GENERALIDADES SOBRE LA DISTRIBUCIÓN $\chi 2$

Isadore Nabi

Como se señala en (Snedecor & Cochran, 1967, págs. 20-21), el investigador a menudo tiene en mente una hipótesis definida sobre la proporción de la población, el propósito del muestreo es obtener evidencia sobre su hipótesis. Así, un genetista que estudia la herencia en el tomate podría tener razones para creer que, en las plantas producidas a partir de cierto cruce, los frutos con interior (en referencia a "la carne" del fruto, puesto que la traducción de "red flesh" es esa) rojo y los frutos con interior amarillo se presenten en una proporción de 3:1. En una muestra de 400 encontró 310 tomates rojos en lugar del hipotético 300. Con su experiencia de variación muestral, ¿aceptaría esto el investigador como verificación o refutación de la hipótesis? En otro ejemplo, un médico tiene la hipótesis de que cierta enfermedad que requiere hospitalización es igualmente común entre hombres y mujeres. En una muestra de 900 casos hospitalarios encuentra 480 hombres y 420 mujeres. ¿Estos resultados apoyan o contradicen su hipótesis? (por cierto, este es un ejemplo en el que la población muestreada puede diferir de la población objetivo); aunque la buena práctica médica puede prescribir la hospitalización, a menudo hay casos que por una razón u otra no llegan a un hospital y por lo tanto no pueden incluirse en su muestra. Para responder a las preguntas anteriores (que pertenecen a la misma familia de preguntas) se necesitan dos resultados, una medida de la desviación de la muestra de la proporción hipotética de la población y un medio para juzgar si esta medida es una cantidad que comúnmente ocurriría en el muestreo, o, por el contrario, es tan grande que arroja dudas sobre la hipótesis. Ambos resultados fueron proporcionados por padre de la estadística matemática y comunista Karl Pearson en 1899. Él ideó un índice de dispersión o criterio de prueba denotado por X^2 (chi cuadrado) y obtuvo la fórmula para su distribución de frecuencia teórica cuando la hipótesis en cuestión es verdadera. Al igual que la distribución binomial, la distribución chi-cuadrado es otra de las distribuciones teóricas básicas muy utilizadas en el trabajo estadístico. El índice de dispersión es, precisamente, el valor de chi-cuadrado.

Naturalmente, las desviaciones de los números observados de los especificados por la hipótesis forman la base del índice. En el ejemplo médico, con 900 casos, el número de casos de hombres y mujeres esperados en la hipótesis es de 450 para ambos sexos. En el cálculo de este valor, en su forma más básica, cada desviación (obtenida como la diferencia entre el valor observado y el valor hipotético o esperado), es elevada al cuadrado y dividida entre el valor hipotético o esperado y tales resultados se adicionan hasta su agotamiento. El valor esperado o esperanza matemática de cada una de las magnitudes estudiadas presente en el denominador en cada uno de los casos ha sido introducido para conseguir que la estimación

tome en consideración el tamaño de la muestra, puesto que el tamaño relativo (en relación al total) es lo importante, pues estadísticamente hablando, en general, de este emana el criterio de representatividad de una muestra en relación a una población. Complementariamente, las desviaciones han sido elevadas al cuadrado porque, según la fuente citada,

Es una práctica común en estadística. Simplemente diremos en la actualidad que se ha encontrado que los índices construidos de esta manera tienen una gran flexibilidad, siendo aplicables a muchos tipos diferentes de datos estadísticos. Tenga en cuenta que el cuadrado hace que el signo de la desviación no sea importante, ya que el cuadrado de un número negativo es el mismo que el del número positivo correspondiente. Está claro que chi-cuadrado sería cero si las frecuencias de la muestra fueran las mismas que las hipotéticas, y que aumentará al aumentar la desviación de la hipotética. Pero no está del todo claro si un valor de chi-cuadrado de, por ejemplo (para el caso planteado en el lugar referido), 4, debe considerarse grande, mediano o pequeño.

Proporcionar una base para juzgar este punto es nuestro próximo objetivo. En general, ¿qué valores de chi-cuadrado deben considerarse inusualmente grandes y qué valores como una variación muestral ordinaria? En el lugar citado se responde la pregunta de manera empírica, a través de algunos ejemplos, mientras que en este documento se resolverá teóricamente, estudiando la investigación fundacional, específicamente (Pearson, 1900, págs. 157-158).

Sea un sistema de desviaciones respecto de las medias de n-ésimas variables aleatorias (aquí se hace referencia al error estándar de la media, es decir, al error estándar de todas las posibles muestras extraíbles de cada una de las n-ésimas variables aleatorias) con desviaciones $\sigma_1, \sigma_2, \dots, \sigma_n$ y con correlaciones $r_{12}, r_{13}, r_{23}, \dots, r_{n-1,n}$. Este sistema de errores estándar genera un subespacio euclidiano de dimensiones [(n-1) × n] (generado empíricamente a partir de los datos muestrales, evidentemente), existirá una superficie del mismo que será la región de dicho espacio que contenga las frecuencias de las variables estudiadas (los errores estándar de las medias de generadas por cada una de las n-ésimas variables aleatorias) y que se expresa matemáticamente a través de la siguiente expresión:

$$Z = Z_{0e}^{-\frac{1}{2}\left\{S_1\left(\frac{R_{pp}}{R}*\frac{x_p^2}{\sigma_p^2}\right)+2S_2\left(\frac{R_{pp}}{R}*\frac{x_p}{\sigma_p}*\frac{x_q}{\sigma_q}\right)\right\}}$$

En la expresión anterior, R es el determinante (lo que implica que es una matriz cuadrada e invertible) de la matriz de covarianzas, R_{pp} y R_{qp} son los menores obtenidos mediante la manipulación algebraica usual (las usualmente denominadas operaciones elementales entre filas) de la p-ésima fila y la p-ésima columna, S_1 es la suma de todos los valores de p, mientras que S_2 es la suma de todos los valores de q.

Ahora, sea χ^2 una constante modelada mediante la siguiente relación:

$$\chi^2 = S_1\left(\frac{R_{pp}}{R} * \frac{x_p^2}{\sigma_p^2}\right) + 2S_2\left(\frac{R_{pp}}{R} * \frac{x_p}{\sigma_p} * \frac{x_q}{\sigma_q}\right)$$

La estructura matemática anterior expresa geométricamente hablando la ecuación de un elipsoide generalizado (es decir, un elipsoide n-dimensional), que geométricamente hablando a su vez es la generalización de una campana, que como el lector sabrá es la forma geométrica de la distribución de probabilidad normal, por lo que la función anterior generaliza la función de distribución normal multivariada y es, algebraicamente hablando, parte de la familia de superficies cuadráticas (superficies modeladas mediante ecuaciones de segundo grado), como se verifica en (Boyd, 2007). Pueden consultarse formas más generales de la función anterior, pero puesto que este documento ha sido elaborado para explicar la distribución de probabilidad y no su estructura matemática en profundidad, se expone de la forma en que vino al mundo en conexión con la teoría estadística.

Así, la distribución chi-cuadrada es la estructura geométrica (un elipsoide generalizado) que se erige sobre el espacio n-dimensional generado por Z y dentro de la cual la frecuencia del sistema de errores es constante (esto significa en términos menos técnicos que existe homocedasticidad); esta estructura geométrica es conocida, a nivel de la teoría estadística, como distribución elíptica. Los valores para los cuales χ^2 (es decir, los valores a evaluar dentro de la función) cubre todo el espacio (el dominio de la función, que en este caso coincide con su suporte -los valores donde la función no se nulifica-) es de cero a infinito.

Ahora supóngase además que el "elipsoide" es deformado tomando de referencia sus ejes principales[1] de tal forma que es convertida en una esfera[2] en la que $X_1, X_2, \dots X_n$ son sus coordenadas; matemáticamente hablando, lo que Pearson hizo ahí fue deformar la estructura geométrica sin alterar las posiciones relativas de los puntos en relación a un centro de referencia, lo cual es precisamente la definición de una deformación topológica. En tal caso, las probabilidades de que el sistema de errores posea una frecuencia igual o mayor que la denotada por χ está dada por:

$$P = \frac{\left[\int\int\int\int \dots e^{-\frac{1}{2}\chi^2} dX_1 dX_2 \dots dX_n\right]_{\chi}^{\infty}}{\left[\int\int\int\int \dots e^{-\frac{1}{2}\chi^2} dX_1 dX_2 \dots dX_n\right]_{0}^{\infty}}$$

En la expresión anterior, el numerador es la integral de superficie n-dimensional del elipsoide de chi hasta infinito, mientras que el denominador es la integral de superficie n-dimensional del elipsoide de cero hasta infinito.

Ahora, supóngase además que ocurre una transformación de coordenadas, de las coordenadas rectangulares antes presentadas a coordenadas polares generalizadas, en las cuales el radio de tales coordenadas es igual a χ y, por consiguiente, se conserva la comunalidad de factores integrales del numerador y del denominador, por lo que tales factores representan los ángulos sólidos de la figura geométrica[3] y tienen idénticos límites (los vectores característicos que describen el nuevo objeto geométrico tienen el mismo comportamiento asintótico).

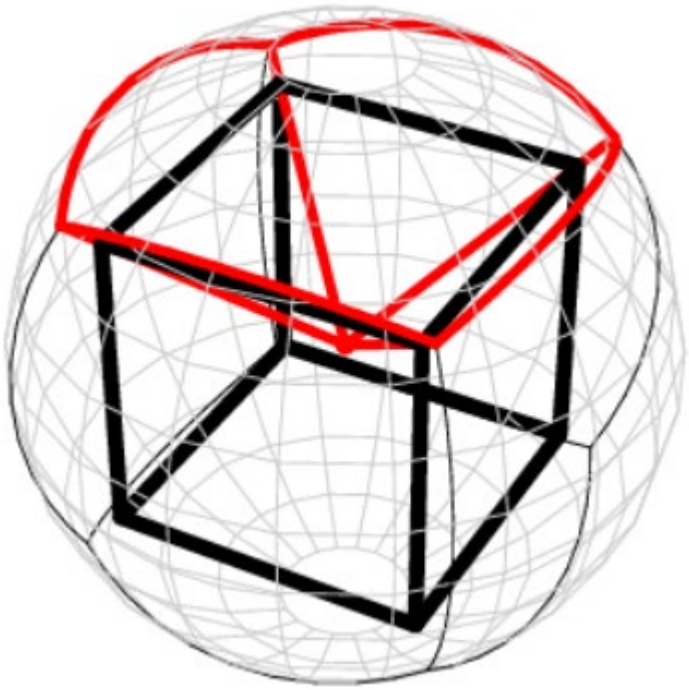
---

[1] Los diámetros principales o ejes principales son los diámetros máximo y mínimo de la elipse, perpendiculares entre sí y que pasan por el centro, como se verifica en (Wikipedia, 2021).

[2] Esto es posible porque existe un difeomorfismo entre una esfera y una elipse. Un difeomorfismo es un isomorfismo [véase (Nabi, 2021)] en variedades suaves, *i.e.*, en variedades $n - ésimamente$ diferenciables (que sus derivadas de orden $n - ésimo$ existen). Por lo anterior, existe una equivalencia esencial (desde la lógica formal, aunque este concepto puede generalizarse bajo las condiciones apropiadas a lógicas más robustas, como se verifica en la última fuente citada) entre tales estructuras geométricas.

[3] Los ángulos sólidos de un objeto geométrico son los ángulos espaciales que abarcan dicho objeto visto desde otro objeto geométrico de referencia (en este caso una n-bola), que se corresponde con la zona del espacio limitada por las rectas proyectantes desde el objeto hacia el observador (puesto que, para calcular el ángulo sólido de una superficie, se proyecta el objeto sobre una esfera de radio conocido). El radio de la n-bola es unitario para este caso, que es equivalente al dominio de un espacio de Lebesgue (espacio medible) en los reales y ello, en conjunto con otras condiciones, permiten la misma representación matemática para ambos.

Ángulos sólidos (cara superior del cubo centrado en el origen de longitud 2a) de una figura geométrica proyectada sobre una esfera de radio 1



Fuente: (Weisstein, 2021).

De lo anterior se desprende que es posible re-expresar P como:

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2} \chi^{n-1} d\chi}{\int_{0}^{\infty} e^{-\frac{1}{2}\chi^2} \chi^{n-1} d\chi}$$

La expresión anterior es la medida de probabilidad dentro de un sistema de números complejos[4] (que son las medidas realizadas sobre las n-ésimas variables chi) de que n errores ocurran con una frecuencia igual o mayor que la frecuencia efectivamente observada dentro del sistema. Además, lo que Karl Pearson denota como P es también la estructura matemática de la magnitud numérica que en la actualidad se conoce como el valor p para una distribución chi-cuadrada.

[4] Esta investigación fue publicada en 1900 y fue hasta en 1933 que el gran topólogo, estadístico matemático y marxista Andréi Kolmogórov construyó los espacios de probabilidad, que al igual como Pearson ve aquí a un espacio de números complejos, pueden ser vistos como sistemas probabilísticos.

REFERENCIAS

Boyd, S. (2007). *Symmetric matrices, quadratic forms, matrix*. Obtenido de Stanford University: Symmetric matrices, quadratic forms, matrix

Nabi, I. (14 de Marzo de 2021). *HACIA UNA INTERPRETACIÓN DIALÉCTICA-MATERIALISTA DE LA TOPOLOGÍA GENERAL: GÉNESIS HISTÓRICA-TEÓRICA DE LA TOPOLOGÍA DESDE LA GEOMETRÍA Y LA TEORÍA DE CONJUNTOS*. Obtenido de El Blog de Isadore Nabi: https://marxianstatistics.com/2021/03/14/hacia-una-interpretacion-dialectica-materialista-de-la-topologia-general-genesis-historica-teorica-de-la-topologia-desde-la-geometria-y-la-teoria-de-conjuntos/

Pearson, K. (1 de Julio de 1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157-175. Obtenido de http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Proportion/Pearson1900.pdf

Snedecor, G. W., & Cochran, W. G. (1967). *Statistical Methods* (Sexta ed.). Iowa: Oxford & IBH Publishing Co.

Weisstein, E. (8 de Septiembre de 2021). *Solid Angle*. Obtenido de MathWorld-A Wolfram Web Resource: https://mathworld.wolfram.com/SolidAngle.html

Wikipedia. (19 de Agosto de 2021). *Elipse*. Obtenido de Figuras geométricas: https://es.wikipedia.org/wiki/Elipse