Cómo medir la validez y fiabilidad de sistemas de análisis forense

Geoffrey-Stewart Morrison

Preocupaciones

- Marco lógicamente correcto para la evaluación de las evidencias
  - ENFSI Guideline for Evaluative Reporting 2015; NCFS Views on statistical statements 2016
- Sin embargo, ¿cuál es la justificación para la opinión? ¿De dónde vienen los números?
  - Risinger a ICFIS 2011
- Demostración de validez y fiabilidad
  - *Daubert* 1993; NRC Report 2009; FSR Codes of Practice 2014; PCAST Report 2016
- Transparencia
  - *R v T* 2010
- Diminuir la influencia potencial de sesgo cognitivo
  - NIST/NIJ Human Factors in Latent Fingerprint Analysis 2012
- Comunicar la fuerza de la evidencia forense al juzgador de los hechos

Paradigma

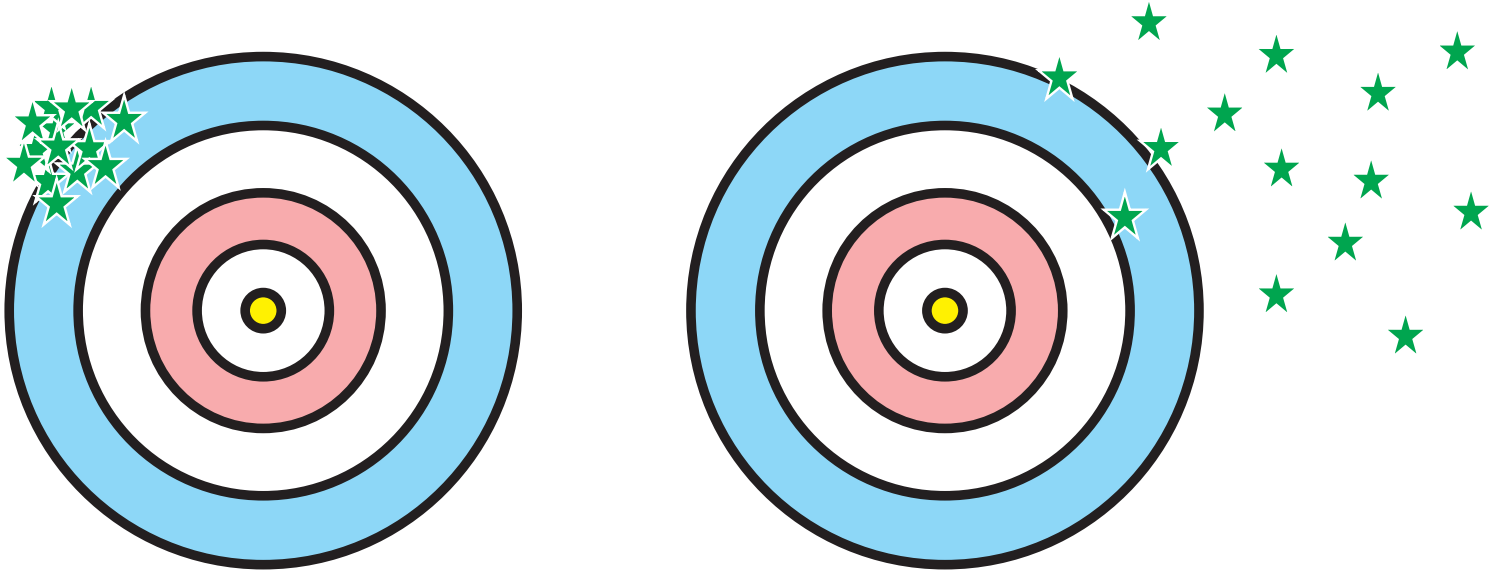
- Uso del marco de relaciones de verosimilitud para la evaluación de las evidencias
  - lógicamente correcta
- Uso de mediciones cuantitativas, datos relevantes (datos representativos de la población relevante), y modelos estadísticos
  - transparente y reproducible
  - relativamente robusto al sesgo cognitivo
- Evaluación empírica de validez y fiabilidad bajo condiciones que reflejan las condiciones del caso bajo investigación, con datos de prueba seleccionados de la población relevante
  - única manera de saber como bien funciona

Validez y Fiabilidad
(Exactitud y Precisión)

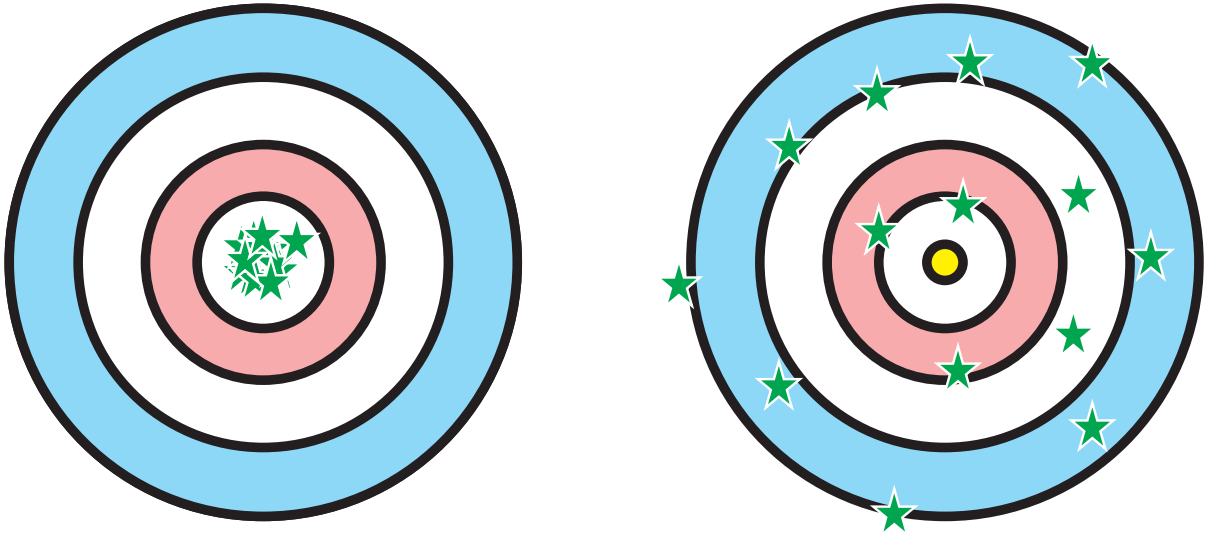
preciso

no preciso

no exacto



exacto



Cómo Medir Validez

Medir Validez

- El conjunto de prueba consiste de un gran número de pares de muestras, unos del mismo origen y otros de diferentes orígenes

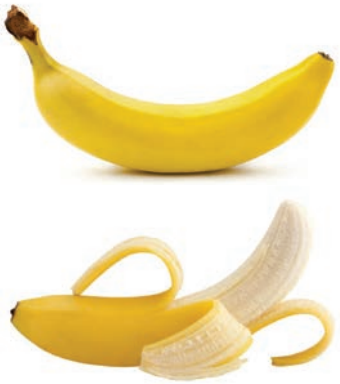
- **El conjunto de prueba debe representar la población relevante y las condiciones del caso bajo investigación**

- Se usa el sistema de comparación forense para calcular una RV por cada par de muestras de prueba

- Para cada par de muestras de prueba, se compara la salida del sistema con el conocimiento que se tiene sobre la entrada

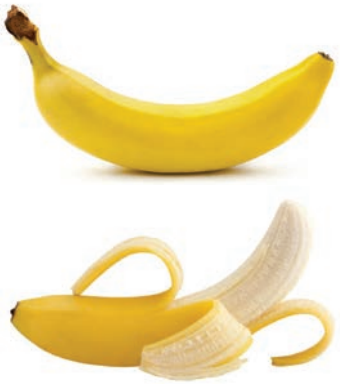

CAJA NEGRA
156


CAJA NEGRA
1/78

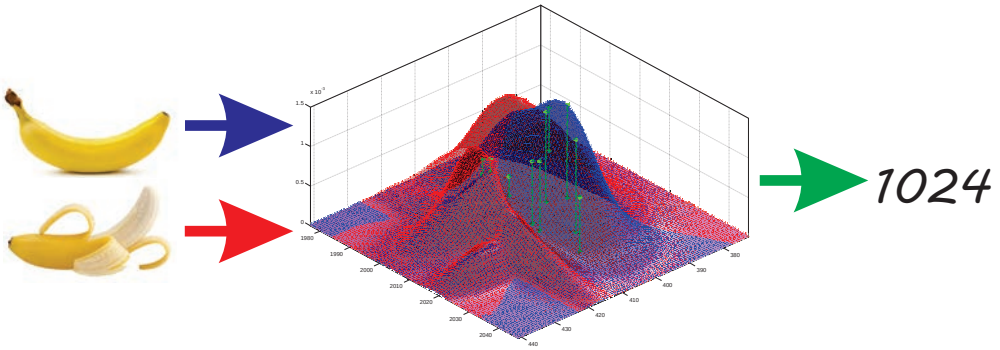


CAJA NEGRA

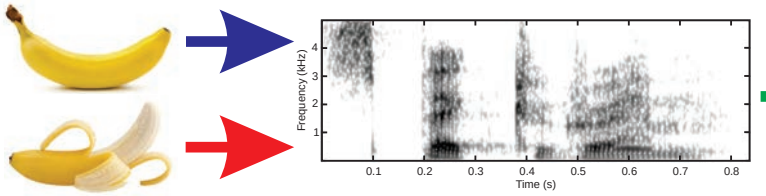
En un lugar de la Mancha, de cuyo nombre no quiero acordarme



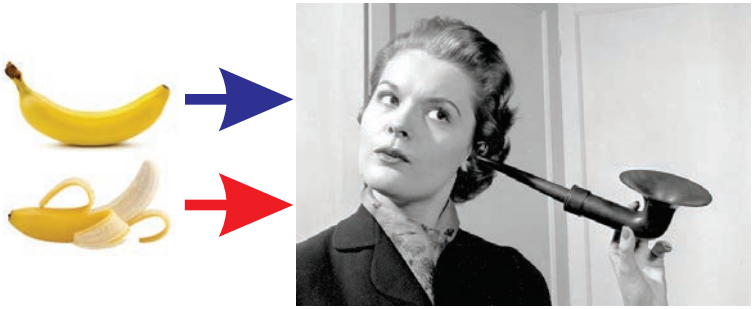
En un lugar de la Mancha, de cuyo nombre no quiero acordarme



1024



1,000,000



42

En un lugar de la Mancha


CAJA NEGRA
1024


CAJA NEGRA
1,000,000


CAJA NEGRA
42


CAJA NEGRA
En un lugar de la Mancha

Medir Validez

- La tasa de clasificación correcta / la tasa de clasificación errónea no es apropiada
  - se basa en probabilidades a posteriori
  - se usa un umbral en vez de presentar un valor gradiente

hecho	decisión mismo	diferente
mismo	aceptación correcta	rechazo falso
diferente	aceptación falsa	rechazo correcto

Medir Validez

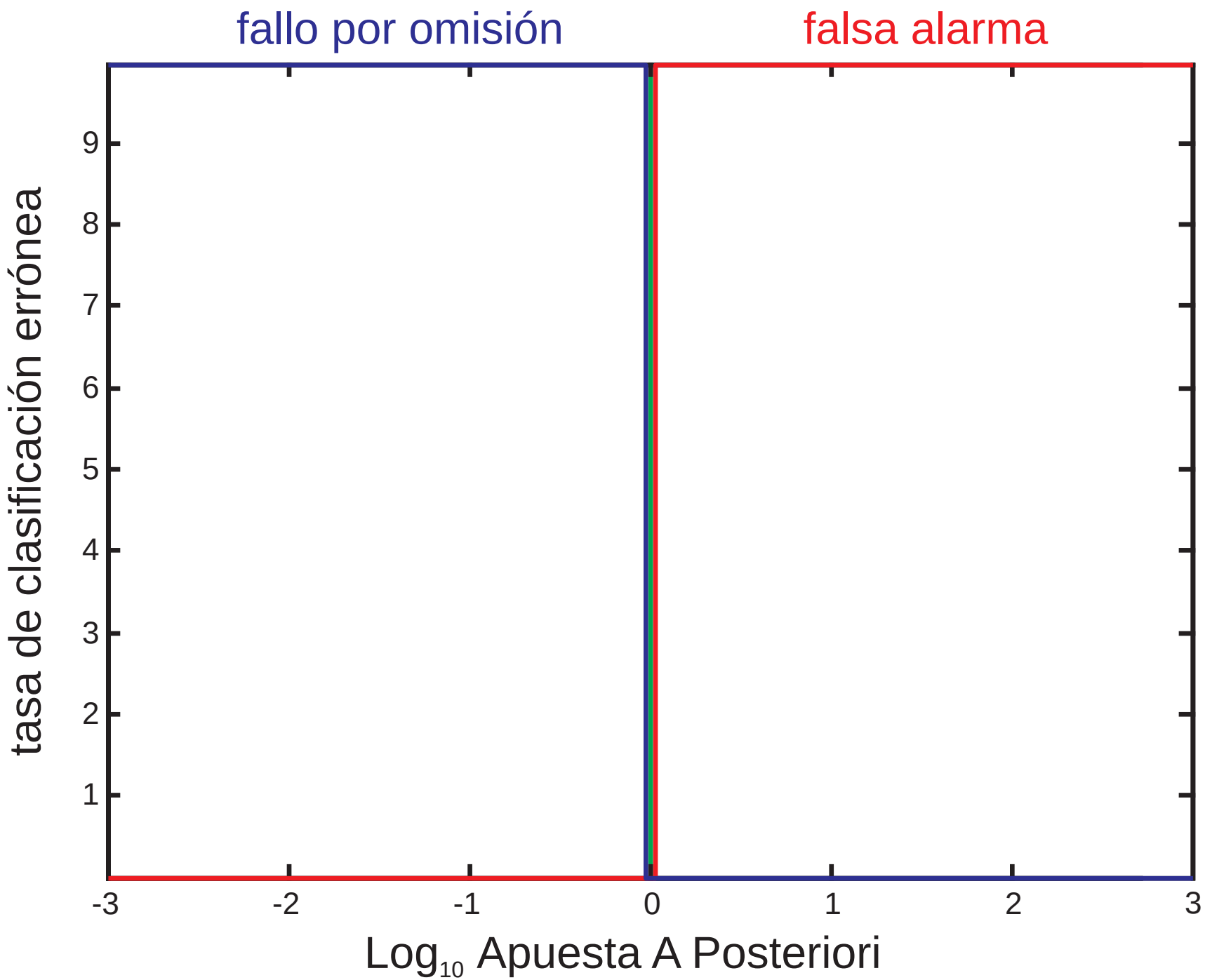
- La tasa de clasificación correcta / la tasa de clasificación errónea no es apropiada
  - se basa en probabilidades a posteriori
  - se usa un umbral en vez de presentar un valor gradiente

hecho	decisión mismo	decisión diferente
mismo		fallo por omisión
diferente	falsa alarma	

Medir Validez

- La tasa de clasificación correcta / la tasa de clasificación errónea no es apropiada
  - se basa en probabilidades a posteriori
  - se usa un umbral en vez de presentar un valor gradiente

hecho	decisión	
	mismo	diferente
mismo	0	1
diferente	1	0


fallo por omisión
falsa alarma
tasa de clasificación errónea
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log10 Apuesta A Posteriori

Medir Validez

- La validez se indica por el **grado** hasta que los de pares de muestras del mismo origen tengan RV > 1, y los de diferentes orígenes tengan RV < 1

- La validez se indica por el **grado** hasta que los pares de muestras del mismo origen tengan log(RV) > 0, y los de diferentes orígenes tengan log(RV) < 0

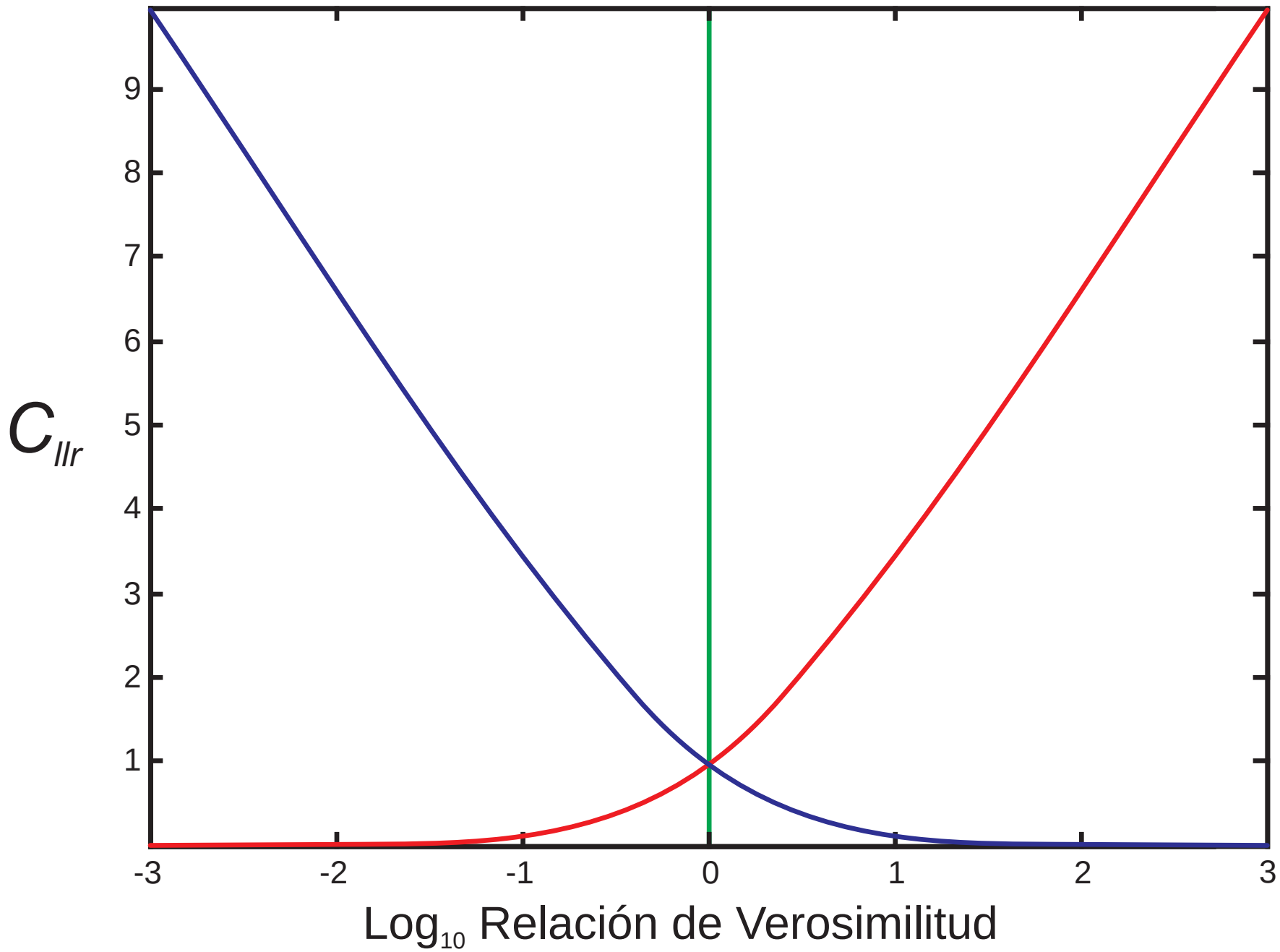
RV						
1/1000	**1/100**	**1/10**	**1**	**10**	**100**	**1000**
-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
$\log_{10}(RV)$						

Medir Validez

- Una medida continua que capta la validez de un conjunto de relaciones de verosimilitud procedentes de datos de prueba es el coste del logaritmo de la relación de verosimilitud, *log-likelihood-ratio cost*, C_{llr}

$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{mo}} \sum_{i=1}^{N_{mo}} \log_2 \left(1 + \frac{1}{RV_{mo_i}} \right) + \frac{1}{N_{do}} \sum_{j=1}^{N_{do}} \log_2 \left(1 + RV_{do_j} \right) \right)$$

Brümmer N, du Preez J (2006). **Application independent evaluation of speaker detection**, *Computer Speech & Language*, 20, 230–275. doi:10.1016/j.csl.2005.08.001

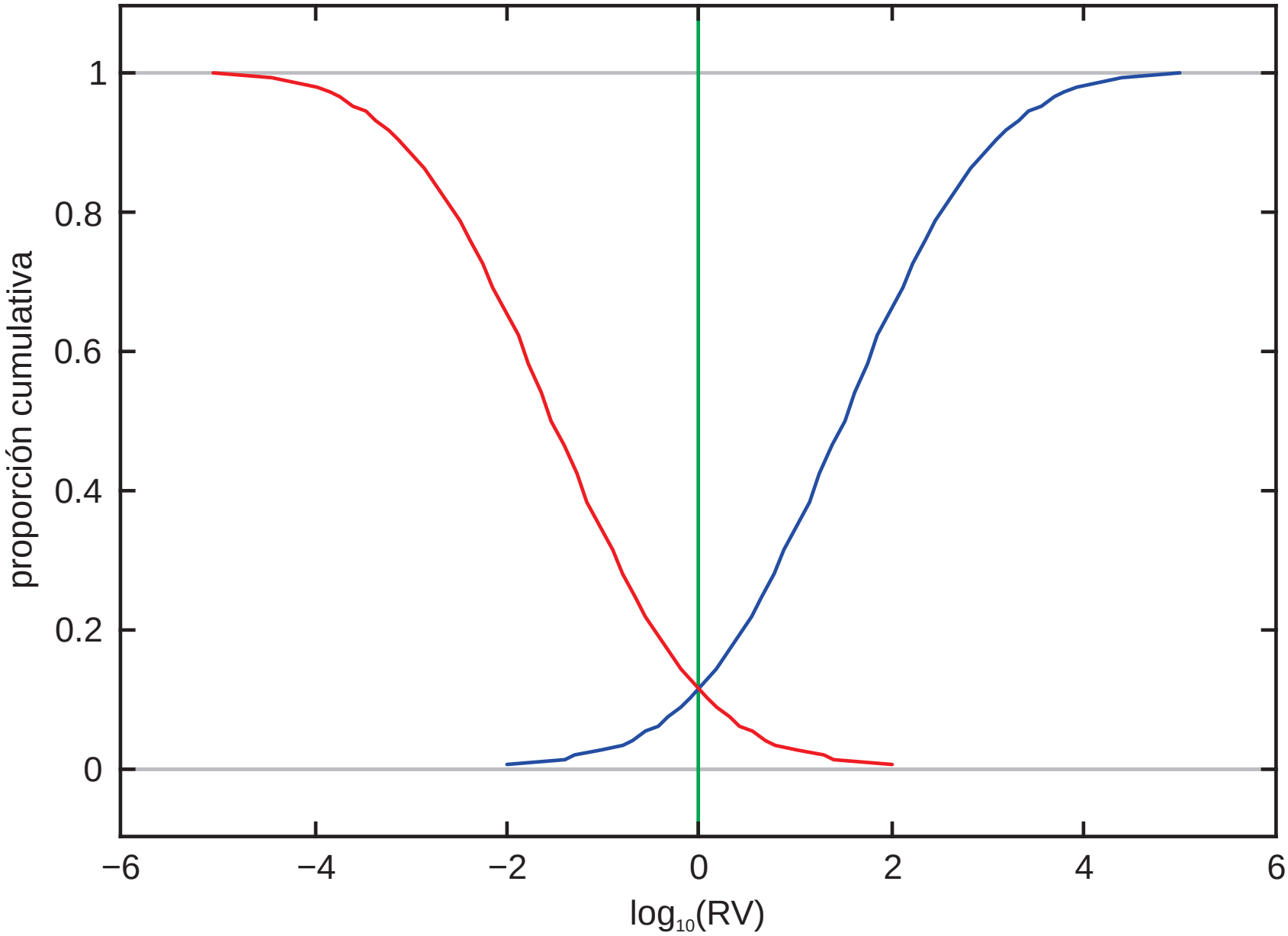

C_{llr}
9
8
7
6
5
4
3
2
1
-3
-2
-1
0
1
2
3
Log_{10} Relación de Verosimilitud

Medir Validez

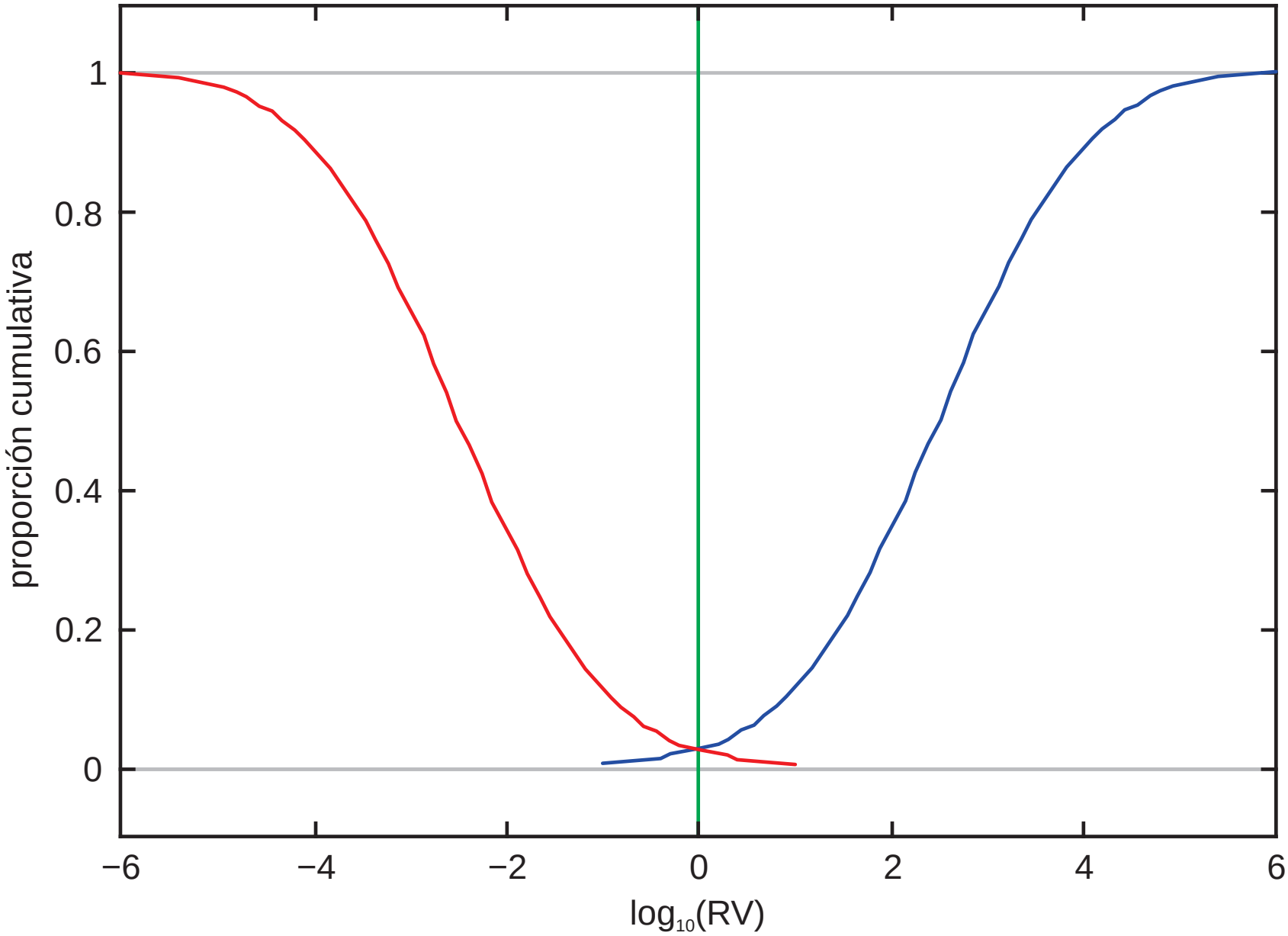
- Sistema A: $C_{llr} = 0.548$
- Sistema B: $C_{llr} = 0.101$
- Sistema C: $C_{llr} = 1.018$

Gráficos Tippett

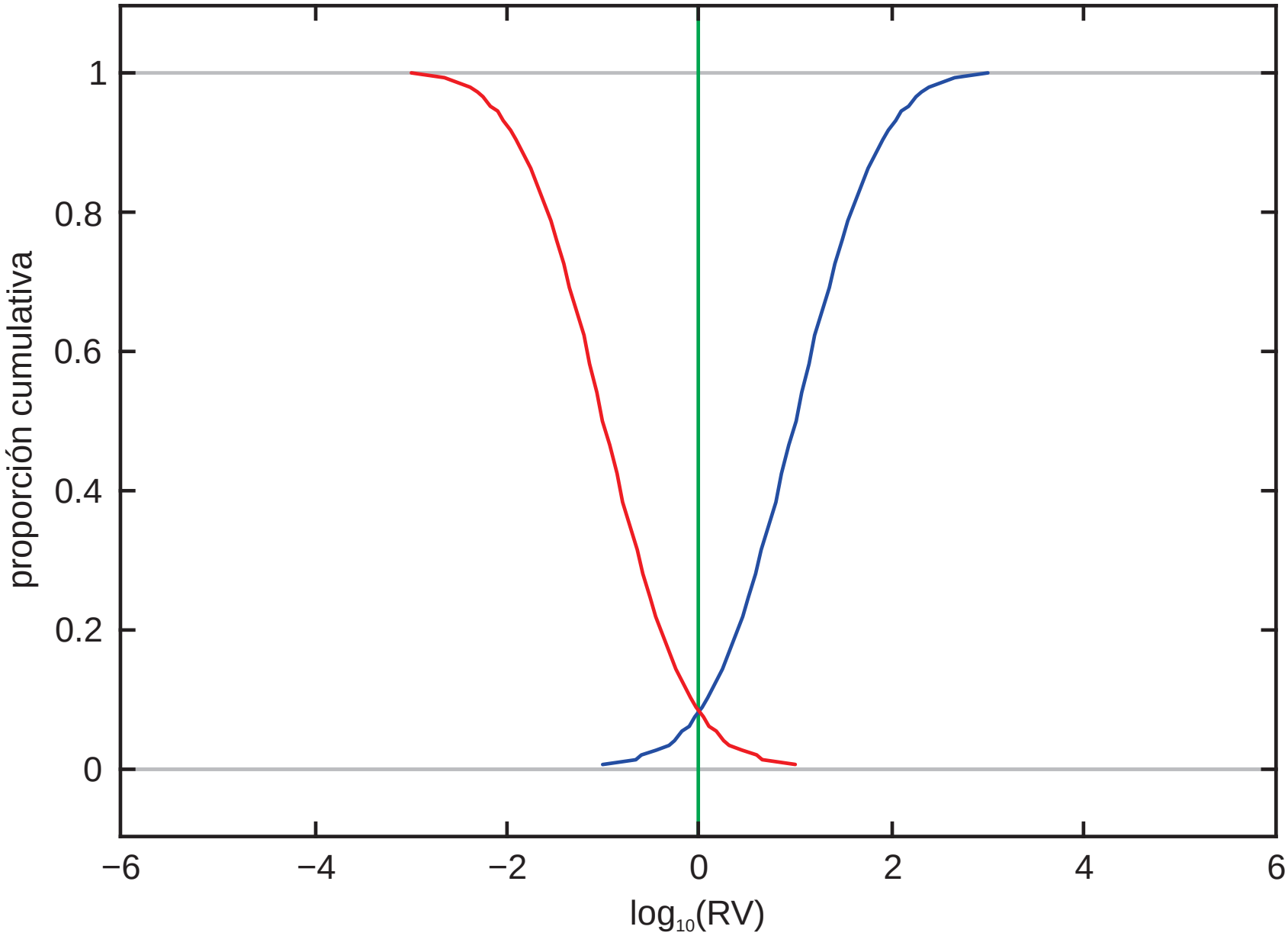
Gráficos Tippett



Gráficos Tippett

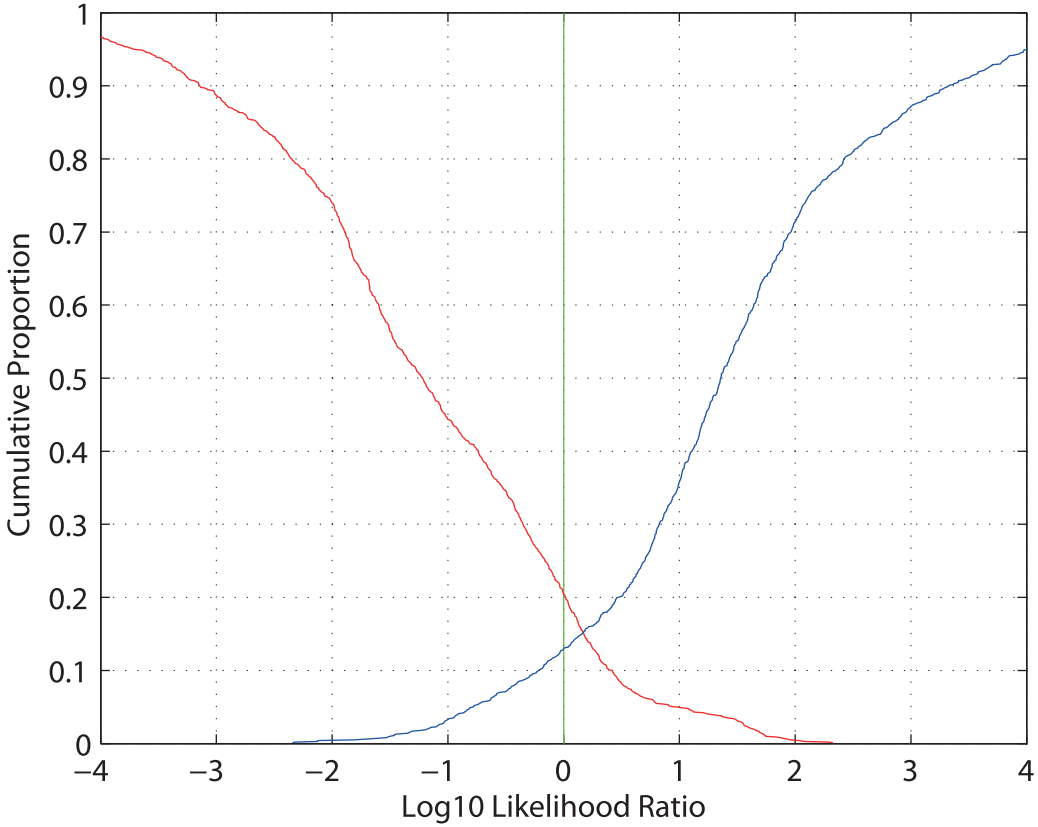


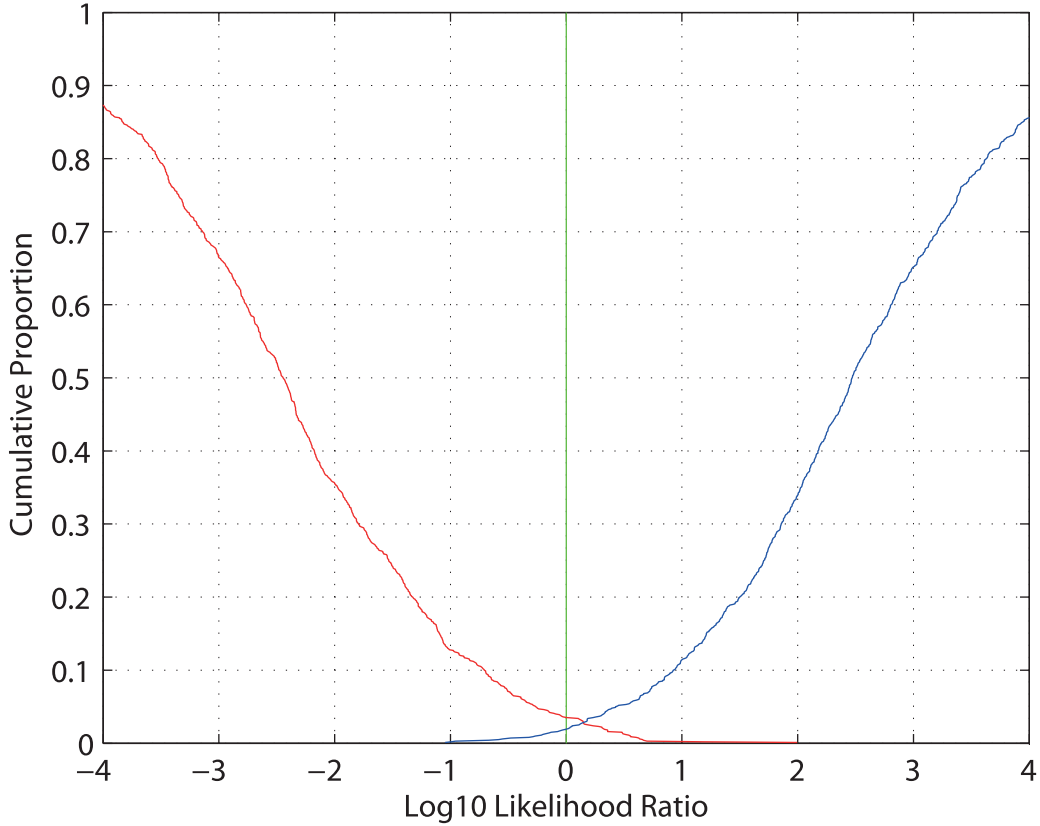
Gráficos Tippett



Gráficos Tippett

- Sistema A: $C_{llr} = 0.548$
- Sistema B: $C_{llr} = 0.101$





Cómo Medir Fiabilidad

Fuentes de imprecisión

- variabilidad intrínseca al nivel del fuente
  - intra-fuente inter-muestra variabilidad
- variabilidad en el proceso de trasferencia
- variabilidad en la técnica de medir
- variabilidad en tomar muestras de la población relevante
- variabilidad en la estimación de parámetros de modelos estadísticos

Morrison, G. S. (2016). **Special issue on measuring and reporting the precision of forensic likelihood ratios: Introduction to the debate**. *Science & Justice*. doi:10.1016/j.scijus.2016.05.002

Medir Fiabilidad

- Imagina que en el conjunto de prueba tenemos tres grabaciones (*A*, *B*, *C*) de cada locutor

- *A* tiene las mismas condiciones (estilo de habla, canal de transmisión, duración, etc.) como la grabación del delicuente

- *B* y *C* tienen las mismas condiciones como la grabación del sospechoso

- Usamos RVs calculados a base de pares *A*-*B* y *A*-*C* para estimar un intervalo de credibilidad (IC) de 95%

Medir Fiabilidad

- Dos pares para cada comparación del mismo locutor

grab. del sospech.		grab. del delicuente	
001	B	001	A
001	C	001	A
002	B	002	A
002	C	002	A
:	:	:	:

Medir Fiabilidad

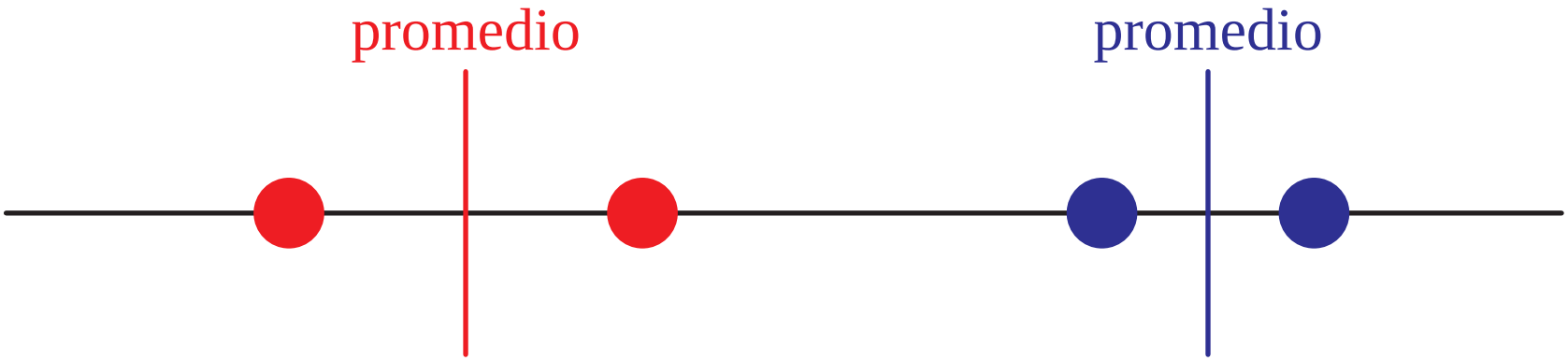
- Dos pares para cada comparación de diferentes locutores

grab. del sospech.		grab. del delicuente	
002	B	001	A
002	C	001	A
003	B	001	A
003	C	001	A
:	:	:	:
001	B	002	A
001	C	002	A
:	:	:	:

Medir Fiabilidad

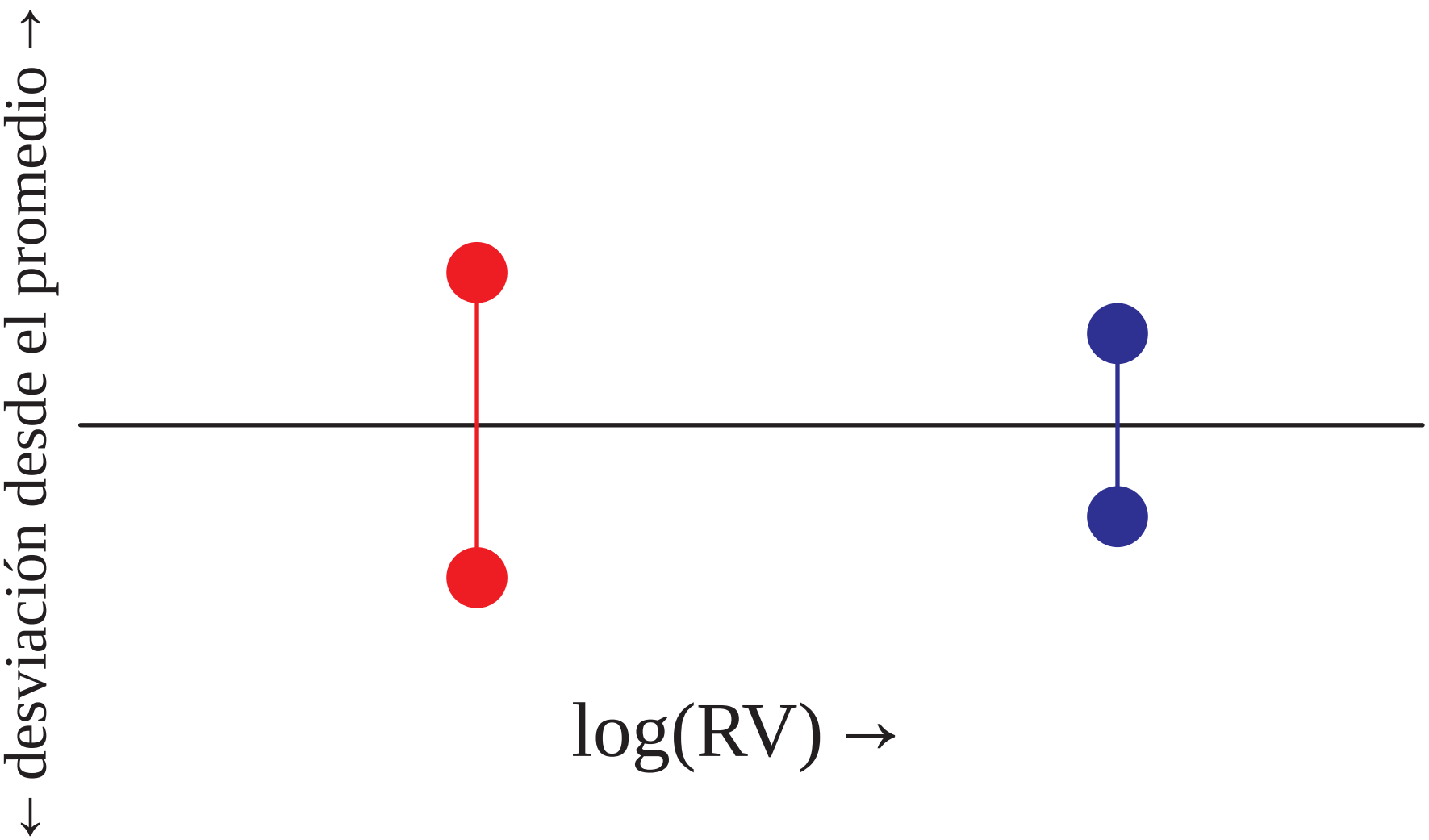
log(RV) →

Medir Fiabilidad



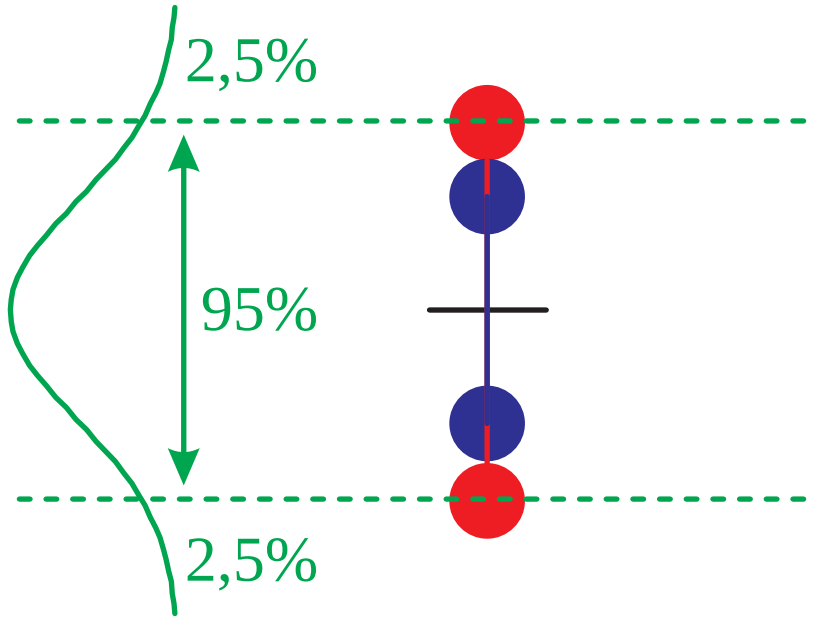
log(RV) →

Medir Fiabilidad



Medir Fiabilidad

← desviación desde el promedio →



Medir Validez y Fiabilidad

- Sistema A: C_{llr} = 0.548 95% CI = ±0.498

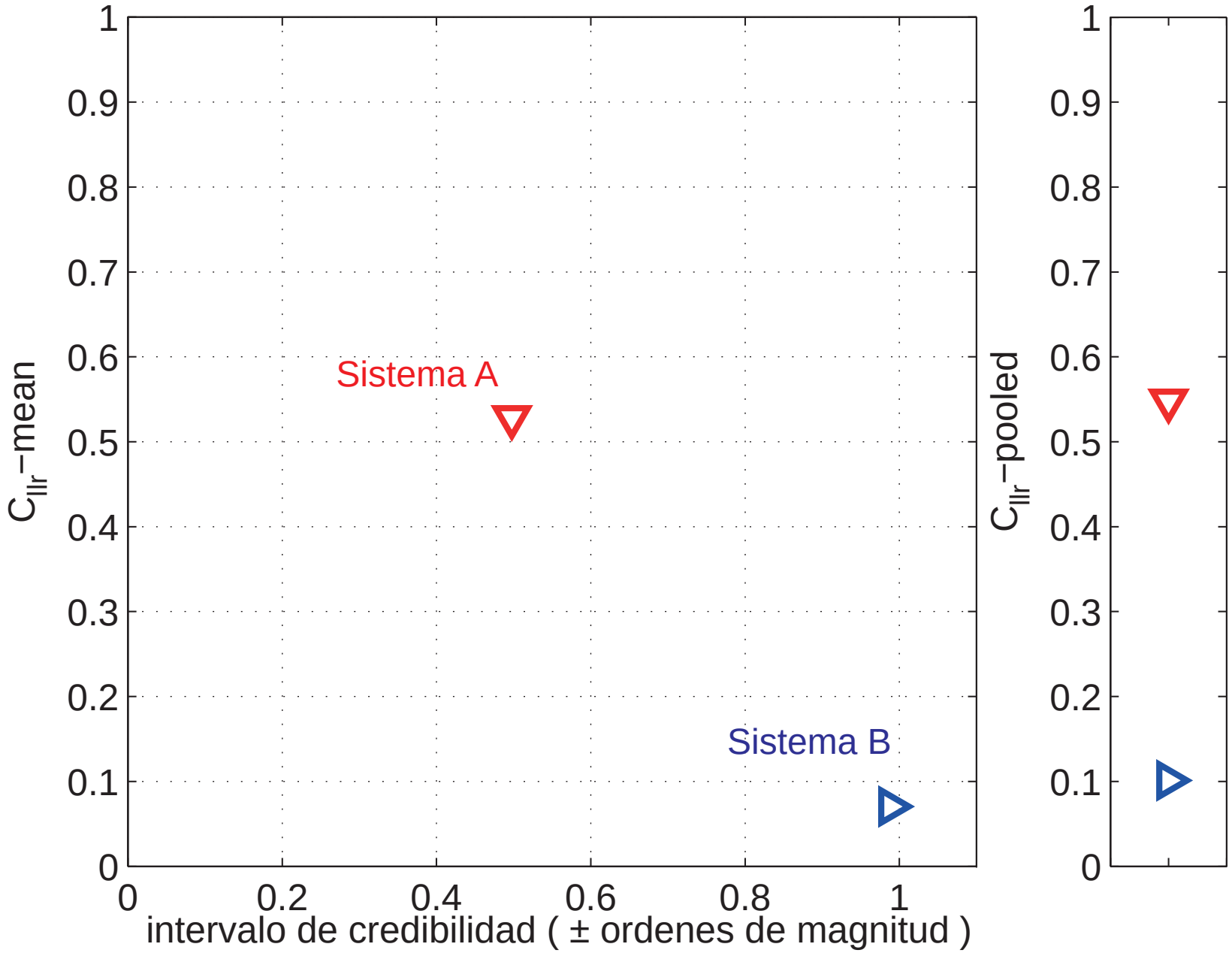
- Sistema B: C_{llr} = 0.101 95% CI = ±0.988

Medir Validez y Fiabilidad

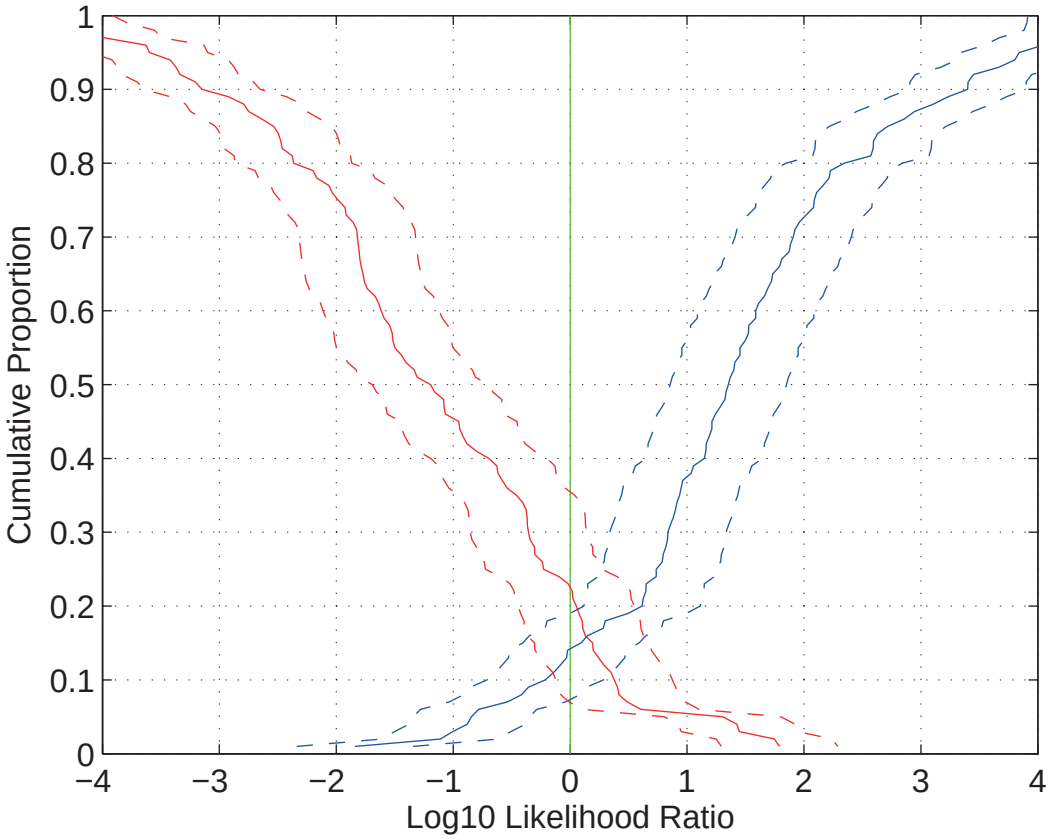
- System A: $C_{llr} = 0.548$ $C_{llr}^{\text{promedio}} = 0.529$ 95% CI = ±0.498

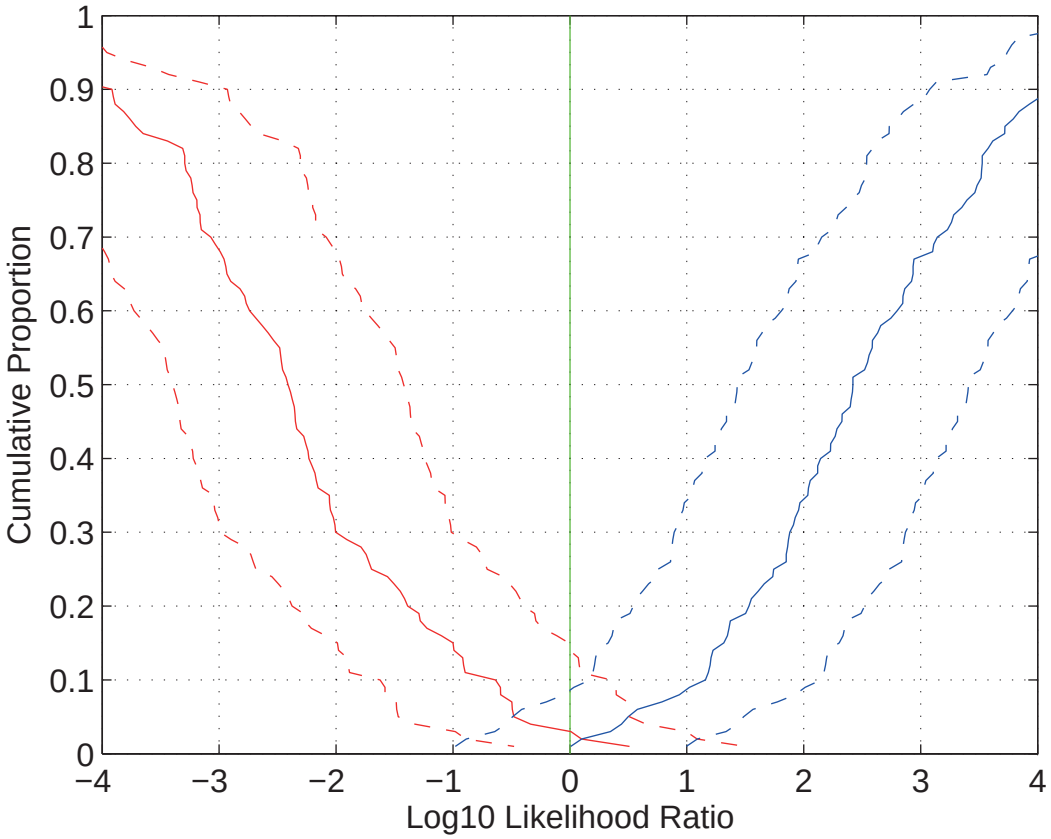
- System B: $C_{llr} = 0.101$ $C_{llr}^{\text{promedio}} = 0.071$ 95% CI = ±0.988

Medir Validez y Fiabilidad



Gráficos Tippett





Sumario

Si fueran consistentes los datos de fondo, calibración, y prueba con las condiciones del caso bajo investigación, y si la comparación entre las grabaciones del delicuente y del sospechoso resultara en una relación de verosimilitud de 100 ($\log_{10}(RV)$ de +2), y la estimación del IC 95% arrojara un valor de ±1 ordenes de magnitud (±1 en $\log_{10}(RV)$), el científico forense podría presentar una declaración como la siguiente:

Basado en mi evaluación de las evidencias, he calculado que las propiedades acústicas de la grabación del delicuente sería 100 veces más probable si la grabación hubiera sido producido por el acusado en contraste de que si hubiera sido producido por otro locutor de la población relevante.

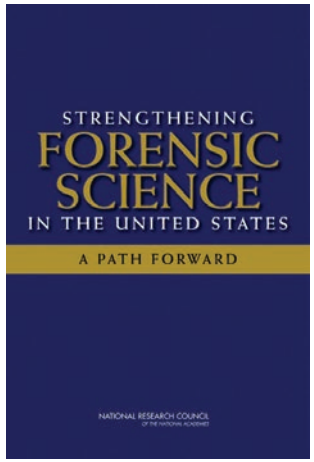
Lo anterior significa que cualquiera que haya sido su creencia previa sobre las probabilidad de que el locutor en la grabación del delicuente fuera el acusado relativa a la probabilidad de que fuera otro locutor, ahora su creencia en el valor de la probabilidad de que sea el acusado relativa a la de que sea otro locutor debe ser 100 veces más de lo que fuera antes.

Basado en mis calculaciones, tengo una certeza de 95% que obtener estas propiedades acústicas es a lo menos 10 veces más probable y no más que 1000 veces más probable si el locutor en la grabación del delicuente fuera el acusado contra de que fuera otro locutor.

Validación Empírica

Validación Empírica

- El Informe al Congreso del National Research Council sobre *Strengthening Forensic Science in the United States* (2009) urgió la adopción de procedimientos que incluyen:
  - "medidas cuantificables de la fiabilidad y exactitud de los análisis forenses" (p. 23)
  - "la presentación de una medición con un intervalo que tiene alta probabilidad de contener el valor verdadero" (p. 121)
  - "la realización de estudios de validación de la eficacia de un procedimiento forense" (p. 121)



Validación Empírica

- Los *Codes of Practice and Conduct* (2014) del Forensic Science Regulator de Inglaterra y Gales requiere:
  - "todos los métodos y procedimientos técnicos utilizados por un proveedor serán validados." (§20.1.1)
  - "Aun cuando un método se considera estándar y se utiliza ampliamente, todavía tendrá que ser demostrado la validación." (§20.1.3)
  - "validación se llevará a cabo utilizando material que simula casos reales ... y ... cuando sea apropiado, con el material de casos reales" (§20.7.3)
  - "demonstrar que pueden proporcionar resultados consistentes, reproducibles, válidos y fiables" (§20.9.1)

Forensic Science Regulator
Overseeing Quality

Validación Empírica

- Tribunal Supremo EEUU: *Daubert v Merrell Dow Pharmaceuticals* (1993)
  - "En un caso relacionado con la evidencia científica, *fiabilidad de evidencia* se basará en la *validez científica*" [énfasis en el original]
  - "evaluación de si el razonamiento o la metodología que subyace en el testimonio es científicamente válida y ... si este razonamiento o metodología puede aplicarse correctamente a los hechos en cuestión."
  - "una pregunta clave que se plantea en la determinación de si una teoría o técnica es el conocimiento científico que ayudará al juzgador de los hechos será si puede ser (y ha sido) sometido a prueba. ... "[L]as declaraciones que constituyen una explicación científica tienen que ser capaz de ser probado empíricamente'."
  - "en el caso de una técnica científica específica, el tribunal normalmente debe tener en cuenta la tasa de error conocida o potencial "

Validación Empírica

- Inglaterra y Gales: *Criminal Practice Directions* (2014)
  - "'el tribunal deberá asegurarse de que existe una base científica suficientemente fiable para que se admita la evidencia.'" (33A.4)
  - "si la opinión tenga adecuadamente en cuenta asuntos, tales como el grado de precisión o el margen de incertidumbre, que afecta a la exactitud o la fiabilidad de los resultados;" (33A.5c)
  - "posibles defectos ... que perjudican ... la fiabilidad, ...
    (a) ... no ... se somete a escrutinio suficiente (incluyendo, cuando sea apropiado, pruebas experimentales ...), ...
    (c) ... datos defectuosos;
    (d) ... se basa en un método o proceso que no se ha efectuado o aplicado correctamente, o que no era apropiado para su uso en el caso particular;" (33A.6)

Validación Empírica

- The President's Council of Advisors on Science and Technology informe *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods* (PCAST, 2016)
  - "Sin estimaciones de exactitud adecuados, la declaración de un examinador que dos muestras son similares, o incluso indistinguibles, carece de sentido científico: no tiene valor probatorio, y tiene un considerable potencial para un impacto perjudicial." (p 6)
  - "el experto no debe hacer afirmaciones o implicaciones que van más allá de la evidencia empírica y las aplicaciones de los principios estadísticos válidos a esa evidencia." (p 6)
  - "Donde no hay adecuados estudios empíricos y/o modelos estadísticos para proporcionar información significativa acerca de la exactitud de un método de comparación forense de características, los abogados del Departamento de Justicia y los examinadores no deben ofrecer un testimonio basado en el método." (p 19)

Experiencia

Experiencia

- Para un perito decir “Creo que esto es verdad porque llevo *x* años ejerciendo este trabajo” no es, en mi opinión, científica. Por otro lado, para un perito decir “Creo que esto es verdad y mi juicio ha sido probado en experimentos controlados” es fundamentalmente científica.

Evett IW (1991) **Interpretation: a personal odyssey**. In C.G.G. Aitken, D.A. Stoney (Eds.), *The Use of Statistics in Forensic Science*. Ellis Horwood, Chichester, UK. pp. 9–22.

Experiencia

- La experiencia en la aplicación de identificación de voz espectrográfica para metas judiciales ha llevado a los proponentes del método de expresar confianza en su fiabilidad. Sin embargo, la base de esta confianza no es accesible para evaluación objetiva.

- La validación de este enfoque para la identificación de voz se convierte en una cuestión de experimentos replicables con el propio experto, considerado como una máquina de identificación de voz. ... La validación requiere una evaluación experimental de rendimiento en tareas pertinentes. ... Se puede objetar que este conjunto mínimo de pruebas es excesivamente difícil. No creemos que lo es. Como científicos no podríamos aceptar menos en la comprobación de la fiabilidad de un “caja negra” que supuestamente realiza la identificación del hablante.

Bolt RA, Cooper FS, David EE Jr., Denes PB, Pickett JM, Stevens KN (1970) **Speaker identification by speech spectrograms: a scientists’ view of its reliability for legal purposes**. *Journal of the Acoustical Society of America* 47, 597–612, http://dx.doi.org/10.1121/1.1911935.

Experiencia

- The President's Council of Advisors on Science and Technology informe *Forensic science in criminal courts: Ensuring scientific validity of feature-comparison methods* (PCAST, 2016)
  - "ni la experiencia, ni el juicio, ni las buenas prácticas profesionales (tales como programas de certificación y acreditación, protocolos estandarizados, pruebas de aptitud, y códigos de ética) puede sustituir a la evidencia real de validez y fiabilidad fundamental. La frecuencia con la que se observó un patrón particular o conjunto de características en diferentes muestras, que es un elemento esencial en la elaboración de conclusiones, no es una cuestión de 'juicio.' Es una cuestión empírica para que sólo la evidencia empírica es relevante. Del mismo modo, la expresión de un experto de *confianza* basada en la experiencia personal profesional o expresiones de *consenso* entre los profesionales acerca de la exactitud de su campo no puede sustituir a las tasas de error estimadas a partir de los estudios pertinentes. Para los métodos de comparación forense de características, el establecimiento de la validez fundamental basada en la evidencia empírica es, pues, una condición *sine qua non*. Nada puede sustituir a ello." (p 6)

Gracias

http://geoff-morrison.net/

http://forensic-evaluation.net/