付録 A 統計的検証で利用される代表的な指標[1]

本テキスト内で利用されている統計的検証の指標について説明する。本テキスト内での利用箇所を、各項の見出しの後に括弧書きで記した。(ただし、平均誤差等、利用頻度の高い一般的なものと、本付録内でのみ利用されているものについては省略している。)

A.1 平均誤差、平方根平均二乗誤差、誤差の標準偏差

予報誤差を表す基本的な指標として平均誤差(Mean Error、一般に ME、バイアスまたは系統誤差と記される)と平方根平均二乗誤差(Root Mean Square Error、一般にRMSEと記される)がある。これらは次式で定義される。

$$ME \equiv \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i)$$

$$RMSE \equiv \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - a_i)^2}$$

ここで、N は標本数、x_i は予報値、a_i は実況値(真値)である(実況値は客観解析値や観測値で近似されることが多い)。ME は予報値の実況値からの偏りの平均である。また、RMSE は最小値 0 に近いほど予報が実況に近いことを示す。なお、RMSE は ME からの寄与を分離して、

$$RMSE^2 = ME^2 + \sigma_e{}^2$$

$$\sigma_e{}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - a_i - ME)^2$$

と表すことができる。ここで σ_e は誤差の標準偏差である。

A.2 カテゴリー検証

カテゴリー検証では、まず、対象となる現象の「あり」、「なし」を判定する基準に基づいて予報と実況それぞれにおける現象の有無を判定し、その結果により標本を分類する。そして、それぞれのカテゴリーに分類された頻度数をもとに予報の特性を検証する。

A.2.1 分割表

分割表はカテゴリー検証において、それぞれのカテゴリーに分類された頻度数を示す表である(表 A.2.1)。各スコアは、表 A.2.1 に示される各区分の頻度数を用いて定義される。なお、以下では全事例数を

$N = FO + FX + XO + XX$

実況「現象あり」の頻度数を

$M = FO + XO$

実況「現象なし」の頻度数を

$X = FX + XX$

で表すこととする。

表 A.2.1 分割表。FO、FX、XO、XX はそれぞれの頻度数を表す。

		実況		計
		あり	なし	
予報	あり	FO	FX	FO+FX
	なし	XO	XX	XO+XX
計		M	X	N

A.2.2 適中率 (第 3.5 節、第 3.9 節)

$$\text{適中率} \equiv \frac{FO + XX}{N} \qquad (0 \le \text{適中率} \le 1)$$

適中率 (Percent Correct、 Proportion Correct) は予報が適中した割合である。最大値 1 に近いほど予報の精度が高いことを示す。

A.2.3 空振り率 (第 3.6 節、第 3.8 節)

$$\text{空振り率} \equiv \frac{FX}{FO + FX} \qquad (0 \le \text{空振り率} \le 1)$$

空振り率 (False Alarm Ratio) は、予報「現象あり」の頻度数に対する空振り(予報「現象あり」、実況「現象なし」)の割合である。最小値 0 に近いほど空振りが少ないことを示す。また、このテキストでの使用例では分母を $FO + FX$ としているが、代わりに N として定義する場合もある。

A.2.4 見逃し率

$$\text{見逃し率} \equiv \frac{XO}{M} \qquad (0 \le \text{見逃し率} \le 1)$$

見逃し率 (Miss Rate、Frequency of Misses) は、実況「現象あり」の頻度数 ($M = FO + XO$) に対する見逃し(実況「現象あり」、予報「現象なし」)の割合である。最小値 0 に近いほど見逃しが少ないことを示す。また、分母を M とする代わりに、N として定義する場合もある。

A.2.5 捕捉率 (第 3.6 節、第 3.7 節、第 3.8 節)

$$\text{捕捉率} \equiv \frac{FO}{M} \qquad (0 \le \text{捕捉率} \le 1)$$

捕捉率(Probability of Detection、Prefigurance、適中率と訳されることもある)は、実況「現象あり」であっ

[1] 美濃 寛士

たときに予報が適中した割合である。最大値1に近いほど見逃しが少なく予報の精度が高いことを示す。ただし、この指標から空振りの頻度（FX）を推定することはできない。ROC 曲線（第 A.3.5 項）のプロットに用いられ、この場合一般に Hit Rate と記される。

A.2.6 False Alarm Rate （第 3.7 節）

False Alarm Rate（ Probability of False Detection とも呼ばれる。誤警報率、空振り率と訳されることもある）は実況「現象なし」であったときに予報が外れた割合であり、第 A.2.3 項の空振り率とは分母が異なる。

$$Fr \equiv \frac{FX}{X} \qquad (0 \le Fr \le 1)$$

最小値 0 に近いほど空振りの予報が少なく予報の精度が高いことを示す。ROC 曲線（第A.3.5項）のプロットに用いられる。

A.2.7 バイアススコア

バイアススコア(Bias Score、Frequency Bias) は実況「現象あり」の頻度数に対する予報「現象あり」の頻度数の比であり、次式で定義される。

$$B \equiv \frac{FO+FX}{M} \qquad (B \ge 0)$$

予報と実況で「現象あり」の頻度数が一致する場合に1となる。1より大きいほど予報の「現象あり」の頻度過大、1より小さいほど予報の「現象あり」の頻度過小である。

A.2.8 気候学的出現率

現象の気候学的出現率 P_c（一般に、現象の出現率、現象の出現相対頻度、Sample Climatology、Sample Climate、Climatological Probability、Sample Relative Frequency、Event Frequency、Base Rate などと呼ばれる）は、標本から見積もられる現象の平均的な出現確率であり、次式で定義される。

$$P_c \equiv \frac{M}{N}$$

この量は実況のみから決まり、予報の精度にはよらない。予報の精度を評価する基準を設定する際にしばしば用いられる。

A.2.9 スレットスコア

スレットスコア（Threat Score、TS、Critical Success Index とも呼ばれる）は「現象あり」の場合の予報適中頻度数（FO）に着目して予報精度を評価する指標であり、次式で定義される。

$$TS \equiv \frac{FO}{FO+FX+XO} \qquad (0 \le TS \le 1)$$

出現頻度の低い現象（$XX >> FO, FX, XO$）について、XX の影響を除いて検証するのに有効である。最大値1に近いほど予報の精度が高いことを示す。なお、スレットスコアは現象の気候学的出現率の影響を受けやすく、例えば異なる環境下で行われた予報の比較には適さない。この問題を緩和するため次項のエクイタブルスレットスコアなどが考案されている。

A.2.10 エクイタブルスレットスコア （第 2.2 節）

エクイタブルスレットスコア（Equitable Threat Score、ETS、Gilbert Skill Score とも呼ばれる）は気候学的な確率で「現象あり」が適中した頻度を除いて求めたスレットスコアであり、次式で定義される(Schaefer 1990)。

$$ETS \equiv \frac{FO-S_f}{FO+FX+XO-S_f} \qquad (-\frac{1}{3} \le ETS \le 1)$$

ただし、

$$S_f = P_c(FO+FX), \quad P_c = \frac{M}{N}$$

である。ここで、P_c は現象の気候学的出現率（第 A.2.8 項）、S_f は「現象あり」をランダムに $FO+FX$ 回予報した場合（ランダム予報）の「現象あり」の適中頻度数である。最大値1に近いほど予報の精度が高いことを示す。ランダム予報で 0 となる。また、$FO = XX = 0$、$FX = XO = N/2$ の場合に最小値 –1/3 をとる。

A.2.11 Heidke のスキルスコア（スキルスコア）（第 3.8 節）

気候学的確率などによる予測の難易を取り除いて、予測の技術力を評価するために考案された指数は一般にスキルスコアと呼ばれている（菊池原 1988）。その代表的なものがここに述べる Heidke のスキルスコア（Heidke's Skill Score）で、本テキストでは、「Heidke の」は省略して用いている。スキルスコアは、気候学的な確率で「現象あり」および「現象なし」が適中した頻度を除いて求める適中率であり、次式で定義される。

$$Skill \equiv \frac{FO+XX-S}{N-S} \qquad (-1 \le Skill \le 1)$$

ただし、

$$S = Pm_c(FO+FX) + Px_c(XO+XX),$$

$$Pm_c = \frac{M}{N}, \quad Px_c = \frac{X}{N}$$

である。ここで、Pm_c は「現象あり」、Px_c は「現象なし」の気候学的出現率（第 A.2.8 項）、S は現象の「あり」を $FO+FX$ 回、（すなわち、「なし」を $XO+XX$ 回）ランダムに予報した場合（ランダム予報）の適中頻度数である。

最大値1に近いほど予報の精度が高いことを示す。ランダム予報で 0 となる。また、$FO = XX = 0$ 、$FX = XO = N/2$ の場合に最小値 –1 をとる。

A.2.12 n×n 分割表とスコア（第 3.8 節、第 3.9 節）

表 A.2.1 の分割表では、事象を 3 個以上のカテゴリーに分類する場合に対応できない。このとき、カテゴリーの個数が n であれば、表 A.2.1 を n×n の分割表に拡張したものを使用する（表 A.2.2）。

表 A.2.2　n×n 分割表。F_iO_j はそれぞれの頻度数を表す。

		実況			
		O_1	O_2	…	O_n
予報	F_1	F_1O_1	F_1O_2	…	F_1O_n
	F_2	F_2O_1	F_2O_2	…	F_2O_n
	…	…	…	…	…
	F_n	F_nO_1	F_nO_2	…	F_nO_n

例えば、天気予報を「晴れ」、「曇り」、「降水あり」の 3 つのカテゴリーに分類して評価する場合は、3×3 分割表(n=3)を用いる。全事例数を

$$Nn = \sum_{j=1}^{n}\sum_{i=1}^{n} F_i O_j$$

とすると、第 3.9 節で利用されている、適中率（第 A.2.2 項）は、n×n 分割表の場合以下のようになる。

$$\text{適中率}(n\times n) \equiv \frac{\sum_{i=1}^{n} F_i O_i}{Nn}$$

また、第 3.8 節で利用されているスキルスコア（第 A.2.11 項）は以下のようになる。

$$\text{スキルスコア}(n\times n) \equiv \frac{\sum_{i=1}^{n} F_i O_i - Sn}{Nn - Sn}$$

ただし、

$$Sn \equiv \sum_{k=1}^{n} \frac{\sum_{j=1}^{n} F_j O_k \cdot \sum_{i=1}^{n} F_k O_i}{Nn}$$

A.3 確率予報に関する指標

A.3.1 ブライアスコア（第 3.2 節）

ブライアスコア(Brier Score、BS) は確率予報の統計検証の基本的指標である。ある現象の出現確率を対象とする予報について、次式で定義される。

$$BS \equiv \frac{1}{N}\sum_{i=1}^{N}(p_i - a_i)^2 \qquad (0 \le BS \le 1)$$

ここで、p_i は確率予報値（0 から 1）、a_i は実況値（現象ありで 1、なしで 0）、N は標本数である。BS は完全に適中する決定論的な（p_i =0 または 1 の）予報（一般に完全予報と呼ばれる）で最小値 0 をとり、0 に近いほど予報の精度が高いことを示す。また、現象の気候学的出現率 $P_c = M/N$ （M は実況「現象あり」の頻度数、第 A.2.8 項参照）を常に確率予報値とする予報（一般に気候値予報と呼ばれる）のブライアスコア BS_c は

$$BS_c = P_c(1 - P_c)$$

となる。ブライアスコアは現象の気候学的出現率の影響を受けるため、異なる標本や出現率の異なる現象に対する予報の精度を比較するのには適さない。例えば上記 BS_c は P_c 依存性を持ち、同じ予報手法（ここでは気候値予報）に対しても P_c の値に応じて異なる値をとる（Stanski et al. (1989) など）。次項のブライアスキルスコアはこの問題を緩和するため気候値予報を基準にとり、そこからのブライアスコアの変化によって予報精度を評価する。

A.3.2 ブライアスキルスコア

ブライアスキルスコア (Brier Skill Score、BSS) はブライアスコアに基づいた指標であり、気候値予報を基準とした予報の改善の度合いを示す。ブライアスコア BS 、気候値予報によるブライアスコア BS_c を用いて

$$BSS \equiv \frac{BS_c - BS}{BS_c} \qquad (BSS \le 1)$$

で定義される。完全予報で 1、気候値予報で 0、気候値予報より誤差が大きいと負となる。

A.3.3 Murphy の分解

Murphy (1973) は、ブライアスコアと予報の特性との関連を理解しやすくするため、ブライアスコアを信頼度 (reliability) 、分離度 (resolution) 、不確実性 (uncertainty) の 3 つの項に分解した。これを Murphy の分解と呼ぶ（高野 (2002) などに詳しい）。

確率予報において、確率予報値を L 個の区間に分け、標本を確率予報値の属する区間に応じて分類することを考える。確率予報値が l 番目の区間に属する標本数を N_l ($N = \sum_{l=1}^{L} N_l$)、このうち実況が「現象あり」であった頻度数を M_l ($M = \sum_{l=1}^{L} M_l$)とすると、Murphy の分解によりブライアスコアは以下のように表される（確率予報値の l 番目の区間の区間代表値を p_l とする）。

BS = 信頼度－分離度+不確実性

$$\text{信頼度} = \sum_{l=1}^{L}\left(p_l - \frac{M_l}{N_l}\right)^2 \frac{N_l}{N}$$

$$\text{分離度} = \sum_{l=1}^{L}\left(\frac{M}{N} - \frac{M_l}{N_l}\right)^2 \frac{N_l}{N}$$

$$\text{不確実性} = \frac{M}{N}\left(1 - \frac{M}{N}\right)$$

信頼度は確率予報値(p_l)と実況での現象出現相対頻度(M_l / N_l)が一致すれば最小値 0 となる。分離度は確率予報値に対応する実況での現象の出現相対頻度(M_l / N_l)が気候学的出現率($P_c = M / N$)から離れているほど大きい値をとる。不確実性は現象の気候学出現率が P_c = 0.5 の場合に最大値 0.25 をとる。この項は実況のみによって決まり、予報の手法にはよらない。また、不確実性= BS_c が成り立つ。これらを用いてブライアスキルスコアを次のように書くことができる。

$$BSS = \frac{\text{分離度} - \text{信頼度}}{\text{不確実性}}$$

A.3.4 確率値別出現率図 （第 3.2 節、第 3.6 節、第 3.8 節）

確率値別出現率図（Reliability Diagram、Attributes Diagram とも呼ばれる）は、予報された現象出現確率 P_{fcst} を横軸に、実況で現象が出現した相対頻度 P_{obs} を縦軸にとり、確率予報の特性を示した図である（図 A.3.1 参照、Wilks (2006) などに詳しい）。一般に、確率予報の特性は確率値別出現率図上で曲線として表される。この曲線を信頼度曲線(Reliability curve)と呼ぶ。

信頼度曲線の特性は、Murphy の分解（第 A.3.3 項）の信頼度、分離度と関連付けることができる。横軸 P_{fcst} の各値について、信頼度（あるいは分離度）への寄与は、信頼度曲線上の点から対角線 $P_{obs} = P_{fcst}$ 上の点（あるいは直線 $P_{obs} = P_c$ 上の点）までの距離の二乗として表現される。P_{fcst} の各値でのこれらの寄与を、標本数に比例する重みで平均して信頼度（あるいは分離度）が得られる。例えば、no-skill line（直線 $P_{obs} = (P_{fcst} + P_c)/2$）上の点では、信頼度と分離度への寄与は等しい大きさを持ち、ブライアスキルスコアへの寄与が 0 となる。また no-skill line と直線 $P_{fcst} = P_c$ との間の領域（分離度への寄与>信頼度への寄与、図 A.3.1 灰色の領域）内に位置する点は、ブライアスキルスコアに正の寄与を持つ。

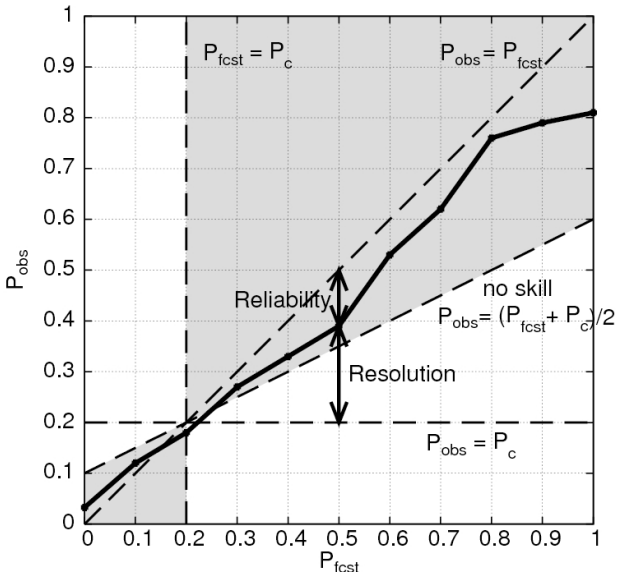

図 A.3.1 確率値別出現率図の模式図。横軸は予報現象出現確率、縦軸は実況現象出現相対頻度、実線が信頼度曲線である。対角線、直線 $P_{obs} = P_c$ からの距離の二乗がそれぞれ信頼度(Reliability)、分離度(Resolution)への寄与に対応している。灰色の領域内の点はブライアスキルスコアに正の寄与を持つ。

特別な場合として、気候値予報（第 A.3.1 項参照）では 1 点 $(P_{fcst}, P_{obs}) = (P_c, P_c)$ が信頼度曲線に対応する。また、次の 2 つの特性を示す確率予報は精度が高い。

- 信頼度曲線が対角線に（信頼度が最小値 0 に）近い。
- 信頼度曲線上の大きい標本数に対応する点が点 $(P_{fcst}, P_{obs}) = (P_c, P_c)$（気候値予報）から離れた位置（確率値別出現率図の左下または右上寄り）に分布する（分離度が大きい）。

A.3.5 ROC 面積スキルスコア （第 3.7 節）

確率予報では、現象の予報出現確率にある閾値を設定し、これを予報の「現象あり」「現象なし」を判定する基準とすることが可能である。さまざまな閾値それぞれについて作成した分割表をもとに、閾値が変化したときの Fr － Hr 平面（ここで Fr は False Alarm Rate（第 A.2.6 項）、Hr は Hit Rate（第 A.2.5 項））上の軌跡をプロットしたものが ROC 曲線（相対作用特性曲線、Relative Operating Characteristic curve、ROC curve）である（図 A.3.2 参照、高野(2002)などに詳しい）。平面内の左上方の領域では $Hr > Fr$ であり、平面の左上側に膨らんだ ROC 曲線特性を持つ確率予報ほど精度が高いと言える。従って、ROC 曲線から下の領域（図 A.3.2 灰色の領域）の面積（ROC 面積、ROC area、ROCA）は情報価値の高い確率予報ほど大きく

なる。ROC 面積スキルスコア（ROC Area Skill Score、ROCASS）は情報価値のない予報（$Hr = Fr$）を基準として ROC 面積を評価するものであり、次式で定義される。

$$ROCASS \equiv 2(ROCA - 0.5) \quad (-1 \le ROCASS \le 1)$$

完全予報で最大値 1 をとる。また、情報価値のない予報（例えば、区間[0,1]から一様ランダムに抽出した値を確率予報値とする予報など）で 0 となる。

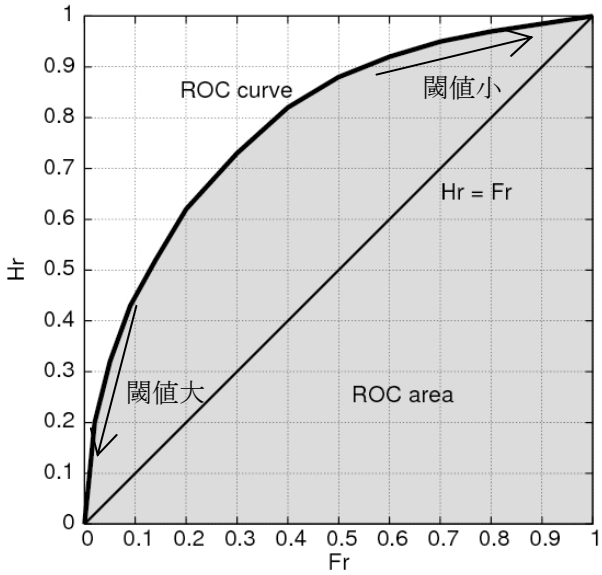

図 A.3.2 ROC 曲線の模式図。横軸は Fr、縦軸は Hr である。灰色の領域の面積が ROC 面積である。

参考文献

菊地原英和，1988：2 カテゴリー予測の検証と評価．気象予測の検証と評価，気象研究ノート，**161**，33-58.

高野清治，2002：アンサンブル予報の利用技術．アンサンブル予報，気象研究ノート，**201**，73-103.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Met.*, **12**, 595-600.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570-575.

Stanski, H. R., L. J. Wilson and W. R. Burrows, 1989: Survey of common verification methods in meteorology. *Research Report No. (MSRB) 89-5*, Forecast Research Division, Atmospheric Environment Service, Environment Canada.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences; Second Edition, International Geophysical Series vol. 91.* Academic Press, 627pp.