Temas de Estadística Práctica

Antonio Roldán Martínez

Proyecto http://www.hojamat.es/

Tema 6

Distribuciones estadísticas teóricas.

El estudio de la probabilidad no entra en los objetivos de estos temas. Por esta razón, de aquí en adelante usaremos la probabilidad como límite de las frecuencias obtenidas en las muestras cuando el número total de datos tiende al infinito*. La Ley débil de los grandes números*, afirma, en efecto, con lenguaje más matemático, que

$$\lim_{n\to\infty} f = p$$

En este sentido usaremos en este curso la probabilidad, aunque no habrá inconveniente en usar hechos derivados de la teoría axiomática correspondiente y quien los conozca podrá seguir este tema con más comodidad.

Variable aleatoria

Llamaremos ***Variable aleatoria simple*** (discreta) a un conjunto de valores $\mathbf{X_1, X_2, X_3, ..., X_n}$ (llamados también *sucesos*) a los que les corresponden unos números (llamados *probabilidades*) , $\mathbf{p_1, p_2, p_3, ..., p_n}$ que cumplen:

a) Todas las probabilidades son positivas o nulas.

b) La suma de todas ellas es igual a la unidad

Como consecuencia de las dos propiedades anteriores se deduce que todas las probabilidades están contenidas entre 0 y 1. En lenguaje menos matemático, diremos que estas probabilidades miden las expectativas que podemos tener o las posibilidades que existen de que ocurra un suceso.

A las variables aleatorias también podemos designarlas con el nombre de *Distribuciones teóricas*.

La media en una distribución teórica viene dada por

$$E(X) = \sum X_i \cdot p_i$$

(en la teoría, la palabra media se sustituye por la de ***Esperanza matemática***)

La varianza, a su vez, viene dada por

$$V(X) = \sum {X_i}^2 \cdot p_i - E(X)^2$$

Una distribución de este tipo se representa mediante una tabla en la que estarán contenidos los valores de X y sus probabilidades. Por ejemplo, la distribución de una tirada de dado viene dada por

X	P
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Llamaremos ***función de distribución F(x)*** de una variable aleatoria, a la formada por las probabilidades acumuladas, es decir:

F(m) = Prob(x≤m)

(El símbolo ***Prob*** designa a la probabilidad de que sea cierta la comparación del paréntesis)

En una hoja de cálculo es imposible distinguir entre frecuencia y probabilidad, por lo que las usaremos de igual forma.

A continuación repasaremos las distribuciones teóricas más importantes usadas en la Estadística. Existen muchas más, cuya inclusión extendería demasiado este documento.

Distribuciones discretas teóricas más usadas

Uniforme

Una distribución se llama **uniforme** cuando todas las probabilidades son iguales. Como todas suman 1, cada una será igual a 1/n. La distribución del dado incluida en el apartado anterior es un caso típico de esta distribución. Otros ejemplos son el modelo de la tirada de una moneda equilibrada:

X	P
Cara	1/2
Cruz	1/2

Todas las extracciones equilibradas en los juegos de azar son de este tipo.

La media y la varianza de esta distribución se calculan del mismo modo que en una distribución de frecuencias relativas.

En el caso particular de una distribución uniforme discreta en la que X abarca el conjunto de números naturales de 1 a n (como las caras de un dado), la media coincide con **(n+1)/2**, y la varianza con **$(n^2-1)/12$.**

Distribución de Bernouilli

Una distribución de Bernouilli se compone de dos sucesos contrarios A y B, a los que se les suele llamar *éxito* y *fracaso*, con probabilidades **p** y **q** respectivamente. Es evidente que **q=1-p**. Si a **p** la llamamos probabilidad **a favor**, a **q** la designaremos por probabilidad **en contra**. Estas palabras son convencionales, pues si se estudia una epidemia, el *éxito* lo constituiría el ver aparecer un nuevo caso de infección.

Su distribución de probabilidad es:

X	P
Éxito	P
Fracaso	q

Todos los trabajos estadísticos efectuados sobre una *variable dicotómica*, con dos resultados A y B dan lugar a una distribución de Bernouilli: Tener o no un accidente en carretera, ganar o perder en el tenis, contraer o no una enfermedad, etc.

La media de una distribución de este tipo coincide con ***p***:

$$E(X) = p$$

y la varianza con

$$V(X) = pq$$

Un hecho que usaremos más adelante es el de que la máxima varianza se obtiene cuando **p** y **q** son iguales.

Distribución binomial

Esta importante distribución se aplica a pruebas repetidas de la ley de Bernouilli, con las siguientes condiciones:

a) Se realizan experimentos repetidos del tipo Bernouilli, **n** en total.

b) La probabilidad **p** permanece constante en todos ellos

c) Cada experimento es independiente del resultado anterior.

Llamamos a **n** el ***número de intentos***. Estamos interesados en estudiar el número de veces que aparece el suceso A (éxito). A su número de ocurrencias le llamaremos ***número de éxitos***.

Por tanto la ley binomial se aplicará cuando repetimos un experimento cumpliendo las condiciones a), b) y c) establecidas y deseamos estudiar el número de éxitos que obtendremos. Son de este tipo las tiradas múltiples de monedas, de dados, de ruleta, etc.

La probabilidad de obtener **r** éxitos en **n** intentos se demuestra que equivale a

$$B(r) = \binom{n}{r} p^r q^{n-r}$$

En la que el paréntesis es el número combinatorio **n sobre r**. Del hecho de que esta fórmula sea muy similar a la del Binomio de Newton proviene el nombre de ***binomial***.

La media (esperanza matemática) de esta distribución viene dada por

$$E(r) = np$$

y su varianza por

$$V(r) = npq$$

Consecuencia de esta es una fórmula que nos será muy útil, y es la de su desviación típica, que viene dada por

$$DESV(r) = \sqrt{npq}$$

La distribución binomial de probabilidad **p** y número de intentos **n** se representa generalmente por ***B(n,p)***

Distribución de Poisson

Esta distribución, llamada de *los sucesos raros*, es el caso límite de la binomial, con las siguientes condiciones:

a) El número de intentos **n** debe tender a infinito.

b) La propiedad **p** debe ser muy pequeña (de ahí el nombre de *suceso raro*)

c) El producto de **n.p** ha de ser constante, y al que llamaremos **m**.

Siguen esta distribución el reparto de estrellas en el firmamento, el cómo cayeron sobre Londres los bombardeos en la Segunda Guerra Mundial, las llamadas a urgencias, las averías de las máquinas de una fábrica, etc.

En general la siguen procesos estables, cuyo promedio de ocurrencias por unidad **m** se mantenga constante. Han de ser procesos aleatorios y las distintas ocurrencias deben ser independientes.

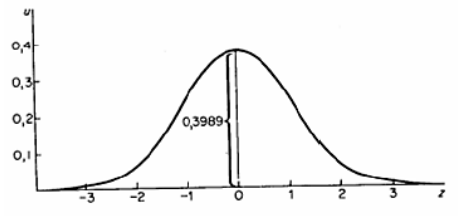
La fórmula de la probabilidad de que aparezcan **x** éxitos viene dada por la fórmula

$$p(x) = \frac{e^{-m}}{x!} m^x$$

La media de esta distribución es **m** y su varianza también vale **m**.

Distribución normal

La distribución ***Normal*** o ***ley de Gauss*** es la más usada de las distribuciones teóricas **continuas**. La popularizaron Gauss, en el estudio de los errores de las medidas, y también Laplace, pero ya la había usado Moivre como límite de la binomial.



Por su característica forma, se la conoce también como *campana de Gauss*. Aquí sólo nos interesa su definición y uso dentro de la Estadística.

La expresión de su función de densidad tiene dos versiones:

1) Normal de media μ y desviación típica σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

A esta distribución la denominaremos con el símbolo N(μ,σ)

2) Normal tipificada, que se aplica a una variable tipificada z =(x-media)/Desv. típ.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2)$$

La distribución tipificada se representa por N(0,1)

La distribución normal aparece en muchos fenómenos y estudios. Podemos destacar:

- Magnitudes que dependen de muchas causas independientes, cuyos efectos se suman y cualquiera de ellas aislada tenga efectos despreciables.

- Distribuciones de errores en las medidas

- Medidas de tipo antropológico (estaturas, pesos, inteligencia...) y biológico (glucemia, nivel de colesterol...)

- Límite de otras distribuciones estadísticas cuando **n** aumenta.

Tiene características matemáticas importantes:

Su media, mediana y moda coinciden en el valor cero.

Es simétrica y mesocúrtica.

Posee un valor máximo en la media, y unos puntos de inflexión en μ ± σ

Es asintótica, es decir, que si x tiende a infinito, su densidad de probabilidad tiende a cero.

El uso generalizado de esta distribución proviene de ser el límite de la **binomial** en virtud del Teorema de Moivre:

Si la variable **x** sigue una ley binomial de probabilidad **p**, entonces se cumple:

$$\lim_{n\to\infty} \frac{x - np}{\sqrt{npq}} = z$$

donde **z** sigue la ley normal N(0,1)

Es decir, que si obtenemos una medida tipificada **z** de una distribución binomial con **n** grande, la distribución de **z** se aproximará a la normal. Esta operación se suele efectuar también en procesos no binomiales: Para ajustar sus datos a una distribución normal, se tipifican en primer lugar y después se tratan como valores en la curva normal N(0,1).

Muchos autores han estudiado en qué circunstancias el *ajuste* entre binomial y normal funciona en la práctica. Algunos consejos son:

- Los productos **np** y **nq** deben ser ambos mayores que 3

- Si p<0,1, debe ser np>5

- Si p>=0,1, aunque np<5, el ajuste es aceptable.

Otras distribuciones continuas

En textos universitarios puedes encontrar muchas más distribuciones derivadas de la Normal. Tres de ellas, la ***chi-cuadrado***, ***T de Student*** y ***F de Snedecor***, son muy importantes en la Inferencia Estadística. Están defibidas, además, la *geométrica, la binomial negativa, las distribuciones Alfa, Beta y Gamma*, etc.

Relación entre frecuencia y probabilidad

El problema más importante que hay que considerar cuando se estudian las distribuciones teóricas es la relación que existe entre la probabilidad definida de forma teórica y las frecuencias observadas. Existe un criterio pragmático, y es que si se define una variable aleatoria y se asignan unas probabilidades, las observaciones posteriores de esa variable han de tener un cierto acuerdo con lo definido. Si se asignan los valores de 1/2 a las probabilidades en una tirada de moneda, sospecharemos que es defectuosa si después las frecuencias se alejan del 50%.

Hay dos metodologías para asignar probabilidades:

A) Se estudian muchas muestras aleatorias de una variable, y se asigna la probabilidad como límite de las frecuencias observadas. Podíamos llamarla probabilidad a posteriori, y se basa en la creencia en que las condiciones del experimento no cambian.

B) Se diseña un modelo teórico, basado generalmente en consideraciones de simetría e igualdad de oportunidades, y después se somete ese modelo a pruebas reales para ver si coinciden con lo previsto.

Podemos especificar más esta relación entre frecuencia y probabilidad mediante **los teoremas de los grandes números**, que aquí incluimos en la versión menos rigurosa.

Teorema central de la Estadística

Dada una variable aleatoria **x**, cuya función de distribución es F(x), en la que se han efectuado **n** observaciones, si se designa como FR(x) a las frecuencias acumuladas de dichas observaciones, se tendrá, para **n** tendiendo a infinito, que será 1 la probabilidad de que la diferencia F(x)-FR(x) sea cero.

De forma más sencilla: *Las frecuencias observadas tienen como límite las probabilidades cuando **n** tiende al infinito*.

Solemos llamar a este hecho la ***Ley de los grandes números***.

Si esta ley falla, es un indicio inequívoco de la probabilidad está mal definida.

Teorema central del límite

Podemos precisar aún más el carácter de límite de frecuencias que posee la probabilidad:

*Si las variables x_1, x_2, x_3,... x_n, tienen todas la misma distribución, con los mismos valores **m** para la media y **s** para la desviación típica, la variable*

$$\frac{x_{1+}x_{2+}x_{3+} \ldots + x_m - nm}{s\sqrt{n}}$$

sigue asintóticamente la distribución normal N(0,1).

Con la palabra asintótica queremos expresar su coincidencia para n tendiendo a infinito.

Consecuencia importante de esto es:

*En toda muestra aleatoria de una población de media μ y desviación típica σ, si llamamos **m** a la media de la muestra, se verifica que la variable*

$$\frac{m-\mu}{\sigma/\sqrt{n}}$$

es asintóticamente normal N(0,1)

Esta convergencia es aceptable a partir de n=30, por lo que este límite se toma para distinguir entre *pequeñas muestras*, en las que la media no se comporta de forma aproximadamente normal, y *grandes muestras*, en las que se sí se puede usar la distribución normal para describir el comportamiento de la media de la muestra.