

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%に取ることになれば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。

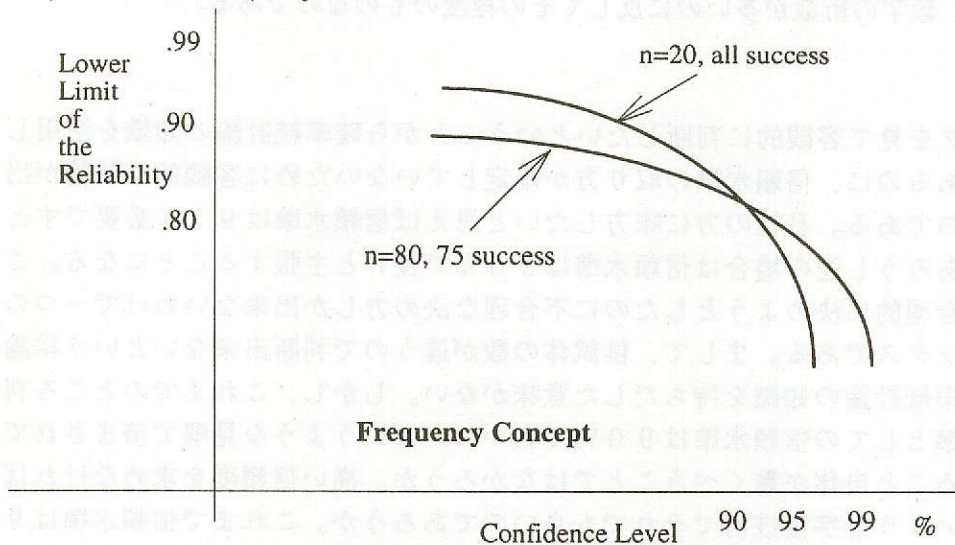


Fig.1 the Reliability (Frequency Concept) depended on the Inspection by Attribute

(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。

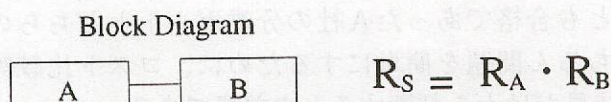


Fig.2 System Reliability

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J I S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。J I S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。N A S A文書ではReliabilityの定義としてJ I S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、J I S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。

N A S A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.

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

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