

## CURSO DE POSTGRADO FANUS

### MÓDULO 7. METODOLOGÍA ESTUDIOS EPIDEMIOLÓGICOS

Dra. D. Corella

#### CONCEPTOS BÁSICOS DE ESTADÍSTICA

Los conocimientos estadísticos son fundamentales en biomedicina. El alumno debe poseer una sólida base bioestadística. Para consolidar dichos conocimientos se presenta una breve introducción teórica antes de realizar los ejercicios de aplicación.

##### Conceptos básicos en estadística:

La estadística tiene dos áreas fundamentales, la llamada **estadística descriptiva** y la **estadística analítica**. El objeto de la estadística descriptiva es describir y resumir la muestra. La estadística analítica o inferencial tiene como objetivo comprobar determinadas hipótesis referentes a una población, a partir del análisis de los resultados obtenidos en muestras.

##### Estadística deductiva o inferencial. Contraste de hipótesis.

Mediante la llamada estadística inferencial, se pretende generalizar a la población los resultados obtenidos en muestras a través de modelos probabilísticos aplicados a los datos. Para ello se recurre al denominado **contraste de hipótesis**. El objetivo del contraste de hipótesis consiste en comprobar determinadas hipótesis (*definición: consecuencias contrastables obtenidas a partir de un cuerpo de conocimientos y de unas normas lógicas*) referentes a una población, a partir del análisis de los resultados obtenidos en muestras.

Existen dos tipos de hipótesis :

##### Hipótesis nula: $H_0$

Se define la hipótesis nula como la hipótesis que se desea contrastar. Habitualmente es la de no diferencia.

Por ejemplo, en el caso anterior, la hipótesis nula sería: "*no existe diferencia de medias de LDL-C entre el grupo de pacientes que recibe un fármaco, y el grupo control al cabo de 1 año de tratamiento*"

##### Hipótesis alternativa: $H_1$

Es la hipótesis contraria a la hipótesis nula.

Siguiendo con el ejemplo, la *hipótesis alternativa* sería aquella que dice que *sí que existe diferencia de medias entre los dos grupos*.

La comprobación de estas hipótesis se lleva a cabo mediante los llamados **tests estadísticos** que más adelante comentaremos. Estos tests, pueden ser de dos tipos:

1.-**Paramétricos**: Se asume una distribución de probabilidad de la variable en la población, y se contrastan valores de parámetros.

2.-**No paramétricos**: No se supone una distribución de probabilidad.

Sea cual sea el test estadístico aplicado, existen unas **reglas de decisión** para aceptar o rechazar la hipótesis nula. A partir de los valores del test estadístico, se divide su distribución en dos regiones:

##### a) Región de aceptación

##### b) Región de rechazo

Los valores que comprenden la *región de rechazo* son aquellos que tienen menor *probabilidad* ( $p$ ) de ocurrir si la hipótesis nula es cierta. Si los valores del test calculado, caen en la región de rechazo, se rechaza la hipótesis nula. En caso contrario, se acepta.

Un test, se dice que es *significativo* si su valor cae en la zona de rechazo.

Las regiones de aceptación o rechazo se eligen en virtud del llamado **nivel de significación  $\alpha$** . Si el test es bilateral, se construye la región de aceptación como el intervalo que deja a su izquierda y derecha una probabilidad  $\alpha/2$ .

Habitualmente, el nivel de significación crítico para aceptar o rechazar la hipótesis nula se establece en  **$\alpha=0,05$** . De esta forma podemos rechazar la hipótesis nula con un 95% de confianza.

En términos probabilísticos, tenemos dos tipos de errores

a) Error **Tipo I o  $\alpha$**  : Es la probabilidad de rechazar la hipótesis nula siendo cierta

b) Error **Tipo II o  $\beta$** : Es la probabilidad de aceptar la hipótesis nula siendo falsa.

Además de con las *pruebas de hipótesis*, también podemos hacer inferencias basándonos en los llamados **intervalos de confianza**. Son intervalos (entre un mínimo y un máximo) que con una probabilidad determinada alojan el valor del parámetro. Normalmente, se utilizan los intervalos de confianza al 95%.

### Principales tests estadísticos:

En función del tipo de inferencia que se desee realizar, habrá que utilizar un test estadístico u otro. Además, habrá que tener en cuenta la distribución de la variable y/o el número de casos en la muestra.

En la Tabla 1, se resumen los principales tests estadísticos aplicados en biomedicina. Al abordar el análisis estadístico hay que delimitar también si se van a aplicar las llamadas técnicas **univariantes** (una sola variable dependiente), o las técnicas **multivariantes** (dos o más variables). También es imprescindible definir cuál es la **variable dependiente** (variable cuyo comportamiento queremos predecir) y cuál o cuáles son las **variables independientes**, llamadas también variables predictoras.

**Tabla 1: Guía para la aplicación de los principales test estadísticos en biomedicina.**

<b>Comparación</b>	<b>Consideraciones</b>	<b>Test más adecuado</b>
Medias de dos grupos independientes	Distribución normal	T de Student con corrección si las varianzas no son homogéneas
Medias de dos grupos apareados	Distribución normal	T de Student
Medias de más de dos grupos	Distribución normal	ANOVA
Varianzas de dos grupos	Distribución normal	Levenne
Medias de dos grupos independientes	No normal	U de Mann-Whitney
Medias dos grupos apareados	No normal	Rangos de Wilcoxon
Medias de más de dos grupos	No normal	Kruskal-Wallis
Porcentajes de grupos independientes	Frecuencias esperadas mayores que cinco	Ji cuadrado de Pearson
Porcentajes de grupos independientes	Frecuencias esperadas menores que cinco	Test exacto de Fisher
Porcentajes de grupos apareados		Test de McNemar
Correlación entre dos variables continuas	Distribución normal	Coefficiente de correlación de Pearson
Correlación entre dos variables continuas	No normal	Rho de Spearman
Asociación entre una variable dependiente y otra independientes	Variable dependiente continua	Regresión lineal simple
Asociación entre una variable dependiente y dos o más independientes	Variable dependiente continua, y las independientes continuas o categóricas (dummy)	Regresión lineal múltiple
Asociación entre una variable dependiente y otra independiente	Variable dependiente dicotómica (2 categorías)	Regresión logística simple
Asociación entre una variable dependiente y dos o más independientes	Variable dependiente dicotómica.	Regresión logística múltiple

## Correlación y regresión. Regresión lineal simple y múltiple.

Cuando se quiere conocer la asociación entre dos variables continuas se calcula el llamado coeficiente de correlación lineal de Pearson o también llamado coeficiente producto-momento. Este coeficiente, expresado como (r) es una medida de la relación lineal entre dos variables continuas. Puede tomar valores entre 0 y 1. Cuanto mayor sea en valor absoluto, mayor será la magnitud de la asociación. Sin embargo, no podemos comparar directamente coeficientes de correlación de estudios con un número de casos muy diferente, ya que, a igualdad de significación estadística, la magnitud de r es menor a mayor número de casos. El coeficiente de correlación puede tomar valores positivos o negativos. Cuando r es positivo, nos indica una relación directa entre las variables. Cuando r es negativo, nos indica una relación inversa.

Si además de calcular el grado de asociación entre dos variables continuas, se quiere establecer cómo se relacionan, el análisis adecuado es realizar una regresión lineal simple. Para ello hay que establecer cuál es la variable dependiente, y cuál la independiente. Mediante el método de regresión lineal o regresión de mínimo cuadrados se estima la ecuación predictiva que relaciona ambas variables, de forma que: **Y=a+bX**.

Y: es la variable dependiente

X: es la variable independiente

a= ordenada en el origen o constante

b= pendiente de la recta o coeficiente de regresión (indica lo que aumenta o disminuye la variable dependiente por cada unidad de cambio de la variable independiente)

Además del cálculo del coeficiente de correlación (r), es muy útil el cálculo del **coeficiente de determinación** (r<sup>2</sup>), que expresado en porcentaje indica la variabilidad de la variable dependiente que es explicado por la combinación lineal de variables independiente.

**Regresión lineal múltiple:** Si tenemos dos o más variables independientes en el modelo de regresión lineal se habla de regresión lineal múltiple. La ecuación predictiva toma entonces la siguiente forma:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_nX_n, \text{ donde:}$$

Y: es la variable dependiente

X<sub>1</sub>: es la variable independiente número 1

X<sub>n</sub>: es la variable independiente número n

a = ordenada en el origen o constante

b<sub>1</sub> = coeficiente de regresión de la variable X<sub>1</sub>

b<sub>n</sub> = coeficiente de regresión de la variable X<sub>n</sub>

En estos casos, también se calcula el coeficiente de determinación global del modelo, así como la significación estadística global y de cada uno de los coeficientes de regresión que se llaman coeficientes de regresión parcial.

## Regresión logística simple y múltiple

Cuando la variable dependiente es dicotómica, no podemos utilizar regresión lineal y es necesario utilizar la denominada regresión logística. En la regresión logística se utiliza una transformación logit:

Habitualmente se toma como F la función de distribución logística, dada por:

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1'x_i)}}$$

Esta función tiene la ventaja de ser continua. Además, como,

$$1 - p_i = \frac{e^{-(\beta_0 + \beta_1'x_i)}}{1 + e^{-(\beta_0 + \beta_1'x_i)}} = \frac{1}{1 + e^{(\beta_0 + \beta_1'x_i)}}$$

resulta que

$$g_i = \log \frac{p_i}{1 - p_i} = \log \left( \frac{\frac{1}{1 + e^{-(\beta_0 + \beta_1'x_i)}}}{\frac{e^{-(\beta_0 + \beta_1'x_i)}}{1 + e^{-(\beta_0 + \beta_1'x_i)}}} \right) = \log \left( \frac{1}{e^{-(\beta_0 + \beta_1'x_i)}} \right) = \beta_0 + \beta_1'x_i. \quad (2)$$

de modo que, al hacer la transformación, se tiene un modelo lineal que se denomina *logit*.

Una de las características que hacen tan interesante la regresión logística es que el coeficiente B de la ecuación para ese factor está directamente relacionado con el odds ratio OR. Es decir que exp(B) es una medida que cuantifica el riesgo que representa poseer el factor correspondiente respecto a no poseerlo, suponiendo que el resto de variables del modelo permanecen constantes:  $OR = e^B$