確率予報の評価について
― 降水確率予報を例にして ―

On the Evaluation for Probability Forecast
― Taking Precipitation Probability Forecast for an Example ―

榛葉　豊*
Yutaka SHINBA

Abstract): We discuss some problems around the evaluation for probability forecast system. The validation method of probabilistic categorical proposition and probabilistic proposition are investigated, with weather forecast as typical example. Behavioral economy point of view is also discussed.

1．はじめに

静岡理工科大学総合情報学部改組検討の中で，人間情報デザイン学科の社会情報系を再編して価値科学コースとする案が出ていたが，筆者は価値科学で取り上げる題材の一つとして確率のキャリブレーションやサヴェッジーワルトの統計的決定理論，ベイズ推定，行動経済学，主観確率の測定問題，情報の価値問題などを検討していた．価値科学コース案は取りやめになったが，前記の主題を含んだ問題の例として，「確率天気予報官ボーナス問題」を考えてみたい．

確率天気予報官の予報がよく当たるのなら，ボーナスは増やさなければならない．それでは「当たる」とはどういうことを言うのであろうか．その程度はどう計量化されるのであろうか．これは，統計的確証問題の良い例である．

天気予報を例にとると，確率的命題の確証問題と天気予報に固有の問題とが絡み合っている．今回は天気予報に固有の問題，すなわち「雨である」というような天気の状態を，どう定義するのか．それから降雨地域の問題，降雨時間の問題などいろいろなカテゴリー分けについての事柄が問題になってくる．気象庁は，これらに対して，受け取る人々の効用にとって必ずしも適切ではないかもしれない定義をして，確率天気予報を実行しているのであるが，このことについては，最後に付録で簡単に触れることにする．次項からは，意思決定理論系の科目の題材として，確率の解釈や行動経済学，統計的決定理論に関した事柄について考えることにする．

2．確率的状況下での最適行動

確率的な命題は，どのような事象が起きようとも（100%か0%という予言以外は）決して反証されない．しかしそれでは何も始まらないので，証拠を発見するたびに，確率の値をベイズの定理によって更新していくという見方がとられる．古くはポパーによって考察された事柄である．

直接的には確証についての問題にはなっていないが，ベイズの定理を実用に供して，最適行動を決める例として，サヴェッジーワルトの統計的決定理論の練習問題を見てみよう．どの教科書にあるのも次のような例題である．

1. 晴れ，曇り，雨といったカテゴリー天気予報があるとして，天気予報と翌日実際にどうなったかの，尤度確率（条件付き確率）の表が与えられる．
2. 統計的な，その地方季節の，天気のカテゴリー別確率，すなわち事前確率が与えられる．
3. 手ぶら，傘を持って出る，さらに長靴を履いて出るなどの雨具の順位についての選択肢と，実際の天気との組み合わせによる，効用の表が与えられる．

この設定の下で，期待効用最大原理によって，予報されたカテゴリーごとの最適行動を決定せよ．

予報を見る人の立場としては，弁当屋という設定が使わ

2015年3月2日受理
*　総合情報学部　人間情報デザイン学科

れたりする．弁当屋がどれだけ弁当を準備するかは，人出によるが，それは天気に依存する，これを「ブライトン（ドーヴァー海峡に臨んだ行楽地）の弁当屋の問題」，個人の外出設定の方は「ヒギンズ氏の雨の問題」という名前がついている．

もう少し複雑にして，気象予報官ボーナス問題に近い設定にすると，次のような問題になる．

2つの天気予報会社がある．それぞれの会社の予報と実際の尤度表が与えられる．事前確率や予報の価格も与えられる．効用表も与えられた上で，どちらの天気予報会社と契約したら良いであろうか決定する．

こちらの問題に近いことを，本年度の卒業研究テーマとして採用してみた．ただし，天気予報会社選択問題は，確率予報の評価の問題と，期待効用最大原理にまつわる問題が複合している．卒論では確率のキャリブレーションのみ調べさせた．

3．2択問題で正解50%の意味

ある予言がどれだけ当たるか，ある解答がどれだけ正解するか．2択問題では，ランダムに答えても，或いはいつでも第一選択肢と答えても正答率の期待値は 50%である．しかしこれを的中度 50%とは言いにくいであろう．点数でいえば，他の記述問題と同次元で，100 点のうちの 50 点とするのは適切ではない．記述問題での 50 点は満点ではないが，なにがしかのことはわかっている．一方2択問題では，何もわかっていなくても 50 点というのでは，点数の加法性の観点からも不具合である．だからといって50%正解した場合に，0 点．100%の正解は 100 点でその間は何らかの方式で補間をするというのは良くないだろう．ランダムではなくて，ちゃんと答えていても偶々50%の正解率になることもあるから，その場合をどう取り込めるかが難点である．

ランダム解答か，考えた末かは本人にしかわからない．1回限りでの試験と，何回も試験が繰り返されるときなど（問題もその都度違うであろうし，受験者の知識も変化して行くであろう）どう考えるか．50%以下，たとえば30%の正解率だったら，まさかマイナス点にするということなどあり得ないだろう．50%以下は全て 0 点なのだろうか．受験者の立場ではなく，出題者側の立場からいうと，情報伝達理論の観点からは，30%の正解率なら，解答の反対が正しい事が多いということになるので，その意味では 0 点ではない．2択ではなく，n 択に拡張した場合の考察も併せてしなくてはならない．

4．カテゴリー天気予報の的中度

さて天気予報の（確率的ではない）晴れ，雨などのカテゴリーを予測する場合について問題を絞ろう．止まっている時計でも，必ず 1 日に二回は合う，・・・等という冗談もあるが，ランダム解答ならぬ，いつも同じ事をいうというのはどうだろうか．統計的にいえば，事象が「晴れ」と「雨」の 2 つだとして，事前確率がたとえば雨が 30%とわかっているとすると，いつも「晴れ」と予言していると，70%は的中する．日本の天気では実際そのようになっている．しかしこれは予報とはいわないであろう．

もう少し改良された方式は，前の日が雨なら，次の日は雨，晴れなら晴れといっていれば，日本の場合的中率は80%にもなるという[1]．

天気の場合，2択問題とは違って各設問の正答選択肢が独立ではない（出題者は独立であるように各問題の正解を配置する）．一日一日の天気は，独立ではなくてその遷移確率はいわゆる多重マルコフ連鎖になっている．言い換えれば自己相関関数がある程度の時間持続するためである．

ところで，まれにしか起こらない事象について的中率（予測と実際の尤度クロス表全体の総和を 100%として，予測と実際が一致しているところの総和）を考えると，いつでも「晴れ」という予言より，ちゃんと予測をした方が的中率は下がる場合がありうる．

天気予報の分野ではスレットスコアといって，珍しいこと（希ではあるが壊滅的な事象）に重点を置いて，警報を出さないときに何も起こらなかったという場合を除外した指標が提案されている．2カテゴリーとして，珍しいことの予報が当たったケース数をそれ以外の 3 つのケースの和で割ったものである．珍しい事象の警報が有効だった割合といえる．スレットスコアなら，何も積極的予報をしない場合には，0 となる．

しかし地震の予報だったらこの数値は我々の実感によくあっているであろうが，天気予報の場合そこまでとはいえない．「雨」が壊滅的ともいえないし，「雨」割合が，地震のように非常に希なわけでもない．最近，イタリアで地震学者が，予測されている地震はたいしたことが無いという見解を出して，その後死者の出る大地震が起こったことがある．地震学者がその廉の殺人罪で起訴されるという事件が起こった．予測の「空振り」「見逃し」にそれぞれどれだけの重きを置くのか，それは社会の構成員全員では決して共有できないことでもあることを，よく認識しなくてはならない．予報を受ける人の効用関数の違いがあるので，誰にでもいい指標というのはあり得ないのである．この辺りの事柄は，クリティカル・シンキングの授業でいい題材になると思う．

実際の天気状況のカテゴリー分けされたケース数の数値セットをエントロピーの観点で見たとき，低エントロピーだと「当たる」のは当たり前である．個人の効用関数の違いが大きな障害となるので，まずは，効用にあまり依存しない指標を追求してみるべきであろう．

5．確率数値予報の確率のキャリブレーション

カテゴリー予報から降雨確率予報になると，降雨確率は（気象庁は10%刻みで発表しているので11カテゴリーで

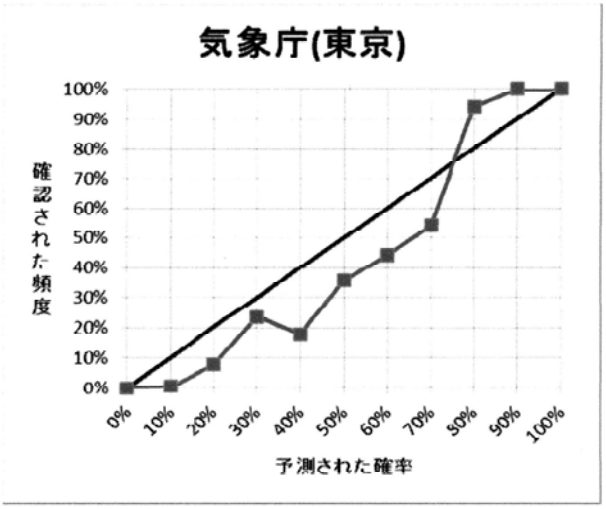

Fig-1 降水確率のキャリブレーション　気象庁，東京

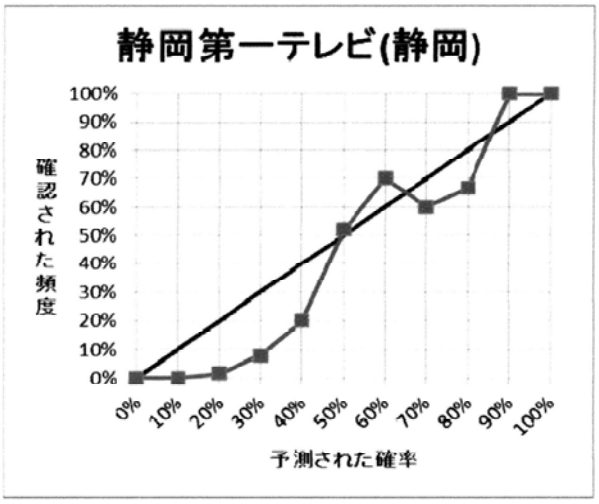

Fig-2　降水確率のキャリブレーション　第一 TV，静岡

あるが）比例尺度といえる．

ここで実際に各々予報された降雨確率ごとの的中率を見てみよう．これは「確率のキャリブレーション」といわれるが，行動経済学や認知心理学でいわれる客観確率と主観確率のキャリブレーションとは少し意味が異なる．しかし形式的には同じようなものである．

ここでいうキャリブレーションとは，降水ありという予報の各確率の値（区間）ごとに，その予報が出された翌日，実際に降雨があったケースは何%かを集計したものある．

上に示す図は，小澤一馬氏が[2] 2014 年 7 月から 12 月にかけて，静岡地方と東京地方の降雨について静岡の民放 2 局（けんみん TV と第一 TV）それに気象庁の降雨確率予報を記録してキャリブレーション図を作成したものの一例である．

小澤氏の作成した図を見ると，どの TV 局でも，どの地方でも，大まかにいって斜め線より下回っている事がわかる．これは，予報でいう降水確率より実際は雨が降らないということで，「空振り」を避けたいというバイアスがかかった心理が見えている．ただし，調査期間が少ないので，0%近くのことについては正確なことはいえない．逆に雨は 80%～90%と予報した場合にはほとんど確実に雨が降るという結果である．このほうは，「雨がほぼ確実に降る」という予報を出して，もし降らなかった場合，雨に対する対策に「余計な」コストをかけてしまった会社などから苦情が多いと考えてのことだろう．

静岡第一 TV の静岡の天気の図で，予報が 60%の時の実際の確率の値が突出していて異様ではあるが，これは 70%という予測値そのものが少なかったことによる揺らぎに原因があると思われる．本当に 70%程度の気象配置と過去の気象が少なかったのか，70 という数値を避ける何かがあるのかはわからない．

一般にいえることではあるが，0%から 100%までの 11 のカテゴリーの降雨確率値が予報される回数は同じではなく，信頼度に差があることに注意しなければならない．

海外の文献を見ると[3]，アメリカ国立気象局のキャリブレーション図は驚くべきもので，上図での斜め右上がりの対角線に，ほとんどぴったり沿ったもので完全キャリブレートに近い．一方，ウェザーチャンネルのものは，上図程度である．

ところがカンザスシティの地方 TV 局は，実際の確率の方が（0%以外では）常に小さく見事なバイアスがかかっている．しかも傾きが水平に近く，どんな予測確率になっても，実際の確率はあまり変化しないという傾向である．特に，上記とは逆であるが 0%と予測したときにも 10%程度は雨が降っているのが特徴的である．

6．いくつかの評価方式の案

キャリブレーション図を見れば，どの予報会社の予報が「当たる」のか判断できるだろうが，定量的に判断する指標としてどのような量をとれば良いだろうか．

関数近似だと思えば，各予報確率値での予報値と実際の値との偏差の 2 乗和というのがすぐに思いつくことだろう．あるいは適合度の検定だと思えば χ 自乗値が候補になるだろう．

6-1　重み付き χ 自乗値：

χ 自乗値だとすると，ケース数が多い場合のウェルドンのサイコロという問題（帰無仮説の棄却傾向）があるが，ケース数を同じにそろえての比較なら問題ないであろう．ただし，一般に対して公表する「当たり」の指標としてはケース数をそろえられないのが一般であるから何らかの対策が必要である．

また，χ 自乗値を用いる場合に，付録 1 に示した，0%100%付近での心理的な重みの違いを考慮して，50%前後の偏差 2 乗値とは，重みを変えた方が良いかもしれない．決定に関する心理だけではなくて，偏差自身も 0%近くで

はマイナスにはなれないから小さい方向には変化が出来ず，100%近くでは 100%は超えられないから大きい方向に変化は出来ない．このことも考慮に入れなくてはならない．χ自乗値は，観測度数と期待度数の偏差の自乗を期待度数で割ったものをカテゴリーに渉って集計したものだから，ある程度補正されているともいえるが，たとえば10%の予測をしていて 9%の偏差が起こって，1%になるのと 19%になるのでは，意味が違うであろう．

本質的に大切であると思われる重み付けは，カテゴリーごとの観測イヴェント数による重みである．観測イヴェントがゼロであるような予報確率値カテゴリーがあったとき，そのカテゴリーについての評価は無意味である．イヴェント数が 1 であったらその予報確率値での観測確率値は 100%かまたは 0%のどちらかになるわけでこれまた無意味に近い．イヴェント数が増えても，少ないうちは信頼度が少ないわけであるから，そのようなカテゴリーでの当たり外れの評価は軽くすべきであると考えられる．χ自乗検定では普通 5~6 イヴェント以下のカテゴリーは検定にかけるのは良くなく，カテゴリーを合併しろ，などといわれる．

一般にχ自乗検定による適合度の検定では期待度数が基準になるが，この問題では，たとえば 70%というカテゴリーの期待度数は全体の 70%ではないのである．70%という予測が予測ケース数全体の何%起こるかは全くわからない．全ケース数のうち何%が，70%といいたくなるような状況かなどという過去の統計からの値などは別の問題である．キャリブレーションで言うと横軸の問題である．カテゴリーの指標としての%と縦地君の問題である的中率の%は混同されやすい．70%などというカテゴリーの%は果たして比例尺度なのか間隔尺度なのか，それとも順序尺度に過ぎないかもしれないのかということも問題である．

これらの問題は，かなり哲学的に難しい問題である．しかし，とにかくイヴェント数の少ないカテゴリーには評価の際の重みを小さく，イヴェント数の多いカテゴリーには重みを大きくしなければならないであろう．

以上述べてきたように，カテゴリーの観測イヴェント数と予想確率値についての心理について何らかの補正がされたχ自乗値は，一つの候補であると思う．

6-2 相互エントロピー，相互情報量などの複雑性指標

エントロピーを変化させるもの，すなわち不確実性を減少させるものが情報で，情報こそ価値なのだとすると，ボーナスはエントロピーの変化に関係するとするのも一案であると思う．

カルバックーライブラー情報量（相対エントロピー）という概念があるが，これは確率分布の間の距離を測るものである．また類似の量として相互情報量という確率分布の相互依存度を測る概念が使われることもある．2 つの確率分布の相互依存度を，それぞれを周辺分布とする同時分布を考えて算出するのである．こちらは結合分布が前提されている [7,8]．他にもエントロピー進化率などという概念も実用化されている．

これらの複雑性指標は，そのままでは降雨確率予報の評価，すなわち完全キャリブレーション状態からのずれの計測には仕えないが，何らかの変更を加えてキャリブレーションの度合い評価に使える量を構成できないか検討中である．

付録 1．行動経済学的側面 トベルスキーの実験

行動経済学のプロスペクト理論で有名なトベルスキーとフォックスは客観確率を横軸にとり，一種の主観確率ともいうべき「決定の重み」を縦軸にとったキャリブレーション [4,5] を示した．それによればキャリブレーション曲線は上記日本の天気予報の場合の図とは逆の，100%から少しでも客観確率が減ると決定の重みは急降下し，0%から少しでも客観確率が増えると決定の重みが急上昇する．20%近辺では，客観確率が変化しても，決定の重みはほとんど 50%のあたりで微増するだけということであった．

この結果は次のように解釈されるだろう．事象が「ある行動をすると大金がもらえる」という場合だと，100%に少しでも欠けると，絶対起こるのではなくて，もしかすると起こらないということもあり得る，という感覚が支配して，その行動を選択する意欲が減る．0%付近でも，0%なら絶対お金にならないのだから，そんな行動はとらないが，0 でないという事はもしかすると奇跡的な事態になるかもしれないからやってみる．

この心理的傾向は，天気予報官の場合にも当然働いている．行楽や用事で外出したい人，洗濯物，会社だったら催し物，農作業などなどいろいろな視聴者の効用，そして実際の効用を超えての不満感．TV 会社の社長や人事関係者の評価，苦情処理係による自分への評価なども頭に浮かぶだろう．小澤氏の調査結果からはリスク回避バイアスが読み取れている．

トベルキーとフォックスの実験に現れている 0%付近と 100%付近の心理的傾向を，天気予報の評価指標に取り入れる必要があるであろう．

付録 2．天気予報固有の問題点

降雨確率予報の場合，「ある地域で雨が降る」という事象の定義の感覚が，予報を受け取る方と発表する側とで乖離しているという問題がある．

気象庁の定義では [6]

1. 雨量について：ある観測点で 1mm 以上の雨が降ったら，その観測点で「降った」になる．それ以上の降水

量の多寡は関係ない.

2. 降雨面積について：該当地域内，総ての観測点で雨が降ってはじめてその地域で「降った」になる.
3. 雨時間について：予報対象時間の中で，少しの時間でも降ったら「降った」になる.

1.〜3.の連言で「降った」は定義される.

そうすると，豪雨なのかどうか，ずっと降っていたのか一瞬降ったのか，あっちの村では降ったが他のこの地方全体では晴れていたのか等区別されない．一般人の，自分を中心とした「雨が降った」という感覚とは大いに異なる.

この多元的なファジー判断を，えいやっと切り分けている（そうせざるを得ない）事から来る違和感はかなりあると思われる．この定義から，ある方向へのバイアスが生じる可能性はある.

謝辞：TV ニュースの天気予報を毎日記録して，キャリブレーション図を作成した図を引用させてくれた，小澤一馬氏に感謝します.

参考文献

1） 立平良三，『気象予報による意思決定 — 不確実性の経済的価値』，東京堂出版，1999 年
2） 小澤一馬，静岡理工科大学卒業論文，2015 年
3） N. シルバー，『シグナル vs. ノイズ』，日経 BP，2013 年
4） Tversky and Fox, 'Weighing risk and uncertainty', *Psychological Review,* ***102,*** (1995)269
5） 広田すみれ他編，『心理学が描くリスクの世界 — 行動意思決定入門』，慶應義塾大学出版会，2002 年
6） 村山貢司，『降水確率 50%は五分五分か』，化学同人，2007 年
7） 榛葉豊，『情報次元による平成 7 年兵庫県南部地震の解析』，静岡理工科大学紀要，第 8 巻，2000 年
8） 榛葉豊，『テクスト間のクラスター解析における Kullback - Leibler Divergence 型距離』，静岡理工科大学紀要，第 10 巻，2002 年