Modelos Lineales Generalizados
Capítulo 11 - págs. 1-19

23 de abril de 2021

Este capítulo trata sobre los valores p en las pruebas de significancia de las hipótesis nulas (NHST). Los valores p son calculados considerando posibilidades imaginarias, cosas que podrían haber sucedido pero no sucedieron. En NHST, el objetivo de inferencia es decidir si un valor particular de un parámetro puede ser rechazado.

Por ejemplo, podríamos querer saber si una moneda es justa, lo que en NHST se convierte en la cuestión de si podemos rechazar la hipótesis nula de que el sesgo de la moneda tiene el valor específico $\theta = 0.5$. La lógica del NHST convencional consiste en calcular las probabilidades exactas de todos los resultados posibles, que a su vez se pueden usar para calcular la probabilidad de obtener un resultado tan extremo (o más extremo que) el resultado realmente observado. Esta probabilidad, de obtener un resultado de la hipótesis nula que sea tan extremo (o más extremo que) el resultado real, se denomina $p-$value. Si el valor p es muy pequeño, digamos menos del 5 %, entonces decidimos rechazar la hipótesis nula.

Una definición más explícita de un valor p es la probabilidad de obtener un resultado de muestra de la población hipotetizada que sea tan extremo o más extremo que el resultado real cuando se utilizan los procedimientos de muestreo y prueba previstos. Es decir, el valor p es la probabilidad de que los posibles resultados cumplan o superen el resultado real. Existen muchos valores p posibles para cualquier conjunto de datos, dependiendo de cómo se genere la nube de resultados imaginarios. La nube depende no solo del criterio de detención previsto, sino también de las pruebas previstas que el analista desea realizar porque la intención de realizar pruebas adicionales expande la nube con posibilidades adicionales de las pruebas adicionales. Además, algo importante es que no se deben sesgar los datos obtenidos.

11.1.1. Definición del valor p

La hipótesis nula en NHST comienza con una función de verosimilitud y un valor de parámetro específico que describe las probabilidades de los resultados para las observaciones individuales. Se trata de la misma función de verosimilitud que en el análisis bayesiano. En el caso de una moneda, la función de verosimilitud es la distribución Bernoulli, con su parámetro θ que describe la probabilidad de obtener el resultado cara en un solo lanzamiento. Normal-

mente, el valor nulo de θ es 0.5, como cuando se comprueba si una moneda es justa, pero el valor hipotético de θ podría ser diferente.

La función de verosimilitud define la probabilidad de una sola medición, y el proceso de muestreo previsto determina la nube de posibles resultados de la muestra. La hipótesis nula es la función de verosimilitud con su valor específico para el parámetro θ, y la nube de muestras posibles está definida por las intenciones de parar y prrobar, denotadas por I.

Cada muestra imaginaria generada a partir de la hipótesis nula se resume con un estadístico descriptivo, denotado $D_{,I}$. En el caso de una muestra de lanzamientos de monedas, el estadístico descriptivo es z/N, es decir, la proporción de caras en la muestra. Ahora, imaginemos que generamos infinitas muestras a partir de la hipótesis nula utilizando la intención de parar y probar I; esto crea una nube de posibles valores resumen $D_{,I}$, cada uno de los cuales tiene una probabilidad particular. La distribución de probabilidad sobre la nube de posibilidades es la distribución de muestreo, se denota como:

$$p(D_{\theta,I}|\theta, I).$$

Para calcular el valor p, queremos saber qué parte de esa nube es tan extrema o más extrema que el resultado realmente observado. Para definir la extremidad debemos determinar el valor típico de $D_{\theta,I}$, que suele definirse como el valor esperado, $E[D_{\theta,I}]$. Este valor típico es el centro de la nube de posibilidades. Un resultado es más extremo cuando se aleja de la tendencia central. El valor p del resultado real es la probabilidad de obtener un resultado hipotético igual o más extremo. Formalmente, podemos expresarlo como:

$$p \text{ value} = p(D_{\theta,I} \succcurlyeq D_{real}|\theta, I),$$

donde $\succcurlyeq$ en este contexto significa tan extremo como o más extremo que, en relación con el valor esperado de la hipótesis. Para el caso de los lanzamientos de monedas, en el que el estadístico resumen de la muestra es z/N, el valor p se convierte en:

$$p(\text{cola derecha}) = p((z/N)_{\theta,I} \geq (z/N)_{real}|\theta, I)$$
$$p(\text{cola izquierda}) = p((z/N)_{\theta,I} \leq (z/N)_{real}|\theta, I)$$

Estos valores p se denominan de una cola porque indican la probabilidad de resultados hipotéticos más extremos que el resultado real en una sola dirección. Normalmente nos interesa la cola derecha cuando $(z/N)_{real}$ es mayor que $E[(z/N)_{\theta,I}]$, y nos interesa la cola izquierda cuando $(z/N)_{real}$ es menor que $E[(z/N)_{\theta,I}]$. El valor p de dos colas puede definirse de varias maneras, pero para nuestros propósitos definiremos el valor p de dos colas simplemente como dos veces el valor p de una cola. Para calcular el valor de p para cualquier situación específica, tenemos que definir el espacio de resultados posibles.

11.1.2. Con la intención de fijar N

En lugar de determinar la probabilidad de obtener exactamente el resultado z/N de la hipótesis nula, determinamos la probabilidad de obtener z/N o un resultado aún más extremo

que el esperado de la hipótesis nula. La razón para considerar resultados más extremos es la siguiente:

- Si rechazamos la hipótesis nula porque el resultado z/N está demasiado lejos de lo que esperaríamos, entonces cualquier otro resultado que tenga un valor aún más extremo también nos haría rechazar la hipótesis nula. Por lo tanto, queremos conocer la probabilidad de obtener el resultado real o un resultado más extremo en relación con lo que esperamos. Esta probabilidad total se denomina valor p. El valor p definido en este punto es el valor p de una cola, porque suma las probabilidades extremas en una sola cola de la distribución muestral. En la práctica, el valor p de una cola se multiplica por 2, para obtener el valor p de dos colas. Consideramos ambas colas de la distribución de muestreo porque la hipótesis nula podría rechazarse si el resultado fuera demasiado extremo en cualquier dirección. Si este valor p es inferior a una cantidad crítica, entonces rechazamos la hipótesis nula.

La probabilidad crítica de dos colas se fija convencionalmente en el 5 %. En otras palabras, rechazaremos la hipótesis nula siempre que la probabilidad total de la z/N observada o de un resultado más extremo sea inferior al 5 %. Observamos que esta regla de decisión nos hará rechazar la hipótesis nula el 5 % de las veces cuando la hipótesis nula sea verdadera, porque la propia hipótesis nula genera esos valores extremos el 5 % de las veces, sólo por azar. La probabilidad crítica, 5 %, es la proporción de falsas alarmas que estamos dispuestos a tolerar en nuestro proceso de decisión. Si consideramos una sola cola de la distribución, la probabilidad crítica es la mitad del 5 %, es decir, el 2.5 %. Por ejemplo, si la probabilidad de una cola es $p = 0.032$, entonces es inferior al 2.5 %, por lo que, no rechazamos la hipótesis nula de que $\theta = 0.5$. En el análisis de NHST, diríamos que el resultado no ha alcanzado la significancia. Pero, esto no quiere decir que aceptemos la hipótesis nula; simplemente suspendemos el juicio sobre el rechazo de esta hipótesis concreta. Con ello, observamos que no hemos determinado ningún grado de creencia en la hipótesis de que $\theta = 0.5$. La hipótesis puede ser verdadera o falsa.

11.1.3. Con la intención de fijar z

En el ejemplo del lanzamiento de una moneda, ¿cuál es la probabilidad de que al realizar N lanzamientos se obtenga z caras? Para responder a esta pregunta, consideramos lo siguiente: sabemos que el $N^{\underline{a}}$ lanzamiento es la $z^{\underline{a}}$ cara porque eso es lo que hizo que se detuviera el lanzamiento. Por tanto, los $N-1$ lanzamientos anteriores tuvieron $z-1$ caras en alguna secuencia aleatoria. La probabilidad de obtener $z-1$ caras en $N-1$ lanzamientos es $\binom{N-1}{z-1}\theta^{z-1}(1-\theta)^{N-z}$. La probabilidad de que el último lanzamiento salga cara es θ. Por lo tanto, la probabilidad de que se necesiten N lanzamientos para obtener z caras es:

$$p(N|z,\theta) = \binom{N-1}{z-1}\theta^{z-1}(1-\theta)^{N-z}\cdot\theta = \binom{N-1}{z-1}\theta^{z}(1-\theta)^{N-z} = \frac{z}{N}\binom{N}{z}\theta^{z}(1-\theta)^{N-z}.$$

(Esta distribución se llama a veces binomial negativa, pero ese término puede ser confuso, por lo que el autor decide no utilizarla aquí). Se trata de una distribución de muestreo, como

la distribución binomial porque presenta las probabilidades relativas de todos los posibles resultados de los datos para el valor fijo hipotético de θ y la regla de detenerse prevista.

El enfoque de esta sección ha sido que el valor p es diferente cuando la intención de detenerse es diferente. Sin embargo, si la recolección de datos se detiene cuando el número de caras (o colas) alcanza un umbral, entonces los datos están sesgados. Cualquier regla de detenerse que se active puede producir una muestra sesgada porque una secuencia accidental de datos extremos al azar hará que se detenga la recolección de datos y, por tanto, impedirá la posterior recolección de datos compensatorios que sean más representativos. Por lo tanto, se podría argumentar que parar en el umbral z no es una buena práctica porque sesga los datos; pero eso no cambia el hecho de que parar en el umbral z implica un valor p diferente al de parar en el umbral N. Muchos profesionales dejan de recolectar datos cuando se sobrepasa un extremo.

11.1.4. Con la intención de fijar la duración

En este apartado consideramos otra variante. Supongamos que, cuando le preguntamos a la asistente por qué ha dejado de lanzar la moneda, responde que ha dejado de hacerlo porque han transcurrido 2 minutos. En este caso, la recolección de datos no se ha detenido por haber alcanzado el umbral N, ni por haber alcanzado el umbral z, sino por haber alcanzado el umbral de duración. La clave para analizar este escenario es especificar cómo pueden surgir varias combinaciones de z y N cuando se realiza un muestreo de duración fija. No hay una especificación única y exclusivamente correcta, porque hay muchas restricciones diferentes en el mundo real sobre el muestreo a través del tiempo. Pero un enfoque es pensar en el tamaño de la muestra N como un valor aleatorio, es decir, si se repite el experimento de 2 minutos, a veces N será mayor y a veces menor. ¿Cuál es la distribución de N? Es la distribución de Poisson. La distribución de Poisson es un modelo de uso frecuente del número de ocurrencias de un evento en una duración fija. Es una distribución de probabilidad sobre valores enteros de N de 0 a $+\infty$. La distribución de Poisson tiene un único parámetro, λ, que controla su media (y también resulta ser su varianza). El parámetro λ puede tener cualquier valor real no negativo; no está restringido a números enteros.

11.1.5. Con la intención de realizar múltiples pruebas

En esta sección se menciona de cómo las intenciones de las pruebas afectan a la nube imaginaria de posibilidades que determinan el valor p. Supongamos que queremos probar la hipótesis de una moneda justa, y tenemos una segunda moneda en el mismo experimento que también estamos probando si es justa. En la investigación biológica real, por ejemplo, esto podría corresponder a la prueba de si la proporción macho/hembra de los bebés difiere del 50/50 en cada una de las dos especies de animales. Queremos controlar la tasa de falsas alarmas en general, así que tenemos que considerar la probabilidad de una falsa alarma de cualquiera de las dos monedas. Así, el valor p para la primera moneda es la probabilidad de que una proporción, igual o más extrema que su proporción real, pueda surgir por azar de

cualquiera de las dos monedas. Así, el valor p de la cola izquierda es:

$$p\big((z_1/N_1)_{\theta_1,I_1} \leq (z_1/N_1)_{\text{actual}} \text{ O } (z_2/N_2)_{\theta_2,I_2} \leq (z_1/N_1)_{\text{actual}} \mid \theta_1, \theta_2, I_1, I_2\big),$$

y el valor p de la cola derecha es:

$$p\big((z_1/N_1)_{\theta_1,I_1} \geq (z_1/N_1)_{\text{actual}} \text{ O } (z_2/N_2)_{\theta_2,I_2} \geq (z_1/N_1)_{\text{actual}} \mid \theta_1, \theta_2, I_1, I_2\big).$$

Dado que esta notación es un poco difícil de manejar, se utilizará la expresión:

$$\text{Extremo}\{z_1/N_1, z_2/N_2\},$$

para denotar la menor de las proporciones al calcular la cola baja, pero la mayor de las proporciones al calcular la cola alta. Entonces el valor p de la cola izquierda es:

$$p\big(\text{Extremo}\{(z_1/N_1)_{\theta_1,I_1}, (z_2/N_2)_{\theta_2,I_2}\} \leq (z_1/N_1)_{\text{actual}} \mid \theta_1, \theta_2, I_1, I_2\big),$$

y el valor p de la cola derecha es:

$$p\big(\text{Extremo}\{(z_1/N_1)_{\theta_1,I_1}, (z_2/N_2)_{\theta_2,I_2}\} \geq (z_1/N_1)_{\text{actual}} \mid \theta_1, \theta_2, I_1, I_2\big).$$

Supongamos que lanzamos ambas monedas $N_1 = N_2 = 24$ veces, y que la primera sale cara $z_1 = 7$ veces. Este es el mismo resultado que en los ejemplos anteriores. Supongamos que tenemos la intención de detenernos cuando el número de lanzamientos alcance esos límites. El valor p para el resultado de la primera moneda es la probabilidad de que z_1/N_1 o z_2/N_2 sean tan extremos o más extremos que $7/24$ cuando la hipótesis nula es verdadera. Esta probabilidad será mayor que cuando se considera una sola moneda, porque incluso si los lanzamientos imaginarios de la primera moneda no superan los $7/24$, sigue existiendo la posibilidad de que los lanzamientos imaginarios de la segunda moneda sí lo hagan. Para cada muestra imaginaria de lanzamientos de las dos monedas, tenemos que considerar la proporción de la muestra, ya sea z_1/N_1 o z_2/N_2, que sea más extrema en relación con θ. El valor p es la probabilidad de que la proporción extrema cumpla o supere la proporción real.

Observamos con ello que no necesitamos conocer el resultado de la segunda moneda (es decir, z_2) para calcular el valor p de la primera moneda. De hecho, no necesitamos lanzar la segunda moneda en absoluto. Todo lo que necesitamos para calcular el valor p de la primera moneda es la intención de lanzar la segunda moneda N_2 veces. La nube de posibilidades imaginarias está determinada por las intenciones de muestreo, no por los datos observados.

11.1.6. Examen de conciencia

Anscombe afirmó que en cualquier experimento o investigación de muestreo en el que el número de observaciones es una cantidad incierta pero no depende de las propias observaciones, siempre es legítimo tratar las observaciones en el análisis estadístico como si su número

se hubiera fijado de antemano. Es importante notar que Anscombe no dijo que es únicamente o sólo legítimo tratar a N como fijo de antemano. De hecho, si tratamos a z como fijo, o si tratamos a la duración como fija, seguimos teniendo distribuciones de probabilidad condicional perfectamente correctas que resultan en un 5 % de falsas alarmas a largo plazo cuando la hipótesis nula es verdadera. Los valores p para datos individuales pueden ser diferentes, pero en muchos conjuntos de datos la tasa de falsas alarmas a largo plazo es del 5 %. Obsérvese también que Anscombe llegó a la conclusión de que todo este problema se evitaría si hiciéramos análisis bayesiano. Anscombe dijo que todo el riesgo de error se evita si el método de análisis utiliza las observaciones sólo en la forma de su función de verosimilitud, ya que la función de verosimilitud (dadas las observaciones) es independiente de la regla de muestreo. Uno de estos métodos de análisis lo proporciona la teoría clásica de la creencia racional en la que una distribución de probabilidad posterior se deduce, por el teorema de Bayes, de la función de verosimilitud de las observaciones y una distribución de probabilidad a priori.

Por otra parte, en el contexto de la NHST, la solución consiste en establecer la verdadera intención del investigador. Este es el enfoque que se adopta explícitamente cuando se aplican correcciones para pruebas múltiples. El analista determina cuáles son las pruebas realmente previstas, y determina si esas intenciones de prueba fueron concebidas honestamente a priori o post hoc (es decir, motivadas sólo después de ver los datos), y luego calcula el valor p apropiado. Por desgracia, determinar las verdaderas intenciones puede ser difícil. Por lo tanto, tal vez los investigadores que utilizan los valores p para tomar decisiones deberían estar obligados a registrar públicamente su regla de detención y pruebas previstas, antes de la recolección de los datos. Pero, ¿qué ocurre si un acontecimiento imprevisto interrumpe la recolección de los datos, o produce una ganancia de datos extra? ¿Y si, una vez recolectados los datos, queda claro que debieron realizarse otras pruebas? En estas situaciones, los valores p deben ajustarse a pesar del registro previo. Fundamentalmente, las intenciones no deberían importar para la interpretación de los datos porque la propensión de la moneda a salir cara no depende de las intenciones del lanzador (excepto cuando la regla de parar sesga la recolección de los datos). De hecho, diseñamos cuidadosamente los experimentos para aislar las monedas de las intenciones del experimentador.

11.1.7. Análisis bayesiano

La interpretación bayesiana de los datos no depende de las intenciones encubiertas de muestreo y prueba del recolector de datos. En general, para los datos que son independientes a través de los ensayos (y no están influenciados por la intención de muestreo), la probabilidad del conjunto de datos es simplemente el producto de las probabilidades de los resultados individuales. La función de verosimilitud captura todo lo que suponemos que incide en los datos. En el caso de la moneda, suponemos que el sesgo (θ) de la moneda es la única influencia en su resultado, y que los lanzamientos son independientes. La función de verosimilitud de Bernoulli recoge completamente estas supuestos.

En resumen, el análisis y la conclusión del NHST dependen de las intenciones encubiertas del experimentador, porque esas intenciones determinan las probabilidades en el espacio de todos los datos posibles (no observados). Esta dependencia del análisis de las intenciones

del experimentador entra en conflicto con la suposición opuesta de que las intenciones del experimentador no tienen ningún efecto sobre los datos observados. El análisis bayesiano, en cambio, no depende de la nube imaginaria de posibilidades. El análisis bayesiano opera sólo con los datos reales obtenidos.

11.2. Conocimiento previo

Supongamos que no estamos lanzando una moneda, sino que estamos lanzando un clavo de cabeza plana. En el ámbito de las ciencias sociales, esto es como hacer una pregunta de encuesta sobre la zurdera o la diestra del encuestado, que sabemos que está lejos del 50/50, en contraposición a hacer una pregunta de encuesta sobre el sexo masculino o femenino del encuestado, que sabemos que está cerca del 50/50. Cuando volteamos el clavo, puede aterrizar con su cola puntiaguda tocando el suelo, un resultado que llamaré cola, o el clavo puede aterrizar equilibrado sobre su cabeza con su cola puntiaguda sobresaliendo, un resultado que llamaré cara. Creemos, sólo con mirar el clavo y pensar en nuestra experiencia previa con los clavos, que no saldrá cara y cola con la misma frecuencia. De hecho, es muy probable que el clavo salga con su punta tocando el suelo. En otras palabras, tenemos una fuerte creencia previa de que el clavo está inclinado hacia la cola. Supongamos que lanzamos el clavo 24 veces y sólo sale cara en 7 ocasiones. ¿Es el clavo justo? ¿Lo utilizaríamos para determinar qué equipo hace el saque de honor en la Superbowl?

11.2.1. Análisis de NHST

Al análisis de NHST no le importa si estamos lanzando monedas o clavos. El análisis procede de la misma manera que antes. Es decir, si declaramos que la intención era lanzar el clavo 24 veces, un resultado de 7 caras significa que no rechazamos la hipótesis de que el clavo es justo (donde $p > 2.5\,\%$). No obstante, tenemos un clavo para el que tenemos una fuerte creencia previa de que está sesgado por la cola. Volteamos el clavo 24 veces y descubrimos que sale cara 7 veces. Concluimos, por tanto, que no podemos rechazar la hipótesis nula de que el clavo puede salir cara o cruz 50/50. Pero, se trata de un clavo. ¿Cómo no se puede rechazar la hipótesis nula?

11.2.2. Análisis bayesiano

El estadístico bayesiano comienza el análisis con una expresión del conocimiento previo. Sabemos por experiencia previa que el clavo de cabeza estrecha está sesgado para mostrar colas, así que expresamos ese conocimiento en una distribución a priori. En el ejemplo ficticio sobre un clavo, supongamos que representamos nuestras creencias a priori mediante una muestra previa ficticia que tenía un 95 % de colas en una muestra de 20. Eso se traduce en una distribución previa beta$(\theta|2, 20)$.