

仮説検定の考え方～導入の一例～

上野 隆彦・聖マリアンナ医科大学
e-mail: t2ueno@marianna-u.ac.jp

序

本稿は高等学校数学Ⅰに含まれる予定の「仮説検定の考え方」とは関係ない。本年度、筆者が講義（立教大学理学部数学科（2年生以上）「確率と統計Ⅰ」および聖マリアンナ医科大学医学部医学科（3年生）「医学情報学」）で行った仮説検定の導入部分の紹介である。

仮説検定は「実験的にとったデータをもとに意思決定をする方法論」だが、昨年度までの講義では学生の十分な理解は得られていないように感じていた。仮説検定は「帰無仮説に基づいて確率分布を検討し棄却域を定め、それに基づいて結果を解釈する」のだが、理解の不十分な学生はデータからの統計量の計算方法とそれをP値に変換する方法までを覚えて「さて結論は？」となっている印象であった。「データをもとに意思決定をする方法論」を教えたはずなのに「データからの統計量の計算方法とP値への変換方法」を覚えて意思決定ができないのでは、仮説検定を学んだとはいえない。そこで

1. 学生に質問し考えさせる
2. 簡単な例で徐々に内容を深める

という2点を意識して講義しようと試みた。まず確率分布を用いずに適当に棄却域を定め、データをもとに結果を解釈するという検定の本質的な部分を紹介し、その後に確率分布を用いて適当な棄却域の意味を考える、というのが本年度の仮説検定の導入部分の流れである。

講義内容（概要）

以下、学生に対する質問はゴシックで表記する。

問。「普通の薬は2人のうち1人に効果がある」とします。新しい薬を作った人は「新しい薬はもっと多くの人に効果がある」と思っています。これを確かめるにはどうしたらいいでしょう？意見のある人は手を挙げて下さい。薬の効果で考えづらければ

普通のコイン：表、裏が出る確率は0.5

新しいコイン：表が出る確率は0.5より大きいと置き換えて下さい。

手を挙げるものはいなかったので「仮に科学的に歪みを分析して、このコインの表が出る確率は0.62だ、といわれても、本当かはわからないので最後は実験して確かめるよりない」と誘導した。

問。10回コインを投げて表が出やすいかを判定することにします。何回以上表が出たら「このコインは表が出やすい」とあなただったら判定しますか？答えづらいと感じる人もいるかもしれませんが、正解があるというものではありません。普通のコインであれば表が出る回数は“5回くらい”。10回中で5回くらいより多いと思う回数はどこからでしょう？

この回答をx, 以下の問のy, zと共に受講者に回答させた。この回答については考察で紹介する。

問。コインの表が出やすいと判定するのは
100回コインを投げたとき、y回以上表が出たら。
1000回コインを投げたとき、z回以上表が出たら。

学生達に「これでコインの表が出やすいか判定する方法が3つ出来ました」と伝えたところ戸惑って

いたように感じたので、その理由を学生に尋ねた。

A. 思いつきで決めた基準で判定してよいのか。

B. 人によって判定法は異なるのではないか。

学生達の疑問は上記 A, B に集約された。A については後で考えると回答し、B については「コインの判定結果と共に自分の判定法を公開すればよい。その結果をどうとらえるかは、それを見た人が判断すること」と回答した。

また「コインをどう投げて結果をどう記録したか」も公開すること、つまり投げ方、記録の仕方も大事と教えた。記録がデタラメでも、投げ方で表、裏を操作できても判定法を決めた意味がないと強調した。

ここで検定を紹介した。「コインを 10 回投げて 8 回以上表が出たら表が出やすいコインと判定する」という判定法のことを検定という。ここで検定の判断基準（この場合 8 回）のことを検定の有意水準と呼ぶ、と説明した。

問. 10 回中 7 回表が出たとき、そのコインは表が出やすいと判定されない。この解釈として「これは普通のコイン（表の出る確率 0.5）である」は妥当か？

この判定法は「表が出やすいことを判定するためのもの」である。その判定基準に満たなかったコインは、あくまでも（この判定法で、今回の実験結果から）「表がでやすいと判定されなかったコイン」となる。「（この判定法で）表が出やすいと判定されない」＝「表の出る確率が 0.5 の普通のコイン」は成り立たない点に注意した。また仮に「10 回中 5 回表が出た」としても同じ解釈となることを注意した。

問. 先程の 3 つの判定法を参考に「裏が出やすい」という判定法を考えよ。また「表または裏が出やすい」という判定法も考えよ。

続けて検定と信頼区間の関係を次のように紹介した。「コインを 10 回投げて 8 回以上表が出たら表が出やすいコインと判定する」場合、「5 回くらい」＝「3 回～7 回」と考えていることになる。つまり 10 回投げるとき ± 2 回の誤差を許容している。そうすると「表が 8 回でた」という結果は ± 2 回を誤差とし

て「表は 6 回～10 回出だろう」、つまり、表は 5 回より多く出そうと解釈できる。同様に「表が 7 回出た」という結果は「表は 5 回～9 回出だろう」、つまりは表の出る回数が 5 回を上回るとまではいえないと解釈される。得られた結果から、観測される結果の存在しそうな範囲を推定することを区間推定という。特に検定と対応する形で考えられる区間を信頼区間という。

次に「人によって異なる、思いつきで決めた判定法」の性能を考えた。この判定法を用いたときに「普通のコインを表の出やすいコインと判定してしまう誤り」を検定の第 1 種の過誤、また「表の出やすいコインを表が出やすいと判定できない誤り」を検定の第 2 種の過誤ということを説明した。

問. 上記の検定で第 1 種の過誤が起きる確率を求めよ（式を答えよ）。

$$\sum_{k=x}^{10} {}_{10}C_k(0.5)^{10}, \sum_{k=y}^{100} {}_{100}C_k(0.5)^{100}, \sum_{k=z}^{1000} {}_{1000}C_k(0.5)^{1000}$$

を答えてもらい、いくつか計算例を示した。また「有意水準」＝「第 1 種の過誤確率」であることを説明した。更に第 1 種の過誤確率をどのくらい小さくしたいかを二項分布を用いて検討しておけば全員が同じ有意水準を選べることを説明した。ここまでの仮説検定の考え方の導入である。

この後に、ここまでの内容を仮説検定の手順（1. 帰無仮説、対立仮説を設定、2. 帰無仮説のもとでの確率分布の検討、3. 実験結果の判断）に当てはめて書き直した。更に続く講義では分布を変えて、同様のことを行うという形で仮説検定を紹介していった。また平均値の検定、平均値の差の検定は分散未知の場合に限るなどの配慮をした。

考察

この導入のメリットは、仮説検定において理解しづらいと思われる以下の点を具体的なイメージと共に学修できることにあると考えている。

1. 有意水準：検定の判断基準で、第 1 種の過誤

を制御.

2. 実験結果を見てから有意水準を設定することがダメな理由.
3. 有意差なし≠差がない.
4. 信頼区間と検定の関係.
5. データをどのようにとるかも考える必要がある.

5 を説明している初心者向けの統計学の本を見たことはないが、仮説検定は「データが間違いやバイアスを含まないようにとられてる」ことが前提であることは、しっかり教えておく必要があると感じている.

次の表は「10 回中 x 回以上表が出たら、表が出やすいと判定する」学生達の判断基準が、試行回数の増加でどう変化するかをまとめたものである.

x	同	小	大	他	計
5	5	0	0	0	5
6	12	1	2	24	39
7	13	16	6	17	52
8	1	16	0	8	25
9	0	5	0	2	7
計	31	38	8	51	128

表の「同」は割合が変化しないもの (例. $x = 7$, $y = 70$, $z = 700$), 「小」は割合が小さくなるもの (例. $x = 7$, $y = 65$, $z = 600$), 「大」は割合が大きくなるもの (例. $x = 7$, $y = 75$, $z = 800$), 「他」はそれ以外のもの (例. $x = 7$, $y = 68$, $z = 680$) と分類した. ただし例外として $x = 6$ の「他」には $x = 6$, $y = 51$, $z = 501$ と回答した 18 名 (全体の 14 %) を含んでいる. この回答は割合が小さくなっていくという意味では小に分類されるべきであるが, 「1 回でも表が多ければ表が出やすいと判定する」というデータのバラツキを許容しないという主旨の回答と考えたため「他」に分類することにした. これは「小」に分類される回答数=大数の法則を何となく知っている学生数としたかったためである.

「小」に分類された学生は 38 名 (全体の 30 %) である. $x = 8, 9$ と回答した 32 名のうち 21 名 (66 %), $x = 7$ と回答した 52 名のうち 16 名 (31 %), $x = 6$ と回答した 39 名のうち 1 名 (3 %) が「小」に分類された. これらの学生は大数の法則を知識また

は経験的に知っていると考えられるが, その多くが $x = 8, 9$ を選択している点は興味深い. 「大」は試行回数が増えるとバラツキも増えるという考え方の学生で 8 名 (全体の 6 %) が該当した. 「他」に分類された多くは, y, z を $y/100, z/1000$ が殆ど同じくらいの割合となるよう選択しているので, 「同」に分類した学生達に近い感覚があるのではないだろうか. [1] において割合を一定, 標本の大きさを変化させたとき 48 人中 22 人 (45.8 %) が, 標本の大きさを考慮しなかった旨等を報告している. 調査方法は異なるが, 今回の調査結果とも類似しているのではないだろうか.

本年度, このような導入を試みた感想は次のようになる. まず導入から二項分布を用いた検定までは学生の様子から察するに理解はしやすかったのではないかと感じている. しかし, その他の検定の説明では, 欠席・遅刻や他の勉強をしている学生が目立ち始めた. 初心者向けの講義では, 学生に主体的に考えてもらう工夫が必要であるように感じている. そのため次年度は学生の出席, 学生同士で議論させることを重視した講義内容にしたいと考えている.

皆様にとって参考になる部分が多少でも本稿に含まれていれば幸甚である.

参考文献

[1] Evans, J. St. B. T. and Dusoir, A.E. "Proportionality and sample size as factors in intuitive statistical judgement", Acta Psychologica, 41, 129-137 (1977).