Cinco Minutos de Lecciones de Ciencia: Estrategias de Evaluación Múltiple en Dos Contextos Distintos (SARS-CoV-2 y Polígrafo)

Raymond Nelson

El propósito básico de cualquier prueba científica es cuantificar, clasificar o predecir un fenómeno de interés, a veces denominado *parámetro desconocido,* que no puede ser sujeto a una observación determinista perfecta o a una medición física directa. El procedimiento de prueba básico de cualquier prueba consiste en obtener algunos datos, a menudo denominados *muestra*, que se pueden utilizar para calcular un *clasificador estadístico* – utilizando algún tipo de función de probabilidad estadística (datos de referencia o distribución de referencia) junto con un proceso estructurado o regla para clasificar como positivo o negativo al resultado de la prueba o experimento científico. Por ejemplo: la regla básica de decisión en la tradición de la estadística de frecuencia es la siguiente: $p < \alpha = sig$. Las muestras de prueba pueden encontrarse en forma de *muestra física,* como cuando se obtienen a través de un hisopo nasofaríngeo o sangre extraída por un técnico médico flebotomista en el caso de pruebas médicas, y también pueden ser en forma de ensayos de *estímulo y respuesta* registrados en pruebas sociales/ conductuales como ocurre en una prueba de polígrafo.

Independientemente del tipo de prueba, los datos de muestra no son en sí mismos el parámetro desconocido o fenómeno de interés, sino que son una *proxy* que está correlacionada con los fenómenos de interés en un grado suficiente que puede ser útil para hacer inferencias estadísticas acerca del parámetro desconocido de interés. Mientras que las pruebas médicas/epidemiológicas, que utiliza muestras físicas, pueden hacer uso de una sola fuente de datos, las pruebas sociales/de comportamiento – incluyendo las pruebas psicológicas y las mediciones de riesgo actuarial - a menudo harán uso de fuentes múltiples de información de las cuales se puedan extraer y combinar las características de respuesta.

Todas las pruebas científicas son fundamentalmente probabilísticas y por esta razón no se espera que sean infalibles – se espera que cuantifiquen la fuerza probabilística o el margen de incertidumbre asociado con el resultado o conclusión de la prueba. Cuando se

Traductor. rodolfo@poligrafia.com.mx

utiliza una prueba para cuantificar, vs clasificar un parámetro desconocido, la información estadística intentará describir la probabilidad estadística de que el valor numérico del parámetro desconocido se encuentre dentro de un cierto rango. Muchas pruebas solamente intentan clasificar (la predicción se puede considerar como una forma de clasificación). Idealmente, aunque no siempre, un clasificador estadístico también proporcionará información acerca de la fuerza práctica de la información o conclusión, o de su margen de incertidumbre.

El clasificador estadístico de algunas pruebas científicas se abstrae del contexto práctico hasta el punto en que, aunque se puede utilizar para clasificar un resultado de prueba, puede que no se encuentre una relación conveniente o intuitiva entre la estadística de la prueba y las consideraciones prácticas tales como la posibilidad real de que los resultados de la prueba sean correctos o incorrectos - a menudo referidos como VP o *sensibilidad* y VN o *especificidad* y tasas FP y FN. Los valores P – utilizados para estimar el error de medición aleatorio – son un ejemplo de esto; se pueden utilizar para clasificar los resultados de una prueba científica o experimento como *estadísticamente significativos* o *no significativos* de acuerdo con un nivel de tolerancia alfa, pero no proporcionan información acerca de las posibilidades prácticas que se asocian con esa clasificación. Con mayor frecuencia los resultados prácticos se describen empíricamente por las tasas de sensibilidad, especificidad y tasas de error de FP o FN que se observan ante determinados umbrales numéricos o de decisión estadística. Se pueden lograr resultados aún más prácticos utilizando métodos Bayesianos o *a-posteriori* que tomen en cuenta tanto una estadística de prueba como la información previa.

Independientemente de si una prueba es médica/epidemiológica o social/conductual/ psicológica o actuarial, los conceptos básicos de las pruebas científicas son similares. También son similares los tipos de preguntas y estrategias que los desarrolladores de pruebas científicas tendrán en cuenta al validar un método de prueba. Otra similitud es que las pruebas son costosas, en términos de costos financieros, actividad humana y tiempo. Cuando es necesario realizar un gran volumen de pruebas, la eficiencia - de tiempo, recursos físicos y de actividad humana – y la necesidad de maximizar los recursos disponibles puede convertirse en una consideración importante. Por ejemplo, ¿cómo *evaluar a un gran grupo de aplicantes de seguridad pública con respecto a su historial por participación en posibles problemas de comportamiento múltiples que los hagan no-idóneos para puestos de confianza pública? O, ¿cómo evaluar a la población de una gran ciudad por SARS-CoV-2 en un intento de aislar y contener la propagación de la enfermedad?*

La evaluación múltiple es una estrategia común que se puede utilizar para incrementar la eficiencia de las pruebas. El uso de estrategias de evaluación múltiple se puede observar en pruebas en diferentes contextos, incluyendo en pruebas de polígrafo asuntos múltiples, y también en forma de pruebas *agrupadas* para SARS-CoV-2, el nuevo coronavirus responsable de la pandemia COVID-19. La evaluación múltiple, en este uso,

se refiere a la evaluación de objetivos múltiples en un solo análisis. En el contexto SARS-CoV-2, las estrategias de evaluación múltiple se conocen como pruebas *agrupadas,* en las que varias muestras se agrupan para su análisis. Aunque otros países ya han hecho uso de estrategias de pruebas agrupadas, en los Estados Unidos la FDA y la CDC han emitido recientemente una guía para el desarrollo y validación de estos métodos para pruebas de diagnóstico y de exploración necesarias por SARS- CoV-2 y COVID-19.

De acuerdo con el sitio web de CDC:

> Las pruebas diagnósticas para el SARS-CoV-2 están destinadas a identificar la ocurrencia a nivel individual cuando hay una razón para sospechar que una persona podría estar infectada por tener síntomas o por sospecha de exposición reciente, o para determinar la resolución de la infección. Algunos ejemplos de pruebas diagnósticas incluyen la evaluación de individuos sintomáticos que se presentan con su proveedor de atención médica, la prueba a individuos mediante esfuerzos de rastreo de contactos, la prueba a individuos que indican que fueron expuestos a alguien que es un caso confirmado o con sospecha de la enfermedad coronavirus 2019 (COVID-19), y la prueba de individuos presentes en un evento donde más tarde se confirmó que un asistente tenía COVID-19.

El sitio web de CDC también proporciona información para diferenciar las pruebas diagnósticas de las pruebas exploratorias:

> Las pruebas de detección para el SARS-CoV-2 intentan identificar la ocurrencia a nivel individual, incluso si no hay razón para sospechar de infección - por ejemplo, cuando no hay exposición conocida. Esto incluye, pero no se limita, a la exploración en individuos no sintomáticos sin exposición conocida con la intención de tomar decisiones basadas en los resultados de la prueba. Las pruebas exploratorias están destinadas a identificar a los individuos infectados sin, o antes del desarrollo de síntomas que pueden ser contagiosos, para que se puedan tomar medidas de prevención de transmisión posterior. Los ejemplos de exploración incluyen planes de evaluación desarrollados en un lugar de trabajo para evaluar a sus empleados, y planes de pruebas desarrollados por una escuela para evaluar a sus estudiantes, profesores y personal. En ambos ejemplos, la intención es utilizar los resultados de las pruebas exploratorias para determinar quién puede regresar y las medidas de protección que serán tomadas.

El concepto general de las pruebas de diagnóstico y exploratorias es esencialmente idéntico al que se describe en los estándares de Práctica APA.

> 1.1.5 Examen diagnóstico: Un examen de poligráfico de evento específico evidenciario o de investigación que se realiza para ayudar a determinar la veracidad de un examinado con respecto a su conocimiento o participación en un

asunto o alegato reportado. Los exámenes diagnósticos pueden abordar un solo aspecto o hechos múltiples de un evento.

1.1.6 Examen exploratorio: Un examen poligráfico realizado en ausencia de un asunto o alegato reportado. Los exámenes exploratorios se pueden conducir como exámenes un solo asunto o de asuntos múltiples.

Lo importante aquí es que las pruebas diagnósticas se lleven a cabo en respuesta a un problema conocido - un incidente o alegato en el contexto poligráfico, y los síntomas de enfermedad o la exposición en medicina y epidemiología. Para muchos, un error tentador y sencillo sería confundir las dos dicotomías: *diagnóstico vs exploratorio* y estrategias de pruebas *individuales vs múltiples*. Los profesionales administrativos y de campo que entiendan correctamente estas diferencias serán más aptos para desarrollar e implementar estrategias y políticas de evaluación que logren sus objetivos.

En el contexto poligráfico, las estrategias de evaluación múltiple se conocen comúnmente como pruebas de asuntos múltiples, y a veces como pruebas de facetas múltiples - siendo la única diferencia si un polígrafo intenta fines de diagnóstico o exploratorio. Para los polígrafos de asuntos múltiples, los estímulos de prueba se evalúan asumiendo una variancia de criterio independiente. Por ejemplo: los asuntos objetivo de polígrafo para la exploración de aplicantes para ser empleados en seguridad pública pueden incluir, su historial de comportamiento con drogas ilegales, la comisión de delitos graves no reportados, abuso de pareja doméstica o íntimas, agresión sexual y delitos de odio o intolerancia social. Es concebible que una persona no se haya involucrado en ninguno, alguno o en todos estos tipos de comportamientos.

Una estrategia de evaluación múltiple tiene la ventaja de aumentar la sensibilidad de la prueba de polígrafo exploratoria ante una gama más amplia de comportamientos de preocupación y considera un uso más eficiente del tiempo y de otros recursos, en lugar de intentar investigar estos distintos comportamientos mediante exámenes independientes. Una desventaja del polígrafo de asuntos múltiples es la reducción potencial de especificidad y de precisión. La heurística para la clasificación de los resultados poligráficos de asunto múltiple es *cualquiera-o-todas*, donde un resultado de prueba se clasifica como positivo si alguna pregunta objetivo produjo un resultado positivo, y se clasifica como negativo si todas las preguntas de la prueba produjeron resultados negativos. (También tome en cuenta que no existe una ventaja empírica conocida de una serie de exámenes de asunto único en comparación con un examen de asunto múltiple. En la medida en que los errores de prueba están en función del error de medición aleatoria, una serie de pruebas de un asunto único pueden estar sujetas a los efectos de la multiplicidad de manera similar a los exámenes de asuntos múltiples.)

Los resultados positivos en un polígrafo de asuntos múltiples podrían o no indicar el área exacta del comportamiento problemático, y por esta razón la consecuencia podría

ser tener que realizar pruebas adicionales al candidato - dependiendo del tamaño del grupo de candidatos, nivel de interés en el individuo, recursos, riesgos y otros factores. También es posible que un candidato pueda simplemente ser ajustado, reducido o eliminado, tomando en cuenta la prioridad o la jerarquía de los candidatos disponibles. Los estándares de práctica de campo del polígrafo basados en evidencia no permiten que los examinadores obtengan clasificaciones positivas y negativas dentro del mismo examen - incluso cuando las preguntas del examen se desarrollaron bajo el supuesto de varianza de criterio independiente, porque hacerlo dañaría la precisión de la prueba (potencialmente crea un contexto para resultados FP y FN en el mismo examen).

El sitio web de la FDA proporciona orientación adicional para los desarrolladores de métodos de evaluación agrupada o múltiple para SARS-CoV-2, con descripción de dos métodos diferentes de combinación de muestras para pruebas múltiples (equivalencia o agrupación de media parcial y agrupación de muestra media):

> En general, la FDA recomienda validar su prueba con cualquiera de los enfoques de agrupamiento de manera que se preserve la sensibilidad de su prueba tanto como le sea posible; es decir, cuando se evalúa individualmente es preferible utilizar un enfoque donde todos los especímenes identificados como positivos, también sean identificados como positivos cuando se haga la evaluación utilizando el enfoque de pruebas agrupadas. Sin embargo, es probable que se reduzca el rendimiento con las estrategias de agrupamiento, debido a la dilución de la muestra clínica primaria. Como se explica en las plantillas, ya que la agrupación de muestras incrementará en gran medida el número de individuos que se pueden evaluar utilizando los recursos existentes, es aceptable tener una pequeña reducción en la sensibilidad dependiendo de la eficiencia del agrupamiento y de otros atenuantes del entorno. Por lo tanto, en general la FDA recomienda que después del agrupamiento, el rendimiento de la prueba incluya un acuerdo positivo del ≥ 85% (PPA) cuando se compare con la misma prueba realizada en muestras individuales. Se pueden recomendar limitaciones adicionales, como considerar que los resultados negativos de las muestras agrupadas son presumiblemente negativos, en función de la población de pacientes incluida en su evaluación clínica y de los datos de rendimiento presentados en su solicitud de EUA [autorización de uso de emergencia – por sus siglas en inglés].

El párrafo anterior es instructivo por varias razones. En primer lugar, reconoce que las estrategias de evaluación múltiple a veces pueden conducir a una reducción en la sensibilidad de la prueba, y que se debe tener cuidado para evitar esto. En el contexto SARS-CoV-2 la sensibilidad de la prueba – la capacidad de la prueba para detectar o identificar cuando están presentes los fenómenos desconocidos de interés – es la métrica de interés primario. En otros contextos, es posible que se prioricen otras métricas; como la especificidad de la prueba - la capacidad de una prueba para determinar correctamente la ausencia del problema de interés. Las pruebas agrupadas de muestras

SARS-CoV-2 difieren de alguna manera del ejemplo del polígrafo (que involucra asuntos objetivo de comportamiento múltiples en un solo individuo) en el que se involucra la toma de una porción de datos de muestra de múltiples individuos y lleva a cabo el análisis en una sola muestra combinada. La clasificación de los resultados de las pruebas agrupadas de SARS-CoV-2 es similar a la clasificación de los resultados de un polígrafo de asuntos múltiples en el sentido de que toda la muestra agrupada se clasificará como positiva o negativa – no es posible lograr resultados positivos y negativos dentro de un solo análisis.

Es importante destacar que este párrafo (FDA) se refiere a que la selección de una estrategia de prueba es inherentemente probabilística y que siempre se hace teniendo en cuenta la capacidad de entender e interpretar el significado científico/probabilístico y práctico de la información obtenida del resultado de la prueba. En este ejemplo (SARS-CoV-2), existe un objetivo claramente declarado de que la concordancia - denominada *acuerdo positivo porcentual* (PPA), está destinado a alcanzar un umbral del 85% al comparar los resultados entre los enfoques de las pruebas múltiples y únicas.

Un riesgo potencial de pruebas agrupadas de muestras SARS-CoV-2 es que la combinación de muestras puede inducir una reducción en la sensibilidad de la prueba en función de la dilución de la señal de interés en las muestras combinadas. Por ejemplo, imagine una muestra combinada de 20 personas de las cuales sólo 1 es positiva. Sería posible diluir la carga viral al grado que caiga por debajo de la puntuación de corte o umbral de la clasificación positiva. Esta necesidad de entender la tasa de la sensibilidad de la prueba es la razón por la que las estrategias de pruebas agrupadas deban estar sujetas a validación, además de la validación misma del método de análisis. La aprobación de una estrategia de prueba agrupada o múltiple solo debe considerarse cuando se pueden mantener métricas aceptables de precisión de prueba.

Un ejemplo de esto en el contexto poligráfico es que los cortes numéricos para la clasificación positiva de los subtotales del ESS-M se calculan sin corrección estadística para las pruebas de asunto múltiple. El uso de una corrección estadística para estas clasificaciones reduciría la incidencia de errores FP, pero lo haría a un costo de reducción de la sensibilidad de la prueba y en un aumento de los errores FN. Para los polígrafos de asunto único, para los que la precisión global es a menudo el objetivo previsto, se espera que la mayoría de las clasificaciones se realicen utilizando la puntuación total, y por esta razón el cálculo de las puntuaciones numéricas de corte del ESS-M para la clasificación positiva incluye una corrección estadística para los subtotales. Para los exámenes de asunto único existe una pérdida de sensibilidad debido a la dependencia de la puntuación total. Para estos exámenes, el uso de subtotales con corrección estadística en la realidad puede incrementar la sensibilidad de la prueba sin el aumento correspondiente en los errores FP.

Otro aspecto interesante de la información del sitio web de CDC es que también hay información disponible en las pruebas de vigilancia:

> La vigilancia del SARS-CoV-2 incluye actividades sistemáticas continuas, que incluyen la recopilación, análisis y la interpretación de datos relacionados con la salud que son esenciales para planificar, implementar y evaluar las prácticas de salud pública. Las pruebas de vigilancia se utilizan generalmente para monitorear una ocurrencia a nivel comunitario o poblacional, como un brote de enfermedad infecciosa, o para caracterizar la ocurrencia una vez detectada, así como examinar la incidencia y prevalencia de la ocurrencia. Las pruebas de vigilancia se utilizan para obtener información a nivel de población, en lugar del nivel individual, y los resultados de las pruebas de vigilancia se pueden devolver en conjunto a la institución que lo requiere. Las pruebas de vigilancia pueden muestrear un cierto porcentaje de una población específica para monitorear el aumento o la disminución de la prevalencia y para determinar el efecto de las intervenciones comunitarias en la población, así como del distanciamiento social. Un ejemplo de pruebas de vigilancia es un plan desarrollado por un departamento de salud pública estatal para seleccionar y tomar muestras aleatorias de un porcentaje de todas las personas de una ciudad de forma continua para evaluar las tasas y tendencias locales de infección.

En este contexto, el término *vigilancia* se refiere a obtener y analizar la información a nivel del grupo o población – sin intentar monitorear o diagnosticar individuos – con el propósito de entender las tasas de incidencia de la enfermedad (también conocidas como tasas base o previas en el análisis Bayesiano). Este uso puede considerarse como una forma de exploración, y no como una forma de pruebas de diagnóstico. El objetivo de las pruebas de vigilancia del SARS-CoV-2 parece ser la obtención de información acerca de la prevalencia de la enfermedad, que puede utilizarse para optimizar la estrategia de evaluación múltiple, y que también puede utilizarse como información previa para calcular las posibilidades Bayesianas posteriores de resultados correctos o incorrectos.

Las pruebas agrupadas fueron descritas por primera vez por el economista Robert Dorman en un artículo de 1943 en los *Anales de Estadísticas Matemáticas,* titulado "Detección de Miembros Defectuosos en Grandes Poblaciones" (un título atroz para la sensibilidad social de hoyen día). El contexto de esa publicación fue el Servicio de Salud Pública de los Estados Unidos y el Sistema de Servicio Selectivo, y la detección de Sífilis en hombres cuando eran canalizados al ejército de los Estados Unidos durante la Segunda Guerra Mundial. En términos económicos, el número de muestras óptimo que deben agruparse para su análisis puede calcularse matemáticamente en función de varios factores, incluyendo la tasa de incidencia cuando se conoce (y si no, la tasa conocida de resultados positivos en las pruebas), el costo de la prueba y el ahorro esperado en costos como resultado de grupos analíticos negativos. En términos prácticos, el costo de la evaluación incluye los costos sociales y económicos asociados con una pandemia no contenida.

En resumen, una ventaja obvia de la evaluación múltiple es que puede reducir

sustancialmente los gastos asociados con los suministros limitados de prueba como hisopos, reactivos y equipos de prueba, así como la demanda de tiempo y carga de trabajo que se impone a los profesionales. Una desventaja potencial de la evaluación múltiple es que pueden restringir los tipos de conclusiones que se pueden realizar. En la medida en que pueda proporcionar una precisión adecuada, en términos de tasas de sensibilidad o especificidad de prueba requerida, o de la capacidad para reducir los errores FP o FN a los niveles requeridos, las estrategias de evaluación múltiple pueden ser un método viable para hacer un uso máximo de los recursos disponibles, incluyendo los suministros materiales, equipos y esfuerzo humano. Como suele ser el caso, el uso exitoso de estas estrategias dependerá, en cierta medida, de los administradores de políticas, de los profesionales de campo y de un público que tenga cierto conocimiento o apreciación de las cuestiones que influyen en las pruebas científicas y en su uso.