誕生月と野球選手の関係

今年のプロ野球選手の誕生月別の人数の表を表示する。(資料提供：井原俊輔先生 (名古屋大名誉教授))

	1月	2月	3月	4月	5月	6月
中日	4	2	4	5	8	6
巨人	3	3	2	15	10	4
阪神	3	2	0	7	3	9
広島	3	1	4	13	6	7
ヤクルト	7	0	3	6	5	10
横浜	3	0	1	9	7	9
合計	23	8	14	55	39	45

	7月	8月	9月	10月	11月	12月	合計
中日	9	6	7	2	3	4	60
巨人	6	0	5	8	3	3	62
阪神	6	4	8	8	6	4	60
広島	4	4	6	8	4	5	65
ヤクルト	4	5	6	6	4	7	63
横浜	9	7	3	4	5	6	63
合計	38	26	35	36	25	29	373

早生まれは、プロ野球選手になることが不利なように見える。しかし、たまたま早生まれの人が少ないだけなのかも知れない。何かの客観的な基準が無いと「プ

ロ野球選手になることが不利」であるかどうか議論を行っても、水掛け論になってしまって、まともな基準で議論に決着が付かない。
このような状況において、判定を下すための客観的な基準を提供する手法が、今日説明する「仮説検定」である。
余談だが、スポーツの種目によって早生まれがプロになるために不利になるものとそうでないものがある。
http://www.volleyball.gr.jp/hayaumare.htm
を参照。

単純な仮説検定の例

実際の例は若干複雑で説明に都合が悪いので、非現実的だが単純で説明に都合がいい例を今後考える。
硬貨を振って賭けをするときに、硬貨の裏と表が出る確率に偏りがあると都合が悪い。そこで「硬貨を振って表が出る確率が 0.5 であるか」という仮説を検定することを考える。
まず硬貨が出る確率は表が出る確率 θ で特徴付けられている。これは分布を特徴付ける母数 (パラメータ) である。一般に仮説検定では、確率分布が有限個の母数で特徴付けられると仮定して、母数の関する条件の

真偽を検定することが多い。硬貨の例では $\theta = 0.5$ という条件を検定している。
硬貨を n 回振って表が出たか裏が出たか記録する。仮説検定のときも n を標本の大きさと呼ぶ。確率変数 X_i を i 回目に表が出たときに 1、裏なら 0 になる確率変数とする。X_i の実現値を標本と呼ぶ。表が出た頻度を

$$T_n = \frac{X_1 + \cdots + X_n}{n}$$

とする。T_n は確率変数である。
T_n が 0.5 よりもあまりにも離れているときには、仮説「$\theta = 0.5$」は正しくないと思われる。そこで T_n がどの範囲にあれば仮説を正しくないと判断すればよいのか？なお、仮説が正しくないと判断することを**「仮説を棄却する」**とよぶ。棄却しないことを**「採択する」**と呼ぶ。
$|T_n - 0.5| \geq c$ となっていれば仮説「$\theta = 0.5$」を棄却することにして、値 c をどのように決めたらいいか考察する。標準的な考え方として、仮説「$\theta = 0.5$」が正しかった場合に仮説を棄却してしまう確率 α が十分小さくなるように c を設定する。この確率 α を**有意水準**と呼ぶ。すなわち

$$P_{\theta=0.5}(|T_n - 0.5| \geq c) = \alpha$$

となるように c を定める。

ここで仮説が採択されたからといって、仮説が必ず正しいことを意味しているわけではない。θ が 0.5 とほんの少し違っても頻度 T_n が 0.5 に近くなることはある。
さていま考えている場合の c の決定であるが

$$\begin{aligned} & P_{\theta=0.5}(|T_n - 0.5| \geq c) \\ = \quad & 2P_{\theta=0.5}(T_n \leq 0.5 - c) \end{aligned}$$

であるから、$P_{\theta=0.5}(T_n \leq 0.5 - c)$ を計算する。n 回の硬貨振った結果は 2^n 種類あり得るが、そのうち表がでた回数が $n(0.5 - c)$ 以下である場合の数は

$$\sum_{0 \leq i \leq n(0.5-c)} \binom{n}{i}$$

で与えられる。また、上記の場合の数の中の一つの場合が生じる確率は $1/2^n$ である。したがって

$$P_{\theta=0.5}(T_n \leq 0.5 - c) = \frac{1}{2^n} \sum_{0 \leq i \leq n(0.5-c)} \binom{n}{i}$$

である。この式を一生懸命計算して c を定めることは可能だが、式をみてもなんだかよくわからない。

正規分布による分布の近似

硬貨の表裏の分布は $\theta = 0.5$ のときに期待値 $E_\theta[X_1] = 0.5$, 分散 $V_\theta[X_1] = 0.25$ になることが簡単な計算で確認できる。n が大きくなると T_n はほとんど 0.5 付近の値をとるようになることが弱大数の法則からわかるが、新しい確率変数

$$S_n = \frac{\sqrt{n}(T_n - E_\theta[X_1])}{\sqrt{V_\theta[X_1]}} = \frac{\sqrt{n}(T_n - 0.5)}{0.5} \qquad (2)$$

は期待値 0 分散 1 の正規分布にだいたい従うことが、中心極限定理からわかる (その説明は演習問題)。この式を逆に解くと $T_n = 0.5S_n/\sqrt{n} + 0.5$ である。
このことを用いると

$$\begin{aligned} & |T_n - 0.5| \geq c \\ \Leftrightarrow\ & |0.5S_n/\sqrt{n} + 0.5 - 0.5| \geq c \\ \Leftrightarrow\ & |0.5S_n/\sqrt{n}| \geq c \\ \Leftrightarrow\ & |S_n| \geq \frac{c\sqrt{n}}{0.5} \end{aligned}$$

n が大きいときに n が変化しても 「$|S_n| \geq$ 定数」が成立する確率はほとんど変化しないから (中心極限定理からわかる)、右辺を一定に保つために n が大きくなるにつれて c を小さく選べることがわかる。

c は n の値に応じて適切に定める必要があるが、統計量 T_n が開区間 $(0.5-c, 0.5+c)$ に含まれていれば仮説を採択する。このような仮説を採択する統計量の実現値の**集合**を**採択域**と呼ぶ。逆に、仮説を棄却する統計量の実現値の集合を**棄却域**と呼ぶ。この場合の棄却域は２つの閉区間の和集合

$$[0, 0.5-c] \cup [0.5+c, 1]$$

である。

両側検定と片側検定、帰無仮説と対立仮説

今、ある年に n 個の製品を出荷したと考え、確率変数 X_i は i 番目に出荷した製品が出荷後 1 年間に故障したときに 1 になり故障しなければ 0 になる確率変数とする。T_n を同じように定義すると T_n は故障率を表わす確率変数になる。X_i は 1 になる確率が母数 θ で決められる分布に従うとし、故障率が 0.01% $(=10^{-4})$ 以下であるかどうか仮説検定したいとする。この場合「$\theta = 10^{-4}$」という仮説を検定しても意味が無くて、その代わりに「$\theta \leq 10^{-4}$」という仮説を検定

するべきである。このとき母数 θ が小さくても仮説は棄却されないが、大きいと仮説は棄却される。
このように母数 θ の値が小さくても (あるいは大きくても) 仮説が棄却されない検定方式を**片側検定**という。一方、硬貨の例のように母数 θ の値が大きくても小さくても仮説が棄却される検定方式を**両側検定**という。
故障率の例に戻り、「故障率が低い」という仮説が棄却された場合には「故障率が高い」という仮説が採択されたと考えることもできる。有意水準 α は「故障率が低い」という仮説が正しいときにあやまって仮説を棄却してしまう確率だが、α を定義するために用いた仮説を**帰無仮説**とよび、帰無仮説が棄却されたときに採択されると考えられる仮説を**対立仮説**と呼ぶ。

片側検定の採択域と棄却域の決め方

故障率の例では帰無仮説は「$\theta \leq 10^{-4}$」であったが、仮説検定の採択域と棄却域を決めるときに母数 θ を適当な値に固定しないと確率を計算できないので、有意水準が α になるように採択域をどのように定めていいのかわからない。
故障率の場合では、$\theta = 10^{-4}$ として採択域を定める。例えば $\theta = 10^{-6}$ として採択域を定めると θ の真の値が 10^{-4} であったときに、帰無仮説「$\theta \leq 10^{-4}$」を棄却してしまう確率が目標としている有意水準 α よりも大きくなってしまい、不適切である。

演習問題

現実の問題を仮説検定してみる。セントラルリーグの選手（２００８年）の誕生月は最初にあげたようになっていた。プロ野球選手になるために早生まれは不利か？

上記のことを検定するために、以下のような確率変数を考える。まず、日本人全体では生まれた月はたいだい均等に分布している。(これは調べればわかることなので、興味がある人は自分で確認して下さい。) 従って、もしプロ野球選手になるために早生まれ (1〜3 月生まれ) が不利でないならば、一人のプロ野球選手が早生まれである確率は 0.25 以上であるはずである。プロ野球選手に適当に番号を 1 から 373 番まで振って、X_i を i 番目の選手が早生まれであるときに 1 になり早生まれではないときに 0 である確率変数とする。T_n を故障率の例と同様に定める。X_i は 1 になる確率が θ で 0 になる確率が $1-\theta$ になる確率分布から得られるとする。

演習問題を解答し、6 月 29 日 (月) の夜までに南 3 号館 3 階 311 号室前の箱に提出すること。なお、T_n の分布を正規分布で近似して解答する場合には、標準正規確率変数 Y が u 以上になる確率は教科書 280 ページに書いてあるので必要に応じて利用すること。なお、280

ページの表を使用しないで棄却域を決めるための計算は人間には不可能なので、計算機プログラムを使用する必要がある。また $\sqrt{373} \approx 19.31320791582796583858$ である。

91. 「プロ野球選手になるために早生まれ (1〜3 月生まれ) が不利でない」という仮説を検定するためには、θ に関する両側検定か片側検定かどちらを用いることがより適切か、理由と共に述べよ。

92. この検定における帰無仮説と対立仮説として、適切だと思う仮説を記述せよ。

93. 有意水準 5% での T_n の棄却域を述べよ。答えを導いた過程を詳しく書くこと。

94. 前問の棄却域に従って仮説を検定した結果をのべよ。

95. 有意水準 1% での T_n の棄却域を述べよ。答えを導いた過程を詳しく書くこと。

96. 前問の棄却域に従って仮説を検定した結果をのべよ。

97. 式 (2) で定義される確率変数 S_n が期待値 0 分散 1 の正規分布に従う理由を述べよ。

98. 式 (2) で定義される確率変数 S_n について、n が大きいときに n が変化しても「$|S_n| \geq$ 定数」が成立する確率はほとんど変化しない理由を述べよ。

99. 授業のわかりにくい点や要望・感想を書いて下さ

い。また前回の答案で×になったところで、理由がわからないところは申し出て下さい。