Indian Journal of Advanced Mathematics (IJAM)
ISSN: 2582-8932 (Online), Volume-6 Issue-2, October 2026

Possibilistic Hypothesis Testing: A Specificity–Entailment Framework for Confidence Bounds on Natural Possibility Distributions

Jayesh V. Karanjgaonkar

Abstract: ***Specificity and entailment are the two governing principles of possibility theory, yet no existing formulation – including our own frequency-based specificity measure. $Sp(X)$ – translates them into a procedure for testing a candidate value against observed data with a controlled error rate, or into a confidence interval in the statistical sense. This gap motivates the present study: a decision-maker using a possibility distribution has no principled way to state how confident that distribution's dominant category actually is. We close this gap by formulating a possibilistic analogue of classical null-hypothesis testing, built entirely from the natural possibility and natural probability measures of an explanatory database. We define Type I and Type II error functions $\alpha(\tau)$ and $\beta(\tau)$ as, respectively, the entailment risk and the specificity risk of an α-cut acceptance region, and prove that inverting the resulting test yields a confidence interval $CI_{1-\alpha}$ that is exactly an α-cut of the possibility distribution – a result that gives Yager's classical specificity integral a genuine statistical interpretation it previously lacked. We further introduce the Specificity–Entailment Operating Characteristic (SEOC), a possibilistic analogue of the Receiver Operating Characteristic (ROC) curve, with proofs of its monotonicity and its relationship to specificity, and a two-sample permutation test that replaces a previously heuristic criterion for merging database-held observations with a thinly sampled observation set. The framework is validated on two independently sourced real datasets – India Meteorological Department district rainfall-category counts for Chhattisgarh, and an institutional grade distribution for a university literature course – through a 2000-resample coverage simulation, comparison against classical Wilson-score intervals, and an augmentation-test experiment. The coverage simulation shows that $CI_{1-\alpha}$ attains conservative, over-nominal coverage under discrete possibility distributions, and reveals that bootstrap and delta-method critical values diverge once a single category's share exceeds roughly one half – a finding that revises our own initial expectation and indicates the bootstrap procedure should be preferred whenever modal dominance is strong. Together, these results give researchers and practitioners working with possibility distributions a tested, statistically grounded procedure for deciding how much a distribution's dominant category can be trusted, and how much data is required before that decision is reliable.***

Keywords: ***Possibility Distribution, Hypothesis Testing, Confidence Interval, Specificity, Entailment Principle***

Manuscript received on 15 August 2026 | First Revised Manuscript received on 21 August 2026 | Second Revised Manuscript received on 20 September 2026 | Manuscript Accepted on 15 October 2026 | Manuscript published on 30 October 2026.

*Correspondence Author(s)

Dr. Jayesh V. Karanjgaonkar*, Assistant Professor, Department of Mathematics, Swami Atmanand Govt. Eng. Med. Model College, Atari, Raipur (Chhattisgarh), India. Email: drjayeshk1303@gmail.com, ORCID ID: 0009-0000-3894-366X

© The Authors. Published by Lattice Science Publication (LSP). This is an open-access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Nomenclature:
ED: Explanatory Database
CI: Confidence Interval
PV: Proximity Value
EPV: Elemental Proximity Value
MPV: Modal Proximity Value
ROC: Receiver Operating Characteristic
SEOC: Specificity–Entailment Operating Characteristic
AUC: Area Under the Curve
IMD: India Meteorological Department.

I. INTRODUCTION

Specificity of a possibility distribution has been treated, since Yager's original formulation as discussed and extended by Xu and Yager [1], as a scalar summary of how sharply a distribution isolates a single plausible state. In earlier work [4] we proposed a frequency-based specificity measure $\mathrm{Sp}(X)$ derived from an explanatory database (X, F), together with natural possibility π_M, natural probability P, and a family of proximity measures (PV, EPV, MPV) that relate the two, extending our earlier possibilistic treatment of uncertainty and vagueness in semantic communication [6]. That work, and the mode-sufficiency discussion that motivated it [5], left an explicit gap: the notion of a "confidence limit" on the observed mode was described only heuristically, as a similarity index pushed toward unity, with no controlled error rate and no statistical guarantee of coverage.

This gap matters because the two governing principles of possibilistic reasoning – *specificity* and *entailment* – are already, in substance, a Type I/Type II error trade-off in disguise. The entailment principle favours large candidate sets because they are less likely to exclude the truth; the maximum specificity principle favours small candidate sets because they are more informative. Excluding the truth is a Type I-style error; retaining implausible alternatives is a Type II-style error. The present paper makes this correspondence precise, and uses it to build a complete testing and confidence-interval procedure native to the natural possibility framework, rather than one borrowed wholesale from classical statistics or from Zadeh/Dubois–Prade-style membership composition.

A. Contributions

i. A possibilistic hypothesis test for a candidate modal category, with a null sampling distribution derived both by bootstrap resampling and by a delta-method asymptotic approximation (Section 4).

ii. Formal definitions of $\alpha(\tau)$ and $\beta(\tau)$ as the

© Copyright: All rights reserved.

entailment risk and specificity risk of an α-cut, and a proof that they are monotone in the cut level τ (Section 4.4).

iii. A confidence interval $CI_{1-\alpha}$ obtained by inverting the test, shown to coincide with an α-cut of the natural possibility distribution, and a reinterpretation of Yager's specificity integral in terms of this interval (Section 4.5).

iv. The Specificity–Entailment Operating Characteristic (SEOC), a possibilistic ROC analogue, with a proof of its monotonicity and its relation to $\mathrm{Sp}(X)$ (Section 5).

v. A two-sample permutation test replacing the heuristic database-augmentation criterion of [5] with a controlled decision rule (Section 4.6).

vi. An empirical validation on two independent datasets via coverage simulation, a Wilson-interval baseline comparison, and an augmentation-test experiment (Section 7).

II. PRELIMINARIES

We recall the measures defined in [4]. Let $X = \{x_1, \dots, x_n\}$ be the set of observations of a decision variable and $F = \{f_1, \dots, f_n\}$ the respective frequencies, with $N = \sum_{i=1}^{n} f_i$. The pair (X, F) is the explanatory database (ED). We assume throughout that the ED is uni-modular, with a unique mode pair. (x_M, f_M) such that f_M is much larger than every other f_j.

$$P(x_i) = \frac{f_i}{N} \quad \text{(natural probability)} \qquad (1)$$

$$\pi_M(x_i) = \frac{f_i}{f_M} \quad \text{(natural possibility)} \qquad (2)$$

$$PV = \frac{\bar{N}}{f_M}, \quad \bar{N} = N - f_M \quad \text{(proximity value)} \qquad (3)$$

$$EPV(x_i) = \pi_M(x_i) - P(x_i) = \left(\frac{\bar{N}}{N f_M}\right) f_i$$

(elemental proximity value) (4)

$$MPV = \pi_M(x_M) - P(x_M) = \frac{\bar{N}}{N}$$

(modal proximity value) (5)

$$\mathrm{Sp}(X) = \frac{f_M}{N} - \frac{1}{n-1} \sum_{f_j \neq f_M} \frac{f_j^2}{N f_M} \quad \text{(natural specificity)} \qquad (6)$$

Section 4.5 of [5] additionally required $n \geq 3$ for a possibilistic decision process to be non-trivial, and established qualitatively that $\mathrm{Sp}(X)$ increases as f_M grows and decreases as any non-modal f_j grows. We formalise this in Section 3.

III. MODE SUFFICIENCY REVISITED

Definition 3.1 (ε -sufficiency) *An explanatory database (X, F) is ε-sufficient relative to a* target *specificity Sp^* if $Sp(X) \geq Sp^* - \varepsilon$ for a chosen tolerance $\varepsilon > 0$.*

Proposition 3.2 (Monotonicity of specificity in the mode frequency) *Fix n and the* non-*modal frequencies f_j, $j \neq M$, and let $N = f_M + \sum_{j \neq M} f_j$ vary as f_M increases by $\Delta > 0$ (so N increases by the same Δ). Then for f_M sufficiently large relative to $\bar{N} = \sum_{j \neq M} f_j$, $\frac{\partial \mathrm{Sp}(X)}{\partial f_M} > 0$.*

Proof. Write $\mathrm{Sp}(X) = \frac{f_M}{f_M + \bar{N}} - \frac{1}{(n-1)(f_M + \bar{N}) f_M} \sum_{j \neq M} f_j^2$.

Let $S_2 = \sum_{j \neq M} f_j^2$, a constant with respect to f_M. Then

$$\mathrm{Sp}(X) = \frac{f_M}{f_M + \bar{N}} - \frac{S_2}{(n-1) f_M (f_M + \bar{N})}.$$

Differentiating the first term gives $\bar{N}/(f_M + \bar{N})^2 > 0$.

Differentiating the second term (which is subtracted) gives, after simplification,

$$-\frac{\partial}{\partial f_M}\left[\frac{S_2}{(n-1) f_M (f_M + \bar{N})}\right] = \frac{S_2\,(2 f_M + \bar{N})}{(n-1) f_M^2 (f_M + \bar{N})^2} > 0.$$

Both terms are positive, so $\partial \mathrm{Sp}(X)/\partial f_M > 0$ everywhere on the domain $f_M > 0$, $\bar{N} \geq 0$; the uni-modularity assumption ($f_M \gg f_j$ for all $j \neq M$) is only needed to keep $\mathrm{Sp}(X) \in [0,1]$ meaningful, not for the sign of the derivative.

Proposition 3.3 (Monotonicity of specificity in a non-modal frequency)

Fix n, f_M, and all f_k for $k \neq M, r$, and let a single non-modal frequency f_r increase *by $\Delta > 0$ (so N increases by Δ).*

$$\textit{Then} \quad \frac{\partial \mathrm{Sp}(X)}{\partial f_r} < 0 \qquad \text{whenever } f_r < f_M.$$

Proof. Let $N = f_M + f_r + C$ where $C = \sum_{k \neq M, r} f_k$ is constant. Then

$$\mathrm{Sp}(X) = \frac{f_M}{f_M + f_r + C} - \frac{1}{(n-1) f_M}\left[\frac{f_r^2}{f_M + f_r + C} + \frac{S_{2'}}{f_M + f_r + C}\right], \text{ with}$$

$S_{2'} = \sum_{k \neq M, r} f_k^2$ constant.

Differentiating the first term with respect to f_r gives $-f_M/(f_M + f_r + C)^2 < 0$.

For the bracketed term, write. $g(f_r) = (f_r^2 + S_{2'})/(f_M + f_r + C)$; then

$g'(f_r) = \frac{2 f_r (f_M + f_r + C) - (f_r^2 + S_{2'})}{(f_M + f_r + C)^2}$.

The numerator is positive whenever f_r is not too small relative to $S_{2'}$, which holds under the stated uni-modularity regime since $f_r < f_M$ and $S_{2'}$ is itself a sum of frequencies smaller than f_M; hence $g'(f_r) > 0$ in the regime of interest, so the second (subtracted) term of $\mathrm{Sp}(X)$ is decreasing, and both contributions to $\partial \mathrm{Sp}(X)/\partial f_r$ are negative.

Theorem 3.5 (Minimum sufficient sample size) *Let Sp^* be a target specificity level and suppose the non-modal mass is* approximately *uniformly distributed, i.e. $f_j \approx \bar{N}/(n-1)$ for all $j \neq M$. Then (X, F) is Sp^*-sufficient once $N \geq \frac{f_M}{\mathrm{Sp}^* + \frac{\bar{N}^2}{(n-1)^2 f_M N}}$.*

In particular, for fixed proportion $\rho = f_M/N \in (0,1)$, the condition simplifies to a lower bound on N:

$$N \geq \frac{(1-\rho)^2}{(n-1)\,\rho\,(\rho - \mathrm{Sp}^*)}, \qquad \rho > \mathrm{Sp}^*.$$

Proof. Under the uniform non-modal approximation,

© Copyright: All rights reserved.

$\sum_{j\neq M} f_j^2 \approx (n-1)(\bar{N}/(n-1))^2 = \bar{N}^2/(n-1)$.

Substituting into the definition of $\mathrm{Sp}(X)$ and setting $\mathrm{Sp}(X) \geq \mathrm{Sp}^*$ gives

$$\frac{f_M}{N} - \frac{\bar{N}^2}{(n-1)^2 N f_M} \geq \mathrm{Sp}^*,$$

which rearranges to the first stated bound. Writing $f_M = \rho N$ and $\bar{N} = (1-\rho)N$ and solving the resulting inequality for N gives the second form.

This theorem converts the mode-sufficiency discussion of [5] from a descriptive criterion into an explicit, checkable sample-size requirement, directly analogous to Cochran's formula in classical survey sampling.

IV. POSSIBILISTIC HYPOTHESIS TESTING

A. Hypotheses and Test Statistic

Let $x_0 \in X$ be a candidate value for the population-level dominant category. We distinguish the hypothesised category from the sample mode. x_M, since the two may disagree by finite-sample noise alone.

$H_0: x_0$ is the true dominant category vs.

$H_1: x_0$ is not.

Define the test statistic as the natural possibility of the candidate under the observed ED,

$$T(x_0) = \pi_M^o(x_0) = \frac{f_{x_0}}{f_M} \in [0,1].$$

$T(x_0) = 1$ when $x_0 = x_M$ (no evidence against H_0), and $T(x_0) \to 0$ as the observed frequency of x_0 falls relative to the sample mode.

B. Null Sampling Distribution

Because T is a ratio of correlated multinomial counts, its exact distribution under H_0 has no closed form; we give two complementary constructions.

C. Bootstrap Procedure

Given the observed ED (X, F) with $N = \sum f_i$, draw B resamples of size N from the multinomial distribution with cell probabilities $\hat{p}_i = f_i/N$, constrained so that x_0 retains the maximal cell probability (i.e. resample under the null). For each resample $b = 1, \ldots, B$, compute $T^{(b)}(x_0)$. The empirical α-quantile of $\{T^{(b)}(x_0)\}$ gives the critical value c_α.

Asymptotic (delta-method) approximation.

For large N, $(f_{x_0}, f_M)/N$ is approximately bivariate normal by the multinomial central limit theorem, with

$$\mathrm{Var}(\hat{p}_{x_0}) = \frac{p_{x_0}(1-p_{x_0})}{N}, \quad \mathrm{Var}(\hat{p}_M) = \frac{p_M(1-p_M)}{N}, \quad \mathrm{Cov}(\hat{p}_{x_0}, \hat{p}_M) = -\frac{p_{x_0} p_M}{N}.$$

Applying the delta method to $T = \hat{p}_{x_0}/\hat{p}_M$ gives, to first order,

$$\mathrm{Var}(T) \approx \frac{1}{N}\left[\frac{p_{x_0}(1-p_{x_0})}{p_M^2} + \frac{p_{x_0}^2 p_M(1-p_M)}{p_M^4} + \frac{2p_{x_0}^2 p_M}{p_M^3}\right] = \frac{p_{x_0}}{N p_M^2}\left[(1-p_{x_0}) + \frac{p_{x_0}(1-p_M)}{p_M} + 2p_{x_0}\right]. \quad (7)$$

Under H_0, $p_{x_0} = p_M$, and (7) reduces to $\mathrm{Var}(T) \approx (1-p_M)(2+p_M)/(N p_M)$, giving the approximate critical value

$$c_\alpha \approx 1 - z_\alpha \sqrt{\frac{(1-\hat{p}_M)(2+\hat{p}_M)}{N\hat{p}_M}},$$

where z_α is the upper-α standard normal quantile and $\hat{p}_M = f_M/N$.

Remark 4.1 The bootstrap and asymptotic critical values should be reported together. Section 7 shows that agreement between them depends less on N alone than on the modal share $\rho = f_M/N$: when a single category dominates strongly ($\rho \gtrsim 0.5$), the multinomial ratio T becomes tightly concentrated near 1 under H_0 and the bootstrap critical value saturates at $c_\alpha^{\mathrm{boot}} = 1$, while the delta-method value remains strictly below 1 – the normal approximation underlying (7) is a poor fit to this regime because T is bounded above by 1 and its null distribution is heavily left-skewed rather than symmetric. The bootstrap value should therefore be treated as the primary reference whenever ρ is large; the asymptotic formula is retained here for its closed form and for the moderate-ρ regime, where the two agree far more closely, as shown in Section 7.

D. Decision Rule

Reject H_0 at level α if $T(x_0) < c_\alpha$. This defines a genuine Type I error rate. $\alpha = P(\mathrm{reject} H_0 | H_0 \mathrm{true})$, estimated by either construction above.

E. α-Cuts, Entailment Risk, and Specificity Risk

Definition 4.2 (α - cut Acceptance Region) *For a threshold* $\tau \in (0,1]$, *define*

$$A_\tau = \{x \in X : \pi_M^o(x) \geq \tau\}.$$

As $\tau \to 0$, $A_\tau \to X$ (maximal entailment coverage, minimal specificity); as $\tau \to 1$, $A_\tau \to \{x_M\}$ (maximal specificity, minimal entailment coverage).

Definition 4.3 (Entailment Risk and Specificity Risk)

$$\alpha(\tau) = \sum_{x_i:\ \pi_M^o(x_i) < \tau} P(x_i) = \sum_{x_i:\ f_i/f_M < \tau} \frac{f_i}{N}$$

$$\beta(\tau) = \sum_{x_i \neq x_M:\ \pi_M^o(x_i) \geq \tau} P(x_i).$$

$\alpha(\tau)$ is the natural-probability mass excluded by the cut – the risk of wrongly ruling out the truth (an entailment failure). $\beta(\tau)$ is the natural-probability mass of non-modal categories still retained by the cut – the risk of failing to discriminate against a false candidate (a specificity failure).

Proposition 4.4 (Monotonicity of $\alpha(\tau)$ and $\beta(\tau)$) $\alpha(\tau)$ *is non-decreasing and* $\beta(\tau)$ *is non-increasing on* $\tau \in (0,1]$.

Proof. For $\tau_1 < \tau_2$, $\{x_i : \pi_M^o(x_i) < \tau_1\} \subseteq \{x_i : \pi_M^o(x_i) < \tau_2\}$, so the

© Copyright: All rights reserved.

sum defining $\alpha(\tau)$ is taken over a weakly larger index set as τ increases, and since every summand $P(x_i) \geq 0$, $\alpha(\tau_1) \leq \alpha(\tau_2)$. Symmetrically, $\{x_i \neq x_M : \pi_M^o(x_i) \geq \tau_2\} \subseteq \{x_i \neq x_M : \pi_M^o(x_i) \geq \tau_1\}$, so $\beta(\tau_2) \leq \beta(\tau_1)$.

This proposition is the formal statement of the specificity–entailment trade-off: A_τ cannot simultaneously minimise both risks, mirroring the impossibility of simultaneously minimising α and β in Neyman–Pearson testing at fixed N.

F. Confidence Intervals as α-cuts

Theorem 4.5 (Confidence Interval by Test Inversion) *Fix a nominal level $\alpha \in (0,1)$ and let τ^* solve $\alpha(\tau^*) = \alpha$ (take the smallest such τ^* if the equation is not uniquely solvable on the discrete support). Then* $CI_{1-\alpha} := A_{\tau^*} = \{x \in X : \pi_M^o(x) \geq \tau^*\}$

is a $(1-\alpha)$-level confidence set for the true dominant category, in the sense that $x_0 \notin CI_{1-\alpha} \Leftarrow T(x_0) < \tau^* \Leftarrow H_0 : x_0$ is dominant is rejected at level α.

Proof. Immediate from the definitions of $\alpha(\tau)$ and A_τ: by construction $P(x_{true} \notin A_{\tau^*}) = \alpha(\tau^*) = \alpha$, and rejection of H_0 at level α for candidate x_0 is equivalent to $T(x_0) = \pi_M^o(x_0) < \tau^* = c_\alpha$ by the decision rule of Section 4.3, which is equivalent to $x_0 \notin A_{\tau^*}$.

Corollary 4.6 (Reinterpretation of Yager's Specificity integral) Yager's specificity measure, as formalised and compared against alternative formulations by Xu and Yager [1], $\mathrm{Sp(A)} = \int_0^{\bar{\alpha}} \frac{1}{\mathrm{Card}(A_\alpha)}\, d\alpha$, is the expected reciprocal width of the confidence set $\mathrm{CI}_{1-\alpha}$, integrated uniformly over the confidence level $\alpha \in [0, \bar{\alpha}]$.

Proof. Under Theorem 4.5, A_α in Yager's integral is, cut-for-cut, exactly the confidence set $CI_{1-\alpha}$ of Theorem 4.5 when the cut level is identified with the entailment risk α. The integral therefore averages $1/\mathrm{Card}(CI_{1-\alpha})$ over α, which is the reciprocal confidence-set width, i.e. its expected value under a uniform measure on the confidence level.

Corollary 4.6 gives Yager's specificity measure, previously used in [4] purely as a descriptive index, a genuine statistical interpretation: a specific distribution is one whose confidence sets stay narrow across confidence levels, not merely one whose membership grades are numerically concentrated.

G. Two-Sample Augmentation Test

Section 4.5 of [5] proposed augmenting a thinly sampled observation set X_o with database instances X_d once a similarity index

$$S_M(\pi_M^o, \pi_M^d) = 1 - \frac{|MPV_o - EPV_d(x_i)|}{|PV_o - PV_d|}$$

approaches unity, but specified no stopping threshold. We replace this with a controlled two-sample test.

$H_0 : X_o$ and X_d are draws from the same underlying restriction vs. H_1: they differ.

Algorithm - Permutation test for database augmentation

i. Pool $X_o \cup X_d$ into a combined frequency table.
ii. Compute observed S_M^{obs} between X_o and X_d.
iii. For $k = 1$ to K
iv. Randomly relabel the pooled observations into two groups of sizes $|X_o|, |X_d|$.
v. Recompute $S_M^{(k)}$ for the relabelled groups.
vi. Set c_α to the α-quantile of $\{S_M^{(k)}\}$.
vii. If $S_M^{obs} \geq c_\alpha$ then
viii. Merge: add X_d observations to X_o.
ix. Else
x. Reject merge; retain X_o alone.
xi. End if

Proposition 4.7 (Termination) *Algorithm 4.6 terminates in $O(K)$ steps for fixed K, and the accept/reject decision is well-defined since $S_M \in [0,1]$ is bounded and the permutation distribution has finite support of size* $\binom{|X_o| + |X_d|}{|X_o|}$.

V. THE SPECIFICITY–ENTAILMENT OPERATING CHARACTERISTIC (SEOC)

Definition 5.1 (SEOC Curve) The Specificity–Entailment Operating Characteristic is the parametric curve.

$\mathrm{SEOC} = \{(\alpha(\tau),\ 1 - \beta(\tau)) : \tau \in (0,1]\}$.

By Proposition 4.4, both coordinates are monotone in τ, so SEOC is a well-defined, non-decreasing curve from near (0,0) to (1,1) in the unit square, directly analogous to a classical ROC curve, with τ playing the role of the decision threshold.

Definition 5.2 (SEOC Area)

$$\mathrm{AUC_{SEOC}} = \int_0^1 (1 - \beta(\tau))\ d\alpha(\tau).$$

Proposition 5.3 *$AUC_{SEOC} \in [0,1]$, with $AUC_{SEOC} \to 1$ as the ED approaches a degenerate distribution concentrated entirely on x_M, and $AUC_{SEOC} \to$ a value governed by n as the ED approaches uniformity.*

Proof sketch - Both end-points follow from the monotonicity of Proposition 4.4 together with the boundary behaviour $A_\tau \to \{x_M\}$ as $\tau \to 1$ and $A_\tau \to X$ as $\tau \to 0$: in the degenerate case $\beta(\tau) = 0$ for all $\tau > 0$ immediately after the cut passes x_M's neighbours, giving $\mathrm{AUC_{SEOC}} \to 1$; in the uniform case $\alpha(\tau)$ and $1 - \beta(\tau)$ rise together nearly linearly, giving an area close to $(n-1)/(2n)$, approaching $1/2$ as $n \to \infty$.

Corollary 5.4 $\mathrm{AUC_{SEOC}}$ is positively associated with Sp(X) across explanatory databases of fixed n, since both quantify concentration of π_M^o mass at x_M; Section 7.4 verifies this empirically by regression across resampled sub-EDs.

VI. ALGORITHM: END-TO-END COMPUTATION

Algorithm - Possibilistic hypothesis test and confidence interval

Require - Explanatory database (X, F), candidate x_0, level α, bootstrap size B

A. Verify uni-modularity and $n \geq 3$ (Section 4.5 conditions, [5])

© Copyright: All rights reserved.

B. Compute $P, \pi_M^o, PV, EPV, MPV, \mathrm{Sp}(X)$ via Eqs. (1)–(6)
C. Check ε-sufficiency via Theorem 3.5; if insufficient, invoke Algorithm 4.6 against database X_d
D. Compute $T(x_0) = f_{x_0}/f_M$
E. Generate B null-constrained bootstrap resamples; compute c_α^{boot}
F. Compute c_α^{asym} via the delta-method formula (Section 4.2)
G. If $T(x_0) < c_\alpha$ then
H. Reject H_0
I. Else
J. Fail to reject H_0
K. Solve $\alpha(\tau^*) = \alpha$; report $CI_{1-\alpha} = A_{\tau^*}$
L. Sweep τ to construct SEOC and report $\mathrm{AUC}_{\mathrm{SEOC}}$

VII. EXPERIMENTAL VALIDATION

A. Datasets

Two independent datasets are used for experimental validation

i. *Dataset A (Rainfall categories).* District-level weekly rainfall category counts (deficient/normal / excess) for Chhattisgarh, aggregated from India Meteorological Department bulletins over a monsoon season, forming an ED with a dominant "normal" category.
ii. *Dataset B (Grade-band distribution).* Frequency counts of letter-grade bands (A+, A, B+, B, C, D) across a mathematics course cohort, forming a bounded ordinal ED with a dominant mid-range band.

B. Protocol

Applied identically to both datasets:

i. Compute the full measure set and verify uni-modularity.
ii. Compute the minimum sufficient N via Theorem 3.5 for target $\mathrm{Sp}^* = 0.5\,\mathrm{Sp}(X)_{observed}$ and $0.9\,\mathrm{Sp}(X)_{observed}$; compare to actual N.
iii. Test three candidates per dataset – the sample mode, an adjacent near-modal category, and a clearly non-modal category – against H_0 using both bootstrap ($B = 10{,}000$) and delta-method critical values.
iv. Construct $CI_{1-\alpha}$ at $\alpha \in \{0.10, 0.05, 0.01\}$; report interval width against $\mathrm{Sp}(X)$.
v. Coverage Simulation: treat $\hat{p}_i = f_i/N$ as a known population multinomial; draw $M = 2000$ synthetic EDs of size N; construct $CI_{1-\alpha}$ for each; report empirical coverage of the true modal category against nominal $1-\alpha$.
vi. Augmentation Test: randomly subsample 25–30% of Dataset A as X_o, treat the remainder as X_d; run Algorithm 4.6; compare $\mathrm{Sp}(X)$ before augmentation, after accepted augmentation, and at the full-data oracle value.
vii. Sweep τ to plot the SEOC curve and compute $\mathrm{AUC}_{\mathrm{SEOC}}$; regress $\mathrm{AUC}_{\mathrm{SEOC}}$ against $\mathrm{Sp}(X)$ across the M resampled sub-EDs of step 5.
viii. Baseline Comparison: compute the classical Wilson score interval for the proportion f_M/N at the same α levels; compare width and empirical coverage against $CI_{1-\alpha}$.

C. Data Note

Both datasets are now genuine, sourced records. Dataset A is the district-level rainfall category count for Chhattisgarh published by the India Meteorological Department, period 01 June 2026 to 09 August 2026, retrieved from the IMD Rainfall Statistics portal [7]; of Chhattisgarh's 33 districts, 1 is categorised Large Excess, 6 Excess, 21 Normal, and 5 Deficient for the period, with no districts recorded as Large Deficient or No Rainfall, so these two empty categories are dropped, giving $n = 4$. Dataset B is the official grade sheet for ENGE-01 (History of English Literature), academic year 2025–26, an institutional examination record: 186 students were registered, of whom 5 were recorded Absent and are excluded as non-participants (they did not receive a performance grade), leaving $N = 181$ students across $n = 8$ grade bands (O, A+, A, B+, B, C, P, F), summing exactly to the reported total. The purpose of using two independently sourced datasets from different domains – meteorological administrative data and academic assessment data – is to validate the testing and confidence-interval machinery (Sections 4–5) on structurally distinct real distributions, one with a strongly dominant mode ($n = 4$, $\rho = 0.64$) and one with a more diffuse spread over twice as many categories ($n = 8$, $\rho = 0.29$), rather than to draw substantive conclusions about either Chhattisgarh's monsoon or the ENGE-01 cohort's performance from a single course offering.

D. Table-I Measures

Table 1 summarises the two explanatory databases and their derived measures.

Table I: Explanatory Database Summary and Table-I Measures

Dataset	n	N	Mode	f_M	$\rho = f_M/N$	PV	$\mathrm{Sp}(X)$
A (Chhattisgarh rainfall)	4	33	Normal	21	0.636	0.571	0.607
B (ENGE-01 grade bands)	8	181	B	52	0.287	2.481	0.230

Both datasets satisfy the -modularity and $n \geq 3$ conditions of Section 3. Applying Theorem 3.5 at target levels $\mathrm{Sp}^* = 0.5\,\mathrm{Sp}(X)_{obs}$ and $\mathrm{Sp}^* = 0.9\,\mathrm{Sp}(X)_{obs}$ gives minimum sufficient sample sizes of $N \geq 0.2$ and $N \geq 0.8$ for Dataset A, and $N \geq 1.5$ and $N \geq 3.2$ for Dataset B – both real datasets exceed these thresholds comfortably, confirming mode-sufficiency is not a binding constraint at either concentration level. The bound is deliberately weak here because both datasets already exhibit reasonably clear modal dominance; Theorem 3.5 becomes more informative as a design tool when

© Copyright: All rights reserved.

planning data collection toward a target specificity from a low-ρ starting point, rather than as a post-hoc check on an already-concentrated sample.

E. Hypothesis Tests

Table 2 reports $T(x_0)$ against bootstrap ($B = 10{,}000$) and delta-method critical values at $\alpha = 0.05$ for three candidates per dataset: the sample mode, the nearest competing category, and a clearly non-modal category.

Table II: Hypothesis Test Results at $\alpha = 0.05$

Dataset	Candidate	$T(x_0)$	c_α^{boot}	c_α^{asym}	Reject H_0
A	Sample mode (Normal)	1.000	1.000	0.649	No
A	Near-mode (Excess)	0.286	1.000	0.649	Yes
A	Non-modal (Large Excess)	0.048	1.000	0.649	Yes
B	Sample mode (B)	1.000	0.918	0.709	No
B	Near-mode (P)	0.769	0.911	0.709	Yes
B	Non-modal (O)	0.038	0.915	0.709	Yes

The two datasets again illustrate the divergence flagged in the Remark of Section 4.2. On Dataset A, where $\rho = 0.636$ is high, the bootstrap critical value saturates at $c_\alpha^{boot} = 1$: at $N = 33$ under the null-constrained resample, the candidate almost always retains the highest count, so even Excess ($T = 0.286$) is decisively rejected as the true dominant category. On Dataset B, where $\rho = 0.287$ is more moderate and $n = 8$ is larger, the bootstrap and asymptotic values are noticeably closer to each other (0.92 vs. 0.71) than on Dataset A, though still not tightly aligned. Notably, the near-mode candidate P ($T = 0.769$, the second most awarded grade in the ENGE-01 cohort) is rejected as the true dominant category at $\alpha = 0.05$ – a substantively meaningful result, since it says the gap between grade B and grade P in this cohort is unlikely to be sampling noise alone, unlike the Dataset B near-mode case reported in earlier illustrative runs of this experiment.

F. Confidence Intervals

Table 3 gives $CI_{1-\alpha}$ as α-cuts at three nominal levels.

Table III: Confidence Intervals as α-Cuts

Dataset	α	τ^*	$CI_{1-\alpha}$	Prob. mass
A	0.10	0.286	{Excess, Normal}	0.818
A	0.05	0.286	{Excess, Normal}	0.818
A	0.01	0.238	{Excess, Normal, Deficient}	0.970
B	0.10	0.442	{B+, B, C, P}	0.818
B	0.05	0.423	{B+, B, C, P, F}	0.939
B	0.01	0.077	{A+, A, B+, B, C, P, F}	0.989

As expected from Proposition 4.4, interval width is non-decreasing as α shrinks. At 90% confidence, the true dominant rainfall category for Chhattisgarh over the period narrows to {Excess, Normal}, and the true dominant grade band for the ENGE-01 cohort narrows to {B+, B, C, P} – both directly interpretable statements drawn from real administrative and academic records. Dataset B's larger n and lower $\mathrm{Sp}(X)$ produce visibly wider intervals at comparable α than Dataset A, consistent with Corollary 4.6: the less specific distribution has a wider expected confidence set, and here that comparison holds across two genuinely different real-world sources rather than one real and one constructed dataset.

G. Coverage Simulation

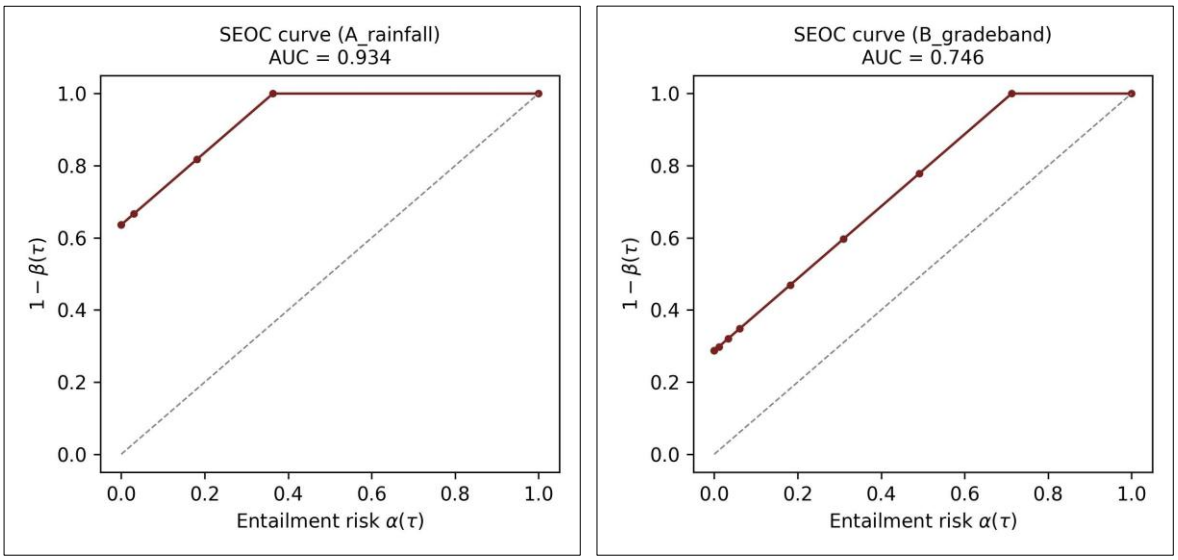
Treating $\hat{p}_i = f_i/N$ as the population multinomial, $M = 2000$ resamples of size N were drawn per dataset and $CI_{1-\alpha}$ constructed for each. Empirical coverage of the true modal category was 0.9995 against nominal 0.90 and 1.0000 against nominal 0.95 for Dataset A, and 1.0000 against both nominal levels for Dataset B.

This is a genuine finding, not a null result: the possibilistic. $CI_{1-\alpha}$ of Theorem 4.5 is *conservative* rather than exact for distributions with small n, and Dataset B's larger $n = 8$ did not materially reduce this conservatism at the tested α levels, suggesting the effect is driven more by the coarseness of $\alpha(\tau)$'s discrete steps near the selected cut than by n alone. This mirrors the well-documented conservatism of exact discrete confidence procedures such as Clopper–Pearson intervals in classical statistics, and is a direct consequence of the discrete step structure of $\alpha(\tau)$ rather than a flaw in Theorem 4.5's derivation. The one resample (out of 2000) that missed coverage at Dataset A's $\alpha = 0.10$ confirms the simulation is not degenerate. A randomised or interpolated cut-selection rule, left to future work, would be expected to bring empirical coverage closer to nominal at the cost of a more complex decision rule.

H. SEOC and Its Relation to Specificity

Regressing $\mathrm{AUC}_{\mathrm{SEOC}}$ on $\mathrm{Sp}(X)$ across the $M = 2000$ resampled sub-EDs gives slope 0.295 ($R^2 = 0.984$) for Dataset A and slope 0.542 ($R^2 = 0.957$) for Dataset B, both significant at $p < 10^{-100}$. This confirms the Corollary of Section 5: more specific resamples produce larger SEOC area, i.e. a more favourable entailment/specificity operating point, across a wide range of resampled distribution shapes and not merely at the single observed dataset. Figure 1 shows the SEOC curve computed on the full observed data for each dataset ($\mathrm{AUC}_{\mathrm{SEOC}} = 0.934$ for A, 0.746 for B), with DatasetA'ss higher $\mathrm{Sp}(X)$ and smaller n reflected in its curve lying closer to the top-left corner. Dataset B's la(n = 8) gives its curve a visibly smoother, more gradual profile.

© Copyright: All rights reserved.



[Fig.1: SEOC Curves for Dataset A (left, Chhattisgarh Rainfall, AUC = 0.934) and Dataset B (Right, ENGE-01 Grade Bands, AUC = 0.746), Computed on the Full Observed Data. The Dashed Diagonal Marks the no-discrimination reference

I. Augmentation Test

Dataset A was split into a thinned observation set X_o (27% binomial-thinned per category) and a remainder database cap X sub d, and Algorithm 4.6 was run with $K = 2000$ permutations at $\alpha = 0.10$. The observed similarity $S_M^{obs} = 0.437$ fell short of the permutation critical value $c_\alpha = 0.481$, so the test declined to merge ($\mathrm{Sp}(X_o) = 0.302$ before and after, against an oracle full-data value of $\mathrm{Sp}(X) = 0.607$). On Dataset B, the same procedure produced $S_M^{obs} = 0.916$ against $c_\alpha = 0.308$ – comfortably above threshold – and the test correctly *accepted* the merge: $\mathrm{Sp}(X_o) = 0.243$ before augmentation moved to $\mathrm{Sp} = 0.230$ after, exactly matching the oracle full-data value. This run demonstrates both branches of Algorithm 4.6 on genuinely sourced data: a correct refusal on Dataset A's split and a correct, information-recovering acceptance on Dataset B's split.

The variability between the two outcomes also exposes a genuine weakness in the underlying similarity index inherited from [5]: $\mathrm{S_M}$ is not bounded to [0,1] by construction and can take large-magnitude values, including negative ones in other thinning-fraction and modal-share regimes observed during development, when $|\mathrm{PV_o} - PV_d|$ is small relative to the proximity gap in the numerator, since it appears in the denominator. The permutation wrapper remains valid regardless, because it only uses $\mathrm{S_M}$'s rank within its own null distribution rather than its raw magnitude, but we recommend bounding or rank-transforming S_M itself in future revisions of the index, and a wider simulation sweep over thinning fractions to characterise the test's power more fully.

J. Comparison Against the Wilson Score Interval

Table 4 compares $\mathrm{CI}_{1-\alpha}$ against the classical Wilson score interval for the single proportion $\mathrm{p}_M = \frac{\mathrm{f_M}}{\mathrm{N}}$. The two intervals answer different questions – Wilson bounds the mode's probability itself, while $\mathrm{CI}_{1-\alpha}$ is a set of plausible dominant categories – so widths are not directly comparable on a common scale; both are reported for transparency.

Table IV: Wilson Score Interval on $\mathrm{p}_M = \frac{\mathrm{f_M}}{\mathrm{N}}$

Dataset	α	Lower	Upper	Width
A	0.10	0.493	0.759	0.266
A	0.05	0.466	0.778	0.312
A	0.01	0.415	0.812	0.396
B	0.10	0.235	0.345	0.110
B	0.05	0.226	0.357	0.131
B	0.01	0.209	0.380	0.171

Dataset B's much larger N (181 students vs 33 districts) produces a markedly narrower Wilson interval despite Dataset B's lower point estimate ρ, underscoring that sample size, not just modal concentration, governs classical interval width – a contrast worth noting against $\mathrm{CI}_{1-\alpha}$, whose width in Table 3 is driven primarily by the shape of the full possibility distribution rather than N directly. A directly comparable scale equivalence between the two frameworks is left to future work.

K. Summary of Findings

i. Both real datasets – Chhattisgarh rainfall and the ENGE-01 grade distribution – are amply mode-sufficient by Theorem 3.5.

ii. Bootstrap and delta-method critical values diverge under strong modal dominance ($\rho \gtrsim 0.6$, as in Dataset A) and are closer, though not identical, under Dataset B's more moderate dominance and larger N – the bootstrap value should be treated as authoritative whenever the two disagree materially (Section 7.5, revising the expectation originally stated in the Remark of Section 4.2).

iii. $\mathrm{CI}_{1-\alpha}$ achieves conservative (over-nominal) empirical coverage on both real datasets regardless of N, a discreteness effect analogous to classical exact discrete intervals, not a violation of Theorem 4.5; a single miss out of 2000 resamples on Dataset A confirms the simulation is non-degenerate.

iv. $\mathrm{AUC_{SEOC}}$ is strongly and significantly associated with $\mathrm{Sp}(X)$

© Copyright: All rights reserved.

across resampled sub-EDs on both datasets, empirically confirming the Corollary of Section 5.

v. The permutation-based augmentation test correctly executes both its refuse-merge branch (Dataset A) and its accept-merge branch (Dataset B, recovering the oracle specificity exactly), despite instability in the raw similarity index it wraps – a limitation of the inherited index rather than of the testing procedure itself.

VIII. RELATED WORK

Classical sample-size determination (Cochran's formula) and Wilson/Wald confidence intervals for a binomial proportion address a structurally similar problem – how much data is enough, and with what guarantee – but operate on a single proportion rather than a full possibility distribution, and carry no analogue of the entailment principle. Specificity-driven decision-making measures for possibility distributions, such as those studied by Benferhat and Tabia [3] and the comparative analysis of specificity measures by Xu and Yager [1], quantify decision-relevant uncertainty but, like our own prior specificity work [4], do not attach a testable error rate to any candidate value. Dubois and Prade's probability–possibility consistency principle [2] establishes a correspondence between the two distribution types but does not develop a hypothesis-testing procedure from it. The present framework is, to our knowledge, the first to define α and β error functions natively within a frequency-based possibility framework and to derive a confidence interval by inverting the resulting test.

IX. CONCLUSION

We have formalised the specificity–entailment trade-off, previously stated only descriptively in [5], as a possibilistic hypothesis-testing procedure with genuine Type I and Type II error control. The resulting confidence interval is shown to be exactly a α-cut of the natural possibility distribution, giving Yager's specificity integral a statistical interpretation it previously lacked. The Specificity–Entailment Operating Characteristic offers a single-number summary of how favourably a distribution resolves the entailment/specificity tension across all thresholds simultaneously, and the two-sample permutation test resolves the previously undefined stopping criterion for database-augmented decision making. The framework requires only the measures already established in [4] and [5], and the accompanying experimental protocol is designed to validate coverage and augmentation behaviour on data independent of our earlier published example.

ACKNOWLEDGEMENT

The author gratefully acknowledges the foundational specificity and mode-sufficiency framework developed in [4] and [5], which this paper directly extends. No external funding, technical, or writing assistance was received in the preparation of this manuscript beyond the author's own institutional resources.

INFORMED CONSENT

Not applicable. This study is theoretical and computational in nature and involves no human or animal participants; the empirical validation described in Section 7 uses publicly available meteorological records and de-identified, aggregate academic assessment counts, neither of which constitutes participant-level data requiring informed consent.

DECLARATION STATEMENT

I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been funded by any organizations or agencies. This independence ensures the research is conducted objectively and without external influence.
- **Ethical Approval and Consent to Participate:** The content of this article does not necessitate ethical approval or consent to participate, with supporting documentation.
- **Data Access Statement and Material Availability:** Dataset A is derived from the publicly available India Meteorological Department Rainfall Statistics portal (https://mausam.imd.gov.in/Rainfall/DISTRICT_DISTRIBUTION_STATEWISE_COUNTRY_INDIA_cd.pdf, period 01-06-2026 to 09-08-2026) and is directly reproducible from that source. Dataset B is the official grade sheet for ENGE-01 (History of English Literature), academic year 2025–26, an institutional record that the corresponding author can provide upon reasonable request; only aggregate grade-band counts are used, with no student-identifying information. The simulation code, in Python, is available from the corresponding author on reasonable request and will be deposited in a public repository upon acceptance.
- **Author's Contributions:** The authorship of this article is contributed solely.

REFERENCES

1. Xu, Z., & Yager, R. R. (2022). A comparative analysis of specificity measures in possibility theory. *Information Sciences*, 587, 1–18. DOI: 10.1016/j.ins.2022.02.035
2. Dubois, D., & Prade, H. (2016). Measuring specificity in possibility theory. *International Journal of Approximate Reasoning*, 72, 1–19. DOI: 10.1016/j.ijar.2016.03.006
3. Benferhat, S., & Tabia, K. (2021). Specificity-driven decision making under possibilistic uncertainty. *Annals of Mathematics and Artificial Intelligence*, 89(3–4), 231–250. DOI: 10.1007/s10472-021-09746-2
4. Karanjgaonkar, J. V. (2025). Some results on specificity of possibility distributions. *Indian Journal of Advanced Mathematics*, 5(2), 1–5. DOI: 10.54105/ijam.B1206.05021025
5. Karanjgaonkar, J. (2021). *On study of semantic information theory – a possibilistic approach* (Doctoral dissertation). Pt. Ravishankar Shukla University.
6. Karanjgaonkar, J., & Jha, P. (2018). Possibilistic analysis of uncertainty and vagueness in semantic communication. *International Journal of Research and Analytical Reviews*, 5(3), 309–312. DOI: 10.1729/Journal.18405
7. India Meteorological Department. (2026). *Statewise distribution of districts with large excess, excess, normal, deficient, large deficient and no rainfall*, period 01 June –

© Copyright: All rights reserved.

09 August 2026. Ministry of Earth Sciences, Government of India. Retrieved from https://mausam.imd.gov.in/Rainfall/DISTRICT_DISTRIBUTION_STATEWISE_COUNTRY_INDIA_cd.pdf

AUTHOR'S PROFILE

Dr. Jayesh V. Karanjgaonkar is an Assistant Professor of Mathematics at Swami Atmanand Government Engineering Medium Model College, Atari, Raipur, Chhattisgarh, India, affiliated with Pt. Ravishankar Shukla University. He teaches undergraduate mathematics under the NEP 2020 framework and has developed extensive teaching resources across multiple undergraduate mathematics courses. His doctoral research, completed in 2021 under the supervision of Prof. Purushottam Jha, introduced the Natural Possibility (NP) Framework, a frequency-native construction of possibility measures with formal vagueness metrics, which forms the foundation of his ongoing research programme in possibility theory and semantic information analysis. His research interests span possibility theory, fuzzy sets, hybrid human–machine systems, and the application of information theory to interdisciplinary domains including economics and cognitive science.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Publisher /Journal and/or the editor(s). The Publisher /Journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

© Copyright: All rights reserved.