ちの平均所得､わがまちの一人あたりの資産額､といった場合、大変なお金持ちに引っ張られる平均値には｢代表｣としての意味があるのでしょうか。大多数が持たざるもの、少数が独占している、そういう格差社会では平均がもつ意味はずいぶん違ってきます。

⑵　データの分散って？

皆さんは平均値についてはよく注意をすると思いますが、データの散らばり具合についてはあまり注意を払ってくれません。例えば小学校の学力のデータを例に取りましょう。平均点の高さで教育熱心な学校の順位をつけるか、それともピンとキリの成績の幅の大小も評価するか、これはとても難しい問題です。義務教育課程で学力の方向を判断するとき、学校全体の平均だけで済む問題ではありません。公教育の使命を考えた場合、ピンとキリの差や点数毎の分布がむしろ重要性を帯びてくる時代に入っています。従来のように家庭が学校の教育を十分に補完し得なくなっている現在、キリの児童数やピンとキリの成績の差にこそ注目すべきです。散らばり具合をあらわす尺度はたくさんありますが、その代表例が原数値と平均との隔たり具合で計算される分散と言われるものです。この平方根が標準偏差です。

同一の教科書を使っていても、行政区域内で土地柄や児童生徒の家庭状況がまったく違っている場合、平均だけでなく分散も計算して、２つの尺度で見ることがとても重要になります。これは５年間ある市の教育委員を務めた経験で実感したことでもあるのです。小中学生はランドセルやカバンの中に教科書ばかりではなく、地域性や家庭環境をそのまま忍ばせて登校してくるのです。

⑶　標準化が必要なわけ？

例えば今年の算数の問題は難しかったが国語は簡単だった、あるいは偶然ある学校区においては優れた算数の先生がいて、児童の算数の成績は上がったが、そのかわり国語の先生がそれほどでもなく国語の成績が振るわなかったとか、いろいろな事情によって児童の成績に関するデータは変化していきます。そこで、学校の成績順位というものを一つの尺度で評価しようというとき、様々な要因から発生するデータの散らばり具合を調整して評価することが重要になります。

実例を挙げましょう。次頁の表をあわせてご覧ください。産官学連携組織である「ネットワーク多摩」では学生たちによるまちづくりコンペティションを今年度発足しました。予選と本戦があります。予選ではグループを３つに分けました。各グループにそれぞれ６人の審査員が張りつきます。不幸にも、とても辛い点数をつける審査員が多数いるグループの参加者はどれも一様に不利になり、甘い点数しかつけない審査員が多いグループの参加者は当然有利になります。ですから、素点そのままで順位をつけると不都合が生じることは納得できると思います。大学の名もエントリー番号に修正した表を見てください。素点の「色」に注目してください。すると、甘い審査員が多いグループは「オレンジ色」だということが分かります。そして「コワーイ、辛い点」をつける審査員が多いグループは「青色」ですね。このまま（素点で）順位をつけると、８番と19番が１位で、２番が３位ということになります。ともに「甘い審査員」達の恩恵に預かっています。これでは「不平等！」という声が上がってきても不思議ではありませんね。そこで、データの標準化が必要になります。一般的に標準得点は、$z=50+10\times\left(\frac{\text{素点}-\text{平均}}{\text{標準偏差}}\right)$で計算します。これを偏差値ともいいます。この値は国民の間ではお馴染みですね。この場合の平均は全体の平均ではなく各グループを構成する審査員たちが各グループの大学につけた平均です。また標準偏差はその平均と素点から求められます。Excelならばすぐに計算できるはずです。

まちづくりコンペの素点の順位と標準化した点数での順位は違ってくることを皆さんにこっそり教えましょう。もちろん、標準化した点数での順位が採用されます。内緒のデータですよ。グループの色に注目して下さい。そして素点の

合計と標準化した得点の順位が入れ替わっていることを確認してください！データを比較するには注意深さが必要だということがお分かりになるでしょう。

素点順位	素点	エントリー番号	偏差値順位	偏差値	エントリー番号
1	259	8番	1	68.4	4番
1	259	19番	2	65.1	5番
3	214	2番	3	60.8	8番
4	212	4番	3	60.8	19番
5	204	5番	5	55.7	13番
6	200	7番	6	54.1	9番
7	188	13番	7	51.0	10番
8	174	3番	8	47.4	3番
9	173	17番	9	46.8	17番
10	172	9番	10	46.8	2番
11	169	10番	11	45.1	15番
12	161	6番	12	45.0	7番
13	153	14番	13	38.8	1番
14	147	15番	14	38.7	12番
15	135	1番	15	35.0	14番
16	129	12番	16	34.5	6番

表　まちづくりコンペの各大学の素点と偏差値

3．データのいたずらとは？

(1)　測り方の間違い？

先ほど体重計が壊れていることを知らないでデータを集めていた例をいいましたが、統計学は測り方の間違いを評価し、それを修正することから始まったと考えていいでしょう。例えば几帳面に小数点２桁までデータを取る人がいると思いますし、大雑把な人は小数点を全部切ってしまう場合もありますし、不規則に四捨五入する不届きも当然あります。それを集めて情報にするわけですから、当然思っても見ないようなばらつきが出てくるということがお分かりになるでしょう。

学問に男性が得意なものと女性が得意なものがあるとすると、統計学は大変几帳面な人が多い女性が得意な学問と言っていいでしょう。事実私も学生時代、統計学は大嫌いでした。統計を学ぶなら遊んでいたほうがずっといい不埒な学生でした。人生は皮肉なもので、逃げれば逃げるほど苦手なものは追ってくるわけですね。測り方の間違いは思い込みや錯覚も影響してきます。しかしその人間の弱さが学問を発展させる原動力にもなるわけです。

(2)　外れ値って？

データを手にしたら、まずグラフにしてみてください。すると飛び抜けたデータというものが時々見受けられます。これを外れ値といいます。先ほど述べたボーナス配給月などはある種の外れ値と言っていいでしょう。あるいは東京のデータ、ニューヨークのデータ、パリのデータ、ロンドンのデータは、メガシティである分それぞれの国の平均からかなり外れた数値を弾き出すので、これらの外れ値をそのまま使うかどうかは、政策的な意図に当然依存します。また外れ値を意識して使う場合もあります。例えば、今年の入学者の水準を前年と比較する場合に、１位や２位の学生のデータ（これも外れ値）を２、３年前と今年で比較し、今年は優秀な学生が入った、入らなかったという判断もできるのです。

(3)　得られなかったデータって？

大学ではそれぞれの年度の受験生の水準を測り、入試の成績と大学教育の成果との関係を測りたいとしましょう。受験時のデータは２つに分かれます。１つは入学者の「１年次の成績を追跡できるデータ」。もう１つは不合格になったり他の大学に流れたグループの「データにすることができないもの」です。そうすると受験生から得られるデータは大学教育の成果を調べる場合にかなり不正確なデータに変わってしまいます。これを切断されたデータといいます。でも、このようなデータが存在することは出入りが激しい地域の住民意識を知る場合に注意が必要なことを暗示してくれます。これから私達はデータが示す情報の不完全さ（これをいたずらと言いましょうか）を十分に意識してデータを取り扱わないといけないということが分かってきます。

まだまだ統計にはおもしろいトピックスがありますが、紙面が尽きましたのでこのお話の続きはまたの機会に。