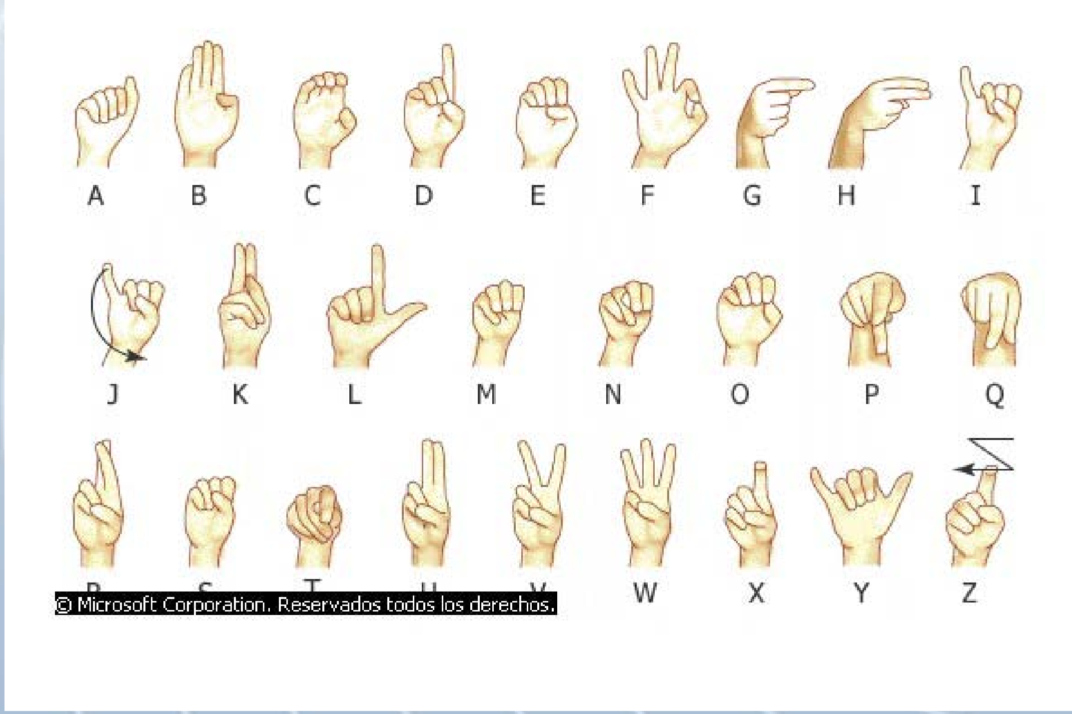
UNIVERSIDAD POLITÉCNICA DE MADRID

conento

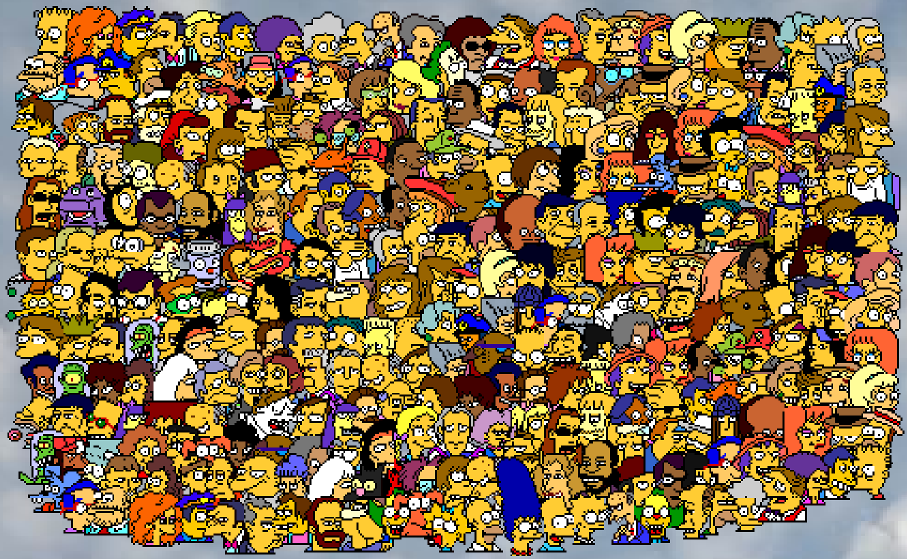
Aprendiendo idiomas VII

María Jesús Vázquez Gallo



Formación Conento: marzo 2009

1. Introducción. Contraste de hipótesis no paramétricas.
2. Contraste de bondad de ajuste.
3. Otros contraste no paramétricos.
4. Continuará…



Ya sabemos algo sobre el contraste de hipótesis, una técnica estadística con la que decidimos aceptar o rechazar una propiedad de una población, a partir de una muestra suya.

¿Recuerdas?

Ejemplo: Tiramos 300 veces un dado, y queremos contrastar si el dado está bien construido. La hipótesis nula H_0 diría que la muestra de tiradas sigue una distribución de probabilidad uniforme (es decir, cada uno de los posibles sucesos: 1, 2, 3 ,4 5 y 6, tiene la misma probabilidad, 1/6).

Esto sería un ejemplo de contraste de hipótesis no paramétrico, en el que no decidimos sobre características numéricas que describan la población.

¿Y por qué no? ¿no sería más preciso?

Puede que:

- queramos determinar precisamente la distribución de probabilidad a la que pueden ajustarse los datos (no se conoce a priori)
- los datos no tengan una interpretación numérica clara (porque son cualitativos, por ejemplo);
- se tengan pocos datos.

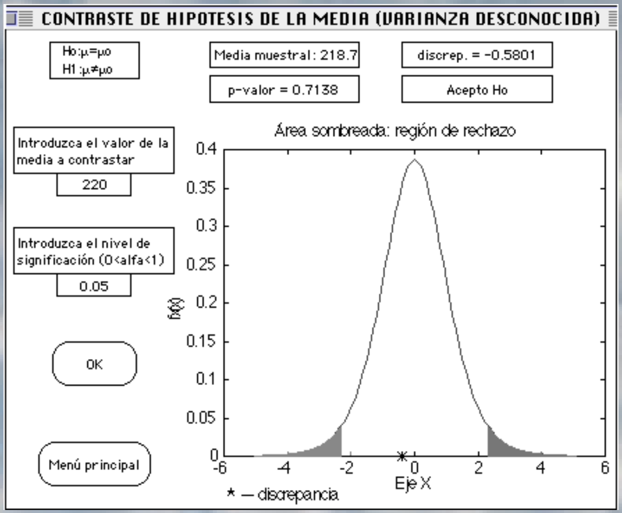
En estas situaciones, los contrastes no paramétricos son rápidos y fáciles de usar, tienen expresiones sencillas y son válidos para muestras pequeñas.

Entonces, ¿por qué se utilizan contrastes paramétricos?

Los contrastes no paramétricos generalmente son menos potentes que sus correspondientes paramétricos.

Moraleja: Preferiremos un contraste no paramétrico cuando aporte una simplificación importante frente al no paramétrico.

Las etapas de un contrastes de hipótesis son comunes a los paramétricos y a los no paramétricos:



Etapas del contraste

1. **Definición de hipótesis.**
2. **Definición de estadístico de contraste.**
3. **Indicar el valor del nivel de significación.**
4. **Decisión sobre las hipótesis según el valor crítico del contraste.**

Algunos contrastes de hipótesis no paramétricos:

- Bondad de ajuste.
- Comparación de poblaciones.
- Independencia entre variables.
- Tamaño muestra.
- Aleatoriedad.

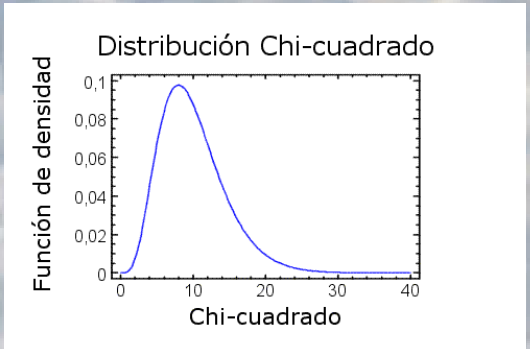
El objetivo de los contrastes de bondad de ajuste es determinar, a través de una muestra, si una variable aleatoria sigue una cierta distribución teórica dada de antemano (normal, uniforme, etc).

Hipótesis nula: la variable se ajusta a una distribución concreta.

¿Recuerdas cómo describe una distribución de probabilidad el comportamiento de una variable aleatoria ? Lo vimos en **Aprendiendo idiomas III**

Función de densidad → le asigna a cada valor posible de la variable la probabilidad de que ocurra.

Función de distribución → le asigna a cada valor posible de la variable la probabilidad de que la variable no lo supere.



Sea cual sea la distribución teórica considerada, siempre existirán diferencias entre los valores teóricos esperados y los observados.

Queremos saber en qué medida dichos valores son debidos al azar (aceptaremos la hipótesis nula: el dado está bien construido) o a que los datos no se ajustan a la distribución teórica considerada (rechazaremos la hipótesis nula: el dado estaba trucado).

Contrastes de bondad de ajuste

1. **Basados en las frecuencias muestrales.**
2. **Basados en estadísticos de posición.**
3. **Contraste de normalidad.**

Los contrastes de bondad de ajuste basados en frecuencias comparan frecuencias teóricas esperadas con las observadas.

1. Contraste Ji-cuadrado (se pronuncia chi-cuadrado):

- Se establece una partición del espacio muestral en k sucesos mutuamente exclusivos.
- Necesita un número suficiente de datos (al menos 30); también es necesario que las frecuencias esperadas sean mayores o iguales que 5.
- Válido tanto para variables discretas como continuas.

El estadístico de contraste es:

$$\sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}$$

frecuencia observada

frecuencia esperada

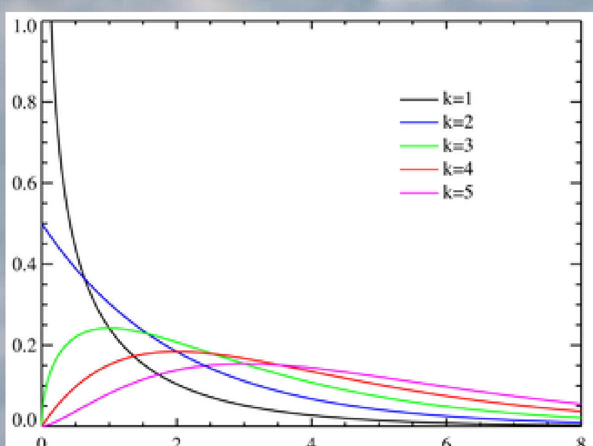
Y sigue una distribución Ji-cuadrado de Pearson χ^2_{k-1}

Distribución Ji-cuadrado de Pearson:

es una distribución de probabilidad continua, con un parámetro *k* que representa los grados de libertad de la variable aleatoria:

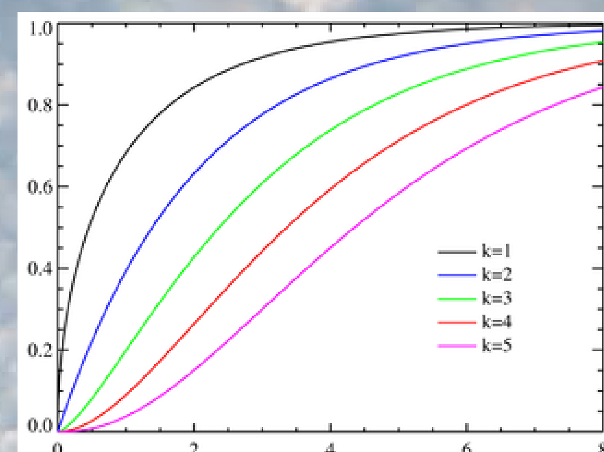
$$X = Z_1^{\,2} + Z_2^{\,2} + \cdots + Z_k^{\,2}$$

donde Z_i son variables de distribución normal, de media cero y varianza uno.



Función de densidad

$$\chi_{k-1}^{\;2}$$



Función de distribución

Ejemplo: Volvemos a la muestra de 300 tiradas de un dado.
La hipótesis nula es: dado bien construido (no trucado).
A partir de la tabla de frecuencias observadas y teóricas esperadas:

	1	2	3	4	5	6	Total
Oi	44	62	52	45	50	47	300
Ti	50	50	50	50	50	50	300

El valor del estadístico de contraste es:

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - T_i)^2}{T_i} = 4.36$$

El valor crítico para la ji-cuadrado con 5 grados de libertad a niveles de significación α del 0,05 y 0,01 es, respectivamente: 11,1 y 15,1. Nuestro estadístico es inferior a ambos valores, por lo que estamos en la región de aceptación y se acepta la hipótesis nula de que el dado es correcto.

Recuerda que: la elección de un nivel de significación α divide en dos regiones el conjunto de posibles valores del estadístico de contraste:

- La región de Rechazo, con probabilidad α, si H_0 es cierta.
- La región de Aceptación, con probabilidad 1- α , si H_0 es cierta.

Recordando lo que vimos en **Aprendiendo idiomas IV:**

En el ejemplo, ¿podría ocurrir que la H_0 fuese falsa?

Si puede ocurrir, ¿con qué probabilidad sucede?

- Contrastes basados en estadísticos de posición.

La idea es que si la muestra procede de la distribución teórica considerada, entonces sus estadísticos de posición (medianas, cuartiles, etc.) no deberían ser muy diferentes.
Por tanto, estos contrastes se basan en la distancia entre los estadísticos de posición de la muestra y los teóricos.

- Contrastes de normalidad.

La distribución teórica considerada es la normal. Para efectuar el contraste, se calculan la media y la varianza muestral y se ordena la muestra en sentido creciente. A continuación, se toman las diferencias entre los elementos mayor y menor, el segundo mayor y menor, etc. Se utiliza un estadístico basado en dichas diferencias, llamado estadístico de Shapiro-Wilk.

1. Los contrastes de homogeneidad entre distribuciones tratan de decidir si dos muestras independientes pueden haber sido extraídas de la misma distribución de probabilidad.

Hipótesis nula: las distribuciones de probabilidad de las que se extrajeron ambas muestras son iguales.

Caso importante: observaciones de varias características de una población, o bien mediciones de una misma característica en distintos instantes de tiempo. En ambos casos, tenemos varias observaciones para cada individuo en la muestra, y queremos contrastar la homogeneidad de las distribuciones de cada una de dichas observaciones.

Ejemplo: un grupo de alumnos realiza dos exámenes de Inglés, antes y después de una estancia en el Reino Unido. Los exámenes tienen la misma dificultad y se pretende analizar la efectividad de la estancia en el extranjero en cuanto al aprendizaje del ingles

2. Los contrastes de independencia entre variables.
Se utilizan para contrastar la posible independencia de dos características de los elementos muestrales.
Válidos para variables discretas cualitativas o cuantitativas.

Ejemplo: podemos contrastar si la edad de un votante esta relacionada con su opción política, el sexo de una persona con las compras que realiza, etc.

Hipótesis nula: ausencia de asociación entre características.

Obs: No depende de una hipótesis de causalidad acerca de la dirección en que se produce la asociación entre dos variables.

En el futuro, aprenderemos como cuantificar el grado de correlación entre variables utilizando modelos econométricos.


Continuará...