Universidad de la República
Facultad de Ciencias Sociales
DEPARTAMENTO DE ECONOMIA

Documentos de trabajo

Estadística para Economistas

Carlos Casacuberta

Nota Docente No. 04

Diploma en Economía 2004
Departamento de Economía
Facultad de Ciencias Sociales

Estadística
Notas de clase

1. Introducción

La estadística y su vinculación con la economía

La Estadística une dos campos de estudio:

1. El estudio sistemático de datos numéricos, el resumen y el análisis de la información contenida en ellos;

2. La teoría del azar y de la incertidumbre, o en otros términos, la teoría de la probabilidad.

Ambos son complementarios, aunque claramente distinguibles. En general los datos describen atributos de interés en un conjunto de objetos de estudio. Podemos considerar dichos datos en sí mismos, y buscar maximizar el uso de la información que nos brindan. Sin embargo, la teoría probabilística va más allá, e implica la utilización de modelos, lo que lleva a ver los datos como realización de una ley más general. A partir de la interacción del análisis de datos y la teoría probabilística surge un tercer campo de estudio, que comprende la prueba de hipótesis a partir de datos muestrales, o inferencia estadística.

Dado que este curso está destinado a economistas, destacamos los siguientes elementos de la vinculación de la estadística con el análisis económico.

En primer término podemos constatar que la metodología estadística interviene en la generación de los datos económicos. Es el caso cuando es imposible observar los actos económicos de la totalidad de los millones de agentes que interactúan en una economía y se debe obtener la información correspondiente por muestreo y realizar inferencias sobre las implicancias de cierta hipótesis a nivel de la población completa. La estadística proporciona herramientas teóricas para abordar este problema.

En segundo lugar, la estadística es el fundamento de la econometría, que puede definirse en forma amplia como el estudio sistemático de los fenómenos económicos utilizando datos observados. Para ello intervienen la metodología estadística y la teoría económica.

Una primera zona de interacción está dada porque en los propios modelos teóricos de la economía se utiliza la estadística a fin de repr3esentar situaciones de incertidumbre en forma probabilística. La teoría económica produce descripciones de los fenómenos económicos en forma de modelos, formulados en forma matemática, que incorporan un conjunto de variables y establecen relaciones entre las mismas buscando explicar y predecir. En los modelos

económicos las relaciones entre las variables no siempre pueden suponerse razonablemente como de naturaleza exacta o determinística.

Dichos modelos incorporan entonces la existencia de incertidumbre sobre los resultados de las acciones de los agentes económicos. Por ejemplo, una empresa que determina su producción, lo hace en condiciones de incertidumbre respecto al nivel de precios agregado o su variación (inflación). Muchas veces importan para los agentes los valores esperados de variables futuras, debiendo hacer el mejor uso posible de la información presente para eliminar al menos una parte de la incertidumbre sobre los períodos siguientes. La estadística es la base de los distintos enfoques utilizados para modelar las expectativas de los agentes económicos sobre hechos futuros.

Un segundo ámbito de interacción entre la estadística y la economía está dado por el análisis econométrico propiamente dicho. Idealmente, los modelos económicos están construidos buscando explicar fenómenos observados, por lo que deberían comprender un conjunto de hipótesis o afirmaciones sobre la forma en que se generan esos datos y sus relaciones. Por tanto, utilizando datos será posible realizar pruebas acerca de las consecuencias de las afirmaciones teóricas en el campo de los datos observables, estudiando así en qué medida la evidencia observable es consistente o no con determinada afirmación respecto al fenómeno que se estudia. El marco en que se realiza esta evaluación es el de considerarla como una decisión en condiciones de incertidumbre, y por lo tanto sujeta a la posibilidad de error. La teoría probabilística intenta unir a una medida de dicho error (precisión) una calificación adicional en términos de confianza, subrayando el compromiso que surge entre ambas, de manera que sólo resulta posible aumentar una a costa de reducir la otra.

En ciertos casos será posible experimentar, tratando de estudiar, en forma controlada, las decisiones económicas de los agentes ante cambios en el entorno, tratando de aislar sus efectos. Sin embargo, la economía es una ciencia en que la experimentación es en la inmensa mayoría de los casos imposible ya que no se puede reproducir las condiciones de la vida económica de manera artificial. En ambos casos, la teoría probabilística proporciona un marco para conceptualizar los procesos de generación de los datos y su utilización para la comprensión de los fenómenos económicos.

En el curso se explora brevemente la estadística descriptiva. A continuación se analizan nociones de probabilidad. Se introducen los conceptos de variable y vector aleatorio, y se presenta un conjunto de distribuciones de probabilidad de interés. Finalmente, se revisan las nociones de inferencia estadística, a través de ejemplos en la estimación puntual y por intervalos y la prueba de hipótesis. Para este curso solamente se requiere elementos básicos de cálculo. En el apéndice se desarrollan algunas herramientas matemáticas adicionales.

2. Estadística descriptiva

El primer tema que consideraremos es el de las técnicas para el resumen de la información contenida en un conjunto de datos acerca de atributos o características de un objeto de estudio. De allí el nombre de estadística descriptiva.

El análisis de datos comienza con una colección de objetos para analizar. Hay distintas formas de aproximarse a este conjunto. Sin embargo, para la exposición que sigue se supone que el conjunto observado nos interesa por sí mismo y no como representativo de un conjunto más amplio de objetos.

Generalmente esta no suele ser la situación, debido a que el número de objetos de interés suele ser muy grande, tal que no es posible examinar cada uno de ellos individualmente, como por ejemplo los estudiantes de la Universidad, los hogares de una ciudad de Montevideo, las empresas productoras de ciertos bienes o servicios. Este conjunto de objetos recibe el nombre de *población*. Solamente a veces es posible estudiar directamente a la población en su conjunto, como en el caso por ejemplo del Censo Nacional de Población y Vivienda. En otros casos, se realiza solamente un *muestreo* de los hogares del país, como es el caso por ejemplo de la Encuesta Continua de Hogares.

De allí surge la necesidad de la inferencia desde un conjunto reducido de objetos (*muestra*) al total de la población que no ha sido observada, que se estudiará más adelante. Sin embargo, abordar los datos en sí mismos permitirá descubrir métodos que serán de utilidad en esta tarea.

Consideramos una población compuesta por N elementos. Cada objeto o elemento en este conjunto está identificado por el índice i, un número entre 1 y N. En cada elemento de la población observaremos un atributo que será un número y que denotamos x, con un subíndice para designar el elemento de la población al que hacemos referencia. La población está representada por el conjunto

$$\{x_1, x_2, x_3, ..., x_N\} \quad \text{o} \quad \{x_i,\ i = 1, 2, ..., N\}.$$

Medidas de posición o tendencia central

Una vez que se tiene el conjunto de los datos, toda la información está contenida en la lista de los N números, y resulta evidente la dificultad de manejarnos cuando N es grande. Podemos representar gráficamente estos números en la recta real y tendremos una idea de cómo se agrupan los datos y dónde se encuentran. Si deseáramos describir este conjunto, una forma razonable de empezar sería intentar en qué porción del conjunto de los reales se encuentran. Podríamos ordenar las observaciones de menor a mayor, y etiquetarlas como $x_1, x_2, x_3,... x_N$, de manera que

$$x_1 \leq x_2 \leq x_3 \leq ... \leq x_N.$$

De esta manera sabemos que en el intervalo comprendido entre el máximo y el mínimo de los datos $[x_1, x_N]$ está toda la información de interés. No obstante, a menos que la distribución de los datos sea muy uniforme al interior del intervalo, el máximo y el mínimo pueden ser valores atípicos, muy poco frecuentes en los datos observados. Surge entonces la pregunta de si es posible obtener una medida igualmente concisa de la posición de las observaciones que

1) esté “cerca” de los datos y 2) en su construcción se utilice la información del conjunto de éstos. Presentamos dos formas de realizar esto.

La media (*aritmética*) se define como

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \sum_{i=1}^{N} \frac{x_i}{N}$$

Se trata de una suma ponderada, en la que todas las observaciones contribuyen a la suma y tienen el mismo ponderador 1/N. El concepto de promedio es una idea familiar, como valor representativo de una colección de números, y su característica principal consiste en que, en algún sentido, "está cerca" del conjunto de números en la población. Ello se puede ver de la siguiente manera: supongamos que hay un número K tal que la suma de las diferencias $x_i - K$, elevadas al cuadrado, sea mínima. Si sumamos las diferencias y nos planteamos las condiciones de primer orden para un mínimo obtenemos:

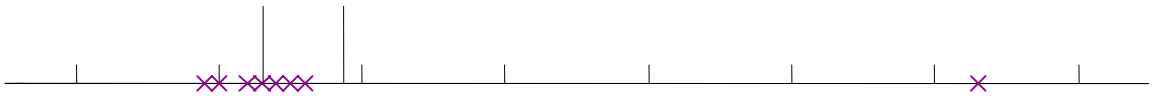
$$\frac{\partial \sum_{i=1}^{N} (x_i - K)^2}{\partial K} = 2\sum_{i=1}^{N} (x_i - K) = 0 \quad \Rightarrow K = \frac{\sum_{i=1}^{N} x_i}{N}$$

Si tenemos las observaciones ordenadas de menor a mayor, la *mediana* se define como

$$\text{Mediana}_X = x_{(N+1)/2} \text{ si N es impar; } \tfrac{1}{2} (x_{N/2} + x_{(N/2)+1}) \text{ si N es par.}$$

Si se tiene un número de observaciones par la mediana es el promedio de las observaciones centrales. Si el número de datos es impar la mediana es la observación central. De modo que el 50% de los datos son menores o iguales que la mediana y 50% de los datos son mayores o iguales que la mediana.

La media y la mediana difieren en la forma en que sus valores son afectados por observaciones ubicadas relativamente lejos de la media (*outliers*, del inglés "caer fuera"). Incluir dichas observaciones afectará en general más a la media que a la mediana, como se muestra en la figura siguiente.



Medidas de dispersión

Una vez que hemos dado una indicación acerca de la posición de los datos, nos interesa conocer si los datos se encuentran agrupados en un entorno vecino de la media o si por el contrario se hallan dispersos y alejados entre sí.

Una serie de medidas de dispersión se basa en las distancias a la media, o *desviaciones* de las observaciones. Como la suma de las desviaciones es cero, las elevamos al cuadrado antes de promediarlas, con lo que se obtiene siempre resultados positivos (enfatizando también la contribución a la suma de las desviaciones mayores en valor absoluto). La media de las desviaciones de la media elevadas al cuadrado es la varianza s^2.

$$s^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N}$$

Hay una forma abreviada conveniente para calcular la varianza, que se obtiene desarrollando el cuadrado de la expresión anterior.

$$s^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^{\,2} - 2x_i\bar{x} + \bar{x}^2\right) = \frac{1}{N}\left\{\sum_{i=1}^{N} x_i^{\,2} - 2\bar{x}\sum_{i=1}^{N} x_i + N\bar{x}^2\right\} = \frac{1}{N}\sum_{i=1}^{N} x_i^{\,2} - \bar{x}^2$$

La varianza es igual a la media de los cuadrados menos el cuadrado de la media.

Al tomar la raíz cuadrada de la varianza obtenemos la desviación standard *s*, que tiene las mismas unidades de media que la media y que las observaciones. Standard quiere decir que este valor es algo con lo que se compara, es un patrón o unidad de mediada de las dispersión de las observaciones con respecto a la media.

Si consideramos nuevamente a los datos ordenados en forma ascendente, podemos definir las medidas de dispersión asociadas a la mediana, que son las siguientes:

El *rango*, que se define como:

$$\text{Rango} = x_N - x_1$$

y el *recorrido intercuartil*, que se define como el intervalo entre el tercer y el primer cuartil. Para definir los cuartiles, supongamos que N+1 es divisible entre 4. Definimos entonces los cuartiles C_1, C_2, C_3, como

$$C_1 = x_{(N+1)/4}$$

$$C_2 = \text{mediana}_X$$

$$C_3 = x_{3(N+1)/4}$$

El recorrido intercuartil queda definido como

$$\text{recorrido intercuartil} = C_3 - C_1$$

y corresponde al rango en que están contenidas el 50% de las observaciones centrales [1].

Datos agrupados

Es frecuente que en una población los atributos que observamos tomen un número reducido de valores posibles, con muchos elementos de la población tomando un mismo valor. Es el caso, por ejemplo del número de miembros como atributo de un hogar, que típicamente será un número entre 1 y 6, con una pequeña proporción de casos por encima. Dichos datos reciben el nombre de *discretos*. En lugar de enumerar la población, la forma más conveniente de presentar los datos es la de una tabla de frecuencias. Para cada valor de los posibles, registramos cuantos casos en la población toman dicho valor.

---

[1] Si N+1 no es divisible entre 4 entonces los valores de los cuartiles deben *interpolarse*. Para ello definimos la parte entera de un número como el entero más cercano menor que un número dado y escribimos [N] = parte entera de N. En este caso los valores de los cuartiles se definen como:

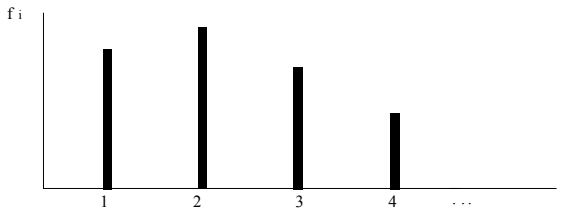
$$C_1 = x_{[(N+1)/4]} + (x_{[(N+1)/4]+1} - x_{[(N+1)/4]})\,((N+1)/4 - [(N+1)/4])$$

$$C_2 = x_{[(N+1)/2]} + (x_{[(N+1)/2]+1} - x_{[(N+1)/2]})\,((N+1)/2 - [(N+1)/2])$$

$$C_{3 =} x_{[3(N+1)/4]} + (x_{[3(N+1)/4]+1} - x_{[3(N+1)/4]})\,(3(N+1)/4 - [3(N+1)/4])$$

Recordemos que los datos están numerados en forma ascendente y que el subíndice nos indica el lugar del dato (por eso se toma la parte entera). Una vez tomada la parte entera de (N + 1)/4 se ubica el dato y se corrige por un factor igual a la diferencia entre este dato y el siguiente multiplicada por la fracción entre (N + 1)/4 y su parte entera.

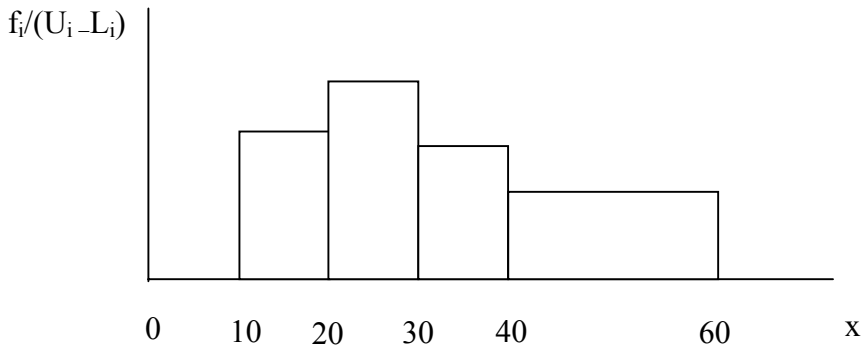
Supongamos que existen k posibles valores diferentes: m_1, m_2, ...,m_k. Denotamos f_1, f_2, ...,f_k a la frecuencia absoluta de cada valor (el número de veces que aparece). Toda la información que necesitamos está en los k pares de números (m_i, f_i). La misma se puede representar en el siguiente diagrama de barras:



Otra forma que toman los datos agrupados es cuando no se informan los valores exactos de los datos, sino se presenta un conjunto de intervalos o clases y la información de cuántos datos caen en cada uno de ellos. Ello puede obedecer a que en una encuesta no se pregunta el valor exacto sino solamente la pertenencia a cierto intervalo, o a que el número de valores diferentes sea tan grande que sea impracticable presentar una tabla. Un ejemplo típico son los datos sobre ingresos de las personas. Los pares son ahora de la forma:

$$\{f_i,\ [L_i\ ,\ U_i)\}$$

donde L_i es la cota inferior del intervalo y U_i la superior, y la notación [,) indica que el intervalo incluye la cota inferior pero no la superior. Estos datos pueden representarse en un *histograma*:



En este caso las frecuencias están representadas por las áreas y no por las alturas de los rectángulos cuya base igual a la amplitud de cada uno de los intervalos o clases.

Podemos definir las medidas de posición y de dispersión para datos agrupados:

Media:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} f_i \cdot m_i$$

Notamos que otra vez se trata de una suma ponderada, no de las observaciones, sino de los valores posibles. El ponderador es f_i/N o la frecuencia relativa. Sin embargo puede mostrarse que es exactamente igual a la media para datos no agrupados, ya que en la suma hay para cada valor xi, f_i sumandos iguales, cada uno con un ponderador igual a 1/N.

Desviación standard:

La fórmula de la desviación standard para datos agrupados según los valores es la siguiente:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{k} f_i \left(m_i - \bar{x}\right)^2}$$

Cuando los datos son discretos, los m_i son valores de los datos, pero cuando los datos están agrupados por intervalos, debe elegirse algún valor que "represente" a los valores observados. Suele tomarse los puntos medios del intervalo:

$$m_i = (L_i + U_i)/2$$

Ello lleva implícita la suposición de que los valores se distribuyen de manera uniforme dentro de cada intervalo o clase. En este caso perdemos la información sobre qué sucede al interior de cada intervalo, y la media ya no es igual a la que se obtendría si se dispusiera de las observaciones individuales sin agrupar.

Con datos discretos, obtener la mediana o los cuartiles no presenta dificultades, ya que puede fácilmente determinarse cuál es el valor hasta el que se acumula el porcentaje deseado de las observaciones, sumando las frecuencias relativas. Para obtener la mediana y las cuartiles en el caso de datos agrupados debemos proceder por interpolación.

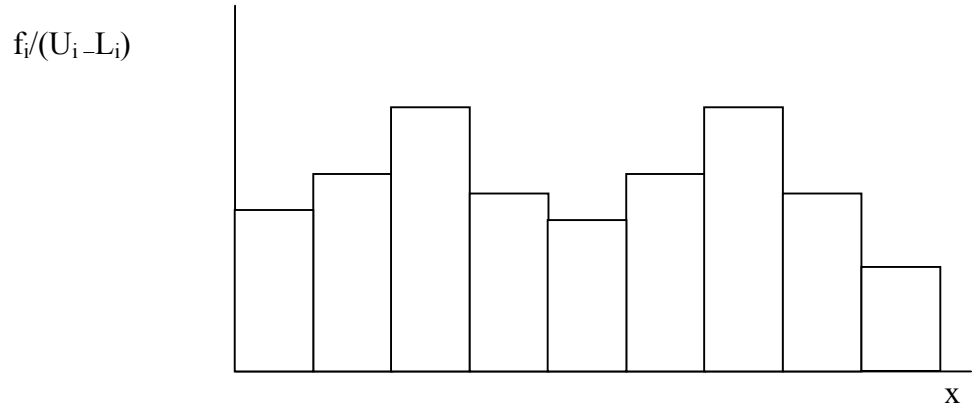
Modo

Una medida de posición adicional es el modo, que es el valor más frecuente en el caso de datos sin agrupar, y la clase con la frecuencia más alta (*intervalo o clase modal)* en el caso de datos agrupados.

Otras características de la distribución de los datos:

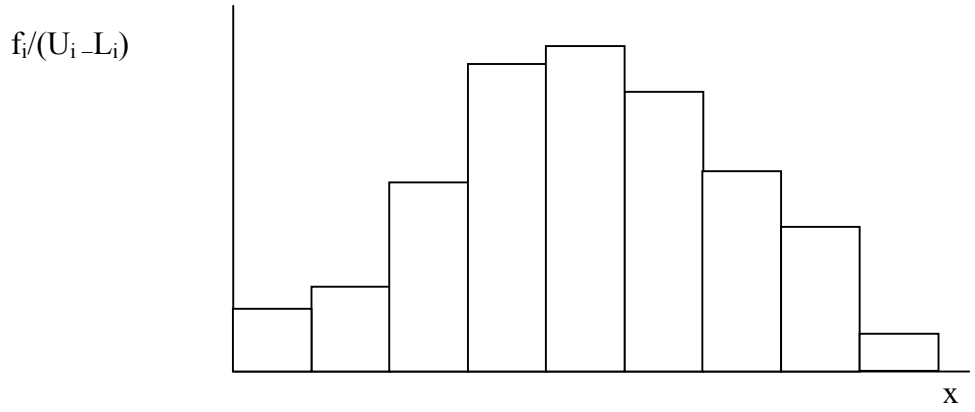
Asimetría

La idea de asimetría surge de la relación entre el "cuerpo" de la distribución (o aquella zona cercana a la media) y las "colas", o valores alejados de la media, donde en general tenemos un número menor de observaciones. Ello implica de algún modo la noción de que las distribuciones tienen *una* clase modal. En el caso de las distribuciones *bimodales,* existen relativamente pocos datos en la vecindad de la media.



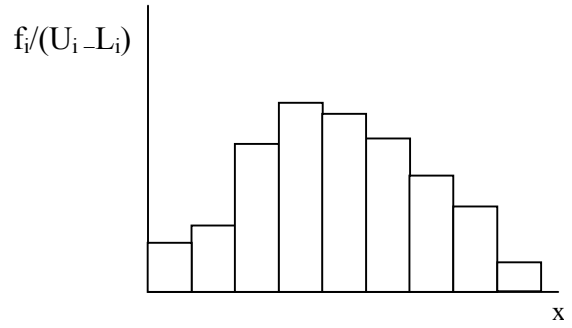
Distribución bimodal

El concepto de simetría se aplica en el caso de una distribución unimodal. En el caso de una distribución unimodal. La simetría indica que no hay una tendencia de los valores lejanos a la media a agruparse en una dirección en particular.

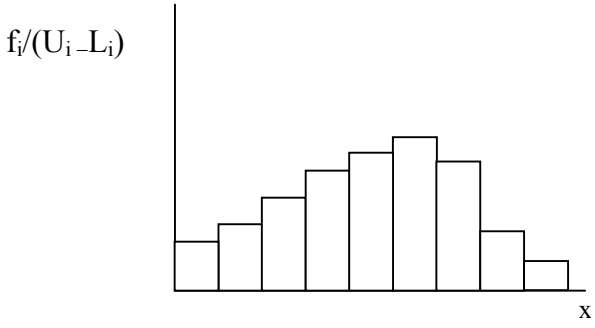


Distribución simétrica

Por el contrario la idea de asimetría se refiere a la tendencia de los valores extremos a agruparse en una dirección particular.



Asimetría a la derecha



Asimetría a la izquierda

La medida de asimetría está dada por la expresión siguiente:

$$\text{Coef. Asimetría} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^3}{s^3}$$

Aquí tomamos como en el caso de la varianza un promedio de las desviaciones, pero elevadas al cubo en vez de al cuadrado, lo que produce que 1) los valores alejados de la media contribuyen a la suma en mayor medida, y 2) las desviaciones conservan su signo original, Si las desviaciones negativas pesan más que las positivas, el coeficiente de asimetría tendrá signo negativo (distribución asimétrica a la izquierda), mientras que valores positivos implican asimetría a la derecha (una distribución simétrica tiene un coeficiente de 0). La división entre la desviación standard elevada al cubo determina que el coeficiente de asimetría no dependa de las unidades de medida empleadas.

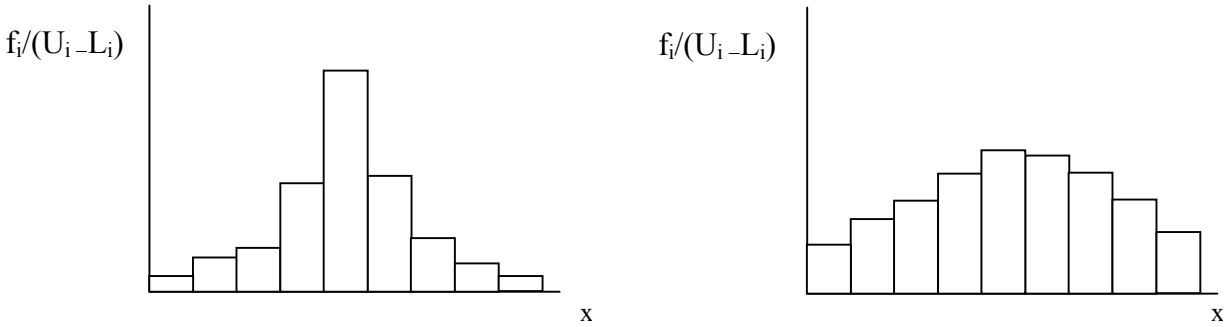
Kurtosis

La kurtosis describe la relación que existe entre el cuerpo de una distribución y las colas. La expresión para el coeficiente de kurtosis es la siguiente:

$$\text{Kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{s^4}$$

Valores reducidos implican que las colas de la distribución pesan poco con respecto al cuerpo (*leptokurtica*). Por el contrario, cuando los valores son altos la distribución tiene una forma más "achatada": las colas tienen un peso importante con respecto al cuerpo de la distribución

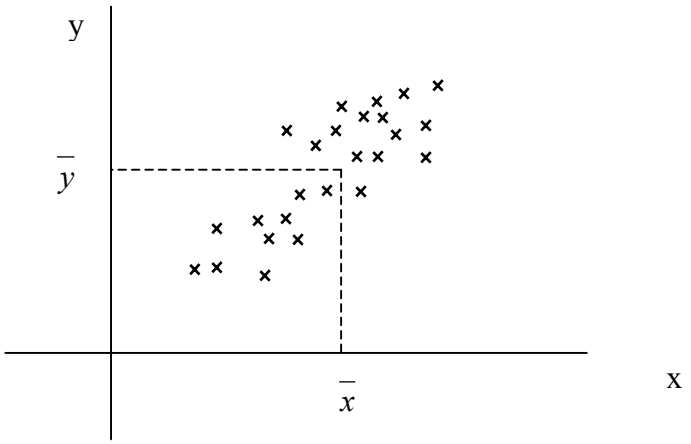
(*platikúrtica*). Por este motivo a veces se menciona el coeficinte de kurtosis como "coeficiente de apuntamiento".



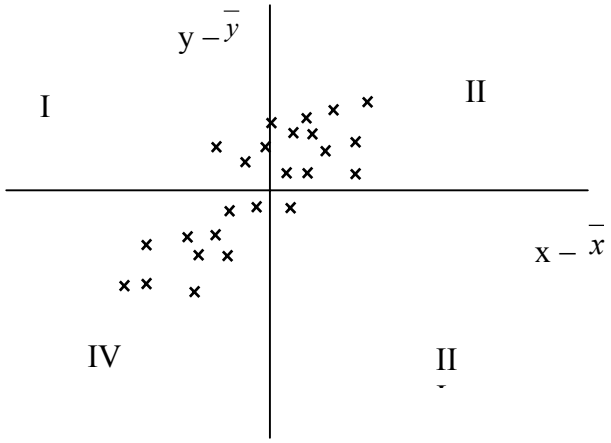
Leptokúrtica Platikúrtica

Nubes de puntos y correlación

Consideremos el caso en que una población genera pares de observaciones, como por ejemplo el consumo y el ingreso mensual de los hogares de Montevideo medido en pesos. El problema que nos planteamos es ver si ambos atributos, que llamaremos X e Y, están relacionadas de alguna manera y en qué forma. Una de las formas de representar los datos que sugiere la relación entre ambos atributos es el gráfico de nube de puntos, en el que hemos representado las medias muestrales de X e Y .



En el ejemplo imaginario del gráfico es evidente la existencia de algún tipo de relación (a valores altos de X corresponden valores altos de Y y viceversa). Esto se puede ver más claro si movemos los ejes del gráfico a los puntos de las medias muestrales de las observaciones. En el siguiente diagrama hemos restado el valor de la media a cada observación de X e Y, es decir las hemos expresado en forma de desviaciones. Todos los puntos en que X e Y exceden la media están en el cuadrante II (ambos positivos); en el cuadrante IV ambas desviaciones son negativas. En los cuadrantes I y III las desviaciones de una y otra variable tienen diferente signo.



Resulta claro aquí que hay más puntos en los cuadrantes II y IV que los que hay en los cuadrantes I y III. En este caso decimos que hay una correlación positiva entre X e Y: valores altos de X están asociados con valores altos de Y y viceversa. Si no hubiera relación alguna entre X e Y podríamos esperar que los puntos estuvieran dispersos alrededor de la media de manera que hubiera más o menos el mismo número de puntos en cada cuadrante.

Ello sugiere una medida de asociación entre ambos atributos, llamada *covarianza*:

$$s_{XY} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

La covarianza es también un promedio, es la media de los productos de las desviaciones. Los puntos en los cuadrantes II y IV contribuyen a la suma con sumandos positivos, mientras que los puntos en los cuadrantes I y III lo hacen con sumandos negativos. La covarianza puede ser positiva y negativa y no está acotada, dependiendo de las unidades de medida de X y de Y que no tienen porqué ser las mismas.

Es más conveniente entones tener un indicador que no dependa de las unidades de medida empleadas, y para ello se utiliza el *coeficiente de correlación*, que está definido como el cociente entre la covarianza y el producto de las desviaciones standard de X y de Y.

$$r_{XY} = \frac{s_{XY}}{s_x s_y}$$

El coeficiente de correlación siempre está comprendido entre −1 y 1 (no lo demostraremos). Cuando el coeficiente de correlación es cercano a cero entonces hay baja o nula correlación entre las variables, en cambio valores cercanos a 1 y a −1 implican alta correlación, positiva y negativa respectivamente. La correlación indica una asociación de tipo lineal entre ambos atributos. Si la relación fuera exactamente lineal tendríamos r_{XY} igual a 1 o a −1.

3. Probabilidad

Enfoques de la probabilidad

Intuitivamente asociamos probabilidad con el grado de verosimilitud o certeza que asignamos a cierto suceso. Al decir "es improbable que Juan venga hoy", o "probablemente mañana llueva", usamos la expresión para referirnos a un suceso que puede ocurrir o no y que nos parece más o menos esperable.

Para precisar la definición comenzamos con una breve discusión de un conjunto de diferentes enfoques acerca de qué es la probabilidad.

El primero que se presenta es el llamado enfoque clásico (o *a priori*) define la probabilidad de la siguiente manera:

Si un experimento aleatorio (definido informalmente como un suceso cuyo resultado se desconoce con anterioridad a que ocurra) puede producir n resultados igualmente verosímiles y mutuamente excluyentes, de los cuales n_A poseen el atributo A, entonces la probabilidad de que ocurra A es igual a la fracción n_A/n (casos favorables/casos posibles).

Ejemplo:

Supongamos que alguien quiere calcular la probabilidad de obtener dos "caras" si lanzamos una moneda dos veces. Los resultados excluyentes e igualmente verosímiles son cuatro: {(C,C), (C,N), (N,C), (N,N)}. Contando, vemos que únicamente en uno de ellos obtenemos dos caras con lo que deducimos que la probabilidad de dicho suceso es igual a 1/4.

Un primer problema de esta definición es que implica restringirse a eventos que tienen un número finito de resultados. Sólo cuando está definido el conjunto de los resultados igualmente verosímiles podemos contar aquellos en que encontramos el atributo A.

El segundo problema que surge es que esta definición necesita que los resultados posibles sean "igualmente verosímiles". Esto vuelve a la definición circular, y no nos permite determinar probabilidades cuando no sabemos a priori si los resultados son igualmente verosímiles.

A diferencia del enfoque clásico, el enfoque *frecuencista* (o *a posteriori*) propone una idea de

probabilidad íntimamente relacionada con la posibilidad de repetir determinado experimento y observar el resultado en un número n, arbitrariamente grande, de realizaciones. Es crucial que esto pueda hacerse en las mismas condiciones cada vez, aunque cada resultado individual siga siendo impredecible. Postulamos que existe un número P(A) que es la probabilidad del suceso A, y lo aproximamos por la frecuencia relativa de su ocurrencia en un número grande de casos.

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}$$

Esta definición permite que la probabilidad no quede restringida a experimentos con un conjunto de resultados equiprobables, pero requiere que sea posible una larga serie de pruebas repetidas. La probabilidad surge de la estabilidad observable de las frecuencias relativas.

La pregunta en este caso es si se puede hablar de la probabilidad de eventos que no caen dentro de la clase de los repetibles en condiciones aproximadamente similares. ¿Cual es la probabilidad que un comando suicida secuestre al profesor cuando abandone el salón de clase? ¿Cual es la probabilidad de que se descubra la cura para el SIDA antes del año 2010?

Estas preguntas pueden legítimamente incluirse dentro de un marco probabilístico dentro del enfoque *subjetivo* de la probabilidad. La definición en este caso es que la probabilidad P(A) representa el grado subjetivo de certeza sobre la ocurrencia del suceso A en un experimento futuro.

Una ilustración sobre cómo puede revelarse esta asignación subjetiva de probabilidades consiste en considerar la pregunta "cuánto me tienen que pagar para que yo acepte apostar un peso si el evento A sucede". Para evaluar la apuesta usamos el criterio de la ganancia esperada, definida informalmente como la suma ponderada de los montos que gano y pierdo en cada caso, cada uno multiplicado por la probabilidad de que ocurra una y otra cosa. Un criterio para decidir aceptar la apuesta podría ser "una apuesta es justa si la ganancia esperada es al menos 0", es decir no hay razones para esperar perder dinero. Partimos de una probabilidad desconocida que llamamos p, y la disposición subjetiva a aceptar cierta apuesta "revela" esta probabilidad.

Por ejemplo, el suceso A podría ser "Peñarol vence a Nacional en el siguiente clásico". Si, por ejemplo, estuviera dispuesto a aceptar 1,25 pesos por cada peso apostado, mis ganancias esperadas equivalen a lo que gano (1,25 centésimos) multiplicado por la probabilidad de ganar (que llamamos p) menos la apuesta (1 peso) multiplicada por la probabilidad de perder (1 – p). La apuesta parecerá justa si mi *ganancia esperada* es positiva.

Ganancia esperada = $1{,}25p - (1-p) = 2{,}25p - 1 \geq 0 \Leftrightarrow \quad p \geq 1/2{,}25 = 0{,}44$

Por lo tanto mi aceptación de la apuesta estaría revelando que yo asigno una probabilidad subjetiva al evento mayor que el 44%. ¿En qué me baso? Puedo haber examinado los registros históricos y calculado la frecuencia relativa de las victorias, pero también puedo modificar mi apreciación de acuerdo a las condiciones del clima, las declaraciones de los directores

técnicos, el hecho de que cierto jugador esté o no lesionado, etc.

Esto muestra que difícilmente puede concebirse una probabilidad subjetiva que no tenga en cuenta experiencia anterior en condiciones aproximadamente similares.

También es útil distinguir entre la total ignorancia y el considerar a los eventos posibles como equiprobables. La ignorancia total se identifica más bien con la imposibilidad de asignar probabilidades, mientras que la equiprobabilidad tiene que ver con que no hay ninguna razón para considerar algún resultado como más probable que otro.

Para el tratamiento matemático de la probabilidad, se comienza por definir un conjunto de axiomas, desarrollando en forma deductiva el conjunto de proposiciones y teoremas utilizando la lógica matemática. Lo interesante es que la definición matemática de probabilidad no requiere en sí misma de ninguna interpretación sobre lo que la probabilidad "es" (clásica, frecuencista o subjetiva), sino que las comprende a las tres y permite en cada caso el cálculo de las probabilidades correspondientes.

Experimento aleatorio

Los axiomas de la probabilidad que desarrollaremos más adelante son una descripción idealizada de un mecanismo aleatorio. El punto de partida es la noción de un *experimento aleatorio.*

Def.: Un experimento aleatorio, denotado por $\mathfrak{E}$, es un experimento que satisface las siguientes condiciones:
i) todos los posibles resultados son conocidos a priori
ii) en una realización particular el resultado no es conocido a priori
iii) el experimento puede ser repetido bajo idénticas condiciones.

Espacio muestral y eventos aleatorios

Podemos definir ahora dos términos esenciales en probabilidad, *espacio muestral* y *evento aleatorio*. Ambos se definen en términos de conjuntos.

El *espacio muestral* es el conjunto de todos los resultados posibles de un experimento aleatorio, y lo denotamos S.

Un *evento aleatorio* es cualquier subconjunto del espacio muestral S. Los eventos aleatorios se definen en términos de los elementos de S. Si denominamos A a un conjunto de elementos de S, decimos "el evento A ha ocurrido" si el resultado del experimento aleatorio fue uno de los resultados contenido en A. El conjunto de los eventos, o *espacio de los eventos*, lo denotamos por $\mathfrak{I}$.

Los elementos de S reciben el nombre de "eventos elementales". Por ejemplo, en el caso del

clásico el conjunto de resultados elementales es:

S = {"gana Peñarol", "gana Nacional" , "empate"}

No siempre en el conjunto S hay un número finito de eventos elementales. Por ejemplo, el experimento aleatorio "tirar una moneda hasta que salga cara" tiene como resultados posibles una serie de n-uplas de la forma:

S = {(C), (NC), (NNC), (NNNC), (NNNNC) ...}

que constituyen un conjunto de infinitos elementos.

Los eventos elementales generan a su vez nuevos eventos, a través de las operaciones usuales entre conjuntos (la unión y la intersección). Por ejemplo, el evento "Nacional no pierde" está formado por la unión de los eventos "gana Nacional" y "empate".

Si A es un evento aleatorio, A^c es el evento "A no ocurre". El espacio muestral S, que contiene todos los resultados del experimento, se puede pensar como un evento, el evento "el experimento tiene un resultado". En este sentido, dado que S ocurre siempre, se puede decir que S es el "evento seguro".

El conjunto vacío, ∅, está contenido en el conjunto de los eventos, como el evento "el experimento no tiene un resultado". De este modo, el evento ∅ es el "evento imposible".

De la misma manera, $A \cup B$ representa el evento "A o B han ocurrido", y $A \cap B$ representa el evento "ambos A y B han ocurrido".

Se puede asimismo definir la idea de eventos excluyentes, que son dos eventos A y B tales que $A \cap B = \varnothing$, lo que puede leerse "si ocurre A entonces B no puede ocurrir y viceversa". Recordando la definición de partición, una colección de eventos es una partición si "uno y sólo uno de ellos puede ocurrir".

σ–álgebra

La teoría impone cierta estructura al conjunto $\mathfrak{I}$ de los eventos. Esta estructura consiste en que $\mathfrak{I}$ sea un *σ–álgebra* asociado a S.

Un conjunto $\mathfrak{I}$ (no vacío) de subconjuntos de S es un σ–álgebra si cumple con :

i) Si $A \in \mathfrak{I}$ entonces $A^c \in \mathfrak{I}$
ii) Si $A_i \in \mathfrak{I}$, i = 1,2,... entonces $\bigcup_{i=1}^{\infty} A_i \in \mathfrak{I}$

De ello se deriva que, para nuestro conjunto $\mathfrak{I}$ de los eventos se cumple que:

i) $S \in \mathfrak{I}$
ii) $\varnothing \in \mathfrak{I}$

Ejemplo. Consideremos el experimento de tirar dos monedas. El espacio de resultados está dado por el conjunto S = {(CC), (CN), (NC), (NN)}. Un σ–álgebra asociado puede definirse como:

$\mathfrak{I}$ = {{S}, {∅}, {(CC)}, {(CN), (NC), (NN)}}.

No hay un único σ–álgebra asociado a cada experimento, y su conformación depende de los eventos de interés. Por ejemplo, en el marco del mismo experimento puede definirse otro σ–álgebra $\mathfrak{I}_2$ de la siguiente manera (comprobar como ejercicio que ambos conjuntos son σ–álgebra):

$\mathfrak{I}_2$ = {{S}, {∅}, {(CC), (NN)}, {(CN), (NC)}}.

En general, partiremos del conjunto S de resultados posibles. Definidos los eventos de interés dentro del experimento, procederemos a completar la clase $\mathfrak{I}$ con las uniones y complementos correspondientes.

Axiomas de la probabilidad

La probabilidad es una función que asocia a cada evento A perteneciente a $\mathfrak{I}$ un número que denotamos P(A), "la probabilidad de que A ocurra". Dicha función queda definida mediante los siguientes axiomas:

Def. : Probabilidad es una función P (·): $\mathfrak{I} \rightarrow$ [0, 1] (definida en $\mathfrak{I}$ y que toma valores en el intervalo cerrado [0, 1] de los números reales), que cumple con los siguientes axiomas:

1) para cualquier $A \in \mathfrak{I}$, P (A) ≥ 0.

2) P (S) = 1

3) Si $\{A_i\}_{i=1}^{\infty}$ es una secuencia de eventos mutuamente exclusivos en $\mathfrak{I}$, entonces

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

La terna (S, P, $\mathfrak{I}$) recibe el nombre de *espacio de probabilidad*.

De los axiomas básicos se deduce un conjunto de reglas:

a) A y A^c son por definición disjuntos, y además $A \cup A^c = S$, de manera que

$P(A) + P(A^c) = P(S) = 1.$

Esto nos da la regla $P(A^c) = 1 - P(A)$

b) $P(\varnothing) = 1 - P(S) = 0$

c) B se puede escribir como la unión de dos conjuntos disjuntos:

$B = (B \cap A) \cup (B \cap A^c)$

de modo que

$P(B) = P(B \cap A) + P(B \cap A^c)$

Al mismo tiempo,

$P(A \cup B) = P(A) + P(B \cap A^c)$

de modo que usando la relación anterior tenemos que

$P(A \cup B) = P(A) + P(B) - P(A \cap B).$

Probabilidad condicional

Consideremos ahora el caso en que la información que tenemos acerca de la ocurrencia de cierto evento se modifica y adquirimos algún conocimiento parcial acerca del mismo. No conocemos si el evento ha ocurrido -si lo supiéramos no tendría sentido hablar de probabilidad- pero tenemos información de que algún otro evento ha ocurrido. ¿cómo son afectadas las probabilidades que asignamos a la ocurrencia del primer evento?

Por ejemplo, supongamos que queremos adivinar cuál es una carta extraída al azar de un mazo de 48 cartas. Si nuestra adivinanza fuera, digamos, que la carta es el rey de espadas y llamamos a ese evento A, la probabilidad que asignaríamos a dicho evento, sin conocer otra información, sería $P(A) = 1/48$.

Supongamos ahora que nos dicen que "el palo es espadas" y llamamos a este evento B. Esto modifica la probabilidad que asignamos a nuestro evento original en dos sentidos. Por una parte, sabemos que, terminantemente, ninguna de las cartas que no son espadas puede salir. Esto reduce nuestro espacio muestral a las 12 espadas. En segundo lugar, el rey de espadas sigue siendo posible, y dado que una de las doce espadas es el evento que nos interesa, podemos obtener la probabilidad de A condicional a que el evento B ha ocurrido:

$$P(A/B) = 1/12$$

o también "probabilidad de A dado B". En el caso particular de nuestro ejemplo $P(A/B^c) = 0$.

Para calcular probabilidades condicionales se utiliza la regla:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Las probabilidades relevantes para cualquier evento A pasan a ser ahora las probabilidades del evento (A ∩ B). El hecho que B ha ocurrido reduce el espacio muestral al evento B, y la división por P(B) introduce este escalamiento. En nuestro ejemplo,

P(rey de espadas /espadas) = P(rey de espadas ∩ espadas)/P(espadas)= (1/48)/(1/4) = 1/12.

Independencia

A partir del concepto de probabilidad condicional, podemos ver que no siempre saber que ocurre B nos va a llevar a modificar nuestra evaluación de las probabilidades que le asignamos a A. Por ejemplo, consideremos al evento A "sacar un as de un mazo de cartas", y al evento B "la carta extraída es una espada". Sabemos que P(A) = 4/48 = 1/12 y nos interesa determinar si conocer la ocurrencia de B modifica la evaluación de la probabilidad de que A ocurra.

Calculando la probabilidad condicional obtenemos que

$$P(A/B) = P(A \cap B) / P(B) = (1/48)/(12/48) = 1/12$$

En este caso conocer la ocurrencia de B no modifica la probabilidad de que ocurra A.

Def. Dados dos eventos A y B, si se cumple que

$$P(A/B) = P(A)$$

ambos eventos son *estadísticamente independientes*. Aplicando la definición de probabilidad condicional obtenemos que, si dos eventos A y B son independientes,

$$\frac{P(A \cap B)}{P(B)} = P(A)$$

de donde deducimos la *regla de la multiplicación* para eventos independientes:

Si A y B son eventos independientes, entonces

$$P(A \cap B) = P(A) \cdot P(B)$$

Por último, nuestro ejemplo muestra que independencia no es lo mismo que exclusión mutua. Los eventos del ejemplo pueden ocurrir ambos a la vez, (la carta puede ser una espada y ser

un as) y son independientes entre sí. Si dos eventos A y B, no vacíos, son mutuamente excluyentes o disjuntos, es decir su intersección es vacía, entonces P(A/B) y P(B/A) son siempre iguales a cero, por lo que no serían independientes. En tal caso, conocer la ocurrencia de A modifica la evaluación de la probabilidad de que B ocurra, hasta el punto de implicar que B es imposible.

Conviene aclarar sin embargo que cuando se consideran más de dos sucesos a la vez, para que sean todos ellos independientes entre sí no basta con que sean independientes dos a dos.

Regla de Bayes

Así como teníamos

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

también se cumple que

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

y por lo tanto

$$P(A \cap B) = P(A/B) \cdot P(B) = P(B/A) \cdot P(A)$$

de lo cual obtenemos la expresión

$$P(B/A) = \frac{P(A/B)}{P(A)} \cdot P(B)$$

conocida como la *Regla de Bayes*. Puede entenderse como una regla para revisar probabilidades a la luz de la incorporación de nueva información. Partimos de un evento B que no ha sido observado pero al que se asigna una probabilidad. Llamamos a esta probabilidad P(B) *a priori*. Si de alguna manera sabemos que A ha ocurrido, si conocemos las probabilidades condicionales P(A/B) y $P(A/B^c)$, podemos revisar nuestra evaluación de P(B) y llegar a P(B/A) o *probabilidad posterior*. Con P(A/B) y $P(A/B^c)$ reconstruimos P(A) utilizando el hecho que $P(A) = P(A \cap B) + P(A \cap B^c) = P(A/B)P(B) + P(A/B^c)P(B^c)$.

Por ejemplo, tomemos el experimento consistente en entrevistar a los integrantes activos laboralmente de un hogar y preguntarles su condición de jefe de hogar o no y su condición de ocupado o desocupado. Tomemos el evento "el individuo está desempleado" y los eventos "el individuo es jefe del hogar" y "el individuo no es jefe de hogar". Supongamos que, observando al conjunto de los desocupados se pueda decir que las probabilidades son:

P(jefe de hogar/ desempleado) = 0,15
P(jefe de hogar/ no desempleado) = 0,85

Supongamos que nuestra probabilidad *a priori* de que un individuo esté desempleado, sin observar su condición de jefe de hogar es de 0,10. Notamos que la probabilidad de "jefe de hogar" es igual a 0,78 = (0,15)·(0,10) + (0,85)·(0,90), dado que los eventos desempleado y no desempleado constituyen una partición de S. La regla de Bayes da las probabilidades a posteriori de estar desempleado dado que el individuo es jefe de su hogar como

P(desempleado/ jefe de hogar) = (0,15)·(0,10)/(0,78) = 0,02

Si consideramos un conjunto de r eventos B_1, B_2, ..., B_r tales que constituyen una partición de S, todos con probabilidades distintas de cero, entonces cualquier evento A contenido en S ocurre a través de la ocurrencia de alguno de los B_i. A su vez las intersecciones de A con cada uno de los B_i son disjuntas dos a dos, de modo que podemos escribir:

$$\text{P (A)} = \sum_{i=1}^{r} P(A \cap B_i) = \sum_{i=1}^{r} P(A/B_i) P(B_i)$$

Para el caso de eventos B_1, B_2, ..., B_r que constituyen una partición de S, todos con probabilidades distintas de cero, entonces la regla de Bayes es:

$$P(B_i / A) = \frac{P(A/B_i)}{\sum_{i=1}^{r} P(B_i) \cdot P(A/B_i)}$$

4. Variables aleatorias

Definiciones

Se introduce a continuación el concepto de *variable aleatoria*, sin duda el más importante en el contexto de este curso. Nos servirá para redefinir nuestro espacio de probabilidad, haciéndolo más flexible, y adaptarlo al conjunto de situaciones en que el resultado de un experimento aleatorio es o puede representarse por un número. Los experimentos aleatorios de interés para la economía suelen casi siempre generar resultados numéricos (por ejemplo, el PBI, la tasa de inflación, tasa de desempleo, etc.). Cuando no es así, la variable aleatoria también permite asignar números a resultados cualitativos, haciendo más manejable nuestro espacio de probabilidad sin cambiar su estructura básica. El espacio (S, P, $\mathfrak{F}$) de probabilidad estudiado hasta aquí presenta el problema de que el dominio de la función P es un σ–álgebra de eventos, lo que hace su manipulación complicada. Para calcular las probabilidades de eventos se deben derivar los elementos de $\mathfrak{F}$, lo cual puede ser una tarea difícil cuando se trata de conjuntos con muchos o infinitos elementos.

Consideremos ahora una función que asigna a cada elemento del espacio muestral S uno y sólo un elemento del conjunto de los números reales:

$$X(\cdot) : S \rightarrow \Re$$

Con esto se establece la base para obtener un mapa que va de los eventos a los números reales. Usamos la variable aleatoria para describir eventos, de modo que la ocurrencia de cierto evento ahora estará representada por la función X tomando valores en un intervalo determinado de los reales. No cualquier función, sin embargo, preservará la estructura de probabilidad de nuestro espacio original. Como las probabilidades están definidas para eventos, necesitamos, para cualquier intervalo de la recta real, que las preimágenes de la función en dicho intervalo sean eventos, es decir, pertenezcan a $\mathfrak{I}$. Por preimágenes entendemos aquellos elementos de S tales que la función X les asocia una imagen en ese intervalo de $\Re$. La definición precisa del conjunto de intervalos que se toman en cuenta hace uso de un concepto nuevo, el de la clase de Borel, que es un conjunto de subconjuntos de $\Re$ que cumple las condiciones de ser un σ–álgebra de intervalos y que incluye los eventos que generalmente serán de interés. Sin embargo, no vamos a estudiar en detalle las clase de Borel y nos vamos a limitar a establecer una condición necesaria para que la función X represente adecuadamente eventos. Si las preimágenes de todos los intervalos abiertos por la izquierda, del tipo $(-\infty , x]$, son eventos, entonces las preimágenes de todos los conjuntos de la clase de Borel serán eventos.

Por lo tanto se pide la siguiente condición adicional para una variable aleatoria: consideremos un número real cualquiera x, y observemos el conjunto de resultados del experimento (los que denotamos con la s minúscula) tales que los valores que la función X asigna son menores o iguales que x. Lo escribimos como el conjunto A_x:

$$A_x = \{s \in S: X(s) \leq x\}$$

La condición indica que el conjunto de los resultados s tal que $X(s) \leq x \in \mathfrak{I}$, es decir, deben pertenecer a la clase de eventos asociada al experimento.

Def. Dado un espacio de probabilidad (S, $\mathfrak{I}$, P), una *variable aleatoria* es una función que asocia a cada elemento del espacio de resultados S un número real, tal que para todo $x \in \Re$, $A_x = \{s : X(s) \leq x\} \in \mathfrak{I}$.

Una variable aleatoria, entonces, sólo tiene sentido en relación a un determinado σ–álgebra de eventos.

Por ejemplo, consideremos el experimento de arrojar dos monedas consecutivas, y definamos la función X = número de caras obtenidas. Para determinar si nuestra función es una variable aleatoria, enunciamos nuestro espacio muestral, que está dado por: S= {NN, NC, CN, CC}, en el cual X está definida como: X (NN) = 0, X (CN) = 1, X (NC) = 1 y X (CC) = 2. Un posible σ–álgebra asociado a este espacio muestral es:

$\mathfrak{I}$ ={ S, ∅, {(CC)}, {(NN)}, {(CN), (NC)}, {(CN), (NC), (NN)}, {(CN), (NC), (CC)}, {(CC), (NN)}}

construido con los conjuntos de resultados que dan 0, 1, o 2 caras, sus uniones y sus intersecciones (verificar que es un σ–álgebra). Para verificar que X es una variable aleatoria,

consideremos los conjuntos de tipo $X(s) \leq x$ para distintos valores de x. Debemos recorrer toda la recta real, pero esta tarea se ve facilitada debido a que la función X toma valores en un número reducido de puntos. Así, construimos la tabla de los posibles intervalos semiabiertos $(-\infty , x]$ y de sus preimágenes en S:

$(-\infty , x]$ tal que	s: $X(s) \leq x$
$x < 0$	$\varnothing$
$0 \leq x < 1$	{(NN)}
$1 \leq x < 2$	{(NN), (CN), (NC)}
$2 \leq x$	S

Como vemos que para cada uno de los intervalos posibles considerados, el conjunto de preimágenes pertenece a la clase de los eventos, podemos concluir que X es una variable aleatoria.

La relación de la variable aleatoria con el espacio de probabilidad original está dada porque podemos considerar si la variable aleatoria toma valores en un intervalo $(-\infty , x]$ y calcular la probabilidad asociada de la siguiente manera:

$$P(X(s) \leq x) = P\{s: X(s) \leq x\}$$

Podemos hacerlo porque siempre el conjunto de los elementos del espacio de resultados a los que la función X asigna un valor menor o igual que x son eventos, para cualquier x real.

En algunos de los ejemplos vistos en la discusión de probabilidad, los resultados de los experimentos aleatorios eran de por sí numéricos, como en el caso del dado. En general en economía será este el caso. Consideremos por ejemplo el experimento aleatorio consistente en "tomar un año dado y observar a las empresas residentes en un territorio dado sumando el total del valor de la producción de bienes y servicios del año a precios corrientes sin incluir los insumos intermedios". El resultado de dicho experimento podría ser cualquier número real, y para describirlo empleamos la variable aleatoria PBI_t, donde el subíndice t denota que está referida a un período de tiempo determinado. Obviamente, esto es así antes de que la producción ocurra y se efectúe la medición. Cuando ésto ya se ha realizado, y tenemos por ejemplo el PBI de 1995, éste ya no tiene nada de aleatorio. Además, no podemos -definitivamente- volver a repetir el experimento.

En experimentos que arrojen resultados de tipo cualitativo, éstos se pueden expresar numéricamente, de la misma manera que etiquetábamos los resultados de arrojar una moneda como "cara" = 0 y "número" = 1.

Funciones de cuantía, densidad y distribución

Al definir las variables aleatorias asociamos probabilidades a los intervalos abiertos de forma $(-\infty , x]$. El siguiente paso que daremos es describir dichas probabilidades mediante el uso de

una función definida en los números reales, la función de distribución.

Def. Sea X una variable aleatoria definida en el espacio de probabilidad (S, $\Im$, P (·)). La función F: $\Re \rightarrow$ [0 , 1] definida por:

$$F(x) = P(X(s) \leq x) = P[\{s: X(s) \leq x\}]$$

se denomina *función de distribución* de la variable aleatoria X.

La función de distribución de una variable X cumple las siguientes propiedades:

1) es no decreciente: para $a < b$ siempre $F(a) \leq F(b)$

2) es continua por la derecha:

$$\lim_{h \to 0} F(x + h) = F(x)$$

3)

$$F(-\infty) = \lim_{h \to -\infty} F(x) = 0$$

4)

$$F(+\infty) = \lim_{h \to +\infty} F(x) = 1$$

Consideremos una vez más) el experimento de lanzar un par de monedas, con X definida como el número de caras. Entonces la función de distribución de X estaría dada por:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/4 & 0 \leq x < 1 \\ 3/4 & 1 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

Variables aleatorias discretas

Consideramos ahora variables aleatorias tales que el número de resultados del experimento al que están asociadas no es necesariamente finito, pero es *contable,* es decir, los resultados pueden ser puestos en correspondencia con los números naturales (enteros positivos).

Para dichas variables podemos enumerar cada resultado del espacio muestral y la probabilidad a éste asociada.

Def. La función de cuantía $P_X(x)$ de una variable aleatoria discreta se define como

$$P_X(x) = P(X = x)$$

para cada x perteneciente al espacio muestral.

Podemos notar que:

$$\sum_x P_x(x) = 1$$

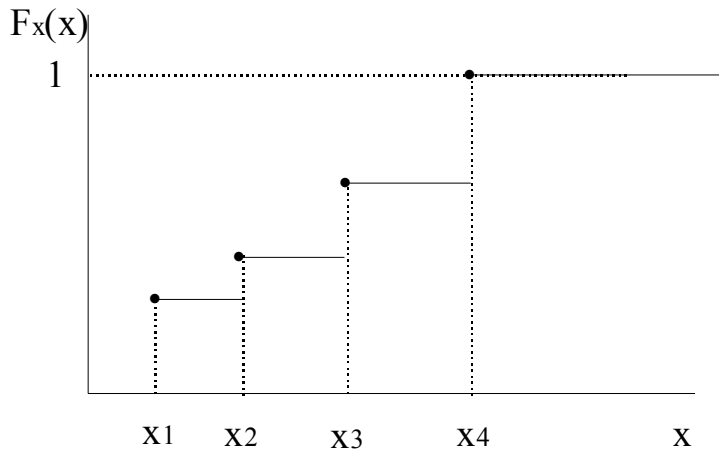
donde la suma se realiza sobre todos los valores posibles de X.

La representación gráfica de una función de cuantía es un diagrama de barras exactamente igual que el que se presentó en estadística descriptiva.

La función de distribución F_x (x) de una variable aleatoria discreta se define como:

$$F_x(x_0) = \sum_{x \le x_0} P_x(x) = P(x \le x_0)$$

Usaremos la letra mayúscula para denotar a las variables aleatorias y la minúscula para cualquiera de los valores que toma. Si graficamos la función vemos que tiene forma escalonada.



Variables aleatorias continuas

Imaginemos un experimento aleatorio tal que su resultado puede ser razonablemente descrito por cualquier número real (como puede ser el caso de muchas variables económicas). Aunque

se puede argumentar que las variables siempre están medidas en unidades de naturaleza discreta (como pesos y centavos), es un hecho que cuando el número de los posibles valores es muy alto, una aproximación continua es mucho más conveniente.

El espacio de posibles resultados es ahora infinito e incontable, y esto cambia la forma en que se realiza la atribución de probabilidades a los eventos y a los valores reales que los describen. La atribución de probabilidades no se realiza a puntos en particular como en el caso de las variables discretas y se realiza sobre intervalos de los reales, utilizando las funciones de *densidad* y *distribución*.

Una variable aleatoria X se define *continua* si existe una función f(x) tal que

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

para cada número real x. La función f(x) recibe el nombre de *función de densidad* de la variable aleatoria X.

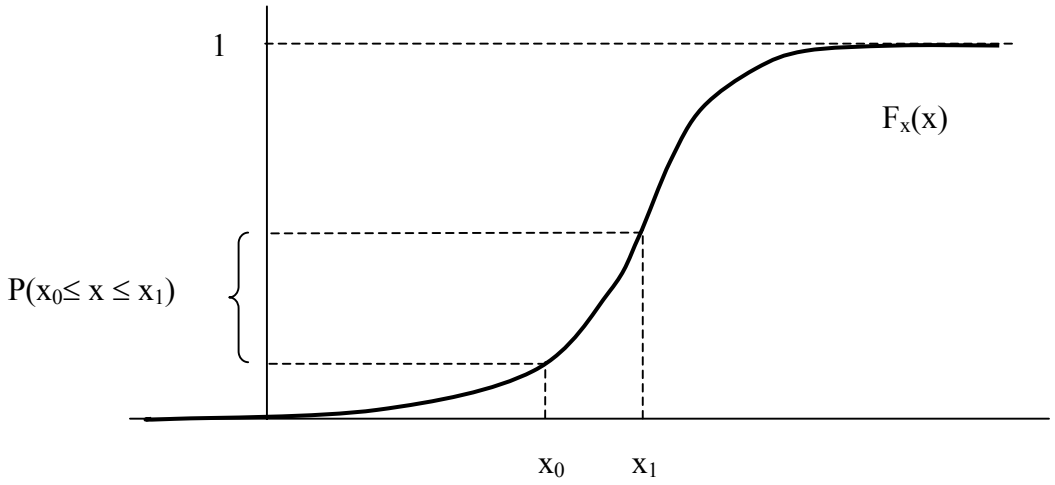
La función de densidad puede verse como una función que distribuye masa de probabilidad sobre los distintos intervalos de la recta real. La probabilidad de un punto en particular resulta ahora irrelevante (de hecho para una variable aleatoria continua es igual a cero) e interesan en cambio las probabilidades de intervalos. La función de densidad nos indica cómo está cambiando la probabilidad acumulada en cada punto, pero no es en sí misma una probabilidad, y puede tomar valores mayores que 1.

La función de densidad cumple con dos propiedades:

1) $f(x) \geq 0$

2) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

Por su parte la función de distribución F(x) es no negativa, no decreciente, y su valor cuando x tiende a $+\infty$ es 1. La función de distribución en el caso de las variables aleatorias continuas también caracteriza la probabilidad de que la variable aleatoria tome valores menores o iguales que los de un valor dado de x, pero ahora, en lugar de construirse como una suma de las probabilidades de puntos, se obtiene como la integral para los valores menores o iguales a cada x de la función de densidad. El gráfico siguiente muestra cómo la probabilidad acumulada crece hasta converger al valor 1. Para cualquier x real, la altura en el gráfico mostrará la probabilidad acumulada $P(X \leq x)$.



La probabilidad de que X tome valores en el intervalo (x_o , x_1) está dada por la resta de los valores de la función de distribución en los extremos del intervalo (que en el gráfico puede medirse como la distancia vertical entre las ordenadas en ambos puntos):

$$P\ (x_o < X \leq x_1) = F(x_1) - F(x_o)$$

Se considera a X mayor estricto que x_0 y menor o igual que x_1 porque el cálculo implica la resta de las probabilidades de X menor o igual que x_1 y que x_0 respectivamente, de modo que el punto específico x_0 no queda incluido. Si se toma en cuenta que en el caso de las variables continuas la probabilidad $P(X = x) = 0$, ello no tiene consecuencias para los cálculos.

Aquí es donde se muestra la importancia de las relaciones entre densidad y función de distribución. Partiendo de que

$$P(X \leq x_0) = F_X(x_0) = \int_{-\infty}^{x_0} f(x)\,dx$$

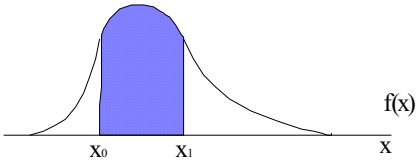
se obtiene que

$$P(X > x_0) = \int_{x_0}^{+\infty} f(x)\,dx = 1 - F_X(x_0)$$

Asimismo, la probabilidad de observar a la variable X en determinado intervalo (x_0, x_1] puede ser vista como la integral de la función de densidad en el intervalo considerado:

$$P(x_0 < X \leq x_1) = F_X(x_1) - F_X(x_0) = \int_{-\infty}^{x_1} f(x)\,dx - \int_{-\infty}^{x_0} f(x)\,dx = \int_{x_0}^{x_1} f(x)\,dx$$

utilizando aditividad respecto del intervalo de integración. La integral equivale al área entre ambos puntos bajo la función de densidad, como se muestra en el siguiente gráfico:



Esta representación gráfica de la probabilidad de un intervalo como un área recuerda los histogramas presentados en el contexto de estadística descriptiva. Esto guarda relación en cierta medida con la motivación del desarrollo de los conceptos de la teoría probabilística, en que las probabilidades pueden verse como surgiendo una idealización de la idea de frecuencias relativas de intervalos.

Como conclusión, partiendo de nuestro espacio de probabilidad original, hemos sustituido a nuestro espacio muestral original S por la recta real, de manera que los eventos quedan representados por distintos valores que toma una variable aleatoria. A su vez, la clase de los eventos $\Im$ ha sido sustituida por la clase de Borel, a la cual nos aproximamos mediante el conjunto de los intervalos semiabiertos $(-\infty , x]$. La función $P(\cdot)$, que asigna probabilidades, está ahora definida sobre intervalos de la recta real. Su dominio sigue siendo un conjunto. Sin embargo, las funciones de distribución, densidad y cuantía son funciones cuyo dominio es el conjunto de los números reales (y por lo tanto permiten un manejo mucho más accesible con los métodos del cálculo) para describir en términos probabilísticos las variables aleatorias.

Momentos

Para el estudio matemático de la distribución de las variables aleatorias, son importantes algunas características de las mismas que llamamos *momentos*. Para su análisis hacemos uso de la noción de *valor esperado*.

Es posible ver al *valor esperado* de una variable aleatoria como una forma idealizada de la media. Para una variable aleatoria discreta el valor esperado se define como:

$$\mu = E(X) = \sum_x x P_x(x)$$

es decir, una suma ponderada de los valores de la variable, en la que los ponderadores son las probabilidades de cada uno de los valores. X puede tomar diferentes valores, pero E(X) es una constante [2].

Para una variable aleatoria continua, el valor esperado se define de la siguiente manera:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx$$

aquí el lugar de la sumatoria lo toma una integral, donde ponderamos a todos los valores reales de x por la densidad.

En muchos casos es de interés averiguar la distribución de probabilidad de una función de una variable aleatoria o de un vector aleatorio. A partir de una variable aleatoria X, a través de una función $g(\cdot)$ obtenemos $Y = g(X)$. Nos interesará saber en qué condiciones $Y = g(X)$ será también una variable aleatoria y si es posible derivar la función de distribución o de densidad de Y a partir del conocimiento de g y de la distribución de X.

No será posible dar una respuesta detallada a ambas cuestiones en el marco de este curso, sino que nos limitaremos a llamar la atención sobre la existencia de este problema y a dar algunas indicaciones sobre resultados útiles que involucran distribuciones de funciones de variables aleatorias y sus momentos. Nos basta con señalar que para que las funciones de variable aleatoria sean a su vez variables aleatorias deben preservar la estructura de eventos en la clase $\Im$ en la que están definidas las probabilidades correspondientes a la variable o vector original. No estudiaremos la técnica para derivar la distribución de Y a partir de $g(\cdot)$ y de la distribución de la X. Sólo mencionaremos que dentro de las funciones, la clase de las funciones monótonas (las que siempre decrecen o siempre crecen con x) permitirán un cálculo más sencillo de las densidades de $Y = g(X)$, ya que admiten en general la función inversa $X = g^{-1}(Y)$ y ello nos ayuda a rastrear las probabilidades de los intervalos correspondientes.

Del mismo modo que calculamos E(X), podemos calcular $E(Y) = E(g(X))$. En lugar de hallar la densidad de Y y promediar con respecto a esta densidad, podemos promediar g(X) con respecto a la densidad de X. En forma general, el valor esperado de una función de una variable aleatoria discreta X, digamos g(X), se define como:

---

[2] El nombre puede no ser del todo claro, ya que el valor esperado no es necesariamente uno de los valores posibles de la variable aleatoria. En el caso de la variable "número de la cara superior de un dado", por ejemplo, E(X) = 3.5, un valor que nunca podemos "esperar" observar. Más confuso aún es el término "esperanza" que a veces se aplica al traducir directamente del inglés *expectation*, que se asocia a la idea de expectativa.

$$E(g(X)) = \sum_x g(x) P_x(x)$$

A su vez, el valor esperado de una función g(X) de una variable aleatoria continua X se define como:

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x)\, dx$$

Una función lineal de X tiene la forma general Y = a + bX, donde a y b son constantes. Una propiedad del valor esperado es la "linealidad":

$$E\,(a + bX) = a + bE(X).$$

La varianza de una variable aleatoria está dada por el valor esperado de la desviación de la media al cuadrado:

$$\sigma^2 = \mathrm{Var}(X) = E(X-\mu)^2$$

con lo cual obtenemos para el caso discreto:

$$\sigma^2 = \sum_x (x-\mu)^2 P_x(x)$$

y para el caso continuo:

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx$$

Una fórmula alternativa surge si desarrollamos el cuadrado, utilizando la propiedad de linealidad y el hecho de que la media es una constante. Para cualquier constante c, E(c) = c.

$$E(X-\mu)^2 = E(X^2 - 2X\mu + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2$$

es decir, la varianza es "el valor esperado del cuadrado menos el cuadrado del valor esperado" de una variable aleatoria. La desviación standard σ se define como la raíz cuadrada de la varianza, en forma semejante a lo definido en el capítulo dedicado a estadística descriptiva.

Utilizando la propiedad de linealidad obtenemos que

$$\mathrm{Var}\,(a + bX) = b^2 \mathrm{Var}\,(X)$$

En general, funciones de la forma:

$$\mu'_r = E(X^r)$$

reciben el nombre de *momentos* de la distribución. La media es el primer momento de la distribución (r = 1). Las funciones de tipo:

$$\mu_r = E(X - \mu)^r$$

son los *momentos centrados* de la distribución, de los cuales la varianza es el segundo (r = 2), en tanto que el tercero y cuarto nos permiten extender las definiciones de asimetría y kurtosis discutidas en el contexto de estadística descriptiva.

Desigualdad de Tschebyscheff

Esta relación establece que para una variable aleatoria X, con media : y varianza Φ^2, se cumple que:

$$P(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}$$

para cualquier $\varepsilon < 0$. Por ejemplo, si tomamos $\varepsilon = \sigma m$, podemos interpretar que la probabilidad de que los valores de una variable aleatoria disten más de m desviaciones standard de la media es menor a $1/m^2$. De allí podemos deducir que la probabilidad de que x diste menos de m desviaciones standard de la media es mayor que $1-1/m^2$. Así, en un radio de 2 desviaciones standard de la media se concentra al menos un 75% de la masa de la distribución, en un radio de 3 desviaciones standard un 89%, etc.. Esto se cumple independientemente de la forma de la función de distribución.

5. Vectores aleatorios

Consideremos un experimento aleatorio en el que los resultados pueden ser descritos por un vector de atributos cuantitativos. En general es el caso que en encuestas como las de hogares se releve información sobre un conjunto de atributos de cada unidad encuestada, como el número de integrantes, edades, escolaridad, ingresos, etc. Aquí nos concentraremos en el caso de dos variables o atributos porque permite comprender los elementos centrales fácilmente generalizables a más dimensiones. Asociamos el par de variable aleatorias (X_1, X_2) a cada resultado perteneciente al espacio S de nuestro experimento.

Def. Un vector aleatorio es una función $\mathbf{X} : S \rightarrow \Re^2$ tal que para cualquier par de números reales (x_1 , x_2) = **x**, el conjunto

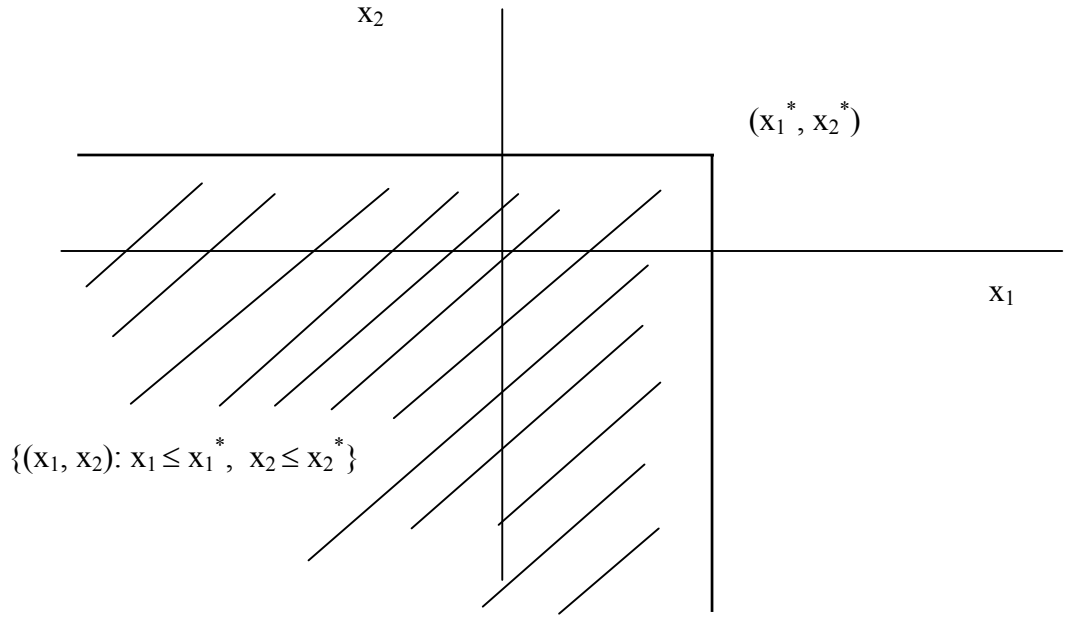
$$\{s: -\infty < X_1(s) \leq x_1 , -\infty < X_2(s) \leq x_2 \} \in \Im$$

Como ejemplo, consideremos el experimento consistente en lanzar dos monedas, con el espacio muestral asociado S= {CC, CN, NC, NN}. Definamos las funciones X_1 como

"número de caras obtenido" y X_2 como "número de números obtenido", de manera que **X** queda definido como:

$$(X_1(CC), X_2(CC)) = (2,0)$$
$$(X_1(CN), X_2(CN)) = (1,1)$$
$$(X_1(NC), X_2(NC)) = (1,1)$$
$$(X_1(NN), X_2(NN)) = (0,2)$$

De modo que a cada elemento de S queda asociado uno y solo un punto de $\Re^2$. Para cualquier punto $(x_1, x_2) \in \Re^2$ consideramos el intervalo $((-\infty, x_1\,] , (-\infty, x_2\,])$ que es un rectángulo semiabierto como se ve en el gráfico:



Siguiendo con nuestro ejemplo, debemos recorrer $\Re^2$ de modo de verificar si los conjuntos de elementos de S que a través del vector nos dan puntos en el plano con coordenadas respectivamente menores o iguales que cada punto (x_1, x_2) son efectivamente eventos. Una forma de verlo es considerar una tabla en la cual se consideran intervalos en el plano de acuerdo a los valores que toma el vector aleatorio, recorriendo todos los posibles valores de x_1 y x_2.

$\{s \in S : X_1(s) \le x_1, X_2(s) \le x_2\}$				
	$x_1 < 0$	$0 \le x_1 < 1$	$1 \le x_1 < 2$	$2 \le x_1$
$2 \le x_2$	$\varnothing$	{(NN)}	{(CN),(NC),(NN)}	S
$1 \le x_2 < 2$	$\varnothing$	$\varnothing$	{(CN),(NC)}	{(CN),(NC),(CC)}
$0 \le x_2 < 1$	$\varnothing$	$\varnothing$	$\varnothing$	{(CC)}
$x_2 < 0$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$

Si consideramos al σ–álgebra de eventos que surge de considerar los conjuntos de resultados que dan diferentes valores a través del vector aleatorio, sus uniones y complementos, obtenemos la clase:

$\Im$ = {{(NN)}, {(CN),(NC)}, {(CC)}, {(CN),(NC),(NN)}, {(CN),(NC),(CC)}, {(NN),(CC)}, ∅, S}

con lo que podemos comprobar que para cada punto (x_1,x_2) en $\Re^2$ el conjunto de preimágenes en S a los que el vector asocia coordenadas menores o iguales respectivamente que x_1 y x_2 pertenecen a la clase $\Im$, de modo que se asegura que son eventos. Ello nos permite transformar a las probabilidades anteriormente correspondientes a los eventos en probabilidades de que el vector aleatorio **X** tome valores en un intervalo dado de $\Re^2$.

Distribuciones conjuntas

Definición: Si $\mathbf{X} = (X_1, X_2)$ es un vector aleatorio definido en $(S, \Im, P(\cdot))$, la función definida por $F(\cdot\ ,\cdot)$: $\Re^2 \rightarrow [0, 1]$ tal que

$F(\mathbf{x}) = F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$

se denomina la *función de distribución conjunta* del vector aleatorio **X**.

La función de distribución conjunta tiene las propiedades de ser monótona y no decreciente en cada variable por separado, así como:

i. $F(x_1, x_2) = \lim_{x_2 \rightarrow -\infty} F(x_1, x_2) = 0$

ii. $\lim_{x_1, x_1 \rightarrow +\infty} F(x_1, x_2) = 1$

Vectores aleatorios discretos

Def. La distribución conjunta de X_1 y X_2 se denomina *distribución discreta* si existe una función $P(\cdot,\cdot)$ tal que:

$$P(x_1, x_2) \geq 0\ ,\ (x_1, x_2) \in \Re^2$$

que toma el valor cero en todas partes excepto en un número finito o infinito contable de puntos en el plano, en los cuales cumple con:

$$P(x_1, x_2) = P(X_1 = x_1, X_2 = x_2)$$

En el caso discreto la función de distribución conjunta se define como:

$$F(x_1^*, x_2^*) = \sum_{x_1 \le x_1^*} \sum_{x_2 \le x_2^*} P(x_1, x_2)$$

Si continuamos con nuestro ejemplo de las dos monedas, podríamos deducir la función de distribución conjunta del vector aleatorio X_1 y X_2 tratando de encontrar para cada posible par (x_1 , x_2) en $\Re^2$ la probabilidad conjunta del evento $\{X_1 \le x_1 , X_2 \le x_2 \}$. Ello nos es facilitado porque ya encontramos los eventos correspondientes cuando comprobamos si nuestra función cumplía con la definición de variable aleatoria. Así obtenemos:

$F(x_1, x_2) = P (X_1 \le x_1, X_2 \le x_2)$				
	$x_1 < 0$	$0 \le x_1 < 1$	$1 \le x_1 < 2$	$2 \le x_1$
$2 \le x_2$	0	1/4	3/4	1
$1 \le x_2 < 2$	0	0	2/4	3/4
$0 \le x_2 < 1$	0	0	0	1/4
$x_2 < 0$	0	0	0	0

Vectores aleatorios continuos

La función de distribución conjunta de X_1 y X_2 se denomina continua si existe una función f $(\cdot,\cdot)$ tal que:

$$f(x_1, x_2) \ge 0 , (x_1,x_2) \in \Re^2$$

que cumple que:

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(u,v)\, du\, dv$$

La función de distribución conjunta, al igual que la densidad conjunta, definen una superficie en $\Re^3$, ya que asocian a cada punto en $\Re^2$ un tercero, que podríamos representar como una altura sobre el plano, y las probabilidades de intervalos quedan representadas por volúmenes bajo la superficie de la densidad. El cálculo de probabilidades a través de la distribución conjunta de vectores bivariados, que para variables aleatorias continuas involucra integrales dobles, va más allá del alcance de este curso.

Distribuciones marginales y condicionales

Cuando analizamos un vector aleatorio $\mathbf{X} = (X_1, X_2)$ surge la pregunta de si podemos separar las variables aleatorias y considerarlas como tales individualmente. Ello nos conduce al

concepto de *distribución marginal*. Las distribuciones marginales de X_1 y X_2 quedan definidas por:

$$F_1(x_1) = \lim_{x_2 \to +\infty} F(x_1, x_2)$$

$$F_2(x_2) = \lim_{x_1 \to +\infty} F(x_1, x_2)$$

Al considerar la probabilidad conjunta del evento $\{X_1 \leq x_1 , X_2 \leq x_2 \}$ cuando una de las dos variables, por ejemplo X_2, tiende a infinito, consideramos la ocurrencia conjunta del evento $\{X_1 \leq x_1 , X_2 \leq +\infty \}$, que es como considerar solamente el caso $\{X_1 \leq x_1\}$, ya que es seguro que $\{X_2 \leq +\infty \}$. Deja de interesarnos en este caso qué ocurre con la componente X_2, perdemos la información de la distribución conjunta y volvemos al caso univariado.

Dado que en la definición de vector aleatorio hemos impuesto la condición que los conjuntos de preimágenes de los intervalos de tipo $((-\infty, x_1], (-\infty, x_2])$ en el plano pertenecieran a la clase de los eventos, puede demostrarse que X_1 y X_2 son efectivamente variables aleatorias, cada una por separado, y puede obtenerse la *cuantía o densidad marginal* en cada caso. En el caso discreto ello equivale a sumar con respecto a la otra variable:

$$P_1(x_1) = \sum_i P(x_1, x_{2i})$$

En el caso continuo la definición se realiza en términos de las funciones de densidad, y podemos definir las *densidades marginales* como

$$f_1(x_1) = \int_{-\infty}^{+\infty} f(x_1, x_2) dx_2$$

y

$$f_2(x_2) = \int_{-\infty}^{+\infty} f(x_1, x_2) dx_1$$

Como ejemplo, consideremos una tabla de frecuencia conjunta de observación de dos variables, en la cual se han tabulado las frecuencias relativas conjuntas de los casos observados en tres tramos de ingresos (X_1) y tres tramos de edad (X_2)

	X_2			
X_1	1	2	3	$p_1(x_1)$
1	0.250	0,020	0.005	0.275
2	0.115	0.325	0.120	0.560
3	0.035	0.055	0.075	0.165
$p_2(x_2)$	0.400	0.400	0.200	1.000

Las cuantías marginales de X_1 y X_2 quedan representadas en los totales por fila y por columna de las celdas de la tabla, que se refieren a la probabilidad de que una persona seleccionada al azar pertenezca a cada grupo de edad o de ingreso.

Independencia

Conociendo la densidad conjunta es posible derivar las marginales. Las distribuciones marginales en general no contienen información referida a la distribución conjunta de ambas variables salvo en un caso particular, en el que es posible derivar la distribución conjunta a partir de las marginales, y es cuando:

$$f(x_1, x_2) = f_1(x_1) \cdot f_2(x_2)$$

o en su caso

$$P(x_1, x_2) = P_1(x_1) \cdot P_2(x_2)$$

en cuyo caso se dice que las variables aleatorias X_1 y X_2 son independientes. En términos de la función de distribución la independencia se expresa:

$$F(x_1, x_2) = F_1(x_1) \cdot F_2(x_2)$$

La interpretación es exactamente la misma que se realizaba al estudiar la independencia de eventos en el marco de la teoría probabilística. Dos variables aleatorias son independientes si la probabilidad de que una de ellas tome valores en determinado intervalo arbitrario no se ve afectada por que la otra lo haga en cualquier otro intervalo.

Finalmente consideramos la posibilidad de simplificar nuestro modelo de probabilidad *condicionando* sobre un subconjunto de las variables aleatorias consideradas. En el caso de la marginalización, toda la información pertinente a otras variables se perdía mientras que aquí aparecerá bajo la forma de valor que toma la variable condicionante. Del mismo modo que en el contexto de un espacio de probabilidad definíamos la probabilidad condicional de un evento A dado otro evento B como

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Podemos considerar:

$$P(X_1 \le x_1 / B) = \frac{P(X_1 \le x_1 \cap B)}{P(B)} = \frac{P(\{s : X_1(s) \le x_1\} \cap B)}{P(B)}$$

En el caso de una variable aleatoria discreta lo anterior es inmediato, pues para un evento B = {s: $X_2(s) = \tilde{x}_2$ } se puede definir en forma análoga la función de cuantía condicional:

$$P_{X1/X2}(x_1 / X_2 = \tilde{x}_2) = P(X_1 = x_1 / X_2 = \tilde{x}_2) = \frac{P(X_1 = x_1, X_2 = \tilde{x}_2)}{P(X_2 = \tilde{x}_2)} = \frac{P(x_1, \tilde{x}_2)}{P_2(\tilde{x}_2)}$$

Usamos el tilde para enfatizar que se trata de un valor particular que toma la variable X_2. En el caso de la variable continua debe notarse que P(X = x) = 0 para todo valor de la variable aleatoria, por lo cual matemáticamente se trata de una cuestión compleja. Por lo mismo solamente se da la definición de densidad condicional para el caso continuo de densidad condicional de X_1 dado $X_2 = x_2$, de la siguiente manera:

$$f_{X1/X2}(x_1 / X_2 = x_2) = \frac{f(x_1, \tilde{x}_2)}{f_2(\tilde{x}_2)}$$

siempre y cuando $f_2(x_2) > 0$. Notamos que a medida que varía X_2, vamos obteniendo una diferente densidad condicional para cada uno de los distintos valores.

Momentos

Cuando se trata de estudiar los momentos de la distribución de un vector aleatorio (X_1 , X_2) el razonamiento es análogo al realizado para momentos univariados, pero la distribución de partida es la distribución conjunta del vector.

En forma general, si tenemos una función de un vector aleatorio g (X_1 , X_2), podemos definir el valor esperado de dicha función en el caso discreto como

$$E(g(X_1, X_2)) = \sum_{x_1} \sum_{x_2} g(x_1, x_2) P_{X_1, X_2}(x_1, x_2)$$

y en el caso continuo como:

$$E(g(X_1, X_2)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x_1, x_2) f_{X_1, X_2}(x_1, x_2) \, dx_1 dx_2$$

En particular, cuando g (x_1, x_2) = x_1, pueden recuperarse los momentos univariados a partir de las distribuciones conjuntas obteniéndose para el caso discreto la media:

$$E(X_1) = \sum_{x_1}\sum_{x_2} x_1 P_{X_1,X_2}(x_1,x_2) = \sum_{x_1} x_1 \sum_{x_2} P_{X_1,X_2}(x_1,x_2) = \sum_{x_1} x_1 P_1(x_1)$$

y la varianza:

$$V(X_1) = \sigma^2_{X1} = \sum_{x_1}\sum_{x_2} (x_1 - \mu_{X1})^2 P_{X_1,X_2}(x_1,x_2) = \sum_{x_1} (x_1 - \mu_{X1})^2 P_1(x_1)$$

En el caso de los momentos de las variables aleatorias continuas, los cálculos involucran integrales en lugar de sumatorias, pero la interpretación de los mismos se mantiene.

De los momentos conjuntos únicamente incorporaremos como concepto nuevo el de covarianza, que en se define como el valor esperado del producto de ambas desviaciones respecto de la media:

$$\text{Cov}(X_1, X_2) = \sigma^2_{X1X2} = E[(X_1 - \mu_{X1})(X_2 - \mu_{X2})] =$$

$$E[X_1X_2 - \mu_{X1}X_2 - \mu_{X2}X_1 + \mu_{X1}\mu_{X2}] = E[X_1X_2] - \mu_{X1}\mu_{X2}$$

En el caso discreto se trata de una suma ponderada de los productos de los desvíos de ambas medias, en que los ponderadores están dados por las cuantías conjuntas.

$$\text{Cov}(X_1, X_2) = \sum_{x_1}\sum_{x_2} (x_1 - \mu_{X1})(x_2 - \mu_{X2}) P_{X_1,X_2}(x_1,x_2)$$

En el caso continuo deberá calcularse la integral doble correspondiente.

La covarianza queda expresada en "unidades de x_1 por unidades de x_2". Al normalizarla dividiendo por las desviaciones standard de ambas variables obtenemos el coeficiente de correlación ρ_{X1X2}, libre de unidades de medida, comprendido siempre entre –1 y 1. La definición de correlación guarda semejanza con la que se estudió en estadística descriptiva. El coeficiente de correlación se define entonces como

$$\text{Corr}(X_1, X_2) = \rho_{X1X2} = \frac{\sigma_{X1X2}}{\sigma_{X1}\sigma_{X2}}$$

y se interpreta como una medida de asociación lineal entre variables aleatorias. Si las variables están en una relación exactamente lineal, el coeficiente de correlación tomará el valor 1 o –1. En ausencia de relación lineal entre las variables el coeficiente de correlación vale cero y las variables están *incorrelacionadas*.

Veremos además como ejemplos los momentos de la suma de variables aleatorias. Dadas dos variables aleatorias X_1 y X_2, con su distribución conjunta, consideremos la suma $X_1 + X_2$. Esta suma será también una variable aleatoria y para el cálculo de sus momentos debemos hacer uso de la distribución conjunta de los sumandos. Se define, para el caso discreto,

$$E(X_1 + X_2) =$$

$$\sum_{x_1}\sum_{x_2}(x_1 + x_2)P_{X_1,X_2}(x_1,x_2) = \sum_{x_1}\sum_{x_2}x_1 P_{X_1,X_2}(x_1,x_2) + \sum_{x_1}\sum_{x_2}x_2 P_{X_1,X_2}(x_1,x_2) = \mu_{X1} + \mu_{X2}$$

El valor esperado de la suma es la suma de los valores esperados. El resultado para variables continuas se omite pero es análogo. A su vez estudiaremos la varianza de una suma de variables aleatorias. Comenzamos notando que:

$$V(X_1 + X_2) = E(X_1 + X_2 - E(X_1 + X_2))^2 = E[(X_1 + X_2)^2] - [E(X1 + X2)]^2 =$$

$$= E[X_1^2 + 2X_1X_2 + X_2^2] - [\mu_{X1}^2 + 2\mu_{X1}\mu_{X2} + \mu_{X2}^2]$$

de modo que

$$V(X_1 + X_2) = E(X_1^2) - \mu_{X1}^2 + 2E(X_1X_2) - 2\mu_{X1}\mu_{X2} + E(X_2^2) - \mu_{X2}^2 =$$

$$= V(X_1) + V(X_2) + 2Cov(X_1, X_2)$$

La varianza de la suma es la suma de las varianzas más dos veces la covarianza. Será igual a la suma de las varianzas *si y solo si las variables están incorrelacionadas*, en cuyo caso la covarianza será igual a cero. Las reglas obtenidas pueden generalizarse de la siguiente manera:

$$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n)$$

Asimismo, si las variables están incorrelacionadas de a pares, también se cumple que

$$V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n)$$

Generalizando las propiedades de linealidad del valor esperado puede obtenerse:

$$E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$$

y a su vez:

$$V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2abCov(X_1, X_2)$$

Valor esperado condicional. Regresión

En el caso de una variable aleatoria discreta, el valor esperado condicional está dado por la expresión:

$$E(X_1 / X_2 = x_2) = \sum_{x_1} x_1 P_{X_1/X_2}(x_1 / X_2 = x_2)$$

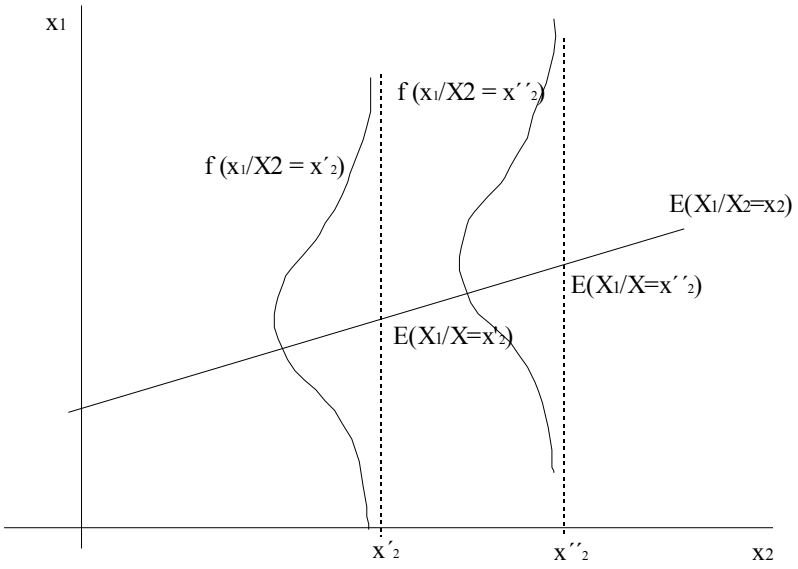
A su vez, cuando se trata de una variable continua, tendremos:

$$E\,(X_1/\ X_2 = x_2) = \int_{-\infty}^{+\infty} x_1 f_{X_1/X_2}\left(x_1 / X_2 = x_2\right) dx_1$$

Como puede verse, en ambos casos tenemos una función de x_2. Con cada valor que toma x_2 hay una nueva distribución condicional de la X_1, que tiene asociado un diferente valor esperado condicional. Esta es una función:

$$E\,(X_1 /X_2 = x_2) = h\,(x_2)\ ,$$

que podemos graficar en el plano, asociando a cada valor de x_2 el valor esperado condicional de X_1 . Esta curva recibe el nombre de *curva de regresión* de X_1 en x_2.



6. Modelos de probabilidad

Una vez definido el concepto de variable aleatoria y estudiada la forma de caracterizar su distribución de probabilidad, vemos que ello nos permite simplificar el manejo de la incertidumbre asociada a los resultados de cierto experimento aleatorio. El problema original, la incertidumbre respecto al resultado de un experimento, se transforma ahora en incertidumbre respecto a los valores que toma una variable aleatoria. Basados en las funciones de distribución y las de densidad o cuantía, hemos descrito en términos probabilísticos el comportamiento de dichas variables y vectores aleatorios.

El paso siguiente consiste en utilizar formas funcionales tipo de distribución, densidad o

cuantía, definibles en forma algebraica, para caracterizar en forma completa *modelos de probabilidad*. Estos modelos también son conocidos como *familias paramétricas de funciones de densidad* (o cuantía), o *distribuciones de probabilidad*. Dichos modelos se conciben como una descripción ideal del proceso aleatorio que genera los datos: cuando se elige determinada familia paramétrica de densidades para construir un modelo de determinado fenómeno, se está suponiendo que los datos observados son generados por el mecanismo aleatorio descrito por dichas densidades (o cuantías). Todas las densidades de una determinada familia comparten una forma funcional común, pero difieren en una serie de valores que las definen en forma completa, y que reciben el nombre de *parámetros*.

Parámetros. Espacio paramétrico

Definir un modelo de probabilidad implica proponer una forma funcional concreta para la función de distribución o densidad. Dicha forma funcional, además de explicitar la dependencia de x, dependerá de un conjunto de cantidades que conocemos como *parámetros*, que usualmente designamos con la letra griega θ. De este modo escribiremos el modelo de probabilidad como:

$$\Phi = \{f(x, \theta), \theta \in \Theta\}$$

Cada modelo de probabilidad comprende a un conjunto de funciones de densidad, que comparten dos características esenciales. En primer lugar, tienen en común una forma funcional dada $f(\cdot)$. En segundo lugar, dependen de un vector de parámetros desconocidos θ. Los parámetros de estas densidades a su vez pertenecen a un conjunto de valores posibles Θ, que denominamos *espacio paramétrico*. La elección de un valor para θ determina en forma única una densidad particular.

Si consideramos al proceso aleatorio que genera los datos como gobernado por la ley probabilística que describe la familia Φ, la incertidumbre sobre los valores que toma una variable aleatoria hace foco ahora sobre la incertidumbre con respecto a cuáles son los parámetros de las funciones de densidad o cuantía. Ello se debe a que, a partir de un modelo de probabilidad dado, con el conocimiento de dichos parámetros podremos caracterizar completamente en forma probabilística el fenómeno estudiado. Sin embargo, los parámetros por lo general no son observados directamente, y esto plantea una serie de problemas que serán abordados por la inferencia estadística.

La elección de una familia paramétrica en particular para modelar determinado fenómeno es crucial. Esta decisión dependerá de la experiencia previa con fenómenos similares o de un análisis preliminar de los datos.

A continuación se presenta un conjunto de familias de distribuciones de probabilidad de uso frecuente.

Distribuciones discretas

El conjunto de distribuciones que se presentan a continuación están caracterizadas por funciones de cuantía discretas y un espacio muestral finito o al menos contable.

Distribución Bernoulli

Se considera un experimento aleatorio en el que hay sólo dos resultados posibles, que convencionalmente denominamos "éxito" y "fracaso", es decir S = {E, F}. Si definimos en S la variable aleatoria X de forma que X(E) = 1 y X(F) = 0, y postulamos las probabilidades de éxito y fracaso como fijas, P(X = 1) = p y P(X = 0) = 1 – p, se puede deducir que la función de cuantía de X está dada por:

$$P_X(x) = \begin{cases} p^x (1-p)^{1-x} & x = 0, 1 \\ 0 & \text{en otro caso.} \end{cases}$$

El único parámetro de esta distribución es p, y escribimos X ~ Bernoulli (p) ("X sigue una distribución de Bernoulli con parámetro p").

El interés de este modelo está en que describe un caso muy simple que sirve de base para estudiar situaciones más complejas y realistas en las cuales el experimento sigue siendo extraer cierto elemento de una población y observar si posee o no un atributo dado.

Utilizando las definiciones dadas pueden obtenerse la media y varianza de esta distribución:

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

$$V(X) = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p\,(1 - p)$$

Si consideramos repetir los experimentos, aproximándonos a una visión frecuencista de la probabilidad, la media tendría la interpretación de la proporción esperada de éxitos.

Distribución Binomial

Para generalizar los resultados del modelo anterior, consideremos ahora una sucesión de n pruebas de Bernoulli, que se realizan en idénticas condiciones. Las pruebas se consideran independientes (en términos de probabilidad, si consideramos el suceso E_i, éxito en la prueba i, entonces $P(E_i/E_j) = P(E_i)$ para $i \neq j$). Nuestro espacio muestral está compuesto de todas las n-uplas posibles compuestas de E y F. La probabilidad de éxito es fija para todas las pruebas e igual a p. Consideremos a una variable aleatoria X = número de éxitos obtenidos.

Para determinar su cuantía, comenzamos por evaluar la probabilidad de una secuencia en

particular de x "éxitos" y n–x "fracasos". Por la independencia, esta probabilidad está dada por la multiplicación:

$$p^x(1-p)^{n-x}$$

La probabilidad de la ocurrencia conjunta es la multiplicación de las probabilidades de los eventos en particular. Sin embargo, esta no es la probabilidad del evento "x éxitos en n pruebas", dado que existen muchas formas en que este evento puede ocurrir, tantas como posibles órdenes de x éxitos y n–x fracasos se pueden construir. La probabilidad del evento "x éxitos en n pruebas" surge como l a suma de las probabilidades de cada una de las formas que tiene de ocurrir, ya que son eventos disjuntos entre sí. Contar el número de estos eventos equivale a determinar de cuántas maneras pueden ubicarse x éxitos en n pruebas, lo que corresponde a C^n_x. Por tanto multiplicamos a la probabilidad de cada uno de ellos por el número total de formas posibles, obteniendo la cuantía

$$P_X(x) = C^n_x p^x(1-p)^{n-x}$$

Los dos parámetros involucrados en esta distribución son n y p. La notación es X ~ B(x,n,p).

Otra forma de verlo es definir en el mismo espacio muestral n variables aleatorias, Y_1, Y_2,..., Y_n de modo que $Y_i = 1$ si se obtiene éxito en la prueba i, 0 si se obtiene fracaso en la prueba i. Podemos expresar a nuestra variable aleatoria como $X = \sum Y_i$. Aplicando los resultados vistos anteriormente sobre media y varianza de una suma de variables aleatorias independientes podemos obtener

$$E(X) = E(\sum Y_i) = \sum E(Y_i) = np$$

$$V(X) = V(\sum Y_i) = \sum V(Y_i) = np\,(1-p)$$

Este modelo se asocia claramente con la extracción de una muestra con reposición de una población dada (la selección de cada individuo para la muestra, verificando si posee cierto atributo es una prueba). Si el muestreo se realiza con reposición (luego de seleccionado un elemento éste es vuelto a considerar parte de la población a muestrear) entonces la probabilidad de observar el atributo p se mantiene incambiada de una prueba a la otra y los resultados de cada prueba (tiene o no tiene el atributo) son independientes entre sí.

Distribución hipergeométrica

Como símil del muestreo en busca de determinado atributo, consideremos n extracciones de un bolillero con 2 clases de bolillas. El modelo binomial resultaba apto para considerar el caso cuando luego de cada extracción se repone la bolilla extraída. Sin embargo si la extracción de n bolillas se realiza sin reposición, cada extracción modifica la proporción de éxitos en las bolillas que permanecen en la caja, de modo que se altera la probabilidad condicional de éxito en una prueba dado el resultado de otra anterior.

Supongamos que tomamos una muestra de tamaño n de una población total de N elementos. En términos de las bolillas, cada una de ellas es una prueba. Del total N, hay S que pertenecen a la clase que consideraremos éxito.

Sea X la variable aleatoria definida X = número de éxitos en n extracciones. Dado que extraemos sin reponer, es lo mismo pensar en extraer las bolillas de a una que las n de una sola vez. Será útil considerar el espacio de resultados del experimento como un conjunto de subconjuntos posibles de tamaño n. El número total de subconjuntos[3] de tamaño n está dado por el número C^N_n. Para encontrar la cuantía de X, procedemos a contar entre los resultados posibles del experimento, cuántos de ellos contienen x éxitos. Entre las S bolillas que se definen "éxitos", hay C^S_x formas de extraer x éxitos. Entre los restantes N-S bolillas que denominamos "fracasos", de la misma manera, pueden extraerse n–x de C^{N-S}_{n-x} formas posibles. El número de muestras conteniendo exactamente x éxitos y n–x fracasos es el producto de dichos números.

Para utilizar la fórmula que calcula la probabilidad como los casos favorables sobre los posibles debemos tener que los resultados son equiprobables. Ello es efectivamente así, ya que es razonable pensar que cada subconjunto de n bolillas tiene la misma probabilidad de ser extraído que cualquier otro. Por lo tanto, la probabilidad de extraer una muestra conteniendo x éxitos y n–x fracasos está dada por:

$$P_X(x) = C^S_x \, C^{N-S}_{n-x} / C^N_n$$

Los parámetros involucrados son tres: S, N y n (número de extracciones).

Distribución Poisson

La distribución Poisson da un modelo realista de diversos fenómenos aleatorios en que está involucrada una cuenta de eventos que ocurren en un continuo, como puede ser el tiempo o el espacio, medible en intervalos dados. Algunos ejemplos, podrían ser el número de accidentes por día en una carretera, el número de pedidos por semana en una fábrica, el número de defectos por unidad de longitud en un cable, etc.

La cuantía de la distribución Poisson es la siguiente:

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Para esta distribución se tiene que $E(X) = \lambda$ y a la vez $V(X) = \lambda$.

Para aplicar la distribución Poisson requerimos saber el número promedio de sucesos que

[3] No nos importa el orden, de ahí la fórmula. Para contar los éxitos y fracasos sólo importa si una bolilla está incluida en la extracción y no el orden en que salió.

ocurren en un intervalo dado, λ, único parámetro de esta distribución, a partir del cual se puede asignar probabilidades a la ocurrencia de cualquier número dado de sucesos.

Se puede observar también que este es un ejemplo de una distribución con un número infinito, aunque contable, de puntos en el espacio muestral. Por supuesto, las probabilidades deberán tender a cero a medida que el número de sucesos considerado se incrementa, de modo de asegurar que la suma de todas las probabilidades no exceda de 1.

Una forma de ver a la distribución Poisson es como una generalización de la distribución binomial, pero en las condiciones particulares en que n es muy grande mientras que p se hace muy pequeño. Consideremos el ejemplo de los vuelos de avión, que son centenares de miles en un intervalo de tiempo dado. La probabilidad de un accidente aéreo es muy, muy baja, pero dado el alto número de vuelos se puede esperar un número pequeño de accidentes en dicho intervalo. Esto sugiere que este modelo es especialmente aplicable a eventos que ocurren en un espacio continuo (tiempo, espacio), en el que en un intervalo finito hay infinidad de puntos, de los cuales sólo un muy pequeño número contienen dichos eventos (y por tanto éstos se conocen como *sucesos raros*). Si bien en principio el modelo binomial podría utilizarse para estos fenómenos, se vuelve complicado debido a los números extremadamente grandes y extremadamente pequeños que se manejarían. La distribución *Poisson* puede verse como el caso límite de la binomial cuando $n \to +\infty$ pero $np \to \lambda$ (fijo) con lo cual $p \to 0$.

Distribuciones continuas

Distribución Uniforme

La distribución uniforme en un intervalo dado [a, b], (con la notación X~ U[a, b]), queda definida por la función de densidad:

$$f(x) = \begin{cases} 1/(b-a) & a \le x \le b \\ 0 & \text{en otro caso} \end{cases}$$

Dentro del intervalo [a, b], la probabilidad de que x tome valores en cualquier intervalo es proporcional a la amplitud del mismo. Ello es así ya que la función de distribución toma la forma:

$$F_X(x) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \le x \le b \\ 1 & x > b \end{cases}$$

de modo que para cualquier $a \le x_0 \le x_1 \le b$ se cumple que:

$$P(x_0 \le X \le x_1) = F(x_1) - F(x_0) = (x_1 - x_0)/(b-a)$$

Esta distribución es el equivalente continuo del caso de los resultados discretos equiprobables. Calculando su valor esperado, obtenemos:

$$\mathrm{E}\,(\mathrm{X}) = \int_{-\infty}^{a} x \cdot 0 \, dx + \int_{a}^{b} \frac{x}{(b-a)} \, dx + \int_{b}^{+\infty} x \cdot 0 \, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

lo cual reafirma que el supuesto utilizado en estadística descriptiva al usar el punto medio como representante de cada intervalo con datos agrupados corresponde a suponer una distribución uniforme al interior de cada uno de ellos.

Distribución Normal

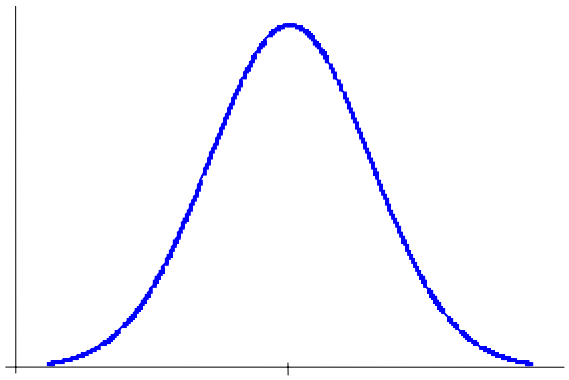
Una variable aleatoria *normal* tiene la función de densidad siguiente:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\omega}\right)^2}$$

donde μ y σ son constantes. Dichos parámetros caracterizan completamente a esta distribución, y corresponden respectivamente a la media y la desviación standard de la variable. Denotamos una variable aleatoria que sigue una distribución normal como

$X \sim N(\mu, \sigma^2)$

"X sigue una distribución normal con media μ y varianza σ^2". La función de densidad tiene la característica forma de campana y es simétrica respecto de la media μ:



La simetría implica que $f(\mu - a) = f(\mu + a)$, y que también las "probabilidades de las colas" son iguales:

$$P\ \{X < (\mu - a)\} = P\ \{X > (\mu + a)\}$$

lo que puede ponerse en términos de la función de distribución como:

$$F\ (\mu - a) = 1 - F\ (\mu + a)$$

Un caso particular de la distribución normal es la variable normal *estandarizada*, con media igual a cero y varianza igual a uno, $Z \sim N\ (0, 1)$. Dado que las expresiones matemáticas de la densidad y de la función de distribución normal son complicadas, es útil contar con las tablas usuales de las probabilidades de intervalos de una variable normal estandarizada. Dichas tablas contienen $F_Z\ (z)$ (o sea la probabilidad de que la variable aleatoria tome valores menores o iguales a dicho z), para una serie de valores reales. Sólo se incluyen los valores de z positivos pues para los valores negativos puede usarse la regla:

$$F_Z\ (-z) = 1 - F_Z\ (z).$$

La tabla puede usarse además para determinar probabilidades de que Z caiga en cualquier intervalo $[z_0, z_1]$, usando la regla:

$$P\ (z_0 \leq Z \leq z_1) = F_Z\ (z_1) - F_Z\ (z_0)$$

Si conocemos media y varianza de una variable normal $X \sim N\ (\mu, \sigma^2)$ mediante las tablas podemos calcular a su vez probabilidades de intervalos, ya que (no lo demostraremos):

$$\frac{X - \mu}{\sigma} \sim N\ (0, 1)$$

por lo que podemos usar la regla

$$P(x_0 \leq\ X \leq\ x_1)\ =\ P\left\{\frac{x_0 - \mu}{\sigma} \leq Z \leq \frac{x_1 - \mu}{\sigma}\right\} = F_Z\left(\frac{x_1 - \mu}{\sigma}\right) - F_Z\left(\frac{x_0 - \mu}{\sigma}\right)$$

Por último consideremos otro uso de las tablas de la distribución normal (0,1). Podemos etiquetar los valores del recorrido de la variable que encontramos en la tabla, que genéricamente hemos denominado z, de acuerdo a la probabilidad correspondiente a cada intervalo de la forma $(-\infty, z]$, de modo que al valor z tal que $P\ (Z \leq z\ ) = \alpha$ lo llamamos z_α. A partir de aquí podemos desarrollar un método de cálculo de las probabilidades del valor absoluto de una variable normal.

Sea $Z \sim N\ (0, 1)$. Sabemos que

$$P(\ |\ Z\ |\ \leq z\ ) =\ P(\ -z\ \leq\ Z \leq\ z)\ =\ F_Z(z) - F_Z(-z) = F_Z(z) - [\ 1 - F_Z(z)] = 2F_Z(z) -\ 1$$

Por lo tanto, si de acuerdo a lo definido tenemos que $P(Z \leq z_{[1-\alpha]}) = F_Z(z_{[1-\alpha]}) = 1-\alpha$, entonces se obtiene que:

$$P(|Z| \leq z_{[1-\alpha]}) = 2(1-\alpha) - 1 = 1-2\alpha$$

esta expresión nos está diciendo que si elegimos un valor α cualquiera, digamos 0,05, la probabilidad de que el valor absoluto de una variable aleatoria normal con media cero y varianza uno sea menor o igual que $z_{0.95}$ es igual a 1 menos 2 veces el valor elegido, es decir 0,90. Otra forma de verlo es a través de la expresión: $P(|Z| \leq z_{[1-\alpha/2]}) = 1-\alpha$.

Una propiedad interesante de las variables aleatorias normales es la siguiente:

Sean X_i , i = 1, ..., n variables aleatorias normales e independientes, $X_i \sim N(\mu_i, \sigma^2_i)$, entonces la variable

$$Y = \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma^2{}_i,\right)$$

La suma de variables aleatorias normales independientes se distribuye también normal, con media igual a la suma de las medias y varianza igual a la suma de las varianzas.

Distribución ji cuadrado

Esta familia de distribuciones está caracterizada por la siguiente forma funcional de la densidad:

$$f(x, n) = \frac{1}{2^{(n/2)}\Gamma(n/2)} x^{(n/2)} e^{-(x/2)}$$

con $x > 0$ y $n = 1, 2, \ldots$. La expresión $\Gamma(\cdot)$ se conoce como la función gamma y está definida de la siguiente manera:

$$\Gamma(n) = \int_0^{\infty} v^n e^{-v} dv$$

La notación que utilizaremos es $X \sim \chi^2(n)$. El parámetro n es conocido como los "grados de libertad" de la distribución. Obviamente el cálculo de probabilidades utilizando dicha densidad es complicado por lo que los valores de las probabilidades acumuladas se encuentran en tablas para diferentes valores de los grados de libertad.

Además, si $X \sim \chi^2(n)$, entonces $E(X) = n$ y $V(X) = 2n$.

El resultado más interesante que utilizaremos en relación con la distribución ji cuadrado está relacionado con que se trata de una familia de distribuciones asociada a la normal, corresponde a la distribución de funciones de variables aleatorias normales.

Si $X_i \sim N(0,1)$ i = 1, 2, ..., n son variables aleatorias independientes, entonces

$$\sum_{i=1}^{n} X_i^{\,2} \sim \chi^2 (n)$$

es decir, la suma de n variables aleatorias normales (0,1) e independientes elevadas al cuadrado se distribuye ji cuadrado con n grados de libertad.

Distribución t de Student

La segunda distribución asociada a la normal que estudiaremos es la *t de Student*, caracterizada por la densidad siguiente:

$$f(x, n) = \frac{1}{\sqrt{(n\pi)}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma(n/2)} \frac{1}{\left(1 + x^2/n\right)^{n+1/2}}$$

con $n > 0$ y $x \in \Re$. Está caracterizada también por el parámetro n, "grados de libertad", y usamos la notación $X \sim t(n)$. Las probabilidades para distintos valores de n se encuentran en tablas. Una variable X con esta distribución tiene $E(X) = 0$ y $V(X) = n/(n-2)$. Cuando n es grande, esta distribución está muy cerca de la normal (0,1).

El único resultado que destacaremos con respecto a esta distribución es el siguiente:

Sea $X_1 \sim N(0,1)$ y $X_2 \sim \chi^2 (n)$, dos variables aleatorias independientes, entonces la expresión

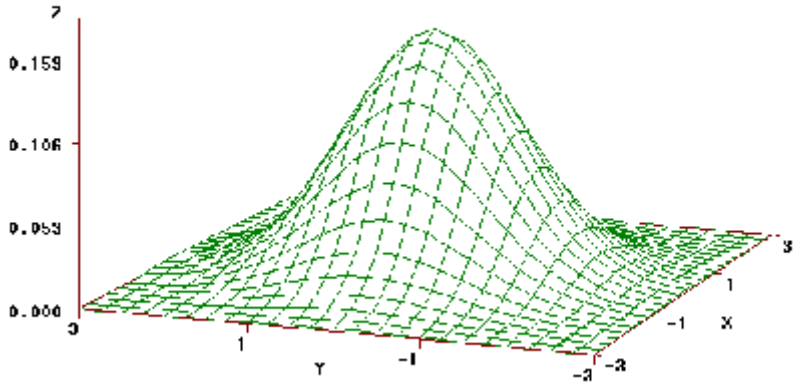
$$t = \frac{X_1}{\sqrt{X_2/n}} \sim t(n).$$

De modo que el cociente entre una normal (0,1) y la raíz de una ji-cuadrado dividida por sus grados de libertad sigue una distribución t con n grados de libertad.

Distribución Normal Bivariada

Consideremos un vector aleatorio (X, Y), que sigue una distribución conjunta normal

bivariada. No detallaremos la forma funcional de la densidad, pero se puede señalar que la distribución está completamente definida por cinco parámetros: μ_X, σ^2_X, μ_Y, σ^2_Y y ρ_{XY}, que corresponden a media y varianza de X y de Y, respectivamente, y al coeficiente de correlación entre ambas. La densidad conjunta bivariada puede representarse gráficamente como una superficie en tres dimensiones, en la que a cada punto de $\Re^2$ le corresponde una altura, de modo que el aspecto del gráfico en tres dimensiones es el de una elevación simétrica, con la cúspide en el punto de las medias (μ_X, μ_Y). En el siguiente gráfico se representa la superficie de una densidad normal bivariada.



Sucesivos cortes horizontales de la superficie dan lugar a elipses, cuya forma depende de las varianzas σ^2_X y σ^2_Y. A su vez, el grado de correlación entre las variables ρ_{XY} , determina el grado de inclinación de la superficie; variables incorrelacionadas dan lugar a elipses paralelos a los ejes.

Las densidades marginales de X y de Y pueden visualizarse proyectando la densidad conjunta sobre los ejes. Son curvas normales, y las variables X e Y tendrán distribuciones marginales normales : $X \sim N(\mu_X, \sigma^2_X)$ e $Y \sim N(\mu_Y, \sigma^2_Y)$.

Por su parte, las densidades condicionales (digamos para Y), $f_{Y/X}$ son los perfiles de cortes verticales en la superficie en el sentido paralelo al eje de las Y, en cada punto x. Tambien se trata de curvas normales, pero sus parámetros son diferentes:

$$E(Y/X = x) = \mu_Y + \rho_{XY} \cdot \sigma^2_Y/\sigma^2_X \cdot (x - \mu_X)$$

y

$$V(Y/X = x) = \sigma^2_Y - \rho_{XY} \cdot \sigma^2_Y/\sigma^2_X$$

de modo que la media condicional de Y dado x depende linealmente de x, a menos que ambas variables estén incorrelacionadas. La varianza condicional es en cambio una constante,

aunque diferente de σ^2_Y, a menos que las variables estén incorrelacionadas.

7. Inferencia estadística

Recapitulando sobre los temas que se han abordado hasta aquí, recordamos que el conjunto de técnicas descriptivas estudiadas en primer término no nos permitían ir más allá del resumen y la descripción del conjunto de datos que estábamos considerando. Dando un paso más adelante, nos dedicamos a continuación a estudiar los fundamentos de un modelo matemático para el proceso de generación de los datos (PGD), hasta llegar a la formulación de modelos de probabilidad (Φ). Este paso consistió en especificar un conjunto de familias paramétricas de densidades.

La diferencia entre el estudio descriptivo de los datos y la inferencia estadística está dada porque en la inferencia estadística se propone a priori un modelo de probabilidad como una descripcion generalizada del proceso que da origen a los datos observados. En el marco de la inferencia estadística, los datos específicamente observados son vistos como una de las muchas posibles realizaciones del PGD. El modelo de probabilidad describe o bien el proceso que da origen a los datos observados, o bien la población de la que los datos observados provienen. De este modo, a diferencia del caso de la estadística descriptiva, que no permite afirmar nada sobre lo que ocurre fuera del conjunto de datos observados, la inferencia estadística permite realizar afirmaciones de tipo probabilístico sobre el PGD o sobre los elementos de la población no observados.

La conexión entre el modelo de probabilidad (Φ) y los datos observados debe establecerse a través de un *modelo muestral*, que es el segundo ingrediente que define a un modelo estadístico.

Modelo Muestral

El modelo muestral describe la relación entre el modelo de probabilidad (Φ) y los datos observados, describiendo la manera en que estos pueden ser vistos en relación a Φ. El concepto fundamental en el modelo muestral es el de *muestra*.

Def. Una *muestra* se define como un conjunto de variables aleatorias $(X_1, X_2, \ldots, X_n)$ cuyas funciones de densidad coinciden con la función de densidad $f(x; \theta)$ postulada por el modelo de probabilidad.

La significación del concepto está dada por el hecho de que en este contexto, los datos observados son considerados una de las muchas posibles realizaciones de la muestra. Esto nos aleja del significado que se da a "la muestra" en el lenguaje de todos los días, en el sentido de designar a cualquier conjunto de datos observados.

Dado que la muestra es un conjunto de variables aleatorias relacionadas con Φ, tienen a su vez una distribución que llamamos *distribución de la muestra*.

Def. La *distribución de la muestra* $X = (X_1, X_2, ..., X_n)$ se define como la distribución conjunta de las variables $X_1, X_2, ..., X_n$, y se denota por

$f(x_1, x_2, ... , x_n ; \theta)$.

La forma de $f(x_1, x_2, ... , x_n ; \theta)$ es crucial en la determinación del modelo muestral. La más simple y más utilizada de las formas que toma está basada en la idea de un experimento aleatorio, y se denomina *muestra aleatoria*.

Un conjunto de variables aleatorias $(X_1, X_2, ..., X_n)$ se denomina *muestra aleatoria* de $f(x ;\theta)$ si las variables aleatorias $X_1, X_2, ..., X_n$ son independientes e identicamente distribuidas (IID). En este caso la distribución de la muestra toma la siguiente forma:

$$f(x_1, x_2, ... , x_n ; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

La igualdad surge de la independencia de las variables aleatorias. Aunque no las estudiaremos aquí, diferentes definiciones de la muestra (muestras no independientes o no idénticamente distribuidas) darán lugar a distintos modelos muestrales.

En general utilizaremos letras mayúsculas para la muestra: $\mathbf{X} = (X_1, X_2, ..., X_n)'$ y letras minúsculas para su realización $\mathbf{x} = (x_1, x_2, ..., x_n)'$. Se supone que $\mathbf{x} = (x_1, x_2, ..., x_n)'$ toma valores en el espacio de las observaciones $\mathcal{X}$ por lo que $\mathbf{x} \in \mathcal{X}$. Usualmente el espacio de las observaciones es $\Re^n$.

Modelo estadístico

En el contexto de la inferencia estadística necesitamos definir ambos modelos, el modelo de probabilidad y el modelo muestral. La unión de ambos define un *modelo estadístico*.

Def. Un *modelo estadístico* se define como la unión de:

(i) un modelo de probabilidad $\Phi = \{f(x ;\theta), \theta \in \Theta\}$
(ii) un modelo muestral $\mathbf{X} = (X_1, X_2, ..., X_n)'$

El concepto de modelo estadístico está en la base de la inferencia paramétrica. Se debe notar sin embargo que hay una rama de la inferencia estadística que es la no paramétrica, y que se caracteriza justamente porque no hay un Φ que se asuma a priori, pero su estudio queda más allá del alcance de este curso introductorio.

Esquema de la inferencia

Partiendo del modelo estadístico definido anteriormente, esquemáticamente definiremos el conjunto de problemas que aborda la inferencia estadística:

1) ¿Son los datos observados consistentes con el modelo estadístico postulado? Este problema se conoce como el problema de la especificación.

2) Suponiendo que el modelo estadístico postulado es consistente con lo datos, ¿qué se puede inferir acerca de los parámetros desconocidos $\theta \in \Theta$?

> a) ¿Es posible elegir dentro de Θ un valor $\hat{\theta}$ como el más representativo para θ? (estimación puntual).
> b)¿Es posible reducir nuestra incertidumbre sobre $\theta \in \Theta$, reduciendo el espacio paramétrico Θ a Θ_0 (un subconjunto de Θ)? (estimación por intervalos de confianza).
>
> c) ¿Es posible considerar la cuestión de si $\theta \in \Theta_0 \subset \Theta$, rechazando o no dicha afirmación en vista de los datos observados? (realización de pruebas de hipótesis).

3) Suponiendo que se ha elegido un valor $\hat{\theta}$ como más representativo para θ, ¿es posible inferir acerca de observaciones adicionales del proceso de generación de datos que describe nuestro modelo estadístico? (posibilidad de la predicción fuera de lo observado).

Estimadores, estadísticos y distribuciones en el muestreo

En el esquema planteado nos referimos al intento de dar un valor numérico al parámetro θ, lo que implica definir la manera en que procesamos la información obtenida de la realización de la muestra para seleccionar dentro del espacio paramétrico un valor en particular. Para ello se construye una función $h(\cdot)$: $\mathcal{X} \rightarrow \Theta$. A dicha función h ($\mathbf{X}$), que asocia a cada muestra un elemento del espacio paramétrico se le llama *estimador* de θ, en tanto que a su valor $h(\mathbf{x})$ se le llama una *estimación* de θ.

En forma general, puede verse que los problemas planteados en inferencia requieren de la construcción de funciones del tipo $q(\cdot)$: $\mathcal{X} \rightarrow \Theta$. que deberán satisfacer distintos criterios de acuerdo a la naturaleza del problema. Dichas funciones recibirán el nombre especial de *estadísticos* (a veces con el adjetivo *muestrales*), y las definiremos de la siguiente manera:

Def. Un *estadístico* es una función de variables aleatorias observables, que es a su vez una variable aleatoria, y que no contiene ningún parámetro desconocido.

Se desprende de lo anterior que un estimador es un estadístico. Por definición los estadísticos, como variables aleatorias que son, tienen sus propias distribuciones. Como variables aleatorias, la discusión relativa a las propiedades y la naturaleza de los estadísticos debe realizarse en términos de sus distribuciones. La inferencia estadística depende crucialmente de

la posibilidad de determinar la distribución de un estadístico dado a partir de la distribución de la muestra. Las distribuciones de los estadísticos reciben el nombre de *distribuciones en el muestreo*.

Las distribuciones en el muestreo deben derivarse de las distribuciones de las muestras que subyacen en ellos, lo cual subraya la importancia de estudiar la distribución de funciones de variables aleatorias. Para ejemplificar, supongamos que el modelo probabilístico que proponemos para estudiar un fenómeno de interés está caracterizado por una variable aleatoria X cuya densidad depende de un parámetro : que representa la media de la densidad propuesta. Si le unimos un modelo muestral basado en una muestra aleatoria, tendremos un conjunto de n variables aleatorias, la muestra, con una distribución dada. Por último, proponemos un estadístico, la media muestral, que consiste en una función de dicha muestra. Como función de variables aleatorias, considerada ex ante, antes de la extracción concreta de una muestra en particular, dicha media muestral es una variable aleatoria, con su distribución en el muestreo asociada.

En un proceso de investigación concreto, usualmente se extraerá una sola muestra, y se realizará el cómputo del valor obtenido de la media muestral. Dicha media muestral obtenida puede considerarse entonces como una realización particular de la distribución en el muestreo de dicho estadístico.

Como ejemplo de distribuciones en el muestreo, nos centraremos en los dos estadísticos mencionados anteriormente, la media y la varianza muestral. la discusión se realizará en términos de una muestra aleatoria, por lo que las observaciones serán independientes e idénticamente distribuidas. Los estadísticos mencionados se definen de la siguiente manera:

$$\overline{X}_n = \sum_{i=1}^{n} \frac{X_i}{n} \text{ (media muestral)}$$

$$S^2 = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}_n\right)^2}{(n-1)} \text{ (varianza muestral[4])}$$

Supongamos que el proceso generador de los datos está representado por una variable X con media μ y desviación standard de σ. En este contexto consideraremos a $\overline{X}_n$ y S^2 como estimadores respectivamente de μ y σ^2.

Si $\overline{X}_n$ es la media muestral de una muestra aleatoria de tamaño n, su propio valor esperado puede obtenerse como:

$$E(\overline{X}_n) = E[1/n(X_1 + X_2 + ... + X_n)] = 1/n[E(X_1) + E(X_2) + ... + E(X_n)] = 1/n \cdot [n\mu] = \mu$$

[4] Notemos que en el caso de la varianza muestral el divisor (n−1) difiere del de la fórmula planteada cuando se estudió la varianza en el contexto de la estadística descriptiva (n).

De modo que obtenemos que el valor esperado de la media muestral es el mismo que el valor esperado de la variable del modelo probabilístico que describe la generación de los datos. Si derivamos la varianza de la media muestral obtenemos

$$\text{Var}(\overline{X}_n) = \text{V}[1/n(X_1 + X_2 + ... + X_n)] = (1/n^2)\,\text{V}[X_1 + X_2 + ... + X_n] = n\sigma^2/n^2 = \sigma^2/n$$

Se ha utilizado el hecho de que la muestra es independiente. Esto indica que a medida que el tamaño muestral se incrementa la dispersión de la media muestral se reduce, lo que puede verse como surgiendo del hecho de que una muestra grande contiene mayor información acerca de la media de la variable original que una más pequeña.

Si la población muestreada está razonablemente descrita por una distribución normal con media μ y varianza σ^2, las X_i a su vez estarán normalmente distribuidas y también lo estará la media muestral (esto surge de los resultados anteriormente enunciados sobre la suma de variables normales independientes). Las observaciones X_i , i = 1,..., n son variables aleatorias normales e independientes, $X_i \sim N(\mu, \sigma^2)$, de modo que la variable

$$\overline{X}_n = \sum_{i=1}^{n} \frac{X_i}{n} \sim N\left(\frac{1}{n}\sum_{i=1}^{n}\mu, \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2\right)$$

es decir, en el caso de muestreo de poblaciones normales, $\overline{X}_n \sim N(\mu, \sigma^2/n)$

Consideremos ahora a la varianza muestral. En primer lugar investigamos su valor esperado, que podemos plantear como:

$$\text{E}(S^2) = \text{E}\left(\frac{1}{(n-1)}\sum_{i=1}^{N}(X_i - \overline{X}_n)^2\right)$$

Si consideramos la sumatoria en esta expresión obtenemos:

$$\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n}\left[(X_i - \mu) - (\overline{X}_n - \mu)\right]^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\overline{X}_n - \mu) + \sum_{i=1}^{n}(\overline{X}_n - \mu)^2$$

El segundo de los términos podemos escribirlo como

$$2\sum_{i=1}^{n}(X_i - \mu)(\overline{X}_n - \mu) = 2(\overline{X}_n - \mu)\sum_{i=1}^{n}(X_i - \mu) = 2n(\overline{X}_n - \mu)$$

con lo que obtenemos que

$$\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X}_n - \mu)^2$$

Y al tomar el valor esperado se obtiene

$$E(S^2) = \frac{1}{(n-1)} E\left[\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X}_n - \mu) \right]$$

El primero de los términos en el paréntesis es $n\sigma^2$, mientras que el último es n veces la varianza de $\overline{X}_n$, que conocemos es igual a σ^2/n. Resumiendo:

$$E(S^2) = \frac{1}{(n-1)} \left(n\sigma^2 - \frac{n\sigma^2}{n} \right) = \frac{(n-1)}{(n-1)} \sigma^2 = \sigma^2$$

La media de la distribución de la varianza muestral es la varianza de la variable aleatoria del modelo [5].

Por último consideraremos el caso en que la variable que muestreamos se distribuye ella misma Normal (μ ,σ^2). Las observaciones X_i , i = 1, ..., n son variables aleatorias normales e independientes y tendremos que la expresión

$$\left(\frac{1}{\sigma^2} \sum_{i=1}^{N} (X_i - \overline{X}_n)^2 \right) = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Recordemos el resultado que establecía que una suma de n variables aleatorias normales independientes al cuadrado seguía una distribución ji-cuadrado con n grados de libertad. La reducción en 1 grado de libertad en la expresión anterior tiene que ver con que se ha utilizado $\overline{X}_n$ en lugar de μ. Ello reduce el número de sumandos independientes a (n − 1), ya que el n-ésimo término de la suma, ($X_i - \overline{X}_n$), está exactamente determinado por los (n − 1) restantes.

Teoremas límite

Bajo el título de teoremas límite se agrupa un conjunto de resultados de la teoría probabilística conocidos con ese nombre porque implican considerar en el límite la distribución de variables aleatorias cuando se toma un número de observaciones arbitrariamente grande. Puede separarse en dos grandes capítulos, bajo los nombres genéricos de "la" Ley de los Grandes Números, por una parte y "el" Teorema Central del Límite, por la otra. Las comillas obedecen a que cada uno de dichos capítulos contiene un amplio número de versiones de dichos resultados, variando las condiciones y supuestos que se requieren para obtener cada una de ellas. Nos limitaremos a enunciar una versión de cada uno de dichos teoremas, que juegan un papel muy importante en la inferencia estadística, ya que nos permiten aproximarnos a la

[5] Esto proporciona un argumento formal para la división por (n − 1).

distribución en el muestreo de estadísticos relevantes.

Ley débil de los Grandes Números

Sea f(·) una densidad con media μ y varianza finita σ^2, y sea $\overline{X}_n$ la media muestral de una muestra aleatoria de tamaño n (recordemos que el requerimiento de una muestra aleatoria implica que las variables X_i son independientes e idénticamente distribuidas). Sean a su vez ε y δ dos números reales tales que $\varepsilon > 0$ y $0 < \delta < 1$. Si n es un entero mayor que $\sigma^2/\varepsilon^2\delta$ se cumple que:

$$P\left[\left| \overline{X}_n - \mu \right| < \varepsilon \right] \geq 1 - \delta$$

Dicho teorema se prueba utilizando la desigualdad de Tschebyscheff.

Supongamos que en el modelo de probabilidad propuesto : denota E(X). Una pregunta crucial es si, utilizando un número finito de observaciones de X, podemos realizar inferencias confiables acerca de μ. En palabras, la ley (débil) de los grandes números, establece que puede establecerse un número n (tamaño muestral) tal que si se toma una muestra aleatoria de ese tamaño o mayor, la probabilidad de que la media muestral $\overline{X}_n$ difiera de μ en una cantidad arbitrariamente pequeña puede hacerse tan cercana a 1 como se quiera. Notemos sin embargo que se requiere contar con conocimiento sobre σ^2, lo cual no necesariamente ocurre.

Teorema Central del límite

Consideremos una sucesión de variables aleatorias $X_1, X_2, ..., X_n$, que supondremos tienen la misma media μ y una varianza finita σ^2. Definamos S_n como la suma de estas variables:

$$S_n = X_1 + X_2 + ... + X_n$$

Dado que son independientes e incorrelacionadas por pares, tenemos los resultados:

$$E(S_n) = E(X_1 + X_2 + ... + X_n) = n\mu$$

$$Var(S_n) = Var(X_1 + X_2 + ... + X_n) = n\sigma^2$$

De esto se sigue que la expresión

$$\frac{(S_n - n\mu)}{\sigma\sqrt{n}}$$

tiene media cero y varianza 1.

TCL: Sea una sucesión de variables aleatorias $X_1, X_2, ..., X_n$ con $E(X_i) = \mu$ y $V(X_i) = \sigma^2$, y

sea $S_n = X_1 + X_2 + ... + X_n$. Cuando $n \to +\infty$, la distribución de la expresión $\frac{(S_n - n\mu)}{\sigma\sqrt{n}}$ tiende a la Normal (0,1), cualquiera sea la distribución original de las variables X.

Esto reafirma la importancia de la distribución normal en estadística. Además de que se ha encontrado que proporciona una aproximación cercana a la distribución de muchas poblaciones en el mundo real, aún cuando las poblaciones de interés no puedan ser razonablemente descritas por la distribución normal, ésta puede utilizarse para determinar probabilidades aproximadas para estimadores como la media muestral.

Una aplicación inmediata es la aproximación de la distribución de la media muestral estandarizada. Sea f (·) una densidad con media μ y varianza σ^2 finita, y $\overline{X}_n$ la media muestral de una muestra aleatoria de tamaño n de f (·). Puede definirse entonces la variable aleatoria Z_n como

$$Z_n = \frac{X_n - E(X_n)}{\sqrt{Var(X_n)}} = \frac{X_n - \mu}{\sigma / \sqrt{n}}$$

Por el TCL, la distribución de Z_n, cuando n tiende a infinito, tiende a la Normal (0,1). Obviamente nunca tendremos una muestra infinita, pero sabemos que en una muestra grande (n > 100, por ejemplo) la normal constituye una buena aproximación a la distribución de la media estandarizada, y aún en una muestra de tamaño moderado, con n = 20, la aproximación es razonable.

Teorema de De Moivre

El teorema de De Moivre es un ejemplo del teorema central del límite, y uno de los primeros resultados de este grupo obtenidos en la historia de la estadística. Consiste en la aproximación de una variable binomial por la normal. Supongamos que X_n es el número de éxitos en n pruebas de Bernoulli independientes. Dado que es la suma de variables Bernoulli independientes, es un caso de la S_n definida en las condiciones del TCL. Su media es np y su varianza np(1−p), donde p es la probabilidad de éxito en una de las pruebas. Entonces la distribución de la expresión

$$\frac{X_n - np}{\sqrt{np(1-p)}}$$

cuando n tiende a infinito, puede ser aproximada por la N(0,1). La aproximación puede ser usada cuando n es un número al menos de 20, y se vuelve muy cercana cuando es mayor que 50. De este modo la probabilidad:

$$P\ (a \leq X_n \leq b)$$

puede ser aproximada por el valor:

$$\Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right)$$

donde $\Phi(\cdot)$ indican los valores de tablas de las probabilidades acumuladas hasta ese punto por la función de distribución de una Normal (0,1).

Propiedades de los estimadores

Una propiedad clave de los estimadores que hemos visto, la media y la varianza muestral, es que en cada uno de los casos la media de la distribución del estimador coincide con el parámetro que con ellos se busca estimar. Los estimadores que tienen esta propiedad se denominan *insesgados*. A su vez, si consideramos un parámetroy su estimador $\hat{\theta}$ podemos definir el *sesgo*, $B(\hat{\theta})$ como la diferencia entre el valor esperado del estimador y el parámetro estimado:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

El sesgo de un estimador insesgado es obviamente cero. La propiedad de insesgamiento para un estimador dado está en relación a un parámetro determinado, y por eso se dice que el estimador $\hat{\theta}$ es insesgado para θ. Así, al hallar el valor esperado de la media muestral demostramos que ésta es insesgada para la media de la población. El estimador tendrá además una determinada varianza, que surge de su distribución en el muestreo, pero la propiedad de insesgamiento es importante pues nos indica que su distribución estará centrada en el valor del parámetro que buscamos. Hay una cierta similitud con tirar al blanco: los tiros tendrán una cierta dispersión alrededor del centro del blanco, pero idealmente se concentrarán en torno a éste. Un estimador sesgado sería como un arma que por algún motivo sistemáticamente apuntara fuera del blanco deseado.

Cuando comparamos dos estimadores insesgados, la elección del mejor estimador dependerá de la comparación de las varianzas.

Cuando se comparan estimadores sesgados, o insesgados contra sesgados, existe la posibilidad de plantearse un compromiso entre el insesgamiento y la varianza del estimador, aceptando un leve sesgo cuando se tiene una reducción significativa en la varianza. Un concepto de utilidad aquí es el de *error cuadrático medio*, que se define como el valor esperado del cuadrado de la diferencia entre el estimador y el parámetro:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

que a su vez puede descomponerse como:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E[\hat{\theta} - E(\theta) + E(\theta) - \theta)]^2 =$$

$$E[\hat{\theta} - E(\theta)]^2 - 2E[\hat{\theta} - E(\theta)][E(\hat{\theta}) - \theta)]^2 + E[E(\theta) - \theta]^2$$

y como

$$E[\hat{\theta} - E(\theta)] = 0$$

obtenemos

$$ECM(\hat{\theta}) = E[\hat{\theta} - E(\theta)]^2 + E[E(\theta) - \theta]^2 = Var(\hat{\theta}) + [B(\hat{\theta})]^2$$

El error cuadrático medio de un estimador es igual a su varianza más su sesgo al cuadrado. En esta situación un criterio para elegir entre varios estimadores podría ser entonces la minimización del ECM.

La segunda propiedad de los estimadores que consideraremos es la *consistencia*. Como se ha expuesto anteriormente, la distribución en el muestreo de los estimadores depende del tamaño muestral. Se ha observado por ejemplo que la varianza de la media muestral (dada por la expresión σ^2/n) disminuye cuando se consideran sucesivamente muestras de mayor tamaño. Esta es la razón de considerar el comportamiento de secuencias de estimadores cuando el tamaño de la muestra crece. En particular se analiza el caso límite cuando $n \to \infty$, obviamente no porque se espere alguna vez observar una muestra de tamaño infinito, sino porque los resultados alcanzados son suficientemente fuertes como para ser muy valiosos en el análisis de muestras grandes.

La definición formal de consistencia es como sigue:

Definición: Dados una variable aleatoria X_n, (que depende de n) y una constante k, si se cumple que

$$P(|X_n - k| > \varepsilon) \to 0 \text{ cuando } n \to \infty \text{ para cualquier } \varepsilon > 0,$$

entonces X_n *converge en probabilidad* a k.

Cuando tenemos un estimador $\hat{\theta}$ para un parámetro θ, si $\hat{\theta}$ converge en probabilidad a θ, entonces se dice que $\hat{\theta}$ es un *estimador consistente* para θ. Si bien estudiar la convergencia en probabilidad involucra alguna complejidad técnica, existe una condición suficiente que ayuda a ver las implicancias de la definición.

Condición suficiente: un estimador cuyo ECM tiende a 0 cuando $n \to \infty$ es consistente.

Ello implica, de acuerdo a la definición que dimos anteriormente, dos condiciones que deben cumplirse cuando $n \to \infty$: la varianza debe tender a 0 por una parte, y por otra el sesgo (si existe) también debe tender a 0. Ello acerca al estimador a una constante (que sería el caso de una variable aleatoria de varianza cero) cuando la muestra crece, e implica que el estimador

coincide en el límite con el parámetro estimado [6].

El ejemplo que tenemos es la media muestral. Es como hemos visto insesgado, de manera que el ECM de la media muestral se reduce a la varianza. Nuestra condición suficiente indica que ésta debe tender a cero cuando $n \to \infty$, lo cual comprobamos que efectivamente ocurre cuando analizamos el comportamiento de la expresión σ^2 / n. De donde concluimos que la media muestral es un estimador consistente para la media μ.

Inferencia. Ejemplos

No se abordará cada uno de los problemas asociados a la inferencia estadística. A través de ejemplos, trataremos algunos de ellos. En particular se tratará la estimación por intervalos y la realización de pruebas de hipótesis.

1. Intervalos de confianza

Un intervalo de confianza para un parámetro θ es una expresión que asigna una cierta probabilidad al evento definido por que cierto intervalo en la recta real contenga a dicho parámetro. Los límites de dicho intervalo dependen de estadísticos, y por tanto son aleatorios. La afirmación probabilística se deriva directamente de la distribución en el muestreo de dichos estadísticos.

Supongamos por ejemplo que una variable aleatoria $X \sim N(\mu, \sigma^2)$. En el caso de muestreo aleatorio, se puede deducir que $\overline{X}_n \sim N(\mu, \sigma^2/n)$, de modo que tendremos que

$$\frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

La distribución de una variable Normal (0,1) está tabulada, de modo que sabemos que para cada valor α, con $0 < \alpha < 1$, tendremos las probabilidades de la expresión:

$$P(| Z | \leq z_{[1-\alpha/2]}) = 1 - \alpha$$

en la que $z_{[1-\alpha/2]}$ es el valor del recorrido de la variable normal (0,1) en el que la función de distribución correspondiente alcanza el valor $1 - \alpha/2$.

Desarrollando la expresión anterior obtenemos:

$$P\left(- z_{[1-\alpha/2]} \leq \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \leq z_{[1-\alpha/2]} \right) = 1 - \alpha$$

[6] Que esta sea una condición suficiente implica que puede no cumplirse en algún caso y el estimador seguir siendo consistente.

y manipulando un poco se obtiene la siguiente expresión:

$$\mathrm{P}\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} z_{[1-\alpha/2]} \leq \mu \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} z_{[1-\alpha/2]} \right) = 1 - \alpha$$

que denominamos "intervalo de confianza al 100(1 – α) % para μ". El valor α es clave en este contexto, y expresa la medida de la confianza que estamos dispuestos a exigir de nuestra estimación.

Hemos partido del supuesto de una distribución normal para la variable que analizamos. La construcción del intervalo de confianza que analizamos incluye un problema adicional, que es la necesidad de conocer el parámetro σ, lo que no tiene porqué ser el caso. En la práctica, σ debe ser estimado, lo cual se realiza a través de la desviación standard muestral *s*.

Ello nos lleva a considerar la distribución de la variable aleatoria siguiente:

$$\mathrm{t} = \frac{\overline{X}_n - \mu}{s/\sqrt{n}}$$

La expresión puede escribirse como:

$$\mathrm{t} = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \frac{1}{\sqrt{\frac{s^2(n-1)}{\sigma^2(n-1)}}}$$

Por una parte habíamos establecido que:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}\ (0,\ 1)$$

y por otra a su vez

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{\ (\mathrm{n}-1)}$$

de modo que la variable buscada puede expresarse como el cociente de una variable normal standardizada, dividida por la raíz cuadrada de una variable ji-cuadrado que a su vez está dividida por sus grados de libertad. La condición adicional que se requiere es que sean independientes, lo cual se cumple en el caso de muestreo de una variable normal, para que la distribución de esta expresión siga una distribución t de Student, con n – 1 grados de libertad. Al igual que la normal, la t de Student es una distribución simétrica. Los valores de la distribución t están también tabulados para diferentes valores de los grados de libertad.

Llamemos $t^{(n-1)}_{[1-\alpha/2]}$ al valor del recorrido de la variable t (con (n –1) grados de libertad) en

el que la función de distribución correspondiente alcanza el valor $1-\alpha/2$. El intervalo de confianza cambia entonces de forma:

$$\mathrm{P}\left(X_n - \frac{s}{\sqrt{n}} t^{(n-1)}_{[1-\alpha/2]} \leq \mu \leq \frac{s}{\sqrt{n}} t^{(n-1)}_{[1-\alpha/2]} \right) = 1 - \alpha$$

Además, debe señalarse que para tamaños de n grandes, la distribución t de Student se acerca a la normal.

3. Prueba de hipótesis

Por último se discute informalmente la realización de pruebas de hipótesis referidas a la distribución de variables aleatorias de interés. El enfoque que utilizaremos es el de considerarlas casos de decisión bajo incertidumbre. Se tiene una afirmación genérica acerca de la distribución de una variable aleatoria, y se utilizarán los datos muestrales para rechazar dicha hipótesis como incompatible con la evidencia muestral, o bien no rechazarla.

En general tendremos una afirmación o hipótesis sobre la distribución de una variable aleatoria, que será la que someteremos a prueba, a la que llamamos *hipótesis nula*, y para la que usaremos la notación H_0. Tendremos además una segunda afirmación o hipótesis contra la que contrastaremos la hipótesis nula, que llamaremos *hipótesis alternativa,* y la denotamos por H_1.

Muchas veces las hipótesis serán afirmaciones sobre cierto parámetro de una distribución. Por ejemplo, "la media de los ingresos de los hogares de Montevideo en pesos es igual a 4764", lo que podemos escribir:

$$H_0: \mu = 4764$$

Existen distintas formas de especificar la alternativa, por ejemplo

$$H_1: \mu \neq 4764$$

Se puede observar que en este caso la hipótesis nula especifica un único punto para un parámetro, en cuyo caso H_0 es una *hipótesis simple*. Al contrario, la hipótesis alternativa es una *hipótesis compuesta*, ya que especifica todo un rango de valores posibles para el parámetro.

La prueba de hipótesis se define mediante la división del conjunto de las muestras posibles en dos categorías: por una parte un conjunto que, si la muestra extraída pertenece a dicho conjunto, conducirá al rechazo de la hipótesis nula, y por otra un conjunto tal que si la muestra pertenece a dicho conjunto, no la rechazaremos. La regla decisión consiste en rechazar la hipótesis nula si la muestra cae en la región de rechazo.

La región de rechazo se define en función tanto de la estimación puntual como de los intervalos de confianza para un parámetro. Intuitivamente podemos estar de acuerdo en rechazar la hipótesis nula sobre el valor de la media si la estimación puntual cae "demasiado lejos" del valor establecido en la hipótesis, y que la medida en que podremos tolerar este alejamiento esté dada por los límites del intervalo de confianza para dicho parámetro, con un nivel de confianza α que nos propongamos.

Las hipótesis nula y alternativa no tienen un papel simétrico. Esto surge de la naturaleza misma del problema de decisión bajo incertidumbre, en el cual pueden producirse cuatro situaciones posibles: dado que H_0 es cierta puedo rechazarla o no rechazarla, y dado que H_0 es falsa puedo rechazarla o no rechazarla. La situación genera dos tipos de error posible: rechazar H_0 cuando es cierta, y no rechazar H_0 cuando es falsa. Se acostumbra comparar esta situación con los errores que puede cometer un juez: absolver a un culpable o condenar a un inocente. En este sentido debe elegirse la formulación de H_0 de modo de controlar el error consistente en rechazar H_0 cuando es cierta. A este error lo llamamos *error de tipo I*.

Se trata de fijar la probabilidad de cometer dicho error de tipo I. Dado que nos interesa la "probabilidad de rechazar H_0 cuando ésta es cierta", debemos determinar la distribución de la variable aleatoria que nos interesa bajo el supuesto de que la hipótesis nula es cierta.. El criterio es determinar la región de rechazo de modo que rechazar la hipótesis nula cuando ésta es cierta sea igual a un cierto nivel α predeterminado, que convencionalmente se fija en un 1% o 5%.

Hay un cuidado especial en no hablar de "aceptar" la hipótesis nula cuando ésta no es rechazada. Al realizar la prueba sólo establecemos que la evidencia muestral es insuficiente para rechazar la hipótesis cuando se ha establecido la probabilidad de erróneamente rechazar la hipótesis cuando es cierta en α, no establecemos que la hipótesis nula es cierta. La hipótesis puede aún ser falsa, y nuestra muestra obtenida no pertenecer a la región de rechazo. El criterio que hemos establecido permite afirmar que solamente $100\alpha\%$ de las veces rechazaremos una hipótesis verdadera. En este sentido nuestra afirmación es que, hasta donde conocemos, la evidencia disponible no es suficiente para rechazarla con dicho nivel de confianza. En el segundo caso, cuando rechazamos la hipótesis nula, estaremos diciendo que la evidencia muestral se desvía lo suficiente de la hipótesis nula para rechazarla, de modo que dicha desviación es "estadísticamente significativa". De allí que al nivel $\forall$, es conocido como el *nivel de significación* de la prueba..

Si tomamos un ejemplo similar al del intervalo de confianza, teníamos una variable aleatoria $X \sim N(\mu, \sigma^2)$, y nos interesa una afirmación respecto a un parámetro de dicha distribución, la media :. Establecemos entonces

$H_0 : \mu = \mu_0$

También nos interesa definir la alternativa. Aunque sea poco realista, imaginemos que nos interesa específicamente el valor μ_0, de modo que nuestro interés está en ver hasta qué punto la evidencia muestral se desvía de este valor, tanto en el sentido de provenir de una distribución con un valor mucho más alto como en el de uno mucho más bajo que μ_0. Esto nos

conduce a una formulación de la alternativa como

$H_1 : \mu \neq \mu_0$

Para realizar la prueba requerimos de la distribución en el muestreo de un estimador para ese parámetro. En nuestro caso sabemos que $\overline{X}_n \sim N(\mu, \sigma^2/n)$, es decir, dicha distribución depende del valor del parámetro μ. Si en dicha expresión sustituimos el valor μ por el valor implicado por la hipótesis nula, obtenemos la distribución de la media muestral *bajo la hipótesis nula*. Habíamos obtenido, para una muestra aleatoria de una variable $X \sim N(\mu, \sigma^2)$ que la expresión:

$$t = \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

por lo que, de ser cierta la hipótesis nula, también se cumple que:

$$t = \frac{\overline{X}_n - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$$

Llamamos a esta expresión el *estadístico de la prueba*.

Por lo tanto, podemos afirmar que

$$P\left(-t^{(n-1)}_{[1-\alpha/2]} \leq \frac{\overline{X}_n - \mu}{s/\sqrt{n}} \leq t^{(n-1)}_{[1-\alpha/2]}\right) = 1 - \alpha$$

Al extraerse una muestra, se pueden calcular la media y desviación standard muestrales, y calcular el valor que alcanza el estadístico de la prueba. El procedimiento de la prueba será calcular dicho estadístico, estableciendo la región de rechazo de manera que rechacemos H_0 cuando éste caiga fuera de los límites del intervalo

$$[-t^{(n-1)}_{[1-\alpha/2]}, t^{(n-1)}_{[1-\alpha/2]}].$$

Dicha región recibe el nombre de *región crítica* de la prueba. Se puede notar el paralelo con la determinación del intervalo de confianza para :. El intervalo de confianza

$$P\left(\overline{X}_n - \frac{s}{\sqrt{n}} t^{(n-1)}_{[1-\alpha/2]} \leq \mu \leq \frac{s}{\sqrt{n}} t^{(n-1)}_{[1-\alpha/2]}\right) = 1 - \alpha$$

puede escribirse como

$$\mathrm{P}\left(\mu_0 - \frac{s}{\sqrt{n}} t^{(n-1)}{}_{[1-\alpha/2]} \leq \bar{X}_n \leq \mu_0 + \frac{s}{\sqrt{n}} t^{(n-1)}{}_{[1-\alpha/2]}\right) = 1 - \alpha$$

y puede observarse que el procedimiento de la prueba resulta equivalente a rechazar la hipótesis nula si el valor calculado de la media muestral cae fuera de dicho intervalo de confianza.

Interpretación de la inferencia estadística

Al inicio del curso se mencionó que el enfoque axiomático de la probabilidad permitía independizarse de la interpretación que se diera a la probabilidad, (frecuencista, clásica o subjetiva) al proporcionar un marco en que todas ellas podrían razonablemente desarrollar el análisis.

Sin embargo, al ingresar al terreno de la inferencia, nuevamente comienza a tener importancia el concepto que se tenga de la probabilidad. Ello es así ya que el concepto de muestra implica que los datos observados son tan sólo una de las muchas posibles realizaciones de un experimento aleatorio. De alguna manera en ello está implícita la posibilidad de la repetición, eventualmente al infinito, de un experimento, que nos permitiría "a la larga" o "en el largo plazo" la reconstrucción del modelo probabilístico Φ. En este enfoque la interpretación que subyace sería la interpretación frecuencista.

La interpretación subjetiva de la probabilidad lleva a un enfoque diferente de la inferencia estadística. En lo que se conoce habitualmente como el *enfoque bayesiano*, el proceso se basa en la revisión de nociones a priori acerca de los parámetros desconocidos θ, a la luz de los datos observados, utilizando la regla de Bayes. La información a priori acerca de θ tiene la forma de una distribución de probabilidad f(θ), es decir, θ es tratado como una variable aleatoria. La revisión de los a priori toma la forma de la distribución posterior f(θ/**x**) a través de la fórmula de Bayes:

$$\mathrm{f}(\theta/\mathbf{x}) = \frac{f(x/\theta) f(\theta)}{f(x)}$$

donde f(**x**/θ) es la distribución de la muestra.

Apéndice

Conjuntos

Como nos referimos una y otra vez a *conjuntos* de resultados de cierto experimento, conviene repasar algunos conceptos sobre teoría de conjuntos. Tomamos un *conjunto* como cualquier colección de objetos. Los objetos en un conjunto son sus *elementos*, y usamos la notación $x \in A$ con el sentido "x es un elemento del conjunto A". Un *subconjunto* de A es un conjunto cuyos elementos son a su vez todos elementos de A, y se escribe $B \subset A$ para denotar "B es un subconjunto de A". En cada aplicación, tendremos en mente un conjunto S del cual todos los conjuntos que consideremos son subconjuntos. Este conjunto "universal" en el contexto de probabilidad se identifica con el conjunto de resultados posibles de un experimento aleatorio.

Nuestro interés estará en distintos subconjuntos de S. En este contexto, el *complemento* de un conjunto A es el conjunto de los elementos que no están en A, pero pertenecen al conjunto S, con la notación $A^c = S - A$. El signo '–' se entiende como "excluyendo a todos los elementos de" A.

El complemento de S es el *conjunto vacío*: $\varnothing = S^c$, el conjunto que no contiene ningún elemento.

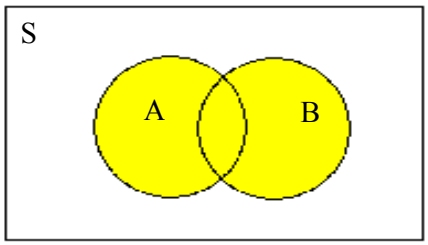
La *unión* de dos conjuntos es el conjunto de elementos que pertenecen a uno, o al otro, o a los dos, y se escribe $A \cup B$.

La *intersección* de dos conjuntos es el conjunto de elementos que pertenecen al mismo tiempo a los dos, y se escribe $A \cap B$.

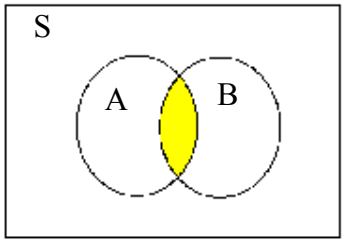
Dos conjuntos A y B son *mutuamente excluyentes* (o *disjuntos*) si no tienen ningún elemento en común, esto es, $A \cap B = \varnothing$.

Una partición del conjunto S es una colección de conjuntos disjuntos cuya unión es S. En símbolos, E_1 , E_2 ,..., E_n es una partición de S si $E_i \cap E_j = \varnothing \ \forall \ i \neq j$, y a su vez
$E_1 \cup E_2 \cup ... \cup E_n = S$.

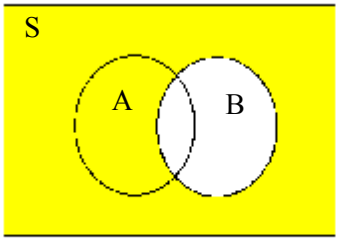
Los diagramas de Venn ilustran estas definiciones:
$A \cup B =$



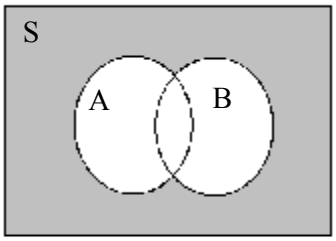
$A \cap B =$



$B^c =$



$(A \cup B)^c =$



Como ejercicio, se pueden utilizar diagramas de Venn para ilustrar dos reglas conocidas como *Leyes de De Morgan*

$A \cap B = (A^c \cup B^c)^c$
$A \cup B = (A^c \cap B^c)^c$

Factoriales, permutaciones y combinaciones

Esta sección está dedicada a la técnica para contar los resultados posibles de diferentes experimentos aleatorios, lo cual será de utilidad en ciertos problemas de probabilidad, por ejemplo cuando se tiene un conjunto de resultados equiprobables. En estos casos hallar las probabilidades de eventos equivale a contar el número de eventos elementales que cada evento contiene. En particular cuando se trata de la extracción al azar de elementos de S, la descripción de los eventos tiene relación con el número de ordenamientos posibles de los elementos de S.

Supongamos que tenemos un conjunto de n objetos. El problema es determinar cuántas maneras hay de ordenarlos. Tomemos por ejemplo las letras A, B, C (n = 3). Los órdenes posibles son seis en total:

ABC, ACB, BAC, BCA, CAB, CBA.

Para llegar a este número, consideremos primero cuántas formas hay de elegir la primera letra. Esta puede ser A, B o C, o sea tres formas diferentes. Para *cada una de ellas* hay luego dos formas de elegir la segunda, y ya no hay ninguna elección cuando se trata de la tercera. Entonces tenemos $3 \cdot 2 = 6$ formas de ordenar A, B y C.

Más generalmente, para n objetos hay n formas de elegir el primero, (n – 1) formas de elegir el segundo, y así sucesivamente, por lo que la fórmula general para contar ordenamientos de n objetos es:

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = n!$$

que se denomina "n factorial". Por convención, 0! = 1.

El siguiente problema que nos planteamos es cómo elegir x objetos en secuencia de un conjunto de n objetos. Es similar al problema de ordenamiento, pero con la diferencia de que ordenamos x objetos y desechamos los restantes n–x. Para contar l número de formas en que se puede hacer, partimos de la fórmula que usamos para ordenar n objetos, pero nos detenemos cuando hemos hecho x opciones. Para elegir el objeto que viene en lugar x, tenemos (n –x +1) opciones. Por tanto la fórmula es:

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n - x + 1) = \frac{n!}{(n-x)!} = P^n_x$$

o "permutaciones de n tomadas de a x".

Consideremos por ejemplo ordenamientos de tres letras extraídos del conjunto {A,B,C,D}. Debemos distinguir dos casos según nos importe o no el orden en que los elementos están colocados en cada ordenamiento. Si nos importa el orden en que los objetos son extraídos, debemos considerar la fórmula de permutaciones.

Si, por el contrario, interesa sólo qué elementos están incluidos y no en qué orden, BAC es equivalente a ABC y no debemos contarlos como ordenamientos diferentes. Si utilizáramos la fórmula de permutaciones para contar, cada ordenamiento obtenido estaría "repetido" tantas veces como sea posible ordenar los elementos que lo integran. Por ejemplo, en este caso tendríamos ABC, ACB, BAC, BCA, CAB y CBA. Si el orden no importa, estos ordenamientos son iguales, son un mismo ordenamiento repetido seis veces. Más generalmente, cada ordenamiento estará repetido x! veces. Para contar los ordenamientos posibles de n objetos tomados de a x *cuando no importa el orden* debemos dividir la fórmula de permutaciones, que contaba ordenamientos en un orden dado, por el número de veces que se repite cada ordenamiento.

Por tanto la fórmula para contar los *ordenamientos de n objetos tomados de a x sin importar el orden* está dada por

$$C_x^n = \frac{n!}{(n-x)!\,x!}$$

o "combinaciones de n tomadas de a x".

Bibliografía

Davidson, J., September course in statistics, class handouts, London School of Economics, 1993.

Spanos, A., Statistical foundations of econometric modeling, Cambridge University Press, 1994

Mood, A., Graybill, F., and Boes, D., Introduction to the theory of statistics, Mc Graw Hill, 1973.