Capítulo 5

Introducción a la inferencia estadística

1. La inferencia estadística. Población y muestra
2. Distribución muestral de un estadístico
3. La distribución de la media muestral
4. Estimación y estimadores
5. El método de los momentos
6. Diagnosis y crítica del modelo
7. El método de máxima verosimilitud

[0] Apuntes realizados por Ismael Sánchez. Universidad Carlos III de Madrid.

5.1. La inferencia estadística. Población y muestra

Como ya se ha dicho en anteriores temas, uno de los principales objetivos de la estadística es el aprendizaje a partir de la observación. En particular, la estadística proporciona el método para poder conocer cómo es el fenómeno real que ha generado los datos observados y que generará los futuros. En estadística, el interés final no está tanto en los datos observados, sino en cómo serán los próximos datos que se vayan a observar. Como ya se ha estudiado anteriormente, consideraremos que la variable que nos interesa es una variable aleatoria X, y que los datos que observamos son sólo una **muestra** (conjunto de realizaciones) procedente de dicha variable aleatoria. La variable aleatoria puede generar un número indefinido de datos. Todos los datos posibles (posiblemente infinitos) serán la **población**. Por eso, muchas veces nos referiremos de forma indistinta a la población o a la variable aleatoria que la genera.

Supongamos, por ejemplo, que queremos saber cómo son los artículos manufacturados por un determinado proceso. Para ello nos concentraremos en algún conjunto de variables medibles que sean representativas de las características de dicho artículo. Por ejemplo, la longitud de alguna de sus dimensiones podría ser una variable que nos interese conocer. La longitud de los posibles artículos manufacturados será una variable aleatoria, pues **todo proceso productivo tiene siempre variabilidad**, grande o pequeña. Las longitudes de los distintos artículos serán, en general, distintas. Diremos entonces que X = longitud de un artículo genérico, es una variable aleatoria de distribución desconocida. Para poder saber cómo es esa variable aleatoria, produciremos una muestra de artículos, y a partir de ella haremos un **ejercicio de induccción, para extrapolar las características de la muestra a toda la población**.

En estadística, al ejercicio de inducción, por el que a partir de la muestra intentamos predecir cómo será el resto de la población que no se ha observado (la variable aleatoria) se le llama **inferencia estadística**, o simplemente **inferencia**. Supondremos que para realizar este ejercicio de inferencia tenemos una conjunto de datos obtenidos al azar de entre la población de posibles datos. A una **muestra** de este tipo se le llamará **muestra aleatoria simple**. Por simplicidad, y mientras no se diga lo contrario, supondremos que las muestras que obtengamos serán muestras aleatorias simples, y por tanto nos referiremos a ellas simplemente como muestras. En una muestra aleatoria simple se tienen dos características importantes:

1. Los elementos de la muestra son independientes entre sí. Por tanto, el valor que tome uno de ellos no condicionará al de los demás. Esta independecia se puede conseguir seleccionando los elementos al azar.

2. Todos los elementos tienen las mismas características que la población.

Sea X nuestra variable aleatoria de interés, y sean $X_1, X_2, ..., X_n$ los elementos de una muestra de tamaño n de dicha variable aleatoria X. Entonces, antes de ver los valores concretos que tomará la muestra formada por $X_1, X_2, ..., X_n$, tendremos que **la muesta $X_1, X_2, ..., X_n$ será un conjunto de variables aleatorias independientes e idénticas a X.**

5.2. Distribución muestral de un estadístico

Volvamos a nuestro ejemplo del proceso productivo, en el que estamos interesados en saber cómo es la variable aleatoria X =longitud de un artículo genérico. Estamos interesados en las características de la población, es decir, en todas las longitudes posibles. Supongamos que tenemos una muestra de $n = 100$ artículos y hemos medido sus longitudes. Supongamos que calculamos un

conjunto de medidas características de dicha muestra de 100 longitudes: la media, la varianza, etc. ¿Son dichos valores los de la población? Está claro que no. La media de las $n = 100$ longitudes no tiene por qué coincidir con la media de toda la población. La media de la población será el promedio de TODOS sus infinitos datos, y sólo por casualidad será igual al promedio de los $n = 100$ datos que hemos seleccionado al azar. El problema es que si no tenemos todos los infinitos datos, no podremos conocer las medidas características de la población. Sólo tendremos lo que podamos obtener de la muestra. Por tanto, una primera conclusión que podemos extraer al intentar predecir (inferir) las medidas características de una población utilizando una muestra es la siguiente:

Conclusión 1: Los valores de las medidas características que se obtienen de una muestra serán sólo una aproximación de los valores de las medidas características de la población.

Otra segunda limitación de intentar averiguar cómo es una población a partir de la información de una muestra, es que nuestras conclusiones dependen de la muestra concreta que hayamos obtenido. Con otras muestras tendríamos otros valores en nuestras medidas características y podríamos sacar conclusiones diferentes sobre la población. Supongamos que tomamos una muestra de $n = 100$ artículos y medimos esas 100 longitudes obteniendo una longitud media de $\bar{X} = 23{,}5$ cm ¿Quiere esto decir que cada vez que tomemos una muestra al azar de 100 de dichos artículos y los midamos, su media muestral será siempre $\bar{X} = 23{,}5$ cm? Pues obviamente no. Será mucha casualidad que dos muestras distintas nos den exáctamente la misma media. Como son muestras de una misma población, las medias será más o menos similares, pero no tienen por qué coincidir. Por tanto, se puede concluir los siguiente:

Conclusión 2: Los valores de las medidas características que se obtienen de una muestra dependen de los elementos que la constituyen. Muestras diferentes darán por tanto valores diferentes.

Como los elementos han sido seleccionados al azar, se dice que **los valores de las medidas características de una muestra dependen del azar del muestreo**. En lugar de hablar de medidas características de una muestra, hablaremos de forma más general de operaciones matemáticas realizadas con la muestra. Vamos a introducir entonces un nuevo concepto: llamaremos **estadístico** a cualquier operación realizada con una muestra. Por ejemplo, la media muestral es un estadístico, así como la varianza, el rango, o cualquier otra medida característica. **El estadístico es la operación matemática, no el resultado obtenido**. El estadístico tomará, de acuerdo a la conclusión 2 anterior, un valor diferente en cada muestra. **Un estadístico será, por tanto, una variable aleatoria** pues el valor concreto que obtengamos dependerá del azar del muestreo. El resultado obtenido con una muestra concreta será una **realización** de dicha variable aleatoria. **Cada vez que realicemos el experimento de computar el valor de un estadístico en una muestra diferente, obtendremos una realización diferente del estadístico.** En la práctica, tendremos una sola muestra, y por tanto una sola realización del estadístico.

Si queremos dar alguna interpretación al valor de la realización de un estadístico en una muestra concreta, necesitamos conocer cómo varía el valor que puede tomar el estadístico de unas muestras a otras (ese será uno de los objetivos de este tema). Por ejemplo, supongamos el caso de las longitudes de los artículos antes mencionados. Supongamos también que sabemos por información histórica que si el proceso productivo funciona adecuadamente, la longitud media (μ) de los artículos que produce debe ser de μ =25 cm. ¿Cómo podemos entonces interpretar que en una muestra de tamaño $n = 100$ se haya obtenido que $\bar{X} = 23{,}5$ cm? ¿Es esa diferencia (25-23.5) evidencia de

que la media se ha reducido y por eso la muestra tiene una media menor? ¿O la media no ha cambiado y esa discrepancia (25-23.5) puede atribuirse a la variabilidad debida al muestreo? Si no conocemos la función de densidad de $\bar{X}$ no podremos valorar si 23.5 es un número muy alejado de 25, y por tanto sospechoso de que algo ha ocurrido en el proceso; o por el contrario que sea muy frecuente que muestras de una población de media $\mu = 25$ den alejamientos tan grandes como 23.5. Supongamos ahora que extraemos una segunda muestra de tamaño $n = 100$ tras haber realizado algunos ajustes a la máquina y que obtenemos que la nueva media muestral es $\bar{X}_{(2)} = 24$. ¿Quiere decir que los ajustes han provocado un aumento en la media (de 23.5 a 24)?¿o ese cambio (23.5-24) puede explicarse simplente por ser muestras diferentes de una misma población de media 25?.

El tipo de preguntas que se plantean en el párrafo anterior son muy importantes en la práctica, y serán las que querramos resolver con la estadística. Para responder a este tipo de preguntas, es necesario conocer las características del estadístico que nos interese (en el caso del ejemplo, la media muestral). A la distribución de un estadístico debido a la variabilidad de la muestra se le denomina, **distribución del estadístico en el muestreo**. Esta distribución dependerá de cada caso concreto.

Es importante darse cuenta de que estamos manejando dos niveles de variables aleatorias. En un primer nivel, más superficial, estaría nuestra variable aleatoria de interés X. En nuestro ejemplo sería X = longitud de un artículo genérico de nuestro proceso productivo. Para conocer las propiedades de dicha variable aleatoria extraemos una muestra aleatoria simple de tamaño n. Las longitudes de esos n elementos serán $X_1, X_2, ..., X_n$. Como hemos dicho anteriormente, antes de extraer la muestra, no sabremos qué valores tomarán $X_1, X_2, ..., X_n$. Y al tratarse de una muestra aleatoria simple, estos n elementos pueden interpretarse como un conjunto de n variables aleatorias independientes, e idénticas a nuestra variable de interés X. Nuestro objetivo es utilizar la muestra para saber cómo es X. Para ello, computamos el valor de un conjunto de estadísticos de interés con los datos de la muestra: $\bar{X}, S_x^2$, etc. Esos estadísticos constituirán, debido a la variabilidad del muestreo, un segundo nivel de variables aleatorias con características diferentes a X. Para fijar mejor estos conceptos, en la sección siguiente veremos el caso de la media muestral.

5.3. La distribución de la media muestral

Supongamos que tenemos una muestra de tamaño n de una variable aleatoria X. Los elementos de dicha mestra serán serán $X_1, X_2, ..., X_n$. La media muestral de esas n observaciones es el siguiente estadístico:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{5.1}$$

Como cada X_i es una variable aleatoria (idéntica a X) tendremos que $\bar{X}$ es también una variable aleatoria. De esta manera, su valor será, en general, diferente en cada muestra. El objetivo de esta sección es analizar cómo es la distribución de $\bar{X}$ en el muestreo.

En primer lugar calcularemos la esperanza matemática de $\bar{X}$. Si llamamos $E(X) = \mu$ tendremos que $E(X_i) = \mu$, $i = 1, 2, ..., n$. Entonces

$$E(\bar{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{E(X_1) + E(X_2) + \cdots + E(X_n)}{n} = \frac{n\mu}{n} = \mu. \tag{5.2}$$

Por tanto, aunque $\bar{X}$ pueda variar de unas muestras a otras, por término medio proporciona el valor de la media poblacional, que es al fin y al cabo nuestro objetivo. Este resultado es muy

importante, pues nos dice que de los posibles valores que podamos obtener al cambiar la muestra, el centro de gravedad de ellos es la media poblacional.

Para ver la dispersión de los distintos valores de medias muestrales alrededor de μ, calcularemos la varianza de $\bar{X}$. Llamaremos $\text{Var}(X) = \sigma^2$. Por tanto $\text{Var}(X_i) = \sigma^2$, $i = 1, 2, ..., n$. Entonces,

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{\text{Var}\left(X_1 + X_2 + \cdots + X_n\right)}{n^2}.$$

Al ser una muestra aleatoria simple, los X_i serán variables aleatorias independientes, y por tanto tendremos que

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n} = \frac{\sigma^2}{n}. \tag{5.3}$$

De este resultado se pueden sacar dos conclusiones que son también importantes:

1. La varianza disminuye con el tamaño muestral. Por tanto cuantos más datos se tengan será más probable que la media muestral sea un valor próximoa μ.

2. La dispersión de $\bar{X}$ alrededor de μ depende de la dispersión de la variable original X. Así, si σ^2 es muy grande hará falta un tamaño muestral grande si queremos asegurarnos que la media muestral esté cerca de la poblacional.

Finalmente, vemos que la media muestral puede escribirse como suma de variables aleatorias. Reescribiendo (5.1) tenemos que

$$\bar{X} = \frac{X_1}{n} + \frac{X_2}{n} + \cdots + \frac{X_n}{n},$$

y por tanto, por el **Teorema Central del Límite** (ver tema anterior) tenemos que, independientemente de cómo sea X, **si el tamaño muestral n es suficientemente grande $\bar{X}$ será aproximadamente una distribución normal**. Se tiene por tanto que **si n es grande** (en la práctica, con más de 30 datos)

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{5.4}$$

Por tanto, la media muestral realizada con un número suficiente de datos es una variable aleatoria simétrica y muy concentrada alrededor de la media poblacional, independientemente de cómo sea la naturaleza de X. De esta forma, la media muestral con un tamaño muestral suficientemente grande proporciona una forma bastante precisa de aproximar la media poblacional μ. No olvidemos que en la práctica tendremos una sola muestra, y por tanto una sola realización de $\bar{X}$. Por esta razón, es muy útil disponer de resultados teóricos tan interesantes como (5.4), que nos ayuden a valorar la fiabilidad de la media muestral. Nótese que el resultado (5.4) es independiente de la distribución que siga X si n es grande. En los próximos temas obtendremos más conclusiones y construiremos procedimientos estadísticos basados en este resultado.

5.4. Estimación y estimadores

Lo que hemos hecho en la sección anterior: utilizar la media muestral para asignar un valor aproximado a la media poblacional se denomina **estimación**. En general, la **estimación** consiste

en asignar valores numéricos a aquellos parámetros poblacionales que no conozcamos. Un ejemplo de parámetros poblacionales desconocidos serían las medidas características de poblaciones de las que sólo observamos una muestra. La estimación se realiza mediante la utilización de un estadístico, que evaluamos con una muestra de datos. **Llamaremos estimador al estadístico que se usa para estimar un parámetro**.

Por ejemplo, supongamos que después de analizar una muestra de longitudes de n artículos hemos concluido que la distribución normal es un modelo de probabilidad adecuado para describir las longitudes de las piezas que salgan de dicho proceso productivo, es decir $X \sim N(\mu, \sigma^2)$. El problema entonces estará en **asignar valores a los parámetros desconocidos** μ, y σ^2 de dicha distribución. El problema de estimación no solo aparece cuando buscamos un modelo de probabilidad. En algunos casos sólo estamos interesados en algunas medidas características poblacionales: media, varianza, etc. En ambos casos, tanto si estamos interesados en encontrar un modelo de probabilidad o sólo algunas medidas características, tenemos que utilizar una muestra de datos para **estimar** un parámetro poblacional desconocido.

Un parámetro es siempre una propiedad de la población, y por tanto será una constante de valor desconocido. El valor del parámetro λ de una Poisson o una exponencial, o el parámetro p de una binomial, etc sólo se podrían conocer con exactitud si tuviésemos acceso a toda la población; es decir, si repitiésemos el experimento aleatorio indefinidamente, lo que en la práctica es imposible.

Para denotar a un estimador usaremos la misma letra que el parámetro pero con el acento circumflejo ($\hat{}$) encima. Para el caso de la media poblacional μ, un estimador será denotado por $\hat{\mu}$. Si usamos la media muestral como estimador de la media tendremos entonces que

$$\hat{\mu} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

El estimador $\hat{\mu} = \bar{X}$ es, como cualquier otro estimador que hubiésemos seleccionado, una variable aleatoria, pues su valor cambia de unas muestras a otras. Vemos entonces que al estudiar una variable aleatoria X nos topamos con otra nueva variable aleatoria $\hat{\mu}$ que será necesario analizar. Las propiedades estadísticas más importantes de un estimador cualquiera son su esperanza matemática y su varianza. Si para estimar un parámetro tuviésemos que elegir entre varios estimadores alternativos, elegiríamos aquél que tuviese mejores propiedades estadísticas. A continuación vamos a enunciar qué propiedades estadísticas de un estimador conviene tener en cuenta.

Supongamos que queremos estimar un parámetro θ y usamos un estimador $\hat{\theta}$ que consiste en cierta operación matemática realizada con una muestra. Las características más importantes de $\hat{\theta}$ son $E(\hat{\theta})$ y $\mathrm{Var}(\hat{\theta})$. En general serán preferibles aquellos estimadores que verifiquen que $E(\hat{\theta}) = \theta$. A los estimadores que verifican esta propiedad se les denomina **insesgados** o **centrados**. Si un estimador verifica que $E(\hat{\theta}) \neq \theta$ se dice que el estimador es **sesgado**. El **sesgo** de un estimador viene entonces definido por

$$\mathrm{sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Por ejemplo, vimos en (5.2) que la media mestral es una variable aleatoria de media μ. Por lo tanto **la media mestral es un estimador insesgado de la media poblacional**. En lo que respecta a la varianza de la variable aleatoria $\hat{\theta}$, diremos que, en general serán preferibles aquellos estimadores que tengan menor varianza, pues será, más precisos en el sentido de que variarán poco de unas muestras a otras.

A la desviación típica de un estimador se le suele denominar **error estándar** del estimador. Por ejemplo, en el caso de la media muestral como estimador de la media poblacional, veíamos en la sección anterior que $\mathrm{Var}(\hat{\mu}) = \sigma^2/n$. Por tanto, el error estándar de $\hat{\mu}$ es $\sigma/\sqrt{n}$.

Vemos entonces que es preferible un estimador insesgado a otro sesgado, y un estimador con poco error estándar a otro con mayor error estándar. ¿Y si dos estimador tiene menos sesgo que otro, pero más varianza?¿Cómo elegimos el mejor? Para estos casos, definiremos un crtiterio que tenga en cuenta tanto el sesgo como la varianza. Definiremos error cuadrático medio de un estimador $\text{ECM}(\hat{\theta})$ a

$$\text{ECM}(\hat{\theta}) = \left[\text{sesgo}(\hat{\theta})\right]^2 + \text{Var}\left(\hat{\theta}\right).$$

En general, entre un conjunto de estimadores alternativos para un mismo parámetro, elegiremos aquél que tenga menor ECM. Al estimador que tiene menor ECM le diremos que es **eficiente**.

Ejemplo 1 *En muestras aleatorias simples de tamaño $n = 3$ de una variable aleatoria normal de media μ y varianza conocida $\sigma^2 = 1$, se consideran los siguientes estimadores de μ:*

$$\begin{aligned} U_1 &= \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3, \\ U_2 &= \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3, \\ U_3 &= \frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3, \end{aligned}$$

donde X_1, X_2, X_3 son las observaciones. Comprobar que son estimadores insesgados y estudiar su error cuadrático medio.

SOLUCIÓN:

$$\begin{aligned} E(U_1) &= \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \frac{1}{3}3\mu = \mu \\ E(U_2) &= \frac{1}{4}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{4}E(X_3) = \frac{1}{4}\mu + \frac{1}{2}\mu + \frac{1}{4}\mu = \mu \\ E(U_3) &= \frac{1}{8}E(X_1) + \frac{3}{8}E(X_2) + \frac{1}{2}E(X_3) = \frac{1}{8}\mu + \frac{3}{8}\mu + \frac{1}{2}\mu = \mu \end{aligned}$$

Al ser centrados, el ECM es la varianza.

$$\begin{aligned} Var(U_1) &= Var(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3) = \frac{1}{9}Var(X_1) + \frac{1}{9}Var(X_2) + \frac{1}{9}Var(X_3) = \frac{1}{3}\sigma^2 = \frac{1}{3} = 0{,}333 \\ Var(U_2) &= Var(\frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3) = \frac{1}{16}Var(X_1) + \frac{1}{4}Var(X_2) + \frac{1}{16}Var(X_3) = \frac{3}{8}\sigma^2 = \frac{3}{8} = 0{,}375 \\ Var(U_3) &= Var(\frac{1}{8}X_1 + \frac{3}{8}X_2 + \frac{1}{2}X_3) = \frac{1}{64}Var(X_1) + \frac{9}{64}Var(X_2) + \frac{1}{4}Var(X_3) = \frac{26}{64}\sigma^2 = \frac{13}{32} = 0{,}406 \end{aligned}$$

luego el más eficiente es U_1.

5.5. El método de los momentos

Existen varios procedimientos para construir estimadores. El más sencillo sea tal vez el método de los momentos. El método de los momentos consiste básicamente en estimar una característica de la población con la respectiva característica muestral. Así, para estimar la media poblacional μ, el método de los momentos consistiría en utilizar la media muetral $\bar{X}$. De la misma forma, el estimador de la varianza poblacional σ^2 por el método de los momentos será la varianza de la

muestra s^2. Si lo que queremos es estimar una probabilidad p de que se observe un suceso en una población, usaremos la proporción muestral, es decir, la proporción de veces que se ha observado el suceso en la muestra analizada.

El nombre de método de los momentos viene de que en estadística se denomina **momento de una población de orden** m **a** $E(X^m)$, mientras que **momento muestral de orden** m es $(1/n)\sum_{i=1}^{n} X_i^m$. Por tanto el primer momento poblacional es $E(X) = \mu$ y el primer momento muestral $\bar{X}$. El segundo momento poblacional será $E(X^2)$ y el segundo momento muestral será $\sum_{i=1}^{n} X_i^2 / n$.

Asímismo, se denomina **momento poblacional de orden** m **respecto a la media** a $E\left[(X-\mu)^m\right]$ y el momento muestral de orden m respecto a la media es $(1/n)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^m$. La media es el momento de primer orden. Por tanto la varianza poblacional σ^2 es el momento de segundo orden respecto a la media, y la varianza muestral s^2 es el segundo momento muestral respecto a la media.

Ejemplo 2 *Sean $t_1, t_2, ..., t_n$ el tiempo de atención a n clientes en cierto puesto de servicio. Si el tiempo de atención es una variable aleatoria exponencial de parámetro λ, estima este parámetro por el método de los momentos.*

SOLUCIÓN:

Como en una exponencial se tiene que

$$E(T) = \frac{1}{\lambda} \Rightarrow \lambda = \frac{1}{E(T)},$$

Entonces, estimaremos $E(T)$ con la media muestral, es decir:

$$\bar{T} = \frac{\sum_{i=1}^{n} t_i}{n},$$

y por tanto el estimador de λ por el método de los momentos será:

$$\hat{\lambda} = \frac{1}{\bar{T}}$$

Ejemplo 3 *Se desea estimar la proporción de artículos defectuosos que produce una máquina. Para ello se analizan n artículos, resultando que d son defectuosos. Estima p por el método de los momentos. Supondremos que la aparición de artículos defectuosos sigue un proceso de Bernoulli.*

SOLUCIÓN:

Si llamamos X a la variable de Bernoulli que vale 1 si el artículo es defectuoso y 0 si es aceptable, entonces

$$E(X) = p$$

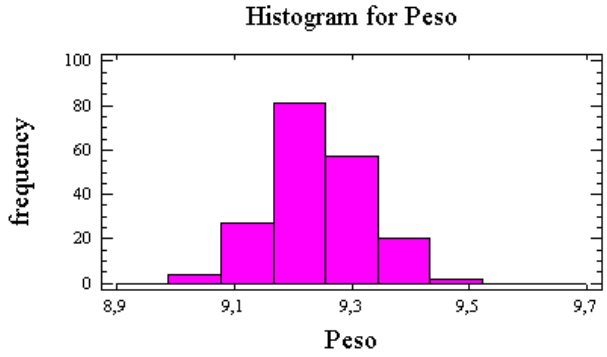
Estimaremos p mediante el promedio de las variables de Bernoullis de los n artículos analizados. Por tanto, si $X_i = 1$ si el artículo i-ésimo es defectuosos y 0 si es aceptable,

$$\begin{aligned} \hat{p} &= \frac{\sum_{i=1}^{n} X_i}{n} = \frac{d}{n} \\ &= \frac{\textit{número de artículos defectuosos}}{\textit{número de artículos analizados}}, \end{aligned}$$

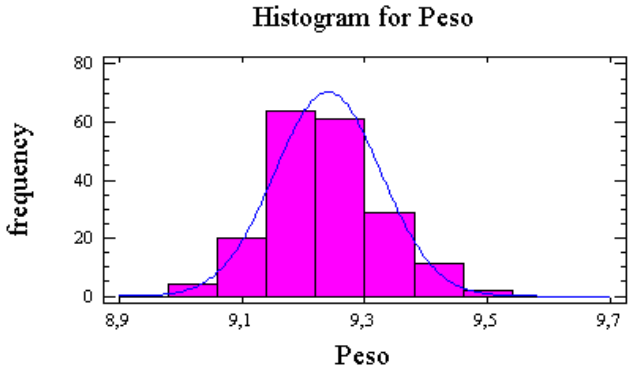
por lo que el estimador de la proporción poblacional es la proporción muestral.

5.6. Diagnosis del modelo

Dado un conjunto de datos $x_1, ..., x_n$, obtenidos al extraer una muestra aleatoria simple de una población, queremos **inferir** qué distribución sigue dicha población. Es decir, queremos saber si siguen una distribución normal, o una exponencial o cualquier otro modelo en el que estemos interesados. Hay muchos procedimientos para hacer ese ejercicio de inferencia. El más utilizado consiste en evaluar la similitud del histograma de los datos con la función de densidad o de distribución seleccionada. Por ejemplo, el siguiente histograma muestra el peso de 191 monedas de 100 pesetas recogidas justo antes de su desaparición y sustitución por el euro (fichero *monedas100.sf3*). El histograma sugiere que el peso de las monedas de 100 pesetas puede ser una normal. Mejor dicho, que la normal es un modelo de distribución adecuado para modelizar la distribución de probabilidades de (todas) las monedas de 100 pesetas. En este caso diremos que los datos se **ajustan** a la distribución elegida.



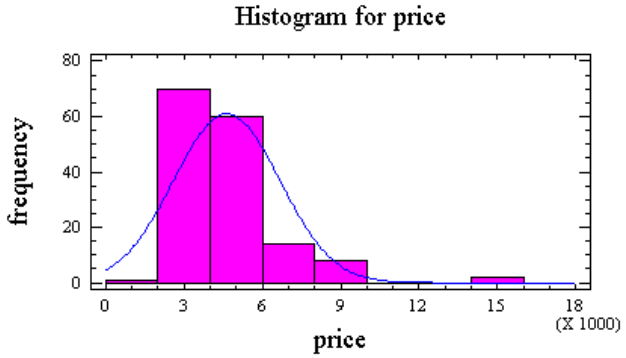
Las medidas características de este conjunto de datos son $\bar{x} = 9{,}23$, $s^2 = 0{,}0075$. Estas medidas características las emplearemos como estimadores de los parámetros poblacionales μ y σ^2. Si superponemos al histograma anterior la función de densidad de una $N(9{,}23; 0{,}0075)$ tenemos el siguiente gráfico.



Lo que vamos a hacer a continuación es una prueba estadística que compare esa curva de la normal con el histograma. Si hay poca discrepancia entre la probabilidad que dice el modelo y la

frecuencia relativa observada en los datos en todos y cada uno de los intervalos, podremos concluir que es muy plausible que dicho histograma pueda proceder de la población representada por el modelo. O bien, que podemos adoptar dicho modelo para representar a la población de la cual proceden nuestros datos.

Existen muchas pruebas estadísticas que comparan los datos observados con modelos de probabilidad. La prueba estadística más popular es **el test la chi-cuadrado**. A este tipo de pruebas se le denomina **pruebas de bondad de ajuste**, pues lo que hacemos es medir **cómo se ajustan los datos al patrón que marca el modelo elegido**. Puede verse que en esta prueba estadística estamos haciendo un ejercicio de **inducción** o **inferencia**, pues vamos de lo particular -los datos observados- a lo general -el modelo que ha generado no solo a esos datos sino al resto de datos que no hemos tomado-. Todas las pruebas de este tipo se basan en algún tipo de comparación de los datos con el modelo elegido. Si la discrepancia es grande se rechaza dicho modelo, si la discrepancia no es grande no se rechaza el modelo, y se considera que es una buena representación de la población que generó los datos. La siguiente figura muestra un conjunto de datos donde la discrepancia con la normal es mayor que en el ejemplo de las monedas de 100 pesetas. Los datos corresponden al precio de unos vehículos (según el fichero *cardata.sf*).



A continuación comentaremos brevemente cómo se realiza la prueba de la chi-cuadrado, aunque en la práctica la haremos siempre con el ordenador. Supondremos que tenemos siempre un mínimo de unos 25 datos, de lo contrario, la prueba es poco fiable. Para hacer el contraste chi-cuadrado se siguen los siguientes pasos:

1. Se hace el histograma, usando más de 5 clases con al menos 3 datos en cada clase.

2. En cada clase del histograma obtenemos la frecuencia (absoluta) observada de individuos, que denotaremos por O_i , $i = 1, 2, ..., k$, siendo k el número de clases del histograma.

3. Con los datos estimamos los parámetros del modelo seleccionado, y calculamos con dicho modelo la probabilidad p_i de obtener valores en cada una de las clases del histograma. Llamaremos $E_i = np_i$ a la frecuencia absoluta esperada de acuerdo al modelo seleccionado O_i

4. Calculamos el siguiente estadístico

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

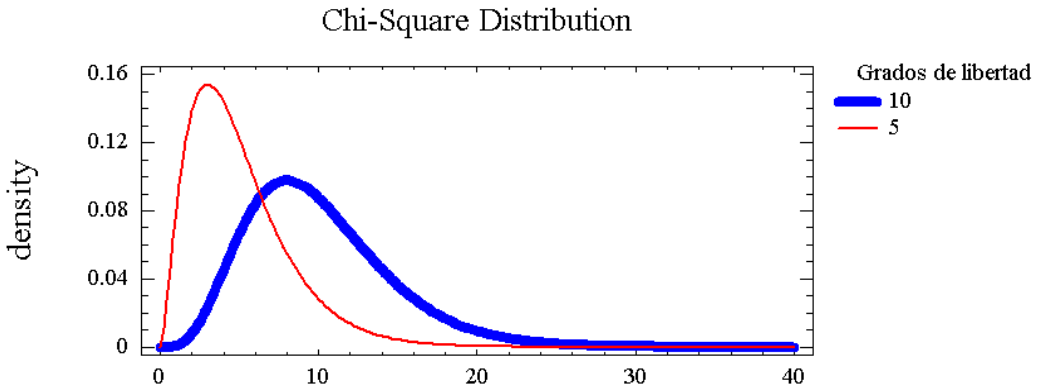
Este estadístico X_0^2 resume toda la discrepancia entre el histograma y el modelo.

5. Si X_0^2 es un número elevado rechazaríamos el modelo, mientras que en caso contrario aceptaríamos el modelo como adecuado para representar a la población que genera los datos. Para valorar X_0^2 es necesario alguna referencia que nos diga cuándo es grande y cuándo no. La obtención de esta referencia se explica a continuación.

La valoración del estadístico X_0^2 se basa en el siguiente resultado:

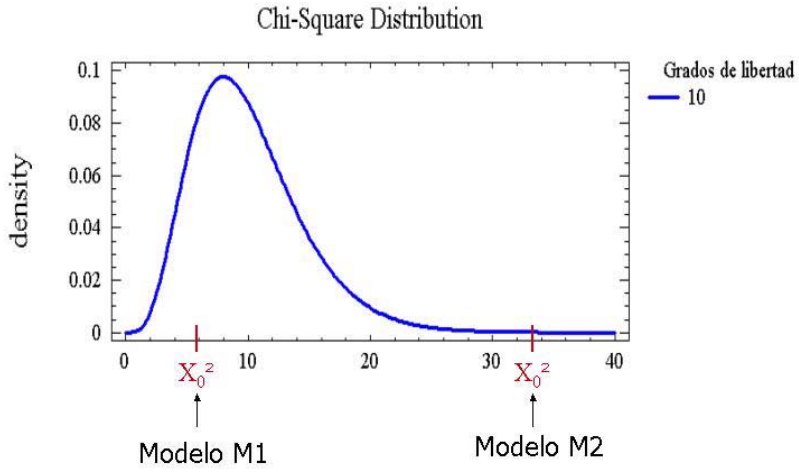
Si los datos pertenecen a una población que sigue el modelo de probabilidad elegido, el estadístico X_0^2 será una variable aleatoria que sigue una distribución denominada Chi-cuadrado, y que se simboliza por χ_g^2, donde g es un parámetro que se denomina grados de libertad. Los grados de libertad son $g = k - v - 1$, donde v es el número de parámetros que hemos estimado en ese modelo. Por el conrario, si la distribución no es adecuada, el estadístico X_0^2 no seguirá dicha distribución, y podrá tomar un valor muy grande con mucha probabilidad

La distribución χ_g^2 es una distribución asimétrica de valores positivos. Es una distribución que se encuentra tabulada en muchos textos de estadística. La siguiente figura muestra dos ejemplos de distribuciones χ_g^2, de 5 y 10 grados de libertad respectivamente.



Consideraremos que el modelo **no** es adecuado para nuestros datos, si X_0^2 está en la zona de la derecha de la distribución, donde la probabilidad de que la distribución χ_g^2 genere valores en esa zona es ya muy pequeña. Por tanto, si X_0^2 está en la zona de la cola de la derecha, será señal de que el modelo elegido no es adecuado. La siguiente figura ilustra esta idea. En esta figura se muestran dos valores de X_0^2 obtenidos al proponer dos modelos diferentes, M1 y M2, a un mismo conjunto de datos. El valor de X_0^2 más pequeño nos llevaría a concluir que el modelo M1 elegido es adecuado, tiene un buen ajuste a los datos; mientras que el mayor nos llevaría a concluir que el

otro modelo, el M2 no se ajusta bien a los datos.



Los programas informáticos proporcionan el área que queda a la derecha de X_0^2 en la distribución de referencia. Ese área recibe el nombre de **p-valor**. De esta forma, **cuanto menor sea el p-valor, peor es el modelo** elegido para representar a nuestra población, pues indicará que X_0^2 está muy a la derecha de la distribución. En general, **rechazaremos un modelo si p-valor<0.05**. Veamos algunos ejemplos.

Ejemplo 4 *El fichero* monedas100.sf3 *contiene el peso de 191 monedas de 100 pesetas. El histograma del peso de estas monedas se ha mostrado anteriormene, sugiriendo que la curva normal puede ser un buen modelo para representar a la población de pesos de esas monedas. Una práctica común en la aplicación del test de la chi-cuadrado es hacer clases de tal manera que la frecuencia esperada E_i en cada clase sea la misma. Para ello, las clases deben ser de longitud diferente. De esta manera se evita que algunas clases tengan pocas observaciones, lo que empeora la fiabilidad de esta prueba estadística. Por tanto, el histograma que nos muestre el programa informático, basado en clases iguales, no coincidirá con el histograma que utilice para hacer el test. A continuación se*

muestra el resultado del test que realiza el Statgraphics:

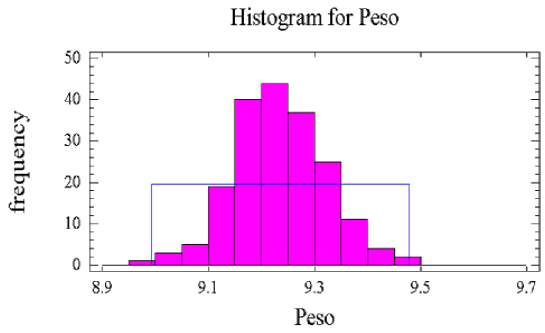
Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	9.10673	9	11.94	0.72
	9.10673	9.13994	11	11.94	0.07
	9.13994	9.16271	11	11.94	0.07
	9.16271	9.18111	22	11.94	8.48
	9.18111	9.19718	7	11.94	2.04
	9.19718	9.21191	16	11.94	1.38
	9.21191	9.22586	12	11.94	0.00
	9.22586	9.23948	12	11.94	0.00
	9.23948	9.25309	12	11.94	0.00
	9.25309	9.26705	10	11.94	0.31
	9.26705	9.28177	13	11.94	0.09
	9.28177	9.29784	7	11.94	2.04
	9.29784	9.31624	11	11.94	0.07
	9.31624	9.33901	13	11.94	0.09
	9.33901	9.37222	12	11.94	0.00
above	9.37222		13	11.94	0.09

Chi-Square = 15.4922 with 13 d.f. P-Value = 0.277639

Puede verse que el p-valor es bastante mayor que 0.05. Por tanto, consideramos que la normal es un modelo adecuado para modelizar los pesos de las monedas de 100 pesetas.

A continuación vamos a intentar ajustar un modelo que sea claramente inadecuado. Si elegimos la distribución uniforme tenemos la siguiente figura:



donde puede verse que el modelo elegido no se ajusta a los datos. El test de la chi-cuadrado que

resulta es

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	9.02063	3	11.94	6.69
	9.02063	9.05125	1	11.94	10.02
	9.05125	9.08188	0	11.94	11.94
	9.08188	9.1125	8	11.94	1.30
	9.1125	9.14312	12	11.94	0.00
	9.14312	9.17375	18	11.94	3.08
	9.17375	9.20438	26	11.94	16.57
	9.20438	9.235	32	11.94	33.72
	9.235	9.26563	22	11.94	8.48
	9.26563	9.29625	20	11.94	5.45
	9.29625	9.32688	16	11.94	1.38
	9.32688	9.3575	16	11.94	1.38
	9.3575	9.38813	7	11.94	2.04
	9.38813	9.41875	5	11.94	4.03
	9.41875	9.44937	3	11.94	6.69
above	9.44937		2	11.94	8.27

Chi-Square = 121.042 with 13 d.f. P-Value = 0.0

que muestra un p-valor=0.0, por lo que resulta evidente que la distribución uniforme no es un modelo adecuado para este tipo de datos

Ejemplo 5 *En este ejemplo vamos a ajustar en primer lugar una distribución normal a la variable precio, del fichero* cardata.sf. *Más arriba se mostró el histograma con la curva normal superpuesta. Dicha curva normal es la que se obtiene al usar la media muestral y la varianza muestral de los datos. El test de la chi-cuadrado da el siguiente resultado:*

Chi-Square Test

	Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
	at or below	1588.53	0	10.33	10.33
	1588.53	2380.17	9	10.33	0.17
	2380.17	2926.07	19	10.33	7.27
	2926.07	3369.63	17	10.33	4.30
	3369.63	3759.45	14	10.33	1.30
	3759.45	4119.22	17	10.33	4.30
	4119.22	4463.4	5	10.33	2.75
	4463.4	4802.73	12	10.33	0.27
	4802.73	5146.91	12	10.33	0.27
	5146.91	5506.68	13	10.33	0.69
	5506.68	5896.5	10	10.33	0.01
	5896.5	6340.06	6	10.33	1.82
	6340.06	6885.96	4	10.33	3.88
	6885.96	7677.6	5	10.33	2.75
above	7677.6		12	10.33	0.27

Chi-Square = 40.3869 with 12 d.f. P-Value = 0.000062011

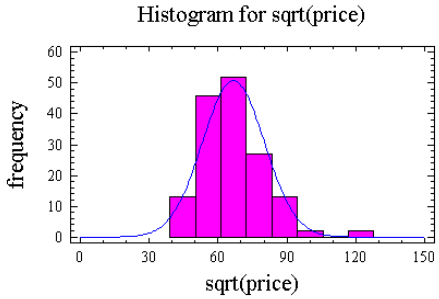
El p-valor es muy pequeño, por lo que se confirma lo que sugería el histograma con la distribución normal: la normal no es una buena representación para esta variable. No obstante vemos que la característica de estos datos es su asimetría positiva. Podemos intentar alguna transformación que haga la distribución de los datos más simétrica (ver Tema 1), y tal vez entonces la variable transformada sí sea normal. A continuación se muestra el resultado de hacer la transformación

$\sqrt{X}$.

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	46.4234	2	10.33	6.72
46.4234	51.7007	14	10.33	1.30
51.7007	55.3398	16	10.33	3.11
55.3398	58.2967	13	10.33	0.69
58.2967	60.8954	14	10.33	1.30
60.8954	63.2937	12	10.33	0.27
63.2937	65.588	7	10.33	1.08
65.588	67.8501	8	10.33	0.53
67.8501	70.1445	12	10.33	0.27
70.1445	72.5428	13	10.33	0.69
72.5428	75.1415	9	10.33	0.17
75.1415	78.0984	12	10.33	0.27
78.0984	81.7374	5	10.33	2.75
81.7374	87.0147	6	10.33	1.82
above 87.0147		12	10.33	0.27

Chi-Square = 21.2264 with 12 d.f. P-Value = 0.0471625

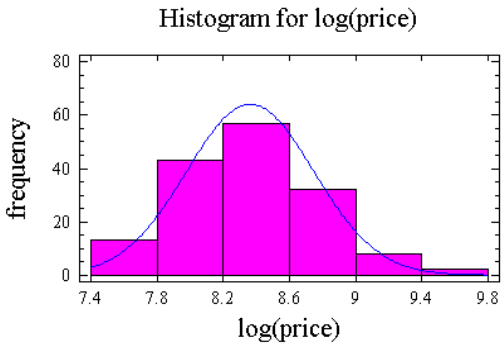


Esta transformación ha conseguido una mayor simetría, pero aún se nota una mayor cola hacia la derecha. El test de la chi-cuadrado aún tiene un p-valor muy pequeño. El ajuste a la normal de $\sqrt{X}$ *no es adecuado. Probaremos entonces otra transformación que sea algo más fuerte que la raíz cuadrada. La figura siguiente muestra el resultado de aplicar la transformación* $\ln(X)$.

Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	7.7837	11	10.33	0.04
7.7837	7.93434	14	10.33	1.30
7.93434	8.03822	7	10.33	1.08
8.03822	8.12262	13	10.33	0.69
8.12262	8.1968	11	10.33	0.04
8.1968	8.26526	7	10.33	1.08
8.26526	8.33075	13	10.33	0.69
8.33075	8.39532	5	10.33	2.75
8.39532	8.46081	12	10.33	0.27
8.46081	8.52927	8	10.33	0.53
8.52927	8.60345	14	10.33	1.30
8.60345	8.68785	14	10.33	1.30
8.68785	8.79173	7	10.33	1.08
8.79173	8.94237	7	10.33	1.08
above 8.94237		12	10.33	0.27

Chi-Square = 13.4848 with 12 d.f. P-Value = 0.334809



En esta ocasión, la transformación sí que proporciona un mejor ajuste a la normal. El p-valor es ya suficientemente grande. Por tanto podemos considerar que el logaritmo de los precios se ajusta a una normal. (Esto es equivalente a decir que los precios siguen una distribución lognormal)

5.7. El método de máxima verosimilitud

5.7.1. Introducción

El método de los momentos proporciona en muchas ocasiones una forma sencilla de obtener estimaciones de los parámetros de dicho modelos. Otro procedimeinto para estimar parámetros a partir de una muestra es el llamado método de **máxima verosimilitud** (MV). El método de MV es más complejo que el de los momentos, pero proporciona, en general, estimadores con mejores propiedades estadísticas. La razón de la superioridad del método de MV está en que se basa en la función de probabilidad o densidad de la población, mientras que el método de los momentos

no aprovecha esa información. En muchas ocasiones MV proporciona los mismos resultados que el método de los momentos. Otra ventaja del método de MV es que permite estimar parámetros que no estén relacionados con ningún momento muestral.

Ejemplo 6 *La velocidad de una molécula, según el modelo de Maxwell, es una variable aleatoria con función de densidad,*

$$f(x) = \begin{cases} \frac{4}{\sqrt{\pi}} \frac{1}{\alpha^3} x^2 e^{-(x/\alpha)^2}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

Donde $\alpha > 0$, es el parámetro de la distribución. Se desea estimar el parámetro α. Si calculásemos $E(X)$ veríamos que no guarda una relación sencilla con α, que permita estimar a a patir de $\bar{X}$. Sería deseable, por tanto, disponer de otra forma de estimar α.

El método de máxima verosimilitud consiste en estimar un parámetro mediante el valor que haga que la muestra observada sea un suceso de probabilidad máxima. Por ejemplo, sea p la probabilidad de observar un suceso al realizar un experimento. Si después de repetir 10 veces el experimento hemos observado el suceso en 5 ocasiones, ¿Cuál será el estimador máximo verosimil de p? Pues, obviamente $\hat{p} = 0{,}5$ será el valor que haga que la muestra observada sea el suceso más probable de entre todos los posibles. En el método de MV vamos a hacer un procedimiento matemático muy general que nos permita precisamente obtener este tipo de estimadores.

El método de máxima verosimilitud tiene los siguientes pasos:

1. Encontrar la función que nos dé, para cada valor del parámetro, o conjunto de parámetros, la probabilidad de observar una muestra dada.

2. Encontrar los valores de los parámetros que maximizan dicha función, lo que se obtiene simplemente derivando

5.7.2. La distribución conjunta de la muestra

Sea una variable aleatoria X de función de probabilidad $p(x)$ o función de densidad $f(x)$ que depende del parámetro (o vector de parámetros) θ. Sean $x_1, ..., x_n$ los valores observados en una muestra aleatoria simple $X_1, X_2, ..., X_n$ de dicha variable aleatoria X. Entonces la función de probabilidad conjunta será la función $p(x_1, ..., x_n)$ que proporcionará la probabilidad

$$p(x_1, ..., x_n) = P(X_1 = x_1; X_2 = x_2; ...; X_n = x_n),$$

es decir, es la probabilidad de obtener unos datos concretos $x_1, ..., x_n$. Esta función dependerá del parámetro θ. Por tanto podemos escribir

$$p(x_1, ..., x_n) = p(x_1, ..., x_n, \theta),$$

que proporcionará, para cada valor de θ la probabilidad de observar $x_1, ..., x_n$. Esta es precisamente la función que necesitamos maximizar en función de θ en el caso de una variable aleatoria discreta. En el caso de una variable aleatoria continua, la función de densidad conjunta será $f(x_1, ..., x_n)$ proporcionará la probabilidad por unidad de medidad de la muestra. Esta función es también función de los parámetros y también puede escribirse como

$$f(x_1, ..., x_n) = f(x_1, ..., x_n, \theta),$$

y ésta será la función a maximizar en funciónde θ. Las funciones $p(x_1, ..., x_n, \theta)$ o $f(x_1, ..., x_n, \theta)$ serán, en general, muy complejas. Sin embargo, en el caso de tener una muestra aleatoria simple, su cálculo es muy sencillo. Al ser una muestra independiente se puede demostrar que la función de probabilidad o densidad conjunta es el producto de las univariantes

$$\begin{aligned} p(x_1, ..., x_n) &= \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} p(x_i, \theta) \\ f(x_1, ..., x_n) &= \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} f(x_i, \theta) \end{aligned}$$

La demostración de este resultado no es sencilla y la omitiremos.

Por ejemplo, usando la información del Ejemplo 6 anterior se tiene

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i) = \frac{4^n}{\pi^{n/2}} \frac{1}{\alpha^{3n}} \prod_{i=1}^{n} \left(x_i^2\right) \exp\left(\frac{-\sum_{i=1}^{n} x_i^2}{\alpha^2}\right),$$

5.7.3. La función de verosimilitud

Las funciones $p(x_1, ..., x_n, \theta)$ y $f(x_1, ..., x_n, \theta)$ pueden utilizarse de dos formas distintas. Por una parte las podemos utilizar para, a partir de unos valores concretos de θ obtener una función que de valores diferentes en muestras diferentes. Es decir $p(x_1, ..., x_n|\theta)$ o $f(x_1, ..., x_n|\theta)$. De esta forma podremos calcular la probabilidad o densidad de unas muestras u otras para cada θ. Esta es precisamente la forma en que hemos utilizado estas funciones hasta ahora. Sin embargo, lo que nosotros necesitamos es justo la utilización inversa: dada una muestra que consideraremos ya fija, observada, obtener el valor de θ que maximiza dichas funciones, es decir $p(\theta|x_1, ..., x_n)$ o $f(\theta|x_1, ..., x_n)$. Estas funciones serán las mismas que las anteriores, sólo estamos cambiando la forma de usarlas. Cuando a la función de probabilidad o de densidad conjuntas las usamos de esta manera les llamaremos **Función de Verosimilitud** $l(\theta)$. La función de verosimilitud nos proporciona, para una muestra dada, la verosimilitud de cada valor del parámetro.

Por ejemplo, usando la información del Ejemplo 6 de la velocidad de una molécula se tiene la función de verosimilitud siguiente

$$l(\alpha) = \frac{4^n}{\pi^{n/2}} \frac{1}{\alpha^{3n}} \prod_{i=1}^{n} \left(x_i^2\right) \exp\left(\frac{-\sum_{i=1}^{n} x_i^2}{\alpha^2}\right).$$

Ésta será la función que tendremos que maximizar para consegir un estimador de α.

5.7.4. El método de máxima verosimilitud

Maximizar $l(\theta)$ a veces puede ser complicado. En el ejemplo anterior tenemos que $l(\theta)$ es una función no lineal de α. Hay α en el denominador, con potencias, e incluso dentro de una exponencial. Con frecuencia suele ser más sencillo maximizar $\ln(l(\theta))$ que $l(\theta)$. Como el logaritmo es una función monótona, el máximo de $\ln(l(\theta))$ y el de $l(\theta)$ se darán para el mismo valor de θ. Llamaremos **función soporte** a

$$L(\theta) = \ln(l(\theta)).$$

Siguiendo con el ejemplo anterior de la velocidad de una molécula, se tiene la siguiente función soporte:

$$L(\alpha) = k - 3n\log\alpha - \frac{\sum_{i=1}^{n} x_i^2}{\alpha^2}. \tag{5.5}$$

Podemos ya aplicar el método de máxima verosimilitud. El estimador de máxima verosimilitud de θ lo denotaremos por $\hat{\theta}_{MV}$ y será el máximo de $L(\theta)$, es decir

$$\hat{\theta}_{MV} = \arg_\theta \max L(\theta).$$

Si lo aplicamos a nuestro ejemplo de la ecuación de Maxwell para la velocidad de una molécula, derivaremos la función soporte (5.5). Se obtiene

$$\begin{aligned} \frac{dL}{d\alpha} &= \frac{-3n}{\alpha} + \frac{2\sum_{i=1}^{n} x_i^2}{\alpha^3} \\ \left.\frac{dL}{d\alpha}\right|_{\alpha=\hat{\alpha}_{MV}} &= 0 \end{aligned}$$

y resolviendo esta ecuación se halla el estimador máximo verosímil de α :

$$\hat{\alpha}_{MV} = \sqrt{\frac{2\sum_{i=1}^{n} X_i^2}{3n}}.$$

De esta forma, a partir de una muestra $X_1, ..., X_n$ podemos tener un valor de $\hat{\alpha}$.

5.7.5. Propiedades de los estimadores de máxima verosimilitud

Los estimadores de máxima verosimilitud verifican:

- Son insesgados, o asintóticamente insesgados
- Son asintóticamente eficientes (de menor error cuadrático medio)
- Son asintóticamente normales
- Su varianza asintótica es

$$\mathrm{Var}(\hat{\theta}_{MV}) = \left[-\frac{\partial^2 L(\theta)}{\partial\theta^2}\right]^{-1}$$

Utilizando el ejemplo anerior tenemos que

$$\mathrm{Var}(\hat{\alpha}_{MV}) = -\left[\frac{d^2 L}{d\alpha^2}\right]^{-1},$$

como

$$\frac{d^2 L}{d\alpha^2} = \frac{3n\alpha^2 - 6\sum_{i=1}^{n} x_i^2}{\alpha^4},$$

$$\mathrm{Var}(\hat{\alpha}_{MV}) = \frac{\alpha^4}{6\sum X_i^2 - 3n\alpha^2}.$$

sustituyendo α por $\hat{\alpha}_{MV}$ y operando se obtiene un estimador de esta varianza asintótica:

$$\widehat{\mathrm{Var}}(\hat{\alpha}_{MV}) = \frac{\hat{\alpha}_{MV}^4}{6\sum X_i^2 - 3n\hat{\alpha}_{MV}^2}$$

$$\widehat{\mathrm{Var}}(\hat{\alpha}_{MV}) = \frac{\hat{\alpha}_{MV}^2}{6n}.$$

Ejemplo 7 *Veamos otro ejemplo. Queremos estimar el parámetro p de una variable aleatoria binomial $X \sim B(n,p)$, a partir de la muestra $X_1,...,X_N$, donde X_i es el número de sucesos observados, en la muestra i-ésima, de tamaño n (tenemos N datos, basados en n elementos cada uno). Por ejemplo, número de artículos defectuosos en el lote i-ésimo de n artículos, de un total de N lotes. La función de probabilidad es*

$$P(X_i = x_i) = \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i}.$$

La función de verosimilitud será

$$l(p) = \prod_{i=1}^{N} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} = \binom{n}{x_1}\binom{n}{x_2}\cdots\binom{n}{x_N} p^{\sum x_i}(1-p)^{\sum (n-x_i)}$$

y la función soporte:

$$L(p) = cte + \left(\sum_{i=1}^{N} x_i\right)\ln p + \left(\sum_{i=1}^{N}(n-x_i)\right)\ln(1-p).$$

Derivando respecto a p tenemos

$$\begin{aligned}\frac{dL}{dp} &= \frac{\sum_{i=1}^N x_i}{p} - \frac{\sum_{i=1}^N (n-x_i)}{1-p} \\ &= \frac{\sum_{i=1}^N x_i}{p} - \frac{Nn - \sum_{i=1}^N x_i}{1-p}\end{aligned}$$

y particularizando para $\hat{p}_{MV}$ tenemos

$$\begin{aligned}\frac{\sum_{i=1}^N x_i}{\hat{p}_{MV}} - \frac{Nn - \sum_{i=1}^N x_i}{1-\hat{p}_{MV}} &= 0 \\ \frac{\sum_{i=1}^N x_i - \hat{p}_{MV}\sum_{i=1}^N x_i - Nn\hat{p}_{MV} + \sum_{i=1}^N x_i\hat{p}_{MV}}{\hat{p}_{MV}(1-\hat{p}_{MV})} &= 0\end{aligned}$$

Por tanto

$$\sum_{i=1}^{N} x_i - Nn\hat{p}_{MV} = 0 \Rightarrow \hat{p}_{MV} = \frac{\sum_{i=1}^N X_i}{Nn}$$

que es la proporción muestral de artículos defectuosos, pues $\sum_{i=1}^N X_i$ es el número de artículos defectuosos encontrados en la N muestras, y Nn es el número total de artículos analizados. Vemos que el resultado es el mismo que con el método de los momentos.