それって，トウケイ？
— 数理統計学入門 —

筑波大学数学系　　狩野　裕[1]

1. 数の歴史

誰かが紙に「7」と書いていたとします．この数字にはどのような意味が隠されているのでしょうか．平成7年度と書きたかったのかもしれませんし，体験学習の受講番号が7だったのかもしれません．筑波大学第一学群に到着するまでに乗り換えた電車の回数かもしれないわけです．いずれにしても，数字として同じでも背負っている歴史は大いに異なります．この「7」という数字からどのようなことがわかるのでしょうか．その歴史を知ることなしには，この問いに答えることはできません．

「キムタク」って知ってる？ と聞いたとき，7 人が知っていたとします．尋ねられた人が 10 人のときと 100 人のときとでは同じ7人でもその意味はまったく違いますね．では，10 人中 7 人が知っていたという場合と，1000 人中 700 人が知っていたという場合を比較してみましょう．どちらの場合も 70 % が「キムタク」を知っていたということですが，その重みが違うことに気づくでしょう．つまり，70 % という数の信頼度に違いがあるのです．

$$\boxed{\frac{7}{10} \neq \frac{700}{1000}} \tag{1}$$

この信頼度の違いは，10 人に聞いたのか，1000 人に聞いたのかという，数 70 % が出てきた経緯，つまり，その歴史に関係してくるわけです．

2. 歴史を記述する

数が背負っている歴史，言い換えると，その数がどのようにして得られたのかを記述したものを確率変数 (random variable) といいます．

女優の畠田理恵と電撃婚約発表した将棋界のプリンス羽生名人は，7 冠制覇を目指して現在，谷川王将と王将戦を戦っています[2]．あなたが将棋に自信があるとします．そして，羽生名人と将棋を指し，10 回勝負して 7 回勝ったとします (実際は勝つことはないと思いますが....)．この数 「7」 の歴史はどのように記述されるでしょうか．いま，7 回勝ったと言いましたが，調子が悪いと 1 回か 2 回しか勝てないかもしれませんし，最悪の場合全敗ということも考えられます．勝負はやってみないとわからないもので

[1] 出身は大阪

[2] 体験学習の頃には決着がついているかも？第 7 局は 3/21〜3/22 です．

す．このように，行ってみないと結果がわからない事柄をランダム（確率的）な現象といいます．ランダムな現象の起こりやすさを 0 から 1 までの実数で表したものが確率です．

羽生名人に勝つ確率を p とします．10 回勝負して羽生名人に勝つ回数を X という記号で表すことにすると，X は 0 から 10 までの値をとります．そして，その確率，つまり，$X=k$ となる確率は

$$\Pr(X=k) = {}_{10}\mathrm{C}_k p^k (1-p)^{10-k} \qquad (k=0,1,2,\cdots,10) \tag{2}$$

となりますね．このように，とり得る値が確率で表された変数を確率変数といいます．羽生名人に 7 回勝ったということは，0 から 10 までの値を生み出すランダムな現象で，確率が

$$\Pr(X=7) = {}_{10}\mathrm{C}_7 p^7 (1-p)^{10-7} \tag{3}$$

である 7 という事象が起こった，ということを示しています．この 7 という数の背後には，このような情報が隠されているのです．この 7 という数にはそれ自身には何の意味もないのですが，その歴史を知っていれば，つまり，10 回対戦して勝った回数であるということを知っているか，もしくは，確率変数（の分布）(2) を知っていいれば，いろんな情報を与えてくれます．

次に，少し違う状況を考えてみましょう．羽生名人と対戦して 7 回目で初めて勝ったとします．羽生名人に勝つ確率を p とすると，6 回連続して負け，7 回目で初めて勝つわけですから，その確率は

$$(1-p)^6 \times p \tag{4}$$

となります．羽生名人と対戦し初めて勝つまでに要した試合数を Y という記号で書きましょう．今回の 7 は，$Y=7$ だったわけです．そして，$Y=7$ となる確率 $\Pr(Y=7)$ は (4) であたえられました．一般に，k 回目に初めて勝つという確率は，

$$\Pr(Y=k) = (1-p)^{k-1} p \qquad (k=1,2,3,\cdots) \tag{5}$$

となります．まとめると，7 回目というのは，(5) で定められるランダムな現象の結果として，たまたま，$Y=7$ となったと考えられます．

同じ「7」という数であっても，「10 回対戦して 7 回勝った」という場合と「7 回目の対戦で初めて勝った」という場合とでは，もちろんその意味はまったく違いますね．言い換えると「7」の歴史が違うということですが，その違いを表したものが，確率変数 X，Y ということができます．

歴史のわからない数はそこから何の情報も得ることができません．上の 2 つの例のように歴史がわかっている数を「(統計) データ」[3]，そうでないものを，単に数値，数字，数などといい，区別することがあります．

---

[3]キャッチコピーの「トウケイ」はこの意味で使っています．

3. 勝利確率 p をあてる

この節では，あたたが羽生名人に勝つ確率 p はいかほどであるかを推定する方法について説明します．コイン投げを無限に行うと，表がでる割合は限りなく 0.5 に近づいていくと考えられます．同じように考えて，羽生名人と無限回対戦すると，勝利する割合はある値 p に近づいていくと仮定します．現実に無限回対戦したり，その間にお互いの実力が変化しないなんてことは考えにくいのですが，そう仮定したときの極限が，現在の確率だと定義するわけです．従って，p の値は「神のみぞ知る」ということになります．そこで，データ（数ではない！）から，p の値はこの位だと推測することになります．これを（統計的）推定 [(statistical) estimation] といいます．

2 節での例，10 回対戦して 7 回勝ったという場合を考えてみましょう．このときは，容易に，p の推定値 $\widehat{p}$ は

$$\widehat{p}=\frac{7}{10}$$

とすればよいことが直観的にも明らかでしょう．

この推定値は，次のような考え方でも導くことができます．(3) の確率を p の関数と考え次のようにおきます：

$$L_X(p)={}_{10}\mathrm{C}_7p^7(1-p)^3 \qquad (0\leq p\leq 1)$$

この $L_X(p)$ を最大にする p を計算してみます．最大値を求めるには微分が便利です．

$$\frac{\mathrm{d}}{\mathrm{d}p}L_X(p)={}_{10}\mathrm{C}_7p^6(1-p)^2(7-10p)$$

となりますから，$L_X(p)=0$ の解から，$p=\frac{7}{10}$ で $L_X(p)$ が最大となることがわかります．つまり，推定値 $\widehat{p}=\frac{7}{10}$ は，$L_X(p)$ を最大にする p の値として導出されるのです．$L_X(p)$ を尤度 (likelihood), このようにして構成される推定値のことを，最尤推定値 (maximum likelihood estimate) といいます．名前の由来は，尤もらしさの程度（尤度）を最大にするところから来ています．

この方法の基本的な考え方は次のようです．羽生名人と 10 回対戦して 7 回勝ったわけですから，そのようなことが起こる確率 $\Pr(X=7)$ は大きいに違いありません．そこで，$\Pr(X=7)$ を p の関数 $L_X(p)$ と考え, これを最大にする p の値を推定値としよう，と考えるわけです．

しかしながら，以上はアイデアです．最尤推定値がそのほかの推定値と比べてどのように良いのでしょうか，また，悪いこともあるのでしょうか，このようなことを調べるのは，数理統計学の重要なテーマのひとつになっています．

では，2 節の 2 番目の例はどうでしょうか．7 回目の対戦で初めて勝ったというわけですから，7 戦 1 勝，従って，$\widehat{p}=\frac{1}{7}$ と予想されます．しかし,「**初めて**勝った」というところに何かひっかかるものを感じてしまいます．そこで，先ほど紹介した最尤推定値を求めてみましょう．尤度は

$$L_Y(p)=(1-p)^6p \qquad (0\leq p\leq 1)$$

のようになります．これを p で微分して

$$\frac{\mathrm{d}}{\mathrm{d}p}L(p) = (1-p)^5(1-7p)$$

が得られます．従って，$L_Y(p)$ を最大にする p として，最尤推定値は

$$\widehat{p} = \frac{1}{7}$$

となり，先ほどの予想が正しかったことがわかるのです．

4. 信頼度ってナニ？

(1) 式で，$\frac{7}{10} \neq \frac{700}{1000}$ ということを強調しました.「キムタク」を知っている割合のことですが，10 人中 7 人知っているという場合ですと，偶然の入り込む可能性がかなり大きいような気がします．つまり，たまたま，音楽好きの若者にインタビューしてしまったという可能性がすてきれません．一方，1000 人にインタビューしたときは，そのような偏りは少なくなっていると考えられます．この違いをどう表したらよいのでしょうか.

今回は，10 人中 7 人が「キムタク」を知っていたということですが，別の 10 人に聞けば，3 人しか知らなかったということが起こり得ますね．つまり，次の比

$$\widehat{p} = \frac{\text{「キムタク」を知っていた人数}}{\text{尋ねられた人数}}$$

が尋ねられた人のグループによってどの程度ばらつくのか，ということが問題です. どの程度 といったって,「$\widehat{p}$ なんて 0 から 1 までの値をとるんじゃないの」という声が聞こえてきそうですが，実は，次のことがわかっています.

$$-1.96 < \frac{\widehat{p}-p}{\sqrt{\widehat{p}(1-\widehat{p})/n}} < 1.96 \qquad \text{となる確率が，およそ 95\% になる}$$[4]

このことを利用すれば,「キムタクを知っている人の (真の) 割合 p が次の区間にはさまれている確率」が，およそ 95% になることが導かれます.

$$\widehat{p} - 1.96\sqrt{\widehat{p}(1-\widehat{p})/n} < p < \widehat{p} + 1.96\sqrt{\widehat{p}(1-\widehat{p})/n} \tag{6}$$

この公式を，今の例 ($n = 10$, $\widehat{p} = 0.7$) にあてはめてみますと,

$$0.42 < p < 0.98 \tag{7}$$

となります．この意味をもう一度確認すると,「公式 (6) によって作られた区間は尋ねられた人ごとに異なるが，それらの区間が本当の p を含んでいる確率は 95% 程度である」となります．この 95 % を**信頼度**, (6) や (7) を**信頼区間**といっています.

[4]1.96 は標準正規分布の上側 2.5 % 点です．詳しくは大学で勉強しましょう.

$n = 1000$, $\widehat{p} = 0.7$ の場合と比較してみるとおもしろいことがわかります．この場合の信頼区間は，

$$0.67 < p < 0.73 \tag{8}$$

となり，先ほどの区間と比べてずっと短い，つまり，精度が高いといえますね．この差は，尋ねれられた人数の違い ($n = 10$ or $n = 1000$) からくるのです．

(7) や (8) のように，未知の量 p を区間で推定する方法を**区間推定** (interval estimation) といいます．これに対して，3 節で勉強した $\widehat{p} = 0.7$ のように，1 点で推定する方法を**点推定** (point estimation) とよんでいます．

5．偶然か，必然か

昨年の日本シリーズは，4 勝 1 敗でセリーグのヤクルトがオリックスを破り日本一になりました．日本シリーズは先に 4 勝したほうが優勝です．従って試合総数は 4 試合から 7 試合あることになります（引き分けは除く）．この試合数の分布を調べてみましょう．優勝が決まるまでの試合数を表す確率変数を Z と書くと，実力が互角だとすれば，

$$\begin{aligned}
\Pr(Z = 4) &= 2\left(\frac{1}{2}\right)^4 = 0.125 \\
\Pr(Z = 5) &= 2\ {}_4\mathrm{C}_1\left(\frac{1}{2}\right)^5 = 0.25 \\
\Pr(Z = 6) &= 2\ {}_5\mathrm{C}_2\left(\frac{1}{2}\right)^6 = 0.3125 \\
\Pr(Z = 7) &= 2\ {}_6\mathrm{C}_3\left(\frac{1}{2}\right)^7 = 0.3125
\end{aligned} \tag{9}$$

と簡単に計算できます．

過去，日本シリーズは 46 回 (1950〜1995) 行われていますから，4 試合で勝負かついたというのは 5 〜 6 回あったと期待されます．これは，

$$E_4 = 46 \times 0.125 = 5.75$$

のように計算したわけです，同様に計算した期待頻度 (E_k) と実際の頻度 (O_k) の分布を表 1 にまとめてあります.[5]

	試合回数				
	4	5	6	7	計
実際の頻度 (O_k)	5	9	16	16	46
期待頻度 (E_k)	5.75	11.5	14.375	14.375	46

表 1．日本シリーズ試合数の分布 (1950-1995)

[5] E_k, O_k は，それぞれ，Expected(期待された)，Observed(観測された or 実際の) の頭文字です．

さて，この表から，期待頻度と実際の頻度とは十分近いといえるでしょうか．ランダムな現象ですから，きっちりと等しくなることは期待できません．期待頻度からのずれが，**偶然**のなせるわざと考えられるのか，それとも，ずれは**必然**なのか，つまり，日本シリーズの試合数は上で計算した確率にしたがっているのかいないのか，検討してみましょう．2 つの状況を仮説にまとめると次のようになります．

H_0 : (9) の確率が正しい　　H_1 : (9) の確率は正しくない

データから，H_0, H_1 のどちらを採択すべきか判断することを**仮説検定** (testing hypotheses) とよんでいます．もちろん，100 % 正しい判断が得られることはありえません．何かしらのリスクはあります．

まず，O_k と E_k の差に基づく次の量（カイ 2 乗統計量とよばれる）を計算します.[6]

$$\chi^2 = \sum_{k=4}^{7}(O_k - E_k)^2/E_k$$

そうして，

$$\begin{aligned} \chi^2 < 7.81 &\implies H_0\text{を採択} \\ \chi^2 \geq 7.81 &\implies H_1\text{を採択} \end{aligned}$$

と判断します.[7] このデータでは $\chi^2 = 1.01$ となり，H_0 を採択，つまり，試合数の分布は (9) に従っていると考えてよく，期待頻度とのずれは**偶然**と考えてよいことになります．

さて，野球といえば大リーグです．ヤクルト対オリックスの日本シリーズに失望した**タイガースファン**の私でも，ドジャースの野茂選手に拍手を送っていました．大リーグのワールドシリーズも 7 戦中 4 戦勝てば優勝です．ワールドシリーズの試合数の分布を表 2 に載せてあります．

	試合回数				
	4	5	6	7	計
実際の頻度 (O_k)	11	11	12	27	61
期待頻度 (E_k)	7.625	15.255	19.0625	19.0625	61

表 2. 大リーグ・ワールドシリーズ試合数の分布 (1922-1982)

こちらの場合，先ほどより少し実際の頻度と期待頻度がずれているような気がします．特に，第 7 戦まで戦う頻度がかなり多いことがわかります．このずれは，偶然なのでしょうか．それとも，このずれが生じる何か必然的な理由があるのでしょうか．今回も仮説検定で調べてみましょう．

[6] χ はギリシャ文字で “カイ” と読みます．
[7] 7.81 は自由度 3 のカイ 2 乗分布の上側 5% 点です．詳しくは大学で勉強しましょう．

カイ２乗統計量 χ^2 を計算してみると次のようになり，今度は，H_1 が支持されます．

$$\chi^2 = \sum_{k=4}^{7}(O_k - E_k)^2/E_k = 8.60 \geq 7.84$$

つまり，試合数の分布は(9)にしたがってはいないという結論になります．

なぜ，日本とアメリカでは結論が違うのでしょうか．実は，シリーズを戦うにあたってのインセンティヴ[8]に大きな違いがあるのです．日本シリーズでは報奨金が第４戦までしか支払われないのに対して，ワールドシリーズでは全試合支払われるのです．従って，大リーグはたとえ優勝できなくても最終戦まで戦おうという気力が涌いてくるのです．大リーグで第７戦まで戦った回数が多い理由はそこにあるのです．

このように，仮説検定は，理屈からのずれが偶然の所産なのか，何か意味のあるずれなのかを明確に判断してくれるのです．

6. エピローグ

90 分としてはやや盛り沢山でしたでしょうか．少し復習をしてみますと，第１節では，数がどのようにして得られたのか — 数の歴史 — が重要であることを指摘しました．第２節では，その歴史を数学的に表現する確率変数を紹介いたしました．第 3, 4 節では，それぞれ，点推定，区間推定の例を紹介し，最後に，第５節で仮説検定の話をしたわけです．実は，第 3〜5 節を理解するためには，標本分布論を勉強しないといけません．脚注にあった正規分布やカイ２乗分布の意味を知らないと，ここで紹介した例を本当に理解したことにはならないのです．これらは，大学へ入学してからの課題としましょう．

高校では，統計的推測を習うことになっているのですがあまり重要視されていないように思われます．実はこれはまちがいです．統計学を勉強することは大きく分けて次の２つの目的があります．ひとつは，今回勉強したように，データをいか解釈し情報をとるかという点です．もうひとつは，数字にダマされないようにすることです．もっともらしい数値を並べ煙に巻こうとする人に対して，その結論はきっちりデータ処理されたものかという意味を込めて「それって、トウケイ？」と問い，また，あなたが何かを主張するとき，疑問を呈する人に対して、これはデータから導かれた客観的な結論であるという意味を込めて,「yes, トウケイ！」と答えられるようになりましょう．

それって，トウケイ？　yes，トウケイ！

体験学習に参加された皆さんは多くの貴重な経験をされたと思います．この経験を活かし，大学を「学歴をつける」ために入学するのではなく,「あるテーマを勉強したいという目的をもって」入学してくださることを期待しています．

[8]incentive: (行動などへの) 刺激，動機

「それって，トウケイ？」演習問題

1. ある人と将棋を指し，10 回対戦して 6 回勝ったとする。勝利確率を p として，尤度 $L_X(p)$ を求め，p の最尤推定値を計算せよ．

2. ある人と将棋を指し，m 回勝つまでに要した試合数を X とする．勝利確率を p として，

   $$\Pr\{X=k\} \quad k=m, m+1, m+2, \cdots$$

   を求めよ．また，3 回勝つまでに要した試合数が 10 であるとき，勝利確率 p の最尤推定値を求めよ．

   研究課題：上の確率をすべて加えると 1 になるはずである．

   $$\sum_{k=m}^{\infty} \Pr\{X=k\} = 1$$

   これを確かめよ．

3. 「小室哲哉」を知っている割合 p を区間推定せよ．(データは講義で指示)

4. 王将戦の歴史を以下に示す．試合数の分布は (9) 式にしたがっていると考えてよいか．仮説検定を行い，その結果を考察せよ．