事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
相関係数を二乗すると説明力になるのはどうしてでしょうか。	相関係数はプラスとマイナスがありますので、二乗する必要があります。また、直観的な理解になりますが、散布図から説明力という情報を得るには、相関係数そのままの数値より、二乗した数値の方が感覚的に合致します。
相関２はひげ図で外れ値となりましたが、相関１の値はなりませんでした。相関１で関東を外れ値と捉えるにはどういった分析が必要になりますか	箱ひげ図で外れ値とする基準も一つの判断方法です。おっしゃってくださっている箱ひげ図は「その数値」における外れ値の判断方法です。関東を外れ値と捉える方法は、単純に散布図から、少し外れている数値に見えるという判断や、例えばほかの地方と比較して関東だけ人口が多いから外れ値とするとか、多面的に考えたうえで、関東が外れているという判断も可能です。
人口の影響を除いたときの広告費と売上高の相関係数0.373とのことですが、これは人口の影響がないということでしょうか。	そのとおりです。人口の影響を除いた場合における相関係数です。
P73、P74でなぜ小売業所と人口の2つの相関がありそうなことが分かるのでしょうか？	2つのグラフの横軸の目盛は同じであり、全体の分布がほとんど同じように見えます。従って、相関があるのでは？という仮説が立てられます。
相関分析でP値を出す必要はありませんか？	P値（有意確率）は、データが多い場合、算出するとすべて小さな値（例：0.00‥1）となってしまいますので算出する意味が無くなってしまいます。
e-StatとRESASは、国としてどのような使い分けをすることを想定しているのでしょうか？	国がどう想定しているかは、e-statは数値データだけになっている状態ですが、jstat-mapでは単純に地図で見える化できるのが違います。それはつまり、自身の営業所なり支店なりがある市町村内だけでなく、周辺の市町村、自動車での30分圏内におけるデータを見える化することによって、特定の地図上の地点の周辺のデータが見える化されることが、使い方だと思います。特定の地点の想定がない場合は（特定の支店の住所がない）、e-statで広くデータを分析できるようにしていると考えることが可能です。
jSTAT MAPに入っていないデータ(独自で把握している地域別の動物の数など）を入れて、地図上に色分けして表示することはできますか。	表示できます。参考URLを入れておきますので参考にしてください。 https://jstatmap.e-stat.go.jp/manual/gis_manual.pdf https://www.stat.go.jp/info/guide/public/kouhou/img/pdf_10/10.pdf
平均値と中央値のどちらを使うのか基準はありますか？	一般的には平均値がよく用いられますが、外れ値のある場合は中央値の方が代表値として優れています。データの分布を把握して使い分けてください。
作成した地図を公開する場合の出典の書き方を教えてください	データそのものは、estatを使っているので、出典はたとえば https://www.e-stat.go.jp/help/surveyitems/search-3-5 にある通りでよくて、それを図示したということでjstatmapを書くと良いと思います。以下例です。 「出典：「政府統計の総合窓口(e-Stat)」、調査項目を調べる－国勢調査（総務省）「住居の種類」を地図で見る統計（jSTAT MAP)にて図示した」という書き方でよろしいかと思います。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
Laxusの例で、「バッグは流行が早く、陳腐化したバッグは家に溜まっていってしまう」とのごLaxus社が保有していたら、不良在庫になってしまうのではないでしょうか？流行があることと、同社がビジネスとして成立している因果関係がわかりません。流行に左右されないものを選抜して保有し貸し出すことによって収益化しているという理由であればまだわかるのですが、流行があるにも関わらず、その点をむしろ強みとできている理由について教えて頂けますでしょうか。	おっしゃる通り、バッグは流行が早く、陳腐化したバッグは家に溜まっていってしまう人のために貸し出しがセールストークです。ただ法人としては、おっしゃる通り不良在庫になるから、流行に左右されないものを選抜して保有し貸し出せるよう、分析してレコメンデーションで出すようにしてみたわけです。それで満足度が低いものなった場合は、さすがに処分（格安販売）などを実施することになるようです。
仮説を持つタイミングについての質問。資生堂のところで話がありましたが、データ分析を進めながら、仮説を持つということでもOKということですね。よく「仮説を事前に準備してからデータ分析をしよう」言われていますが、最初から持つのはハードルが高いと感じています。	おっしゃる通りです。分析を進めながら、こうかもしれない！と思ったことがあれば、それが仮説です。
事業コンセプトが当初の「家にある不良在庫バッグを共通資産化して多重利用してもらう」から、単に「ニーズが高いバッグを集めて貸し出す」に変わった、と理解すればよいでしょうか？	セールストークとしては、前述の内容で今もセールスすべきで、会社としては後述の内容にした方がコストカットにつながると思います。究極流行に左右されない商品が個人に知られたら、個人はそれを買えばいいから、知られたくはないですよね？そういう意味で、セールストークとしては前述、会社としては後述にしているわけです。
相関係数に関しての質問です。相関係数を算出した後に二乗すると、解釈が容易になるとのことで、統計では二乗を活用する機会が多いような印象を受けておりますが、そもそもなぜ二乗なのでしょうか？	世界は二乗でできている、という本があるくらい、二乗というのは色々なものを説明可能とします。二乗すると情報量が多くなり、人間にとって解釈がしやすくなるため、二乗が多用されます。
棒グラフや折れ線グラフだと、グラフ上の個別データをクリックすると、表中の該当データが示されます。しかし、散布図では一つのデータをクリックしても、表全体が示されてしまい、データが探しにくいです。これを解決する方法はありますでしょうか？	元データを並び替えてみることが多いです。図から知りたいほとんどのデータは、とても高いか低いかのデータです。そのため、データを並び替えて、小さい順または大きい順にすると、該当データを見つけやすくなります。
交絡要因を偏相関で計算する方法と、売上と広告費をそれぞれ人口で除して一人当たりにしてから相関を計算するのでは、どちらが良いでしょうか。また、それぞれメリデメはありますか？	とても良い視点です。ただ、一人あたりの広告費および売上を算出したとき、人口多いところは一人あたりの金額が少ないことになりますね？それは本当にそうなのでしょうか。この計算で算出された指標は、人口が多いところは金額が少なく、人口が少ないところは金額が多くなるという法則のもとによる、相関ということになります。これは正しいでしょうか？正しいかどうかは分かりませんよね？そうなると、単純に一人あたりの金額を算出して相関を出すよりは、広告費と売上の関係の内部に人口という要因が及ぼしている影響を排除するだけの方が、より信頼性は高そうです。このような考えから、通常は偏相関を使うことが多いです。
近似曲線の書式設定の中で「切片」というところを選択すると、回帰式が変わります。これはどういう風に解釈すればよろしいでしょうか？	切片を選択すると、切片がないものとして考えるようになり、いつでも原点0からの傾きを計算することとなります。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
先ほどのなぜ決定係数は2乗を使う方が感覚に合致するか？についての説明を、もう一度、より詳しくご説明いただけないでしょうか？	これは相関係数を二乗の部分でしょうか。感覚に合致という点では、ウェーバーフェヒナーの法則にあるとおり、私たち人間は、ミクロな変化についてよりも、マクロな傾向を読み取ることが得意です。散布図における各点から計算する結果は、ミクロな変化を表象しているだろうなと感じていただければ、二乗することや対数をとることが、感覚に合致しやすくなることに合意いただけると思います。
回帰分析のところで、"P値 0.2069" となっていましたが、ここの数値の大小は評価にどのように考慮すべきなのでしょうか？	P値の部分は、基本マーケティングや企業様の分析においては、あまり気にしなくても良いです。
交絡要因という言葉を初めて知りました。図のなかで「人口も関係あるよね」という話がでていましたが、これは「仮説」とは異なるものでしょうか。仮説と交絡要因とは違うものでしょうか。	人口も関係あるよね、というのは仮説になります。そして人口の要因を取り除いた偏相関の結果、相関が高くなる、または低くなるような変化が生じるならば、仮説とした人口が関連するというのを、確かめたともいえます。交絡要因というのは、単純に関係がありそう（またはなさそう）に見えていたものに影響を与えている要因のことをいいます。
先ほどのデータを切片ゼロにすると、R2が0.98になってむしろ当てはまりが良くなってしまいます。制度が良いように見えるのですが、どう解釈せばよいでしょうか？	切片がゼロというのは想定上考えられない状態となるため、口頭でもお伝えしたとおり、使うことがありません。確かに確認しましたら、R2乗はおっしゃる通り高くなってました。そういう意味では口頭での説明に間違えがありまして申し訳ありません。ただ精度が高まる低まるから使わないのではなく、切片がないことを想定することができないのです。なぜならば、切片というのは、ｘである説明変数の値が「0」だったときの、目的変数の値のことなのです。ｘの値が0であったらｙの値は必ず0だとは想定できないですよね？広告費と売上というのを考えたとき、広告費が0円だったときにも、恐らく売上というのは生じるはずです。切片をなくすということは、広告費が0だったら、売上も0である、という想定の下に生じる計算になります。通常、人の社会生活において、片方が0ならもう片方も必ず0になるというのは想定しづらいです。そのため、回帰分析には切片を0として分析することがない、と覚えておいていただければ良いと思います。
決定係数R2が2乗である理由として、負の値を消すという教科書説明に比べ、２乗の方が感覚的にわかりやすい、という説明はとてもユニークでなるほどと思いました。できれば聞き漏らしなきよう、もう一度、より詳しくR2(2乗)で確認する方が良い理由を教えていただけないでしょうか？	上述させていただきましたので、そちらをご確認いただけますと幸いです。
CSVよりもXSLS形式での出力のほうが使いやすいですか？使い分けはありますか？	CSVでも良いです。今回はExcelで操作することを想定していたので、吐き出されるデータがそのまま操作しやすい形式で出力しましたが、CSVの方が使われることが多いかな、と思います。
単純な質問です。estatからのダウンロードの際、「桁区切りを使用しない」にするのはなぜでしょうか？使用したほうが便利な気がしますが。	桁区切りを使うと、文字列としてExcel上で出るときがあるので、ただの数値とした方がエラーが生じづらいために今回そのようにしました。
ロジスティック回帰分析について教えて頂きたいです。	ご意見、ありがとうございます。今後のセミナーでロジスティック回帰分析も行えるようにしていきたいと思いますので、貴重なご意見として関係各所にお伝えいたします。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
jSTAT MAPで色々なデータを出力した場合、それぞれのデータの所得年度は分かりますか？	とても良い質問で、基本そのまま記載されていないと思います。最新のデータで作ってくれます。確認方法としては、統計グラフ作成で、年度が出ますので、そこで選択範囲におけるデータが何年度のものかの確認をいただくと良いかと思います。また、データそのものはestatのデータですので、estatで確認いただくのも良いかもしれません。
前段の講義での質問ですが，p56の相関係数（0.373）とp57の相関係数（0.362949）が違うのはなぜですか（計算式は同じはずですが）？	これは単純に私が手計算したからです。つまり、ほかの取り除きたい相関などの数値において私が勝手に小数点第3位くらいで四捨五入して計算したからです。P57の方はExcelで計算させたので、細かい小数点も全て含めて計算した結果、多少数値が異なっています。説明不足で申し訳ありません。
回帰式で、売上高と乗降客数の一次関数式を求めたのち、補正R2などの係数を用いて、意思決定のための売上高を予測する手法などはあるのでしょうか？それとも〇%の説明力という前提で一次関数で求めた値を採用して、意思決定するものなのでしょうか？	一時間数式が予測する式になります。補正R2はその式の精度を判断するための指標です。もしその精度が低いときには、例えば今回途中で行っていたような、データの中で、外れ値（関東のデータ）になっているところのデータを取り除いて分析したら、補正R2が上昇したら、そっちの式を使います。または、予測したい状況が、関東と環境が近いというならば、入れたデータの式で予測した方が良いと思います。
相関係数のところで、「外れ値」かどうかは、目で見て判断、でしょうか？	おっしゃる通り、目で見て判断です。「実際にはどの程度のずれていると外れ値とみなしますか？」の質問の回答も参考にしていただければと思います。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
今回の資料上の外れ値はわかりやすい値でしたが、実際にはどの程度のずれていると外れ値とみなしますか？	外れ値の選定については個々のケースの事情を加味して考えることが多く、そのための絶対的な根拠がございません。 ただ、多くの場合では四分位数か標準偏差という指標をもとに考えることが多いです。 まず、四分位数と呼ばれる、データの値を小さい順に並べていき、データ数を4等分(0～25%、25～50%、50～75%、75～100%)にして区切る境界の指標があります。この四分位数を用いると「箱ひげ図」と呼ばれるグラフを作成することができます。この箱ひげ図の両端についたひげの範囲を確認し、そこから外れた値を外れ値とみなす方法がよく使われます。 また、分析の際に平均値とともに用いられる指標に、データのばらつき(個人差)などを表す標準偏差(SD:Standard Deviation)というものがあります。平均値に対してこの標準偏差の値の2倍、または3倍以上差がある場合は、そのデータを外れ値とみなすというケースもあります。 たとえば、平均値が50で標準偏差が5であれば、40以下または60以上（2倍のケース）、35以下または65以上（3倍のケース）の回答を外れ値とみなすという形になります。 ただ、それぞれの値がもつ意味はどのようなデータを分析しているかによって変わりますので、このような統計的な基準とそのデータの背景(ビジネス分野など)・特徴などから総合的に判断すると、より適切な選択ができるかと思います。 外れ値の考え方やExcelでの箱ひげ図の作成の資料のリンクを共有いたします。こちらもご参考にしていたければと思います。 https://www.onoff.ne.jp/blog/?p=7502
疑似相関と同じでしょうか	疑似相関というのは、相関が高いけども、どう考えても関係があるとは思えないときのものです。たとえば、アイスクリームの販売数と溺死数のように。アイスクリームの売上げがあがると、溺死が増えるというのは相関を調べると高い相関となりますが、疑似相関です。単純に暑いからアイスクリームを食べる人が増えるという状態と、暑いからプールや海で泳ぐ人が増えている状態、という二つが相関係数が高くなってしまっていて、考えてみると直接関係がないと思えるものを、疑似相関と呼びます。講義で取り扱った「広告費」と「売上げ」は関係がないことは考えられません。ただ、その相関の高さ（または低さ）は、ほかの要因によって生じているかもしれない、というのが今回のお話しでした。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
講座の内容とはズレますが、散布図をポジショニングマップ的に解釈することは良くないことでしょうか。相関係数に現れないものを散布の形状で判断することが分析手法的に許されるものなのか知りたいです	散布図は2つの変数の関係性を可視化し、その特徴をつかむための分析ツールですので、必ずしも相関係数のような直線的(1次関数的)関係のみを対象にする必要はございません。 相関係数のような関係性の強さの指標を用いる必要性がないのであれば、散布の形状からそのデータ間の関係が意味するものをポジショニングマップ的に解釈することは、特に問題とはならないかと存じます。 特に層別に散布図を作成した際に、ポジショニングマップのように層(たとえば、若者・中年・高齢)ごとにその散布の範囲が分かれるようであれば、マーケティングを行う上での有用な分析になるかと思います。
相関関係と似た言葉に「因果関係」がありますが、どのような違いがありますか？ また、これは相関関係と同様に分析する手法がありますか？	相関関係と因果関係の違いについては、たとえばこちらのサイトをご参照ください（https://www.nttcoms.com/service/research/dataanalysis/causation-correlation/）。 因果関係を分析する手法はあります。ただし、それは因果関係を調べる目的で収集されたデータがないと困難です。今回の資料にあります「資生堂」様の事例は、因果関係を調べている事例です。「仮説」をたてて、それを検討しているということは、因果関係を調べているといえます。すなわち、「この広告戦略」が「売上げ」に影響するというのが、原因が広告戦略で、結果が売上げとなっているので、因果関係を調べています。因果関係までしっかり調べられると、資生堂様の事例のように、しっかりとした売上げや効率化の結果につながります。ただしお金と時間がかかるというのが難点です。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
estatのデータの年度が合致しないケースも多いと思います。そのようなときはどう考えるのが良いでしょうか？ 人口は2020年か2025年 課税対象所得データは2023年のデータがある。 例えば、このような場合は、人口と課税対象所得の関係を見るときどちらの年の人口を選ぶのが妥当でしょうか？	ご質問のとおり、分析のためのデータの出典年度などが完全一致しないというケースは実際の分析場面ではよくありえることかと思います。しかし、分析対象とするデータの選出は、分析の目的やデータの背景・特徴を加味して考えることが多く、絶対的な原則のようなものがございません。ただ、いくつかそのための方針や基準はあるのでいくつかご紹介いたします。 まず、単純な基準としては、2つのデータ間の時間差が小さくなるような組み合わせを選ぶという考え方があるかと思います。 ご質問で挙げられていたケースだと、課税対象所得のデータは2023年のみですので、2023年にもっとも近い2025年のデータの方が適切と判断することができるかと思います。 次に分析の目的を基準にするという方針です。たとえば、2020年の日本(の人口）、あるいは2000年以前の日本（の人口）に興味があるというように、分析したいデータの背景が明確な場合もあるかと思います。このように、分析を通して知りたいものとそのためのデータの前提が具体的な場合には、その目的に合致する(または近い)ようなデータを基準にして、使うデータ(セット)を選択するという考え方もあるかと思います。 最後に、2020年の人口や2025年の人口など、複数のデータセットのどちらも利用可能な場合であれば、それぞれで分析を行い、その結果を比較するという方法もあります。 たとえば、人口(2020年/2025年)と所得(2023年)の相関関係を調べたときに、2020年のデータでは相関係数が0.79、2025年のデータでは相関係数が0.81というような結果が得られたとします。この場合、どちらのデータでも傾向に大きな差はありませんので、人口と所得にはかなり強い相関があるという結論を下すことが妥当にみえます。 反対に、2020年のデータでは相関係数が0.61、2025年のデータでは相関係数が0.31というような結果が得られたとします。だとするとこの5年で人口(と所得)の実態に大きな変動があったことが推察できます。分析の目的にもよりますが、人口と所得には年代により弱い相関や中程度の相関があり、安定するような強い関係性はないのかもしれない、などと考察することなどもできるかもしれません。 上記の考え方などを参考に、分析の目的とその目的にどのくらい合致するデータを使うことができるのかなどを考慮して、自分なりにデータセット選出の根拠を設けていただくと、分析の妥当性が高まるものと思います。
estatで相関等を見たいデータの年度が合わない時は、どう考えるのが適当でしょうか？ 人口データは　2020年と2025年しかない 課税対象所得データは　2023年しかない この二つの数値の関係を見る際は、どちらの年の人口データを見るのが適当でしょうか？使用したいデータによって異なることもあるかと思いますので、原則がわかるとありがたいです。	
エラー値に1をかけて解除する方法をもう一度教えてください	どこか適当なセルに1を入れて、それをコピーし、形式を選択して貼り付けを選択します。その中で「乗算」または「除算」で貼り付けを、エラーが出ているセルに実施してください。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
私はサプライ品のメーカーで営業部門の管理を行っております。 本日の相関と回帰を使って営業系の数値の分析をするのはどのようなことがありますか？	関係性を調べるという相関は、どの場合でも使用可能です。営業であるならば、どのサプライ品が、どの会社でいつくらいに売れるというのがあると思いますが、その関係が強いかどうかを数値で知っておくこと、または予測しておけるようにしておくことは、重要だと思います。または、営業で大成功、成功、不成功ということがあるならば、その結果と関係が強いのはどれであるかを明確にすることで、今後の営業戦略に活かせると思います。また、講義でも話しましたが、効率化という点で、移動経路と時間と成功について関係性を調べておくと、仕事の効率化につながると思います。
層別散布図の２番目の作成方法で、自動的にタイトル（商品満足度）挿入する方法はありますか？	2つ目の方法での層別散布図では、自動的に適したタイトルを入れることは困難で、ご自身でタイトルを入れるしかないかと思います。基本的な散布図であれば、タイトルは「列名」を含めて図を作成することできます。左と右の2列を選択して図を作成すると思いますが、右の列の列名が自動的に図のタイトルして認識されます。
3次エリアの数は、1次と2次を含んだ数字でしょうか？	3次エリアは、おっしゃる通り、1と2次のエリアのデータを含めた数字です。
グラフは全エリアの割合でしょうか、それとも1次エリアだけでしょうか	基本的にはグラフにエリア内と書いてあるときには、3次エリアまでのデータです。あとは「市」と「県」となるのがグラフの横に書いてあります。
データ分析をエクセルで行う時、どのような考え方でシートを分けていくと良いでしょうか？あるいはファイル分けをどのような考え方で整理していけば良いでしょうか	講義中に回答させていただきましたが、まずは原本というシートは必ず用意してください。そして加工用のシートを作成し、あとは行う分析ごとにシートを用意しておくと良いかと思います。
「散布図で」とか「相関係数で」とか、分析の手法を指定したうえで、「作業」自体を生成AIにやってもらう、というのは専門家の立場からするといかがなものでしょうか？	個人的には効率化という観点で、生成AIを利用することは大いに賛成です。ただときどき間違えますので、しっかり確認をすることが重要だと思います。簡単に言えば、誰かに作業をお願いしたとして、その作業結果を必ず確認しているかと思いますが、それと一緒です。
STAT MAPにデータのインポート可能とありましたが、セキュリティ面に不安があります。インポートしたデータはどのように扱われるのでしょうか？	インポートすると、一旦ログインしているIDにそのデータが必ず保存されています。保存されている先は政府のサーバですが、あとはそれを信用できるかどうかの話になりますので、個々または企業様でご判断いただくしかないことかと思います。
重回帰分析をする際には、様々な変数から分析対象となる変数をピックアップすることになると思いますが、どういった変数を分析対象としたらよいか選ぶコツなどありますでしょうか。	回帰分析であれば「目的変数（従属変数」を決めて、それ以外色々な「説明変数」を選ぶと思います。ご質問は、説明変数の選び方ということですね？コツとしては、目的変数と相関が高くない説明変数では、当然関係性が低いということですので、良い予測は立てられません。したがって、目的変数と関係が高い説明変数を選ぶというのが1つ。もう一つは説明変数同士の相関が高いと、良い予測が立てられません。これは、たとえば目的変数が売上げで、それを説明したいと思っているときに、その支社のセールスマンの数と従業員数という2つの説明変数では、関係性が高く相関が高いものになりますよね？こういうときには良い予測が立てられません。 似たような概念で一つのことを説明するよりは、色々な角度、色々な立場から説明させる方が良いと感じませんか？回帰分析も一緒で、できるだけ目的変数に対して、色々な角度から説明するように変数を選択すると高い精度の予測ができるようになります。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
e-Statでスーパーマーケットもダウンロードすることは可能ですか？	本日のセミナーで用いたホームセンター販売に関するデータと同様に、商業動態統計調査のデータベースから、スーパーマーケットに関するデータもダウンロードすることが可能です。 ホームセンターのデータ入手と同様の手順でe-statから「スーパー」等でご検索いただくと、商業動態統計調査のデータセット一覧のなかに「百貨店・スーパー販売」のデータセットをご確認いただけるかと思います。 下記のリンクからもアクセスできると思います。 https://www.e-stat.go.jp/stat-search/database?page=1&layout=datalist&toukei=00550030&tstat=000001081875&cycle=1&tclass1=000001081878&hclass=100%2C150&tclass2val=0&metadata=1&data=1 ホームセンターがページ下部だとすると、スーパーのデータはページの上部側にございます。 本日の復習や、今後の分析の練習等ご活用いただければ幸いです。

事例から学ぶビジネスパーソン向け統計データ利活用セミナー

受講者からの質問事項及び講師の回答

受講者からの質問事項	講師の回答
仮説を立てて分析を進める際にその仮説は正しくないと見極めるポイントなどがあれば教えてください	分析プロセスで仮説の妥当性を判断するポイントですが、仮説に対応するような分析を行い、その指標をもとに考えることが要点になるかと思います。 たとえば、「顧客の年齢が上がるほど来店頻度は低くなる」というような仮説を立てた場合は、年齢と来店数(頻度)の関係性を検討することが分析の目的となります。 この場合は、この2つの変数の相関係数を算出し、その値から仮説の妥当性を判断します。相関係数が-0.78であった場合は、強い負の相関があるので、仮説の妥当性は高いと考えることができます。一方、相関係数が-0.11であった場合には、ほとんど関係性がないので、仮説が正しくなかったと判断できるかと思います。あるいは、相関係数が0.42であった場合には、関係性はありそうだが正の相関であるため、仮説で想定していたものではないと判断することができます。 絶対的な基準ではありませんが、相関係数の絶対値が0.7以上でかなり強い相関、0.4~0.7を中程度の相関、0.2~0.4を弱い相関、0.2以下を無相関、と考える見方もあるので、こうした基準と照らし合わせて判断するというのも有効だと思います。 このような形で平均値や中央値、相関係数などの代表値や、回帰係数などの指標を算出し、具体的・客観的な数値から仮説の妥当性を判断するということがポイントになります。 また、このような分析で検証ができるレベルまで仮説を具体的にする、ということも分析プロセスでは非常に重要になります。そのため、分析するデータや分野について詳しくなるということも、良い仮説を立てること、そしてその妥当性を考えることにとても役立ちます。