推測統計入門*

今野　良彦†

1　はじめに

天気予報に「確率」なる言葉が使われ出してからかなり経ちますが，登場当時には降水確率という耳慣れない言葉に違和感を感じた人が多いという新聞記事を見かけたように記憶しています．この言葉もすでに定着したかもしれません．このように学校の勉強以外の日常生活においても「確率」という言葉を耳にすることはあると思います．では，いったいこの言葉はどのような場面に登場するのかということをまず考えてみましょう．別の例を考えてみます.「タバコを吸うと肺ガンになりやすい」といったとき，もちろん，タバコを吸っている人は肺ガンに必ずなるということではありません．しかし，肺ガンになる可能性が吸わない人よりは高いという具合に理解したほうがよさそうです．このように，確定的には予見できない現象を扱うときに「確率」という言葉がもちいられます．偶然性や不確実性をともない，実際の結果をみてみないとわからない現象のことを**でたらめ**もしくは**ランダム**な現象とよぶことにしましょう．ランダムな現象を数学的に表現したのが「確率論」です．そして，確率をもちいてでたらめな現象を記述したものを**確率モデル**と呼びましょう．

日常生活では，偶然性に左右されることは避けたいものですが，なかなかそうはいきません．しかし，長い時間の間，繰り返しランダムな現象を観測し，そこになんらかの傾向や法則性を見いだすことができる場合があります．たとえば，野球の打率や防御率といったものがその例にあたります．不確定な現象を繰り返し観察したり，集団を構成する個体を大量に調べてたりして，そこに働く法則性や傾向を見いだせるならば，それを**統計的法則**とよぶことにしましょう．さきほどの例の「タバコを吸えば肺ガンになりやすい」は典型的な統計的法則です．この統計的な法則を確率を用いて表現したものを**統計モデル**と特によぶことにしましょう．

これから，確率やでたらめとはどんなものかをお話し，簡単な確率モデルが実際の現象をうまく表現しているかを観測されたデータから調べることを考えてみます．

2　確率とは

不確定な現象が起こる可能性もしくはその確からしさを 0 から 1 の数値にあらわしたものが確率です．簡単な例を通してもう少し詳しく観ていきましょう．

サイコロを投げる例を考えましょう．出る目はいくつかといえば，それは 1 から 6 のどれかです．実際にどの目が出るかは不確実ですが，起こる可能性がある結果の範囲は確定しています．このように起こり得る結果のすべてを集めたものを**標本空間**と呼びます．起こる得る結果の集まりを「できごと」といいます．確率はこの「できごと」に対して考えます．堅い言葉でいえば，標本空間の部分集合である「できごと」を**事象**と呼び，確率は事象に 0 から 1 の数値を対応させます．たとえば，

「今度出る目は 1 」

「今度出る目は丁（偶数)」

というものが事象の例になります．そして，これらの事象が起こる確率を

$$P(\text{「今度出る目は丁（偶数）」})$$

---

* これは千葉大学の高校生向けのサマースクール―「いろいろな数」(2000 年 8 月 8 日) で行った講演の原稿である．

† 千葉大学大学院自然科学研究科
email: konno@math.s.chiba-u.ac.jp
http://www.math.s.chiba-u.ac.jp/ konno/talk.html

のように書きます．すなわち，

$$P(\text{「事象」})$$

です．事象は言葉でも数式でも起こり得る結果を示すものであればよいとします．

コインを投げる例を今度は考えてみましょう．可能な結果のすべての集まりである標本空間は

$$\{\text{表, 裏}\}$$

となります．すべての事象を書けば，

$$\{\text{表}\},\{\text{裏}\},\{\text{表，裏}\},\emptyset$$

となります．したがって，確率はこれらそれぞれの事象に 0 から 1 の数値が対応させます．投げるコインに細工がしてなければ，表の出る確率は 1/2 ですから

$$P(\text{表})=\frac{1}{2}$$

という式が得られます．

次に,「表」と「裏」の事象 [3] に数値を対応させることを考えます．たとえば，その対応を X と書き,「表」と「裏」という事象にそれぞれ 1 と 0 を対応させましょう．すなわち，

$$X: \begin{array}{ccc} \text{表} & \to & 1 \\ \text{裏} & \to & 0 \end{array}$$

です．すると上の式は

$$P(X=1)=\frac{1}{2}$$

という具合に数式らしく書けます．このように根元事象に数値を対応させるものを**確率変数**と呼びます．

[3] このふたつの事象はこれ以上分解できない事象である．このような事象を**根元事象**とよび，根元事象に数値を対応させることを考えるのである．

3　二項分布

今度はコインを 3 回投げることを考えます．表が x 回でる確率は

$${}_3C_x\left\{\frac{1}{2}\right\}^3$$

となります．ただし，

$$\begin{aligned} {}_3C_x &= \frac{3!}{x!(3-x)!}, \\ x! &= x\times(x-1)\times\cdots\times 1, \\ 0! &= 1 \end{aligned}$$

です．これは

$$\begin{aligned} P(\text{表が 3 回}) &= \underline{1} \times\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^0 \\ P(\text{表が 2 回}) &= \underline{3} \times\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^1 \\ P(\text{表が 1 回}) &= \underline{3} \times\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^2 \\ P(\text{表が 0 回}) &= \underline{1} \times\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^3 \end{aligned}$$

のことです．では，実験回数[4] が 3 回のときの $\{\underline{1}, \underline{3}, \underline{3}, \underline{1}\}$ はどのように求められたのでしょう．表が 3 回のときは毎回表が出なければならず，$\underline{1}$ 通りしかありません．また表が 2 回のときは，3 回のうちどこかで 1 回は裏となるからその組合せ数は $\underline{3}$ となります．以下，同じような計算をすれば，上の式が出てきます．では，実験回数を 4 回にしたらどうなるでしょうか？表 1 から

$$\begin{aligned} P(\text{表が 4 回}) &= \underline{1} \times\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 \\ P(\text{表が 3 回}) &= \underline{4} \times\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 \\ P(\text{表が 2 回}) &= \underline{6} \times\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 \\ P(\text{表が 1 回}) &= \underline{4} \times\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 \\ P(\text{表が 0 回}) &= \underline{1} \times\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 \end{aligned}$$

[4] すなわち，コインを投げる回数が 3 回．

となります．一般に，n 回の実験中で k 回表が出るいう組合せは $(n-1)$ 回の実験中で表が $(k-1)$ 回と k の組合せの合計となります．すなわち，

$$ {}_{n-1}C_{k-1} + {}_{n-1}C_k = {}_nC_k $$

という関係式です．

つぎに，1回の実験の可能な結果がふたつでその一方が起こるという事象の確率が p であるような実験を同一条件で前後の結果が影響しない状況のもとで繰り返し n 回行います．これは偶然性がともなう現象です．どうなるかわからないのでこの回数を確率変数 X と書きます．この X がたまたま $x(x=0,1,\ldots,n)$ になる事象，すなわち n 回中に $x(x=0,1,\ldots,n)$ 回起こるいう事象の確率は

$$ P(X=x) = {}_nC_x p^x (1-p)^{n-x} \tag{1} $$

となります．ただし，

$$ \begin{aligned} {}_nC_x &= \frac{n!}{x!(n-x)!}, \\ x! &= x \times (x-1) \times \cdots 2 \times 1, \\ 0! &= 1 \end{aligned} $$

とします．確率変数 X の確率（分布）が (1) であたえるものを**二項分布**とよびます．

4　確率を求めよう

大相撲の巴戦と日本シリーズの試合数を例にとり，すこし複雑な確率計算をしてみましょう．

プロ野球の日本シリーズでは，セリーグとパリーグのリーグ優勝球団が対戦を繰り返し，先に 4 勝した球団が日本一になります．したがって，引き分けの試合を除いた勝敗の決した試合数は 4 から 7 になります．対戦する球団の実力が互角とし，さらに前の試合の結果はつぎの試合の結果に影響を与えないと仮定[5] します．このとき，シリーズで行われる試合数を表す確率変数を Z とし，その確率を調べてみましょう．まず，対戦チームの名前を A と B としましょう．まず，A が 4 連勝する確率を求めます．4 連勝とは第 1 試合から第 4 試合まで続けて勝つことですから,「 A が 4 連勝」という事象は「A が第 1 試合勝ち，かつ第 2 試合勝ち，かつ第 3 試合勝ち，かつ第 4 試合に勝つ」というように言い換えられます．前の試合の結果は次の試合に影響を与えない（すなわち，独立である）ので，

$$ \begin{aligned} &P\{\text{「A が 4 連勝」}\} \\ &= P\{\text{「A が第 1 試合に勝つ」}\} \\ &\times P\{\text{「A が第 2 試合に勝つ」}\} \\ &\times P\{\text{「A が第 3 試合に勝つ」}\} \\ &\times P\{\text{「}A\text{が第4試合に勝つ」}\} \end{aligned} $$

ということから

$$ P(\text{「A が 4 連勝」}) = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} $$

となり，

$$ \{Z=4\} = \{\text{「A が 4 連勝」または「B が 4 連勝」}\} $$

であり，A と B は同時に 4 連勝することはないので，

$$ P\{Z=4\} = P\{\text{「}A\text{が4連勝」}\} + P\{\text{「}B\text{が4連勝」}\} $$

から $P(Z=4) = 1/8$ がわかります．

つぎに，シリーズが 5 戦目で終了する確率を求めます．そのために，A が 4 勝 1 敗でシリーズを終わり確率を求めしょう．A が最初の 4 戦を 3 勝 1 敗で戦い 5 戦目に勝てばいいわけです．すなわち，

$$ \begin{aligned} &P\{\text{A が 5 戦目で優勝}\} \\ &\quad = P\{\text{A が最初の 4 戦を 3 勝 1 敗}\} \\ &\quad \times P\{\text{A 5 戦目に勝つ}\} \end{aligned} $$

[5] したがって，それぞれの試合で勝つ確率は 1/2 と等し，それぞれの試合の独立な試行ということです．

です．最初の 4 戦を 3 勝 1 敗になる確率は

$$ {}_4C_1\left(\frac{1}{2}\right)^4 $$

なので，

$$
\begin{aligned}
&P\{Z=5\}\\
&\quad = 2P\{\text{ A が 5 戦目で優勝 }\}\\
&\quad = 2{}_4C_1\left(\frac{1}{2}\right)^4 \times \frac{1}{2}
\end{aligned}
$$

からわかります．あとは同じように考えれば，

$$
\begin{aligned}
P(Z=4) &= 2\left(\frac{1}{2}\right)^4 = 0.125 \\
P(Z=5) &= 2{}_4C_1\left(\frac{1}{2}\right)^5 = 0.25 \\
P(Z=6) &= 2{}_5C_2\left(\frac{1}{2}\right)^6 = 0.3125 \qquad (2)\\
P(Z=7) &= 2{}_6C_3\left(\frac{1}{2}\right)^7 = 0.3125
\end{aligned}
$$

となります．

実際に，4 試合から 7 試合になったシリーズ数の分布（相対的頻度）はこの理論値に近いかどうかが気になるところですね．この点についてはあとでみることにしましょう．

大相撲の本場所は 15 日間の勝ち星の多さでその場所の優勝者を決めます．非常に混戦した場所では，15 日間の最高の勝ち星数をあげた力士がひとりではない場合[6] が起きます．相星[7] がふたりの場合，そのふたりの間で優勝決定戦を行います．相星が三人のときには，トーナメント方式では組み合わせに公平さを欠きます．そこで，巴戦をするわけです．これは，くじで最初の取り組みを決めます．この取り組みの勝者が控えの力士と対戦し，最初の取り組みの勝者がここでも勝てば，連勝した力士を優勝とし，最初の取り組みの勝者が負ければ，2 回目の取り組みで勝った方が 1 回目の取り組みで負けた力士と対戦するといった具合に，連勝する力士が現れるまで取り組みが繰り返されます．はたして，巴戦は公平でしょうか？それとも，組み合わせにより，有利不利があるのでしょうか？

話を簡単にするために，A, B, C の力士がそれぞれお互いに互角としましょう．したがって，それぞれの取り組みで勝つ確率がどの力士も 1 / 2 ということです．A と B が最初の取り組みで対戦するとします．最初の取り組みで A も B も勝つ確率は同じだから，優勝する確率も同じになることがわかります．最初の取り組みで控えにまわる C の優勝する確率をもとめれば，

$$ P(\text{「 A が優勝」}) = P(\text{「B が優勝」}) $$

および

$$
\begin{aligned}
&P(\text{「A が優勝」}) + P(\text{「B が優勝」})\\
&+P(\text{「 C が優勝」}) = 1
\end{aligned}
$$

から

$$ P(\text{「 A が優勝」}) = \frac{1}{2}\{1 - P(\text{「C が優勝」})\} \qquad (3) $$

となり，すべての力士について優勝する確率がもとまるわけです． さらに話を簡単にするために，取り組みの疲れもふくめて前の取り組みの結果はつぎの取り組みの勝敗の確率に影響を与えないとします．C が優勝するためには巴戦の番数が 3 の倍数にならなければなりません．$3m$（m は自然数）回目の取り組みで， C が負ける確率と優勝する確率は同じで，これを v_m と記すことにしましょう．$3m+2$ 回目の取り組みまで巴戦の決着がつかず，$3(m+1)$ 回目の取り組みで C が優勝する確率と負ける確率は

$$ v_{m+1} = \frac{1}{8}v_m $$

となります．したがって，C が優勝する確率は

$$ \sum_{m=1}^{\infty} v_m = \lim_{m\to\infty}\frac{v_1(1-(1/8)^m)}{1-(1/8)} = \frac{8}{7}v_1 $$

[6] これを事象と呼ぶといった．しかし，この言葉はどうも日常では使いなれないですね．

[7] 勝ち星が同じであること．

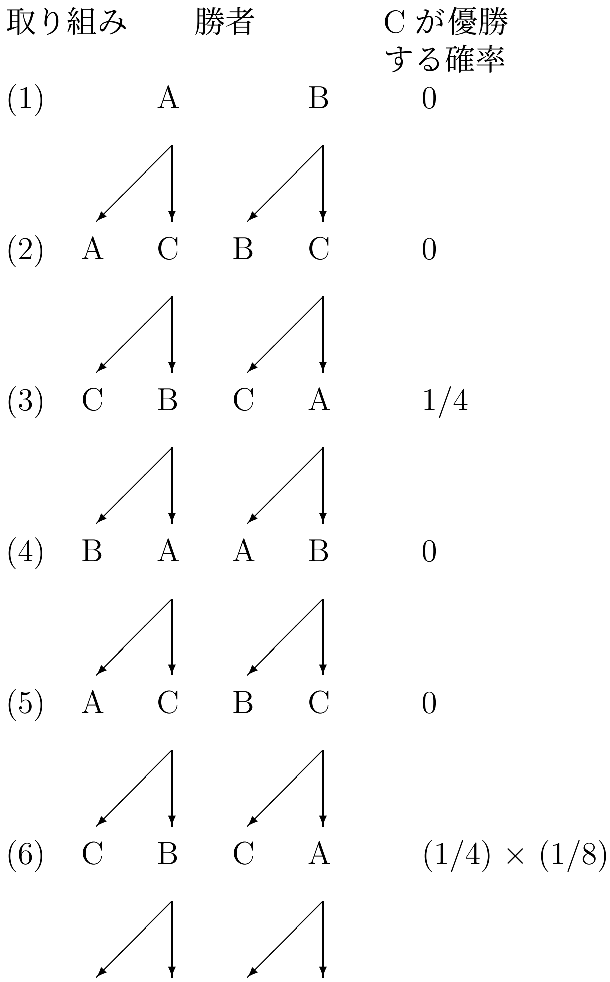

図 1: 巴戦の確率計算

と表現できます．．図 1 から $v_1 = 1/4$ となるので，C の優勝する確率は

$$\frac{4}{14}$$

になります．(3) から A と B の優勝する確率はともに

$$\frac{5}{14}$$

になり，C は A と B にくらべてやや不利であること[8] がわかります．相星の力士が４人以上[9] のときには，巴戦によって優勝力士をきめるのではなく，トーナメント方式になります．４人決定戦は公平ですが，５人決定戦では組合せで２力士の優勝する確率が 1/ 8 で，のこりの３力士が 1 /4 となり，ひどく公平性を欠くもの[10] になります．

5　打率 p をあてる

ここでは，前節で導入した二項分布を用いて打率の確率モデルにおける打率の推定を考えてみましょう．

イチロー選手がヒットを打つ確率を p とします．10 回打席[11] に立ってヒットを打った回数を X としましょう．ただし，それぞれの打席でイチロー選手がヒットを打つ確率 p は一定とし，前の打席の結果がつぎの打席の結果に影響しないと仮定[12] します．現実には，このような仮定するのは不自然で本来はもっと複雑な現象ですが，隅から隅まで逐一観察することは不可能です．だから，ある程度（かなりの程度）まで単純化し，端的に観ていくことが必要です．もちろん，どう単純化するかは自明の場合ばかりとは限りません．単純化された現実の代用品を「モデル」と呼びます．さらに，実際に手元にある観測結果から確率モデルや統計モデルを特定したり，考えているモデルが観測さ

[8] ここでは，3 力士が互角としましたが，互角でなくともやはり控えの力士が不利になることがわかります．組み合わせによる不公平をなくすためには，3 力士で総当たり戦をやり，総当たり戦で 2 勝する力士がでるまで，総当たり戦を繰り返し行えばよいことがわかります．総当り戦が巴戦より優勝者が決まるまでに必要な番数の期待値が多くなることが予想されます．興味のある人は計算してみましょう．

[9] 優勝決定戦の制度ができた 1947 年以降に，巴戦が 6 回，４人決定戦が 2 回，５人決定戦が 1 回あったようです．

[10] 優勝決定戦の取り組みの組み合わせをより公平にするためには，巴戦とすればいいはずです．しかし，力士が互角としたとき，優勝者が決まるまでに期待される番数が 15 番程度になり，本割がすんだあとにこのような仕組みで決定戦をおこなうのは興行上無理なわけです．[1] を参照．

[11] 四死球をのぞいたもの．

[12] 対戦相手の投手が変われば，ヒットを打つ確率は変わるでしょうし，イチロー選手の調子によってもこの確率は変化するかもしれません．

れたデータにうまく合っているかをしらべることが必要となります．この一連の手続きの集まりとそれにまつわる理論の体系を（推測）統計学です．

打率のモデルを数式で書いてみましょう．X は 0 から 10 の値をとります．x を 0 から 10 のどれかとして，$X = x$ なる事象がおこる確率は

$$P(X = x) = {}_{10}C_x p^x (1-p)^{10-x} \qquad (4)$$

となります．ただし，$x = 0, 1, \ldots, 10$ です．上の式で p の値をデータから推し量ることをしましょう．これを**推定**といいます．

たとえば，実際にイチロー選手が 10 回の打席でヒットを 4 回放った場合を考えましょう．このとき，p の推定値 $\hat{p}$ は

$$\hat{p} = \frac{4}{10}$$

とすればよいことが直観的にわかるでしょう．実際に，打率はこのように計算されます．

直観的に正当化するだけでは芸がないので，この 0.4 という推定値がどのように導かれるかをすこしだけ統計学的に考えてみましょう．実際に 10 打席中に 4 本のヒットを打ったので，このような「できごと」が起こる確率である $P(X = 4)$ は大きいに違いありません．そこで，(4) の式に $x = 4$ を代入したものを考えましょう．そして，これを p の関数とみなしたのも

$$l(p) = {}_{10}C_4 p^4 (1-p)^6, \qquad (0 \le p \le 1).$$

考え，これを最大化する p の値を推定値とすればよいでしょう．これは現実におこった「できごと」の確率を最大にするようにモデルを選ぶということを意味します．$l(p)$ のことを**尤度**[13]，そして，この尤度を最大にする p の値のことを**最尤推定値**と呼びます．また，この手続きに基づいて推定値を求める手法のことを最尤法とか最尤推定量[14] と呼ぶことにします．なんだか尤もらしいことをいっているようですが，本当かなと思いませんか？やっぱり，これだけではよくわかりませんね[15]．最尤法についてはとっても深い深い話があります．そして，尽きない話でもあるので，最尤法に関してはほんのさわりだけにしておきます．

[13] 尤もらしさの程度

6　でたらめと信頼度

前節の議論をそのまま適用するとおかしな結論を誤って導いてしまうことがあります．そのような例に新聞や雑誌の記事を読んでいるとお目にかかることがあります．たとえば，

パソコン職場「女児出産ラッシュ」の真相

といったようにとてもセンセーショナルな見出しの記事がその例です．この記事によれば，総合家電メーカー開発部門で，ここ 3 年間に職場の男性が授かった子ども 20 人のうち，14 人が女の子であった．パソコンから出ている電磁波のせいではないか....　といった具合です．先ほどの最尤法を用いれば，0.7 が尤もらしい女児が生まれる比率の推定値になります．最尤法を知らなくとも，やぱり 0.7 と考えるでしょう．男女比はだいたい 1:1 なのだから，どうもパソコンをいじってばかりいる男性が授かる出生児は女の子になりやすいのだろうか?でも，ちょっと待てよ．すこし変な感じがしませんか．そこで,「でたらめ」ということがどんなものかコインを投げる実験を 20 回してみました．コインの表と裏の出る確率はそれぞれ 0.5 と考えていいので，男女の出産比率が 1:1 の場合を想定

[14] 推定値と推定量の使いわけがきちんとできれば，推測統計学のきほんがわかったかなということになります．

[15] 最尤推定法や尤度の正当化は推測統計学の重要なテーマで，そして，尤度をきちんと理解することを重要な目標として現在も研究を続けけている流派がいます．

して 実験をしたことになります．実際に試してみると

男	男	女	女	男	女	女	女	男	男
女	男	女	女	女	女	男	女	女	女

となりました．男の子が 7 人で，女の子が 13 人になりました．この場合の比率は理論比率の 10 人，10 人にはぴったりと一致しません．どうもでたらめの影響で理論比率と実験で得られる比率とでは違い[16] が生じるようです．では，男女の出生比率が等しいと仮定したときに，でたらめの影響のみで男女の比率が 7:14 になる確率はどのくらいあるのでしょうか？X を女の子が生まれる人数とすれば，X は 2 項分布に従うので，女の子が 14 人以上生まれる確率は

$$
\begin{aligned}
P(X \geq 14) &= \sum_{x=14}^{20} {}_{20}C_x (0.5)^x (0.5)^{20-x} \\
&= 0.05766
\end{aligned}
$$

となります．同じ調査を 100 回[17] やれば，20 人の出生児中に女の子が 14 人になることはでたらめの影響で 100 回の調査中に 5 回ないし 6 回ぐらいは偶然におこる結果といえます[18]．だから，この記事にあることだけでパソコンをいじってばかりいると女の子が生まれやすいなんていう結論を導くのは早とちりの可能性があるわけです．たとえ，男女の出産比率が 1:1 であっても 20 人の出生児の性別を調べれば，すこしすこしずれてくるわけです．では，この記事による出生比率 $\hat{p}$ から実際の比率（理論値）p がどのような範囲にあるかを考えてみましょう．n を調査数としたときに，

$$P\left\{-1.96 \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq 1.96\right\} \approx 0.95$$

ということが知られています[19]．したがって，理論比率 p がつぎに区間

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \tag{5}$$

にはさまる確率が 0.95 になることがわかります．この公式に $n = 20$，$\hat{p} = 0.7$ を代入すれば，

$$(0.499, 0.901)$$

になります．これは，公式 (5) によって得られる区間は調査ごとに異なるが，それらの区間が理論比率 p を実際に含んでいる確率は 0.95 程度であるということです．この区間のことを 95 % の**信頼区間**といいます．

公式 (5) をどのように導出する[20] かはさて置き，信頼区間の計算の仕方はわかりましたか？もし，100 人に対して調査をしたらところ 70 人が女の子であったら，比率 p の信頼区間はどうなるか計算してみましょう．(5) に数値を代入すれば，

$$(0.672, 0.728)$$

になります．こうなってくると影響があると考えた方がよさそうです．20 人中 14 人であっても 100 人中 70 人であっても比率は同じ 0.7 ですが，$\frac{14}{20}$ と $\frac{70}{100}$ では意味が違うことがわかると思います．(5) の平方根の中の分母をみれ

[16] 逆に，

男	男	女	女	男	女	女	女	男	男
女	男	女	男	女	男	男	女	女	男

という結果は偽造の疑いがもたれます．実験データをねつ造するときには，あまり理論値と一致しないほうがそれらしく見えます．

[17] すなわち，20 人の出生児の性別を調べて，そこに何人の女の子がいるかを数える調査を 100 回繰り返すことです．

[18] でたらめのせいで理論値と観測値のずれに説明が付きそうなとき「**統計的に有意でない**」といい，逆にでたらめの影響だけではその差を説明することができず，別の理由に起因する差であると考えた方が妥当なとき「**統計的に有意である**」といいます．注意深くテレビや新聞を読んだりみたりしているとこの言葉がときどき登場していることに気が付きます

[19] 1.96 は標準正規分布の上側 2.5 % 点のことです．

[20] 導出については大学の統計学の講義で学びましょう．

ば，調査数 n がおおきくなればなるほど信頼区間は $\hat{p}$ を中心に短い区間になることがわかります．だから，調査数を増やせば，$\hat{p}$ の信頼性が増すわけです．すなわち，調査数をふやせば，観測される女児の比率は理論比率に近づいて行くわけです．このことを**大数の法則**と呼びます．

7　データと較べる

最後に確率法則と現実のデータがうまく適合しているかをデータから調べてみます．でたらめの要素があるので現実のデータは理論値に一致するのではなく，適当にずれているのです．だから，すれがでたらめだけによるものか理論や法則が現実には当てはまらないのかを調べるということです．

先ほど，プロ野球の日本シリーズの試合数の確率分布を調べてみました．では，1950 年から 1999 年までの日本シリーズ 50 回の結果と理論値を比較することにしましょう．試合数が 4 になる確率は 0.125 だったので，50 回中で 4 試合でシリーズがおわってしまう期待値は

$$E_4 = 50 \times 0.125 = 6.25$$

となります．つまり，4 試合でシリーズの勝敗が決まってしまうことが 6 回くらいあると期待されるわけです．同じように 5 試合，6 試合，7 試合でシリーズの勝敗が決するシリーズ数の期待値を計算することにします．この結果をまとめたものが表 2 の左側の柱です．

先ほど来の議論から，実現値は理論値とはぴったりと一致はしないことは予想されます．しかし，このずれはでたらめのためでしょうか？それとも別の要因のためでしょうか？すなわち，試合数が 4 から 7 になるシリーズ数は上で計算した確率にしたがっているかを検討しましょう．まず，何を調べるかをきちんとまとめましょう．

$$\begin{aligned} H_0 &: (2) \text{ の確率は正しい} \\ H_1 &: (2) \text{ の確率は正しくない} \end{aligned}$$

です．観測したデータから H_0 か H_1 のどちらかを受け入れる（採択）べきかを判断する推測統計の手続きを**仮説検定**と呼びます．どちらかの仮説を採択するにあたり，正しい判断が必ずしも可能ではなく，H_0 が本当は正しいのもかかわらず，H_1 を採択したり，逆に H_1 が本当は正しいのもかかわらず，H_0 を採択をするような誤り[21] を犯す危険があります．

O_k と E_k の差に基づく量（**カイ２乗統計量**と呼ばれます）を計算します．

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

上の式[22] の分子をみると実現値と理論値の差の２乗を取っているので，その差がひろがればひろがるほど χ^2 の値は大きくなります．でたらめの影響で χ^2 の値はある程度の範囲の数値になるが，でたらめ以外の影響をうけ理論値と実現値がひどくかけ離れれば χ^2 の値は大きくなるはずです．では，実現値が理論値からひどく離れているとどこ[23] からみなすのでしょうか？これはでたらめの影響だけで起きた差であったとしても起きる可能性がとても低いところからとします．たとえば，実現値と理論値が離れ具合がでたらめの影響のみによって引き起こされる可能性が 5 % はあることを許すせば，

$$\begin{aligned} \chi^2 < 7.81 &\implies H_0 \text{ を採択} \\ \chi^2 \geq 7.81 &\implies H_1 \text{ を採択} \end{aligned} \tag{6}$$

[21] 前者のような誤りを**第一種の誤り**，後者を**第二種の誤り**と呼んで区別をします．この命名は覚えにくいですね．統計的検定では，前者n誤りの危険を一定水準以下にして，後者の誤りをできるでけ低くするような検定方式を見つけようとします．

[22] χ は文字で「カイ」と読みます．

[23] この境のことを閾値と呼びます

という具合[24] にします．このデータに基づいて χ^2 の値を計算してみます．

$$\frac{(5-6.250)^2}{6.25}+\frac{(12-12.500)^2}{12.5} +\frac{(17-15.625)^2}{15.625}+\frac{(16-15.625)^2}{15.625} = 0.4 < 7.81$$

となります．したがって，H_0 を採択，すなわち，シリーズ数の分布は (2) にしたがっていると考えてよいわけです．偶然のずれにしては理論値と実現値が近すぎるぐらい似ている結果になりました．

すこしデータが古いですが，大リーグの 1922 年から 1982 年までのワールドシリーズについて同じ分析をしてみましょう．ワールドシリーズの場合は試合数が 7 の場合のシリーズが多いことが一見してわかります．この実現値と理論値の違いはでたらめのために生じたものでしょうか，それともなにか別の理由によるものかと考えた方がいいのかを検定手続き (6) を用いて調べてみましょう．カイ 2 乗統計量の値を計算してみると

$$\chi^2 = \frac{(11-7.625)^2}{7.625}+\frac{(11-15.25)^2}{15.25} +\frac{(12-19.0625)^2}{19.0625}+\frac{(27-19.0625)^2}{19.0625} = 8.60 \geq 7.81$$

となります．試合数の分布は (2) にしたがっていないと考えた方いいようです．同じような対戦方式であっても日米ではすこし様子が違うようです．大リーグと日本のプロ野球は同じ野球でも違うもの[25] であるとよくいわれますが，このようなところにも違いが現れたのかもしれません．

[24] 7.81 は自由度 3 のカイ 2 乗分布の上側 5 % 点です．

[25] 大リーグの方が最終戦までシリーズの勝敗がもつれる傾向にはなにか誘因があるかもしれません．そのひとつの可能性として，シリーズの報奨金の制度の違いです．日本シリーズでは報奨金が 4 戦分までしか支払われませんが，ワールドシリーズでは全試合分支払われます．この違いから最終戦までがんばろうという気が両チームともわいてくるのでしょうか．

8　終わりにかえて

この時間では，でたらめや偶然性をあつかう確率の話から始まり，野球や相撲の例をとり，確率の計算や確率モデルの確率をデータから推定すること（点推定法のひとつの考え方である最尤法や区間推定）およびデータが確率モデルにうまく適合しているかをみてみました．

ここでのお話は推測統計のほんのさわりでしかありませんので，統計学および推測統計学から，でたらめの影響を考慮しつつデータをいかに解釈するするのかということに対する指針を学ぶことが大切であるかをほんのすこしでも理解できれば十分だと思います．それから,「数値で説明されることが常に客観性をもち，間違いがない」というようには考えてはいけないことを理解し,「数値にだまされない」こころ構えだけでも今日ここでのお話を通じて身につけてもらえれば，この時間でのお話の初期の目標は達成できたと思います．

参考文献

[1] Hayter, A.J., Some Probability calculations concerning the playoff system in Sumo tournaments. ISM Symposium, Recent advances in Statistical Data Analysis.

[2] 狩野裕,「それって，トウケイ？」筑波大学（高校生）数学体験学習（DVI file の配布資料），1996.3　(http://koko15.hus.osaka-u.ac.jp/%7Ekano/application/index.html).

[3] 鈴木義一郎，先をよむ統計学　「情報量基準」とは何か，講談社ブルー　バックス　B-855，1991.

[4] 清水誠，推測統計　はじめの一歩，講談社ブルー　バックス　B-1283，2000 年．

[5] 吉村　功,　平均　順位　偏差値 – データに負けるな–, 岩波ジュニアシリーズ 72, 1984 年.

3 回まで		4 回目	4 回まで		小計
表が 3 回	$\underline{1}$	表	表が 4 回	$\underline{1}$	$\underline{1}$
表が 3 回	$\underline{1}$	裏	表が 3 回	$\underline{1}$	
表が 2 回	$\underline{3}$	表	表が 3 回	$\underline{1}$	$\underline{1}+\underline{3}=\underline{4}$
表が 2 回	$\underline{3}$	裏	表が 2 回	$\underline{3}$	
表が 1 回	$\underline{3}$	表	表が 2 回	$\underline{3}$	$\underline{3}+\underline{3}=\underline{6}$
表が 1 回	$\underline{3}$	表	表が 1 回	$\underline{3}$	
表が 0 回	$\underline{1}$	表	表が 1 回	$\underline{1}$	$\underline{3}+\underline{1}=\underline{4}$
表が 0 回	$\underline{1}$	裏	表が 0 回	$\underline{1}$	$\underline{1}$

表 1: 二項確率の計算

	シリーズ数	
試合数	実現値（O_k）	理論値（E_k）
4 試合	5	6.250
5 試合	12	12.500
6 試合	17	15.625
7 試合	16	15.625
合計	50	50

表 2: プロ野球日本シリーズの試合数 (1950-1999)

	シリーズ数	
試合数	実現値（O_k）	理論値（E_k）
4 試合	11	7.625
5 試合	11	15.25
6 試合	12	19.0625
7 試合	27	19.0625
合計	61	61

表 3: 大リーグのワールドシリーズの試合数 (1922-1982)