仮説検定

2018 年 前期講義資料 （作成者：大塚 淳）

1 Neyman-Pearson の仮説検定

ネイマン・ピアソンの仮説検定では、帰無仮説 H_0 と対立仮説 H_1 があるとき、与えられたデータをもとに、帰無仮説を棄却する（対立仮説を採択する）ことができるかどうかを判定する。ここでのポイントは、仮説が蓋然的である限り、そうした判定は常に誤る可能性があるということである。NP はこれを受け入れたうえで、データをもとに検定を行ったとき、それが平均してどれだけ誤るのかを計算し、それをもとに仮説の棄却・採択を行う。

授業でも述べるように、ここには二種類の誤りの可能性がある：

第 1 種の誤り H_0 が真であるのに、それを誤って棄却してしまう（偽陽性）
第 2 種の誤り H_0 が偽であるのに、それを棄却しそこなう（偽陰性）

どのようなテストにも、この二つの誤りの可能性はつきまとう。テストが第 1 種の誤りを犯す確率を α、第 2 種の誤りを犯す確率を β とする。後で見るように、多くの場合この二つはトレードオフの関係にある。そこで NP はまず α を低い値に定め、その基準を満たすテストの中から、もっとも低い β を持つテスト選ぶ、というようにしてこの問題を解決する。以下では、簡単な例にもとづいて、検定のアイデアと、α, β の計算方法を示す。

2 有意水準と第 1 種の誤り

ある国で発行されているコインは、淵が丸まっているので、立てようとしてもどちらかに倒れてしまう。肉眼では見分けはつかないが、実はこの丸み処理には 2 種類あって、最近発行されたものは 1/4 の確率で表を上にして倒れるが、旧式のものは 3/4 で表がでる。旧式のものは今やレアで、コレクター市場ではかなりの値が付くそうだ。さて、あなたの叔父さんがこの国に旅行して、おみやげとしてこの旧式コインをくれた。しかしあなたはどうも確信がもてない。これは本当にレアな旧式コインなのか？

そこであなたは、そのコインを 10 回立てる実験を行って判断することにした。実は旧式ではないという帰無仮説は $H_0 : p = 0.25$、正真正銘のレアものであるという対立仮説は $H_0 : p = 0.75$ である。もし H_0 が正しかったら、10 回のうち何回表がでると期待できるだろうか？表の回数を X とし

たとき、$P(X=x)$ の確率は、$p=0.25, n=10$ の二項分布から次のように求まる。

$$P(X=x) = {}_{10}\mathrm{C}_x (0.25)^x (0.75)^{10-x}$$

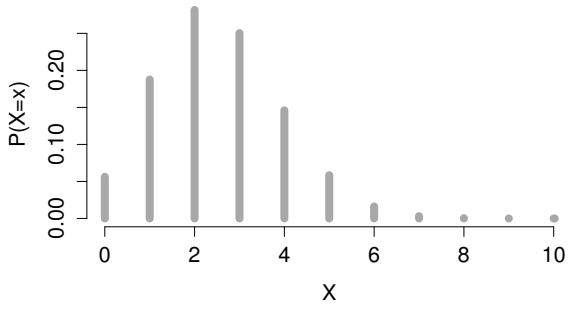

図 1　H_0 のもとでの $P(X=x)$

図から明らかなように、帰無仮説は右に行くほど（表が多いほど）起こりにくいと予想する。そこで、このように予測から外れる結果が実際に出た際に帰無仮説を棄却する、という方針が考えられる。では、何回以上表が出たら「予測が外れた」と判断すべきだろうか？同じ分布から：

1. $P(X=10) = 0.00000095$
2. $P(X \geq 6) \sim 0.020$
3. $P(X \geq 5) \sim 0.078$

これをもとに、3 つのテストを考えてみる

i. いま、1 のときだけ、つまり全部が表だったときだけ H_0 を棄却すると決めたとしよう。このとき実は H_0 が真なのに棄却されてしまう、つまりこのテストが第 1 種の誤りを犯すのは、せいぜい 100 万回に 1 回ということになる。
ii. 次に、2 のとき、つまり 6 回以上表だったら H_0 を棄却するテストを考えると、これが誤って H_0 を棄却する確率は 2% 弱である。
iii. 最後に、半分以上表が出たら H_0 を棄却するテストを考えると、3 よりこの第 1 種の誤りの確率は 8% 弱となる

i のテストは極めて慎重だが、ほとんど帰無仮説を棄却できず感度が悪い。一方 iii を用いれば、比較的わずかなズレでも H_0 を棄却できるが、しかし誤って棄却してしまうリスクもそれなりにある。よって上の 1,2,3 の確率は、i, ii, iii のテストによる帰無仮説の棄却しやすさと同時に、それが正しい時に誤って棄却してしまう第 1 種の誤りの確率 α を与える。これを**有意水準** (significance level) といい、有意水準以下になるような（つまり「予測から外れた」）データの範囲を**棄却域** (critical region)

という。仮説検定ではまず有意水準を定め、実際の結果がその棄却域に入っていたら帰無仮説を棄却する。有意水準の設定は分野や文脈によって異なるが、慣習的に 5% とされることが多い。ここでこの基準を採用するなら、コインが 6 回以上表がでたとき、H_0 を棄却すれば良いことになる。

α が第 1 種の誤りの確率を示しているとすれば、$1-\alpha$ はテストの厳しさ、つまり H_1 が偽であるときそれをどれくらいの確率で見抜けるか、を表している。これをテストの**サイズ**という。サイズが高いテストほど（有意水準が低いほど）、H_0 の棄却／ H_1 の採択が偶然の産物である可能性は低くなるわけで、その意味でその結果はより有意である (significant) といえる。

3　検出力と第 2 種の誤り

次に、第 2 種の誤り、つまり対立仮説 H_1 が真であるときに、帰無仮説 H_0 を棄却しそこねる確率を考えてみる。H_1 が真であるとは、つまり $P(X=x)$ は本当は $p=0.75, n=10$ の二項分布に従う、つまり真実は図 1 でなく以下の図 2 のようである、ということだ。

$$P(X=x) = {}_{10}\mathrm{C}_x (0.75)^x (0.25)^{10-x}$$

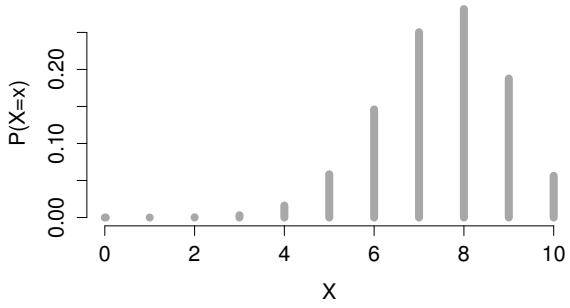

図 2　H_1 のもとでの $P(X=x)$

もちろん我々には、図 1 と 2 のどちらが真実なのかはわからず（わかっていたら検定などする必要がない)、粛々と上で選んだテストを行うだけである。さて、そのテストを行ったところ、帰無仮説を棄却できなかったとしよう。このとき気になるのは、この結論がどのくらい信頼のおけるものなのか、ということだ。つまり、実は H_0 は偽（H_1 が真）なのにそれを見過ごしてしまっている確率、テストが第 2 種の誤りを犯す確率を知りたい。これは図 2 の分布をもとに、以下のように計算できる。

1'. $P(X<10) = 0.944 \sim 100$ 回に 95 回強
2'. $P(X<6) = 0.078 \sim 100$ 回に 8 回弱
3'. $P(X<5) = 0.020 \sim 100$ 回に 2 回

これがそれぞれ上で考えた i, ii, iii のテストが第 2 種の誤りを犯す確率 β となる。その理由を考えてみよう。その際、以下がヒントとなる。

- 第 2 種の誤り（H_0 が偽/H_1 が真であるときに H_0 を棄却しない）を問題にしているのだから、真実として想定すべき確率分布は図_____である
- i, ii, iii のテストが H_0 を棄却しないのはそれぞれ_____, _____, _____のときである
- よって第 2 種の誤りの確率はそれぞれ_____, _____, _____である

第 2 種の誤りとは、本当は H_1 が正しいのに、それを見逃してしまう、ということだ（偽陰性）。よってその確率 β が低いほど、感度の高いテストだと言える。ここから、$1-\beta$ を、そのテストの**検出力** (power) という。通常、我々の興味ある仮説が対立仮説として設定されるので、検出力はテストが興味ある結果をどれだけしっかり検知できるか、ということを表している。

このように、Neyman-Pearson の検定理論では蓋然的な仮説をテストするに当たり 2 種類の誤りを考慮する。どちらをより重視すべきか？ NP では、このうち第 1 種の誤りの方を重視する。つまり、帰無仮説を棄却するには、それを採択するよりもより慎重であるべきとされる。よって検定では、まず有意水準 α を定め、その基準をクリアするテストのうち最も β が低いような（検出力が高いような）ものを選ぶ、というのが一般的な戦略である。

4　仮説検定まとめ

■ポイント

- テストは 2 つの指標、α と β、ないしはサイズ $(1-\alpha)$ と検出力 $(1-\beta)$ で特徴づけられる
  - α（有意水準）は H_0 が真のときそれを誤って退ける第 1 種の誤りの確率を表す
  - β は H_1 が真のときそれを見逃す第 2 種の誤りの確率を表す
- NP では、一定の α の基準を満たすテストのうち、β を最小にする（検出力を最大にする）ようなテストが推奨される

■検定のステップ

1. 帰無仮説と対立仮説を決める → 両仮説のもとでの分布がそれぞれ定まる
2. どれくらいの「ありえなさ」だったら帰無仮説を棄却するか（有意水準）を定める
   → これが同時にテストの第 1 種の誤りの確率 α となる
3. 有意水準と帰無仮説 H_0 のもとでの分布から、棄却域が定まる
4. 実験を行い、結果が棄却域内であれば、帰無仮説を棄却する
5. 棄却できなかった場合、対立仮説 H_1 のもとでの分布をもとに第 2 種の誤りの確率を求める

5　サンプルサイズの重要性

ベイズ推論のところで我々は、観測を重ねるほど結論が精確になっていく過程を見た（swamping of priors）。仮説検定においても、サンプルサイズは重要な役割を持つ。しかしその意味合いは異なる。ベイズ推論では大サンプルは結論の精確さに寄与するのに対し、検定ではテストの精確さに関わる。これを見るために、上と同じ状況だが、コインの試行回数を 10 でなく 20 回に増やした実験を考えてみよう。

$n = 20$ としたときの H_0, H_0 のもとでのコインの表がでる回数 X の確率分布は以下のとおり

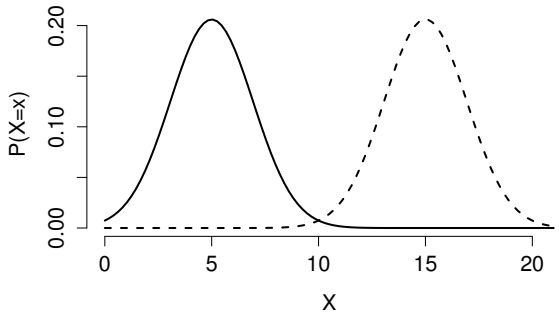

図 3　$n = 20$ としたときの $P(X = x)$。ただし見やすいようにプロットを曲線化し、H_0 のもとでの分布を実線、H_1 を点線で表している。

ここで有意水準を 5% とするテストを考えると、帰無仮説 H_0 のもとでの棄却域は $P(X \geq 9) = 0.041$ より、そのテストは表が 9 回以上でたとき、H_0 を棄却することになる。つまり、表がでるのが半分以下でも帰無仮説を棄却できるという意味で、$n = 10$ のときの同水準のテストよりも感度が高くなっている。

また、そのテストが仮説を棄却できなかった場合、つまり表が 8 回未満しかでなかった場合を考えよう。対立仮説 H_1 のもとで、$P(X < 9) = 0.00094$。したがって仮に帰無仮説が偽であったときに、そのテストがそれを見逃してしまう確率は 1000 回に 1 回程度である。

我々は図 1 と図 2 で、テストの感度（サイズ）と検出力がトレードオフの関係にあることを見た。棄却域を右にずらすほど、テストの感度・サイズ（図 1 で棄却域の左側にくる確率）は上がるが、検出力（図 2 で棄却域の右側にくる確率）は下がる。しかしサンプルサイズを増やすことができれば、H_0, H_1 の分布の分散を減らすことで、テストの感度と検出力を共に上げることができる。こうした意味で、検定理論においても沢山のデータを集めることはより確実な推論を可能にするのである。

練習問題

1. 2 節で述べた実験について、有意水準を 10% としたときの棄却域を図 1 より求めよ。またそのときの検出力を求めよ。
2. 上のテストで、めでたく帰無仮説を棄却できたとしよう。そのとき、この結果が意味することとして正しいものをすべて選べ
   （a）これが旧式のレアコインである確率は 10% 以下だ
   （b）これが旧式のレアコインである確率は 90% 以上だ
   （c）もしこれがレアコインでなかったとしたら、こんな結果は 10 回に 1 回もでないだろう
   （d）もしこれがレアコインでなかったとしても、10 回に 9 回はこのような結果がでただろう
3. ベイズ主義者だったら、叔父さんがくれたコインがレアかどうかを決めるためにどうするだろうか。また、そのときに必要となる情報はなんだろうか。