TEMA 5

MÓDULO:
TÉCNICAS AVANZADAS DE DATA MINING

DISTRIBUCIONES DE PROBABILIDAD Y EL TCL

FERRAN ARROYO

Licenciado en Empresariales y en Ciencias Actuariales y Financiaras por la UB. Máster Executive en Data Science por la MBIT School. Data Scientist.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

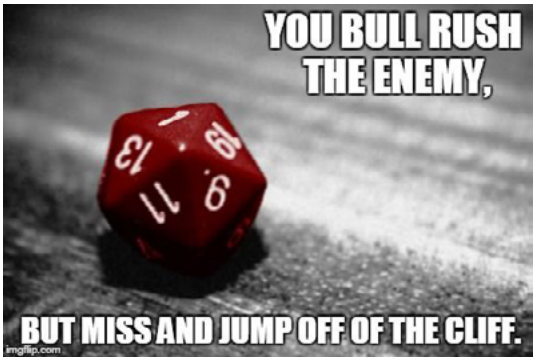
ÍNDICE

OBJETIVOS ESPECÍFICOS

- Conocer lo que es una distribución de probabilidad.
- Diferenciar entre función de distribución vs función de densidad.
- Conocer las principales distribuciones de probabilidad discretas y contínuas.
- Familiarizarse con el Teorema Central del Límite así como con sus principales utilidades prácticas.

1. LA DISTRIBUCION DE PROBABILIDAD

Para entender el concepto de distribución de probabilidad, es necesario conocer lo que es una **variable aleatoria.**



Fuente: https://logowiki.net/star-wars-episode-iv-logo.html

Una variable aleatoria no es más que una función que asigna un determinado valor al resultado de un experimento aleatorio. Por ejemplo, los posibles resultados de tirar un dado de 20 caras. En este caso concreto, nuestra variable aleatoria podría tomar los valores de 1 a 20, y la probabilidad de obtener un 1, sería de 1/20.

Dicho esto, **la distribución de probabilidad es una función que asigna, a cada suceso definido sobre la variable, la probabilidad de que dicho suceso ocurra.**

La distribución de probabilidad está definida sobre el conjunto de todos los sucesos y cada uno de los sucesos es el rango de valores de la variable aleatoria.

También puede decirse que tiene una relación estrecha con las distribuciones de frecuencia. De hecho, una distribución de probabilidades puede comprenderse como una frecuencia teórica, ya que describe cómo se espera que varíen los resultados.

La distribución de probabilidad está, completamente, especificada por la **función de distribución,** cuyo valor en cada x real es la probabilidad de que la variable aleatoria sea menor o igual que x. Es decir, **esta función acumula probabilidades.**

Así pues, las distribuciones de probabilidad nos permitirán saber la probabilidad de un suceso. La notación básica para saber la probabilidad en un determinado punto es la siguiente:

P(X) = probabilidad de que una variable aleatoria tome un valor específico de X.

P(X>2) = probabilidad de que una variable aleatoria tome un valor mayor a 2.

La función que utilizaremos para conocer la probabilidad de un suceso en un punto se llama **Función de Densidad.** La función que utilizaremos para saber la probabilidad de que un suceso sea mayor o menor que un determinado valor se llama **Función de Distribución.**

Por ejemplo, la probabilidad que un cliente que entra por la puerta me compre un producto exactamente, en este caso sería P(1).

La suma de todas las probabilidades para todos los valores posibles debe ser igual a 1.

Además, la probabilidad de un determinado valor o gama de valores debe estar entre 0 y 1.

Ten en cuenta esto, es más importante de lo que parece. Si en tu ejercicio, te sale una probabilidad superior a 1 o inferior a 0, dale varias pensadas y sobre todo, ni se te ocurra entregarlo:

Fuente: http://memegenerator.net/instance/51875086/sixth-sense-boy-i-see-negative-probabilities

Las distribuciones de probabilidad se pueden dividir en dos tipos:

- Distribuciones discretas.
- Distribuciones contínuas.

1.1 DISTRIBUCIONES DISCRETAS

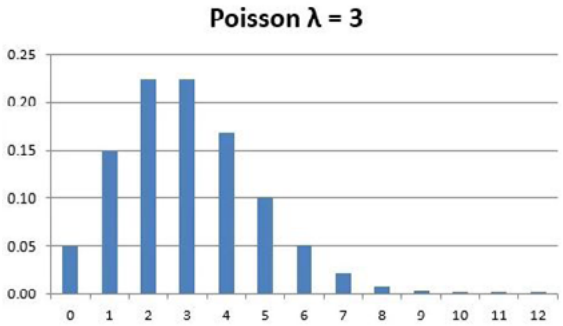
Las distribuciones discretas pueden asumir un número discreto de valores. Por ejemplo, el lanzamiento de monedas y el recuento de eventos son funciones discretas.

Este tipo de eventos pueden ser definidos dentro de distribuciones discretas porque no hay valores intermedios. Por ejemplo, en un lanzamiento de monedas sólo puede haber cara o cruz. Del mismo modo, si estás contando el número de coches que hay aparcados en la calle en un determinado momento, podrás contar 10 o 11, pero nunca 10.5.

Para las funciones de distribución de probabilidad discreta, cada valor posible tiene una probabilidad distinta de cero. Es decir, **su función de densidad siempre será mayor a 0.**

Por ejemplo, la probabilidad de lanzar un número específico en un dado es de 1/6. La probabilidad total para los seis valores es igual a uno, mientras que la probabilidad de lanzar un número par es de 3/6, es decir, P(2) + P(4) + P(6).

En la siguiente imagen puedes ver un ejemplo de distribución discreta:



Fuente: https://www.real-statistics.com/binomial-and-related-distributions/poisson-distribution/

En el gráfico de arriba, puedes ver la clásica distribución Poisson de media 3. Tal y como puedes ver, la probabilidad que una observación sea igual a 1 es del 0.15, es decir, del 15%. Este es el modo en que se calculan las probabilidades en las distribuciones continuas, mediante sumatorio. Formalmente:

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_probabilidad

1.1.1 DISTRIBUCIÓN BINOMIAL

La distribución binomial es una distribución de probabilidad discreta que cuenta el **número de éxitos en una secuencia de n ensayos de Bernoulli independientes entre sí,** con una probabilidad fija p de ocurrencia del éxito entre los ensayos.

Un experimento de Bernoulli se caracteriza por ser dicotómico, esto significa que, únicamente, dos resultados son posibles. A uno de estos resultados lo llamaremos éxito y tiene una probabilidad de ocurrencia **p,** mientras que al otro resultado lo llamaremos **fracaso,** y tendrá una probabilidad de **1-p.**

En la distribución binomial, el anterior experimento se repite n veces, de forma independiente, y se trata de calcular la probabilidad de un determinado número de éxitos. Para n = 1, la binomial se convierte, de hecho, en una distribución de Bernoulli.

Para representar que una variable aleatoria X sigue una distribución binomial de parámetros n y p, se representa del siguiente modo:

$$X \sim B(n, p)$$

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_probabilidad

Existen muchas situaciones en las que se presenta una experiencia binomial.

Cada uno de los experimentos es independiente de los restantes (la probabilidad del resultado de un experimento no depende del resultado del resto). El resultado de cada experimento ha de admitir sólo dos categorías (a las que se denomina éxito y fracaso). El valor de ambas posibilidades ha de ser constante en todos los experimentos, y se denota como p y q, respectivamente, o p y 1-p, de forma alternativa. Se designa como X a la variable que mide el número de éxitos que se han producido en los n experimentos.

Cuando se dan estas circunstancias, se dice que la variable X sigue una distribución de probabilidad binomial.

La función de probabilidad para la Distribución Binomial es la siguiente:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq p \leq 1$$

donde x=

$$\{0, 1, 2, \ldots, n\},$$

Siendo

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

las combinaciones de n sobre x (n elementos tomados de x en x),

donde:

- n: número de ensayos.
- p: probabilidad de éxito.
- X: variable aleatoria binomial.

EJEMPLO

Vamos a imaginar que se lanza un dado (con 6 caras) 51 veces y queremos conocer la probabilidad de que el número 3 salga 20 veces.

En este problema, un ensayo consiste en lanzar el dado una vez. Consideramos un éxito si obtenemos un 3. Encaso contrario, si no sale 3, lo consideramos un fracaso.

Definimos X como el número de veces que se obtiene un 3 en 51 lanzamientos.

En este caso, tenemos una X ~ B(51, 1/6) y la probabilidad sería P(X=20):

$$P(X = 20) = \binom{51}{20}(1/6)^{20}(1 - 1/6)^{51-20} = 0.0000744$$

A continuación, puedes ver sus principales estadísticos:

Media	np
Mediana	Uno de $\{\lfloor np \rfloor, \lceil np \rceil\}$[1]
Moda	$\lfloor (n+1)\,p \rfloor$
Varianza	$np(1-p)$
Coeficiente de simetría	$\frac{1-2p}{\sqrt{np(1-p)}}$
Curtosis	$\frac{1-6p(1-p)}{np(1-p)}$

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_binomial

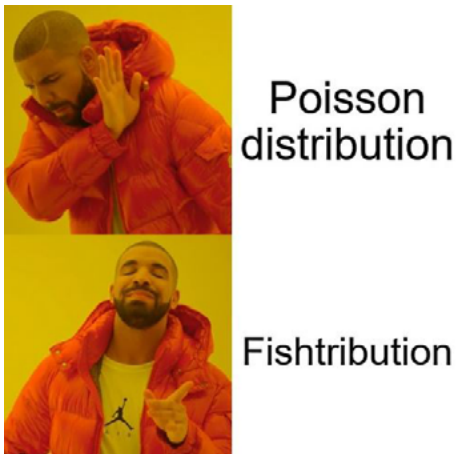
1.1.2 DISTRIBUCIÓN DE POISSON

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, **la probabilidad de que ocurra un determinado número de eventos durante cierto periodo de tiempo.**

Esta distribución es muy útil para trabajar con la probabilidad de ocurrencia de sucesos con probabilidades muy pequeñas.

Fue propuesta por **Siméon-Denis Poisson,** que la dio a conocer en 1838 en su trabajo Recherches sur la probabilité des jugements en matières criminelles et matière civile (Investigación sobre la probabilidad de los juicios en materias criminales y civiles).

Y respondiendo a la primera pregunta que habrás tenido al ver su nombre, te anticipo que sí. Probablemente, esta es una de las distribuciones con más memes:



Fuente: https://imgflip.com/i/3cn73w

La función de probabilidad para la Distribución de Poisson es la siguiente:

$$f(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!}$$

donde:

- k: es el número de ocurrencias del evento o fenómeno (la función nos da la probabilidad de que el evento suceda, precisamente, k veces).
- λ: es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado.
- e: es la base de los logaritmos naturales (e = 2,71828...).

EJEMPLO

Vamos a imaginar que el 2% de los libros encuadernados en cierto taller tiene encuadernación defectuosa.

Para obtener la probabilidad de que 5 de cada 400 libros encuadernados en este taller tengan encuadernaciones defectuosas, usamos la distribución de Poisson.

En este caso concreto, k es 5 y λ, el valor esperado de libros defectuosos, es el 2% de 400, es decir, 8.

Por lo tanto, la probabilidad que estamos buscando es:

$$P(5;8) = \frac{8^5 e^{-8}}{5!} = 0,092.$$

A continuación, puedes ver sus principales estadísticos:

Media	λ
Mediana	usualmente cerca de $\lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$
Moda	$\lceil \lambda \rceil - 1$
Varianza	λ
Coeficiente de simetría	$\lambda^{-1/2}$
Curtosis	$3 + \lambda^{-1}$

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_de_Poisson

1.2 DISTRIBUCIONES CONTINUAS

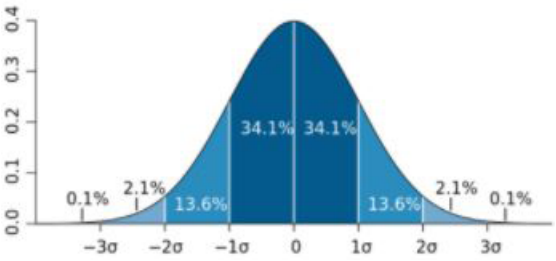
Las distribuciones continuas pueden asumir un número infinito de valores. Por ejemplo, la temperatura en grados de una determinada región.

A diferencia de las distribuciones discretas, en las que cada valor tiene una probabilidad distinta de cero, **los valores específicos en las distribuciones continuas tienen una probabilidad de cero.** Por ejemplo, la probabilidad de medir una temperatura que es exactamente 32 grados es cero.

¿Por qué?

Vamos a considerar la temperatura como un número infinito de otras temperaturas que son infinitesimalmente más altas o más bajas que 32. Al considerar la probabilidad que la temperatura sea exactamente 32, estamos estableciendo una probabilidad estricta en un valor infinitesimalmente pequeño, por lo que es equivalente a cero.

Así pues, las probabilidades de las **distribuciones continuas se miden sobre rangos de valores,** en lugar de puntos individuales. Una determinada probabilidad en una distribución continua está indicando la probabilidad de que un valor caiga dentro de un intervalo. Esta propiedad es sencilla de demostrar utilizando un gráfico de distribución de probabilidad:



En el gráfico de arriba, puedes ver la clásica **distribución normal.** Tal y como puedes ver, la probabilidad que una observación esté situado entre 0 y 2 sigma es de 34.1% + 13.6%. Este es el modo en cómo se calculan las probabilidades en las distribuciones continuas, mediante **integrales definidas:**

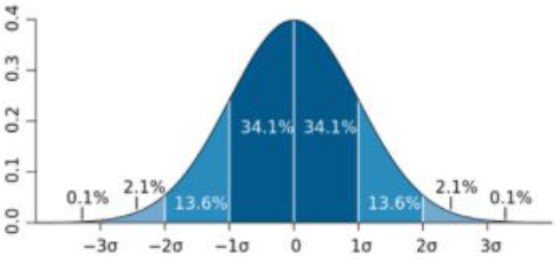
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Fuente: https://es.wikipedia.org/wiki/Distribuci%-C3%B3n_de_probabilidad

1.2.1 DISTRIBUCIÓN NORMAL

Esta distribución es una de las que más aparece en estadística y en teoría de la probabilidad. También, podrás encontrarla definida como **Distribución Gaussiana.**

La gráfica de su función de densidad tiene una forma acampanada y **es simétrica respecto de un determinado parámetro estadístico.** Esta curva se conoce como campana de Gauss y es el gráfico de una función gaussiana:



La importancia de esta distribución radica en que permite modelar numerosos fenómenos naturales, sociales y psicológicos.

Mientras que los mecanismos que subyacen a gran parte de este tipo de fenómenos son desconocidos, por la enorme cantidad de variables incontrolables que en ellos intervienen, el uso del modelo normal puede justificarse asumiendo que cada observación se obtiene como la suma de unas pocas causas independientes.

La distribución normal, también, es importante por su relación con la estimación por mínimos cuadrados, uno de los métodos de estimación más simples y antiguos.

Sin lugar a dudas, la distribución normal es la más extendida en estadística y muchos test estadísticos están basados en una “normalidad” más o menos justificada de la variable aleatoria bajo estudio.

La función de probabilidad para la Distribución Normal es la siguiente:

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\,du$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du, \quad x \in \mathbb{R}.$$

donde:

- μ es la media (también puede ser la mediana, la moda o el valor esperado, según aplique)
- σ es la desviación típica [estándar es un anglicismo]
- σ^2 es la varianza
- φ representa la función de densidad de probabilidad

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

También podemos definir la Distribución Normal a través de la función de densidad:

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

Fuente: https://es.wikipedia.org/wiki/Distribuci%-C3%B3n_normal

IMPORTANTE

La función de distribución normal estándar es un caso especial de la función donde la media (mu) es igual a 0 y la desviación estándar (sigma) es igual a 1:

$$\Phi(x) = \Phi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\,du, \quad x \in \mathbb{R}.$$

Fuente: https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

A continuación, puedes ver sus principales estadísticos:

Media	μ
Mediana	μ
Moda	μ
Varianza	σ^2
Coeficiente de simetría	0
Curtosis	0

Fuente: https://es.wikipedia.org/wiki/Distribuci%-C3%B3n_normal

2. TEOREMA CENTRAL DEL LÍMITE

El Teorema Central del Límite establece que si se tiene una población con media μ y desviación estándar σ, y se toman muestras aleatorias, suficientemente grandes, de la población con reemplazo, entonces la distribución de las medias de la muestra se distribuirá, aproximadamente, como una Distribución Normal.

En el caso de las muestras aleatorias de la población, podemos calcular la media muestra como:

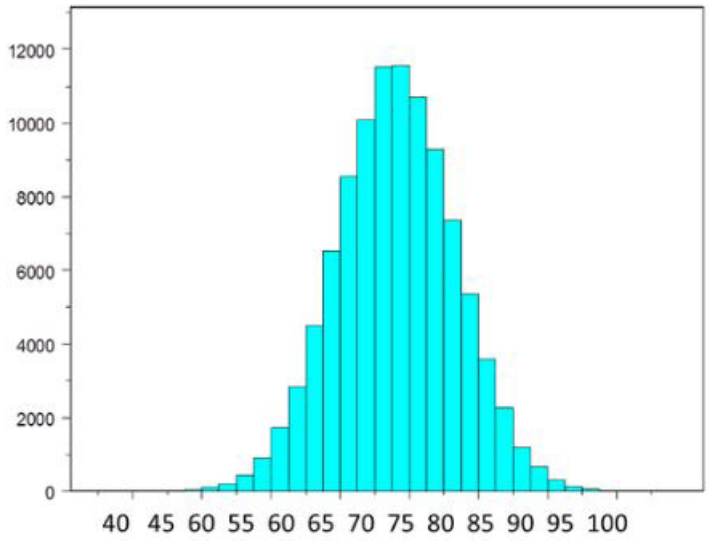
$$\mu_{\bar{x}} = \mu$$

Y la desviación estándar de la muestra:

$$(\sigma_{\bar{x}}) = \frac{\sigma}{\sqrt{n}}$$

2.1 TCL PARA UNA POBLACIÓN NORMAL

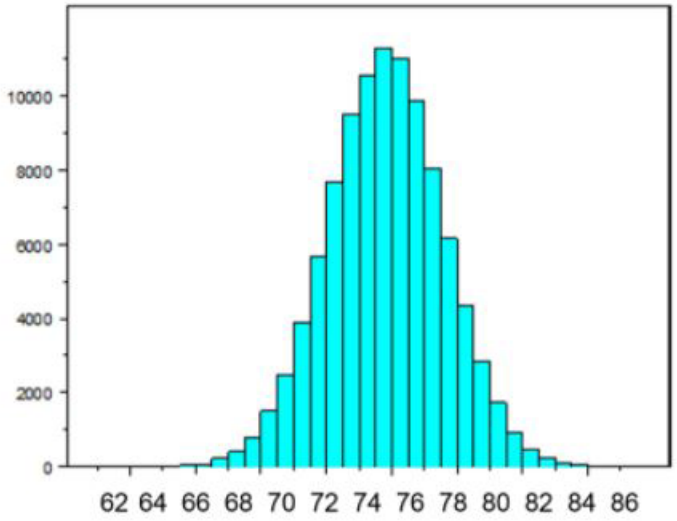
En la siguiente imagen, puedes ver población normalmente distribuida cuya media es de 75, y su desviación estándar es de 8:



Si tomamos muestras aleatorias simples (con reemplazo) de tamaño n=10 de la población y calculamos la media de cada una de las muestras, la distribución de las medias de las muestras debe ser, aproximadamente, normal según el **Teorema Central del Límite.**

Pese a que el tamaño de las muestras aleatorias simples es inferior a 30, dado que la población de origen está distribuida normalmente, esta limitación no es un problema.

En la siguiente imagen, puedes ver la distribución de una de las muestras. Pese a no ser exactamente igual que la distribución de la población general, se parece bastante, tal y como podrás comprobar:



Si te fijas, podrás comprobar que el rango de la distribución ha variado sensiblemente, ya que mientras la distribución poblacional tiene, aproximadamente, un rango de 60-100, el rango de la distribución de la muestra es de 66-84.

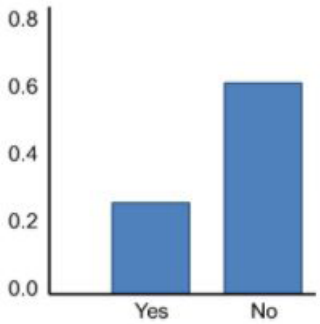
Si calculamos la desviación estándar de la muestra, obtendremos 2.5:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{10}} = 2.5$$

Este es un valor distinto al de la desviación estándar poblacional que es de 8, por eso, puedes observar un rango más estrecho.

2.2 TCL PARA UNA DISTRIBUCIÓN DICOTÓMICA

Ahora, vamos a suponer que queremos medir la característica X en una determinada población, y esta característica es dictónomica (por ejemplo: el éxito de un determinado tratamiento médico: SI o NO) con un 30% de éxito (p=0.30):



El Teorema Central del Límite se puede aplicar, incluso, a poblaciones dicotómicas como ésta, siempre que el mínimo de **np** y **n(1-p)** sea por lo menos 5, donde n se refiere al tamaño de la muestra, y **p** es la probabilidad de éxito en un ensayo determinado.

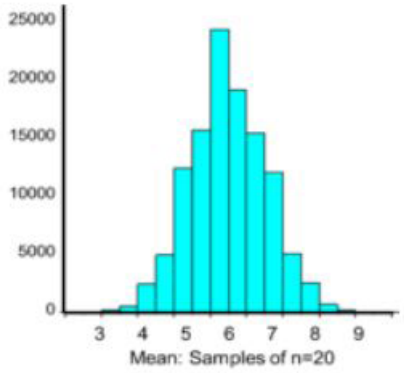
En este caso concreto, tomaremos muestras de n=20 con reemplazo, por lo que min(np, n(1-p)) = min(20(0,3), 20(0,7)) = min(6, 14) = 6. Por lo tanto, el criterio se cumple.

Anteriormente, vimos que la desviación estándar y la media de una Distribución Binomial es:

Media	np
Mediana	Uno de $\{\lfloor np \rfloor, \lceil np \rceil\}$[1]
Moda	$\lfloor (n+1)\,p \rfloor$
Varianza	$np(1-p)$
Coeficiente de simetría	$\frac{1-2p}{\sqrt{np(1-p)}}$
Curtosis	$\frac{1-6p(1-p)}{np(1-p)}$

(toma en consideración que la desviación estándar no es más que la raíz cuadrada de la varianza).

La distribución de la muestra poblacional basada en n=20 es la siguiente:



Así pues, la media de la muestra la podemos calcular como:

$$\bar{X} = np = 20\left(0.3\right) = 6$$

Y su desviación estándar como:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{n(p)(1-p)}}{\sqrt{n}}$$

$$\sigma_{\bar{X}} = \frac{\sqrt{20(0.3)(0.7)}}{\sqrt{20}} = 0.46$$

Ahora, en lugar de tomar muestras de n=20, supongamos que tomamos muestras aleatorias simples (con reemplazo) de tamaño n=10.

En este escenario, no cumplimos el requisito de tamaño de la muestra para el TCL (es decir, min(np, n(1-p)) = min(10(0.3), 10(0.7)) = min(3, 7) = 3). **El tamaño de la muestra debe ser mayor para que la distribución se aproxime a la normalidad.**

PARA SABER MÁS

Los siguientes papers adjuntos pueden serte de utilidad para profundizar más en el Teorema Central del Límite:

- *TCL_Demostración.pdf:* donde podrás ver la demostración del Teorema Central del Límite.
- *TCL_Ejemplos.pdf:* donde encontrarás algunos teoremas relacionados con el TCL y ejemplos resueltos.

3. ACTIVIDAD GUIADA

El ejercicio propuesto consiste en una pequeña demostración práctica de cómo funciona el TCL mediante un elemento común que conocemos todos/as: el dado de 6 caras.

Todos/as sabemos que puede tomar valores entre 1 y 6 con una probabilidad de 1/6 para cada cara.
Lo primero que haremos es simular los resultados de un dado creando una función llamada roll:

```
# Carga paquetes necesarios
require(plyr)
require(ggplot2)
require(stringr)
```

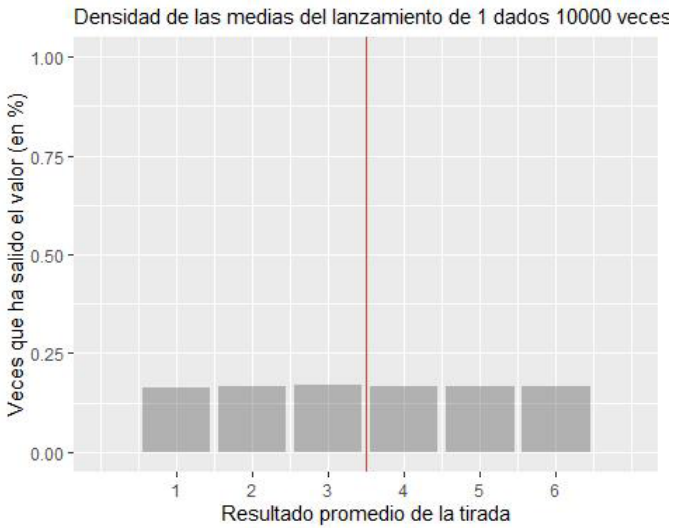
```
# m = numero de veces lanzado el dado
# n = numero de dados lanzados
roll <- function(m, n){
  set.seed(1234)
  means <- plyr::ldply(1:m, function(x){
    return(mean(sample(1:6, n, replace = TRUE)))
  })
}
```

Ahora que ya tenemos la función **roll()** creada, podemos llamarla para ver los resultados que arrojan 10.000 lanzamientos de 1 único dado.

Graficamos el resultado con un gráfico de barras:

```r
# Lanzamiento de 1 dado 10.000 veces
n_ <- 1
m_ <- 10000

g<-ggplot(roll(m = m_, n = n_),
          mapping = aes(x = V1)) +
  geom_vline(xintercept = 3.5, colour = "tomato3") +
  labs(
    subtitle = str_interp('Densidad de las medias del lanzamiento de ${n_} dados ${m_} veces'),
    x = 'Resultado promedio de la tirada',
    y = 'Veces que ha salido el valor (en %)'
  ) +
  geom_bar(aes(y = ..prop..), alpha = 0.4) +
  scale_x_continuous(
    breaks = 1:6,
    lim =  c(0, 7)
  ) +
  lims(
    y = c(0, 1)
  )
g
```

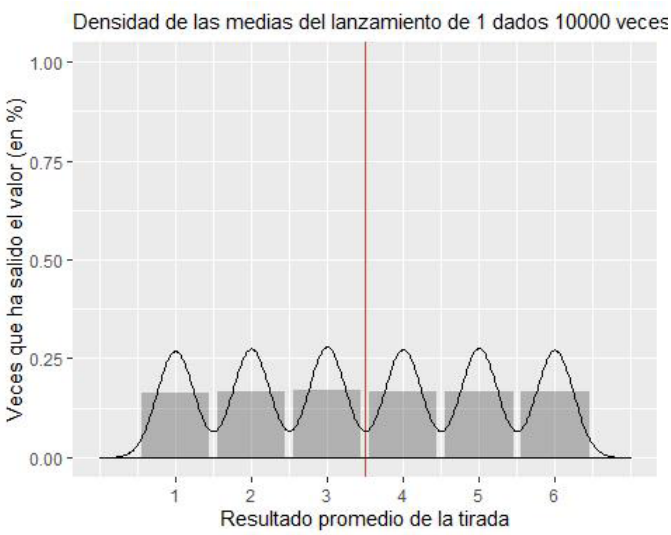


Estos resultados no deberían sorprendernos demasiado salvo que el dado esté trucado.

Fuente: https://g2-collectibles-hobbies.myshopify.com/collections/dice/products/loaded-dice-pair

Vamos a añadir el argumento **geom_density():**

```r
# Añadimos la función de densidad
g <- g +
  geom_density()
g
```



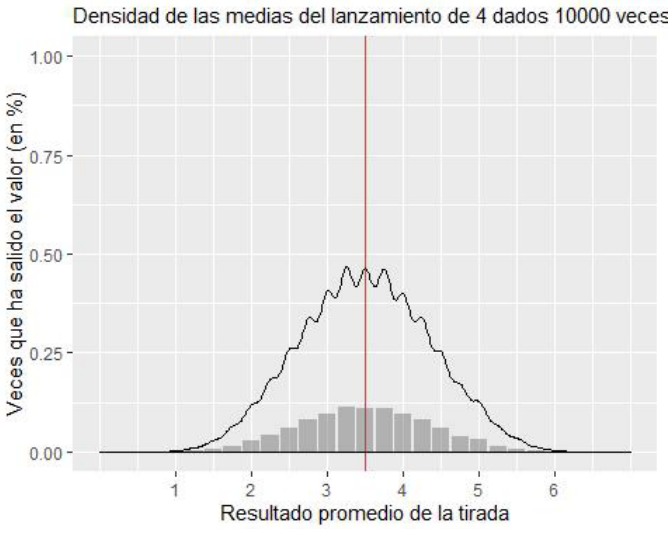
No te preocupes al ver que la función de densidad no está en 0 entre los valores 1,2,3,4,5 y 6. Esto se debe a que la curva se ha suavizado. En realidad, hemos puesto **una función de densidad continua sobre algo que es discreto.**

Técnicamente, acabamos de producir 10.000 muestras de tamaño 1 (hemos tirado 1 dado 10.000 veces).

Vamos a incrementar el número de dados lanzados a 4:

```
# Lanzamiento de 4 dados 10.000 veces
n_ <- 4
m_ <- 10000

g<-ggplot(roll(m = m_, n = n_),
          mapping = aes(x = V1)) +
  geom_vline(xintercept = 3.5, colour = "tomato3") +
  labs(
    subtitle = str_interp('Densidad de las medias del lanzamiento de ${n_} dados ${m_} veces'),
    x = 'Resultado promedio de la tirada',
    y = 'Veces que ha salido el valor (en %)'
  ) +
  geom_bar(aes(y = ..prop..), alpha = 0.4) +
  scale_x_continuous(
    breaks = 1:6,
    lim =  c(0, 7)
  ) +
  lims(
    y = c(0, 1)
  ) +
  geom_density()
g
```



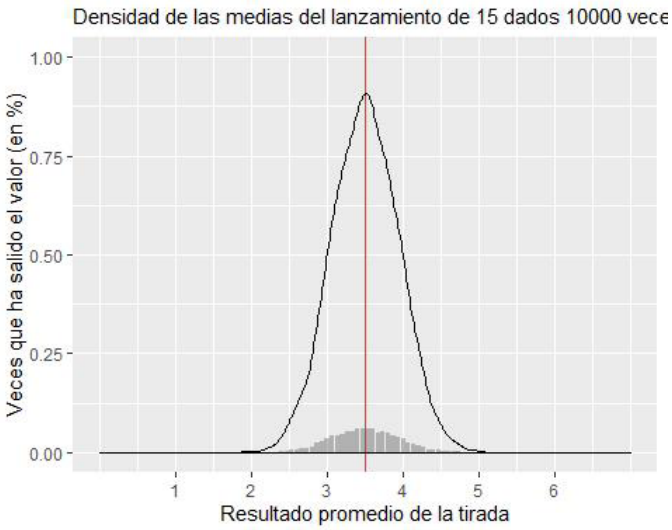
No solo la curva empieza a parecerse cada vez más a una Distribución Normal, si no que se puede observar que las colas de la distribución empiezan a situarse en 1 y 6, convirtiéndolos en los valores menos probables.

Esto tiene sentido, ya que para que el resultado promedio de la tirada sea 1 o 6, es necesario que en los 4 dados salga un 1 o un 6 a la vez.

Vamos a aumentar el número de dados lanzados a 15:

```r
# Lanzamiento de 15 dados 10.000 veces
n_ <- 15
m_ <- 10000

g<-ggplot(roll(m = m_, n = n_),
          mapping = aes(x = V1)) +
  geom_vline(xintercept = 3.5, colour = "tomato3") +
  labs(
    subtitle = str_interp('Densidad de las medias del lanzamiento de ${n_} dados ${m_} veces'),
    x = 'Resultado promedio de la tirada',
    y = 'Veces que ha salido el valor (en %)'
  ) +
  geom_bar(aes(y = ..prop..), alpha = 0.4) +
  scale_x_continuous(
    breaks = 1:6,
    lim =  c(0, 7)
  ) +
  lims(
    y = c(0, 1)
  ) +
  geom_density()
g
```



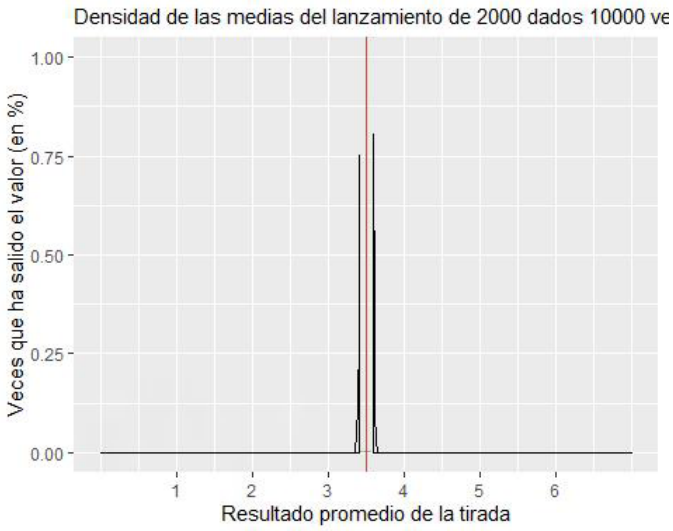
Ahora ya podemos ver la clásica campana que caracteriza a la Distribución Normal. En este caso, podemos ver que el pico de la distribución es mucho más elevado que en el caso anterior (es una distribución más apuntalada).

En este caso, los valores extremos todavía parecen más improbables que en el caso anterior. Concretamente, 2 y 5 ya son eventos que suceden raras veces.

Vamos a simular 10.000 tiradas de 2000 dados:

```r
# Lanzamiento de 2000 dados 10.000 veces
n_ <- 2000
m_ <- 10000

g<-ggplot(roll(m = m_, n = n_),
          mapping = aes(x = V1)) +
  geom_vline(xintercept = 3.5, colour = "tomato3") +
  labs(
    subtitle = str_interp('Densidad de las medias del lanzamiento de ${n_} dados ${m_} veces'),
    x = 'Resultado promedio de la tirada',
    y = 'Veces que ha salido el valor (en %)'
  ) +
  geom_bar(aes(y = ..prop..), alpha = 0.4) +
  scale_x_continuous(
    breaks = 1:6,
    lim =  c(0, 7)
  ) +
  lims(
    y = c(0, 1)
  ) +
  geom_density()
g
```



Tal y como puedes ver, esta es una distribución muy muy delgada, donde la mayoría de valores están rondando el valor promedio (3.5). Tal y como puedes ver, la dispersión, prácticamente, ha desaparecido, por lo que, utilizando la fórmula que hemos visto anteriormente:

$$\left(\sigma_{\overline{x}}\right) = \frac{\sigma}{\sqrt{n}}$$

donde **n**= 2000 y **mu** es 3.5,

podemos deducir que, manteniendo la media población, un incremento de n supone una reducción de la desviación estándar con límite en 0.

Así pues, podemos concluir que **cuanto mayor sea el tamaño de la muestra, más delgada y apuntalada será la Distribución Normal resultante.**

IDEAS CLAVE

- La distribución de probabilidad es una función que asigna, a cada suceso definido sobre la variable, la probabilidad de que dicho suceso ocurra.

- La función que utilizaremos para saber la probabilidad de un suceso en un punto se llama Función de Densidad.

- La función que utilizaremos para saber la probabilidad de que un suceso sea mayor o menor que un determinado valor se llama Función de Distribución. Es decir, esta función acumula probabilidad.

- La suma de todas las probabilidades para todos los valores posibles debe ser igual a 1.

- La función de densidad siempre será mayor a 0 en distribuciones discretas, e igual a 0 en distribuciones continuas.

- El Teorema Central del Límite establece que si se tiene una población con media μ y desviación estándar σ y se toman muestras aleatorias suficientemente grandes de la población con reemplazo, entonces la distribución de las medias de la muestra se distribuirá, aproximadamente, como una Distribución Normal.