Semana 4

Tablas de contingencia *Valentín Vergara Hidd*

Consideremos este documento como una especie de *pausa* de las pruebas de hipótesis que estamos revisando hasta ahora. Esto no implica que utilice una lógica diferente, sino que corresponde a una prueba de hipótesis que se utiliza únicamente cuando tenemos los datos grupados en una tabla de contingencia. Es decir, es una tabla con mínimo dos filas y mínimo dos columnas, que sintetiza información de un grupo más grande de casos. Dicho de otra forma, son tablas que cuentan datos.

Si bien esta es una forma muy común de encontrarse con datos, no es muy útil para la estadística inferencial. Lo único que podemos saber al utilizar tablas de contingencia es si existe independencia entre la información que se presenta en las filas y la que se presenta en las columnas.

Para trabajar con estas Tablas, presentaremos otra distribución de probabilidad: χ^2 (se pronuncia de forma similar a la primera sílaba de la palabra *kilo*). Su utilidad va más allá de las Tablas de contingencia, por lo que es aún de mayor importancia revisar este tipo de pruebas antes de continuar.

---

1. Tablas de contingencia

Es importante calrificar algunos conceptos antes de seguir. El primero de ellos es a qué nos referimos con una *tabla de contingencia*.

Definición 1: Tabla de contingencia

Una **tabla de contingencia** o tabla de doble entrada, es una forma de ordenar y sintentizar información, que consiste en mostrar dos variables categóricas al mismo tiempo: una de ellas ocupará las filas y otra las columnas. Su principal ventaja es que permite ver todas las psoibles interacciones entre la variable de las filas y de las columnas al mismo tiempo.

El uso que le vamos a dar las tablas de contingencia en este documento va más allá de simplemente mostrar información. Utilizaremos la lógica de la prueba de hipótesis para identificar dependencia entre la variable de las filas y la variable de las columnas. Llamaremos a esto un **test de independencia**.

Definición 2: Test de independencia

Un test de independencia es una prueba estadística en la que su hipótesis nula se plantea en términos de la independencia de la variable de las filas y la variable de las columnas.

La distribución de probabilidad que se utiliza en este caso corresponde a χ^2. Vale la pena analizar con detalle en qué consiste esta distribución.

2. Distribución χ^2

Consideremos un conjunto $\{Z_1, Z_2, Z_3, \ldots, Z_k\}$, donde cada elemento Z_i corresponde a una variable de una distribución normal estándar (ver documentos previos). Si a partir de estas k variables creamos una nueva variable aleatoria Q:

$$Q = \sum_{i=1}^{k} Z_i^2, \tag{1}$$

podemos afirmar que Q es una variable aleatoria con distribución χ^2 de parámetro k. Formalmente:

$$Q \sim \chi^2(k) \tag{2}$$

La función de probabilidad para esta variable es:

$$f(x;k) = \begin{cases} \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, & x > 0 \\ \\ 0, & \text{en cualquier otro caso} \end{cases} \tag{3}$$

Igual que en casos anteriores, $\Gamma()$ es la función del mismo nombre, que se *parece* a la función factorial. Al graficar la función de probabilidad se obtiene

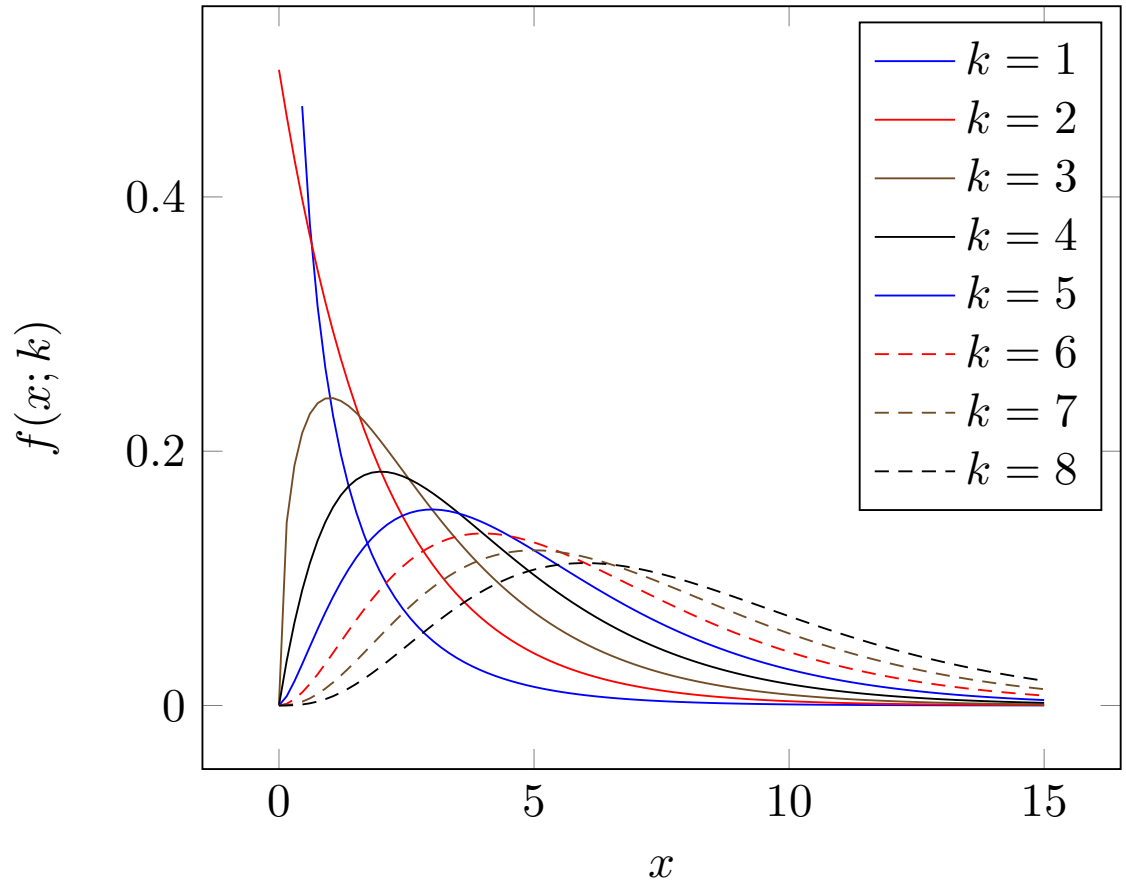

Figura 1: Función de probabilidad para la distribución χ^2 con distintos parámetros

El parámetro k es responsable de la forma que toma la distribución. Si lo pensamos en términos de la función de probabilidad 3 y su relación con la definición de la variable aleatoria que se presenta en 1, significa que a mayor cantidad de muestras, más se va a parecer a una curva gaussiana (como la distribución normal).

3. Pasos de la prueba de hipótesis

Tal como en documentos anteriores, para mostrar cómo se ejecuta una prueba de hipótesis, lo más pertinente es presentar un ejemplo. En este caso, tomaremos un ejemplo de **?**:

Ejemplo 1: Tratamientos para una fractura de pie

La tabla a continuación contiene información sobre dos variables: en las filas se encuentran distintos tratamientos para una fractura de pie; mientras que en las columnas se encuentra una variable que identifica aquellos casos que se recuperon exitosamente y aquellos que no.

	Recuperación exitosa	No recuperado
Cirugía	54	12
Yeso con peso	41	51
Yeso sin peso (6 semanas)	70	3
Yeso sin peso (<6 semanas)	17	5

Para este caso, una pregunta interesante sería: ¿Existe una relación entre el tipo de tratamiento y la recuperación de una fractura? Otra forma de plantear esta pregunta es. ¿son independientes ambas variables?

Los cuatro pasos de la prueba de hipótesis son:

3.1. Establecer hipótesis

En este caso, la formulación es bastante simple:

$$H_0 \text{ :La recuperación es independiente del tratamiento}$$
$$H_1 \text{ :El tratamiento está relacionado con la recuperación}$$

Noten que deliberadamente invertí las variables entre una hipótesis y otra. Esto es para hacer énfasis en que en una tabla de contingencia, no existe variable dependiente e independiente. Simplemente estamos revisando si existe algún tipo de relación entre ellas, sin siquiera establecer cuál es esa relación.

3.2. Establecer α y valores críticos

Como siempre, utilizaremos $\alpha = 0,05$. Para determinar el valor crítico de la distribución de probabilidad[1], los grados de libertad son $(\text{filas} - 1)(\text{columnas} - 1)$. Para encontrar el valor crítico, se puede utilizar esta tabla. Al igual que con el resto de las distribuciones, se puede calcular el valor de la tabla a partir de la expresión 3 y utilizando el hecho que:

[1] Este valor crítico, que en este caso llamaremos χ^2*, corresponde a aquel que separa la región aceptación de H_0 con la de no-aceptación.

$$\int_0^{\infty+} f(x;k)dx = 1 \tag{4}$$

En cualquier caso, se define el valor crítico en la distribución de probabilidad utilizando como grados de libertad $gl = (4-1)(2-1) = 3$, lo que arroja

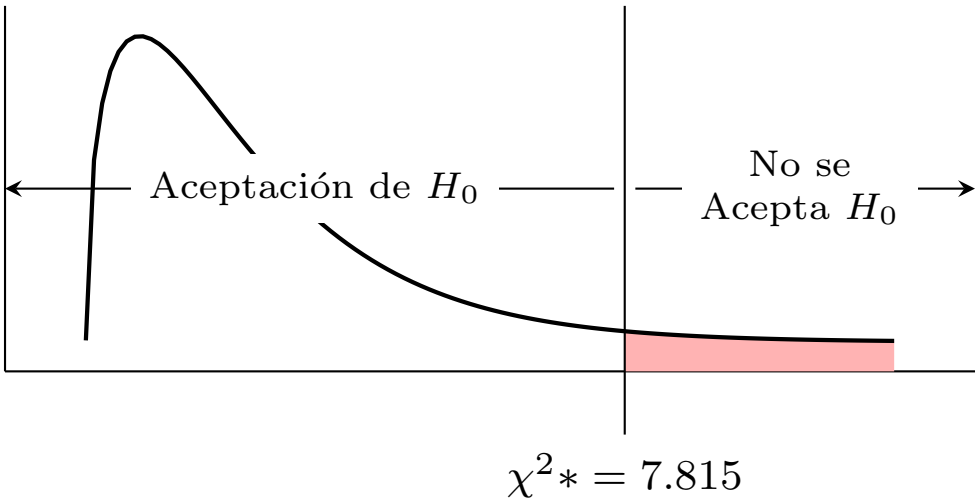

Figura 2: Valor crítico de la distribución $\chi^2(k = 3)$

3.3. Calcular el valor de la prueba estadística

En este caso, la pruerba estadística se calcula con

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{5}$$

donde O_i y E_i representan el valor observado y el valor esperado de la celda i en la tabla de contingencia, respectivamente. Para calcular el valor esperado de cada celda:

$$E_i = \frac{(\text{total fila})(\text{total columna})}{(\text{total general})} \tag{6}$$

Por tanto, para la tabla con los datos del ejemplo 1

	Recuperación exitosa	No recuperado
Cirugía	$O = 54;\ E = 47,478$	$O = 12;\ E = 18,522$
Yeso con peso	$O = 41;\ E = 66,182$	$O = 51;\ E = 25,818$
Yeso sin peso (6 semanas)	$O = 70;\ E = 52,514$	$O = 3;\ E = 20,486$
Yeso sin peso (<6 semanas)	$O = 17;\ E = 15,826$	$O = 5;\ E = 6,174$

Luego, para calcular la prueba, se reemplazan los valores de la tabla en la expresión 5:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(54 - 47,478)^2}{47,478} + \cdots + \frac{(5 - 6,174)^2}{6,174} = 58,393 \tag{7}$$

3.4. Decisión

Posicionando el valor calculado en 7 en la Figura 2; se llega a la decisión de **no aceptar** H_0. Es decir, podemos afirmar con un 95 % de confianza que existe una relación entre la recuperación de una fractura de pie y el tipo de tratamiento.

4. Para practicar

Para continuar practicando el análisis de tablas de contingencia, además de replicar el ejemplo que presento en el video, una idea es replicar el ejemplo presentado en el video anterior. Es decir, ¿existe relación entre la el nivel educacional y la región de residencia? Algunas recomendaciones para el análisis

- Utilizar sólo tres regiones: Biobío, Araucanía y Metropolitana
- Limpiar las categorías que no son relevantes del nivel educacional

Una vez que hayan hecho el análisis, recomiendo comparar los resultados con aquellos obtenidos en la prueba que utiliza análisis de varianza.