2. DISTRIBUCIONES PROBABILÍSTICAS CONTINUAS

2.1 Distribuciones de variable continua

Las distribuciones teóricas continuas ocupan, sin lugar a dudas, el centro del análisis estadístico moderno.

"Es un hecho destacable que una ciencia que empezó analizando juegos de azar acabe convirtiéndose en el más importante objeto del conocimiento humano". Pierre-Simon Laplace.

Dentro del estudio de la teoría de la Estadística, el descubrimiento de las distribuciones probabilísticas continuas fue el inicio de una avalancha de posibilidades de aplicación que hoy en día aún sigue creciendo. En esta lectura se abordan los fundamentos necesarios para introducir las distribuciones probabilísticas continuas, así como la evolución de algunos conceptos expuestos en la lectura anterior acerca de las distribuciones de variable discreta. Adicionalmente, se presenta la distribución de variable continua más importante, la distribución normal, y su papel dentro del análisis de las distribuciones. La lectura culmina con algunas de las aplicaciones más naturales de las distribuciones probabilísticas (entre las cuales se encuentran las pruebas de bondad de ajuste) y deja el camino preparado para el estudio de una de las técnicas de análisis estadístico más común: las pruebas de hipótesis.

2.1.1 El paso de lo discreto a lo continuo

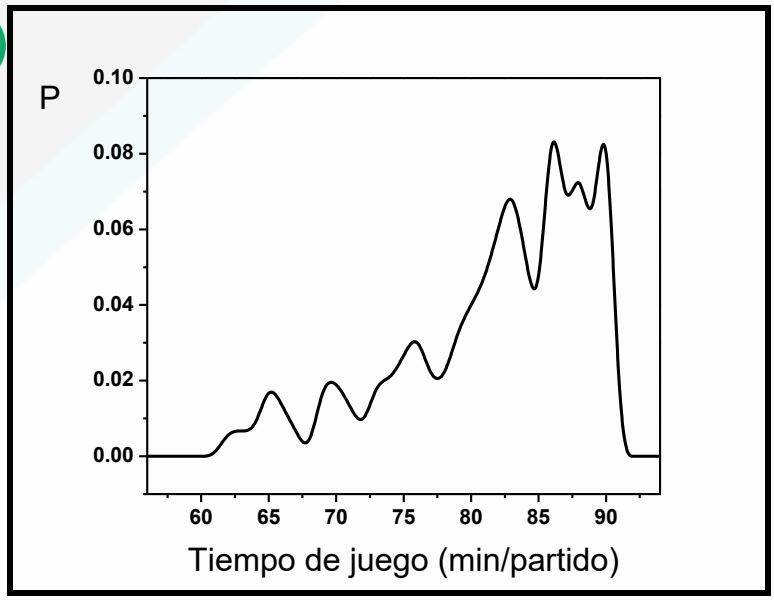
Las distribuciones probabilísticas introducidas en la lectura anterior surgen a partir de análisis discretos, para los cuales el espacio de posibilidades se compone por un número finito de sucesos y, por tanto, las variables aleatorias definidas en dicho espacio de posibilidades toman *valores discretos*. Muchos de los sucesos de la vida cotidiana cumplen adecuadamente con modelos del tipo discreto, como, por ejemplo, el número de *caras* que se obtienen al lanzar una moneda varias veces o la cantidad de partidos de fútbol que un determinado equipo gana durante una temporada; en estos casos, las variables son descritas mediante números enteros. Sin embargo, podrían existir variables cuyas mediciones no puedan ser expresadas en términos de números enteros. Existen múltiples situaciones en las cuales las distribuciones involucran un rango continuo de valores de mediciones posibles, y para estos casos la variable aleatoria se define como *variable continua*.

Por ejemplo, supongamos que estamos interesados en el tiempo que juega por partido cada jugador de una liga determinada de fútbol. En este caso, el espacio de posibilidades estaría representado por todos los jugadores de la liga, mientras que la variable aleatoria haría referencia al tiempo de juego por partido (variable continua) de cada suceso del espacio (jugador). Sobre un determinado rango continuo de valores de la variable aleatoria existirá una distribución de probabilidades, tal como ha sido descrito para el caso discreto. Sin embargo, esta vez no resulta conveniente dibujar cientos de barras por cada valor posible de la variable aleatoria, debido al carácter continuo, por lo que es preferible representar la distribución mediante una *función continua*.

Lo discutido anteriormente se ejemplifica en la **Figura 1**, en la que se muestra la distribución de probabilidades del tiempo de juego por partido obtenida a partir de datos de 150 jugadores de la *Major League Soccer* (MLS)[1]. Esta función que describe una distribución probabilística de una variable continua es conocida como **función de densidad probabilística**.

Figura 1: Función de densidad probabilística para la distribución del tiempo de juego por partido de un jugador de futbol.

[1] **Liga principal de fútbol de Estados Unidos y Canadá**



Fuente: elaboración propia a partir de datos de 150 jugadores de la MLS, obtenidos de http://www.mlssoccer.com/stats

Una de las nociones que resulta relevante para el paso de lo discreto a lo continuo consiste en la concepción de sumas infinitas, dada por el cálculo integral. En este sentido, las definiciones dadas para distribuciones de probabilidad discretas son extrapolables a las distribuciones de probabilidad continuas al sustituir las sumas sobre intervalos discretos por integrales definidas sobre intervalos continuos. Es decir:

Discreto **Continuo**

$$\sum_{i=0}^{n} x_n \longrightarrow \int_{0}^{n} x dx$$

La función de densidad probabilística, que se puede denotar como $p(x)$, cumple con determinadas propiedades que son consecuencia de su propia definición. Una de estas propiedades es que toma valores que están entre 0 y 1, debido a que se trata de valores probabilísticos. Por otra parte, tal y como ha sido descrito para las distribuciones discretas, la probabilidad total debería ser igual a 1, con la salvedad de que la probabilidad total es ahora la suma infinita de todos los valores, es decir, la integral:

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

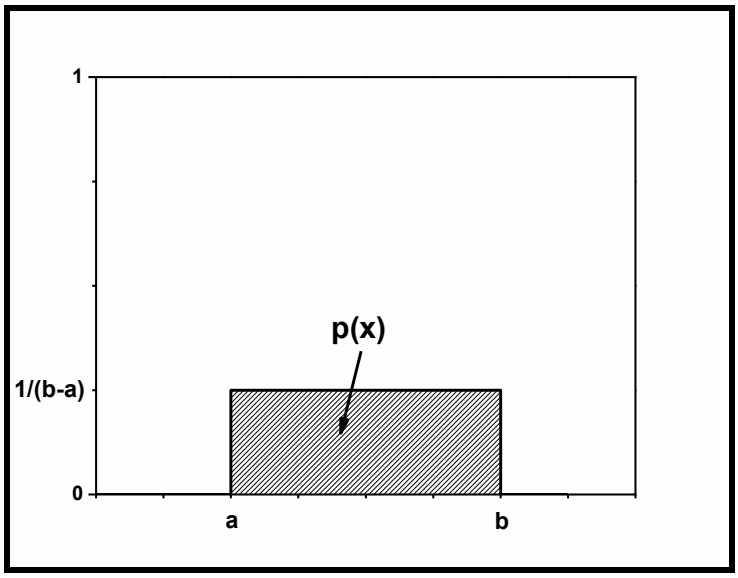
Por lo tanto, si se tiene cualquier intervalo específico de valores de la variable aleatoria, digamos $[a, b]$, entonces la probabilidad de que dicha variable tome valores dentro de ese intervalo estaría dada por:

$$prob(x \in [a, b]) = \int_{a}^{b} p(x)dx$$

El ejemplo más sencillo de una distribución continua es la **distribución uniforme** (**Figura 2**). La definición de esta función de densidad de probabilidad es relativamente trivial, pues se trata del caso en el que todos los sucesos posibles dentro de un intervalo de valores continuos de la variable aleatoria tienen la misma probabilidad. Lo importante para destacar es que dicho valor de probabilidad no es completamente arbitrario, depende del intervalo en que esté definida la función, puesto que debe cumplirse que la probabilidad total sea 1. Para un intervalo $[a, b]$, esta función se define como:

$$p(x) = \begin{cases} 0 & si\ x < a \\ \dfrac{1}{b-a} & si\ x \in [a, b] \\ 0 & si\ x > b \end{cases}$$

Figura 2: Ejemplo general de distribución uniforme



Fuente: elaboración propia.

Es posible encontrar en la literatura una infinidad de funciones de densidad probabilística definidas para distintos casos generales y prácticos. Entre ellas, destacan la *distribución beta*, la *distribución exponencial*, la *distribución F*, la *distribución gamma*, la *distribución* χ^2, la *distribución normal* y la *distribución t de Student*; las últimas tres funciones se abordarán en esta lectura.

2.1.2 Media y varianza en distribuciones continuas

Como ya se mencionó, los mismos fundamentos desarrollados para las distribuciones discretas pueden ser extrapolados al caso continuo. Entre ellos, se encuentra la definición de media, de varianza y de desviación estándar, para las cuales solo es necesario reemplazar las sumatorias de las definiciones en el caso discreto por integrales para el caso continuo. La **Tabla 1** resume estas definiciones.

Tabla 1: Definiciones de media y de varianza para distribuciones continuas

	Caso discreto	Caso continuo
Media (μ)	$\mu = \sum_i x_i p_i$	$\mu = \int_{-\infty}^{\infty} xp(x)dx$
Varianza (σ^2)	$\sigma^2 = \sum_i (x_i - \mu)^2 p_i$	$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$

Fuente: elaboración propia.

El caso de la desviación estándar mantiene exactamente su definición, tanto para el caso discreto como para el continuo, siendo $\sigma = \sqrt{\sigma^2}$.

Una forma de comprender mejor el significado de la media y de la varianza en el caso de distribuciones continuas consiste en asociar dicha distribución con una función de distribución de peso o de densidad de masa a lo largo de un cuerpo (de allí el nombre de densidad probabilística). En este caso, la media representaría el punto para el cual dicha distribución de pesos se balancea, es decir, es el punto de equilibrio o centro de masas de dicho cuerpo. Por otra parte, la varianza representaría cuán extendida se encuentra la masa a lo largo del cuerpo, es decir, sería una medida de la homogeneidad o la heterogeneidad de la masa del cuerpo.

Transformaciones afines sobre variables aleatorias continuas

Una de las ventajas que presenta el trabajo con variables aleatorias continuas es que es posible realizar transformaciones continuas sobre ellas. Estas transformaciones no son más que el uso de los valores de la variable aleatoria para la construcción de otra variable aleatoria a partir de una determinada regla. Dentro de estas transformaciones, las que resultan de mayor utilidad son las transformaciones afines, como la *traslación* y la *dilatación* o *contracción*.

Dada una variable aleatoria X, se define la traslación como la construcción de otra variable aleatoria Y de la forma:

$$Y = X + b$$

Donde b representa una constante conocida. El caso especial de $b = 0$ constituye la transformación trivial de equivalencia, donde los valores de Y son iguales a los de X.

Por otra parte, se define la dilatación o contracción como:

$$Y = aX$$

Donde a es una constante que define si la transformación es una dilatación ($a > 1$ o $a < -1$) o una contracción ($-1 < a < 1$). Para los casos en que $a < 0$, la transformación incluye una reflexión adicional a la dilatación o contracción; con el caso especial de $a = -1$, la transformación es solamente una reflexión. Los casos de $a = 0$ o $a = 1$ representan las transformaciones triviales anuladoras (todos los valores de Y son 0 sin importar el valor de X) y de equivalencia, respectivamente.

Resulta interesante saber qué ocurre con la distribución probabilística de una variable aleatoria sometida a alguna de estas transformaciones, para lo cual

es válido analizar cómo se transforman la media y la varianza. Utilizando las definiciones respectivas del valor esperado (media), varianza y desviación estándar es fácil demostrar que:

Valor esperado	Varianza	Desviación estándar
$\mu(X+b) = \mu(X) + b$	$\sigma^2(X+b) = \sigma^2(X)$	$\sigma(X+b) = \sigma(X)$
$\mu(aX) = a \times \mu(X)$	$\sigma^2(aX) = a^2 \times \sigma^2(X)$	$\sigma(aX) = \lvert a \rvert \times \sigma(X)$

Estas transformaciones resultan útiles para estandarizar las distribuciones probabilísticas y, de esta forma, reducir muchos casos de trabajo a una misma problemática. Esto lo veremos, en especial, para el caso de la *distribución normal*.

2.1.3 Distribución normal

A La *distribución normal* es, probablemente, la distribución probabilística continua de mayor uso y más conocida dentro de la Estadística. Esto se debe, en parte, a su constante aparición en casos de estudios reales de la vida cotidiana. Por ejemplo, supongamos que se lanza una pelota en línea recta y hacia arriba, y se mide la distancia desde el punto en que la pelota cae al suelo hasta el punto de lanzamiento. Si realizamos este experimento múltiples veces, se observa que esta variable muestra una distribución normal. Por otra parte, en muchos casos se utiliza esta distribución como criterio de evaluación de ciertos procesos. Por ejemplo, un profesor de una determinada asignatura espera que la distribución de las notas obtenidas por sus estudiantes cumpla con una distribución normal. Si se observasen desviaciones considerables respecto a esta distribución, puede ser un indicativo de que el proceso de enseñanza se encuentra polarizado hacia una subpoblación del grupo de estudiantes y, por tanto, no sea homogéneo. Sin embargo, la razón principal por la cual la distribución normal consiste en la distribución más importante en la estadística está dada por el *teorema del límite central*, que será discutido más adelante.

La **distribución normal estandarizada** se denota por $N(0,1)$ y se define a partir de la relación:

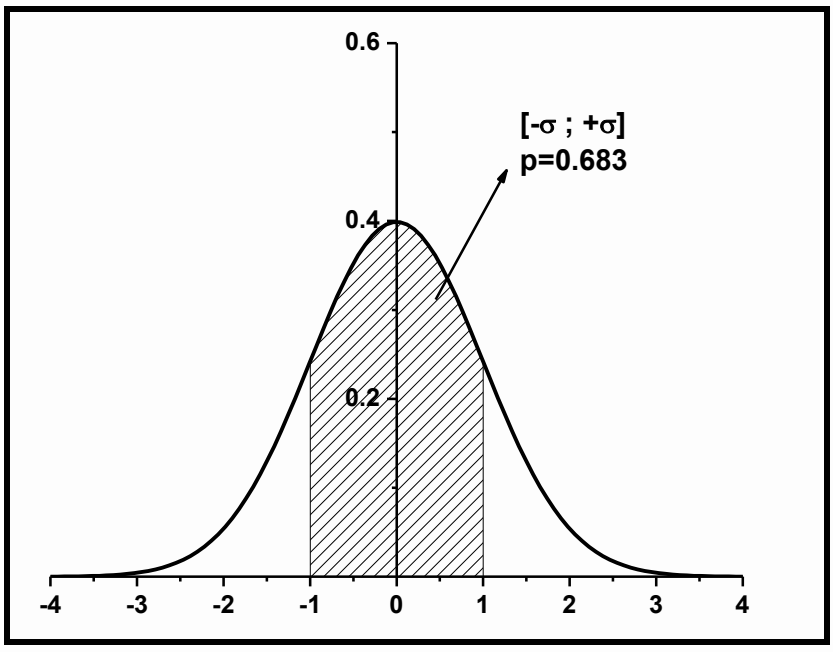
$$\phi(x) = e^{-\frac{x^2}{2}}$$

Esta expresión es comúnmente conocida como *función campana* o *función gaussiana*, debido a la forma del gráfico obtenido (**Figura 3**). El término "estandarizada" se refiere a la obtención a partir de esta expresión de una distribución de probabilidades que cumpla con $\mu = 0$ y $\sigma^2 = 1$. Para ello, es necesario realizar un proceso de normalización para garantizar que la probabilidad total sea 1. De esta forma, la expresión de la distribución normal estandarizada queda:

$$\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

Donde $\sqrt{2\pi}$ representa la constante necesaria para la normalización.

Figura 3: Gráfico de la distribución normal estandarizada



Fuente: elaboración propia.

De la **Figura 3** se desprende que esta función presenta un máximo aproximado de 0,4 y tiene carácter simétrico en torno a 0, debido a que es

una función par. Como se ha mencionado, esta función se utiliza para determinar valores de probabilidades, y resulta muy común representarla con la letra Z a las variables aleatorias que cumplen con esta distribución (estandarizada). De esta forma, la probabilidad de que la variable aleatoria tome valores en el intervalo $[a, b]$ está dada por:

$$p(Z \in [a,b]) = \int_a^b \phi(x)dx$$

Existen algunos intervalos $[a, b]$ que resultan interesantes dentro de la estadística, como por ejemplo, el intervalo $[-\frac{2}{3}; +\frac{2}{3}]$, para el cual se demuestra que la probabilidad es del 50%. Adicionalmente, es posible definir intervalos que están basados en el valor de la desviación estándar, como por ejemplo, $[-1; +1]$ y $[-2; +2]$, que para el caso de la distribución estandarizada representan los intervalos de una desviación estándar y dos desviaciones estándar, respectivamente. La utilidad de esta forma de definir los intervalos radica en que los valores de probabilidad obtenidos en cada caso (0,683 y 0,956, respectivamente) se mantienen para la distribución normal generalizada, que se discutirá más adelante. En la **Figura 3** se representa el intervalo correspondiente a una desviación estándar.

Otro de los intervalos que es muy utilizado en el análisis estadístico es el intervalo $[-1{,}96; +1{,}96]$, para el cual la probabilidad es del 95%. Este alto valor de probabilidad es realmente arbitrario, sin embargo, existe el consenso en la estadística de utilizar el 95% como valor de corte para denotar confiabilidad; es decir, expresar que el resultado de un estudio cualquiera es confiable significa que la probabilidad de acierto es igual o mayor al 95%.

En muchos casos de estudio en los que la distribución normal juega un papel importante, no necesariamente se cumple que $\mu = 0$ y $\sigma^2 = 1$, por lo que es preciso extender la función de densidad de probabilidad para un caso más general. En este sentido se define la distribución normal generalizada $\boldsymbol{N(\mu, \sigma^2)}$ mediante la función:

$$p(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

En este caso, el término constante necesario para la normalización de la probabilidad total es de $\sqrt{2\pi\sigma^2}$. Las variables aleatorias que cumplen con

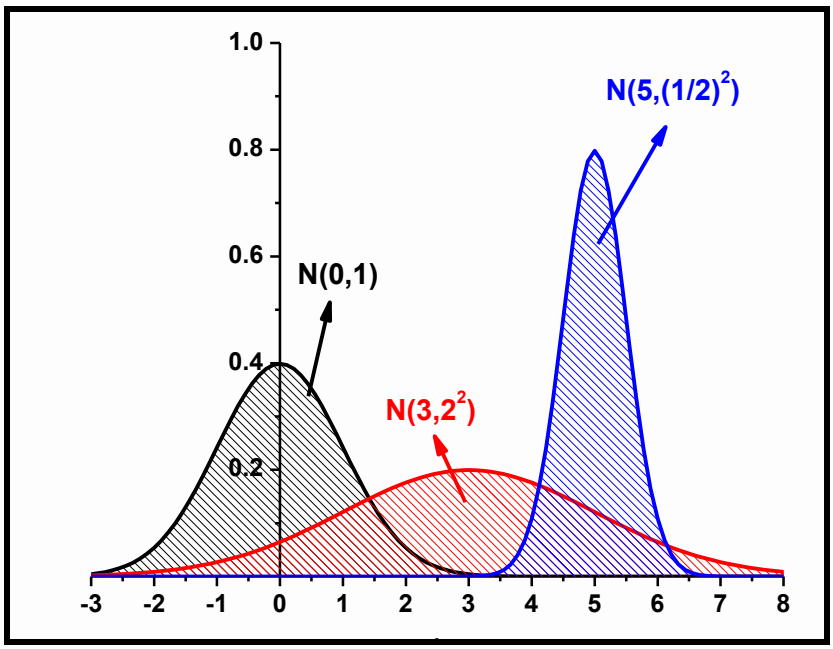
esta distribución suelen tener una denominación típica (X, Y…), y el uso de la letra Z para denotar la variable suele reservarse para la distribución estandarizada. La **Figura 4** muestra algunos ejemplos de distribuciones generalizadas.

La ventaja de la distribución normal estandarizada consiste en la posibilidad de ser utilizada para cualquier variable que cumpla con una distribución normal generalizada haciendo una transformación de variables. Supongamos que la variable aleatoria *X* cumple con una distribución normal $\boldsymbol{N(\mu, \sigma^2)}$ con media μ y con varianza σ^2, se define entonces la variable Z:

$$Z = \frac{X - \mu}{\sigma^2}$$

Utilizando las transformaciones explicadas anteriormente, es fácil mostrar que la variable Z cumple con una distribución estandarizada $\boldsymbol{N(0,1)}$.

Figura 4: Ejemplos de distribución normal generalizada



Fuente: elaboración propia.

2.1.4 Distribución t de Student y distribución de Pearson

En adición a la distribución normal existen otras distribuciones continuas muy utilizadas para diferentes análisis estadísticos. Entre estas distribuciones, se encuentra la *Distribución χ^2 de Pearson* y la *Distribución t de Student*. La deducción y la definición de la función de densidad probabilística de estas distribuciones reúnen conceptos matemáticos avanzados que escapan al objetivo de este curso. Sin embargo, debido a su importancia, haremos una breve discusión del significado y el uso de estas distribuciones.

Distribución χ^2 de Pearson

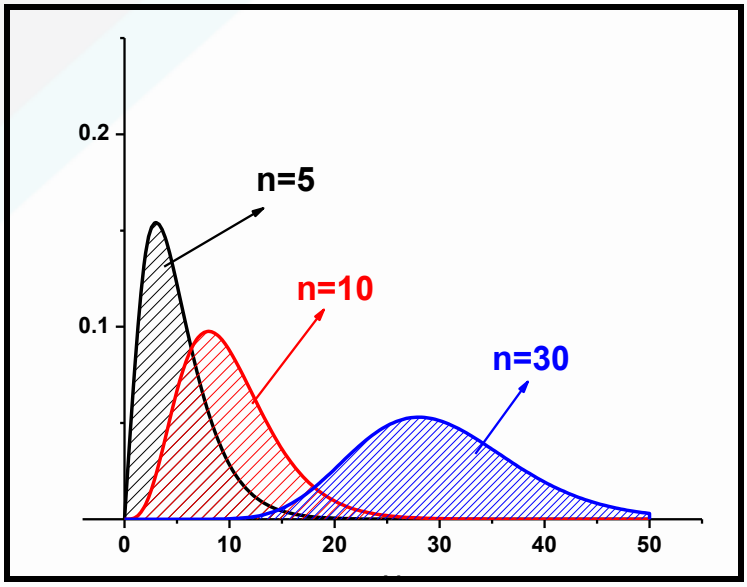
La Distribución χ^2 de Pearson tiene una importancia fundamental para el desarrollo de la inferencia estadística y su definición está relacionada con la distribución de cierta propiedad de datos obtenidos a partir de una distribución normal. Sean n variables aleatorias $Z_1, Z_2, \ldots, Z_n$, donde todas cumplen con una distribución normal estandarizada, se define la distribución χ^2 de Pearson como la distribución de la variable X dada por:

$$X = \sum_{i=1}^{n} Z_i^2$$

El número de variables aleatorias con distribución normal estandarizada (n) es conocido como grados de libertad de la distribución χ^2, y consiste en el parámetro que define dicha distribución.

Por estas razones, esta distribución consiste realmente en una familia de funciones de densidad probabilística que varían para cada grado de libertad. La Figura 5 muestra algunos ejemplos para distintos grados de libertad.

Figura 5: Algunas distribuciones χ^2



Fuente: elaboración propia.

Esta distribución es muy utilizada en pruebas de independencia y de bondad de ajuste, que se discutirán brevemente más adelante. Para resolver el problema de estimar la media de una población distribuida normalmente o en la estimación de la pendiente de una recta de regresión lineal, esta distribución resulta igualmente importante.

Distribución t de Student

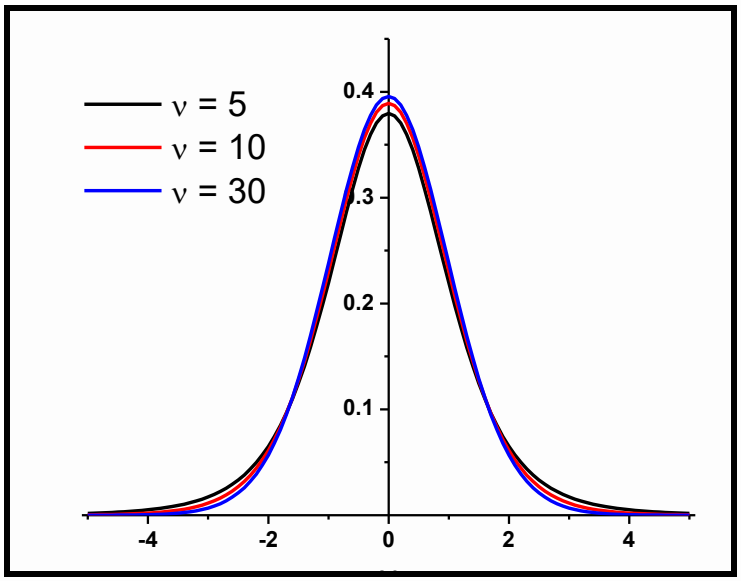
La Distribución t de Student surge de la problemática de estimar la media de una población distribuida normalmente para el caso en el que la muestra es pequeña y la varianza, desconocida. Para una muestra de tamaño n, la distribución t tiene $\nu = n - 1$ grados de libertad y, por lo tanto, existe una distribución t diferente para cada tamaño, lo que resulta entonces una familia de distribuciones, al igual que el caso de χ^2.

Siendo Z una variable aleatoria que sigue una distribución normal estandarizada y V otra variable aleatoria que cumple con una distribución χ^2 con ν grados de libertad, se define entonces la **distribución t de Student** como la distribución probabilística de la variable T dada por:

$$T = Z\sqrt{\frac{\nu}{V}}$$

La **Figura 6** muestra algunos ejemplos de distribuciones t para diferentes grados de libertad. Las curvas de estas distribuciones son de tipo campana, simétricas con $\mu = 0$ y $\sigma^2 > 1$. Mientras mayor es el grado de libertad más próxima a 1 es la varianza, siendo la distribución normal estandarizada el caso extremo cuando $\nu \to \infty$. Para n mayor que 30, la diferencia entre la distribución normal y la distribución t no se considera muy importante.

Figura 6: Algunas distribuciones t de Student



Fuente: elaboración propia.

Esta distribución tiene gran utilidad para determinar las diferencias entre dos medias, así como para construir los intervalos de confianza cuando no se conoce la desviación estándar de la población y es necesario estimarla a partir de una muestra.

Como dato curioso adicional, esta distribución fue descrita por William S. Gosset en 1908, quien la publicó con el seudónimo de Student, debido a que la fábrica de cerveza para la cual trabajaba prohibía la publicación de artículos científicos

2.2 Análisis de distribuciones probabilísticas

Para hacer uso de las distribuciones estadísticas en la estimación de probabilidades, o bien en la resolución de algún problema estadístico, es necesario comprender antes algunos resultados importantes que surgen del análisis de las distribuciones. En esta unidad se discuten algunos de estos resultados, que permiten armarnos de técnicas para utilizar las distribuciones de probabilidades. Adicionalmente, se discute la razón principal de por qué la distribución normal resulta de gran importancia para cualquier estudio estadístico.

2.2.1 Función de distribución acumulada

Supongamos que una variable aleatoria X tiene una distribución de probabilidades $p(X)$ y que estamos interesados en determinar la probabilidad de que X tome valores en un intervalo $[a, b]$, para ello sería necesario resolver la integral:

$$\int_a^b p(x)dx$$

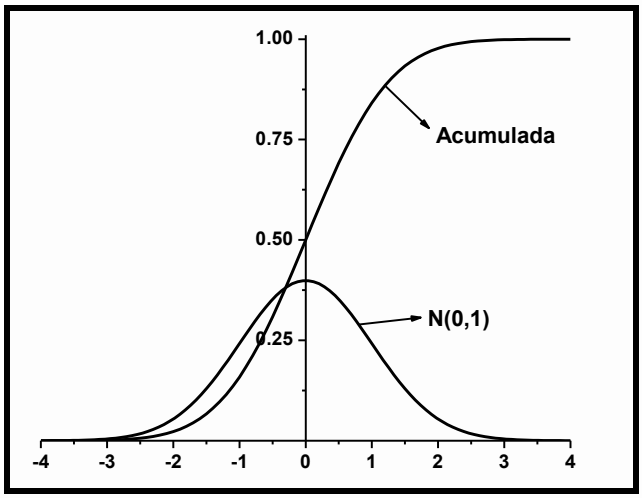
Sin embargo, en muchas ocasiones no resulta trivial resolver este tipo de expresiones. Por ello, existe una técnica útil para determinar fácilmente valores de probabilidades a partir de las funciones de densidad probabilística, que es la construcción de la *función de distribución acumulada*. Se define **la función de distribución acumulada** como la integral de $p(X)$, de la forma:

$$F(x) \equiv \int_{-\infty}^{x} p(t)dt$$

Esta función representa, como bien indica su nombre, la suma de todas las probabilidades para valores menores que un valor de x dado. Es por ello que esta función siempre será creciente a medida que aumente el valor de x, con los extremos de $F(x) \to 0$ cuando $x \to -\infty$ y $F(x) \to 1$ cuando $x \to +\infty$.

La **Figura 7** muestra la función de distribución acumulada correspondiente a la distribución normal estandarizada.

Figura 7: Función de distribución acumulada correspondiente a la distribución normal estandarizada



Fuente: elaboración propia.

La utilidad de la definición de esta función acumulada radica en el conocido **teorema fundamental del cálculo,** a partir del cual se puede obtener el siguiente resultado:

$$\int_a^b p(x)dx = F(b) - F(a)$$

La ventaja de este tipo de función consiste en que solo es necesario tabular un único set de datos para cada distribución, sin tener que resolver la integral para cada problema en particular. Los valores de la función de distribución

acumulada de las distribuciones más importantes se encuentran en tablas cuidadosamente construidas. De esta forma, para saber la probabilidad en un intervalo cualquiera, solo es necesario buscar en estas tablas los valores de los extremos del intervalo y hallar la diferencia. A pesar de que la resolución numérica de integrales no resulta muy complicada con los avances informáticos actuales, las tablas de las distribuciones de probabilidades más importantes continúan siendo muy utilizadas para la resolución de problemas estadísticos. Este tipo de tablas se encuentran en casi la totalidad de los libros que tratan la temática, aun en aquellos que son referenciados al final de esta lectura.

2.2.2 Teorema del límite central

Uno de los resultados más importantes de la matemática estadística en el análisis de las distribuciones probabilísticas es el **teorema del límite central**, que explica por qué la *Distribución Normal* resulta tan natural y aparece continuamente en la experiencia cotidiana. Existen varias formas diferentes de plantear este teorema, algunas matemáticamente más formales y cuya demostración escapa a los objetivos de este curso. Por ello, enunciaremos este importante teorema de una forma más natural y comprensible:

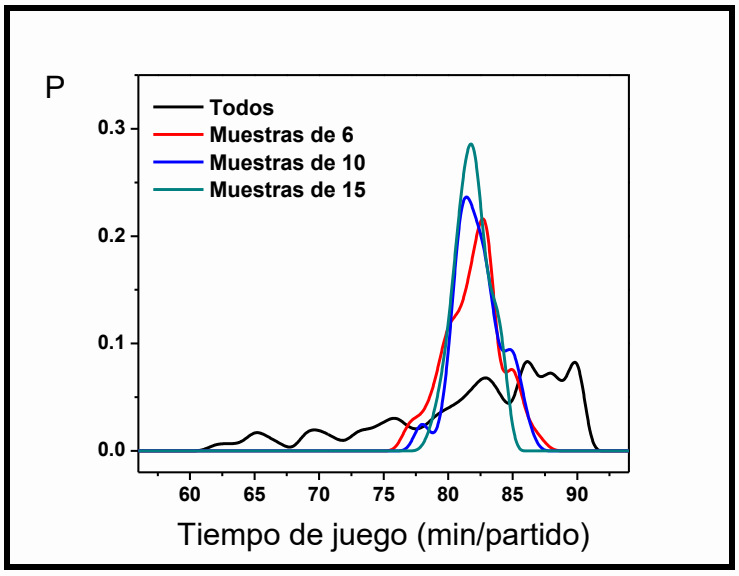
> Si para una variable aleatoria dada, que cumple con tener una distribución probabilística cualquiera, se toman muestras n veces y se promedian dichas muestras, entonces la distribución probabilística de estos promedios se acerca a una distribución normal a medida que n aumente (siendo el caso extremo de $n \to \infty$).

La importancia de este teorema se explica en el hecho de que se cumple sin importar la forma de la distribución probabilística inicial, o cuán diferente sea esta de una distribución normal. Es por este resultado que se plantea en estadística la diferencia entre población y muestra. La población de valores que puede tomar una variable podría seguir cualquier distribución de probabilidades, no obstante, el promedio de los valores de muestras de esta misma variable se acerca siempre a una distribución normal a medida que el tamaño de las muestras sea mayor.

Este teorema consiste en el resultado matemático que explica por qué la distribución normal aparece comúnmente en casos cotidianos de la vida real, debido a que la gran parte de las cosas que vemos o medimos son realmente promedios resultantes de efectos más pequeños.

Podemos poner a prueba este teorema utilizando los datos mostrados en la **Figura 1**, en la que se muestra la distribución probabilística correspondiente al tiempo de juego por partido de un jugador de futbol obtenida a partir de datos de 150 jugadores; es posible notar a simple vista que dicha distribución difiere considerablemente de la normalidad. Si agrupáramos los datos en grupos de un tamaño dado (muestras) y determináramos un promedio de estos grupos, el teorema nos plantea que dicho promedio se aproximará a la normalidad a medida que el tamaño de los grupos se haga mayor. La **Figura 8** muestra el resultado de este análisis para los datos de la **Figura 1**, en la que las distribuciones se obtienen promediando los datos en muestras de 6, 10 y 15 jugadores. Es posible observar cómo la distribución va cambiando y se aproxima a una forma de campana.

Figura 8: Utilización del Teorema del Límite Central para los datos de tiempo de juego por partido de un jugador de futbol.



Fuente: elaboración propia a partir de datos de 150 jugadores de la MLS, obtenidos de http://www.mlssoccer.com/stats

2.2.3 Pruebas de bondad de ajuste

Las distribuciones probabilísticas discutidas hasta el momento son distribuciones teóricas que podrían tomar los valores de variables aleatorias. No obstante, la medición real de los valores de la variable podría diferir de la distribución teórica perfecta, por ejemplo, debido a diversas fuentes de error sobre la medición de la variable. Es por ello que resulta importante poseer métodos que permitan conocer cuánto puede acercarse una variable a cumplir con una determinada distribución, y para ello existen las pruebas de bondad de ajuste o de significación.

En su generalidad, estas pruebas están basadas en alguna forma de comparación entre las probabilidades (o frecuencias) observadas y las probabilidades esperadas según la distribución que se desea comprobar. Esta forma de comparación normalmente implica la construcción de algunas variables que seguirán alguna distribución hipotética, a partir de las cuales se calcula un estadístico determinado que se utiliza para establecer el criterio de significación. Esto se evidencia en una de las pruebas de bondad de ajuste muy utilizada en la estadística: la **prueba** χ^2.

Si las probabilidades observadas y teóricas de una variable difieren solo por el azar, entonces estas diferencias constituirán variables que siguen una distribución normal (debido al teorema del límite central) y serían variables independientes para cada observación. Con estas premisas es posible definir el estadístico χ^2 de la forma:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - F_i)^2}{F_i}$$

Donde F_i son las frecuencias observadas y F_i las frecuencias esperadas según la distribución teórica a comprobar.

Este estadístico debe seguir una distribución χ^2 de Pearson con $k - m - 1$ grados de libertad, donde m es el número de parámetros de los que depende la distribución por estimar. Por lo tanto, existe una probabilidad asociada a este estadístico que puede ser determinada a partir de las curvas de distribución; no obstante, en la práctica resulta más simple establecer un valor de significancia (α) para afirmar o descartar la posibilidad de que la variable por analizar se ajuste a la distribución deseada. Para este caso, solo sería necesario comparar directamente el valor del estadístico χ^2

calculado con el valor teórico de la distribución χ^2 para los grados de libertad específicos y la probabilidad específica (significancia). El valor de $\alpha = 0{,}05$ (95%) resulta el más utilizado en la literatura; y en este caso, si el valor del estadístico calculado es menor que el correspondiente valor teórico, significa que hay un 95% de confianza en que el estadístico cumpla con una distribución χ^2, y por tanto que la variable analizada se acerque a la distribución deseada.

Existen otras diversas pruebas de bondad de ajuste que varían en la definición de los estadísticos utilizados, así como en los casos de aplicación. Entre ellas se encuentra la prueba de Shapiro-Wilk, utilizada para comprobar normalidad en muestras de pequeño tamaño; la prueba de Kolmogórov-Smirnov (o prueba K-S); la prueba de D'Agostino, útil en casos en los que hay valores repetidos, entre otras. Muchos *softwares* estadísticos se encuentran equipados con herramientas para realizar fácilmente este tipo de pruebas, como, por ejemplo, el programa **STATISTICA** o la herramienta de Análisis de Datos, de **Microsoft Excel**. Las pruebas de bondad de ajuste son, en realidad, una parte específica de un grupo de pruebas de mayor amplitud, las pruebas de hipótesis, que serán tratadas posteriormente.

2.2.4 Aplicaciones de las distribuciones de variable continua

Como se ha mencionado con anterioridad, la distribución normal tiene mucha importancia, puesto que aparece constantemente en la vida real. Es por ello que una de las aplicaciones más básicas que presenta consiste en el cálculo de probabilidades para eventos que cumplen con seguir esta distribución. Por ejemplo, supongamos que el tiempo de espera en el aeropuerto es una variable aleatoria *X* que se distribuye normalmente, con valor esperado, o media, de 2,5 horas y con desviación estándar de 0,34 horas. Podríamos entonces plantearnos la pregunta: ¿cuál es la probabilidad de tener que esperar más de tres horas en el aeropuerto?

Para ello es necesario determinar $p(X > 3)$, pero como la variable sigue una distribución normal, es posible transformarla para que siga una distribución normal estandarizada y así poder determinar la probabilidad

deseada mediante el uso de las tablas de la función de distribución acumulada. Se tiene, entonces:

$$p(X > 3) = p(X - 2{,}5 > 3 - 2{,}5) = p(\frac{X - 2{,}5}{0{,}34} > \frac{3 - 2{,}5}{0{,}34}) \cong p(Z > 1{,}5)$$

Según la definición de la función de distribución acumulada se tiene que:

$$p(Z > 1{,}5) = 1 - F(1{,}5) \approx 0.066$$

Se ha mencionado ya que las distribuciones probabilísticas pueden ser utilizadas para estimar con qué probabilidad una variable *X* cumple con una distribución específica (pruebas de bondad de ajuste), y que esta idea es extrapolable a otros tipos de estimaciones más generales (pruebas de hipótesis). Veamos un ejemplo de uso de la Prueba de χ^2 para comprobar que los datos de la **Figura 8** se aproximan en efecto a una distribución normal, como nos plantea el teorema del límite central.

Para este caso se determinaron las frecuencias observadas en el intervalo de 55 a 90 minutos/partido de cada una de las distribuciones y las frecuencias esperadas fueron estimadas a partir de asumir una distribución normal con μ y σ calculadas de la distribución real (ver definición en **Tabla 1**). Se calcula entonces el valor del estadístico χ^2 definido arriba y se determina el nivel de significancia según la distribución χ^2 con 32 grados de libertad (se utilizaron 35 $(90 - 55)$ categorías de frecuencias y la distribución normal tiene dos parámetros). Este valor de probabilidad puede obtenerse de las tablas de la distribución χ^2 y representaría la probabilidad de que la variable se acerque a una distribución normal.

La **Tabla 2** resume los resultados obtenidos.

Tabla 2: Resultados de la Prueba χ^2 sobre la normalidad de los datos de la Figura 8

	μ	σ	χ^2	P
Todos los datos	**81,81**	**7,01**	**675,090**	**0**
Muestras de 6	**81,98**	**2,15**	**43,883**	**0.0786156836148**
Muestras de 10	**82,26**	**1,78**	**12,397**	**0.9992854233183**
Muestras de 15	**81,86**	**1,30**	**4,074**	**0.9999999993783**

Fuente: elaboración propia.

Como se observa en la Tabla 2, el agrupamiento de los datos en muestras de 10 futbolistas es suficiente para que la distribución se acerque a una distribución normal con un 99,93% de probabilidad, de acuerdo con el planteamiento del teorema del límite central. El valor de $\chi^2_{(32)0,05}$ es de 46,194, por lo que el valor de corte del 95% es superado en todos los casos ($\chi^2 < \chi^2_{(32)0,05}$), salvo cuando los datos no son agrupados en muestras.

Ejemplos más generales, así como los métodos y las herramientas de uso de este tipo de pruebas, serán abordados posteriormente en el curso.

Referencias

Ambroggio, E. E., y Pérez Socas, L. B. (2016). *Certificado Estadística XXI: Estadística para la toma de decisiones* (curso 1, módulo 3, lectura 3). Córdoba: Córdoba: Universidad Empresarial Siglo 21.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2011). *Statistics for Business and Economics* (11ª ed.). Mason, OH: Cengage Learning.

Bertsekas, D. P., & Tsitsiklis, J. N. (1997). *Introduction to Probability* (2ª ed.). USA: Wiley.

Levine, D. M., Krehbiel, T. C., & Berenson, M. L. (2014). *Estadística para administración*. México: Pfarson Educación.

Major League Soccer. (2017). Major League Soccer Stats. Recuperado de: http://www.mlssoccer.com/stats

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to Probability and Statistics*. Boston, MA, USA: Brooks/Cole, Cengage Learning.

Rumsey, D. J. (2006). *Probability for Dummies*. USA: Wiley.