8

Distribuciones de probabilidad. El teorema central del límite

Neus Canal Díaz

8.1. Introducción

La distribución de frecuencias es uno de los primeros pasos que debemos realizar al inicio del análisis estadístico, conjuntamente con la aplicación de las medidas descriptivas, y refleja cómo se reparten los individuos de una muestra según los valores de una variable. Cuando se trata de poblaciones, la probabilidad de observar los diferentes valores de una variable aleatoria pueden expresarse como una función de probabilidad. La mayoría de los fenómenos de interés en investigación científica, como pueden ser la talla y la presión arterial, siguen unas leyes o distribuciones de probabilidad teóricas, especificadas matemáticamente en las que se basan la mayoría de los métodos estadísticos. La distribución más conocida es la distribución Normal o de Gauss. Muchos de los procedimientos estadísticos habitualmente utilizados asumen la normalidad de los datos observados. Aunque muchas de estas técnicas no son demasiado sensibles a desviaciones de la distribución normal, y en general esta hipótesis puede obviarse cuando se dispone de un número suficiente de datos (teorema central del límite), resulta recomendable contrastar si se puede asumir o no una distri-

bución Normal. Para decidir si nuestra muestra procede o no de una distribución normal existen gráficos (gráficos P-P y Q-Q) y contrastes de hipótesis (test de Kolmogorov-Smirnov) que pueden ayudarnos. Cuando los datos no son normales pueden transformarse o emplearse otros métodos estadísticos que no exijan este tipo de restricciones, llamados los métodos no paramétricos.

8.2. Concepto de función de distribución

Siempre que se quiera realizar un estudio, debemos medir la(s) variable(s) que caracterizan los resultados del mismo. Tales variables se conocen como **variables aleatorias**. Decimos que una variable es continua si puede tomar cualquier valor en un intervalo conocido (por ejemplo, TAS) y es discreta si sólo puede tomar algunos valores (respuesta completa, respuesta parcial, enfermedad estable).

Imaginemos que obtenemos una muestra de los valores de TAS de 100 pacientes; si los agrupáramos en pequeños grupos de igual rango de valores de presión arterial, es decir, un grupo cada 5 mmHg y contásemos cuántos hay en cada grupo, podríamos dibujar un gráfico o histograma como el que se muestra en la Figura 29. Si cada vez hiciéramos los intervalos más estrechos, así como también aumentáramos el tamaño de muestra veríamos que el histograma tiende a estabilizarse llegando a convertirse su perfil en la gráfica de una función. De esta forma, las distribuciones de probabilidad de variables continuas se definen mediante una función y=f(x) llamada **función de probabilidad** o **función de densidad** y asocia valores de una **variable aleatoria** con sus **respectivas probabilidades.**

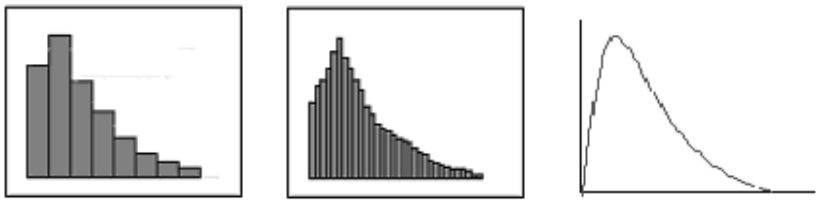

Figura 29. Histograma de una variable aleatoria y evolución a la función de probabilidad

La función de densidad de una variable aleatoria cumple que es positiva en todo su dominio, que toma valores entre 0 y 1 y que permite obtener la probabilidad de que un valor de la variable aleatoria se encuentre entre dos puntos, siendo esta probabilidad el área bajo la curva. El área bajo la curva de cualquier función de probabilidad es 1.

A partir de aquí observamos que sería útil para cada tipo de variable conocer la probabilidad de obtener un valor concreto. Imaginemos que queremos conocer la probabilidad de encontrarnos con un valor de TAS de entre 100 mmHg y 110 mmHg, indicados en la figura anterior. Como la figura indica la función de probabilidad, sabemos que será el área de la curva existente entre los dos puntos. En este concepto es cuando debemos utilizar la **función de distribución de probabilidad acumulada** o simplemente **función de distribución**, que determina para cada valor de nuestra variable, la probabilidad de obtener un valor menor a él, tal como se muestra en la siguiente anotación:

$$P(X \leq x) = F(x)$$

Evidentemente, el valor de la función de distribución es igual a la suma de todos los valores de la función de probabilidad desde el extremo inferior del dominio de la variable hasta x inclusive.

8.3. Distribuciones más utilizadas en estadística

En estadística, todos los sucesos se intentan definir como variables aleatorias. Existen muchos tipos de ellas según sea su distribución de probabilidad. Dado que estas distribuciones tienen una relación directa con los datos por si mismos y con las pruebas que deben realizarse para su análisis, comentaremos las distribuciones más comunes como son: la distribución normal, la binomial, la Ji-cuadrado y la F de Fisher, entre otras. Existen otras distribuciones más complejas y otras que son más específicas de sectores propios, como la Poisson en la teoría de colas. Sin embargo, estos conceptos sobrepasan nuestro propósito.

8.3.1. La distribución Normal

En muchas distribuciones de probabilidad de las variables cuantitativas se observa una tendencia de los valores alrededor de la media y menos observaciones a medida que nos acercamos a los extremos del rango de valores. Si el número de observaciones es grande la distribución adopta una forma de campana: campana de Gauss o distribución Normal. La curva en forma de campana viene dada por la ecuación:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \qquad -\infty < x < \infty$$

Donde x representa un valor de la variable, μ su media y σ la desviación

estándar. La distribución normal o gaussiana viene determinada por dos parámetros, su media y si desviación estándar, denotadas generalmente por μ y σ. Así se dice que una característica X sigue una distribución Normal de media μ y varianza σ^2, y se denota como $X \sim N(\mu, \sigma)$.

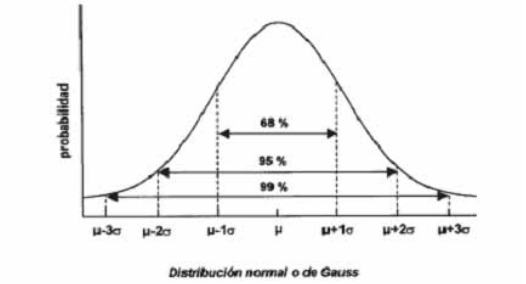

Figura 30. Distribución de probabilidad normal o de Gauss y simetría respecto a μ

La distribución Normal se define por presentar ciertas propiedades importantes. Es una función continua que tiende asintóticamente a infinito por los extremos, es simétrica con respecto a su media μ, por lo que existe una probabilidad de un 50% de observar un valor mayor a la media y la misma probabilidad de observar un valor menor. Dadas las características de esta distribución, la media, la mediana y la moda coinciden siempre.

La forma de la campana depende de los parámetros μ y σ. La media indica la posición de la campana, de modo que para diferentes valores de μ la gráfica se desplaza a lo largo del eje horizontal. El parámetro σ indica la dispersión de los datos entorno a la media, así cuando mayor sea σ mas plana será la curva, tal como se indica en la siguiente figura. De un modo gráfico, podemos decir que σ es la distancia que existe entre la media μ y el punto de inflexión de la curva.

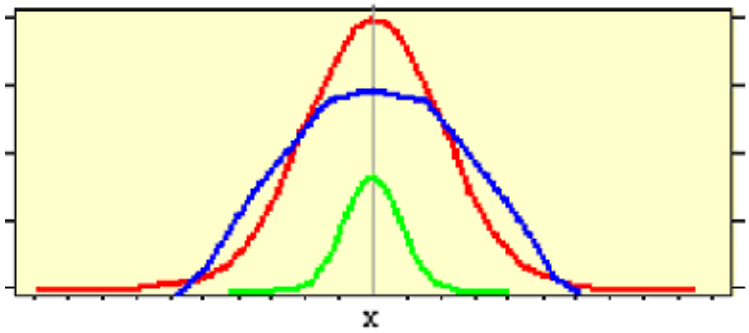

Figura 31. Diferentes formas de funciones normales según el valor de su parámetro σ.

8.3.2. Distribución Normal estándar

Los valores de una población normal (μ, σ) pueden transformarse en valores de una población estandarizada. Este proceso de normalización transforma la función original N(μ, σ) en una función estandarizada N(0,1), resultando interesante en la práctica, ya que de esta distribución existen tablas publicadas a partir de las que se puede obtener la probabilidad de observar un dato menor o igual a un cierto valor z. La ecuación de estandarización es:

$$Z = \frac{X - \mu}{\sigma}$$

Veamos un ejemplo: Supongamos que observamos una característica que sigue una distribución normal con μ igual a 50 y σ^2 igual a 81 (el valor σ es 9), siendo su notación N(50,9). Nuestro objetivo es conocer cuál es la probabilidad de que una observación obtenida al azar tome un valor mayor a 63. Mediante la estandarización obtendríamos:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 50}{\sqrt{81}} = \frac{X - 50}{9}$$

De modo que $P(X > 63) = P[(X - 50)/9 > (63 - 50)/9)] = P(Z > 1{,}4) = 1 - P(Z \leq 1{,}4)$. Esta probabilidad se encuentra a partir de las tablas de la N (0 , 1) que se encuentran en el Anexo. Situándonos en el valor de Z correspondiente al valor 1,4, observamos que el valor de p que le corresponde es de $P(Z \leq 1{,}4) = 0{,}925$. Así obtenemos que la probabilidad de elegir un valor mayor a 63 (el complementario de que sea menor o igual) es de 1-0,925 = 0,075, es decir, del 7,5%.

Existen unos valores de Z que son muy conocidos por su amplia utilización en cualquier prueba en la que se utilice esta distribución. Dichos valores corresponden a los puntos de la función de densidad que dejan una probabilidad acumulada conocida. Tal como se indica en la siguiente figura, el punto z=1,96, deja una probabilidad acumulada de 0,975 y su simétrico (z=-1,96), deja una probabilidad acumulada de 0,025. Esto implica que el área que se establece entre un punto y otro de la curva, corresponde a la resta de sus dos probabilidades, siendo por tanto de 0,95. Esta probabilidad está asociada a lo que veremos como intervalo de confianza al 95%.

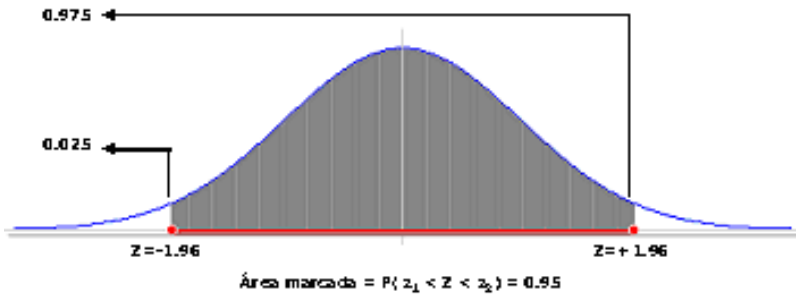

Figura 32. Distribución normal (0,1) y área bajo la curva

Otros valores críticos de la distribución normal estandarizada son los valores z=1.64 y z=2.58 que se representan a continuación y que serán útiles en el capítulo 9.

Z	1.64	1.96	2.34	2.58
Probabilidad acumulada p(z<Z)	0.950	0.975	0.990	0.995

8.3.2.1. Contrastes de normalidad

Los gráficos de probabilidad normal son una herramienta gráfica para averiguar si un conjunto de datos procede de una distribución Normal. Básicamente consiste en enfrentar en un gráfico los datos observados frente a los datos que se obtendrían en una distribución normal. Si la distribución coincide con la Normal, los puntos se concentrarán entorno de una línea recta. En los gráficos P-P se comparan las proporciones acumuladas de nuestra variable con las de la distribución normal, mientras que los gráficos Q-Q representan los cuartiles de nuestra variable respecto los cuartiles de la distribución normal.

En el gráfico P-P, se representa para cada probabilidad acumulada de los valores de la variable (Prob acum. observada) la probabilidad acumulada esperada si la muestra perteneciera a una distribución normal. Si la variable sigue una distribución Normal, los puntos estarán dispersados a lo largo de la línea diagonal, que representa la igualdad entre las probabilidades observadas y esperadas. El gráfico Q-Q representa los valores observados de la variable peso tipificados frente a los valores normales esperados de una distribución N(0,1). La variable estudiada seguirá un distribución Normal si los puntos de la figura se encuentran alrededor de la línea diagonal.

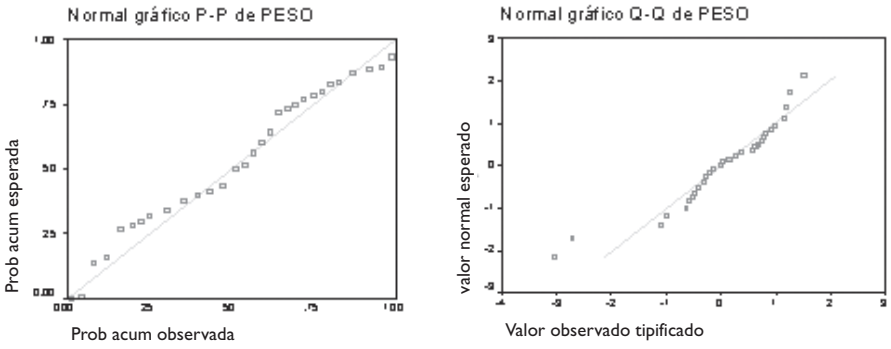

Figura 33. Gráficos de normalidad P-P y Q-Q

También existen test estadísticos para contrastar si una variable sigue una distribución normal. El test más extendido es el test de Kolmogorov-Smirnov, que consiste en comparar la función acumulada de los datos observados con la de la distribución normal, midiendo la distancia entre ambas curvas. Veamos un ejemplo de aplicación del test de Kolmogorov-Smirnov. Nuestro objetivo es contrastar si el peso de los pacientes de nuestra muestra sigue una distribución normal.

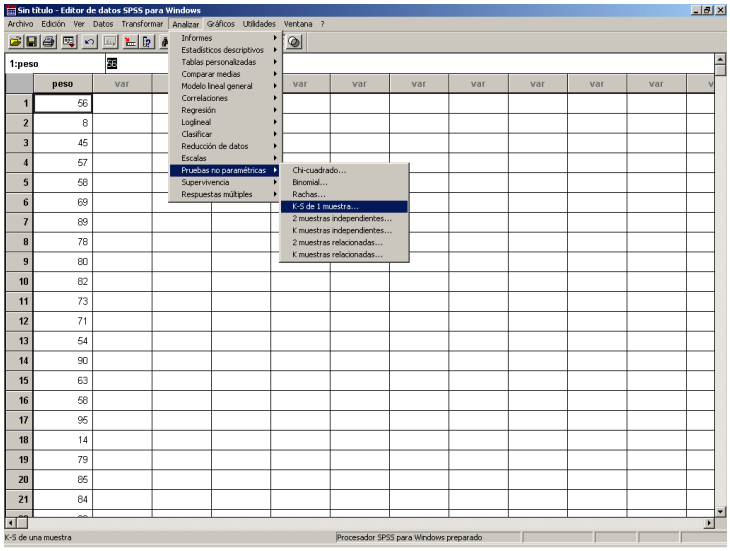

Figura 34. Prueba de normalidad de una variable mediante el test de Kolmogorov-Smirnov en SPSS

Nos adentraremos un poco en la prueba de normalidad, aunque las pruebas y los contrastes de hipótesis se desarrollarán más adelante. Como en cualquier test de hipótesis, la hipótesis nula se rechaza cuando el p-valor es inferior al nivel de significación fijado (se suele utilizar p=0,05). En nuestro caso el p-valor es de 0,515, por lo que no podemos rechazar la hipótesis nula y podemos asumir la distribución normal.

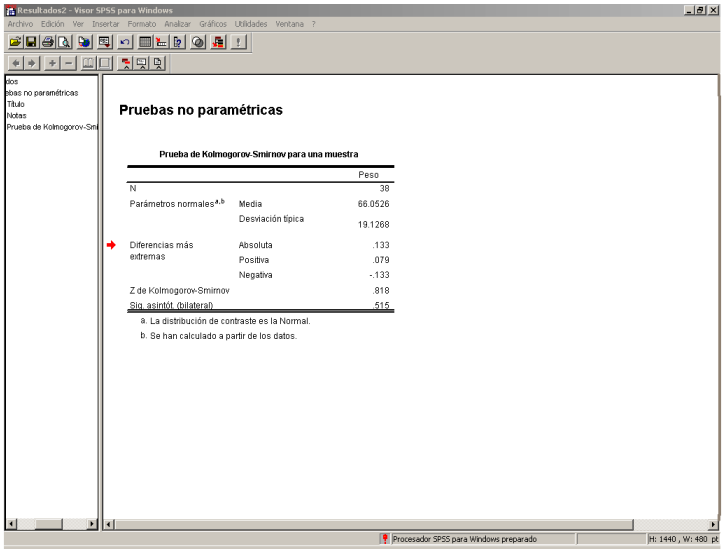

Figura 35. Resultados obtenidos en la prueba de normalidad mediante el test de Kolmogorov-Smirnov en SPSS

8.3.3. La distribución t de Student

La distribución de la t de Student es una distribución de probabilidad teórica, muy similar a la distribución normal estándar. Uno de los parámetros necesarios en la distribución normal es la desviación estándar poblacional, pero en el caso de no disponer de ella, puede utilizarse la desviación estándar muestral. En este caso ya no se trata de una distribución normal, sino que se conoce como la distribución t de Student. Esta distribución se diferencia de la distribución normal ya que en ella aparece un parámetro, llamado grados de libertad (df 'degrees of freedom'). Esto significa que para cada medida de la muestra n, en realidad tenemos una distribución diferente. La distribución t de Student con n grados de libertad, denotada como t_n, es muy parecida a la distribución normal N (0,1):

- Es simétrica respecto al 0 y se extienden desde menos infinito a más infinito.
- Cuanto mayor es el número de df, más se aproxima la distribución t de Student a la distribución normal (0,1).
- Puede considerarse aproximar la t_n por una normal estándar para n>100.

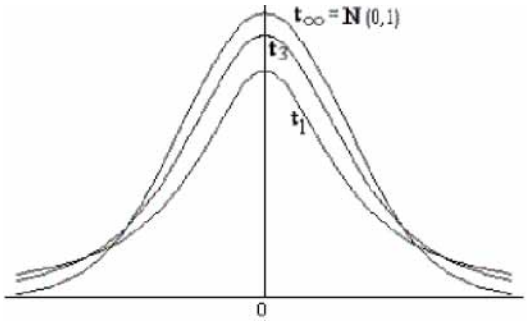

Figura 36. Distribución t de Student para diferentes grados de libertad

Generalmente, este modelo se aplica al caso de la media, proporciones y sus diferencias o sumas. Para una estimación con 30 o más grados de libertad, se pueden usar tanto el modelo de Gauss, como el de Student, aunque los intervalos obtenidos con Student son más anchos que sus equivalentes gaussianos. Por eso, se dice que el modelo Student tiene menor precisión que el de Gauss. Los casos más frecuentes en la práctica son:

Prueba	Utilidad	Cálculo del Estadístico
T de Student para medias muestrales	Comparar una variable continua con un valor de referencia. Ejemplo: ¿Es el peso medio de mi muestra igual a 70 Kg?	$t = \frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$
T de Student para proporciones	Comparar una variable proporción muestral con un valor porcentual de referencia. Ejemplo: ¿El 60% de los pacientes de mi muestra son diabéticos?	$t = \frac{(p - \pi)}{\sqrt{\pi(1-\pi)/n}}$
T de Student para dos muestras independientes	Comparar dos muestras de variables continuas aleatorias e independientes. Ejemplo: ¿Los niveles de creatinina obtenidos con el tratamiento A son iguales a los obtenidos con el tratamiento B?	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}}$

Prueba	Utilidad	Cálculo del Estadístico
T de Student para dos proporciones	Comparar dos proporciones provenientes de variables aleatorias e independientes. Ejemplo: ¿La proporción de pacientes curados es la misma para el tratamiento A y para el B?	$t = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\left(\frac{p_1(1-p_1)}{n_1}\right) + \left(\frac{p_2(1-p_2)}{n_2}\right)}}$
T de Student para dos muestras apareadas	Comparar dos muestras de variables continuas aleatorias relacionadas. Ejemplo: ¿Existen cambios en los valores de creatinina tras 3 meses de ser dializados?	$t = \frac{\bar{d}}{\sigma_d / \sqrt{n}}$

Tabla 2. Cálculo de diferentes estadísticos donde se usa la distribución t de Student

Donde: $\bar{\chi}$ y σ es la media y desviación estándar de la muestra.

μ y π es la media poblacional y la proporción poblacional de referencia , respectivamente.

$\bar{\chi}_1$ y $\bar{\chi}_2$ es la media muestral para el tratamiento A y el tratamiento B, respectivamente.

μ_1 y μ_2 es la media poblacional de referencia para el tratamiento A y B, respectivamente (generalmente 0).

σ_1 y σ_2 es la desviación estándar para el tratamiento A y el tratamiento B, respectivamente.

p es la proporción muestral

p_1 y p_2 es la proporción muestral para el tratamiento A y el tratamiento B, respectivamente.

n_1 y n_2 es el tamaño muestral para el tratamiento A y el tratamiento B, respectivamente.

π_1 y π_2 es la proporción poblacional para el tratamiento A y B, respectivamente (generalmente 0).

Los valores de probabilidad según los grados de libertad de la distribución de la t de Student, se presentan en el Anexo.

8.3.4. La distribución Ji-Cuadrado

Esta distribución de probabilidad, a diferencia de las dos anteriores, no es simétrica respecto el valor 0, sino que es asimétrica positiva, es decir, solo toma valores positivos. Como en la distribución t de Student depende de los gados de libertad. Se denota como X^2 con n grados de libertad. Esta distribución se hace más simétrica al aumentar los grados de libertad. La prueba X^2 asociada a dicha

distribución se utiliza para comparar variables de tipo ordinal o nominal, lo que es lo mismo, comparaciones de frecuencias observadas contra las frecuencias esperadas, con datos de recuento. Más adelante, se desarrolla mejor este tema, lo mismo que su uso para testear la independencia de dos o más factores en una tabla de contingencia. Los grados de libertad se calculan como (número de filas -1) x (número de columnas -1) y, a medida que aumentan los grados de libertad, tiende a una distribución normal. En el apartado 11. Comparación de proporciones veremos su aplicación.

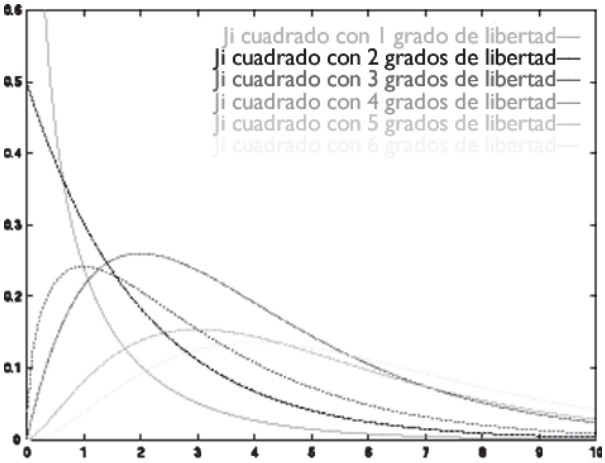

Figura 37. Distribución Ji-cuadrado para diferentes grados de libertad

8.3.5. La distribución F de Snedecor

Principalmente esta distribución de probabilidad se caracteriza por ser totalmente asimétrica y depender de dos parámetros o grados de libertad. Si de dos poblaciones normales, o aproximadamente normales, se extraen dos muestras aleatorias e independientes, y a cada una se le calcula su respectiva varianza, el cociente de ambos valores F tendrá una distribución de Fisher, cuyos valores críticos fueron obtenidos por W. Snedecor. Esta tabla se caracteriza por tener dos grados de libertad: el correspondiente al numerador n1 - 1 y el del denominador n2 - 1. La tabla de distribución F se presenta en el Anexo según los grados de libertad.

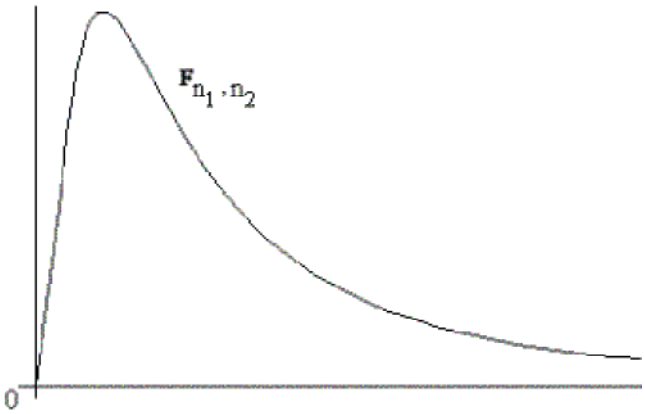

Figura 38. Distribución F de Fisher con n1 y n2 grados de libertad

El principal uso de esta función asociada a la comparación varianzas es el Análisis de Varianza, que también se verá más adelante en el capítulo 12. Comparación de medias, para cuando se necesita comparar más de dos medias muestrales a la vez. En estos casos la idea es detectar si el efecto de uno o más tratamientos afecta a las muestras testeadas. En cambio, cuando se tiene el caso de dos muestras, la idea es testear si hay homoscedasticidad (igualdad de varianzas) en las dos poblaciones en estudio. Una vez verificado este supuesto, se puede avanzar más verificando si hay diferencia entre las medias muestrales, y así verificar si ambas muestras tienen igual media y varianza, porque eso significa que en realidad provienen de la misma población normal.

8.3.6. La distribución Binomial

La distribución binomial se utiliza cuando la variable sólo tiene dos valores posibles, obtener un éxito con probabilidad p o un fracaso con probabilidad 1-p. Por tanto, en cada prueba del experimento sólo son posibles dos resultados: el suceso A (éxito) y el suceso B (fracaso). Además, la probabilidad (p) del suceso A es constante e independiente de los resultados obtenidos anteriormente, y la probabilidad del contrario del suceso A es 1-p, representado por q. El experimento consta de un número de pruebas n. Su variabilidad se estima como p x q, o lo que es lo mismo p x (1-p). Por ejemplo, si 50 de cada 1000 personas presenta una determinada enfermedad ((p=50/1000=0,05, prevalencia de la enfermedad), la distribución tendrá como parámetros p=0,05 y de variabilidad q=p x (1-p) = 0,047. La probabilidad depende de dos parámetros, n y p, y se calcula como:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

La distribución multinomial es una generalización de la distribución binomial. En vez de tener dos alternativas en cada experimento, se tienen k alternativas (k > 2). Por ejemplo, hay seis resultados posibles cuando se tira un dado. Si el "1" tiene probabilidad p_1, el "2" tiene probabilidad p_2,...,el "6" tiene probabilidad p_6, y si hacemos *n* lanzamientos independientes, los números M_1 de "1", M_2 de "2",..., M_6 de "6" constituyen un vector aleatorio *M* con una distribución multinonimal de parámetros $n,p_1,p_2,...,p_6$.

8.4. El teorema central del límite

El teorema central del límite es uno de los resultados fundamentales de la estadística. Este teorema nos dice que si una muestra es lo bastante grande (generalmente cuando el tamaño muestral (n) supera los 30), sea cual sea la distribución de la media muestral, seguirá aproximadamente una distribución normal. Es decir, dada cualquier variable aleatoria, si extraemos muestras de tamaño n (n>30) y calculamos los promedios muestrales, dichos promedios seguirán una distribución normal. Además, la media será la misma que la de la variable de interés, y la desviación estándar de la media muestral será aproximadamente el error estándar.

Un caso concreto del teorema central del límite es la distribución binomial. A partir de n=30, la distribución binomial se comporta estadísticamente como una normal, por lo que podemos aplicar los tests estadísticos apropiados para esta distribución.

La importancia del teorema central del límite radica en que mediante un conjunto de teoremas, se desvela las razones por las cuales, en muchos campos de aplicación, se encuentran en todo momento distribuciones normales o casi normales.

8.5. Consideraciones importantes

Las distribuciones de probabilidad se distinguen entre las variables discretas y las continuas, distinción que se basa en el tipo de valores que puede tomar la variable: numerable (normalmente finito) o innumerable. Entre las primeras, la más importante es la distribución binomial (particularidad de la multinomial), con un buen número de aplicaciones de carácter práctico. Y entre las segundas, la más importante es la distribución normal, a la cual se ajustan fenómenos de

carácter biológico, psicológico, económico, etc. Las distribuciones más frecuentemente utilizadas en la investigación además de la distribución normal y la binomial, son la F de Snedecor, la t de Student y la Ji-Cuadrado, entre otras.

La mayoría de valores observados sobre variables continuas a nuestro alrededor suelen aproximarse a una distribución normal. Esta es una función de distribución que ofrece un gran interés por las múltiples aplicaciones que presenta. Por ejemplo, el área bajo la curva normal está tabulado y se interpreta en términos de probabilidad, proporción o porcentaje. Los manuales de estadística suelen incluir tablas estadísticas de las distribuciones más importantes, a pesar de aparecer tanto los valores de los test, como los de su probabilidad asociada en cualquier programa de análisis estadístico que facilitan su computación e interpretación. En nuestro caso, el Anexo contiene las tablas estadíticas que se han comentado durante este capítulo.

Antes de realizar pruebas estadísticas se debería comprobar que la variable de interés procede de una distribución normal (supuesto de normalidad), para poder aplicar posteriormente pruebas paramétricas o no paramétricas.