



Lección de Cinco Minutos de Ciencia: Métricas de Precisión de las Pruebas

Raymond Nelson

El objetivo práctico de cualquier prueba científica es clasificar, predecir y cuantificar un fenómeno de interés, a menudo denominado parámetro desconocido, para el que no son posibles ni la observación determinista (teóricamente perfecta) ni la medición física (sujeta únicamente a errores de medición aleatorios). A menudo, como ocurre con el comportamiento humano y los fenómenos sociales, el parámetro desconocido de interés es amorfo. Es decir, no hay sustancia física y, por tanto, no hay unidad física de medida. Por este motivo, todo uso de métodos de comprobación científica para los procesos de clasificación, predicción y cuantificación se basará en la teoría de la probabilidad y la estadística. Por este motivo, resulta útil comprender el vocabulario conceptual de las métricas de precisión de las pruebas.

La finalidad de cualquier prueba científica es cuantificar un fenómeno de interés que no puede someterse fácilmente a una medición física directa. La medición real requiere tanto un fenómeno físico como una unidad de medida

normalizada para ese fenómeno. Los científicos y los ingenieros pueden definir una unidad de medida normalizada para cualquier fenómeno físico. Un aspecto interesante de la validez científica lo describe el *teorema de Buckingham pi*, para el que una consecuencia del teorema es que las conclusiones sobre cuestiones cuantitativas, si se responden mediante procedimientos basados en pruebas matemáticas y lógicas, serán invariables en torno a la unidad de medida. En otras palabras, las conclusiones válidas serán exactamente las mismas independientemente de la unidad de medida. Por ejemplo, utilizando el sistema de medida imperial, familiar para los adultos en EE.UU., un niño que mide 39,37 pulgadas es exactamente la mitad de alto que un adulto que mide 78,74 pulgadas. Utilizando las medidas métricas, comunes en la mayoría de los países, ese mismo niño mide 100 cm y el adulto de 200 cm sigue midiendo exactamente el doble que el niño. Otro ejemplo se refiere a las unidades comunes de temperatura, para las que una

temperatura diaria de 15 grados centígrados no indica exactamente la mitad de la actividad termodinámica (calor) de 30 grados centígrados. Utilizando la escala de temperatura Fahrenheit, esas mismas temperaturas son 59 grados y 86 grados. Las inferencias matemáticas que utilizan las escalas Celsius y Fahrenheit no son válidas. (Esta es la razón por la que los físicos utilizan la escala Kelvin para modelar matemáticamente la transferencia de energía termodinámica). Es importante que todos los expertos que utilicen pruebas científicas conozcan los principios básicos de medición y los conceptos básicos de probabilidades y resultados de las pruebas.

Las pruebas científicas, al ser fundamentalmente probabilísticas, no se espera que sean infalibles o deterministas. Cualquiera que esté interesado en afirmar algún nivel de pericia en las ciencias sociales -especialmente cuando intente explicar el significado real de una prueba o experimento científico- se verá obligado a aprender a entender algo de estadística y probabilidad. (Curiosamente, los científicos también han descubierto, a nivel atómico y subatómico, que intentar comprender algunos fenómenos físicos también depende de la comprensión de la teoría de la probabilidad). Desde un punto de vista más práctico, un indicio fácilmente observable cuando se trata de pseudociencia es la tendencia a afirmar con excesiva confianza que las propias conclusiones son prácticamente infalibles, lo que sirve para intentar causar una poderosa impresión social / psicológica, y también alivia

la pesada tarea de cuantificar realmente el nivel de confianza y el margen de incertidumbre asociado a una conclusión. La ciencia real suele estar plagada de incertidumbre. Se espera que tanto los científicos teóricos como los aplicados y los profesionales expertos sepan que su tarea consiste en intentar encontrar formas de reducir esa incertidumbre, y que sepan que esto empieza por cuantificar realmente la incertidumbre. Desde un punto de vista práctico, comprender en cierta medida la probabilidad y los resultados probabilísticos de las pruebas permitirá un debate más útil y realista sobre el valor real del resultado de una prueba.

En torno a nuestras conclusiones y conocimientos científicos han surgido una serie de conceptos útiles, como las discusiones sobre el margen de error o incertidumbre, el grado de fuerza, el intervalo de confianza, los intervalos creíbles, las probabilidades, los cocientes de probabilidad, las probabilidades condicionales, etc. La forma más habitual de enmarcar nuestros debates sobre el grado de certeza asociado a las pruebas científicas y los experimentos científicos es hablar de precisión, que, en realidad, puede significar varias cosas distintas en función del problema práctico de interés. A la hora de hablar de la precisión de una prueba científica, resulta útil disponer de un vocabulario conceptual coherente y bien organizado que resulte fácilmente reconocible para otros profesionales. En la figura 1 se muestra un gráfico que ilustra las relaciones entre una serie de términos conceptuales útiles para comprender la precisión de las pruebas.



Figura 1. Términos conceptuales para comprender la precisión.

		Estado de Criterio Externo (Realidad)		
		Estado Positivo	Estado Negativo	
Prueba o Resultado Experimental	Resultado Positivo	VP (Verdadero Positivo)	FP (Falso Positivo) Error Tipo 1	Valor Predictivo Positivo (VPP) = VP / Todos los Resultados Positivos
				Indice de Falso Positivo (IFP) = FP / Todos los Resultados Positivos
	Resultado Negativo	Fn (Falso Negativo) Error Tipo 2	VN (Verdadero Negativo)	Valor Predictivo Negativo (VPN) = VN / todos los Resultados Negativos
				Indice de Falso Negativo (IFN) = FN / Todos los Resultados Negativos
		Prueba de Sensibilidad = VP / todos los casos de Estado Positivo	Prueba de Especificidad = VN / todos los casos de Estado Negativo	Exactitud

Estado de criterio (criterio externo): es cualquier fenómeno de interés que deseamos cuantificar, clasificar o predecir. A veces se denomina *parámetro desconocido* o *fenómeno de interés* desconocido.

Resultado de la prueba: suele referirse al resultado categórico de la prueba. Los resultados de las pruebas científicas no se discuten en términos de pasó o falló -aunque en algunos contextos puede haber una tendencia a las interpretaciones prácticas a este nivel-, sino que se utilizan los términos *positivo* y *negativo* para significar si la probabilidad es suficiente para apoyar una conclusión categórica sobre la presencia o ausencia del parámetro desconocido.

Resultado positivo: es un término que significa cuando el resultado de una prueba probabilística apoya una conclusión categórica de que el parámetro desconocido o el fenómeno desconocido de interés está *presente* en un caso.

Resultado negativo: significa cuando el resultado de una prueba probabilística apoya una conclusión categórica de que el parámetro desconocido o el fenómeno desconocido de interés está *ausente* en un caso

Estado positivo: se refiere a si el parámetro desconocido o el fenómeno desconocido de interés está realmente *presente* (en la realidad) para un caso.

Estado negativo: se refiere a si el parámetro desconocido o el fenómeno desconocido de interés está realmente *ausente* (en la realidad) para un caso.

Verdadero positivo (VP): describe un resultado positivo que coincide con un estado positivo. Una prueba ha identificado correctamente la presencia del fenómeno desconocido de interés para un caso.

Verdadero negativo (VN): describe un resultado negativo que coincide con un estado negativo. Una prueba ha identificado correctamente la ausencia del fenómeno desconocido de interés para un caso.

Falso positivo (FP): describe un resultado positivo que coincide con un estado negativo. Una prueba ha identificado incorrectamente la presencia del fenómeno desconocido en un caso para el que el estado del criterio externo (realidad) es en realidad negativo (el fenómeno desconocido de interés está en realidad ausente). El FP se calcula a veces como 1-especificidad. Sin embargo, el cálculo de la 1-especificidad será incorrecto para las pruebas que incluyen el uso de clasificaciones no concluyentes (cuando un resultado probabilístico no es estadísticamente significativo ni para la clasificación positiva ni para la negativa).

Falso negativo (FN): describe un resultado negativo que coincide con un estado positivo. Una prueba ha identificado incorrectamente la ausencia del fenómeno desconocido en un caso para el cual el estado del criterio externo (realidad) es realmente positivo (el fenómeno desconocido de interés está realmente presente). FN a veces se calcula como 1 - sensibilidad. Sin embargo, el cálculo de 1 - sensibilidad será incorrecto para las pruebas que incluyen el uso de clasificaciones no concluyentes (cuando el resultado de una prueba no es ni positivo ni negativo).

Valor predictivo positivo (VPP): se refiere a la proporción de resultados TP a resultados (TP + FP) para un grupo de casos para los cuales se conoce el estado positivo o negativo real. (Los científicos de datos a veces se refieren a estos como casos etiquetados). Puede ser útil para estimar la probabilidad de que un resultado

positivo sea correcto para un caso desconocido. Sin embargo, el VPP no es resistente a las diferencias en la proporción anterior de casos en estado positivo y en estado negativo (tasa base o tasa de incidencia). Es decir, el VPP –la probabilidad de que un resultado positivo sea correcto– estará en función de la proporción previa de casos de estado positivo y de estado negativo, además de estar influenciada por la sensibilidad de la prueba y las tasas de FP.

Valor predictivo negativo (VPN): se refiere a la proporción de resultados VN con respecto a los resultados (VN + FN) para un grupo de casos etiquetados (es decir, casos para los que se conoce el estado positivo o negativo real). Puede ser útil para estimar la probabilidad de que un resultado negativo sea correcto para un caso desconocido. El VPN tampoco es resistente a las diferencias en la proporción previa de casos de estado positivo y estado negativo (tasa base o tasa de incidencia). El VPN –la probabilidad de que un resultado negativo sea correcto– variará tanto con la tasa de especificidad de la prueba como con la proporción de casos de estado positivo (es decir, probabilidad previa, tasa base o tasa de incidencia) en un grupo de casos.

Índice de falsos positivos (IFP): se refiere a la proporción de resultados FP con respecto a todos los resultados positivos (FP + VP) de un grupo de casos identificados. Puede ser útil para estimar la probabilidad de que un resultado positivo sea incorrecto para un caso desconocido. El IFP no es resistente al desequilibrio de grupo (es decir, a las diferencias en la proporción de casos de estado positivo y de estado negativo), y variará tanto con la tasa de base previa o tasa de incidencia como con la tasa de sensibilidad de la prueba.



Índice de falsos negativos (IFN): se refiere a la proporción de resultados FN con respecto a todos los resultados negativos (FN + VN) de un grupo de casos identificados. Puede ser útil para estimar la probabilidad de que un resultado negativo sea incorrecto para un caso desconocido. El IFN no es resistente al desequilibrio de grupo. El IFN variará tanto con la tasa de especificidad de la prueba como con las diferencias en la proporción de casos de estado negativo y de estado positivo (tasa de base previa o tasa de incidencia).

Sensibilidad: se refiere a la proporción de un grupo de casos de estado positivos para los que una prueba puede identificar correctamente la presencia del fenómeno

desconocido de interés. La sensibilidad de la prueba, al calcularse sólo dentro del subgrupo de casos de estado positivos, es resistente a las diferencias en la tasa de incidencia previa. Es decir, la sensibilidad de la prueba será invariable al desequilibrio del grupo.

Especificidad: se refiere a la proporción de un grupo de casos de estado negativo para los que una prueba puede identificar correctamente la ausencia del fenómeno desconocido de interés. La especificidad de la prueba se calcula dentro del subgrupo de casos de estado negativo y es resistente a las diferencias desequilibrio del grupo (probabilidad previa, tasa base o tasa de incidencia).

