平成 10 年度「数学への誘い」
(筑波大学数学体験学習教材)

統計学は未来を予測できるか？

赤平昌文 (筑波大学数学系)

1. はじめに

最近の世の中の状況をみると, 従来のシステムがうまく機能しなくなり, 大幅な改革を余儀なくされている. このような混沌として先の見えない時代に統計学は有効であろうか. 統計学は, 本来, データに基づいて, そのデータの源泉である母集団の特性について推測するものであるが, さらに拡張することによって, 統計的予測問題を考察することもできる. その際, 観測されたデータ X に基づいて未来 (未観測) の確率変数 Y を予測することになるが, 通常, X, Y の分布が未知の母数 (パラメータ) に依存するので, そのことを十分考慮しなければならない. このことは, ある時点までに得られたデータに基づいて未来を予測できることを示している. そのようなことは, 実際に可能であろうか?

本論においては, まず, ベルリンの壁の崩壊年の予測について述べ, 1999 年の日本のプロ野球チームのある時点での勝数 X に基づいて, 残り試合での勝数の予測問題について考え, さらに 1999 年の米国の大リーグにおいてホームラン数の新記録達成を競った選手のある時点までのホームラン数 X に基づいて, 残り試合でのホームラン数の予測問題について考察する. そして, 現実のデータに基づいて予測した結果と実際に起こった結果を照らし合わせて, 本論の統計的定式化の下での予測法の妥当性を確かめる.

2. ベルリンの壁の崩壊年の予測

1969 年に米国の青年が夏休みを欧州で過ごしていたとき, ふとベルリンの壁を訪れた. 東西冷戦の象徴であるベルリンの壁は 1961 年に築かれていて, ちょうど 8 年が経過していた. その青年はその壁に対峙(たいじ)して, この壁の存在する状況はあとどのくらい続くかを考えた. そのとき, その青年は築年数だけでその壁の存続期間を見積もる方法を思いついた. まず, その訪れた時点は壁が存在する期間の任意の時点であり, 何か特別な時点ではないと考えた. そして, 壁が築かれてから存続する期間を θ 年とすれば, 区間 $[1961 + (\theta/4), 1961 + (3\theta/4)]$ の中に, 訪れている時点 $(1961 + X)$ 年が入る可能性が確率 50% であると考えた. 実際, 壁がその時点 $(1961 + X)$ 年から存続する期間は $(\theta - X)$ 年であるから, $\theta/4 \le X \le 3\theta/4$ より

$$\frac{X}{3} \le \theta - X \le 3X$$

となり, 壁を訪れた時点からその壁が存続する期間が築年数の 1/3 倍から 築年数の 3 倍になる確率が 50% となる. 今の場合, $X = 8$ であるから, その青年がその壁を去る前に, この壁の存続期間は 50% の確率で 2 年 8ヶ月以上 24 年以下であろうと友人に予言した. 実際, この壁は 20 年後の 1989 年 11 月に崩壊した. その崩壊は突然ではあったがその予言はほぼ予想通りであった. 同様の考え方で, 壁を訪れた時点からその壁が存続する期間は築年数の 1/39 倍から築年数の 39 倍になる確率は 95% であることが示される. なお, この青年は後に, 米国のプリンストン大学教授 (宇宙物理学) になった.

3. 統計的予測問題と準備

本論では 2 つの具体的な予測問題について考えてみよう.

[**問題 1**] 1998 年の日本のプロ野球では, シーズン後半にセ・リーグの横浜, 中日, 巨人の 3 チームが優勝争いを展開した. このとき, 各チームは残り試合において何勝するだろうか?

[**問題 2**] 1998 年の米国の大リーグではホームラン数の新記録達成の可能性について, 多くの人達の注目を集めた. 残り試合数が少なくなった時点で, マグワイア選手やソーサ選手があと何本ホームランを打つのかという話題で興奮するのは当然である. さて, 残り試合で両選手のホームラン数は何本であろうか?

そこで, 上記のような問題を取り組む際に統計的定式化が必要になるが, そのための準備をしよう.

3.1. 2 項分布

まず, 1 枚のコインを用意し, コインを投げる実験をするときに, このコインは表か裏のいずれかが出るとし, また, 表が出る確率が p $(0 < p < 1)$ とし, さらにこの実験を通してこの確率は変わらないとする. このとき裏が出る確率は $q = 1 - p$ になる. 一般に, このような (2 値性, 定常性をもつ) 実験を行うことを 2 項試行 (またはベルヌイ試行) という. つぎに, 上のコイン投げ実験の試行を 1 回行ったときの結果を X とし

$$X = \begin{cases} 1 & (\text{表が出たとき}), \\ 0 & (\text{裏が出たとき}) \end{cases}$$

とする. このとき, $X = 1$ をとる確率 (probability) は p になり, これを記号で, $P\{X = 1\} = p$ で表わし, 同様に $P\{X = 0\} = q$ で表わす. ここで, X は変数で, X のとり得る値に対してその確率が定まるとき, X を確率変数という. さら

に, この試行を n 回独立に繰り返したときの結果を $X_1, \cdots, X_n$ として, その和を $Y = \sum_{i=1}^{n} X_i$ (n 回のうち表が出る回数) とすれば, Y は確率変数で, その確率分布は

$$f_Y(k) = P\{Y = k\} = {}_nC_k p^k q^{n-k} \qquad (k = 0, 1, \cdots, n) \tag{1}$$

になり, これを 2 項分布 (binomial distribution) といい, 記号で $B(n, p)$ で表す (図 1〜3 参照). 上記の X の確率分布は, (1) で $n = 1$ の場合であり,

$$f_X(x) = P\{X = x\} = p^x q^{1-x} \qquad (x = 0, 1)$$

になり, これは 2 項分布 $B(1, p)$(またはベルヌイ分布) である. なお, (1) は $(p + q)^n$ の 2 項展開の $p^k q^{n-k}$ の項になっていることに注意. また, Y が 2 項分布 $B(n, p)$ に従うとき, Y の平均 μ, 分散 σ^2 はそれぞれ

$$\mu = \sum_{k=0}^{n} k f_Y(k) = np, \qquad \sigma^2 = \sum_{k=0}^{n} (k - \mu)^2 f_Y(k) = npq$$

になる.

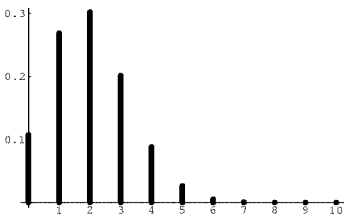


図 1. 2 項分布 $B(10, 0.2)$

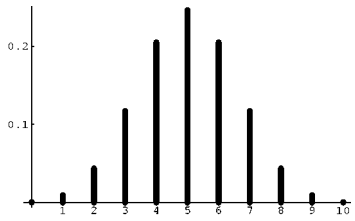


図 2. 2 項分布 $B(10, 0.5)$

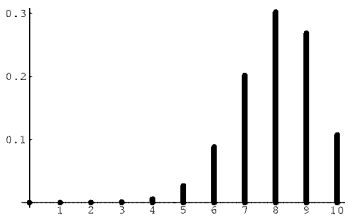


図 3. 2 項分布 $B(10, 0.8)$

3.2. ポアソン分布

上記において, 離散的な時点でランダムに起こる現象に注目したときに 2 項試行列でとらえたが, 連続な時点の場合にはどうであろうか. 区間 $(0, t)$ を幅 $h = t/n$ の n 個の小区間に分割して, 各小区間ではある事象が 2 回以上起こり得ないほど n を十分大きくとるとする. 各小区間においてその事象が 1 回起こる確率を p とし, 区間 $(0, t)$ においてその事象が起こる小区間の数を X とし, X が n 回の独立な 2 項試行の結果の和と見なせるとき, X は $B(n, p)$ に従う. いま, 区間 $(0, t)$ におけるその事象の平均生起数を λ で一定とすれば, $np = \lambda$ となる. このとき, $n \to \infty$ (すなわち $p \to 0$) とすれば

$$\begin{aligned} f_X(k) = P\{X = k\} = {}_nC_k p^k (1-p)^{n-k} &= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &\to \frac{\lambda^k e^{-\lambda}}{k!} \end{aligned}$$

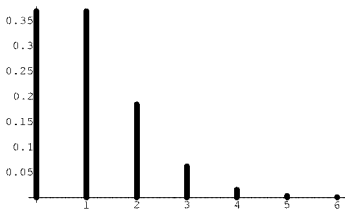
になる. ただし, $\left(1+\frac{1}{n}\right)^n \to e$ $(n \to \infty)$ とする. そこで, 確率変数 Y が確率分布

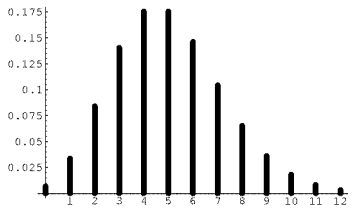
$$f_Y(k) = P\{Y = k\} = \frac{\lambda^k e^{-\lambda}}{k!} \quad (k = 0, 1, 2, \cdots ; \lambda > 0)$$

に従うとき, これをポアソン分布 (Poisson distribution) といい, 記号で $Po(\lambda)$ で表わす (図 4, 5, 6 参照). 実際には, 一定の時間間隔内における機器の故障数, 電話がかかってくる回数, 交通事故数などがポアソン分布に従うことが知られている. また, Y がポアソン分布 $Po(\lambda)$ に従うとき, その平均 μ, 分散 σ^2 はそれぞれ

$$\mu = \sum_{k=0}^{\infty} k f_Y(k) = \lambda, \qquad \sigma^2 = \sum_{k=0}^{\infty} (k-\mu)^2 f_Y(k) = \lambda$$

になる.





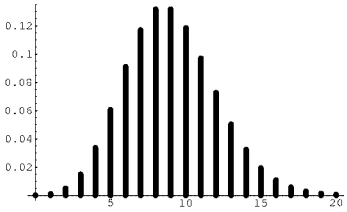


図 4. ポアソン分布 $Po(1)$ **図 5.** ポアソン分布 $Po(5)$ **図 6.** ポアソン分布 $Po(9)$

4. 予測問題の統計的定式化とその解決

観測データ (確率変数) を X, 未観測確率変数を Y とし, X, Y が未知の母数 θ をもつある確率分布に従っているとする. このとき, 任意の α $(0 < \alpha < 1)$ に対して, X に基づく区間 $[a(X), b(X)]$ をとって, Y がこの区間に入る確率が $1-\alpha$ 以上になる, すなわち, すべての θ について

$$P_\theta\{a(X) \le Y \le b(X)\} \ge 1 - \alpha \tag{2}$$

となるとき, この区間 $[a(X), b(X)]$ を**信頼度 $1-\alpha$ の予測区間**という. また, X が実現値 x をとるとき, 区間 $[a(x), b(x)]$ を**信頼係数 $100(1-\alpha)\%$ の予測区間**という (図 7 参照).

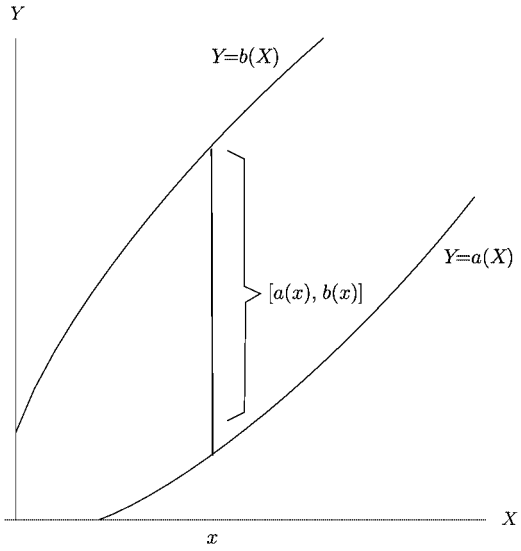


図 7. 信頼係数 $100(1-\alpha)\%$ の予測区間 $[a(x), b(x)]$

4.1. [問題 1] の統計的定式化とその解決

まず, プロ野球で, あるチームが m 試合消化した段階で X 勝しているとき, 残り n 試合での勝数 Y を区間予測しよう. このとき, 試合の結果は勝ちか負けかの 2 通りで引き分けは除いて考え, そのチームの 1 試合当り平均勝率を p とすれば, X, Y はそれぞれ 2 項分布 $B(m,p)$, $B(n,p)$ に従うと考えられる. ただし, p は $0<p<1$ で未知とする. また X, Y は互いに独立, すなわち任意の x, y について $P\{X=x, Y=y\} = P\{X=x\}P\{Y=y\}$ と見なせるから, $T=X+Y$ とおくと, T も 2 項分布 $B(m+n,p)$ に従う. さらに, $T=t$ を与えたときの Y の条件付確率分布 $f_{Y|T}(y|t) = P\{Y=y|T=t\} = P\{Y=y, T=t\}/P\{T=t\}$ $(P\{T=t\}>0)$ は, 超幾何分布になり, これは p に無関係になる. そして $f_{Y|T}(y|t)$ は図 8 のように与えられる. そこで, m, n が大きいとき, $f_{Y|T}(y|t)$ を正規分布によって近似して, (2) より各 t について分布の両裾の確率 (面積) がそれぞれ $\alpha/2$ となるように定められた予測曲線 $Y=a(X)$, $Y=b(X)$ を近似的に求めることができ, 信頼度 $1-\alpha$ の予測区間も得る.

次に, その応用をもっと具体的なデータに基づいて考えてみよう

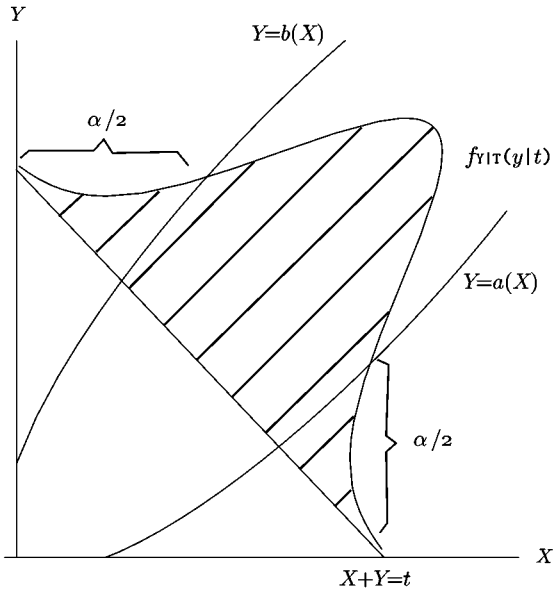


図 8. $T = t$ を与えたときの Y の条件付確率分布

例 1 (ミラクルは起こるのか?). 日本のプロ野球も大詰めを迎えた (1999 年 9 月 8 日) 現在, セ・リーグにおいて巨人は 2 位であるが, 果たしてミラクルは起こるのか? そこで, 横浜, 中日も含めた残り試合での勝数 Y の区間予測を行うと, Y の信頼係数 $100(1-\alpha)\%$ の予測区間と予測曲線を得る (表 1, 図 9 参照).

信頼係数 (%)	中日	巨人	横浜
99	[6.699, 19.861]	[5.444, 18.403]	[5.870, 19.924]
95	[8.349, 18.437]	[7.018, 16.952]	[7.533, 18.308]
90	[9.194, 17.686]	[7.829, 16.191]	[8.395, 17.466]
80	[10.167, 16.802]	[8.767, 15.301]	[9.396, 16.484]
70	[10.821, 16.195]	[9.400, 14.693]	[10.074, 15.816]
60	[11.340, 15.708]	[9.904, 14.206]	[10.615, 15.282]
50	[11.783, 15.286]	[10.335, 13.785]	[11.079, 14.822]

表 1. 残り試合での各チームの勝数の予測区間

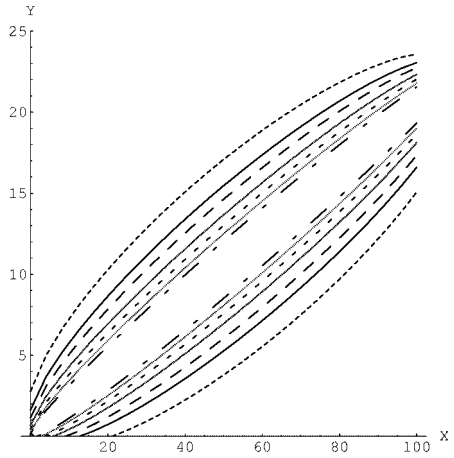


図 9. 中日の勝数 Y の予測曲線

信頼係数99%　信頼係数95%　信頼係数90%
信頼係数80%　信頼係数70%　信頼係数60%
信頼係数50%

また, 前半が終了した 1999 年 7 月 24 日現在のセ・リーグの上位 3 チームの成績は次表のようであった.

チーム	試合数	勝数	負数	引分	残り試合数
中日	83	50	33	0	52
巨人	81	44	37	0	54
横浜	81	40	41	0	54

表 2. 1999 年 7 月 24 日現在の 3 チームの成績

このとき, 各チームの後半での勝数の信頼係数 $100(1-\alpha)\%$ の予測区間と予測曲線を得る (表 3, 図 10 参照).

信頼係数 (%)	中日	巨人	横浜
99	[19.430, 42.233]	[17.159, 41.077]	[14.697, 38.698]
95	[22.285, 39.784]	[20.025, 38.387]	[17.471, 35.899]
90	[23.750, 38.488]	[21.510, 36.976]	[18.918, 34.441]
80	[25.439, 36.959]	[23.232, 35.324]	[20.606, 32.743]
70	[26.576, 35.909]	[24.398, 34.196]	[21.755, 31.588]
60	[27.477, 35.064]	[25.327, 33.292]	[22.673, 30.667]
50	[28.247, 34.333]	[26.123, 32.512]	[23.463, 29.875]

表 3. 後半戦における 3 チームの勝数の予測区間

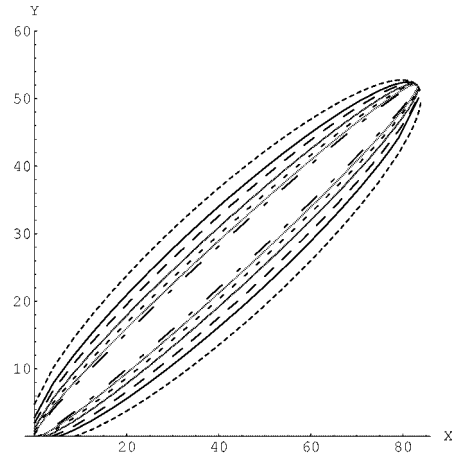


図 10. 中日の勝数 Y の予測曲線

信頼係数99%　信頼係数95%　信頼係数90%
信頼係数80%　信頼係数70%　信頼係数60%
信頼係数50%

4.2. [問題 2] の統計的定式化とその解決

プロ野球で, ある選手がある時点で, それまでに打ったホームラン数 X に基づいて残り試合におけるホームラン数 Y を区間予測しよう. このとき, その選手の 1 試合当たりの平均ホームラン数を λ とすれば, X, Y はそれぞれポアソン分布 $Po(m\lambda)$, $Po(n\lambda)$ に従うと考えられる. ただし, $\lambda > 0$ で未知とする. また, X, Y は互いに独立と見なせるから, $T = X + Y$ とおくと, T もポアソン分布 $Po((m+n)\lambda)$ に従う. さらに, $T = t$ を与えたときの Y の条件付確率分布 $f_{Y|T}(y|t) = P\{Y = y|T = t\} = P\{Y = y, T = t\}/P\{T = t\}$ $(P\{T = t\} > 0)$ は 2 項分布 $B(t, n/(m+n))$ になり, これは λ に無関係になる. そして, $f_{Y|T}(y|t)$ も図 8 と同様の形で与えられる. そこで, [問題 1] の場合と同様にして, m, n が大きいとき, $f_{Y|T}(y|t)$ を正規分布によって近似して, (2) より各 t について分布の両裾の確率 (面積) がそれぞれ $\alpha/2$ となるように定めた予測曲線 $Y = a(X)$, $Y = b(X)$ を近似的に求めることができ, 信頼度 $1 - \alpha$ の予測区間も得る.

次に, その応用をもっと具体的なデータに基づいて考えてみよう.

例 2 (マグワイアとソーサはあと何本ホームランを打つか). 米国の大リーグのソーサ選手とマグワイア選手は, 1999 年 9 月 6 日現在, ソーサ選手は 136 試合消化した時点で 58 本のホームランを打って, 残り試合は 26 試合である. マグワイア選手は 139 試合消化した時点で 54 本のホームランを打って, 残り試合は 23 試合である. そ

の時点での各選手のホームラン数を X とするとき各選手の残り試合でのホームラン数 Y の区間予測を行うと, Y の信頼係数 $100(1-\alpha)\%$ の予測区間と予測曲線を得る (表 4, 図 11, 12 参照).

信頼係数 (%)	ソーサ	マグワイア
99	[3.174, 22.060]	[2.016, 18.798]
95	[4.814, 19.132]	[3.428, 16.147]
90	[5.712, 17.711]	[4.207, 14.864]
80	[6.799, 16.134]	[5.154, 13.445]
70	[7.564, 15.108]	[5.824, 12.523]
60	[8.190, 14.313]	[6.374, 11.811]
50	[8.740, 13.646]	[6.858, 11.214]

表 4. ソーサ, マグワイア両選手の残り試合でのホームラン数の予測区間

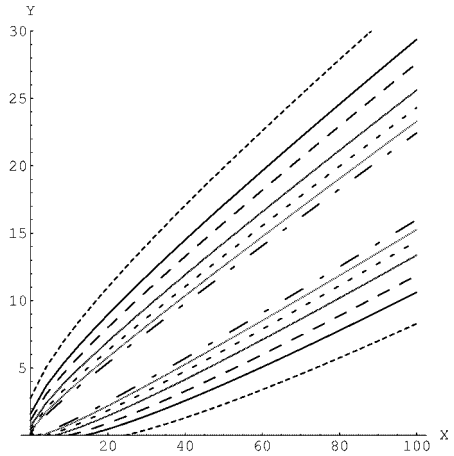
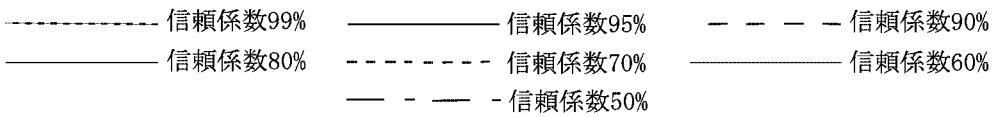


図 11. ソーサのホームラン数 Y の予測曲線



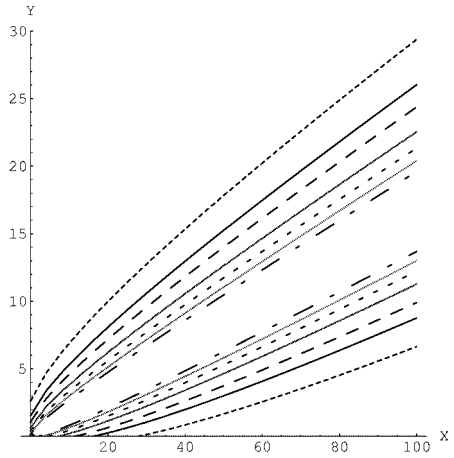


図 12. マグワイアのホームラン数 Y の予測曲線

信頼係数99%　信頼係数95%　信頼係数90%
信頼係数80%　信頼係数70%　信頼係数60%
信頼係数50%

5. おわりに

本論の統計的定式化による区間予測は, 第 4 節の現実の問題への適用結果からみて妥当なものと思われるであろう. ここでは, プロ野球の話題について考察したが, 気候の予測や経済予測の問題などにも適用可能であろう. なお, 本論の第 2 節は [1], 第 4 節は [2], [3] を参照した.

参考文献

[1] Gott III, J. R. (1997). A grim reckoning. *New Scientist*, 36–39, Nov. 15 (邦訳: サイアス (朝日新聞社), 1998 年 1 月, 78-79).

[2] 飛田英祐, 赤平昌文 (1999). 離散指数型分布族における区間予測とその応用. 京都大学数理解析研究所講究録 (掲載予定).

[3] 竹内啓 (1975). 統計的予測論. 培風館.