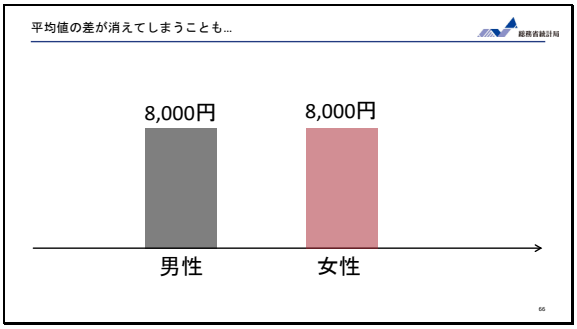


集計をして差が付くのを確認できましたが、よくあるのがその差は「たまたま」ではないかという指摘です。

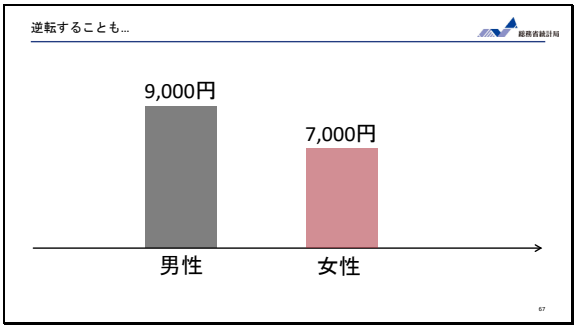


仮にこのようでデータがあったとします。

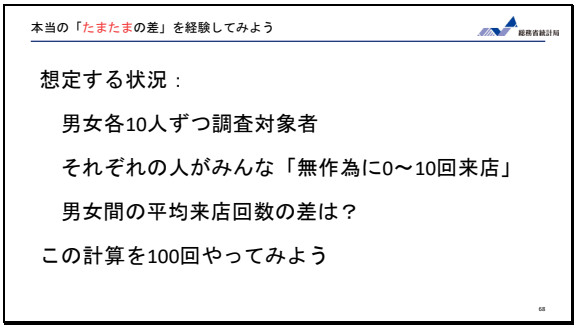
たまたまであったとすると



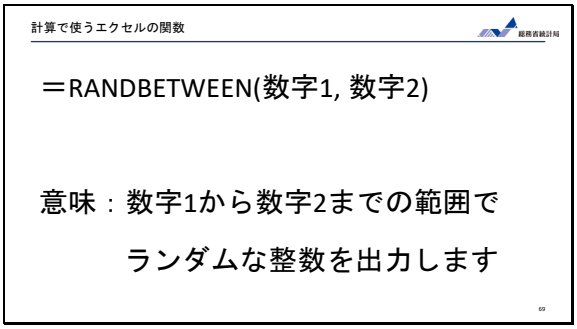
一年後に男性と女性の購入金額の平均が同じになることや



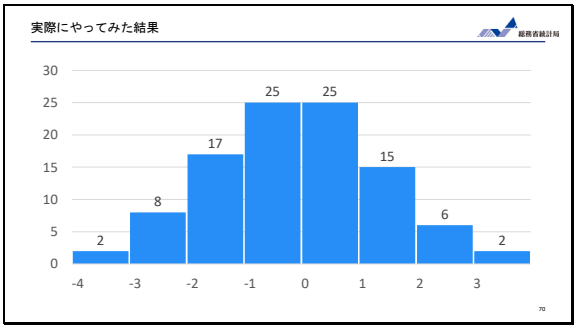
男性が逆転することがあるかもしれません。
このようなことがあるのであればそれは「たまたま」ついた差かもしれません。



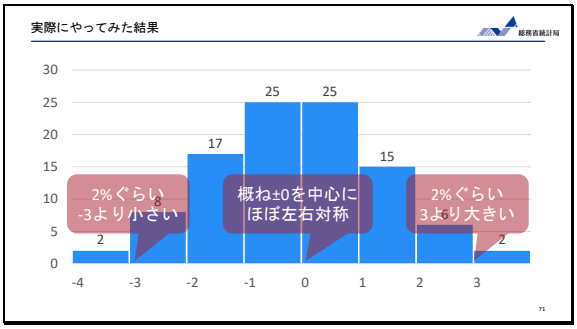
このようなデータを用いて実際に「たまたま」の差を経験します。



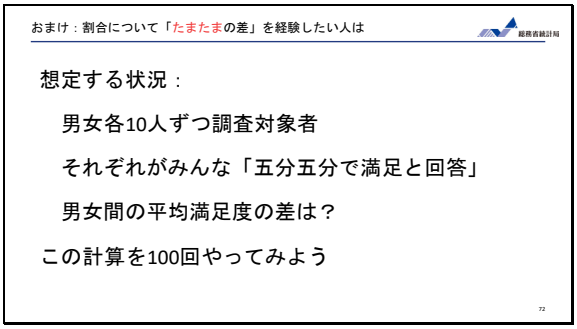
エクセルのRANDBETWEEN関数でランダムな整数を生成して計算します。



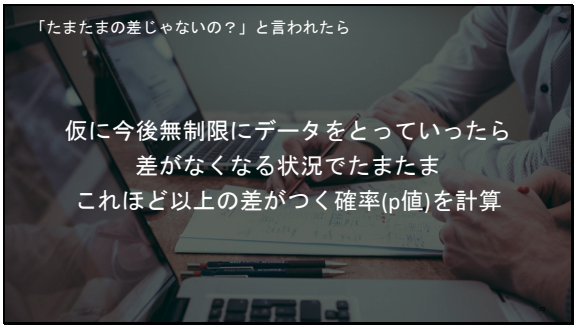
実際に手元で実行した結果です。
100回実行したところ男女差は概ね±1の範囲に収まりますが、たまに±3を超える場合があります。



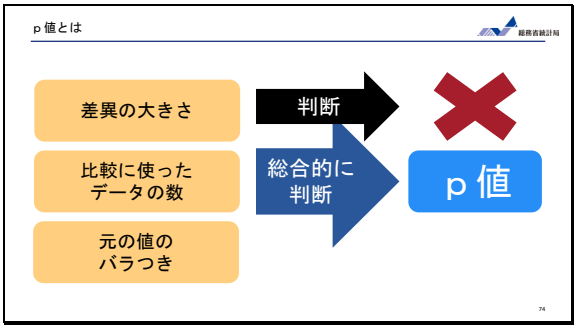
表を見ると、概ね±0を中心に左右対称になりますが、ごくまれに±3になる確率がそれぞれ100回中2回ほど発生します。



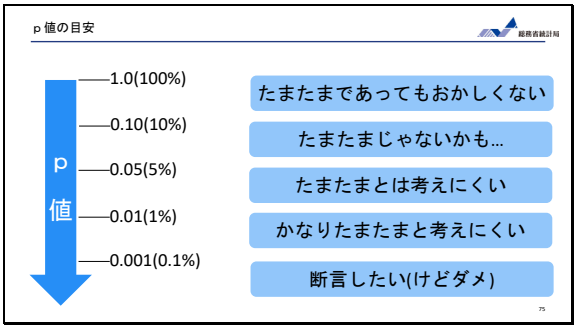
もし割合について同様にたまたまの差を経験したい場合は、
1=満足、0=満足でない
このような内容で乱数を利用して、男女間の満足度の割合の差を算出してください。



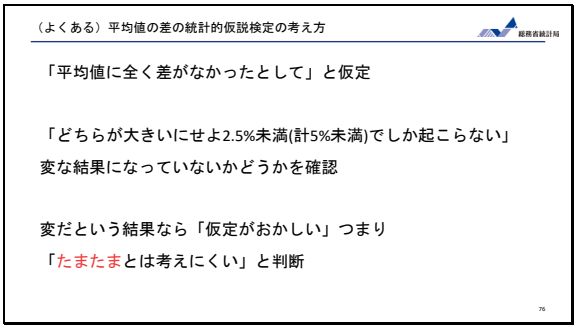
仮に今後無制限にデータをとっていったら差がなくなる状況でたまたまこれほど以上の差がつく確率(p値)を計算し、それがもし小さい場合「たまたま」と考えにくいと考えます。



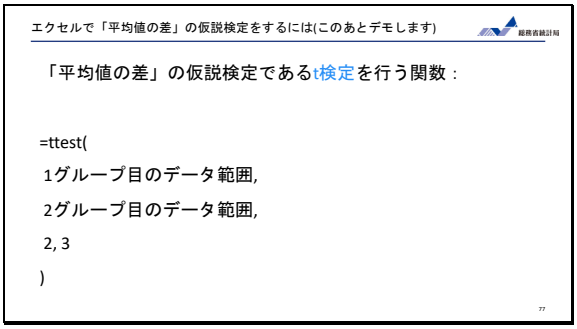
p値とは、平均や割合の差だけでなく、どれくらいのデータを使ったか、元々どの程度データがばらついていたかを総合的に判断して計算します。



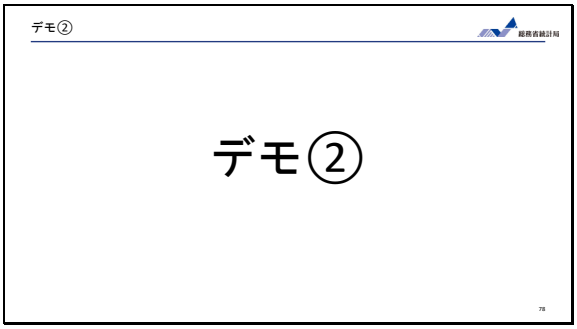
p値が、小さければ小さいほどたまたまと言い難くなります。
10%程度だともしかするとたまたまである可能性があるので判断を保留します。
５％以下になると20回に1回の確率となるので、さすがにたまたまと考えにくいと判断します。1000回に1回になるとたまたまでないと言いたいところですが、ばらつきの可能性があるため0.1％未満と表現します。



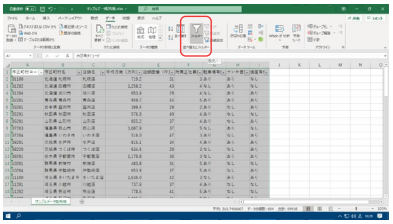
平均値の差の統計的仮説検定の考え方となります。



「平均値の差」をエクセルで計算するためにはttest関数を用いたt検定を行います。

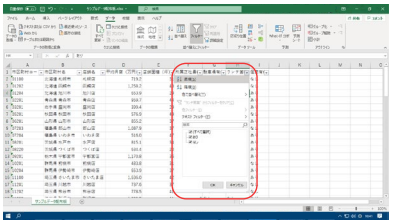


デモ②



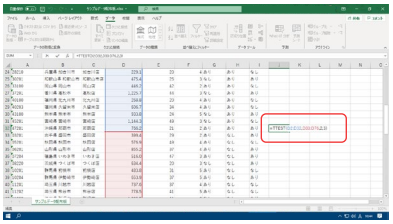
①データ全体を選択して、データ＞フィルタを押下します。

デモ②



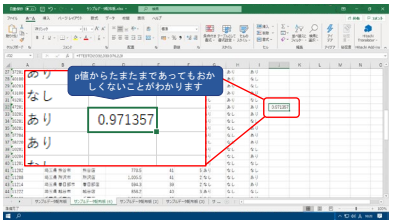
②「ランチ営業」列のフィルタメニュー▼から「昇順」を選択すると、あり→なしの順に並びが変わります。

デモ②



③ 任意のセルに、=ttest(1グループ目のデータ範囲, 2グループ目のデータ範囲, 2, 3)と入力します。

デモ②



④結果が表示されます。これで「平均値の差」の仮説検定をすることができました。

補足1:データ範囲の指定方法

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

補足1:データ範囲の指定方法

一個のセルだけの範囲なら
直接セル名を指定

A4

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

補足1:データ範囲の指定方法

「開始セル：終了セル」で範囲指定

A2:A4

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

85

補足1:データ範囲の指定方法

同じ行や列でなくても
長方形に指定ができる

A2:C4

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

86

補足1:データ範囲の指定方法

開始位置と終了位置の順番は不問
(お作法としては「左上：右下」)

C5:B2

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

87

補足1:データ範囲の指定方法

A1	B1	C1
A2	B2	C2
A3	B3	C3
A4	B4	C4
A5	B5	C5

コンマで区切れば
複雑な範囲指定もできる(非推奨)

B1,A3:B4,C2:C5

88

補足2:片側か両側か

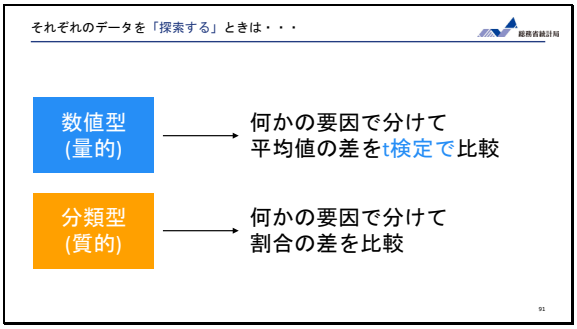
片側	「男性側が多いか」など 一方向の違いのみに興味
両側	「両者に差があるか」 両方向の違いに興味

89

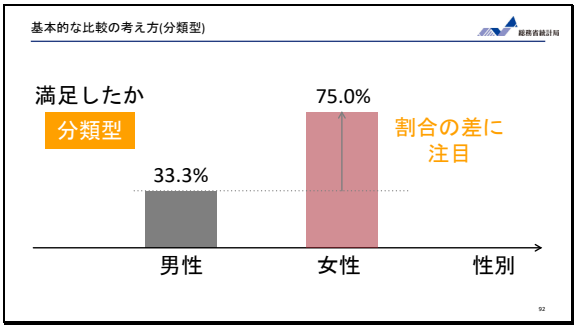
補足3: t検定の細かい部分

対	「同じ人の二時点の値」など データ間に対応がある場合
等分散	グループごとのバラつき方が 同じであると仮定
非等分散	グループごとのバラつき方が 同じであると仮定しない

90



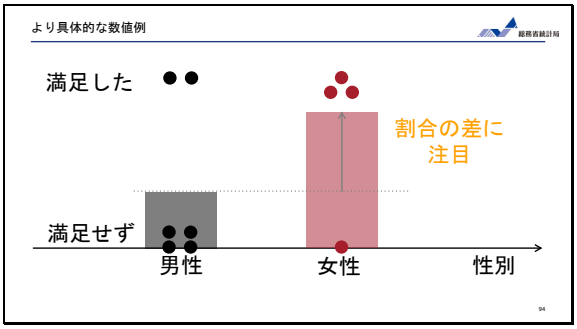
数値型でデータを探索する際は、何かの要因で分けて平均値の差をt検定で比較します。
また、分類型でデータを探索する際は、何かの要因で分けて割合の差を比較しますが、数値型と同様にたまたまの差と指摘されないように進めます。



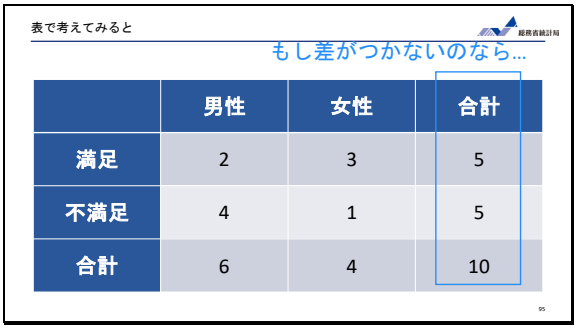
先ほどと同じように男女で分けられたグラフの差が意味があるのか考えます。



たまたまと指摘された際の反論を考えます。

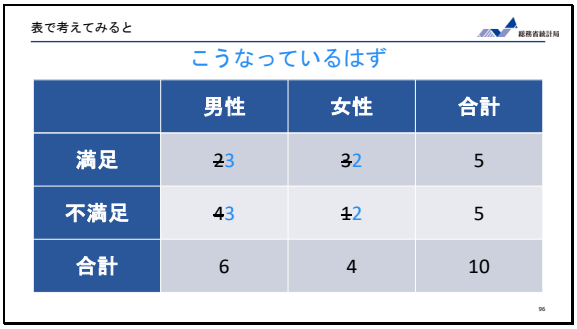


割合の差をグラフに表しましたが、まず手始めに男性、女性それぞれ「満足した」「満足せず」が何人含まれているかを考えます。
本グラフでは、男性4人が「満足せず」、2名が「満足した」と読み取れます。



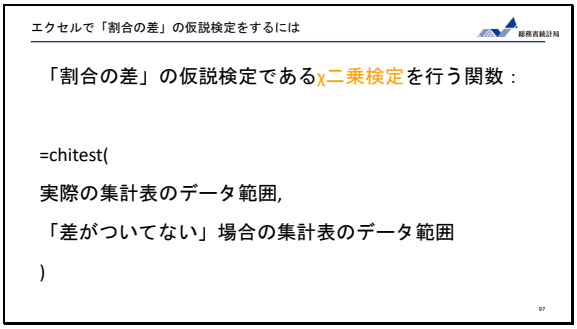
	男性	女性	合計
満足	2	3	5
不満足	4	1	5
合計	6	4	10

改めて集計表で考えてみます。満足、不満足について男女を合計するとそれぞれ5名となります。つまり満足50%、不満足50%になるので、もし男女間の性別で差がつかないとすると男女の値は同じになるはずです。

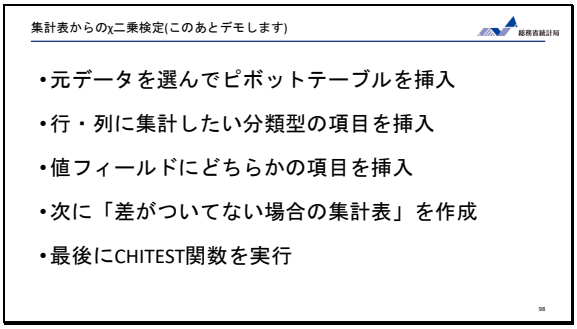


	男性	女性	合計
満足	~~2~~3	~~3~~2	5
不満足	~~4~~3	~~1~~2	5
合計	6	4	10

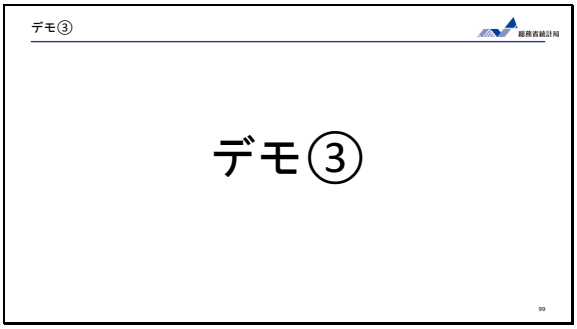
つまり性別の差が無いとすると男性の6名はそれぞれ3名ずつ、女性は２名づつに分かれるはずです。しかし実際のデータでは、男性の満足は1名少なく、女性の満足は1名多くなっています。



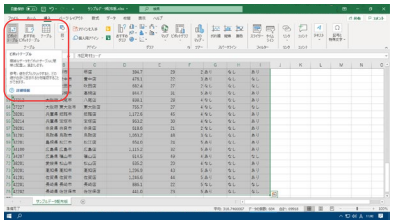
先ほどの集計表を元に、「割合の差」の仮説検定をするにはchitest関数を用いてχ二乗検定を行います。



エクセルで集計表からのχ二乗検定を行います。



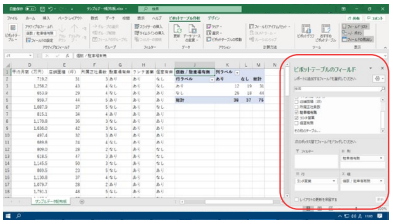
デモ③



①データをすべて選択し、挿入＞ピボットテーブルを選択します。

100

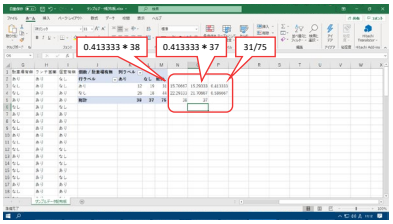
デモ③



②ピボットテーブルのフィールドで、行・列に集計したい分類型の項目を、また挿入値フィールドにどちらかの項目を設定します。

101

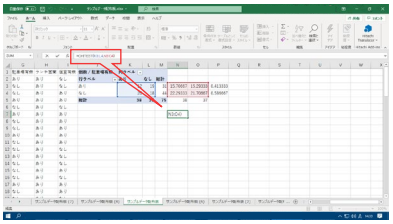
デモ③



③「差がついてない場合の集計表」を作成します。

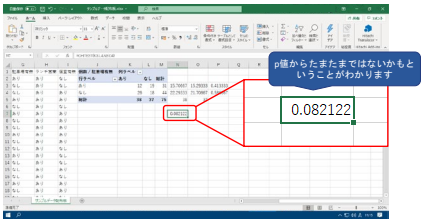
102

デモ③



④任意の場所に　=chitest(実際の集計表のデータ範囲,「差がついてない」場合の集計表のデータ範囲)を入力します。

デモ③

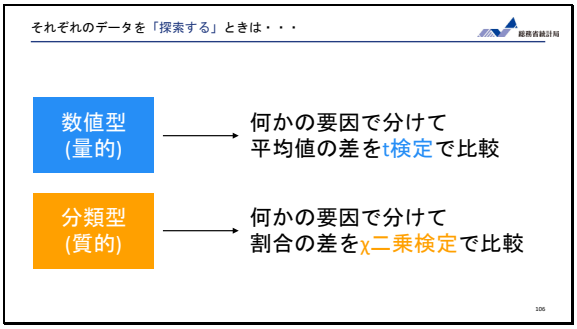


④結果が表示されました。これで「割合の差」の仮説検定ができました。

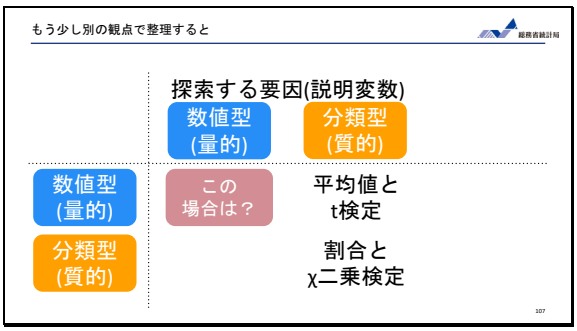
4)数値データ同士の関係性の見方（単回帰分析）

次に「グループ分けする」という質的な要因を探索するだけではなく、数値と数値の関係性をみるための単回帰分析の方法を学びます。

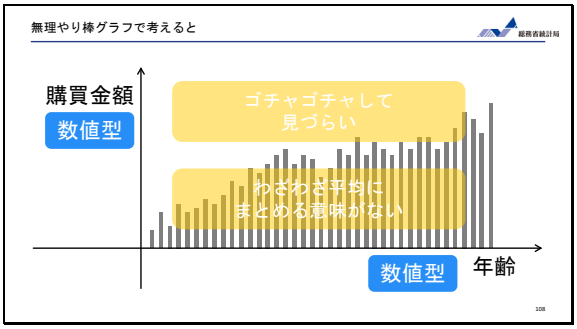
まずエクセルのグラフツールで散布図を描き、その中の回帰直線および数式を確認する方法を示すほか、ここでもやはり「たまたまのトレンド」と言えるのかどうか、分析ツールメニューから得られるp値の確認方法を示します。



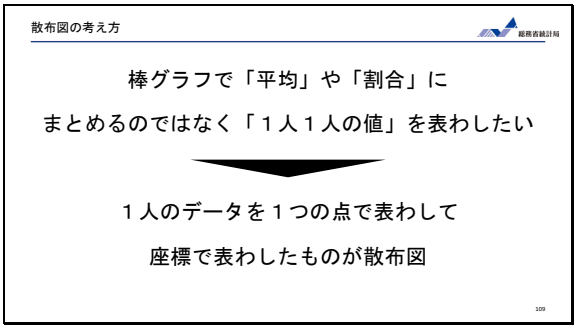
数値型のデータは何かの要因で分けて平均値の差をt検定で比較します。
分類型のデータは何かの要因で分けて割合の差をχ二乗検定で比較します。



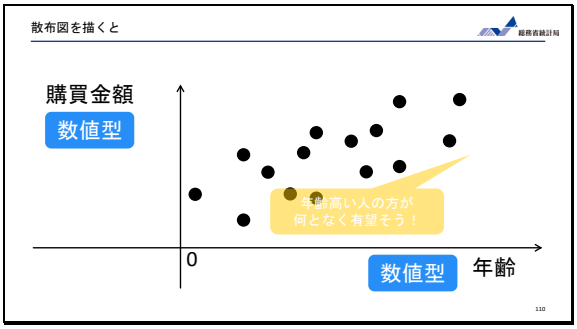
ここまでは、分類型データをグループ分け（駐車場やランチの有り無しなど）したものに対する平均値、割合の差を見てきましたが、ここからは数値型データを探索する場合について考えます。



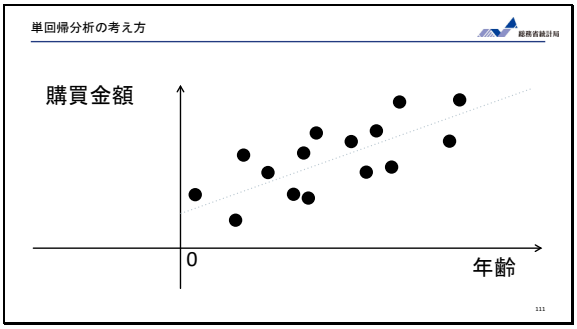
仮に横軸に年齢、縦軸に購買金額といった棒グラフで表現すると煩雑で見にくくなります。また平均値でまとめる意味がありません。



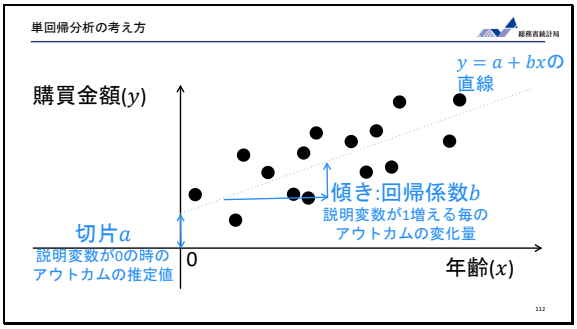
１人のデータを１つの点で表わして座標で表わしたものが散布図です。
散布図を用いて考えます。



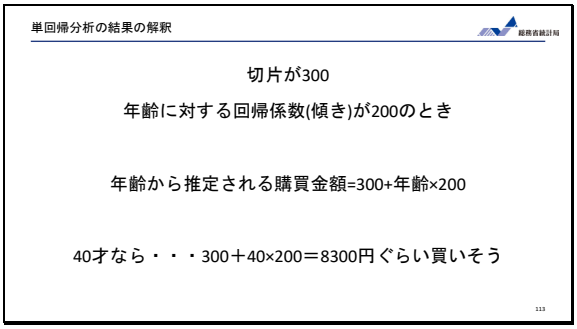
何歳の人がいくら買っているかというのを点（一人ずつ）で確認できます。このグラフでは年齢が高い人が買っているというトレンドが見えます。



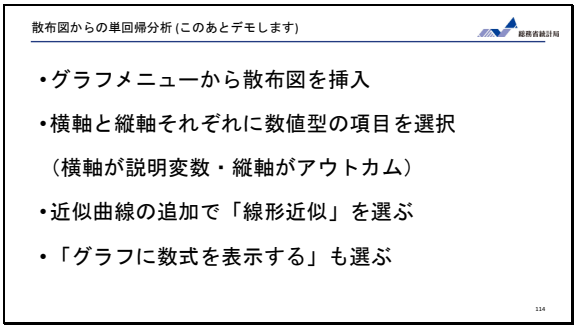
こういったトレンドが見えると値の真ん中に通る線を引くことができます。



この直線を回帰直線といいます。
直線はy=a+bxの式に当てはめることができます。

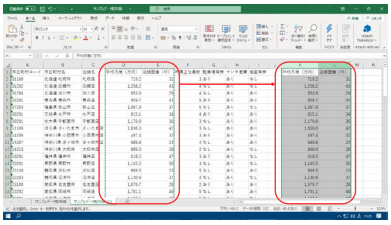


先ほどのようなグラフに直線を引いて、あてはまりが良いものを考えることを単回帰分析といいます。
例えば、購買金額=切片(300)+年齢×回帰係数(200)の場合では、もし顧客の年齢が40才なら・・・300＋40×200＝8300円買いそうということになります。

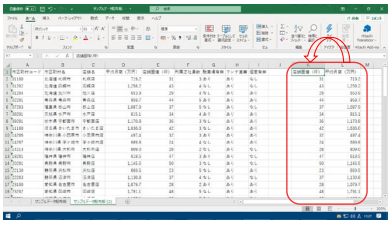


エクセルで散布図を用いて単回帰分析を行います。

デモ④

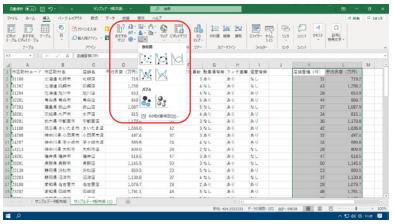


①表中から「平均月商」「店舗面積」をコピーして抜き出します。



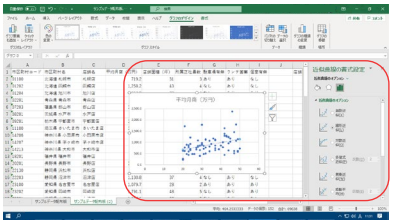
②左側に「店舗面積」がくるように並び替えます

デモ④



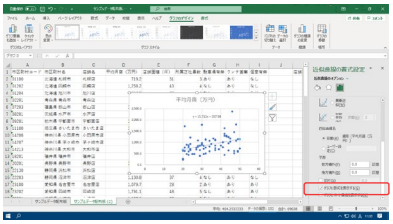
③「店舗面積」列と「平均月商」列を選択し、挿入＞散布図を選択すると散布図が表示されます。

デモ④

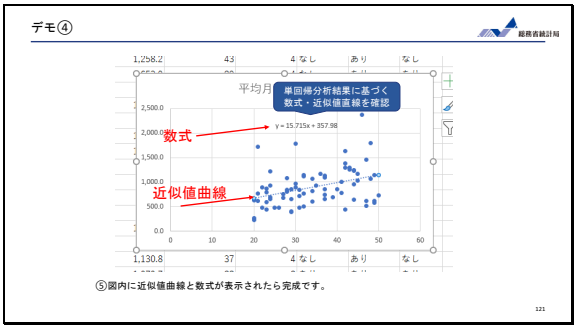


④散布図内で、右クリック＞近似値曲線を表示　を選択し、書式設定で「線形近似」を選択すると図内に近似値曲線が表示されます。

デモ④

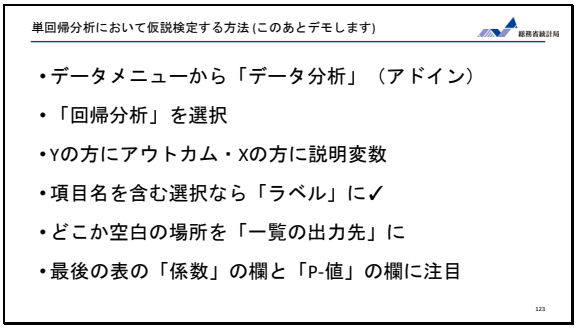


④書式設定下部にある「グラフに数式を表示する」にチェックを入れます。



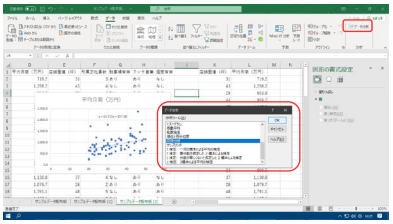


もちろんこの場合も、そのトレンド「たまたま」じゃないという指摘があります。

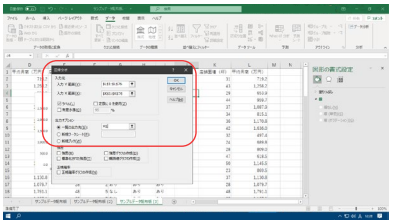


先ほど、t検定、χ二乗検定でp値を計算しましたが、同様にエクセルを用いて単回帰分析において仮説検定をすることが可能です。

デモ⑤

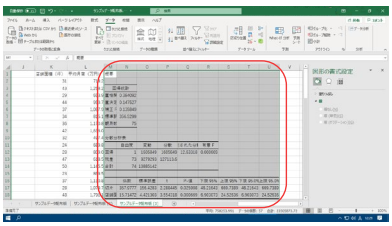


①データ＞データ分析＞回帰分析を選択します。
（データ分析メニューはアドインから追加できます）



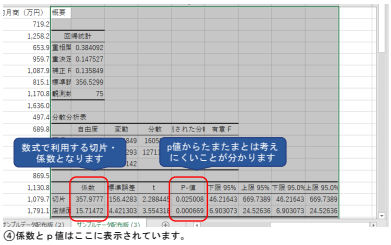
②Yには「平均月商」列、Xには「店舗面積」列を指定します。
選択した「平均月商」「店舗面積」は一行目がラベルなので「ラベル」にチェックを入れます。出力オプションは任意の位置を指定します。

デモ⑤



③回帰分析の結果が表示されます。

デモ⑤

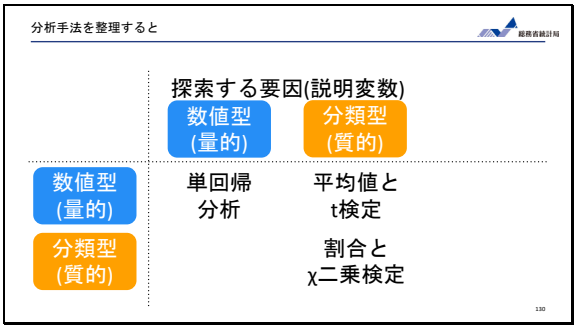


④係数とｐ値はここに表示されています。

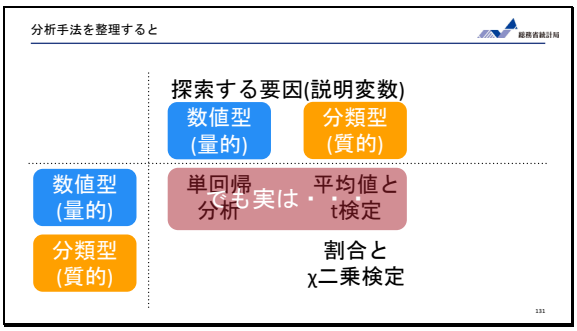
5)複数の説明変数を同時に使う方法（重回帰分析）

ここではまず、「ダミー変数」という考え方を通して単回帰分析とt検定が本質的には同じようなことをやっているということを示した上で、単回帰分析の限界と、それを重回帰分析で解決できることを説明します。

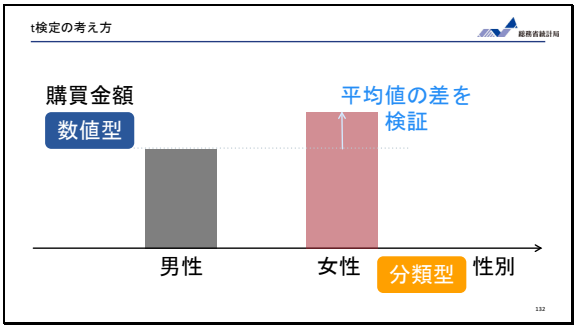
次に先程と同じ分析ツールメニューを使って重回帰分析を実行し、その結果の解釈方法について学びます。



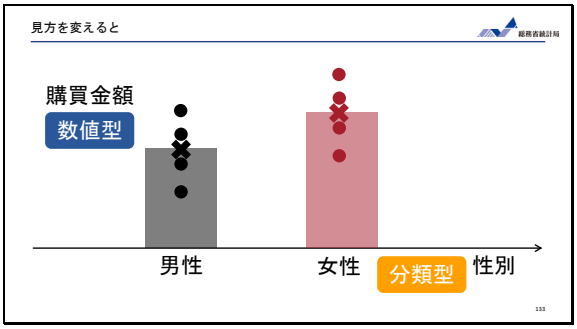
探索する要因が数値型であった場合、単回帰分析を実行することで数値型と数値型の関係を説明することができます。またそのトレンドが偶然であるかということを確認することもできます。



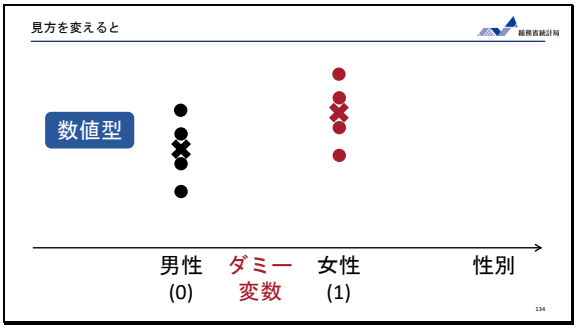
でも実は単回帰分析とt検定は内部では同じようなことを行っています。



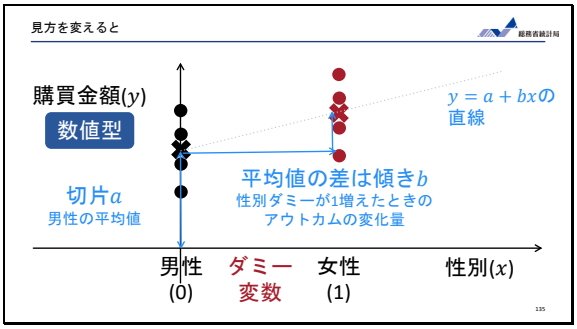
これまで男性、女性で分けて平均値の差を見ていました。



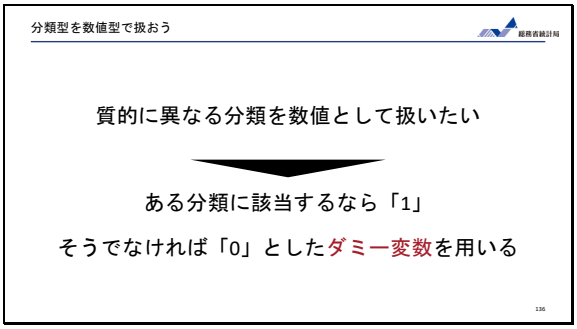
グラフの元データを点に置き換えてみると、平均値を中心にそれぞれのデータが散らばっているのがわかります。



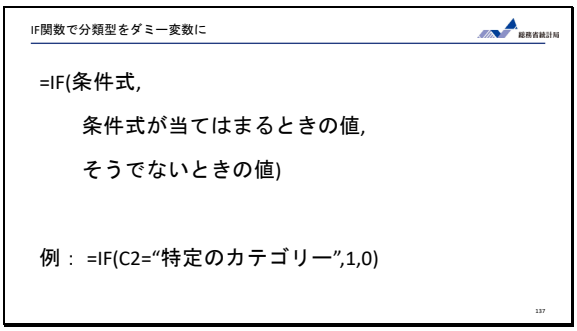
棒グラフを消して見方を変えます。



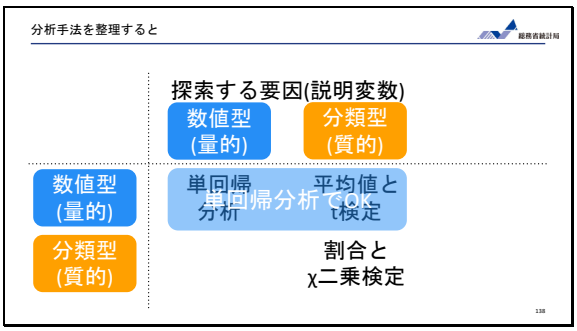
実は単回帰分析は平均値を通るように直線を引いているのがわかります。この場合傾きが何を示しているかというと男女間の平均値の差を示しています。また切片は男性の平均値となります。



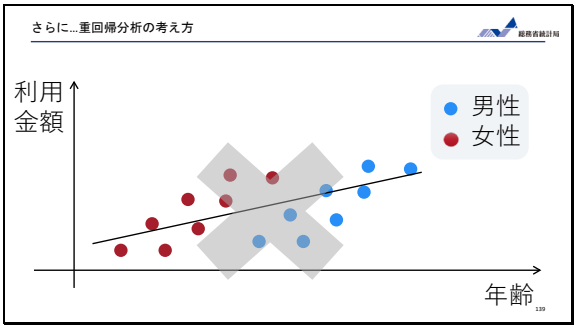
例えば、男性、女性や駐車場の有無のような質的に異なる分類を数値として扱う場合は、ダミー変数を用います。



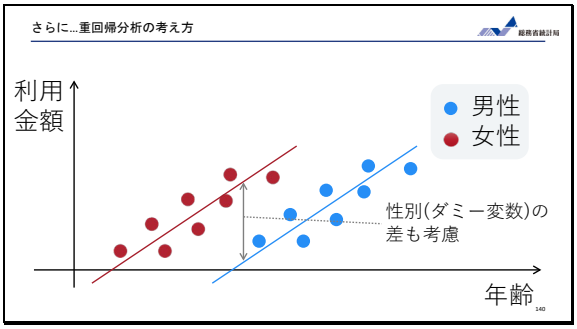
ダミー変数を使うと平均値（t検定）の差も回帰分析で行えます。エクセルのIF関数を用いることでダミー変数を利用できます。



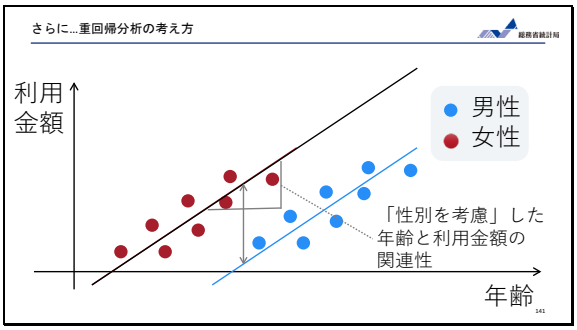
アウトカムが数値型の分析は、説明変数が数値型、分類型どちらであっても単回帰分析を用いることできます。



単回帰分析は一つのアウトカムと説明変数の関係について分析する手法でしたので、このような複数（年齢・性別）の説明変数についての分析することはできません。
男性、女性それぞれで年齢と利用金額に対する傾きが違うため線の傾きが一致しないことがお分かりいただけると思います。



変数が増えてくるとそれぞれの変数を色分けするのが難しくなります。そこで重回帰分析では平行線が引けるという一つ仮定をおきます。
平行線が引けるということはどういうことかというと男性、女性それぞれで1歳増えるごとに購買金額の上がり方が男女ともに一定だった仮定とします。そうすることで平行線が引けるようになります。それにより性別の差が考慮できるようになります。
この男女間の差が回帰係数となります。



男女間で平行線が引けたことで、もし男性が女性だった場合について利用金額を計算することが可能になります。
このように複数説明変数がある場合、それらをまとめて使って分析を行うことができるのが重回帰分析となります。

重回帰分析のやり方(このあとデモします)

- 予め用いたい説明変数列を横並びに整形
- データメニューから「データ分析」（アドイン）
- 「回帰分析」を選択
- Yの方にアウトカム・Xの方に説明変数の範囲
- 項目名を含む選択なら「ラベル」に✓
- どこか空白の場所を「一覧の出力先」に
- 最後の表の「係数」の欄と「P-値」の欄に注目

142

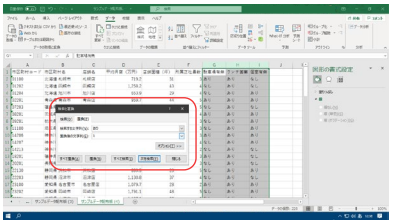
エクセルで重回帰分析を行います。

デモ⑥

デモ⑥

143

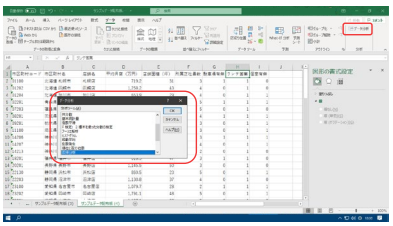
デモ⑥



①検索と置換機能を使用して、「駐車場有無」列、「ランチ営業」列、「個室有無」列をダミー変数（あり＝１／なし＝０）に置き換えます。（検索と置換機能を使用）

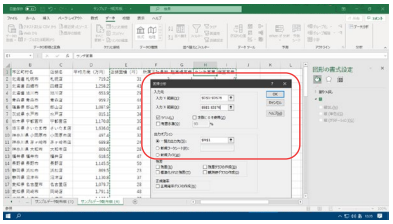
144

デモ⑥



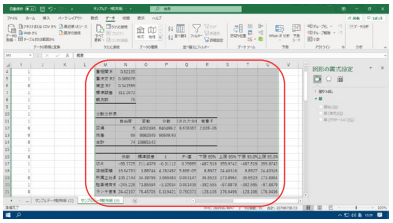
②データ＞データ分析＞回帰分析　を選択します。

デモ⑥



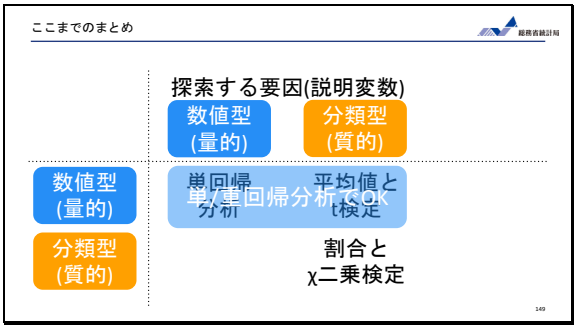
③Yには「平均月商」列、Xには「店舗面積」列、「所属正社員数」列、「駐車場有無」列、「ランチ営業」列「個室有無」列を指定します。出力オプションは任意の位置を指定します。

デモ⑥

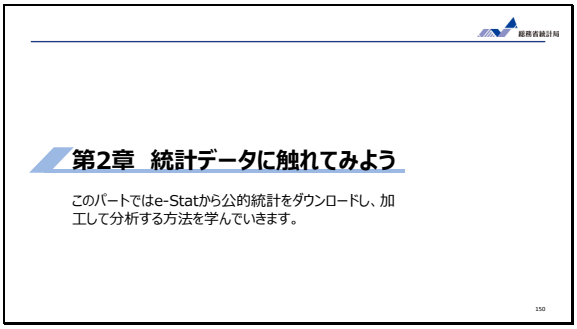


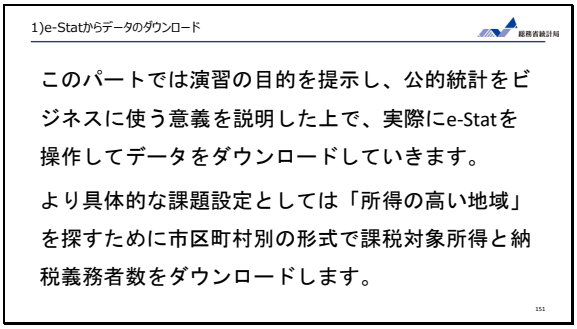
④回帰分析の結果が表示されます。

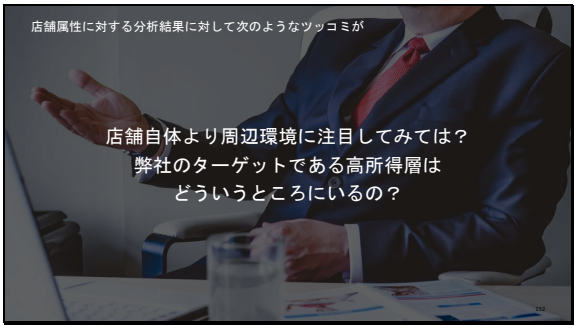




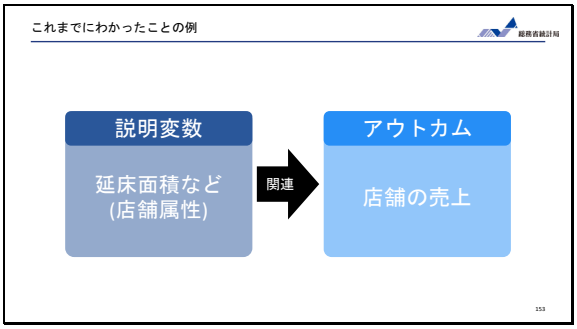
アウトカムが数値型の場合は、単/重回帰分析を使うことで説明変数が、数値型、分類型に関わらず分析を行うことができます。



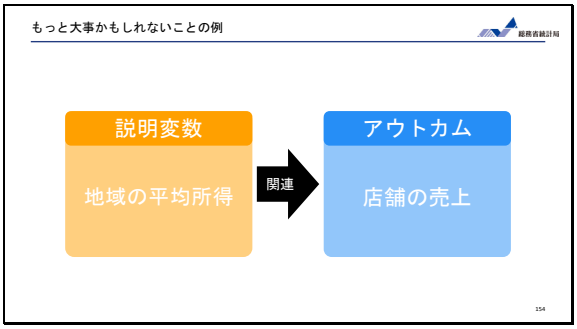




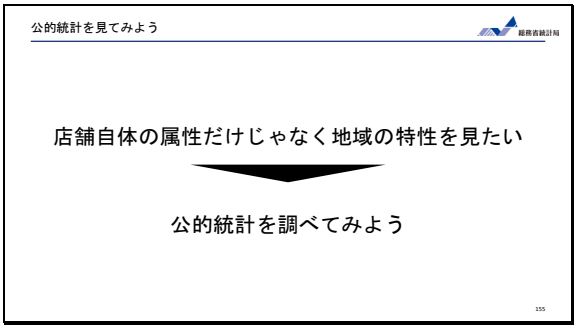
店舗データを分析することで、どういった店舗が売上が高いのかはわかりました。
ただ店舗データ以外で何かあるのではないかという指摘を受けることが有ります。
例えば今回のシナリオである焼肉店の場合、自社のターゲットはどこに住んでいるのかなどが考えられます。



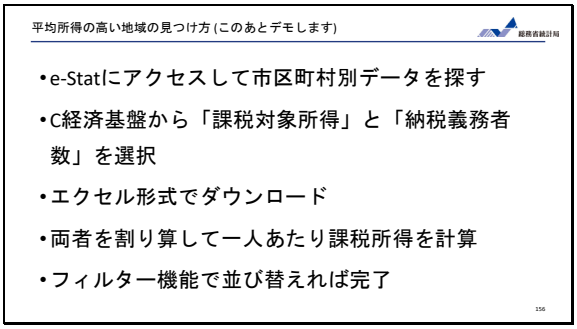
つまりアウトカムとして、店舗の売上を考えて分析を実行した結果、延床面積などが関係していることがわかったとします。



ところがそれ以外にももしかすると関係のあることが有るかもしれません。例えば地域の平均所得などが考えられます。



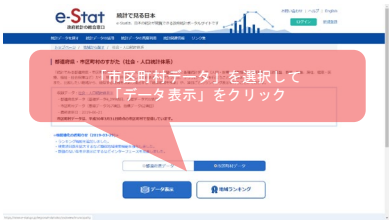
自社が保有している以外のデータを分析に利用したい場合、公的統計を活用します。

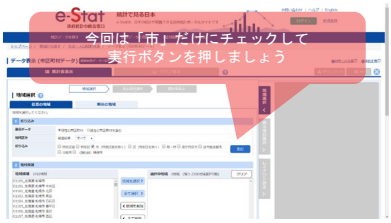


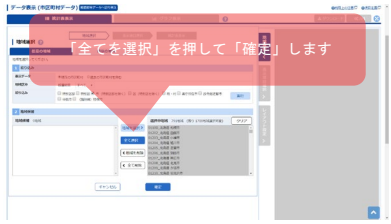
e-Statを利用してデモを行います。

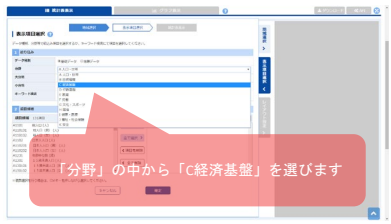
デモ⑦


https://www.e-stat.go.jp/にアクセスして
「地域」のボタンをクリック

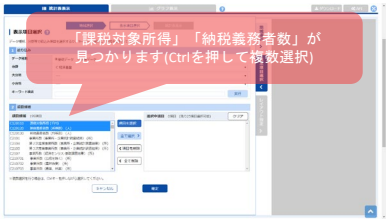

「市区町村データ」を選択して
「データ表示」をクリック


今回は「市」だけにチェックして
実行ボタンを押しましょう

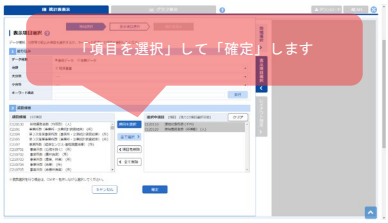

「全てを選択」を押して「確定」します


「分野」の中から「C経済基盤」を選びます

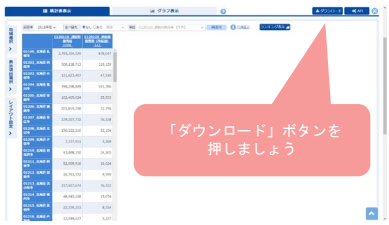
e-Statでデータを探そう



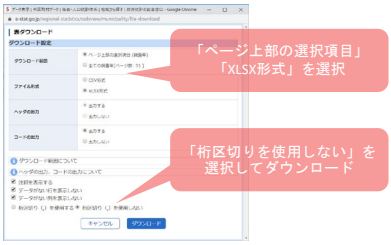
e-Statでデータを探そう



e-Statからダウンロードしよう

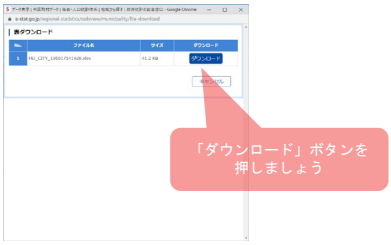


e-Statからダウンロードしよう



166

e-Statからダウンロードしよう



167

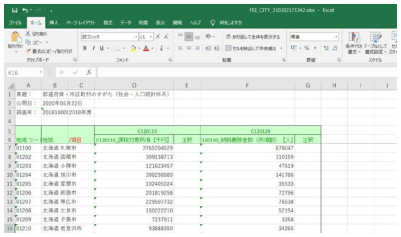
2)エクセル上でのデータ加工

次にダウンロードしたデータをエクセル上で見やすく加工していきます。

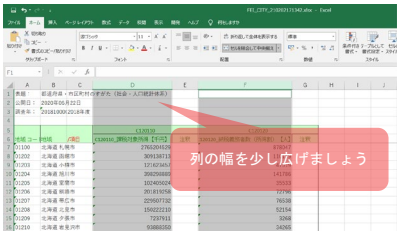
不要な行や列の削除、エクセル上での計算により「納税義務者数あたりの課税対象所得」という割り算を実行し、最後にフィルター機能を適用して降順に並べ、（一人あたりの平均で）所得が高い地域から順に並べていきます。

168

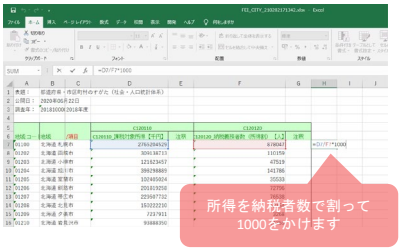
エクセルでデータを触ってみよう



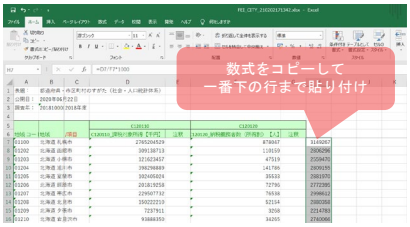
エクセルでデータを触ってみよう



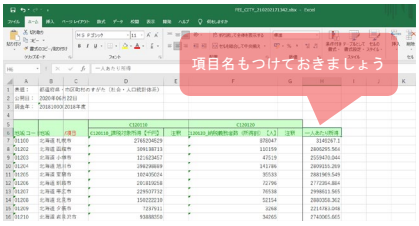
エクセルでデータを触ってみよう



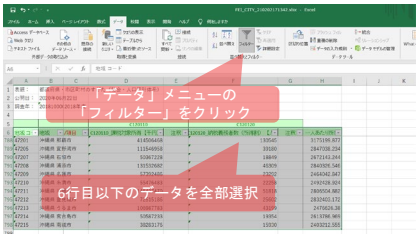
エクセルでデータを触ってみよう



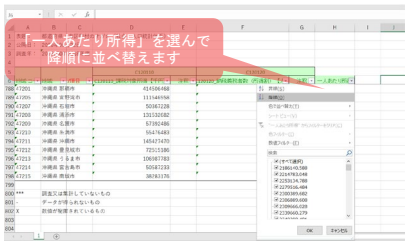
エクセルでデータを触ってみよう



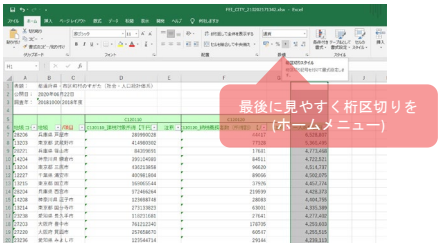
エクセルでデータを触ってみよう



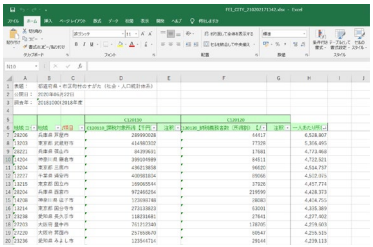
エクセルでデータを触ってみよう

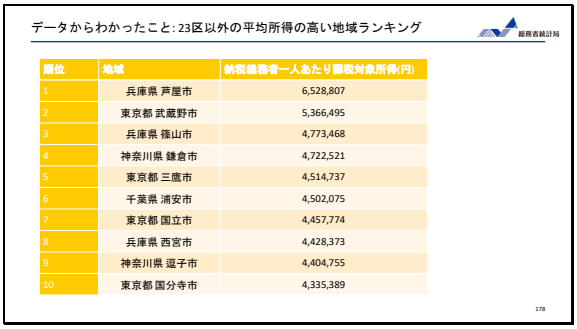


エクセルでデータを触ってみよう

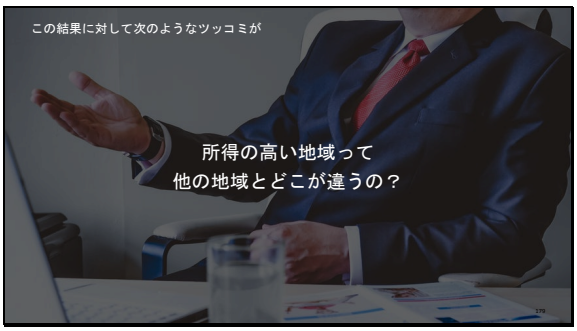


エクセルでデータを触ってみよう

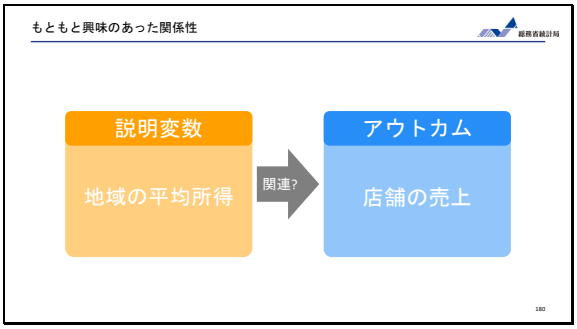




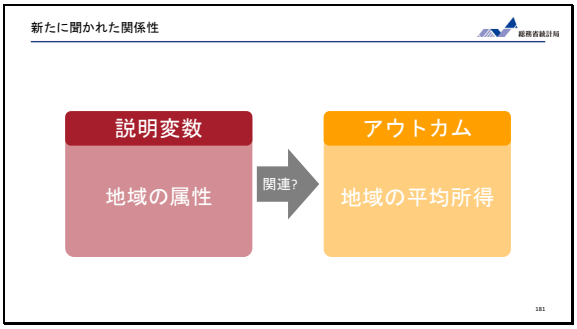
順位	地域	納税義務者一人あたり課税対象所得(円)
1	兵庫県 芦屋市	6,528,807
2	東京都 武蔵野市	5,366,495
3	兵庫県 篠山市	4,773,468
4	神奈川県 鎌倉市	4,722,521
5	東京都 三鷹市	4,514,737
6	千葉県 浦安市	4,502,075
7	東京都 国立市	4,457,774
8	兵庫県 西宮市	4,428,373
9	神奈川県 逗子市	4,404,755
10	東京都 国分寺市	4,335,389



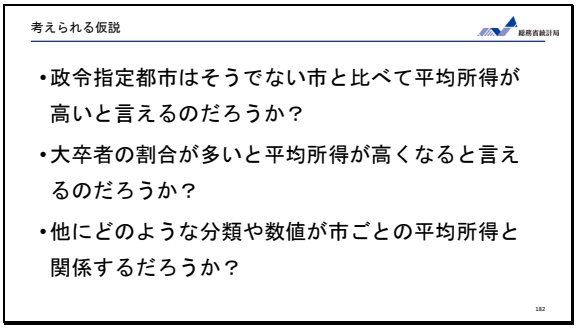
所得の高い地域が公的統計データからわかりましたが、次にこのような指摘を受けることがあります。



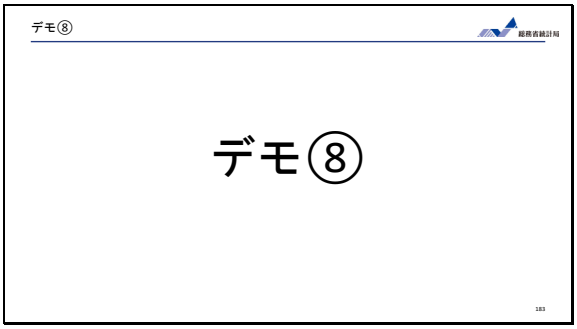
当初は、店舗の売上をアウトカムに分析を行いました。
その結果地域の平均所得が関係していると分析結果が出たとします。



次に当初の分析結果で得られた結果を深堀します。アウトカムを店舗の売上でなく、先ほどの分析結果である地域の平均所得とします。
そうすると地域の平均所得に関係のある地域の属性を分析結果から得ることができます。



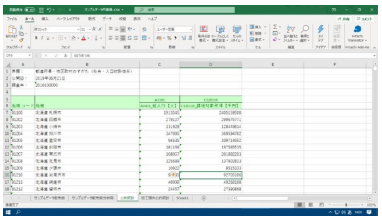
この場合、仮説としてこのようなものが考えられます。
演習で確認していきましょう。



デモ⑧-1

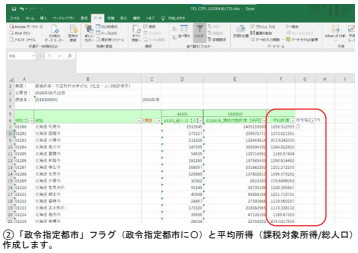
1. 政令指定都市かどうかを自分で入力した上で政令指定都市か否かで平均所得に差があるかどうか考えてみましょう
2. e-Statから市ごとの課税所得、納税義務者数に加えて総人口と大学卒業者数をダウンロードして先ほどの仮説について考えてみましょう

デモ⑧-1



①e-Statからデータ（表示項目A人口・世帯から総人口、C経済基盤から課税対象所得、調査年2010年を選択）をダウンロードします（詳細はP158-167の手順を参考）。

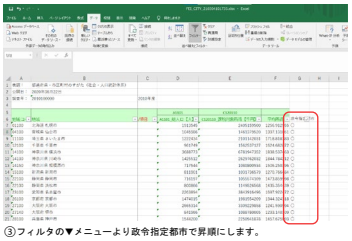
デモ⑧-1



②「政令指定都市」フラグ（政令指定都市に○）と平均所得（課税対象所得/総人口）を作成します。

デモ⑧-1

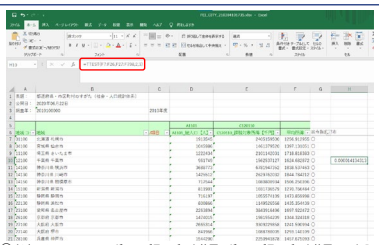
総務省統計局



③フィルタの▼メニューより政令指定都市で昇順にします。

デモ⑧-1

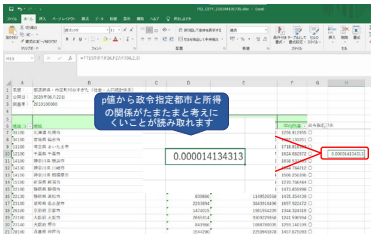
総務省統計局



③ 任意のセルに　=ttest(1グループ目のデータ範囲,2グループ目のデータ範囲, 2, 3)と入力します。

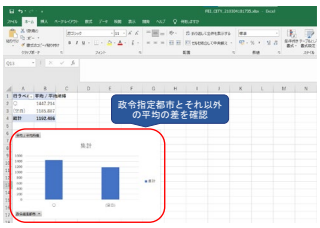
デモ⑧-1

総務省統計局



④「平均値の差」の仮説検定の結果が表示されました。

デモ⑧-1



⑤参考値としてそれぞれの平均値を比較するピボットを作成してもよいでしょう。

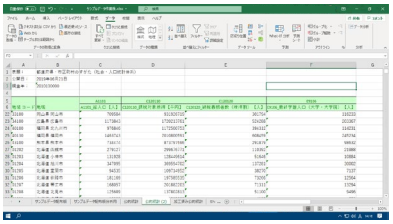
190

デモ⑧-2

1. 政令指定都市かどうかを自分で入力した上で政令指定都市か否かで平均所得に差があるかどうか考えてみましょう
2. e-Statから市ごとの課税所得、納税義務者数に加えて総人口と大学卒業者数をダウンロードして先ほどの仮説について考えてみましょう

191

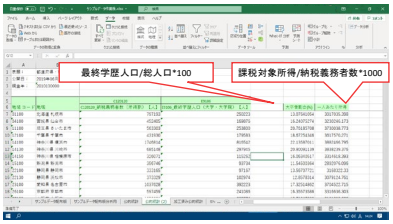
デモ⑧-2



① e-Statからデータ（表示項目A人口・世帯から総人口、C経済基盤から課税対象所得、納税義務者数（所得割）、E教育から最終学歴人口（大学・大学院）、調査年2010年を選択）をダウンロードします（詳細はP158-167の手順を参考）。

192

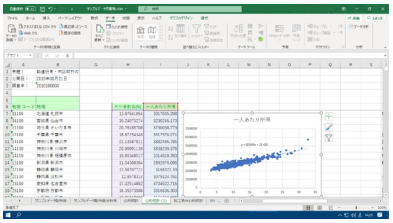
デモ⑧-2



②「大卒者割合」と「一人当たり所得」を算出します。

193

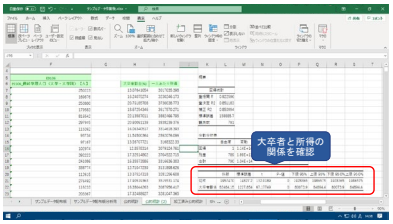
デモ⑧-2



③「大卒者割合」と「一人当たり所得」で散布図を表示させます。

194

デモ⑧-2



④ p 値を確認したい場合は回帰分析を行ってみましょう。

195

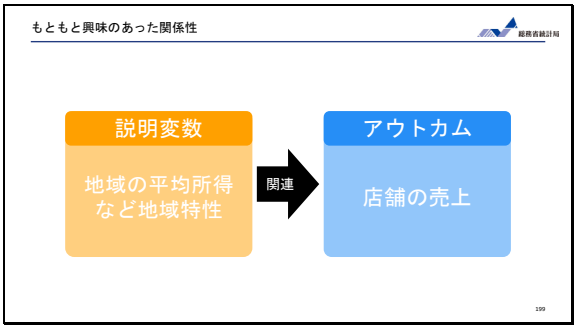
前節で用いたデータ以外にもe-Statからダウンロードできる統計データには様々なものがあります。

「グループ分けして平均値を比べてt検定」か「単回帰分析でp値も確認」というこれまでに学んだ分析方法を使って「どのような特性の地域において（一人あたりの平均で）所得が高いと言えるか」を探索してみましょう。

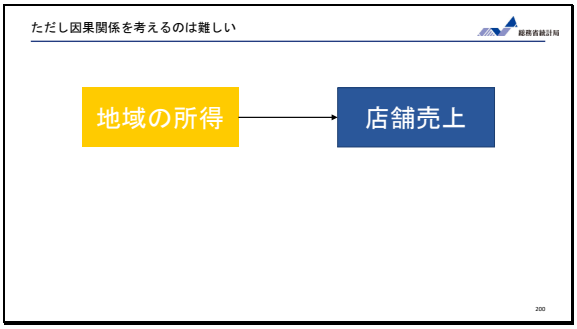
第3章　統計データを組合わせよう

配布されたサンプルデータと公的統計を組み合わせて分析する方法を実践していきます。

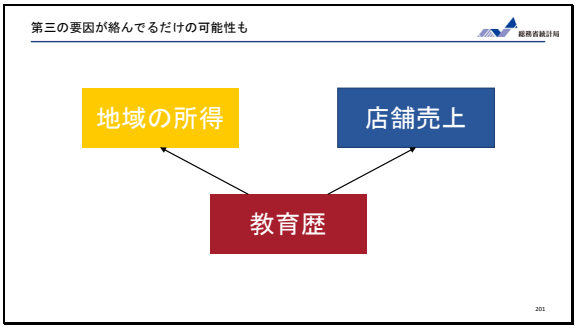
第一章で触れたサンプルデータ(飲食チェーンの店舗ごと属性と売上)とe-Statからダウンロードした公的統計を結合して分析するために、エクセルのvlookup関数の使い方と重回帰分析の考え方を学びます。



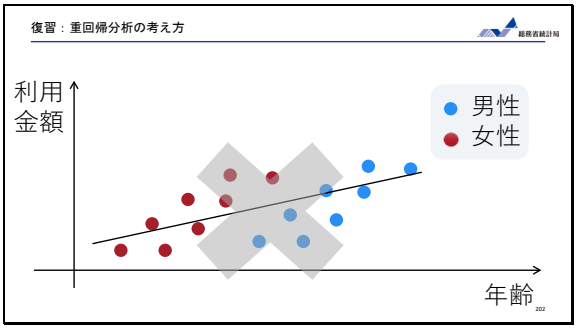
当初の分析は自社データを用いてアウトカムが店舗の売上に対して行っていました。
次に公的データを活用し地域特性が説明変数として分析結果が得られるように進めてきました。



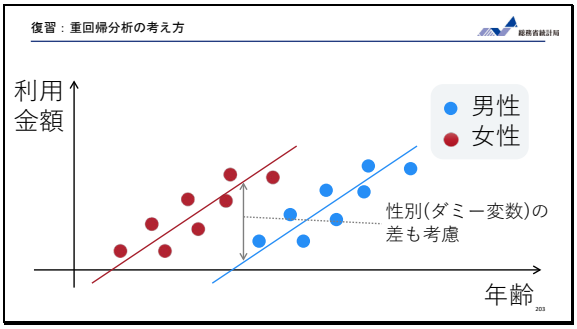
一見すると地域の所得と店舗売上は関係しているように見えますが、もしかすると他の要因があるかもしれません。



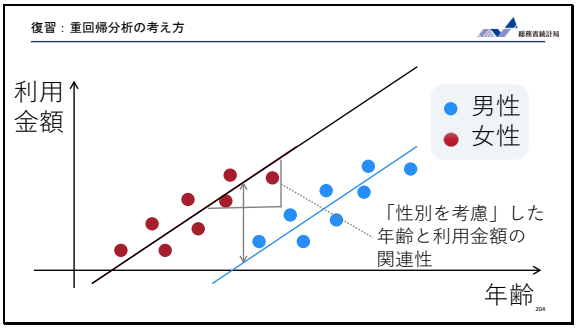
例えば、教育歴などです。
教育歴と地域の所得の関係が強い場合、店舗売上に関係しているのは教育歴の可能性があります。



重回帰分析で様々な条件を考慮します。
すでに説明したように複数の説明変数が存在する場合、一本の線だけで説明することはできません。



それぞれの項目で平行線を引いて、その差を考慮します。



性別、年齢と同じように、店舗属性や地域属性も重回帰分析を用いることでアウトカムとの関係を知ることができます。

補足：データを紐付ける際に注意すべきこと

- 異なるデータソースだと地域名の表記が曖昧
- 今後もミスタイプなどがあるかも
- コピーアンドペーストする際にもズレると
- 市区町村名は「市区町村コード」で管理して
  VLOOKUP関数を使うと安心です

205

VLOOKUP関数の使い方

=VLOOKUP
(検索したいIDの値,
「$」をつけて検索するデータ範囲,
データ範囲の左から何列目の
値を抜き出したいのか,
近似一致をするか[通常False])

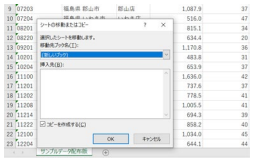
206

VLOOKUP関数を実際に使う際の流れ

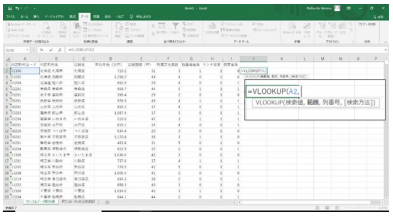
1. エクセルシートを1枚のブックにまとめる
2. 店舗ごとのデータの横にVLOOKUP関数を入力
3. 別シートのデータ範囲を選択
4. A1をA1というようにセルを絶対参照で固定
5. この式をコピーして一番下まで貼り付け

207

デモ⑨

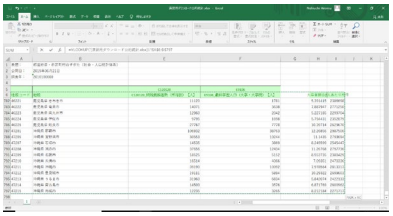


①まず使用する複数のエクセルシート(店舗データおよび公的統計)のシート名を右クリックし「移動またはコピー」から1つのエクセルブック(ファイル)にコピーしてまとめる(なお使うデータはダミー変数化済みのものであることに注意)



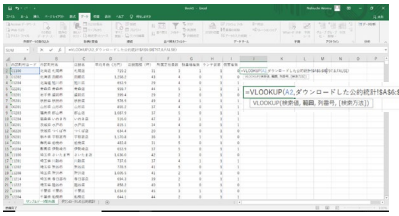
②店舗ごとのデータの右端にVLOOKUP関数を入力(説明変数の列がひとかたまりになるよう位置に注意)

デモ⑨



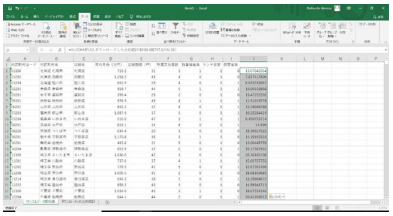
③別シートにある公的統計のデータ範囲を選択

デモ⑨



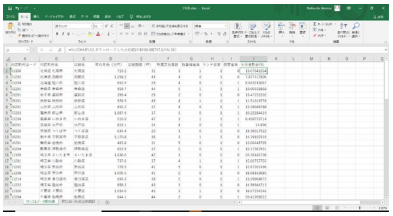
④「A1」なら「A1」といったように検索する範囲を指定するセルを絶対参照で固定するように注意

デモ⑨



⑤この式をコピーして一番下まで貼り付け

デモ⑨



⑥ 最後に項目名を1行目にコピー(複数列を分析したい場合はここまでの工程を繰り返し)

デモ⑨



⑥無事データが結合できたら分析ツールで回帰分析

デモ⑨



⑦説明変数が複数列(矩形で選択)なところ以外は単回帰分析のときと同じ

デモ⑨

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-1874.03	604.7151	-3.09903	0.002838	-3081.05	-667.015	-3081.05	-667.015
店舗面積（坪）	20.39036	3.683005	5.536337	5.56E-07	13.03905	27.74166	13.03905	27.74166
所属正社員数	148.7026	33.62291	4.422656	3.67E-05	81.59094	215.8142	81.59094	215.8142
駐車場有無	-195.056	68.53643	-2.84602	0.005869	-331.856	-58.257	-331.856	-58.257
ランチ営業	78.75383	71.04517	1.108504	0.271608	-63.0529	220.5606	-63.0529	220.5606
個室有無	174.4024	68.95625	2.529174	0.013793	36.76513	312.0396	36.76513	312.0396
一人あたり課税対象所得	0.000463	0.000239	1.937186	0.05694	-1.4E-05	0.000941	-1.4E-05	0.000941
大卒者割合	-0.37078	21.0968	-0.01758	0.98603	-42.4802	41.73861	-42.4802	41.73861

⑧回帰係数やp値などの解釈も基本的に単回帰分析と同じ

217

おまけ：余力のある人向けのチャレンジ課題

今回用いたもの以外にもe-Statからダウンロードできる統計データには様々なものがあります。

VLOOKUP関数と重回帰分析を用いてどのような条件が店舗ごとの売上に関係しているか探索してみましょう。

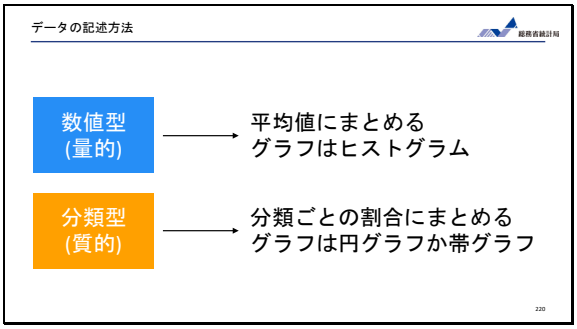
218

今日やった内容：「まとめる」ことのメリット

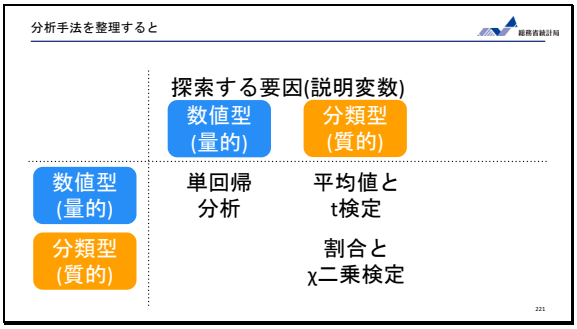
1. 現状を把握するための記述ができる　主にここ
2. 現状を左右するような要因の探索ができる
3. 今後限られた情報だけで予測ができる
4. その要因が本当に原因なのか検証できる

219

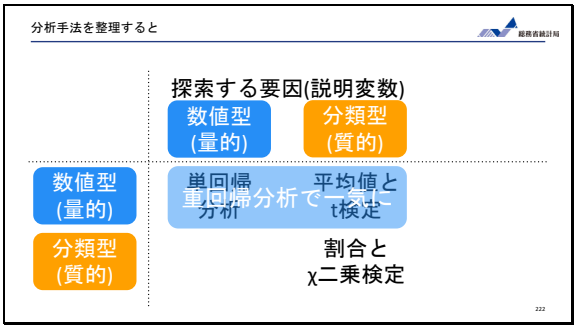
本セミナーでは主に記述と探索について説明を行いました。



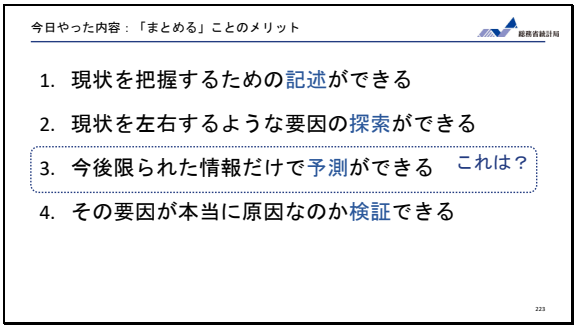
数値型については平均値にまとめ、グラフはヒストグラムで記述します。
分類型については、分類ごとの割合にまとめて、グラフは円グラフか帯グラフで記述します。



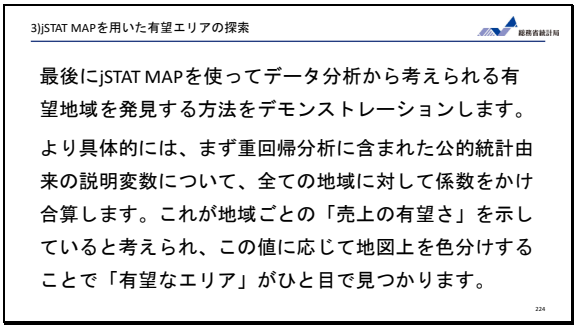
アウトカムと説明変数の型（数値、分類）の組み合わせで分析手法が異なります。

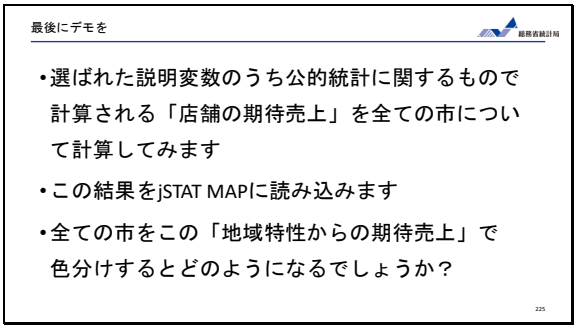


アウトカムが数値型のものについては、重回帰分析を使用することで一気に分析をすることができます。

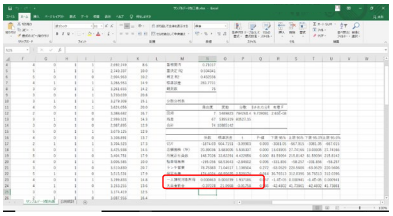


最後に補足として、予測についてデモを行います。

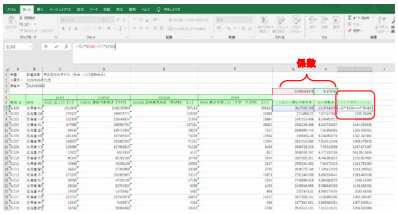




デモ⑩

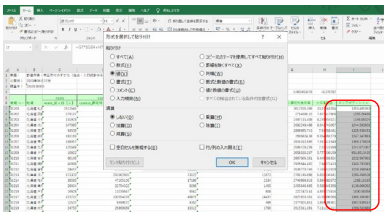


①回帰分析で算出された「係数」を使用してエリアポテンシャルを作成します。



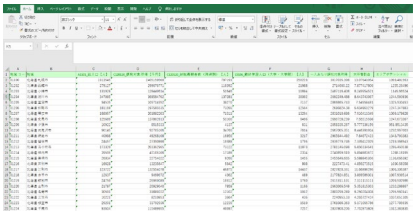
②係数にそれぞれの説明変数の値を掛け算して足し合わせた数式(絶対参照に注意)をコピーして一番下まで貼り付けます

デモ⑩



③計算できた値を「形式を選択して貼り付け」から値だけを同じ領域に貼り付けて値を固定します

デモ⑩



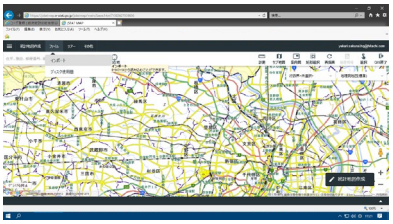
④1行目に項目名が来るように不要な行を削除してcsv保存を行いましょう

デモ⑩



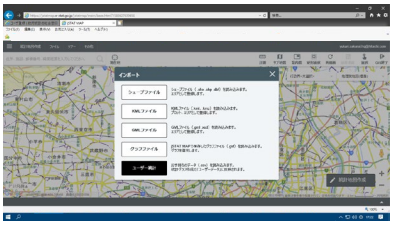
⑤jSTAT MAPにログインします。
※jSTAT MAPの「ユーザ登録」を行ってください。
（https://www.e-stat.go.jp/mypage/user/preregister）

デモ⑩



⑥ファイル>インポートを選択します。

デモ⑩



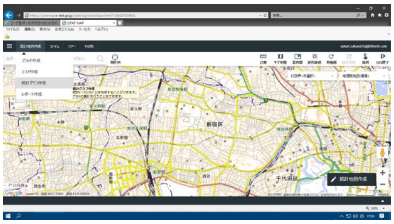
⑦「ユーザー統計」を押下します。

デモ⑩



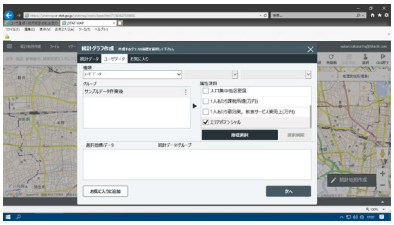
⑧種類は「国勢調査」「2015年」「市区町村」にし、アップロードするファイルを参照から選択します。

デモ⑩



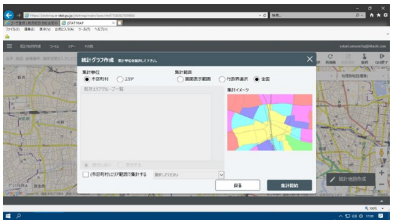
⑨統計地図作成＞統計グラフ作成　を選択します。

デモ⑩



⑩ユーザーデータタブを選択＞種類のプルダウンからアップロードしたファイルを選択＞属性項目でエリアポテンシャルにチェックを入れる＞指標選択を押下＞次へを押下します。

デモ⑩



⑪集計単位「市区町村」＞集計範囲「全国」＞集計開始を押下します。

デモ⑩



⑫地図にエリアポテンシャルがプロットされたら完成です。

238

4)まとめとビジネスへの活かし方

最終的にここまでの内容のまとめを行うとともに、データ分析の結果をビジネスに活かすやり方として「変える」「狙う」「大丈夫にする」という３つの切り口を提示します。

すなわち、発見された説明変数について有利な方向に物事を「変える」か、説明変数が有利なところを「狙う」か、逆に不利なところを「大丈夫にするか」という観点で考えられる施策例を具体的に示します。

239

データ分析からのアクションの考え方①：『変える』

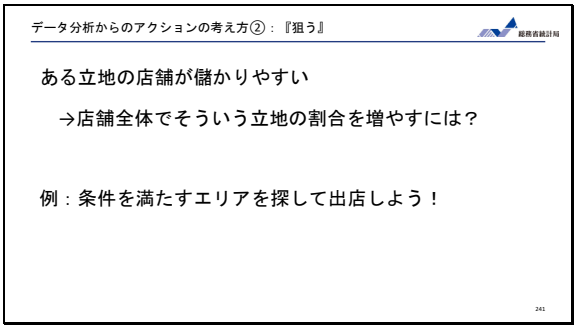
ある属性の店舗が儲かりやすい

→今ある店舗をそのように変えたらどうだろう？

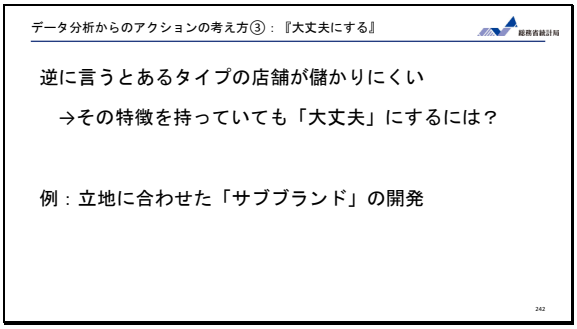
例：店舗を改装してみよう！

240

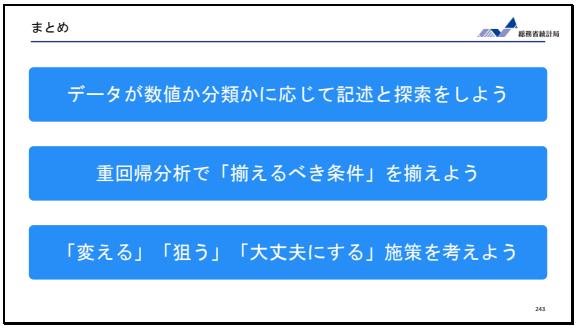
データ分析の結果をアクションにつなげるための考え方を解説します。
大きく3パターンに分けられます。
まずは、「変える」ですが、データ分析の結果から特定の属性がある店舗が売上が高いと出たとします。
もしそれが個室の有無や、スタッフ数など変更可能なものであれば是非変えてみてください。



次に「狙う」という方法です。
データ分析の結果を参考に自社の顧客層にマッチするエリアを公的統計データなどから探して新店舗を出店してみましょう。



最後に「大丈夫にする」となります。
データ分析の結果、あるタイプの店舗は儲かりにくいということが分かったとします。
そういった店舗を抽出し立地に合わせてリブランド、店舗改装などを行います。



まとめとなりますが、データ利活用はまず、データが数値か分類か理解して記述と探索を行います。
その後重回帰分析でそろえるべき条件を揃えて分析を行います。そうすることで関係している説明変数だけでアウトカムの値を求めることができます。
最後に分析結果の説明変数を活用し「変える」「狙う」「大丈夫にする」施策を考え、実行します。