信頼度の意味するもの

原　宣一

1．はじめに

もう四半世紀以上も昔のことであるが「後悔をしない決断をするには如何にすべきか」について漠然と考えていた時期があった。人生はやり直しが効かないし、適切な判断を下すよりどころとなる数学的論拠があるものならばそれを知りたいと思ったのである。これは数学を駆使して最適戦略を求めるオペレーションズ・リサーチ(OR)が目的とするところでもある。数年後、幸いなことに留学のチャンスを与えられ、フロリダ工科大学(FIT)修士課程でORを専攻し、最適値を求めるいくつかの手法についてある程度の勉強をすることが出来た。統計的意志決定もORの一つのテーマである。

意思決定を合理的にしたいと本気で考え出したそもそものきっかけは航空機の疲労寿命(Ref.1)を決める手法を再検討するように当時の上司から指示されたことである。疲労破壊はばらつきが大きいので寿命に対する安全率を大きく取り、安全に使える寿命を短めに設定せざるを得ないのであるが、その安全率があまりにも大きすぎる(Ref.2)のでもう少し何とかならないだろうかということであった。寿命を長くして商業的価値を高めたい(Ref.3)という上司の意図はもちろん理解できることである。それよりも実際に航空機を運航する人に対して明確な説明が必要だと考えたのである。例えば、着陸装置(Landing Gear)の安全寿命(Safe Life Limit)は着陸６０００回と決めてしまうと、毎日整備していて見掛け上錆びたり摩耗していなくても６０００回の着陸回数に達したら棄てなければならないのである。何故、この脚は今日から使ってはいけないかの論理を説明するのは設計者（会社）の義務であろう。規則により航空局の認可を得たものであるという答えだけでは理由にならない。

さて安全寿命の決め方がどうなっているか、即ち、何が仮定で、試験データのばらつきはどのようなもので、統計的に何が言えるのかを検討した。すると統計的意思決定につきものの信頼水準の設定があまりにも恣意的で根拠がないことに不満を持つに至った。さらに確率論の基礎的な考え方についていくつかの文献をあたった結果、確率統計の歴史は古典的確率論に始まり、大論争時代も経ていることが判った。結局、ベイズ流統計学と言われている方法(Ref.4, 5)が

最も合理的であると確信を抱くに至り、この考え方に沿って疲労寿命の決め方をよりすっきりした形で提案することができた(Ref.6)。破壊確率の許容値や他の仮定は同じにしたままで論理的に安全寿命は伸ばせることが分かった。後に、問題の本質をより明確にするために属性試験結果から信頼度を決める方法も提案した(Ref.7)。

しかし、「・・・流」という形容詞がつく間は学問としては主流でない響きを持つらしく、学会誌への投稿も採用されるところとならず、シンポジウムでの発表も何ら関心を引くことは出来なかった。大学に残り純粋数学を専攻している友人は「確率論に問題があるなら面白い」と言って喜んで頭を突っ込んでくれた。彼は集合論との関係、変数変換と分布の対応関係をすぐ説明してくれたが、「応用分野で確率にどのような意味を持たせるかは数学として何の興味もないよ」とのことで、「工学分野での確率統計論の応用はおかしいところがある」という私の主張に引き込むことは出来なかった。数学的には確率は公理で定義される抽象的な数値であり、公理から出発して導かれる演繹的な定理群にはいささかも欠陥がないし、その公理についても測度論により揺るぎないものになっているのでもう研究テーマにならないということであった。

このような状況のまま転勤等で仕事の内容が変わり２０年近くこの話題を放置してしまった。この間、信頼度に関する状況は殆ど進歩がなかったように思われる。そして３年程前に信頼性管理部に配属されると、信頼性管理の理論的背景として信頼度の定義に使われている確率の捉え方が気になり出した。信頼性管理部での業務上、最も基本的な信頼度に係わる数学的考え方をはっきりさせておきたいのである。統計的意思決定方法として現在一般的に行われていることの何が不合理であるか、また、その打開策はどうあるべきかについてもう一度、ここに示したい。

２．属性試験による信頼度の決め方とその問題点

論点を出来るだけ明確にするため簡単な例として属性試験結果から信頼度を決める方法について考える。属性試験(Inspection by Attribute)とは試験結果が合格か不合格か、あるいは○か×かで表されるような試験を言う。試験結果が数値で表される測定値のようなものでも合格ラインを設定しておき、それ以上か以下かのみを着目すれば、属性試験となる。属性試験としては、例えば分離ボルト等の火工品がある。これらは作動、不作動の結果がはっきりしている。

このような試験において、１個ずつ試験をしていきｎ個の試験したとしてその結果の全体を $\{x_i : i = 1 \text{からｎ}\}$ 、またはベクトル表記で x と書く。合格の場合は $x_i = 1$ 、不合格の場合は $x_i = 0$ とすると約束する。このように決めた x は０か１の値を取るが試験をしてみないと分からない。そして、ある確率ｐで $x = 1$ を取ると考えられる。このような x は確率変数(random variable)と呼ばれるものの最も簡単な例である。

確率統計論では、上記のように確率ｐで１を取り、従って、確率１－ｐで０を取るような確率変数 x はｎ個試験すると結果が確率的にどうなるかが分かっている。例えば、５個試験して５個とも１が得られる確率、５個の内、４個が１になる確率、等が計算できるということである。これを確率変数 x はパラメータをｎ，ｐとする２項分布(Binomial Distribution)に従うと言い、習慣的に $B(n, p)$ と表記する。

さて、あるＡという会社が新しく分離ボルトを開発したので使ってみたいが、重要な場所に使う部品なのでその分離ボルトが確かに作動するものであるかどうかを知りたいという状況にあるとする。実際は、その会社から開発状況や会社の過去の実績等の関連情報が大きく信頼感に作用するのであるが、これらの情報に頼らず何個かの試験結果からのみで判断しなければならないとする。即ち、属性試験結果からその分離ボルトの信頼度について客観的にどのようなことが言えるのかを問題とする。

今、２０個の分離ボルトを試験したところすべて合格であったとする。すなわち、 $\{x_i = 1 : i = 1 \text{から} 20\}$ のデータを得たわけである。つまりＡ社の分離ボルトは合格確率ｐであると考える。しかし、ｐの真の値は永久に未知なパラメータであるものの、試験結果からある程度の推定ができると考えるのである。このｐが分離ボルトの信頼度に他ならない。ｐが０．５ぐらいであれば２０個連続して合格するようなことはめったに起こり得ないことは丁半賭博で丁が２０回も続くようなことが殆ど起こり得ないことを想起すればすぐ納得できる。もし、そのようなことが起これば１００万回に１回程の極めて稀なことが起こったとは考えずに「いかさま」があったに違いないと推察することになる。それではｐが０．８とか０．９ならば２０回も続けて成功することが普通に起こることかどうかということが知りたいところである。

前述のように２項分布の知識からｐがどんな値であろうとｎ個取った場合の

結果がどうなるの確率は計算できるのである。計算するとｐが０．８６１であったとしても５％はこのようなことが起こり得ることがわかる。つまりｐが０．８６１であったとしても、２０個試験するという試行を１００回行ったとして平均として５回の試行で２０個とも合格するという結論が得られる。

（信頼水準の取り方に決めようがない）

そこで現在の確率統計論では「信頼水準を９５％に取るとｐは０．８６１以上であると言える」という表現になるのである。このことから、技術者は「安全側に」ｐ＝０．８６１と推定することになる。

この信頼水準(Confidence Level)とは歴史がある割に全く恣意的で根拠が無いものである。２０個とも合格したというデータはさらに「信頼水準を９０％に取れば０．８９１以上である」とも言えるし「信頼水準を９９％に取れば０．７９５以上である」とも言える。一般に信頼水準としての数値は９０％、９５％、９９％の３種類をよく見かけるが、文書によっては信頼水準６０％という数値も用いられている。信頼水準を何％に取るとの必然性を明確に示した文献は皆無である。ｐの値と信頼水準の値とは性格が異なるのでこのような表現で表すしか方法がないのだとされている。

ある決断を下すに際して、その決断に係わる多くの要素があろうとも、即ち多次元で考えなくてはならない状況であろうとも、常に一次元に変換して何らかのしきい値を越えているかどうかで決断を下すことになる。実際問題の多くは多次元から一次元への変換式も分からないし、しきい値もふらふら定まらない場合が多いことであろう。しかし、状況が同じであれば同じ決断を下す、即ち、首尾一貫した決断を下すためにはこれらを明確に定める必要がある。

「Ａ社の分離ボルトを採用する」という決断を下すためにデータを取ったら信頼度と信頼水準という二つの要素を考慮する必要に迫られたわけである。現在の統計的意思決定では、しきい値の定め方を教えてくれていないわけである。「信頼水準を９０％に取る」という根拠が全くないことに不満を抱かざるを得ない理由である。

実は同じようなことを考えた人はいたのである。これら信頼水準と信頼度の二つの数値を組み合わせて一つにすることを試みたマクダネル・ダグラスの技術者が書いた文献(Ref.8)があったが大方の賛同を得ていない。

（信頼水準付きの信頼度は比較が出来ない）

さらにＢ社が同じ分離ボルトを開発していることが分かって８０個の試験を行ったとする。そしてその結果、７５個合格したが５個不合格であったとしよう。すると２０個とも合格であったＡ社の分離ボルトとどちらの物を採用すべきであろうか。もちろん問題を簡単にするために、コスト比較等他の考慮事項を一切省いて試験結果だけから判断するとの前提である。

Ｂ社の分離ボルトは「信頼水準を９５％に取れば０．８７３以上である」と言え、「信頼水準を９０％に取れば０．８８７以上である」と言えるわけである。従って、この下限値を取って信頼度と推定する方法では、信頼水準を９５％に取ることにすればＢ社の製品の方がＡ社の製品より信頼度が高いという結論になるが、信頼水準を９０％に取ればＡ社の製品の方が信頼度が高いことになる。この関係をFig.1に示す。果たして人は直感的にはどちらを採用したいと考えるのであろうか。

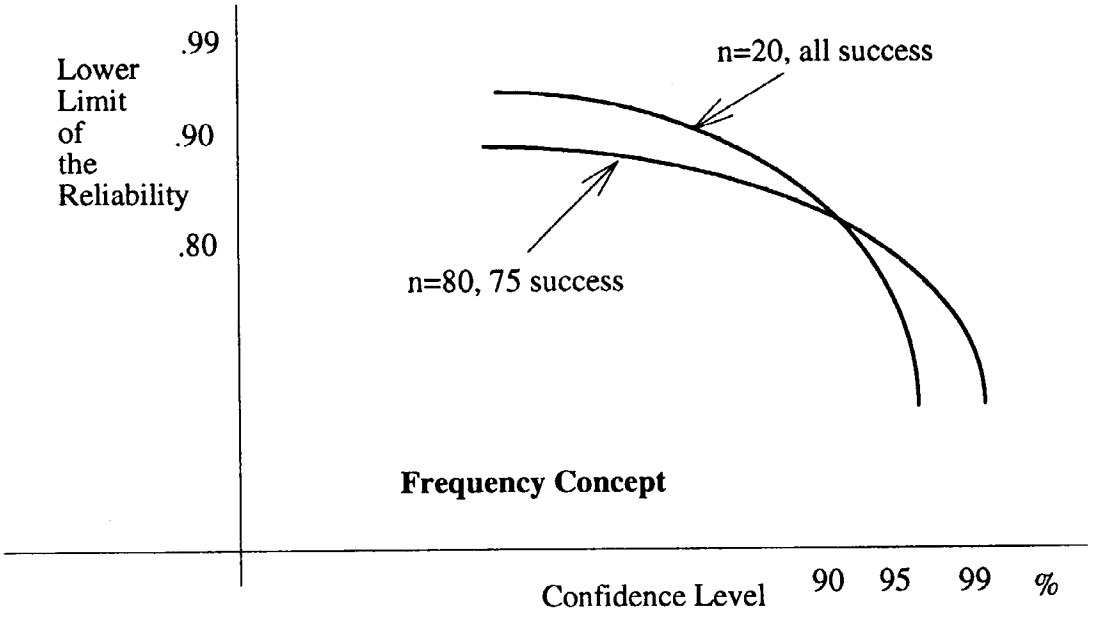

Fig.1 the Reliability (Frequency Concept) depended on the Inspection by Attribute

（信頼水準付きの信頼度は合成が出来ない）

最も簡単なシステム（Ｓ）は二つのコンポーネント（Ａ），（Ｂ）から構成されているシステムである。そのシステムブロック図(Fig.2)がシリーズになっているとするとシステムの信頼度R_Sは二つのコンポーネントの信頼度R_AとR_Bの積で表される。このとき、R_AとR_Bが信頼水準９０％や９５％で推定されたものであるならば積$R_A \cdot R_B$は信頼水準何％であると言えるのか答えようがな

い。仮に両方とも９０％であるとしても積$R_A \cdot R_B$はもはや信頼水準９０％ではない。

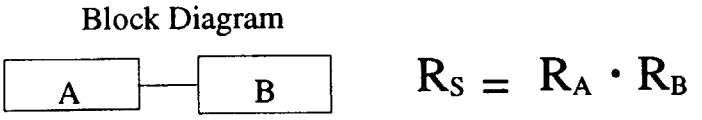

Fig.2 System Reliability

もとの試験データに戻って、２次元の同時分布を考えて計算すれば求められなくもないが、多数のコンポーネントの場合はすぐお手上げになる。システムの信頼度をブロック図を通じて計算するメリットは無くなる。実際は、信頼度計算を行う時に信頼水準は無視されているので問題にならないだけである。信頼度計算に伴う多くの仮定の一つに過ぎないと考えられる。一般に、信頼度計算とは、数字の桁数が多いのに反してその程度のものなのである。

データを見て客観的に判断したいということから確率統計論の知識を活用した筈であるのに、信頼水準の取り方が確定していないために客観的な結論が出せないのである。Ｂ社の方に味方したいと思えば信頼水準は９５％必要ですというであろうし逆の場合は信頼水準は９０％で良いと主張することになる。これでは合理的に決めようとしたのに不合理な決め方しか出来ないわけで一つのパラドックスである。まして、供試体の数が違うので判断出来ないという結論では確率統計論の知識を持ちだした意味がない。しかし、これまでのところ判断の基準としての信頼水準は９０％で良いのだというような見解で済まされてきていること自体が驚くべきことではなかろうか。高い信頼度を求めなければいけないような宇宙技術でそれでも良いのであろうか。これまで信頼水準は９０％とすることで良かったからということであれば、少なくとも何故９０％で良かったかの説明が欲しいものである。

３．確率の定義

上述のように「信頼水準Ｃ％で信頼度はＲ以上である」という控えめな表現しか出来ない理由は確率の定義に問題があるのである。前節で信頼水準を付けた厳密な表現をしてみても３つの問題があることを指摘した。「信頼度は０．９９９」とか「信頼度は０．９５」とかの一つの数値でないと実際上の役に立

たないのである。

さて、信頼度(Reliability)の定義はＪＩＳによると「アイテムが与えられた期間与えられた条件下で機能を発揮する確率」となっている。ＪＩＳでは信頼性(Reliability)も定義されていて「アイテムが与えられた期間与えられた条件下で機能を発揮する性質」となっている。ＮＡＳＡ文書ではReliabilityの定義としてＪＩＳの二つの定義を合わせたようなものになっている。むしろ、ＪＩＳの定義では定量的な信頼度と定性的な信頼性という二つの用語を使い分けるために米国の定義を分けたものと思われる。信頼度の定義はどの文書でも同じように使われており特に問題とするところはない。
ＮＡＳＡの定義(SSP 30000 S.9)では、
Reliability: A characteristic of a system or an element thereof expressed as a probability that it will perform its required function under condition at designated times for specified operating periods.

ところが確率の定義には問題もあり論争の歴史もある。大きく見ると4種類の定義がある。まず数学者が不確定の事象を取り扱うことを始めたのはラプラス(Laplace)が賭けの問題を相談されてからだとされている。ラプラスは取り得る状態（事象）がどれも同じような(equally likely)場合にはそれぞれの事象に全事象の数の逆数を先見的に(a priori)割り当てることを基礎にして求める条件に合うものの割合を確率と定義した。物理学の世界ではラプラスの定義で議論されおり、取り得る状態の数を数えることが重要な課題である。

ラプラスの定義で最も批判をされたのは equally likely という表現であって何を持って同様に確からしいと見るかであった。もっと客観的な確率の定義が望まれたわけである。フォン・ミーゼス(von Mises)は n 回の試行で条件に合う場合が n_A 回あったとした場合に、相対頻度 n_A / n の n を大きくしたときの極限値を確率の定義とした。英語の表現では次のようになる。
Probability is a limit value of the relative frequency, which is assumed to exist.

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

確かにこれで客観的な確率を議論出来るように見えるし、n が十分大きいとき、即ち大数の法則が成り立つような場合は実用上の問題はなく使われてきた。選挙のとき等、無作為抽出により標本(sample)調査で十分全体の結果を早期に予測することに成功している。Ｒ．Ａ．フィシャー等の統計学の大家やネイマ

ン・ピアソンの仮説検定理論等が出され、実務の確率の定義として確立しているかのように見えるのである。

しかし、このような極限値が一定の値に近ずくことが数学的に保証されているわけではないし、この定義の確率を知るすべはなく前述のように信頼水準を設定して推定するしかない。また、例えば大量生産される電球等の信頼度ならまだしも、人工衛星やステーション等のように通常１機しかないものに無限に多くの同一の衛星を想定して相対頻度の極限値を持って定義することにかなりの無理があるように思われる。

純粋数学者は確率の定義として実世界で何を意味しようと関係ないという立場を取って理論の世界だけ完ぺきなものにした。これを確立したコルモゴロフ(Kolmogoloff)は３つの公理を満足する数$P(A)$を事象Ａの確率と定義した。それは、

Ⅰ．　$P(A)$は非負の数である。　$P(A) \geq 0$

II.　確実に生起する事象の確率は１である。　$P(S) = 1$

III.　もし、事象ＡとＢが相互に背反であるならば、

$$P(A+B) = P(A) + P(B)$$

確率の公理を満たすものはラプラスの定義であっても、フオン・ミーゼスの定義を取ろうとも全て確率論の結果を応用出来るのである。この公理の基礎を固めている数学のキーワードは次のようなものである。

・完全加法族 completely additive class
・測度空間 measure space
・有限加法的測度 finitely additive measure
・可測集合 measurable set

フィシャー達と大論争があったとされている一派はベイズ流統計学派と呼ばれている人達である。人により少し趣が異なるが、サベジ(Savage)による定義は「命題の真偽に関する確信の度合い(degree of belief)に対して割り当てる数値」とするものである。同じ命題、例えば「Ａ社の分離ボルトは作動する」に対して、試験データを見た人と何もデータを見ていない人とで割り当てる数値は異なってくる。このことから、主観確率(subjective probability)と呼ばれることがある。この派の人はすべての確率とは主観的なものであるという立場を取っていることもあって主観確率であることを否定しない。このことが「主観的なものは学問的でない」という風潮によって不当に評価されてきたきらいがある。

データを見てどのように確率を割り当てるかの割当方が数学の定理であるトーマス・ベイズの定理を使うという意味で規定されているのでむしろ客観的な方法なのである。情報が無いときにどのように事前分布を考えるか、定性的な情報をどのように組み入れるか等数式に乗りにくい情報の生かし方には研究の余地があるし、数学的でないと批判される余地がある。全く情報が無い状態についてはラプラスの定義のように一様分布を仮定することが合理的であると考えられている。意志決定を合理的に行うことを目的とする分野、例えば経済分野では殆どベイズ流統計学が使われている。

4．確信の度合い

2節で客観的な確率を推定するというアプローチではパラドックスに陥ることを示した。これと同じ例を使って主観確率、即ち、確信の度合いとしての確率ならどのように決定を下せるかを示す。

「A社の分離ボルトは作動する」という命題が真である確信の度合い（確率）をpとする。

（1）A社の製品を試験する前の信頼度、即ち成功確率pに関する事前情報が何も無い状態として0から1までの一様分布を仮定する。即ち、pは0から1までのどの数値になるか分からないのでどの数値になるかは同じ割合であると見込むことが合理的であると考えるのである。事前分布は仮定により既知となる。

（2）次に20個の試験結果が全て成功であったというデータを考慮するとベイズの定理によりpに関する事後分布を求める事が出来る。ベイズの定理は事後分布は事前分布とデータのpに対する尤度の積に比例するというものである。

（3）pに関する知識が分布の形でよりはっきりしたわけであるが、分布を代表させる一つの数値は期待値である。事後分布は2項分布になるのでその期待値は$\frac{r+1}{n+2}$となる。ここでnは試験個数、rは成功数である。

従って、20個の成功を見た後では確信の度合いとしての確率は21／22＝0．955となる。80個試験して75個合格であったB社の製品の信頼度は0．927となる。

２０個の成功データを見ただけでは信頼度は０．９５５に過ぎないのである。もし、信頼度は０．９９９即ち千に一つしか失敗が許されないような部品が要求されており、１０００個も試験する余裕はないような場合には、開発を管理して十分審査する以外に成功の確信を高める、即ち、信頼度を確保する手立てはないのである。これが信頼性管理活動が必要な理由でもある。

因に試験データが何も無いときには成功か不成功かの２つの状態が等しく起こりうると考えることが合理的であり確率はどちらも０．５である。これは一様分布を仮定した事前分布(Fig.3)の期待値でもある。確率を確率分布で表現したとき、その確率分布から求められる期待値が確率に他ならないのである。コイン投げで表が出るか裏が出るかという試行で「表が出る」という命題に対しても確信の度合いは０．５であるが、この場合の事前分布としてはデイラックのデルタ関数を借用して表現出来る。このような分布は強い確信の分布(Fig.4)であり、データがどのようなものであっても事後分布は事前分布と同じになる。

弱いｐ＝０．５の確信

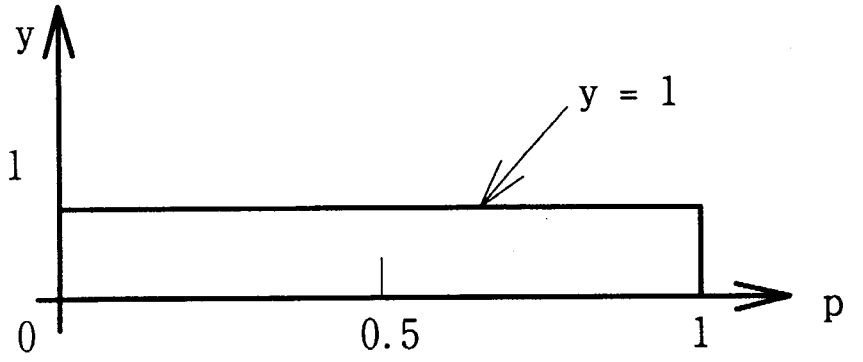

Fig.3 Dencity of Weak Belief

強いｐ＝０．５の確信

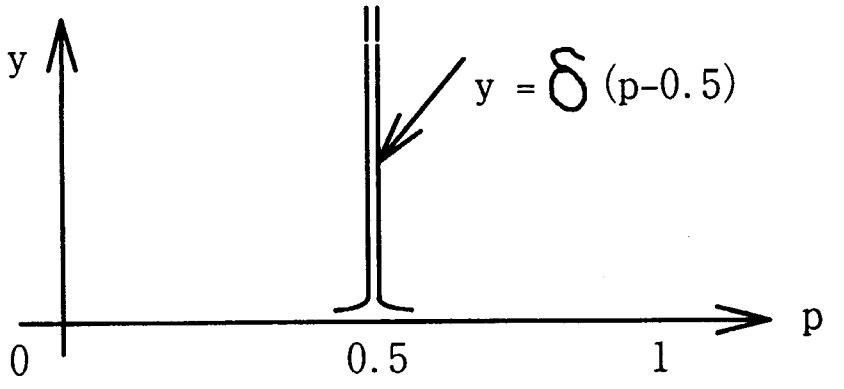

Fig.4 Dencity of Strong Belief

5．あとがき

私が実世界における確率の定義は変えるべきであるとの信念を抱いてから、あるいは取りつかれてから、ざっと２０年間経ってしまった。要約すると、確率の定義は本質的に異なるものが４種類あり、このうち公理に基づくものは数学の世界でありなにも問題はない。実世界の現象を説明するために客観的な確率を想定した相対頻度の極限値という解釈は成功を納めたといって良い程の実績があり世の中に広く使われている解釈であるが、データが十分多い時にだけ使えるもので、よく考えると矛盾を含むものである。データが少なくても、少ないなりに何が言えるかを厳密にするためには、確率の解釈として確信の度合いを採用すべきである。この解釈は確率の古典的解釈である等確率の原理を含むより統一した立場の解釈である。

このような話を議論するならどこかのシンポジウムが適切であるとの上司のご意見は正当であるのを承知の上で無理矢理お願いし、年に一度のＮＡＳＡ／ＮＡＳＤＡの会議の場でＮＡＳＡに対して問題提起をさせて頂いた。日本のシンポジウムでは、現行の方式をいささかでも否定する話は極めて受け入れられ難いことを思い知らされていたし、ベイジアン学派は米国に多い(Ref.14)と聞いていたからである。影響力の大きいＮＡＳＡが問題点に気がついてくれさえすれば定義の見直しは簡単に実現するであろう。

今回の会議のＮＡＳＡ側の出席者は３名で、その団長はＮＡＳＡ本部のミッション・サクセス保証局の局次長 (Deputy Associate Administrator)のDr. Greenfieldであった。彼は自らＮＡＳＡのアクション・アイテムとして設定し、持ち帰って検討することを約束してくれた。このような会議の宿題は次回にまで持ち越されるものだが２カ月足らずで返事がきた。Dr. Greenfield 自身の手紙と共に、彼の部下である Mr. Bob Weinstock が検討結果の第一報を書いてくれた。

Mr. Weinstock の手紙を要約すると、
（１）信頼水準を設定する現行方式には制約があり、信頼水準を設定する方策は何もないものである、従って、この点で原が言っていることに同意する、
（２）ＮＡＳＡでは "uncertainty distribution" なる概念が state of the arts の方法として使われ出しているが、原が提案している内容はこのことではないか。これに関して２、３の資料から抜粋した部分を同封するので検討してみると良い、というものであった。

（２）に関して同封してくれた資料はミサイルを含む固体ロケットの失敗例の統計データを検討したもので非公開資料の一部らしきものであった。この中に、横軸に頻度(frequency)縦軸に(probability density)を取った図があり、考え方に近いものが有るようであるように思われるものの、まだこれまでの確率の定義に引きずられているようにも見受けられる。明確に確率の定義を変えることには抵抗があるのか、その意義に気がついていないかのどちらかである。

確率は命題の真偽に対する確信の度合い(degree of belief)に対して割り当てる数値であるとするサベジ流定義を受け入れた時、**確率は絶対的未来予測が出来ない人間の知識の程度を表現する道具に他ならない**ことがわかる。

６．文献

(1) 軽構造理論とその応用、下、林毅編、日科技連、1966

(2) 航空機構造の疲れ寿命の安全率、上山忠夫、日本航空宇宙学会誌、Vol.9 No.88, 1961.5

(3) ＹＳ１１の悲劇、山村堯、日本評論社、1995.4

(4) Introduction To Probability and Statistics from a Bayesian Viewpoint, D.V.Lindley, 1965

(5) 確率統計入門１及び２、竹内啓・新家健精、培風館、1969

(6) ベイズの方法による疲労寿命とばらつき係数、原宣一、第１４回構造強度に関する講演会、福岡市、1972

(7) 属性試験と信頼度、原宣一、宇宙開発事業団第５回社内技術成果発表会、11/15/1978

(8) L.F. Impellizzeri, ASTM CTP404, p.136, 1966

(9) Probability, Random Variables, and Stochastic Processes, Papoulis, 1965, McGraw-Hill

(10) Mathematical Methods of Statistics, Harold Cramer, 1966, Asian Text Eddition, Overseas Publication, First Published in Sweden, 1945

(11) Introduction to Mathmatical Statistics, Hogg, et al, Third Edition 1970

(12) Quantity Control and Industrial Statistics, Duncan, Forth Edition 1974

(13) Journal of American Statistical Association, Vol. 59,1964, p353 J.W. Pratt, H.Raffia and R. Schlaifer

(14) Quantitative Analysis for Business Decisions, Bierman, et al, Forth Edition 1973