統計数理研究所　2013年度子ども見学デー

BB弾サンプリング実験　解説

2013年10月19日土曜日、統計数理研究所では「子ども見学デー[1]」を開催いたしました。当日は曇りがちで気温も低い1日でしたが、多くの方に研究所までお越しいただきました。お立ち寄りいただいた皆様、どうもありがとうございました。

子ども見学デーでは、白と黒のBB弾がたくさん（全部で10万個）入った水槽を用意して「BB弾のサンプリング実験」を行いました。はたしてこの水槽の中に黒玉はいくつ入っていたのでしょうか。

サンプリング実験

参加者のみなさんには、この水槽からコップを使って取り出した300個の玉のうち、黒玉がいくつ入っていたかを数えてもらいました。もしこの300個が水槽の中からかたよりなく取られたものだとすれば（黒ばかり選んで集めたりしなければ）、300個のうちの黒玉と白玉の割合は、水槽の中の黒玉と白玉の割合と似たものになっていると考えられます。ですので、300個のうちいくつが黒玉だったかを数えて黒玉の割合を調べれば、水槽の中の黒玉の数が分かりそうです。300個のうちの黒玉の数をもとに、水槽の中の黒玉の数を以下の計算によって予想してみましょう。

[1] http://www.ism.ac.jp/events/kodomo2013/index.html

$$\text{（水槽の中の黒玉の数の予想値）} = \frac{\text{（300 個のうちの黒玉の数）}}{300} \times 10\text{ 万}\quad[\text{個}]$$

ここで、[（300 個のうちの黒玉の数）/300] という量は 300 個のうちの黒玉の割合です。この割合を水槽に入っている BB 弾の数（10 万個）に掛けることで、水槽の中の黒玉の数を予想するのです。例えば、300 個のうち 80 個が黒玉だった場合の予想値は

$$\begin{aligned}\text{（水槽の中の黒玉の数の予想値）} &= \frac{80}{300} \times 10\text{ 万} \\ &= 26667\quad[\text{個}]\end{aligned}$$

となります。

参加者のみなさんの予想はどうだったのでしょうか。今回の実験では次ページの写真のように、参加者のみなさんに数えてもらった黒玉の個数をシールに書いて、グラフになるように貼ってもらいました。[2] グラフの横軸は 300 個のうちの黒玉の数（“48〜52” のように適当な範囲で区切られています）を表しており、縦軸は頻度を表しています。シールが縦に長く積み重なったところは、その範囲の黒玉の数がよく観察されたことを表していると言えます。

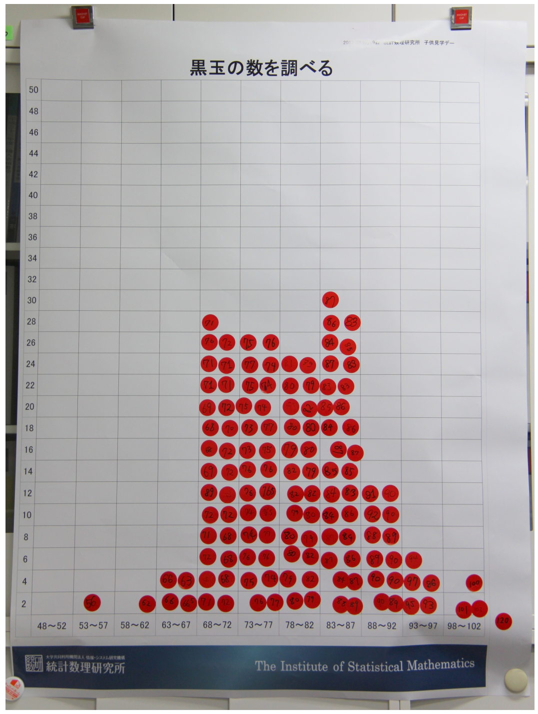
[2] 実験で得られた黒玉個数の結果は以下の通りです。

(56, 62, 63, 66, 66, 66, 68, 68, 68, 68, 68, 68, 68, 69, 69, 70, 70, 71, 71, 71, 71, 71, 71, 71, 72, 72, 72, 72, 72, 72, 72, 72, 72, 73, 73, 74, 74, 74, 74, 74, 74, 75, 75, 75, 75, 75, 76, 76, 76, 76, 76, 76, 76, 77, 77, 77, 77, 79, 79, 79, 79, 79, 79, 79, 79, 80, 80, 80, 80, 80, 80, 80, 80, 80, 81, 82, 82, 82, 82, 82, 82, 83, 83, 83, 83, 83, 84, 84, 84, 84, 84, 84, 85, 85, 85, 85, 86, 86, 86, 86, 86, 87, 87, 87, 87, 87, 87, 87, 88, 88, 89, 89, 89, 89, 90, 90, 90, 90, 90, 90, 91, 92, 93, 94, 95, 96, 97, 100, 101, 103, 120)

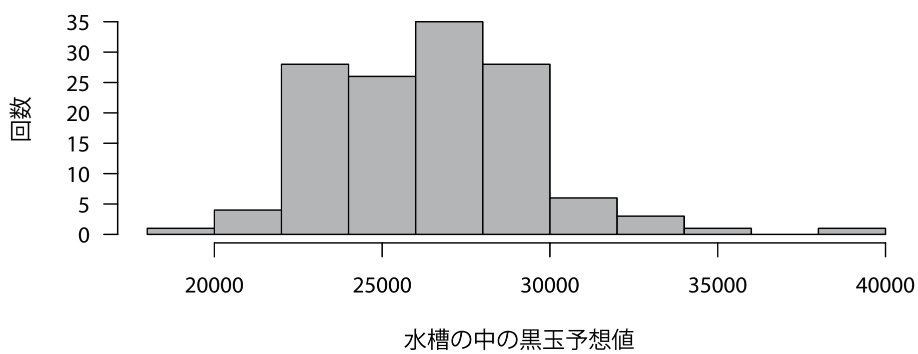
上記の結果の他に、2 回遊んだ場合の平均値が 3 つありました。

(66.5, 76.5, 76.5)

このレポートではこのデータも含めた結果を報告しています。



このデータをもとに、みなさんの水槽の中の黒玉予想値を計算して上と同じような図にすると、以下のようになりました。



グラフは一つの山のような形になっています。少ないと 2 万個以下、多いと 4 万個に近い予想値が出る場合もありましたが、だいたい 2 万個から 3 万個の間で山が高くなっていることから、実験された方の多くの予想値が 2 万個から 3 万個くらいの範囲に収まっていたことがわかります。

実は、黒玉は水槽の中に**2万5千個**入っていました。みなさんの予想はかなり良いものだったのではないでしょうか。10万個のうちの300個、全体のごくわずかを調べただけなのに、それなりの正確さで水槽の中の黒玉の数を予想できてしまったのです。

便利なサンプリング法

水槽の中の全てのBB弾を調べなくても、抜き出した300個から黒玉の割合を調べれば水槽の中の黒玉の数がある程度予想できることが感じられたと思います。この実験で体験していただいたように、大きな集団の特徴（ここでは水槽の中の黒玉の数）を知るために、集団の一部を抜き出して全体の特徴を推測することをサンプリング法と言います。数が多過ぎるなどの理由で全体を調べることが難しい場合には、サンプリング法はとても便利です。BB弾を10万個調べなくてはならない状況を想像してみてください。とても大変で1日では終わらないかもしれませんね。[3]

サンプリング法は社会の様々な場面で用いられています。日本人全体の特徴を知りたいという状況を考えてみましょう。特定の時間に見ていたテレビ番組や支持する政党を全ての人に聞くのは無理なので、テレビの視聴率や政党の支持率などはサンプリング法によって調べられています。統計数理研究所が実施している日本人の国民性調査[4]でも、時代ごとの日本人のものの見方や考え方を捉えるためにサンプリング法が用いられています。社会を対象とした調査だけでなく、自然を対象とした自然科学研究でもサンプリング法は重要な役割を果たしています。

また、サンプリング法の基礎は統計学の理論が支えています。前のページの、みなさんの予想値のグラフをもう一度見てみましょう。300個を取り出すサンプリングを何度も繰り返すと、予想値が一つの山のように分布することがわかると思います。山の中央はだいたい、本当の値である25000個のあたりにあり、予想値が10万個になるといった極端

[3] 10個数えるのに10秒かかるなら、全て数えるのに27時間以上かかります。子ども見学デーに関する平野勝臣先生のコラムもご覧ください。http://www.ism.ac.jp/ism_info_j/labo/column/090.pdf

[4] http://www.ism.ac.jp/ism_info_j/kokuminsei.html

なことはほとんど起こりません。サンプリング法による調査はこのような性質があるおかげでなり立つと言えますが、統計学にはこのような性質が現れることを説明する、重要な数学の定理があります。その定理を用いることによって、サンプリング法による予想値の精度なども知ることができます[5]。統計学のおかげで、BB 弾を 10 万個調べなくても済むのです。

分析と執筆：統計数理研究所　統計思考院

特任助教　深谷肇一

上で述べた定理とは「中心極限定理」と呼ばれるもので、（厳密でない説明になりますが）かたよりのない BB 弾サンプリングを繰り返した時に、得られる黒玉の個数はおおよそ正規分布という、つりがね型の分布にしたがうことを主張するものです。ここで「かたよりなく」とは、水槽中の全ての玉が等しい確率でサンプルされることとします（このようなサンプリングを「無作為抽出」と言います）。少し高度な内容になりますが、この定理を用いて今回の実験について 2 つのことを考えてみたいと思います。

まず、BB 弾のサンプリングが毎回かたよりなく行われていたのかどうかについて考えてみましょう。中心極限定理から言えることは、もしサンプリングが毎回かたよりなく行われているならば、実験で得られた黒玉個数データはおおよそ正規分布にしたがうということです。逆に、もしデータが正規分布にしたがうとは考えられないような分布になっているならば、それはかたよったサンプリングが行われていたことを暗示していると言えるでしょう。

黒玉個数データが正規分布にしたがうとみなせるかどうかは、「統計学的仮説検定」と呼ばれる手続きによって確かめることができます。ここでは説明を省略しますが、統計学的仮説検定では、データが正規分布にしたがうと仮定した場合に今得られているデータがめったに得られないものであると分かれば、その時はデータが正規分布にしたがうとはみなさない、と考えます。今回の実験データで確かめてみた結果、得られた黒玉個数の分布は正規分布とはみなせないことが分かりました。どうやら毎回かたよりなくサンプリングが行われていたわけではないようです。

[5] サンプリングによる予測の精度については、ぜひ昨年の子ども見学デーの解説もご覧ください。http://www.ism.ac.jp/events/kodomo2012/20120804_sampling.pdf

ここで3ページの予想値グラフを見てみると、グラフの右側に一つだけ山から外れたデータがあることがわかります。実験中、1度だけ黒玉が120個も得られるという珍しいことがあったのです。試しにこのデータを除いて、改めて同様の仮説検定を行なったところ、データの分布はおおよそ正規分布と考えて良いということが分かりました。

以上のことから今回の実験では、少なくとも1度は黒玉ばかりがサンプルされるような、かたよったサンプリングが行われていたことが示唆されます。120個もの黒玉が得られた回では、水槽の中のBB弾がよく混ざっていなかったか、あるいは黒玉ばかり狙って取ってしまうことがあったのかもしれません。黒玉の数を知りたいときには黒玉をたくさん集めてしまいたくなるかもしれませんが、かたよったサンプルは水槽中の黒玉の割合を正しく反映しないため、サンプリングによって集団全体の特徴を知りたいときには無作為抽出を行うことがとても重要です。

最後に、水槽の中にある黒玉の割合についてもう一度考えてみましょう。本文で、水槽の中には白玉と黒玉合わせて10万個が入っており、そのうち黒玉は2万5千個あると述べました。水槽中の黒玉の割合は$25000/100000 = 0.25(25\%)$です。ところが、この割合が本当に正しいものかどうか、実は私たちにもわかりません。というのも、このBB弾サンプリング実験は毎年の子ども見学デーで行なっており、10万個のBB弾を用意したのはもう何年も前のことだからです。

実験や後片付けの途中でBB弾が無くなってしまうことがあるので、実際にはBB弾の個数は10万個より少ないと考えられます。この時もしも、白玉と黒玉の間で無くなりやすさが異なっていれば、どちらか一方の色の玉が多く無くなるので、水槽中の黒玉と白玉の割合が時間とともに変わっていくことが予想されます。もしかすると、水槽中の黒玉の割合は10万個のBB弾を用意した当時と今では違っているのではないでしょうか。

この「仮説」は、統計学の知識があれば今回の実験データを用いて検証することができます。みなさんの実験データから先ほどの「外れ値」を除いたものの平均値を求め、これをもとに水槽中の黒玉の割合を求めると、本来の25%よりも少し高い26.5%であると計算されます。(説明は省略しますが) ここで再び中心極限定理を用いると、もしも水槽中の黒玉の割合が25%だったとすれば、実験データから1.5%も高い推定値を得ることはめったに起こることではないということが分かります。つまり、今回の実験データから26.5%という高い推定値が得られたのは、黒玉の比率が25%という状況から偶然生じた結果というよりは、黒玉の比率がBB弾を準備した当時と比べて高くなっているから生じた結果だろうと考えられるのです（ここでは先に述べた統計学的仮説検定の考え方が用いられていることに注意してください)。

このようにみなさんの実験データからは、長年実験を繰り返しているうちに白玉が減って、黒玉の割合が高くなってきていることが示唆されます。なにかの理由

で白玉の方が無くなりやすいようです。統数研の床は白っぽいので、床に落ちてしまった白玉は見つかりにくいのかもしれません。

果たして現在、水槽の中に白玉と黒玉はそれぞれいくつあるのでしょうか。今回の実験から、水槽中の黒玉の「割合」は以前よりも高くなっていることが分かりました。しかし、玉の「数」は実際に数えてみないと分かりません。玉の数を一つ一つ手で数えるのは大変ですが、文字認識などに使われている画像解析の技術を駆使すれば、全ての白玉と黒玉の個数を自動計測できるかもしれません。統計数理研究所の統計的機械学習研究センター (http://noe.ism.ac.jp/sml-center/research/) の先生に相談してみようと思います。

今年の後片付けで床から見つかった BB 弾は全て白玉 (7 個) でした。