**計量社会科学**　　　清水　剛

**この講義の目的**

　ザックリ言うと「社会についてデータをとるときにどうしてもブレが生じる。そのブレがなぜ生まれるのか、どうすれば対応できるのか」ということです。

～本題に入る前にちょっと用語の整理を。～

☆分析法

**量的分析**：数量化されたデータを得ること

　　　　　所得・支持率・収穫量など数字で表されるやつ

　　　　　計算できるものはこれになるので、授業の後半で扱っているのはほぼこれです。

　　　　　科学的・客観的なようですが様々なブレが生じてそう簡単ではありません

**質的分析**：その逆

　　　　　哲学・歴史学・文学など言葉で分析するやつ

　　　　　作文を「自分のありのままを書こう」と言われても「ありのまま度5％UP」というわけにはいかないが、言ってることは分かるって感じです

現状を示しているようでそうとも言えません。今まで見てきたものにバイアス（詳しくは後で）がかかるので。

☆尺度水準

何を基準にとるかということは重要です。ALESSでいうとIndependent。

次の4つは集合でいうと、上の方が下の方を包む関係にあります。つまり下がより詳しいくくりになってます

**名義尺度**：単に分類をしてるだけ　男性・女性など順序の比べようのないもの

変数間のデータとしては最頻値くらいしか使えません

**順序尺度**：分類の順序を示している　優・良・可・不可など「順位はあるけど優ひく良って何？」となるもの

中央値も使えるようになります

**間隔尺度**：測定値の差が意味を持つもの　摂氏何℃など

気温40℃の日は10℃の日より4倍暑いというわけではないので。

　　　　　平均が使えるようになります

**比尺度**：数値そのものが意味を持つもの　時間・長さなど

　　　　　比・四則演算も使えます

1. つぎの尺度はどの尺度水準にあたるか。
   1. 収入
   2. 日付（何月何日など）
   3. 絶対温度
   4. モース硬度（ダイアモンドは10で一番堅い）
   5. 「センター数学はどの数字を書けば一番当たるか」という時の「数字」
   6. テストの得点（83点など）

　　　　　　　　　　　　　　　　　　　　　　　　※問の答えは全て最後にあります

では、本題に入ります

～調査～

☆調査方法

　社会調査のために多くのデータを取らなくてはなりません。しかし、調査員を雇うと（調査票につき給料を変えるとかにしがちなので）でっちあげる可能性があるということは頭に留めておきましょう。

**面接調査**：調査員が調査対象と直接面談。　　駅前でインタビューなど

**留置法**：調査票を配布し、後で回収。

**郵送法**：アンケートを郵送し、送ってもらう。　　本に挟まっているハガキなど

**電話調査**：ランダムに電話をかけ、応えてもらう。

**インターネット調査**：ネット上にアンケート。

いろんな方法がありますが、それぞれに問題もあります。

テレビのインタビュー：プライバシーの問題・40代会社員などに偏りがち

留置法：高コスト・本人が本当に回答した？

電話調査：家に固定電話が無い人は対象外になる　　　　などなど

テレビで「調査の結果…」とか言い出したら、調査の偏りを考えてみましょう。

☆調査票の設計

　調査票の質問文の言葉遣い（**ワーディング**）に気を付けましょうってことです。どういうことに気をつけたらいいか分類します

「社員」など**多様な解釈が可能な言葉**（「社員」は会社法で従業員ではなく株主などの意味）。

専門用語など**意味の分からない言葉**。「つてを頼る」or「コネを使う」（ものは言い様）など**一定の評価を示唆する言葉**。「喫煙は健康によくないのでやめるべきだ」など二つにするべき質問が混ざる**ダブル・バーレル**（喫煙が健康に良くないことが前提になっている）。「世間では～と言われていますが、あなたはどう思いますか？」など**インパーソナルとパーソナル**。

ワーディングだけじゃなく選択肢によっても回答が変動します。

授業評価アンケートみたいな「１．とても思う　２．少し思う　…　５．全く思わない」という形式のやつは**リッカード尺度**といい、極端な１・５が選ばれにくくなります。

※「次の質問文にはどのような欠点があるか」といったような問題が予想できます。

調査倫理に関してはプリントを見てください（多分でないとは思うんだけど）

一応、「インフォームドコンセントに反する研究」は重点的に説明していた気がします。

何もナシってのもアレなので1つだけ

人間の思い込みの力は強く、「薬を飲んだ」と思うだけで病気が治っちゃうときがあります。これをプラシーボ効果といいます。さらにその薬を被験者に渡す人が結果を知っていると、無意識のうちにそれを感じ取って結果が変わるかもしれません。そこで新薬などの調査をするとき、それを渡す人にも効果がある薬かどうか教えないで調査することあります。これを**二重盲検法**と言います。

　しかし、どういう効果がある薬か教えないのは倫理的にどうなの？という問題があります。

～サンプリングとバイアス～

　データをとるとき基本的に全部を調べることは難しいので、その一部をサンプルとして取り出して調べることになります。ただ、いい加減にサンプリングすると偏り（バイアス）がかかって正しい調査ができなくなります。つまりランダムに選ばなきゃダメってこと。

結果的にサンプルが偏っても、それは起こりうる範囲内くらいなので大きな問題ではない。

☆サンプリング手法

**単純無作為抽出**：観察対象全体に対して等確率でサンプルをとる

　　　　　　　　できるなら一番理想的だが、調査が大変なうえにそもそもリストがあるの？という問題が

**系統的抽出**：最初のサンプルから等間隔に

調べてみたけどメリットが見つからなかった…

　　　　　　意味あるのかな？

**多段階抽出**：一旦大きなくくりでサンプリングする

全国規模でサンプリング調査すると遠い島に何度も行かなくてはならなくなって大変なので、まず調査する市区町村を選びそれから人を選ぶ

**層化抽出**：東大生に対して調査するとき、理Ⅰから○人…とするみたいな

☆**バイアス**の種類（重要）

選出バイアス：選ばれた人に偏り

　　　　　　　上述のインターネット調査なら若者中心になってしまう

無回答バイアス：選んだ人が平等でも答える人の割合が違う

　　　　　　　　「お客様満足度９７％」みたいなやつは次の注文をするついでにアンケート回答するので必然的に高くなる

回答のバイアス：ある質問の回答が、ほかの質問や社会同調のせいで歪む

　　　　　　　　例は「妨害効果　誘因効果」って検索して

問2．次の調査にはどんなバイアスがかかると考えられるか

①　学校の授業評価アンケート

②　選挙の出口調査

～データの分析～

データをいい加減に分析すると誤った結論がでるので気を付けましょう

☆相関係数

　まずは単純に

問3．次のデータの相関係数を求めよ

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 5 | 4 | 3 | 7 | 2 | 9 | 6 | 5 | 8 |
| Y | 3 | 4 | 6 | 1 | 8 | 5 | 5 | 7 | 4 | 7 |

ちなみに、上の図を散布図と言います。

しかし、単純に相関係数を出すだけでは結論を間違えることもあります。

例えば、2次関数的な関係には相関係数は大きくならないし、１個だけ**外れ値**があるとそれに大きく影響されます。

また、本来別々のものとして考えられるべきデータを一緒にしたり、バイアスがかかって全体のデータが得られないと相関係数が歪みます。（多分、ここ重要）

例はネットで「アイリス・データ」と検索すると、本来は正の相関係数になるデータなのに、一緒にデータとったせいで負の相関係数が計算される例が出てきます。

それを踏まえて予想問題を

問4．次のデータ分析の誤りはなぜ起こったか。また、どういうデータが（正の相関係数・無関係・２次関数・円などなど）どういうデータとして観測されてしまったと考えられるか。

1. 上述の「アイリス・データ」
2. 小学生の身長と３０秒で解ける足し算の数について調べると強い正の相関係数が出たので、学力を上げるには背を伸ばすといいと結論付けた。
3. とある大学のとある学部の大学生について、センターの英語と数学の得点を調べたら強めの負の相関係数が出たので、英語力を上げたければ数学の勉強をしなければよいと結論付けた。
4. 何人かの人に一週間の飲酒量と脳年齢を調べたら負の相関係数が出たので、飲酒をすると頭がよくなると結論付けた。

（注・あくまで例なので事実か分かりません。一般的に推測してってことで。）

あとは計算問題ばっかり

基礎統計でやったとこは教科書を見てください（中央値・最小２乗法など）

次の問題は授業プリントを見ながらで

問5．問４のデータについて**ケンドールのτb**を求めよ

問6．次のデータについて**クラメールのV**を求めよ

（この2つは、先生が√2しか出さないと言ってたので出ない気も…）

|  |  |  |  |
| --- | --- | --- | --- |
|  | 高 | 中 | 低 |
| 高 | 12 | 14 | 5 |
| 低 | 3 | 16 | 33 |

**答え**

問1．①300万円+400万円＝700万円のように足し算できるので**比尺度**

②3月5日+2月10日＝5月15日とはできないが、

　　　　各日付の間隔（1日間）は共通なので**間隔尺度**

　　　③300Kは150Kの2倍のエネルギーのように掛け算できるので**比尺度**

　　　④モース硬度が大きくなると固くなるが、その間隔が一定であるわけではないので**順序尺度**

　　　⑤数字なので一見間隔尺度のようにも見えるが、各数字の順番に確率が上がったりしないので**名義尺度**

　　　⑥80点の人が40点の人の2倍賢いわけではないので比尺度ではない。

　　　　点数の順番に賢いので**順序尺度**ではある。

　　　　各得点の間隔が均等かは微妙。英単語テストなら均等かも。よって**間隔尺度**と言えるかは場合による。

　　　　でも、入試では単純に合計するし、進振りも単純平均なんだよね…

ただし、間隔尺度ならば順序尺度かつ名義尺度なので一番詳しいやつを書いてます。

問2．①授業に出る人にしか配られないという選出バイアス

　　　　めんどくさがりは答えないという無回答バイアスは…あるのかな？

　　　　周りの友達の目があるから5にしづらいという回答のバイアス

　　　　リッカード尺度なので1・5を選ばないという回答のバイアス

　　　②期日前投票はカウントされないという選出バイアス

　　　　プライバシーに厳格な人は答えないという無回答バイアス

　　　　記者の目を気にする回答のバイアス（勝ちそうな方へと偏りやすいそうです）

他にも見つけたという人は多分それも正解です

問3．Xの平均＝5、Yの平均=5なので

　　　X：(1-5)2+(5-5)2+……+(8-5)2＝60

　　　Y：(3-5)2+(4-5)2+……+(7-5)2=40

　　　XY：(1-5)(3-5)+(5-5)(4-5)+……+(8-5)(7-5)=29

　　　29÷（√60×√40）≒0.59

問4．①ネット見たら書いてあったと思いますが

　　　　2つの異なる種類のあやめのデータを混ぜたので、正の相関が負の相関になった。

　　　②「学年が上→身長が高い・学力がある」という関係なのに見かけの相関に騙されてそのまま処理したので、無関係が正の相関になった。

　　　③「ある大学のある学部→その集団のセンターの合計点はほぼ同じ」です。センターの合計点が同じということは、いわばｘ＋ｙ＋…＝一定の関係にあるのでxとyに負の相関係数が出てしまうのは当然。よって（おそらく）正の相関にあるものが負の相関になった。

　　　④「実年齢が若い→脳年齢も若い・飲酒量も多い」という関係なのに見かけの相関に騙されてそのまま処理したので、（おそらく）負の相関にあるものが正の相関になった。

問5．データをXの小さい順で並び替えると

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 5 | 6 | 7 | 8 | 9 |
| Y | 3 | 5 | 1 | 6 | 4 | 4 | 7 | 8 | 7 | 5 |

その正順・逆順は

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 正順 | 8 | 4 | 7 | 3 | 4 | 4 | 1 | 0 | 0 | 0 |
| 逆順 | 1 | 3 | 0 | 3 | 0 | 0 | 1 | 2 | 1 | 0 |

よってG-H＝20

　　　　n=10,Tx=1,Ty=3なので

τb＝20÷（√（45-1）×√（45-3））≒0.465

問6．各項目の総計を出すと

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高 | 中 | 低 | 計 |
| 高 | 12 | 14 | 5 | 31 |
| 低 | 3 | 16 | 33 | 52 |
| 計 | 15 | 30 | 38 | 83 |

この場合最も関連のないケースは

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高 | 中 | 低 | 計 |
| 高 | 5.60 | 11.20 | 14.19 | 31 |
| 低 | 9.40 | 18.80 | 23.81 | 52 |
| 計 | 15 | 30 | 38 | 83 |

これらの差は

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 高 | 中 | 低 | 計 |
| 高 | 6.40 | 2.80 | -9.19 | 0 |
| 低 | -6.40 | -2.80 | 9.19 | 0 |
| 計 | 0 | 0 | 0 | 0 |

よって、χ2検定統計量は、(6.40)2/5.60+…+(9.19)2/23.81≒22.3

φ係数は、√(22.3/83)≒0.518

クラメールのVは、これを√（2-1）=1で割って0.518

テストは用語問題2題・論述問題2～3題・計算問題1題って言ってたような…

予想としては、

用語：「留置法とはどういう意味で、どのような問題点があるか」

　　　「次の質問文にはどのような問題点があるか」　など

論述：問2・4

計算：問3・5・6　かな？

あくまで予想なので各自で考えてね