研究員の眼

# 平均値の信憑性

## 平均値は、その集団を代表しているか?

保険研究部 主任研究員　篠原　拓也
(03)3512-1823　tshino@nli-research.co.jp

どんな統計でも同じだが、個々の数量データをそのまま表示するのは冗長でわかりにくい。そこで、何らかのデータ加工が必要となる。例えば、データを棒グラフや円グラフにまとめて、分布の様子を見れば、どのような傾向があるのかをより早く読み取ることができる。しかし、データを用いて数量的な評価をしようとするときには、いちいちグラフを描いて分布をみることは面倒でもある。そこで、個々のデータを表示する代わりに平均値を計算し、その値が集団の平均的な姿を表しているものとみなして、議論を進めることが多い。

生命保険の保険料計算はまさにその典型で、例えば、万一の際に保険金が支払われるような死亡保障保険では、死亡率を用いて保険料を計算する。死亡率は過去に販売した保険の死亡実績や、国民全体の人口動態から算定される平均値である。そして、これに一定の調整をした上で、保険料はもとより、保険会社が将来の保障のために積み立てる責任準備金や、保険の途中で解約した場合に支払われる解約払戻金など、様々な形で、保険価格の計算のベースとして使われている。

しかし、平均値には、注意しておかなくてはならない落とし穴がいくつかある。

まず1つ目に、一部データが他の大多数のデータから乖離しているために、平均値が影響を受けることがある。例えば、健康な40歳男性の100人の集団が1年の間に何回入院するか(入院率)を平均値として計算してみる。過去の実績をもとに、集団全体ののべ入院回数を100で除して入院率を求めることとなる。過去の実績を見ると、おそらく大多数の人は1回も入院をしていないか、せいぜい1回の入院にとどまっているはずである。しかし、その中に体調を崩して何回も入退院を繰り返すような人が混じっていれば、それが1人なのか、2人なのかによって集団の平均値は大きく違ってくる。

2つ目に、集団の設定に偏りがあるために、その平均値も歪んでしまうということがある。アメリカの事例だが、治療過程で男性の性機能障害を引き起こしやすい前立腺がんについて、治療法別の性機能回復率を調査した結果が新聞で報じられた。小線源治療、放射線治療、外科手術の順に回復率が高かったとの結果だった。このことから小線源治療が最も性機能障害を起こしにくい、と結論づけてしまいそうだが、そもそも小線源治療という治療法は若くて体調の良い男性に多く用いられるため、この結果は至極当たり前のものといえる。平均値のもととなる集団の設定に注意して、データが均質

と言えるかどうかを考える必要がある。

3つ目に、集団の総数が少ないために、平均値から離れた部分にもある程度データが存在していて、平均値が集団を代表しきれないことがある。集団の規模が大きくなると、分布が安定してきて平均値からある程度離れた部分にはデータの存在が限られるということが数学の定理として示されている。逆に言えば、規模が小さい集団ほど分布が収れんしていないため、平均値が変動しやすいことになる。例えば、ある集団のうち1年の間に何人が風邪をひくかというような推定を行う場合、風邪をひく、ひかないという分布が収れんして平均値を安定させるためには、通常は、集団内に30人以上のサンプルを設定して計算をする必要があると言われている。

それでは、平均値の代わりに何か集団を代表する手頃な指標はないだろうか。集団のちょうど真ん中のデータの値である中央値とか、もっとも多く発生したデータである最頻値がその候補になり得る。これらの値は、平均値の持つ弱点を補って集団を代表する指標となり得る。しかしながら、平均値が持つ数学的性質が成り立たないため、処理がしにくく統計の実務ではあまり使われていないようである。しかし、「統計のウソ」、「数字のカラクリ」云々といった話題が巷間で盛り上がる中、日頃統計を読み解く際に当たり前に使っている平均値について信憑性を気にしつつ、中央値や最頻値を使った多面的な捉え方もできるように、理解の幅を拡げておいてもいいのではないだろうか。

 