ソフトウェア活用法

ヒストグラムと分散、条件式

2010/6/21 長島雅裕

1. サイコロの目の分布： ヒストグラム

前回、サイコロの目をランダムに出す方法を学んだ。[0,1)の一様乱数を発生させる関数、rand()を使えば、=int(rand()*6)+1 で 1～6 までの整数をランダムに出せることがわかった。

さて、正しく作られたサイコロならば、どの目が出る確率も 1/6 である（場合の数が 6 通りで、どれも等確率なので）。そこで、実際に 1/6 の確率になっているか、図示してみよう。

まず手始めに、サイコロを 6 回ふり、そのヒストグラムを書いてみよう。

1. A1～A6 の 6 つのセルに、上と同じく=int(rand()*6)+1 と書こう（一つだけ書いて、あとはコピーすればよい）。6 回ふれば、原理的にはすべての目が 1 回づつ出る可能性はあるが、たった 6 回では偏りが生じているのがわかるだろう。
2. 次に、ヒストグラムの横軸を明示しておく。つまり、サイコロの目の値、1～6 を書いておく。今回は、C1～C6 に 1～6 の数字を書いておこう（場所はどこでもよい）。
3. 「分析ツール」を使ってヒストグラムを作る。「データ」タブの右端にある「分析」をクリックする
   ※もし分析ツールがなければ、「Office ボタン」を押し、下に出てくる「Excel のオプション」をクリック、「アドイン」を選び、下にある「管理」から「Excel アドイン」を選び「設定」ボタンを押す。出てきたパネルの中の「分析ツール」にチェックを入れ「OK」を押すと使えるようになる。
4. 「ヒストグラム」を選ぶ。「入力範囲」は、サイコロの目の範囲、すなわち A1～A6 である。「データ区間」はさきほど入力したヒストグラムのデータ範囲、C1～C6 である。「出力オプション」ではどこにヒストグラムを出力するか指定できる。ここでは E1 を指定しておこう。これで OK をクリックすると、出力オプションで指定したセルにヒストグラムの数値データが出力される（OK するたびに新しい乱数に変わってしまうので、最初に出した乱数の分布を覚えておき、正しいかチェックしよう）。
5. グラフウィザードを使い、今度は棒グラフで示してみよう。「項目軸ラベルに使用」で「データ区間」のセルが選ばれるようにしておくこと。正しく図がかけたか、数値と照らし合わせてみよう。
6. F9 を押してサイコロを振り直し、上と同様の操作を繰り返してみよう。

6 回ふっただけでは、「確率が 1/6」かどうかなんて確かめられないことがわかったと思う。1/6 の割合で目が出るためには、もっと沢山ふらなければならない。そこで、今度は **60 回ふってみて、そのヒストグラムを書いてみよう**。

課題：ヒストグラムの図を印刷して提出する。

2. ズレの度合い

次に、確率 1/6 からのズレを調べてみよう。

60 回サイコロをふったら、それぞれの目が出る「期待値」は 10 回づつである。これは、（全試行回数）×（その事象が実現する確率）であらわせられる。そこで、それぞれの目が出た回数と、「10」とのズレを見る。

1 の目が出た回数は E2 に出力されていると思う（違う人は、以下の説明も適切に読み替えること）。そこで、その数値の隣(F2)に、=E2-60*(1/6)と書いてみよう（=E2-10 でも同じことである）。これを他のすべての目についても行う（セルの右下をドラッグして、下にコピーすればよい）。これが、期待値からのズレである。

ズレの平均はいくらだろうか？F8 セルに、F2～F7 セルの値の平均を書いてみよう（ヒント：average 関数）。結果はいくらになっただろうか？0になったはずだ。なぜ0になるかはよく考えてみよう。

さて、これでは、どれくらいズレるのかの見通しがつかめない。そこで、ズレを自乗してすべて正の値にしよう。さらに隣の G2 セルに、=F2^2 と書いてみよう。これで、F2 セルの値の自乗が書き込まれたはずだ。これを他のすべての目に付いても実行し、さらに F8 セルに F2～F7 までの平均値を計算してみよう。

最後に、その平均値（ズレの自乗の平均値）の平方根をとってみよう。これが、**確率分布から期待される平均値に対する典型的なゆらぎの大きさ**となる。これを**標準偏差**と呼ぶ（**自乗したものは分散**と呼ばれる）。

60 回ふった場合は、各目が出る期待値が 10 なので、標準偏差はおおよそ√10≒3 程度になるはずだ（なぜ√10 で導けるのかは来週学ぶ）。

3. 標準偏差の試行回数依存性

さて、直感的には、試行回数を増やすほど、まんべんなくそれぞれの目が出現し、ズレの度合いは相対的に小さくなっていくと予想されるだろう。実際そうなっているか調べてみる。

まず、60 回のときの相対的な「誤差」を求めておこう。これは（標準偏差）／（期待値）で見積もられる。なんどかサイコロを振りなおしてみて、大体どれくらいの値になるか調べてみよう。

次に、600 回ふってみよう。期待値は 100 になる。この場合の、誤差を何回か求めてみよう。

結論を先に述べると、だいたい 1/√N 程度になるはずだ。つまり、期待値が 10 ならおよそ 0.33, 100 なら 0.1、もし 10000 ならば、0.01 となる。パーセントで表すと、それぞれ 33%, 10%, 1%となる。従って、誤差 1%で物事を言うならば、サンプルの大きさは 10000 程度ないといけないことがわかる。

4. 練習問題:血液型の分布

日本人の場合、血液型分布はおよそ A:O:B:AB=4:3:2:1 になっていることが知られている(これは民族によってかなり異なる)。では、10 人の集団があったら、それぞれの血液型の人が 4 人、3 人、2 人、1 人となるであろうか?そうはならないだろうということは、今までの結果からわかるだろう。では、少人数集団の場合、だいたいどれくらいのズレが生じるだろうか?

そこで、全体では上のような血液型分布になっている集団の中から、10 人をランダムにピックアップしたとしよう。今回は、ランダムではあるが、期待値が一様ではなく、血液型によって違いがあることに注意する。

まず、エクセルでは文字列よりも数値データが扱いやすいので、A を 1、B を 2、O を 3、AB を 4 と番号をつけておこう。次に、rand 関数で生成される[0,1)一様乱数を用いて、期待値としては 4:2:3:1 になるようにうまく値を変換する。もっとも単純な方法は、出てきた乱数が 0.4 未満なら A(1)、そうではなくて 0.6 未満なら B(2)、さらにそうではなくて 0.9 未満なら O(3)、さらにそうでなければ AB(4)とする。これならば、上記の期待値を実現できるであろう。

1. まず、**10 人分の乱数を発生**させよう。A1 セルに=rand()とやって、[0,1)一様乱数を発生させる。次に、それを 10 人分になるように下にコピーしよう。
2. 次に血液型に変換するため、数値を読み取り、条件によってどの値(1,2,3,4)を出力するかを判別しなければならない。そのために使えるのが **IF 文**である。

   IF 構文は次のようになっている。=IF(A,a,b)、この意味は、「A が真なら a を、偽なら b を実行せよ」ということである。試しに B1 セルに次のように書いてみよう。=IF(A1<0.4,1,0)、すると、A1 の値が 0.4 未満なら B1 には 1 が表示され、0.4 以上ならば 0 が表示されるはずである。いま、0 は「A 以外」に対応する。

   「A 以外」を具体的にしていこう。IF 文の中に IF 文を入れることも可能である。そこで、以下のようにすれば、4つの血液型に分けられることになる、=IF(A1<0.4,1,IF(A1<0.6,2,IF(A1<0.9,3,4)))。これによって、A1 の値が 0.4 未満なら 1,0.4 以上 0.6 未満なら 2,0.6 以上 0.9 未満なら 3、0.9 以上なら 4 が出力されることになる。この 1,2,3,4 が A,B,O,AB に対応していると読めばよい。
3. では、これらのヒストグラムを作ってみよう。やり方は3と同様であるのでやってみてほしい。ただし、図にする際は、横軸が数字ではわかりづらいので、どこかに(例えば D1～D4、どこでもよい)A,B,O,AB と書いておき、それを横軸に使用するようにしよう。

   これも何度も乱数を振りなおして図を書いてみて欲しい。10 人の集団では、血液型分布に偏りの出る場合がかなりあることがわかるだろう。
4. 次に、もっとサンプルを増やして、**100 人でやってみよう**。するとどうなるか。10 人のときよりも 4:2:3:1 に近づき、偏りの目立つ場合が少なくなっていることがわかると思う。

この場合も、典型的には√N 人程度の「誤差」がある。相対誤差にすれば、√N/N=1/√N となるので、サンプルが大きいほうがより期待値に近づくのである。

逆に言えば、少人数のサンプルから全体を推測することは極めて危険であることがわかる。「歴代首相には○型が多いから○型はリーダー向き」のような主張の無意味さを読み取って欲しい[1]。また、高々数百人のサンプルで得た視聴率では 0.1 ポイントの違いは完全に誤差の範囲で比較する意味がないこともわかるだろう。

課題:100 人の集団に対するヒストグラムの図を印刷して提出する。

5. 課題(再掲)

1. サイコロを 60 回ふったときのヒストグラムのグラフ
2. ランダムに選んだ 100 人の血液型分布(ヒストグラム)のグラフ

これらを印刷して提出せよ(番号・名前を忘れずに)。

---

1 1 万人以上のサンプルで慎重に調べた結果、血液型と性格の間には強い相関(血液型から性格を当てる、あるいはその逆が可能なくらいの強さ)はないということが既に明らかになっている。つまり、仮になんらかの相関があったとしても、それは 1%程度のごくわずかな違いしかない、ということであり、日常生活のレベルでは血液型と性格は無関係であると言ってよい。