Distribución Normal.

Distribuciones Continuas de Probabilidad.

Muchos experimentos dan como resultado una variable aleatoria continua. Es decir: que puede tomar cualquier valor en un intervalo, no sólamente un conjunto discreto de ellos. Un ejemplo: elegimos una persona al azar y medimos su altura. Esta altura se puede medir con toda la precisión que queramos. ¿Cuál es la probabilidad de que esta altura sea exactamente $1'781254$ cm.? Pues muy muy baja. Pero la probabilidad de que tenga entre $1'78$ y $1'79$ cm. es ya algo más razonable de manejar.

Una variable puede ser discreta y tomar valores decimales. Por ejemplo, imagina que un curso tiene 10 asignaturas y consideramos la variable aleatoria: «fracción del curso que tiene aprobada un alumno al azar». Está claro que un resultado de $0'15$ no tiene sentido. Sólo cuentan valores de $0'1$ en $0'1$. Una variable es continua cuando absolutamente todos los valores del intervalo son posibles.

Llamaremos **función densidad de probabilidad**, o meramente **función de probabilidad** a una función $f(x)$ que nos permite calcular la probabilidad de que x esté en cualquier intervalo $[a, b]$. Para ello nos basta con calcular el **área** de la función comprendida entre el eje X y las líneas $x = a$ y $x = b$.

Un ejemplo. Fíjate en la figura 1. Nos da una función densidad de probabilidad. Si queremos la probabilidad de que el resultado caiga en el intervalo $[2, 3]$ sólo tenemos que calcular el área sombreada.

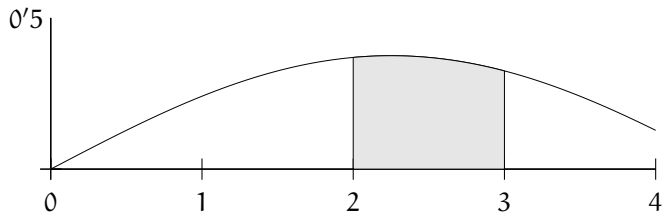

Figura 1. Área y probabilidad.

El área de una función entre dos puntos se llama la **integral** de dicha función entre esos dos puntos. En símbolos se suele escribir así:

$$P[a, b] = \int_a^b f(x)dx$$

Para nosotros el problema se reduce al cálculo de áreas. Desafortunadamente, hay muchos casos en los que no sabemos hacerlo...

De cualquier modo, es seguro que el área *total* bajo la gráfica de una función de probabilidad tiene que ser 1. En otras palabras: la probabilidad de obtener «cualquier valor» tiene que ser 1.

Un ejemplo importante es la distribución *uniforme*, en la que $f(x)$ toma un valor constante h en un intervalo y vale cero fuera de él:

$$f(x) = \begin{cases} h & \text{si } x \in [a, b] \\ 0 & \text{en caso contrario} \end{cases}$$

Sea el intervalo, p.ej., $[3, 6]$. En la figura 2 tenemos un bosquejo de la función.

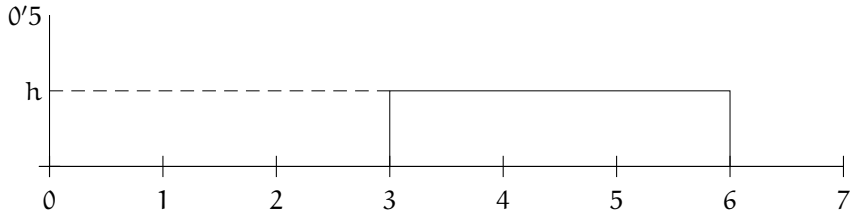

Figura 2. Distribución uniforme.

¿Cuál será el valor de h sabiendo que el área total bajo la función tiene que ser 1? Veamos: el área de un rectángulo es base por altura. La base mide 4, así que la altura tiene que medir $h = 1/4$.

Si ahora nos preguntan por la probabilidad de que el resultado salga entre $3'5$ y 4 tendremos que calcular el área:

$$P[3'5, 4] = \int_{3'5}^{4} f(x)dx = (4 - 3'5) \cdot \frac{1}{4} = 1/8$$

E1. Sea $f(x)$ una distribución uniforme sobre el intervalo $[0, 10]$. Calcula la probabilidad de obtener un resultado en el intervalo $[3, 4]$.

E2. Considera la siguiente función densidad de probabilidad:

$$f(x) = \begin{cases} x/2 & \text{si } x \in [0, 2] \\ 0 & \text{en caso contrario} \end{cases}$$

Calcula la probabilidad de obtener un resultado en el intervalo $[0'75, 1'25]$.

Función de Distribución. Media y Varianza.

A veces nos dan, en lugar de la función densidad de probabilidad una función mucho más cómoda. La función de distribución $F(x)$ está calculada de tal manera que la probabilidad de obtener un resultado en un intervalo determinado viene dada por:

$$P[a, b] = \int_a^b f(x)dx = F(b) - F(a)$$

Mientras que la integral (área) puede ser muy difícil de calcular, la resta es siempre una pavada.

La función de distribución $F(x)$ se define como la probabilidad de obtener un resultado menor o igual que x:

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

donde puedes ver que el área entre $-\infty$ y x corresponde realmente con la probabilidad de obtener un valor $\leq x$. Como es lógico, cuanto más pequeño sea x menor tiene que ser el valor de $F(x)$. En concreto, si $x \to -\infty$, entonces $F(x) \to 0$ (es *imposible* obtener un valor menor que $-\infty$). Y si $x \to +\infty$, entonces $F(x) \to 1$ (es *seguro* obtener un valor menor de $+\infty$).

E3. Un experimento aleatorio viene determinado por la función de distribución de probabilidad siguiente:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x/5 & \text{si } x \in [0, 5] \\ 1 & \text{si } x > 5 \end{cases}$$

Calcula la probabilidad de obtener un resultado en el intervalo $[2, 3]$.

Una variable aleatoria continua también tiene asociados una *media* y una *desviación típica*[1]. Tienen el mismo significado que en el caso discreto: la media es el valor más esperable y la desviación mide «cómo de dispersa» es la distribución.

[1] Aunque no te podemos decir aquí cómo calcularlas exactamente, te podemos dar la siguiente pista. Una distribución continua se puede «aproximar» por una distribución discreta si nos fijamos tan sólo en una serie de valores. Por ejemplo, contando de $0'1$ en $0'1$.

La Distribución Normal.

De todas las distribuciones continuas posibles la más importante es la llamada **normal** o **gaussiana**[2]. Surge cada vez que hay un valor «especial» (que será la media) y muchas pequeñas deviaciones independientes sumadas. Por ejemplo, cuando un arquero lanza una flecha, hay un valor especial (el centro), y luego muchas pequeñas desviaciones (el pulso, el viento, irregularidades de la flecha...) que, sumadas, dan lugar a la desviación final. Si el arquero tiene buena puntería, estas desviaciones serán más pequeñas que si es malo.

¿Qué forma tiene esta distribución? Mira la figura 3.

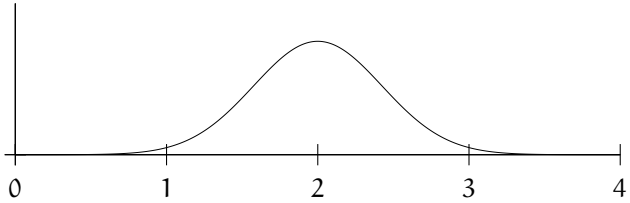

Figura 3. Distribución normal o gaussiana.

La media μ de una distribución normal marca el valor especial. La desviación típica σ marca la «falta de puntería» del arquero. Observarás que la función es simétrica respecto de la media: las desviaciones «por arriba» son igual de frecuentes que las desviaciones «por debajo».

En la figura 4 puedes ver dos distribuciones normales con la misma media y distinta desviación típica:

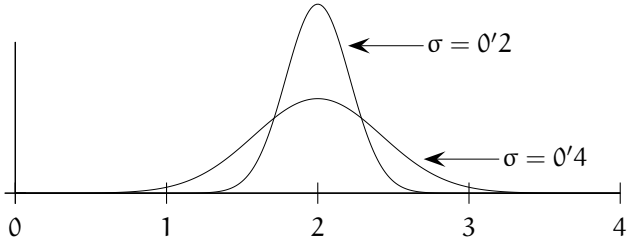

Figura 4. Dos gaussianas con misma media y diferente desviación.

Con estos valores, la función densidad de probabilidad toma la forma:

[2] Llamada así por C.F. Gauss (sí, el del método de Gauss para resolver sistemas), aunque la había utilizado ya De Moivre.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

La expresión es complicada y aparecen dos números «especiales» de las matemáticas: el número π y el número e. La razón por la que aparecen es interesante y profunda, pero no la podemos explicar aquí... De todas formas, nosotros no usaremos nunca esa expresión. Veamos cómo vamos a trabajar.

Uso de Tablas de la Distribución Normal.

En hoja adjunta a estos apuntes tienes una tabla con unos curiosos números. Es una tabla de la distribución normal, que nos permite estimar el valor de la función de distribución $F(x)$. Te recuerdo que con esta función, calcular la probabilidad de obtener un resultado en un intervalo determinado se reduce a una resta.

Llamaremos $\mathcal{N}(\mu, \sigma)$ a la distribución normal con media μ y desviación típica σ. Llamaremos **normal estándar** a la distribución $\mathcal{N}(0, 1)$, es decir: la que tiene media $\mu = 0$ y desviación $\sigma = 1$. La tabla se refiere a esta distribución.

Supón que queremos calcular $F(2'37)$. Buscamos primero en el eje vertical. Los valores van avanzando de $0'1$ en $0'1$. Llegamos a ver el $2'3$. Ahora echamos un vistazo al eje horizontal hasta llegar al $0'07$. El número que está en la intersección de las dos líneas es $F(2'37) = 0'9911$. Es decir, que en la normal estándar la probabilidad de obtener un valor $< 2'37$ es de $0'9911$.

¿Cuál es la probabilidad de obtener, en la distribución normal estándar, un resultado entre 2 y $2'2$?

Fácil. Calculamos a partir de los datos de la tabla:

$$P[2, 2'2] = F(2'2) - F(2) = 0'9861 - 0'9772 = 0'0089$$

Pero la tabla no trabaja con valores menores de 0. ¿Qué hacer si nos piden la probabilidad del intervalo $[-1, 1]$? Buscamos $F(1) = 0'8413$. También sabemos que $F(0) = 0'5$. Por tanto, $P[0, 1] = F(1) - F(0) = 0'3413$. Por simetría, $P[-1, 1]$ debe ser el doble: $0'6826$. Observa la figura 5.

E4. Estima el valor de $F(2'4785)$ a partir de la tabla. Pista: recuerda cómo funciona la interpolación lineal.

Tipificación de una Variable.

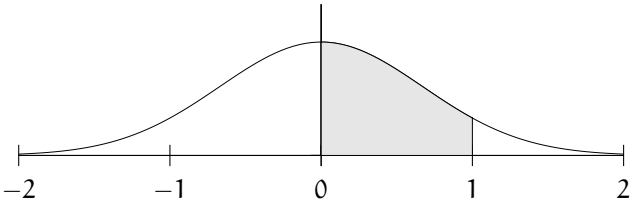

Figura 5. Calculando P[-1,1]. El área sombreada es $P[0, 1] = F(1) - F(0)$.

Ahora imaginad que no tenemos una distribución normal estándar $\mathcal{N}(0, 1)$, sino una $\mathcal{N}(\mu, \sigma)$. ¿Cómo usar las tablas de la normal?

Imagina que te piden calcular la probabilidad del intervalo $[12, 15]$ bajo la normal $\mathcal{N}(9, 3)$.

Como $\mu = 9$ y $\sigma = 3$, nos damos cuenta de que el valor $x = 12$ está a una distancia de la media de «una desviación típica». Por tanto, «es equivalente» al punto $x = 1$ de la normal estándar. En cambio, el valor $x = 15$ está a «dos desviaciones típicas» de la media. por tanto, «es equivalente» al punto $x = 2$. La probabilidad pedida es, por tanto, $P[12, 15] = F(2) - F(1) = 0'1526$.

La idea básica es transformar la variable de manera que el valor $x = \mu$ se convierta en el valor 0 y los valores $x = \mu + \sigma$ y $x = \mu - \sigma$ se conviertan en 1 y -1 respectivamente. ¿Cómo hacerlo? Mediante la siguiente transformación:

$$z = \frac{x - \mu}{\sigma}$$

que llamaremos **tipificación** de la variable. La z de un valor significa «a cuántas desviaciones estándar de distancia está de la media». En el problema anterior, con $\mu = 9$ y $\sigma = 3$,

$$x = 12 \quad \rightarrow \quad z = \frac{12 - 9}{3} = 1 \qquad\qquad x = 15 \quad \rightarrow \quad z = \frac{15 - 9}{3} = 2$$

E5. En una distribución normal de media 16 y desviación típica 4 calcula la probabilidad de obtener un resultado en el intervalo $[12, 24]$.

E6. El tiempo que tarda Maripuri en beberse un cubata es una variable aleatoria que, según sesudos experimentos llevados a cabo por sus amigos, se corresponde con una distribución gaussiana de media 10 segundos y desviación 3 segundos. Calcula la probabilidad de que tarde más de 15 segundos en beberse un cubata.

E7. La esperanza de vida entre los yanomamo del Brasil tiene una media de 40 años y una desviación típica de 8 años. Calcula qué fracción de la población vivirá más de 50 años.

E8. La vida media de un coche en una ciudad determinada es de 10 ± 2 años. Calcula, haciendo las suposiciones que creas convenientes, la probabilidad de que un coche viva entre 9 y 11 años.

Q1. En una distribución gaussiana cualquiera, calcula la probabilidad de que el resultado esté a una distancia de la media menor de (a) σ (b) 2σ (c) 3σ.

Intervalos Característicos.

Dada una distribución normal y una probabilidad p, se llama **intervalo característico** a un intervalo centrado en la media tal que la probabilidad de extraer un valor que pertenezca a él sea p. Llamaremos «radio característico» al radio de dicho intervalo[3].

Parece difícil, así que veamos un ejemplo. Se nos pide calcular, dada la normal $\mathcal{N}(4, 1'5)$, el intervalo característico para $p = 80\%$. En otras palabras, se nos pide averiguar x tal que la probabilidad del intervalo $[4 - x, 4 + x]$ sea $0'8$.

La probabilidad de pertenecer a nuestro intervalo es $0'8$, así que la probabilidad de *no* pertenecer a él será de $0'2$. Un intervalo característico siempre deja dos «colas» iguales, una a cada lado de la media. Cada cola reúne una probabilidad de $0'2/2 = 0'1$. ¿Cuál es el valor z tal que la probabilidad de obtener un resultado mayor que él es sólo de $0'1$? Buscamos en la tabla de la normal dónde cae el valor $1 - 0'1 = 0'9$ y encontramos el valor $z \approx 1'28$. ¿Qué hacemos con él?

Ese valor está tipificado, así que tenemos que «destipificarle». Recordamos que $z = 1'28$ significa que el valor está a $1'28$ «desviaciones» de distancia de la media. Como cada desviación vale $1'5$, el *radio* del intervalo será $1'5 \cdot 1'28 = 1'92$. Por tanto, el valor real es:

$$x = \mu + 1'28 \cdot \sigma = 4 + 1'92 = 5'92$$

Ése es el valor «por exceso», pero el intervalo característico está centrado en la media, así que calculamos el valor «por defecto»:

$$x = \mu - 1'28 \cdot \sigma = 4 - 1'92 = 2'08$$

[3] Es decir, la mitad de su longitud.

Así que el intervalo centrado en la media para el que la probabilidad es del 80% es el $[2'08, 5'92]$.

En general, para determinar el intervalo característico de una determinada probabilidad dados μ, σ y p se debe:

- Sumar a la probabilidad del intervalo la de la «cola izquierda»: $p + (1-p)/2$.
- Buscar $p + (1-p)/2$ de manera inversa en la tabla. Obtenemos el z característico para esa probabilidad.
- Multiplicamos ese z por la desviación, obteniendo el radio característico.
- El intervalo de confianza es la media $\pm$ dicho radio.

E9. Calcula el intervalo característico de la normal $\mathcal{N}(10, 5)$ para $p = 0'6$.

E10. Los tornillos que produce una máquina tienen un diámetro de $2 \pm 0'1$ milímetros. Calcula qué tamaño mínimo tendrá que tener un orificio para que quepa en él el 90% de los tornillos.

E11. En una distribución normal nos dicen que $[3, 5]$ es un intervalo de confianza con probabilidad del 90%. Obtén la media y desviación de dicha normal.

E12. Considera la normal $\mathcal{N}(10, 2)$. Se nos pide dibujar la gráfica del radio de confianza en función de la probabilidad p, cuando dicha p va del 80% al 100%.

Teorema Central del Límite.

La magia de la gaussiana viene del hecho de que es la distribución de probabilidad que aparece siempre que consideramos como variable «la suma de muchas otras variables aleatorias independientes». Esto se puede formular como teorema, aunque su demostración es muy complicada, y se llama **teorema central del límite**.

Un instante: la distribución binomial *también* es la distribución de probabilidad para la suma de muchas variables independientes. El número de caras que se obtienen tirando una moneda veinte veces es la suma de veinte variables independientes, cada una de las cuales vale 0 ó 1 dependiendo de que salga cruz o cara. Por tanto, ¿hay alguna relación?

Por supuesto que la hay: si el número de experimentos es muy grande, la distribución binomial se parece cada vez más a una distribución normal. Observa la figura 6, que muestra las distribuciones binomiales con $p = 0'25$ para $N = 5, 10, 15$ y 20.

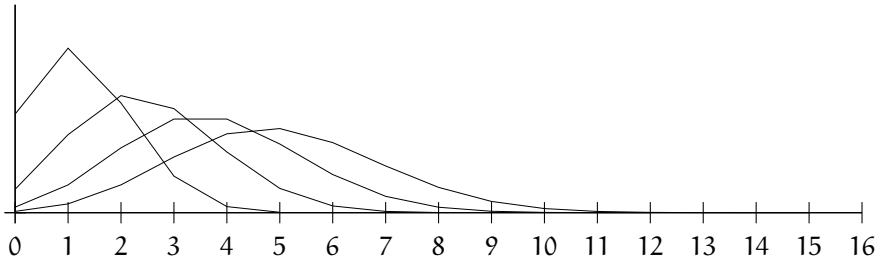

FIGURA 6. De cómo la binomial, cuando N se hace grande, va tendiendo hacia la normal, confirmando el teorema central del límite.

La última de las gráficas ya tiene un aspecto bastante «acampanado». La gente práctica ha inventado la siguiente regla del «dedo gordo»: si tanto Np como Nq son mayores que 5, una binomial se puede aproximar por una normal con misma media y desviación. Veamos para qué nos puede servir esto.

En una gran ciudad se sabe que el 30% de la población aborrece a su alcalde. Si extraemos una muestra de 100 personas, se pide la probabilidad de que 40 de esas personas lo hagan.

En principio se trata de una distribución binomial con $N = 100$ y $p = 0'4$. Así que la probabilidad será:

$$P(40) = \binom{100}{40} (0'3)^{40}(0'7)^{60} \approx 0'00849$$

Cálculo precioso salvo por el hecho de que el número combinatorio ha tenido que ser calculado por ordenador... ¿Qué podríamos hacer si no contamos con uno –como suele suceder en los exámenes?
Comprobamos si el criterio «del dedo gordo» para aproximar la binomial por la normal se cumple: Np y Nq valen 30 y 70 respectivamente, que son $\gg 5$. Por tanto es válido.
¿Cuál es la media y varianza de la distribución original? $\mu = Np = 30$ y $\sigma = \sqrt{Npq} = 4'58$, así que aproximaremos la distribución por una $\mathcal{N}(30, 4'58)$. Bien. Nos piden la probabilidad de obtener como resultado 40. ¡Mucho ojo! La distribución original era discreta y esta es continua, así que hay que tener un poco de cuidado. Calcularemos la probabilidad de resultado en el intervalo $[39'5, 40'5]$. Para ello tipificamos la variable:

$$x = 39'5 \;\to\; z = \frac{39'5 - 30}{4'58} = 2'07, \qquad x = 40'5 \;\to\; z = \frac{40'5 - 30}{4'58} = 2'29$$

Ahora sólo resta ir a la tabla de la normal y comprobar:

$$P[39'5, 40'5] = F(2'29) - F(2'07) \approx 0'00809$$

Cuando el resultado exacto era $0'00849$... Parece que la aproximación es mala, ¿no? Bueno... en la binomial exacta, para 41 personas la probabilidad es de $0'0053$, y para 39 personas de $0'01299$. Teniendo en cuenta que hemos tomado por buenos los valores entre $39'5$ y $40'5$, el resultado no está mal. La precisión es la esperable.

E13. Calcula la probabilidad de que se produzcan menos de 150 fracasos escolares en un centro de 800 alumnos, sabiendo que el porcentaje de fracaso en la ciudad es del 20%. Supón que cada alumno es un caso independiente.