歯科疫学統計
－第1報　各種統計分布の相互関係と利用の潮流－

瀧　口　　徹

A review of oral epidemiological statistics
－ Part I: Trends in the interrelationship and application of various types of statistical distribution. －

Toru Takiguchi

要旨：とかく統計学は難解な数式と専門用語が先行し医療関係者を困惑させ効率的な利用を阻害している感がある。さらに各種研究をまとめるには統計処理は必須であるが、本来個別性の高い保健・医療の世界を従来型、すなわち正規分布を大前提とし統計分布の中心部分（平均値の周囲）での議論から脱却しないでいることにより近未来の個別医療へのニーズの変化への専門的対応が遅れる懸念がある。そこで第1報では各種統計分布の相互関係と利用の潮流を、1．代表的各種統計分布の相互関係、2．ロジスティック分布の有用性、3．ポアソン分布の有用性、および4．各種統計分布使用の時代変遷、に分けて概説した。

キーワード：統計分布、ロジスティック分布、ポアソン分布、適合度検定、統計モデル

はじめに

医歯学の（臨床）疫学領域の統計解析において仮定する分布（distribution）を何にするかということは他の多くの分野同様極めて重要である。分布が不明な場合やあやふやな場合、ましては間違って適用した場合は研究者が日常的に使う有意差はもはや飾りにしか過ぎなくなる。ただでさえ難しい予測も当たるも八卦当たらぬも八卦の科学的根拠の薄い単なる予想にしか過ぎなくなる恐れがある。時代は個々個別を問題とするレディメードの保健・医療からオーダーメードの保健・医療に向かいつつある端境期（はざかいき）の今日、この統計分布選択の最適化の問題は予防手段や治療法を選択する場合にも（臨床）疫学の研究をする場合にも避けて通れない問題という認識から本稿を起こすこととした。既に他誌[1), 2), 3)]において基本的な統計の概念と処理法および統計学的な検証に基礎を置くEBMの概念について総説として示してあるので、今回は有意性とは、分散とは、検定とは、共分散とは、相関とはということには触れず統計分布およびそれを利用したモデル解析に絞って概説することにした。

周知のごとく代表的な連続分布は正規分布、t分布、χ^2分布、F分布等であり、離散分布では二項分布を代表とし、出現頻度が希な事象ではPoisson分布が、弱伝播性の事象にはPolya-Eggenberger分布がより適切であるとされている。しかし、こ

【著者連絡先】
〒341-0003　埼玉県三郷市彦成3-86
深井保健科学研究所
主席研究員　瀧口　徹
TEL&FAX：048-957-3315
E-mail：taki8020@mth.biglobe.ne.jp

うした教科書的な統計学と実際の使用とは多分に乖離があることが予想されたので、分布に関してその基本を押さえた上で現在の統計モデル分析の潮流を把握し今後の歯科疫学研究の統計学的な示唆を得ることを本総説の主眼とした。

なお、重要であるが本論の趣旨に関して補足的な情報は本文中ではなく極力図表の注釈に解説したので参考にしていただきたい。

1．代表的各種統計分布の相互関係

図1をみていただきたい。ここには20の代表的

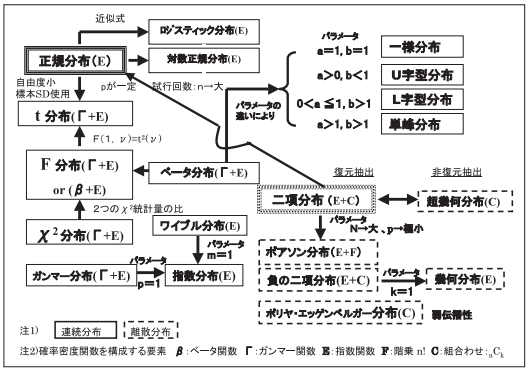

図1　代表的各種統計分布の相互関係

図１の注釈

Note:

指数分布：Exponential dist.
指数関数で表される分布。　正規分布、t分布、F分布、χ^2分布等の確率密度関数の構成要素となる。医学領域においての応用は指数関数とともに生存率解析の基礎分布となる。

ベータ分布：β dist.
a,b2つのパラメータを持ち、その採り方によって一様分布、U字型分布、L字型分布、単峰分布など様々な分布になる。　医学領域においての応用は指数関数とともに生存率解析の基礎分布となる。

ガンマ分布：Gamma dist.
1からnまでの正数を掛け合わせた階乗**n！**(足し算の和であるシグマ：Σに対して掛け算の和はパイ：Πで表す)を実数まで拡張したもの。ガンマ：**Γ(x)**で表す。t分布、F分布、χ^2分布等の確率密度関数の構成要素となる。

ワイブル分布：Weibull dist.
電化製品等の機械の故障や破壊データの分布は 正規分布にならない。故障、破壊現象は、材料の最も弱いところにできた損傷が一気に拡大するというメカニズムのため材料の平均的な耐久性とは関係なく、最弱点だけで決まるためである。時間経過に伴う瞬間故障率は，形状母数 mによって変化する（αは尺度母数）。m＜1 のときには単調減少で初期故障型，m=1 ならば一定で偶発故障型，m＞1 のときには単調増加で摩耗故障型となる。なお，m=1 の場合はワイブル分布は指数分布に一致する。
医学領域においての応用は指数関数とともに生存率解析の基礎分布となる。

一様分布：Uniform dist.
別名「矩形分布」。すべての結果が等確率で起きる分布でβ分布の特殊型。データ数値を四捨五入したときの誤差分布等。

幾何分布：Geometric dist.
ベルヌーイ試行において x回の失敗のあと x+1 回目に初めて事象 A が起こる確率を考える。すなわち，x 回は事象 A が起こらないのでその確率は q x であり，x+1 回目に事象 A が確率 p で起こるので，その積が求める確率となる．式が幾何級数になっているためこう呼ばれる。

超幾何分布：Hypergeometric dist.
有限母集団から標本を非復元（母集団に戻さない）で抽出する場合に厳密には二項分布ではなく超幾何分布になる。二項分布と平均値は同じだが、分散が異なる。多くの標本調査では二項分布と見なして統計を行うが、厳密さを要求される製品の不良品の抜き出し調査では本分布が多用される。

ロジスティック分布：Logistic dist.
説明変数の線形合成式：Zを用いて確率（密度）関数：F(x)=1/(1-exp(-Z))の分布。1948年の米国フラミンガムコホート研究において開発された。正規分布との近似性が高く、またその累積分布は成長曲線に近似し生物学的妥当性が高くなること、および計算のし易さから医学領域の研究において重回帰分析に取って代わりつつある。Z＝1．7xとしたとき標準正規分布に近似する。

1．代表的各種統計分布の相互関係

な統計分布の相互関係を類型化して示した。連続分布の代表である正規分布と離散分布の代表である二項分布を主軸とし、元となる一般分布とそのパラメータが特殊な場合に発生する分布との関係を矢印で示した。例えば汎用されているt分布は小数例の場合で、かつ標本のSD：標準偏差を用いることによる正規分布の変形型ということがわかる。さらに、F分布との関係では1つの自由度が1の場合のF値=t値の2乗となり、t分布はF分布の特殊な場合の分布であると言える。また図1において各種分布の確率（密度）関数を構成する基本的な関数を各分布名の（　）内に、β、Γ、E、F、Cの略語で示しており、連続確率（密度）関数を構成する基本関数はγ関数（Γ）と指数関数（E）が、離散型の確率関数を構成する基本関数は指数関数（E）と${}_nC_k$：組合わせ（C）であることが理解できる。ここでΓ関数は階乗：n!を実数にまで拡張したものであるが独立した事象が連続して起きる場合、その確率は足し算ではなく掛け算になることがこの関数が構成要素の主役になる所以である。連続分布においては周知のごとく正規分布が基本であり、いわば王様である。これに対してβ分布[4]は2つのパラメータの違いによって一様、U字型、L字型、単峰分布、さらにF分布と自在に変化し、いわば変化自在な魔女的存在である。U字型、L字型等の分布は医学領域においてもしばしば登場[5),6)]する応用範囲の広い分布であることがわかる。またγ分布[4]とワイブル分布4)もパラメータによってL字型と単峰型の分布に変化する。離散分布においては有限母集団の場合、抽出データをその都度元に、すなわち母集団に戻す（復元抽出）か、戻さない（非復元抽出）かによって二項分布か超幾何分布かが決まり、「無作為に選ばれた5人とも血液型がAB型である確率計算」等、厳密には後者なのに標本サイズ：nが大きいので前者と見なして計算されることが多いことが気づかれる。しかし母集団が有限で比較的小さい場合（不良品検査等）はその確率分布の違いを注意しなければいけない。

図2は図1の補足であり4種の代表的離散型統計

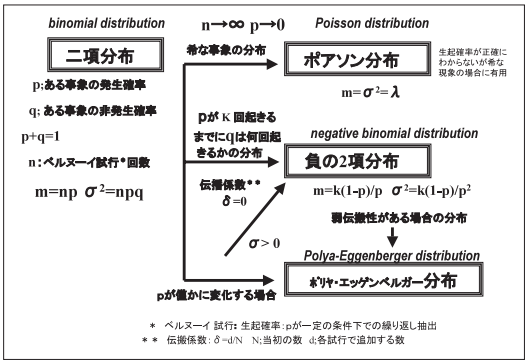

図2　4種の離散型統計分布の相互関係

分布の相互関係を示してある。二項分布とポアソン分布の関係は後者は標本サイズ：nが大きい場合で、かつ生起する確率：pが極めて小さい場合という説明が教科書的[4), 7)]であるが、よく考えると「pが極小であろうと二項分布なんだからそのまま二項分布で計算すればいいじゃないか」という素朴な疑問がでると思われる。ポアソン分布の存在意義はまさにそこにある。この点は3項で詳説する。

図2のポリア・エッゲンベルガー分布[8)]は細菌汚染等の弱伝播性がある場合に利用される分布であり、病原性が高まったり希釈されたり母集団比率が変化する場合である。その伝播力の強さをパラメータを伝播係数：δで表す。この分布は微生物の感染の場合のみに適用されるのではなく、使用例は非常に少ないものの本邦において病院の入院および外来患者の疾患について医師が使用する約1万8千の疾患名の10年間の使用状況がポアソン分布ではなく本分布に適合したという興味深い報告[9)]があり、行動科学に関連する情報伝達を扱うときに考慮すべき分布であろう。

2. ロジスティック分布の有用性

図3にロジスティック分布を用いた回帰モデル分析（=logit分析）を医学領域で用いる場合の有用性を示す。強調されるべきは、多くの医学データがカテゴリーデータ（離散量）で、しかも目的変数は疾病か健康か、手術か経過観察かというような意志決定を必要とする判別分析的な特性に本分布が向いているという点である。不連続のデータには厳密には連続分布である正規分布の適用ができない。また線形の重回帰分析では特に確率分布の両端において生物学的妥当性の問題が生じる。これらのことを解決するために説明変数の線形合成式：Zを指数関数に組み込んだ確率密度関数：p(x)=1/(1-exp(-Z))が開発された。これをロジット変換：log　p(x)/(1-p(x)))すると図3の④に示すような線形式が得られる。本分布は図4、図5に示すように正規分布との近似性が非常に高く、またその累積分布は成長曲線に近似（図5）し生体事象を扱う研究の多くで生物学的妥当性が高くなること,および計算のし易さから医学領域の研究において重回帰分析に取って代わりつつある。なお、Z＝1.7xとしたとき標準正規分布（平均値0、分散1）に近似する。より詳細な本分布の解説お

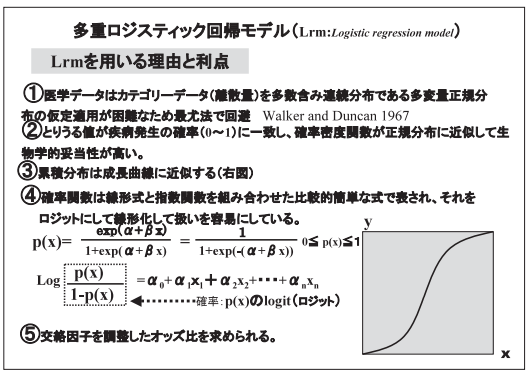

図3　多重ロジスティック回帰モデルの有用性

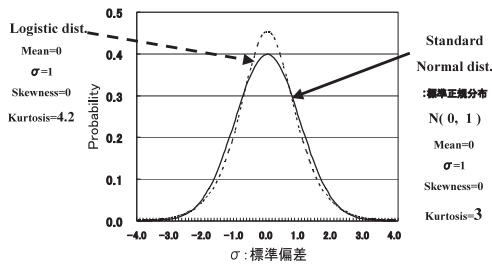

図4　正規分布とロジスティック分布の違いNo.1
－確率密度関数－

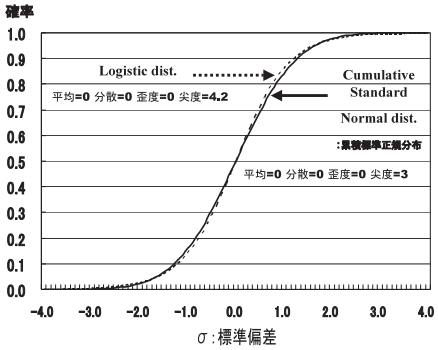

図5　正規分布とロジスティック分布の違いNo.2
－分布関数（別名　累積分布関数）－

よびその活用の実際については丹後らの教科書10）を参照されたい。

3. ポアソン分布の有用性

図1、図2においてポアソン分布の位置づけを示したが本項においてはその有用性について言及する。まず導入編として表1を見ていただきたい。これは著者個人のアドレスに各種企業から30日間に届いたメール数の分布である。交換台にかかってくる電話の回数がポアソン分布すると統計の教科書[11]で言われているので、少し目先を変えてメール数を調べてみたというのが理由である。複数のウエッブサイト関連会社、カード会社、航空会社、鉄道会社、各種製品のネット販売会社等々からのメール（いわゆる迷惑メールを含む）は一月で総計681、一日平均23通と予想以上に多かった。余談であるが先頃のニュースによるとマイクロソフト社のオーナーで世界一の富豪であるビル・ゲイツ氏の場合、1日約400万通だそうであるから著者の場合その20万分の1でとても足下にも及ばない数である。

まず、得られた表1の数値をみて何がわかるであろうか。ポアソン分布を特徴付けるのは確率：pが極めて小さい場合の二項分布であり、その分布を識別する母数：parameter*1の第1の特徴[4), 7)]は平均値＝分散＝λである。この点についてみると平均値が22.70に対して21.18と（主観的にみて）かなり近似している。また尖度も近似している。しかし、歪度は符号が逆である。すなわち母数の比較だけでは断定的なことはいえない。次に図6をみていただきたい。これは実際の度数分布（棒）とポアソン分布（折れ線）および正規分布（折れ線）を比較したものである。かなり適合がいいように見えるが、これでも断定はできない。つまり適合度検定が必要である。図6右端に示す適合性の検定(χ^2検定)を理論値1未満を統合する通法に従って行ったところ実際値とポアソン分布予測値との間で差が無いという帰無仮説はp=0.7672という高い値で棄却できなかった。また正規分布に関

表1　各種企業からの一日当たりメール数の基礎統計量

		Poisson 分布の場合	
例数:**N**	681		
平均:**MEAN**	22.70	22.70	λ
標準偏差:**SD**	4.60	4.76	√λ
分散:**VARIANCE**	21.18	22.70	λ
歪度:**SKEWNESS**＊1	-0.24	0.21	1／√λ
尖度:**KURTOSIS**＊2	3.07	3.04	3+1/λ

＊1:歪度はプラス値の場合、平均値からみて右側に裾が広がり(右側の面積が大きい)、マイナスの場合は左側に裾が広がる。分布の頂部はこの逆の位置にあるので解釈時に注意。

＊2:本総説では尖度を正規分布の場合3としているが教科書、解釈上の容易さの観点から統計ソフトによっては値から3を減じて正規分布では0にしている場合があるので注意。

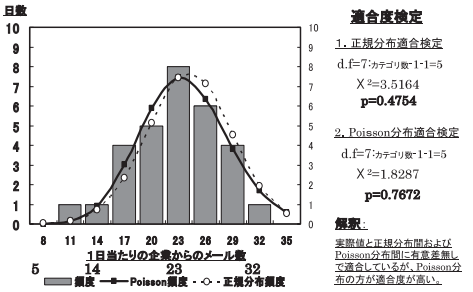

図6　Poisson分布を示す各種企業からの一日当たりメール数
－2004/10/19 11/17までの30日間－

注）適合度検定は理論値の大きさが5未満の周辺度数を統合して10カテゴリーを4カテゴリーとしてχ^2検定。自由度：dfはカテゴリー総数：4から標本から推定した母数（平均値）の数を減じて3とした。

しても p=0.4754 で棄却できなかった。通常の検定で議論する p<0.05,0.01,0.001 を考えたときしっくりこないかも知れないが解釈は同じである。つまり仮に「差あり」と扱った場合、つまりポアソン分布と不適合とすると通常の有意水準で使用する5%の危険率を遙かに凌ぐ77%の危険率（type Ⅰ error）が、また正規分布と不適合とすると48%の危険率があるという意味である。むろん扱っているデータは連続量ではなく不連続量であり、この意味からも厳密には正規分布として扱えない。従って「ポアソン分布でないとは言えない、踏み込んで言えばポアソン分布である」と結論付けられる。図1、図2で示したようにポアソン分布の本家は二項分布である。本件の場合、未知であるが生起する確率：pは以外に高いかも知れない。その場合はポアソン分布よりも二項分布の適合がいいはずであると考えるのがむしろ自然である。しかし、実はこの考えには落とし穴がある。ある企業が有る特定のメールアドレスにメールを出す確率は全くわからない。さらに多数の企業全体ではほとんど検証不可能である。すなわち本事例の場合生起する確率：pが皆目不明であり二項分布は計算できない。さらにもっと厳密に思考すればメールを発信する企業の数と企業戦略は一定でなく時々刻々変化するはずであるから、むしろ図1に紹介した特殊な分布である超幾何分布が近いかも知れない。このように様々の複雑な背景要因がある事象にポアソン分布は極めて有用となる。その有用性をより明確に理解いただくため次の事例を紹介する。

次の事例は毎年甚大な自然災害をもたらす台風に関してである。周知のごとく2004年はかってないほどの数の台風が本土に上陸し、その数は10回を数えた。地球温暖化の影響だという話がマスコミを中心に流布された。しかし真実は全くの偶然で実はそう珍しくない現象かも知れない。そこで何らかの科学的推論をするために門外漢であるがチャレンジしてみた。公表されている気象庁のデータを基に1951年からの台風発生数、本土上陸数および上陸率%のデータを図7に示した。一見して本年に限り発生数が多いため結果的に上陸数が多くなったという仮説は成り立たないことがわかる。また地球温暖化が事実としてもそのため台風発生数あるいは本土上陸数が年々増加しているという傾向も否定されるだろう。要は2004年だけ、しかも本土上陸数だけが異常に多いと考えられるのである。ここで1951年から2003年(2004年を除く)の年間台風発生数の平均値および標準偏差は26.72±4.75である。2004年は12月10日現在27個の発生があったがこれは平均的な発生数である。しかし本土上陸数は10回と極端に多い。上陸率も37％と群を抜いて高い。表2に年間本土上陸台風数の分布が何かを検証するための比較を示した。いずれも異常値と思われる2004年を外した53年分の実績である。実際値の分布の4つのパラメータ（平均値、分散、歪度、尖度）からは正規分布

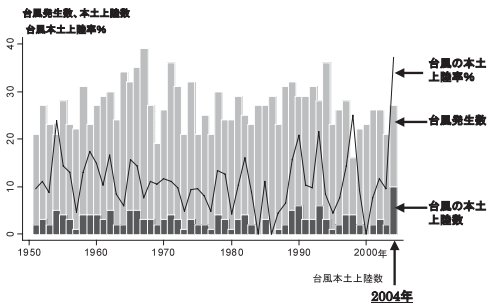

図7　台風の発生数と日本への上陸数および上陸率の年変化
1951年～2004年（12月10日）出典：気象庁

表2　年間の台風本土上陸数の分布は何か？

	実際値*1	正規分布*2	二項分布*3	Poisson分布*4
例数	150			
平均	2.83	2.83	2.83	2.83
標準偏差	1.45	1.45	1.59	1.68
分散	2.11	2.11	2.53	2.83
歪度	0.11	0.00	0.50	0.59
尖度	2.72	3.00	3.17	3.35

＊1:実際値の基礎統計量は異常値と考えられる2004年を外して1951～2003年の総計53年分の気象庁公開データから算出
＊2:正規分布は実際値の平均と標準偏差(分散)を一致させて算出
＊3:二項分布の計算で上記表の値は試行数を26.717とし、推定値の計算は年平均発生数を正数27として計算。確率:pは1951～2003年の平均上陸率の0.1059として計算
＊4:Poisson分布の計算は実際値の平均を一致させて算出(自動的に分散=平均値となる)

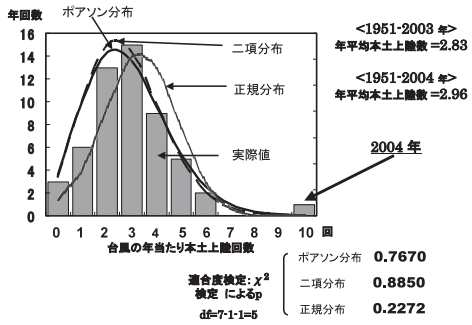

図8　年間本土上陸台風数の分布は何か？
※折れ線で示す3種の分布は2004年を除いて1951-2003年の53年分のデータから求めている。

表3　台風の上陸回数の予測確率と実際値

1. 台風本土上陸数>=10の場合の確率		
	1951-2003年	1951-2004年
10回のz値	4.94	4.05
正規分布	0.00000039	0.00002549
二項分布	0.00004461	0.00006869
ポアソン分布	0.00017931	0.00026360
実際値	0回/53年=0.0000	1回/54年=0.0185

2. 台風本土上陸数=0の場合の確率		
	1951-2003年	1951-2004年
0回のz値	1.50	1.17
正規分布	0.02555288	0.04403122
二項分布	0.05002902	0.04300611
ポアソン分布	0.05900172	0.05166561
実際値	3回/53年=0.0566	3回/54年=0.0556

と比較してやや右裾が広がり、尖度が低い分布をしていることがわかる。これに対して二項分布、ポアソン分布との比較ではともに分散、歪度および尖度がやや小さい分布をしていることがわかる。図8は実際値（棒グラフ）と正規分布、二項分布およびポアソン分布の曲線を示している。いずれの分布も適合度検定（χ 2検定）で有意差はなくいずれの分布であっても不適合とは言えない。しかしながら正規分布は図8で視覚的にも確認できるように他2つの分布と比較して適合が悪い。すなわちp値（p=0.2272）が他の2分布より有意水準に近い。pの大きさから判断して3つの分布のうち最も適合度が高いのは二項分布（p=0.8850）、これに近似してポアソン分布（p=0.7670）である。ポアソン分布が二項分布の適合度よりやや低いことは台風の上陸が「希な事象」とは必ずしも言えないことを示している。事実、平均の本土上陸率は約10％であり、ポアソン分布の条件であるpが極小という点には若干の問題がある。しかしながら二項分布の計算に用いた台風発生数は平均値を整数化した27であるが毎年同じ数の台風が発生するわけではなく、発生数のばらつきが大きいと二項分布の推定誤差を引き起こす。この点台風発生率：pと試行回数：nに無関係な分布であるポアソン分布は頑健かつ母数がシンプル（平均値＝分散＝λ）であり、本例でも台風発生数の情報がパラメータとして全く考慮されていない（台風本土上陸数の平均値2.83のみを使用）にもかかわらず二項分布と極めて近似した分布を得ることができている。これがポアソン分布の有用性である。

ここで本題の2004年の10回の来襲は確率的にどの程度異常かということを表3に示す。まず比較として年に1回も本土上陸が無い確率をみていただきたい。正規分布と仮定すると確率0.026、すなわち100年に2,3回程度発生する現象となり、二項分布では0.050、ポアソン分布でみると0.059、すなわち100年に5,6回発生するという確率になる。実際値は53年間のうち3回で確率は0.057であるからポアソン分布の予測確率がよく当てはまっている。つまり50年、100年の間に年に1回も本土上陸がない年が数回あっても異常とは言えないということになる。これに対して年10回以上の上陸はどうだろうか。2004年までの実際値の確率は1回/54年=0.0185、すなわち100年に2回ほどであるが、年10回以上本土上陸する確率は正規分布の場合は2004年を含めて計算しても10万年に2,3回の確率、二項分布は10万年に6,7回、確率の裾野が広いポアソン分布でも1万年に2,3回の確率にしかならない。従って本データで見る限り2004年の年10回の本土上陸はいままでの気象環境では本来は起こりえない極めて異常な現象といえるだろう。しかもそれは前述のように発生数の増加と関連がなく、経年的な傾向でもない。専門ではないし本総説の趣旨からはずれるのでこれ以上の議論は避けるが海流、気流等台風のコースをコント

表4　ポアソン回帰モデル分析の有用性

Poisson regression model analysis

ポアソン回帰モデル分析

①ある事象の発生率；λがポアソン分布をすると仮定され

②その発生の平均値が要因β_1～β_nで説明されると仮定される場合

③対数を掛けることにより線形予測モデルができる

$$\mathrm{Log}(\lambda)=\beta_0+\beta_1+\beta_2+\cdots+\beta_n$$

1)歯科クリニック保健・医療水準の関連要因分析　患者要因、口腔状態、地理的要因

2)齲蝕の多寡の社会経済学的背景要因分析

3)口腔ガンの増加要因分析（スペイン1975-1994）

表5　離散型の統計分布を用いた統計分布論文の時代変遷（1970年代-2000年代）

－メディカル・データベースPub Medによる－

分布の種類 / 年代		Binomial Dist. 二項分布	Negative Binomial Dist. 負の二項分布	Poisson Dist.# ポアソン分布	Polya-Eggenberger Dist. ポリヤ・エッゲンベルガー分布
規登録文献数の年平均	1970年代	2.5	0.5	3.5	0.0
	1980	13.8	3.1	26.4	0.1
	1990	47.4	6.4	113.4	0.3
	2000※	69.7	24.0	175.8	0.0
文献数の実数※		968	214	2,268	4

※ 1970年1月1日－2004年9月30日　＃ Poisson regression を除く

ロールする要因系の偶然ではない変化が背景にあるのかも知れない。

ポアソン分布はフランスの数学者S.D.Poisson(1781-1840)が1837年に発見して発表し、それから60年もたってからプロシアの陸軍連隊の「毎年、馬に蹴られて死亡する兵士の数別の連隊数の分布がポアソン分布する」という奇妙な初適用(Bortkiewicz 1898)[7]で重要性が世に知られるようになった統計分布であるが、医学領域における使用状況については後段4．で解説する。

以上、ポアソン分布は非常に有用性の高い分布であるが、さらに4項で言及するように近年ポアソン回帰モデル（Poisson Regression Analysis）の応用が世界的に普及してきた。ポアソン分布する事象（例えば患者の受療行動等）を多要因で説明する場合に対数処理することによりあたかも重回帰分析のような線形解析を行うことができる手法である。表4にこのモデル分析の有用性を示したので参照にされたい。

4．各種統計分布使用の時代変遷

図1に示す連続型分布である正規分布、t分布、F分布およびχ^2分布等は研究者にとっていわば日用品で使用頻度は極めて多く、あらゆる科学研究の基礎中の基礎であるため逆に文献検索データベースのアブストラクトに記載されていない場合も多いと考えられる。従って時代変遷を調べる意味合いと信頼性は殆どないと考え、本論ではこれら連続型の分布を単独では議論の対象にせず、比較的使用頻度の低い二項分布、ポアソン分布等の離散型分布および前提とする分布がそれぞれ異なる各種多変量解析を検索調査の対象とした。

1）離散型の統計分布利用の時代変遷

医歯学・疫学領域における各種離散型分の使用状況の時代変遷をみる目的で代表的な離散分布である二項分布（BD；binomial dist.）、負の二項分布（NBD；negative binomial dist.）、ポアソン分布（PD；Poisson dist.）およびポリア・エッゲンベルガー分布（PED；Polya-Eggenberger dist.）の4種の年平均の出現数を1970年から2004年9月30日までを10年刻みで医学洋文献データベース：PubMed検索を行い、年代ごとの年間平均新規掲載論文数を検索した。

表5に4種の統計分布使用頻度の年代別変遷の結果を示す。1970年から2004年9月30日までの総計で分布解析のBD数が968、NBDが214、（ポアソン回帰分析を除く）PDが2,268、PEDが4であった。圧倒的にポアソン分布が多く、かつ年代が進むにつれ急速に増加する傾向が明らかであった。このポアソン分布を使用した研究の領域は表6に示すように29.6%が死亡関係、10.0.%が事故、傷害関連であり本分布が医学領域においても多要

*1：母数（parameter）：平均値、分散、歪度、尖度等の母集団の分布を特徴づけている指標の総称。

表6　Poisson分布、Poisson regression modelを用いた文献のテーマの内訳

1970年１月１日～2004年9月30日

死亡率関係	→1,061(29.6%)
事故、傷害関係	→ 359(10.0%)
歯科保健・医療（7領域）	→ 129(3.6%)
その他	→2,226(62.1%)

注）該当文献総数 3,586 うち 41 文献がテーマの複合がある。

表7　メディカル・データベースPubMedで歯科保健・医療関連文献を検索するための71のキーワード

1.GENERAL WORDS 23 items
dental, "oral health" ,tooth,teeth, "tooth brushing",toothpaste","tooth pain",enamel, dentin, cementum, dentition, "root canal", "dental pulp", "tooth loss", "missing teeth", maxilla, mandiblar "alveolar bone", saliva, taste, "dental occlusion",masticatory ,"chewing ability"

2.CARIES(prevention & treatment) 12 items
caries, "dental plaque","oral biology","dmf, def, DMF, fluoride, fluoridation,", "sugar consumption","tooth restoration",amalgam,resin

3.PERIODONTAL DISEASES(prevention & treatment) 8 items
"CPITN", "pocket depth", "attachment loss",gingivitis,"gum disease", periodontal,periodontitis , "PMTC"

4.ORAL SURGERY 8 items
"oral surgery", "tooth extraction", "oral cancer", "oral cartinoma" "salivary gland", ameloblastoma, "maxillary sinusitis", tongue

5.ORTHODONTICS 6 items
orthodontics, "Angle classifficaion", crowding, overbite, overjet, cephalometric

6.PROSTHESIS 7 items
articulator,"dental bridge", denture, prosthesis,"dental implant", "porcelain crown","metal bond"

7.SOCIO-ECONMIC STUDY 7 items
"dentist","dental practitioner","dental hygienist", "dental technician", "dental school","dental clinic","dental patient"

表8　各種多変量分布を用いた統計分析論文の時代変遷（1970年代-2000年代）

－メディカル・データベースPub Medによる－

多変量解析法の種類 ＼ 年代		Multiple r.m.	Logistic r.m.	Probit r.m.	C. log-log m.	Poisson r.m.
新規登録文献の年平均	1970年代	37.6	12.5	0.0	1.5	0.0
	1980年代	234.4	243.1	0.6	8.3	2.0
	1990年代	655.1	2,188.8	2.2	12.3	57.9
	2000年代 ※	957.7	5,047.4	5.7	16.0	151.4
実際の文献総数 ※ 1970年1月1日-2004年9月30日		13,820	48,395	55	297	1,318
歯科保健・医療関連文献の総数		392	1,160	2	1	31

Multiple r.m.; Multiple regression model ：重回帰モデル
Logistic r.m.; Logistic regression model ：ロジスティック回帰モデル
Probit r.m.; Probit regression model ：プロビット回帰モデル
C. log-log m.;Complementary log-log model ：（相補二重対数モデル）＝二重指数関数モデル
Poisson r.m.; Poisson regression model ：ポアソン回帰モデル

因で確率の極めて低い現象、特に事故を統計学的に捕らえるために使用されていることがわかる。基礎医学の分野では例えば細菌学において繊毛細菌の運動時間の分布[12]、赤血球計単位面積当たりの細胞数[13]等がポアソン分布することが報告されている。歯科領域での使用は3.6%（129件）であった。

なお、歯科保健・医療に関連した文献を効率よく抽出する目的で、表7の7領域総計71のキーワードからなる検索表を独自に作成し使用した。関連語を広く抽出するPubMedのMesh機能をほとんど使用していないので71のキーワードそのものだけが掲載されている文献を収集していることになる。従って採りすぎは心配ないであろうが類似語、近似語の採り漏らしの問題がある程度内在していることをご承知おきいただきたい。

2）回帰分析モデル利用の時代変遷

また代表的な多変量回帰分析モデルである重回帰モデル（Mrm ； Multiple regression model）,ロジスティック回帰モデル（Lrm ； Logistic r.m）、ポアソン回帰モデル（Prm ； Poisson r.m）等5種の論文の出現数についても同様な検索調査を行った。表8に各種回帰モデル使用頻度の年代別変遷の結果を示した。直線回帰を前提とするMrmと累積正規分布に近似した曲線を前提とするLrmの出現数は両者とも年代とともに大幅に増加しているが1980年代にLrmが逆転して主流になったことがわかる。またPrmを応用した研究が増加傾向にある。これらのうち歯科領域の実績に関してはMrmが全体の.2.8%、Lrmが.2.4％、Prmが0.2％であった。なお、probit回帰モデル[10]は標準正規分布を用いたモデル、またcomplementary log-logモデル[10]は対数変換を二重に行うモデルであり第2報で詳述するGLIM：一般化線形モデル[14]においてLrmとともに統一的に取り扱われている。

1970年代から現在に至るまでコンピュータのハードとソフトの両面の成熟度と普及度を背景として分布解析が多様化しポアソン分布利用が急増し、回帰モデル解析においてはMrmからLrmに主流が移っている。Lrmは前述のように疾病のリスクファクター分析において独立変量である各指標が連続量、離散量、カテゴリーデータ等多種多様な尺度で表される場合の解析に極めて有用であ

る。また予測と要因分析に多用されてきた回帰分析についても同様に機械的な直線回帰でなく、前提とする分布の生物学的妥当性の向上を求めてPrmが増加してきている。このように解析の前提となる分布および関連モデルの選択が厳密化している時代的傾向が確認された。こうした潮流が1989年に第2報で扱う一般化線形モデル法：GLIM法を結実させていくことになる。

文　献

1）瀧口　徹：EBMのための（臨床）疫学・統計学的基礎（1）第1章　統計の基礎；一部のデータから全体を推定する．障害者歯科学雑誌　23：1-10，2002.

2）瀧口　徹：EBMのための（臨床）疫学・統計学的基礎（2）第2章　疫学の基礎；流行病の法則性を見つけ予防する．障害者歯科学雑誌　23：89-98，2002.

3）瀧口　徹：EBMのための（臨床）疫学・統計学的基礎（3）第3章　EBMの基礎；臨床疫学の最近の潮流とポイント．障害者歯科学雑誌　23：443-458，2002.

4）池田　央：統計ガイドブック，新曜社，東京，第4刷，1997，62-67頁.

5）Gaziano JM, Gaziano TA, Glynn RJ et al.：Light-to-moderate alcohol consumption and mortality in the Physicians' Health Study enrollment cohort. J Am Coll Cardiol. 35(1)：96-105，2000.

6）Li Z, Lew NL, Lazarus JM, Lowrie EG：Comparing the urea reduction ratio and the urea product as outcome-based measures of hemodialysis dose. Am J Kidney Dis. 35(4)：598-605，2000.

7）編集　柳川　洋，執筆　橋本　勉，中村好一　ほか：疫学マニュアル，南山堂，東京，第5版，2000、136頁.

8）鈴木義一郎：現代統計学小辞典，講談社，東京，第1刷，1998，215-222頁.

9）Lin ZJ, Yamamoto K, Kamae I, Sasagawa N et al.：The frequencies of disease names with the natural language used in the hospital information system. J Med　Syst. 19(5)：381-385，1995.：

10）丹後俊郎，山岡和枝，高木春良：ロジスティック回帰分析，朝倉書店，1996，14-15頁.

11）木下宗七：入門統計学，有斐閣，東京，第9刷，2004，98-101頁.

12）Paul Lewus, Roseanne M. Ford：Temperature-Sensitive Motility of Sulfolobus acidocaldarius Influences Population Distribution in Extreme Environments. Journal of Bacteriology．181(13)：4020-4025，1999.

13）John M. Last編，重松逸造，春日　斉，柳川　洋監訳：疫学事典，日本公衆衛生協会，東京，1987，129頁.,

14）P. McCullagh, J.A. Nelder：Generalized Linear Models. Chapman & Hall/CRC.., USA., 2nd ed., 1999，pp．30-31.

A review of oral epidemiological statistics

－ Part I: Trends in the interrelationship and application of various types of statistical distribution.－

Toru Takiguchi
（Fukai Institute of Health Science）

Abstract: There is a general perception that statistics is a particularly difficult field, probably due to the plethora of complex formulas and technical terminology. Therefore, medical personnel easily become frustrated and are unable to achieve optimal and effective application of statistics in their research.

A strong grasp of statistical techniques is indispensable for researchers as they try to analyze and understand the results of their research. Fortunately, high-powered computers and statistical software have spread around the world in the past two decades, so younger researchers can take advantage of the relative ease of statistical analysis with this technology.

Nevertheless, my experience has given me strong misgivings about the current use of statistics, especially among younger dental personnel and researchers. I have noticed that their interests and discussions still tend to focus only on the central area of disease distributions (around the population mean), while ignoring cases that are far removed from the mean. Farthest from the population mean on one side, there are special persons who can maintain healthy lives without a sustained daily effort to improve their lifestyle. On the other side we see exceptions to the rule in the opposite direction-people who are unable to sustain healthy lives in spite of their best efforts.

In the near future, probably within this 21st century, the concepts of prevention and treatment will undergo a dramatic transition, from “ready-made medicine” to “order-made medicine”. Therefore, new classifications of disease and preventive methods based on strict analyses of statistical distributions are essential.

From this standpoint, the most common statistical distributions (e.g. normal, binomial, Poisson, Polya-Eggenberger, β, and γ) and the interrelationship among 20 typical distributions are discussed in Part 1 of this review.

Key words: statistical distribution, logistic distribution, Poisson distribution, test for goodness-of-fit, statistical models

Reprint requests to T. TAKIGUCHI, Fukai Institute of Health Science, 3-86, Hikonari, Misato-shi, Saitama 341-0003, Japan
TEL:048-957-3315/FAX:048-957-3315/E-mail:taki8020@mth.biglobe.ne.jp