統計学はお嫌いですか?ー政策に統計学のすすめー

中央大学大学院　公共政策研究科教授　**細　野　助　博**

皆さんは行政に入られて、まちの情報がこんなにも市役所には集まるのかと驚かれた経験があると思います。そうなんです、行政には自然と情報が集まります。ですから、その情報をできるだけ行政活動に活用するということが住民参画の時代、地方分権の時代、地域間競争時代になって一層求められているのです。これから、皆さんがデータを十分に活用して業務に役立てようとするときの、基本的な知識を述べたいと思います。

1．データって何？

(1)　データは数？符号？どっち？

皆さんのところに集まってくる情報をデータと呼びましょう。普通考えられるデータとは数字である場合が多いのですが、数字でないデータもあります。たとえばアンケートでは統計をとる場合に、１を男性、２を女性という区別をしますが、欧米では一般に１を女性、２を男性とします。慣習ですからどちらでも構いません。つまり皆さんが使われるExcelデータで性別を表すときの１、２は単に符号でしかないのです。順序は関係ありません。

今度はそこに順序を付け加えてみましょう。例えば学歴というものがあります。１を小学校卒、２を中学校卒、３を高校卒、４を大学卒以上としましょう。確かに教育年限の長短の順序がつきますが、たとえば小学校は６年、中学・高校は３年、大学は４年とその間隔は等間隔ではありません。旧制と新制の区別もあり順序には関係しますが、その間隔は便宜的なものです。等間隔でも気温なども同類と考えていいでしょう。人間の感覚を単に便宜的に温度という等間隔にしただけであり、21度から22度になる間隔と30度から30.5度の差はどちらの方が暑いと感じるでしょうか。ですから、気温や学歴のようなものは間隔よりも順序のほうにずっと意味があると考えたほうがいいでしょう。そして皆さんが日常手にする数字のデータに人口数、課税対象所得、行政面積などがあります。なぜデータの種類を厳密に区別するのか。それは名前くらいは聞いたことのある相関分析や、回帰分析などちょっとカッコイイ分析にはどのデータなら使って良いか、あるいは悪いかを判断する必要があるからです。データの種類はとても重要なのです。これは統計学を学ぶ大きな理由の一つと言ってよいでしょう。

(2)　データは時間に乗るって？

住民基本台帳人口は１ヶ月とか１年毎に集計されますが、それを10年、20年の長期にわたって手にすることもできます。時間に伴う変化を見るために時間軸で収集されるデータですから時系列データといいます。物価もそうですね。物価は季節の規則的な変動を伴います。例えばボーナス時期、消費額は集計するととても大きくなりますから、ビジネスはこの時期に価格を一斉に上げようとします。これが物価変動として現れてきます。

また消費者意識（物価、賃金や暮らし向きに対する感じ方）も時系列データの代表例ということができます。時系列データは季節変動、長期的な構造変化を伴う変動、そしてちょっとした天候不順や事件をもとにした不規則変動というものにどんどん分解することができます。株価を見るとずいぶん変動します。誰も予測できない分、株屋さんが儲かるのでしょうね。彼らは予想屋さんですからね。変動が激しいことを歓迎します。

(3)　データは空間を走るって？

さて、皆さんにとって、もっと大切なものが

あります。それが空間データ、あるいは横断面データ（クロスセクションデータ）です。皆さんはお隣の自治体の人口がどうなっているか気になりますよね。新しい大規模マンションができて若い世帯がたくさん住みだすらしい、新しい事業所が進出してきたらしい、という情報に神経を尖らせるでしょう。隣よりも少しでも新鮮なよい情報が詰まったデータがほしいと思うでしょう。ですから、年や月を定めて、一定の期間で収集される横断面データがとても気になると思います。例えば昨年度23区の転出入人口と多摩・島しょ地域の転出入人口などを比較すると、おおよそ人口増加は10倍の開きがあります。都心回帰が既に定着していることをこのデータから実感すると思います。

マクロで見ると、人口は魅力を求めて合理的に移動していきます。情報や交通のネットワークが充実すればするほど人口は空間を走っていきます。特に交通利便性の高い首都圏をはじめ大都市圏では顕著に現れます。その結果が人口変動というデータに表れるのです。

⑷　必要なデータ数って？

皆さんは「ビッグデータ」という言葉を聞いたことがあると思います。昔はデータの数は多ければ多いほど正確な情報が得られると「勘違い」していました。実はビッグデータの「ビッグ」は数の多さよりも内容の豊富さのほうが重要なのです。例えばコンビニでカードで支払うと、どこに住んで、家族が何人いて、誕生日がいつで、昨日は何を買ったか、今日は何を買ったか分かる訳です。

それが何万人、何十万人積み重なったとしても、確かに予測の精度は上がりますが、平均値の精度は「ルートの法則」に従います。例えば100倍のデータは10倍の精度、10,000倍のデータは100倍の精度になる訳ですね。これはアメリカの有名な実例ですが、大統領が誰になるか予測をした時、データの数は少ないのですが、まんべんなく有権者の気持ちを収集できたデータは、それより何十倍もたくさんのデータを取りはしたが偏った階級の有権者の気持ちしか収集できなかったデータよりずっと予測の精度が高かったのです。つまり必要かつ十分な数のデータがあれば、それで十分なのです。むしろそれよりも、データだけに頼るのではなく、統計学を使用しようとする時の注意深い観察とそれを裏付ける理論的な分析のほうがずっと重要なのです。

2．比較はむずかしい？

⑴　データの平均って？

先ほどデータは必要十分な数だけあれば十分という話をしました。それは、本気になってたくさんのデータを集めようとするとコスト（お金と時間）がかかるからです。データの売買をめぐる犯罪が世間を賑わせていますが、それほどデータというものには価値があるのです。では、データからどのような情報が得られるのでしょうか。それは調べようとしている調査対象の平均的な姿を推測すること、それと同時にデータから計算して求めた平均値が果たして調査対象全体を代表しているかどうかを検定するという２つの作業を通じてデータの教えてくれる情報の信頼性を確認することです。なぜ、このような注意深さが必要かというと、データが教えてくれる平均値は調査対象の姿を如実に教えてくれる可能性が高いからです。

皆さんも「お味噌汁」の味を吟味する時に、「よくかき混ぜる」でしょう。あれも「全体の味を知るために平均を採る」ための動作です。平均値はデータの集まりの中心、あるいはデータ全体を代表する値です。例えば小学１年生の平均体重や平均身長などがその代表です。皆さんは小学校の児童なら全数（悉皆）調査だから、その平均は全体の平均と思われるかもしれません。しかしちょうど風邪をひいて休んでいる児童もいるかもしれません。体重計のいくつかが故障して正確な数値を表していないかもしれません。ですから厳密には「標本」として捉えるべきです。得られたデータについては注意深い検討が必要ですね。もっと重要なのは、今、格差ということが話題に上っていますが、わがま