仮説検定

1　ネイマン・ピアソン流の考え方

統計的仮説検定とは、１つあるいは複数の確率データの分布に関する命題について、真偽の判断を下そうとするもの。命題を棄却できるかどうかの規則、手順を定めることが目的である。分布に関する命題を仮説とよび、帰無仮説 (null hypothesis) とは、もう一方の対峙する対立仮説 (alternate hypothesis) を支持して、棄却されるあるいは破棄されることを意図して立てられるもの。(参照: 8 章　検定　115 ページ「統計学演習」村上/安田、培風館)

仮説の検定である判断を下すならば、2 種類の誤りが必ず生じる。第 1 種の過誤と第 2 種の過誤である。ネイマン・ピアソン流の考えでは第 1 種の過誤の確率に対して上限（有意水準とよぶ）を決めておき、第 2 種の過誤の確率を最小とする、あるいは過誤を犯さない確率（検出力とよび）を最大にするような判断の方法を求める。

2　尤度比検定

[**ネイマン・ピアソンの補題**]　帰無仮説 $H_0 : f = f_0$, 対立仮説 $H_1 : f = f_1$ において、有意水準（サイズともいう）α とするとき、第 2 種の過誤の確率を最小とする検定は次で与えられる：

$$C = \{x = (x_1, x_2, \cdots, x_n) : \frac{\Pi_i f_1(x_i)}{\Pi_i f_0(x_i)} > k\}$$

の形であり、ただし k は $\alpha = \mathbb{P}(X \in C|H_0) = \int_C f_0(x)dx$ を満たすとする。簡略化のために $(f_0(x) = \Pi_i f_0(x_i), dx = dx_1 dx_2 \cdots dx_n)$ としている。C は棄却域であり、帰無仮説と対立仮説を比較 (尤度比) して分子にある対立仮説のほうが可能性が高ければ帰無仮説を棄却するという意図である。

標準的な仮説検定問題では、どのような仮説（帰無仮説、対立仮説）について、どんな検定統計量、棄却域が用いられるかという問いに対しては、このネイマン・ピアソンの補題によって解いた結果を表にまとめられている。(参照: 8 章　検定　116 ページから 118 ページまでの一覧表「統計学演習」村上/安田、培風館)

問１（本文資料 28 ページ）正規分布 $N(\mu, \sigma^2)$ における $\sigma^2 = 1$ の標本 $X_1, \cdots, X_n$ について $H_0 : \mu = 5$ とし、$H_1 : \mu = 6$ に対し、サイズ 0.05 の検定をおこなう。(1) $x = (5.1, 5.5, 4.9, 5.3)$ ではどうなるか？ (2) 同じ標本が得られたとするが、帰無仮説と対立仮説を逆にしたら、結論はどう変わるか？

（注意）この問題は、優先権をもつ帰無仮説の選択が重要であることを示している。

3　適合度検定

いま n 回の独立試行を行ったとき、可能な k 種類の事象が結果データ：$(x_1, \cdots, x_k), \sum_j x_j = n$ として得られたものとする。また結果 i の起こる確率は p_i であるとする。結果データと起こる確率を比較して、実際にこのような生起確率にもとづく試行であったのだろうかを検定する。

[**ピアソンのカイ 2 乗統計量**] 適合度検定に用いる統計量：近似的に

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

は自由度 $k-1$ のカイ 2 乗分布に従う。ここで可能な事象 $i = 1, 2, \cdots, k$、観測値 $o_i = x_i$、期待値 $e_i = np_i$。

問2（本文資料 37 ページ）適合度検定 $H_0 : p_i = p_i(\theta), \theta \in \Theta_0$ $H_1 : p_i unrestricted$ において尤度比 $L_x(H_0, H_1) = \dfrac{L_x(H_1)}{L_x(H_0)} = \dfrac{\sup_p f(x|p)}{f(x|p = (p_i(\theta)))}$ を考える。

(1) $2 \log L_x(H_0, H_1) = 2 \sum_{i=1}^{k} x_i \log \left(\hat{p_i} / p_i(\hat{\theta}) \right)$ ここで $\hat{p_i} = x_i/n$, $\hat{\theta}$ は H_0 のもとでの θ の MLE.

(2) $2 \log L_x(H_0, H_1) \fallingdotseq \sum_{i=1}^{k} \dfrac{(o_i - e_i)^2}{e_i}$ を導け、ただし $o_i = x_i$, $e_i = np_i$.

(ヒント) 対数関数 $\log(1 + x) = x - x^2/2 + x^3/3 - \cdots$ と近似され、$\sum_i (o_i - e_i) = 0$.

（本文資料 38 ページ）属性が 2 つで分類されている場合を考える。m 行 n 列の長方形にデータを当てはめて、この枠組みの 1 つのセル (i, j) にはデータ X_{ij} があるものとする。縦および横の計、さらに総計は次で表す。

$$X_{i\cdot} = \sum_{j=1}^{n} X_{ij}, \quad X_{\cdot j} = \sum_{i=1}^{m} X_{ij}, \quad X_{\cdot\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}.$$

4 分布同一性の検定

いま各行における分布は同じであるという仮説を検定する。ある確率 $\{p_j\}$ が与えられたとする。

帰無仮説　$H_0 : p_{ij} = p_j, \forall i$, 各 $j = 1, 2, \cdots n$.
対立仮説　$H_1 : p_{ij}$ 制約なし.

同様に計算すると $2 \log L_x(H_0, H_1) \fallingdotseq \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij}$ を得る。したがって、帰無仮説のもとでは、自由度 $(n-1)(m-1)$ のカイ 2 乗分布にしたがうとして、検定を行う。

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

問3（本文資料 39 ページ）アスピリンを服用している集団には心臓発作を起こすことがあると危惧されている。この関係を調べるためにプラシーボ（偽薬）をもちいて調査した。次のデータを得たと仮定し、この関係を検定せよ。

	心臓発作あり	なし
アスピリン	104	10,933
プラシーボ	189	10,845

5 独立性の検定

つぎに独立性の検定を考える。2 つの属性の間には従属関係がないことを検定する。データがあるの行と列のセルに含まれることが独立となるかどうかを調べる。この表は分割表とよばれる。

帰無仮説　$H_0 : p_{ij} = p_i q_j, 0 \le p_i, q_j \le 1, \sum_i p_i = 1, \sum_j q_j = 1.$
対立仮説　$H_1 : p_{ij}$arbitrary$, 0 \le p_{ij} \le 1, \sum_{i,j} p_{ij} = 1.$

問4（本文資料 40 ページ）ある人はエレベータのなかで、ハンカチをわざと落とし、隣り合わせた人物が果たして親切に拾い上げてくれるかどうか、性別の違いによる調査結果を得た。

	拾ってくれた	くれない
男性	370	950
女性	300	1,003

このデータでは、拾ってくれるかどうかは性別には因らないといえるだろうか？

問5（参照: 8 章　検定　133 ページ「統計学演習」村上/安田、培風館）2 行 2 列の分割表では、

	B_1	B_2	計
A_1	a	b	$a+b$
A_2	c	d	$c+d$
計	$a+c$	$b+d$	n

つまり、対応 $(i,j=1,2)$ が $o_{ij} => a,b,c,d$; $o_{i\cdot} => a+b, c+d$; $o_{\cdot j} => a+c, b+d$; $n=a+b+c+d$ である。このとき

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

が成立することを確かめよ。

(補足) 条件付き最大最小問題：よく用いられる解法のひとつには、ラグランジュ法 (Lagrangian method) がある。

問題：与えられた集合 X, n 変数の実数値関数 $f(x), x=(x_1,x_2,\cdots,x_n)$, n 変数の m 次元ベクトル関数 $h(x)$、m 次元ベクトル b について $\min(\max)$ $f(x)$, subject to $h(x)=b$, $x\in X$ を解け。

いまベクトル (ラグランジュ乗数) として、λ を導入し、ラグランジュ関数

$$L(x,\lambda) = f(x) - \lambda^\top (h(x)-b)$$

このような条件がない場合の関数 $L(x,\lambda)$ の最小化に帰着される。偏微分を計算し、$n+1$ 個の連立方程式 $\dfrac{\partial L}{\partial \lambda}=0, \dfrac{\partial L}{\partial x_i}=0, i=1,2,\cdots,n$ の意味である。これは必要条件であり、この連立方程式の解の中に ” 実際の解 ” が含まれている。十分条件としては、2 階の微分係数を計算しなければならない。

例題：2 つの変数 $x_i, i=1,2$ において、$a_1x_1+a_2x_2=b, x_1,x_2>0$ のとき、$x_1^2+x_2^2$ を最小化せよ。

(解) $L=x_1^2+x_2^2-\lambda(a_1x_1+a_2x_2-b)$ で、連立微分方程式は $2x_1-\lambda a_1=0, 2x_2-\lambda a_2=0$ $a_1x_1+a_2x_2=b$. 前式 2 つから $x_i=\lambda a_i/2, i=1,2$. これを 3 番目の式に代入すると $\lambda = 2b/(a_1^2+a_2^2)$. 実際 $\dfrac{\partial^2 L}{\partial x_i^2}>0$ $\dfrac{\partial^2 L}{\partial x_1 \partial x_2}=0$ から最小値になっていることがわかる。