INTRODUCCIÓN A LAS DISTRIBUCIONES CONTINUAS Y EL TEOREMA CENTRAL DEL LÍMITE (10 MINUTOS)

Hemos aprendido a identificar distribuciones discretas y también a reconocer algunas de sus características más importantes, como valor esperado y varianza. Sin embargo, no siempre podremos usar las distribuciones discretas para modelar las cosas que nos interesan. Por ejemplo, ¿cuál es la edad promedio de la sala de clases? ¿Cuánto tardan en atender a una persona en un banco? ¿Cuánto es el ingreso de un recién egresado?

Todas las variables de las preguntas anteriores tienen algo en común y es que son continuas. Hasta ahora solamente habíamos estudiado las probabilidades para variables discretas, como una elección (sí o no) o un número de eventos. Sin embargo, muchas variables son continuas, es decir, son fraccionables y entre dos valores siempre podremos encontrar un tercero. Por ejemplo, entre dos personas con 21 y 22 años podemos encontrar (aunque no necesariamente esté en nuestra muestra) a alguien con 21 años y 3 meses. El ejercicio anterior lo podemos seguir replicando siempre, hasta encontrar a personas que tienen horas o segundos de diferencia en su edad.

Lo mismo pasa con muchas otras variables. Tomen el salario como ejemplo. En marketing puede ser necesario modelar el ingreso de las personas para saber cuánto están dispuestas a pagar por un producto y a que público apuntar. En finanzas, el salario puede determinar si un grupo de individuos comprarán o no instrumentos financieros. Para profesionales trabajando en RRHH es importante saber cómo se mueven los salarios en el mercado. Y para los economistas la distribución de ingresos es uno de los temas más importantes de políticas públicas.

En esta clase, aprenderemos a modelar variables continuas para luego terminar con una aplicación muy importante que nos permitirá simplificar los cálculos: el Teorema Central del Límite.

Ejercicio Variables Continuas: Indique al menos 3 ejemplos de variables aleatorias continuas.

DISTRIBUCIONES DE PROBABILIDAD CONTINUA (15 MINUTOS)

Como dijimos anteriormente, las variables aleatorias continuas no pueden ser modeladas por distribuciones discretas como Bernoulli, Binomial o Poisson. Para poder modelar variables continuas tenemos que usar distribuciones de probabilidad continuas. Antes de trabajar con ellas definiremos dos conceptos importantes: función de densidad y función de distribución de probabilidad.

Función de densidad: nos indica la forma en que se reparten las probabilidades de una determinada variable continua. Por ejemplo, las horas de estudio de los alumnos probablemente se concentran más en las 5 horas a la semana, que en las 25 (aunque a los profesores les gustaría que fuera lo contrario). Si definimos la función de densidad como $f(x)$, entonces este ejemplo se vería así:

$$f(25) < f(5)$$

Sin embargo, la función de densidad no nos dirá la probabilidad de ocurrencia de la variable sino cómo ellas se reparten. Para poder saber las probabilidades, que es lo que nos interesa, necesitamos la función de distribución de probabilidad.

Función de distribución de probabilidad: nos indica la probabilidad de que un evento suceda menos veces que algún número. Formalmente, vamos a definir la función de distribución de probabilidad (f.d.p.) como:

$$F(x) = \Pr(X \leq x)$$

Esta fórmula indica la probabilidad de que una variable aleatoria sea menor a un valor x.

Además, existe una relación entre la f.d.p. y la función de densidad, que es la siguiente:

$$\Pr(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$$

Es decir, la probabilidad de que una variable sea menor a un valor x es igual a la integral desde menos infinito hasta x de la función de densidad, en otras palabras, la f.d.p. se puede describir como el área bajo la curva de la función de densidad. Por ejemplo, si queremos calcular la probabilidad de que un alumno estudie menos de 30 minutos a la semana, tendríamos que calcular:

$$\Pr(X \leq 0{,}5) = F(0{,}5) = \int_{-\infty}^{0{,}5} f(t)\mathrm{d}t$$

Es importante notar que en las probabilidades continuas no podemos calcular un punto, o sea, $\Pr(X = x)$, ya que ese valor es cero. La intuición es que, por ejemplo, la probabilidad de que alguien estudie justo 30 minutos es despreciable ya que lo más probable es que si pudiésemos medir el tiempo de estudio de los alumnos, no encontraríamos 30 minutos exactos.

FUNCIONES DE DISTRIBUCIÓN CONTINUA (35 MINUTOS)

¿Cuáles son los tipos de distribución de probabilidad continuas? ¿Qué formas puede tomar $f(x)$? Así como teníamos funciones de distribución con aplicaciones bien precisas en el caso de distribuciones discretas, lo mismo sucede cuando tenemos distribuciones continuas. Las dos funciones de distribución continua más conocidas son la distribución exponencial y la distribución normal

Distribución Exponencial

La distribución exponencial nos sirve para medir tiempos o distancias entre dos sucesos. Tomando el ejemplo de la introducción, podemos utilizar la distribución exponencial para modelar cuánto tiempo tarda una persona en el banco en ser atendida. La forma de la **función de densidad exponencial** es:

$$f(x) = \lambda \cdot \exp(-\lambda \cdot x)$$

O sea, la función de x es lambda, por el exponencial elevado a menos lambda por x. La pregunta clave aquí es qué es λ. La respuesta corta es que lambda nos indica cuántos eventos suceden en una unidad de tiempo.

Algo fundamental para caracterizar las funciones es conocer su esperanza y varianza. La **esperanza y varianza exponencial** tiene las siguientes características:

$$\mathrm{E}(x) = \frac{1}{\lambda}$$

$$\mathrm{V}(x) = \frac{1}{\lambda^2}$$

Supongamos, por ejemplo, que una persona tarda en promedio 30 minutos en ser atendida en el banco. Entonces, en este caso lambda sería:

$$\lambda = \frac{1}{30} = 0{,}033$$

Es decir, lambda nos dice a cuantas personas atienden por minuto. Si lo generalizamos, lambda nos indica cuantos eventos suceden en una unidad de tiempo.

Sin embargo, recordemos que esta función solo nos dice la densidad de los valores. Para conocer la probabilidad, usamos la f.d.p:

$$F(x) = \int_{\infty}^{x} f(t)\mathrm{d}t = 1 - \exp(-\lambda x) = 1 - e^{-\lambda x}$$

Es decir, la f.d.p se puede expresar, genéricamente, como 1 menos el exponente de menos lambda por x. Lo anterior significa que no tendremos que calcular siempre la integral, ya que existe una fórmula general para la probabilidad, que es uno menos el exponencial elevado a menos lambda por x. Siguiendo con el ejemplo, esto significa que para saber cuál es la probabilidad de que me atiendan en un máximo de 20 minutos en el banco, tendría que calcular:

$$\Pr(X \leq 20) = 1 - \exp(-20 \cdot 0{,}033) = 0{,}48$$

Es decir, existe un 48% de probabilidades de que me atiendan en máximo 20 minutos.

Ejercicio Distribución Exponencial

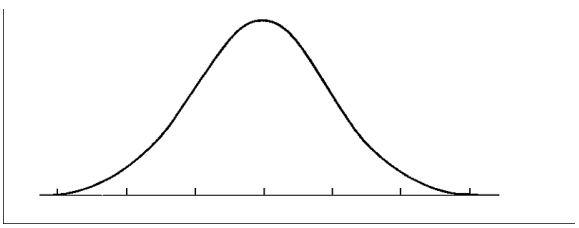
a. ¿Cuál es la probabilidad de que me atiendan en menos de 5 minutos?
b. ¿Cuál es la probabilidad de que me atiendan en más de 10 minutos?

Distribución Normal

La distribución normal es probablemente la distribución más famosa de todas. Su uso es diverso, pero en general, obedece a la siguiente regla: una variable aleatoria se distribuye de forma normal cuando sus valores tienden a concentrarse en torno a un número central. Por ejemplo, imaginemos una línea de producción de tuercas (a pesar de que las tuercas no son continuas, podemos definir la variable como “miles de tuercas” y tendríamos una variable bastante próxima a una continua). Dicha línea de producción debiese producir al día 30 mil tuercas. Sin embargo, puede que un día empiece a funcionar un minuto antes o uno después que el día anterior. Puede que un día el operario haga el trabajo un poco más rápido y otro día un poco más lento. Por lo tanto, al final de una semana, observamos que la producción fue: 30.001, 30.020, 29.998, 30.000 y 29.990.

Supongamos ahora, que seguimos guardando datos para un mes. Un día el trabajador, por error, puso la máquina a funcionar a su máxima potencia, por lo que la producción llegó a 32.000. Al día siguiente hizo lo mismo, por lo que produjeron 31.823. Sin embargo, al tercer día la máquina falló por tanta carga. Ese día, y el siguiente, tuvo que trabajar menos tiempo, por lo que la producción fue de 27.465 y 27.832.

Si replicamos el ejercicio para un semestre, o para un año, tendremos que, la mayor parte del tiempo, la máquina funcionará bien y estaremos cerca de las 30 mil tuercas diarias. Algunos días sucederán cosas excepcionales como las descritas y tendremos valores muy altos o muy bajos. Si lo graficamos, tendríamos una función de densidad como la que aparece en la figura. La parte más alta estaría sobre los 30.000.



Formalicemos lo anterior. La **función de densidad normal** se representa de la siguiente manera:

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

En palabras, es el exponente de menos x menos mu al cuadrado dividido por dos veces sigma cuadrado, dividido por la raíz de dos veces pi por sigma al cuadrado.

A su vez, **la esperanza y varianza normal** son:

$$E(x) = \mu$$

$$V(x) = \sigma^2$$

Es decir, el valor esperado es mu y la varianza es sigma al cuadrado.

En el ejemplo anterior sobre las tuercas definimos que $\mu = 30.000$. Para el ejemplo que sigue asumamos también que la varianza es de 90.000. Supongamos que queremos saber la probabilidad de que no logremos cumplir con alguna meta de trabajo, que diremos es de 30.050 tuercas. Al igual que con la distribución exponencial, en este caso necesitamos conocer la f.d.p para poder calcular probabilidades. Sin embargo, dado que la función normal es bastante compleja, los estadísticos han desarrollado formas más fáciles de hacer el cálculo.

Estandarización

¿Alguien sabe lo que es la estandarización? Una propiedad muy útil que tienen las variables normales, es que podemos sumarles y multiplicarles valores, y seguirán siendo una variable normal. ¿Por qué es una propiedad útil? Porque podemos hacer que cualquier variable normal tenga los mismos parámetros y en muchos libros hay tablas con todas las probabilidades que esta distribución normal puede tener. Este proceso se conoce como estandarización.

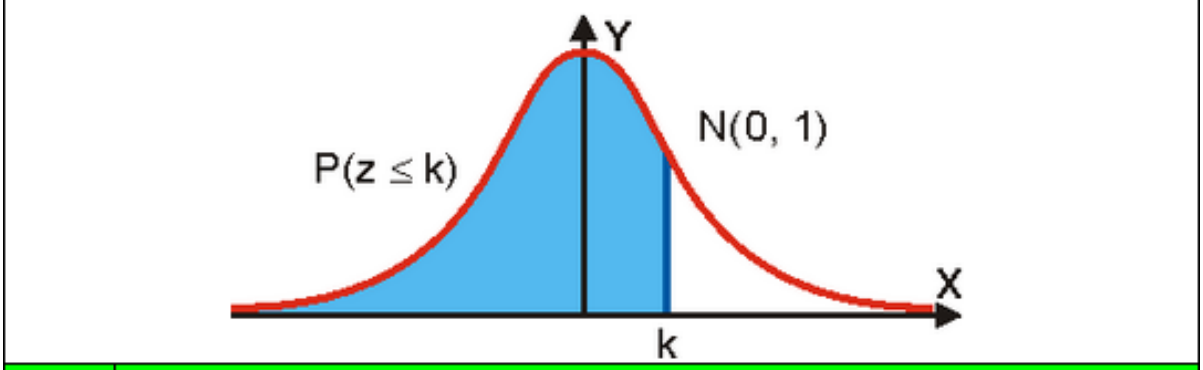
Formalmente, supongamos que tenemos una variable aleatoria X normal, con promedio μ y varianza σ^2. Entonces, podremos crear una variable Z con promedio 0 y varianza 1 si hacemos lo siguiente:

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

Es decir, Z es igual a x menos mu, dividido por sigma. Dicho de otra forma, Z es igual a x menos la esperanza de x, dividido por su desviación estándar. Esta variable sigue lo que se conoce como una distribución normal estándar. Su principal gracia está en que todas sus probabilidades ya han sido calculadas y existen tablas que las resumen. Estas tablas están disponibles en todos los libros de estadísticas

Si tenemos una variable normal estandarizada y queremos saber cuál es la probabilidad de que dicha variable sea menor a 2,54, tenemos que buscar en la primera columna de la tabla el valor 2,5. Luego, en la primera fila buscamos la centésima, en este caso 0,04. Si intersectamos la fila con la columna encontrada, el valor será 0,9946, que corresponde a la probabilidad de que la variable sea menor a 2,54.

ÁREAS BAJO LA DISTRIBUCIÓN DE PROBABILIDAD NORMAL ESTÁNDAR, N(0, 1)



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4,0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Ahora que sabemos esto, podemos continuar con la pregunta sobre las tuercas. Lo que queremos es buscar la probabilidad de que no alcancemos a cumplir la meta de 30.050 tuercas, siendo la varianza de 90.000. Para esto tenemos que calcular:

$$\Pr(X \leq 30.050)$$

Entonces, procedemos a estandarizar nuestra variable:

$$\Pr\left(\frac{X - 30000}{300} \leq \frac{30050 - 30000}{300}\right) = \Pr\left(Z \leq \frac{50}{300}\right) \approx \Pr(Z \leq 0{,}17)$$

Ahora, en vez de hacer cualquier cálculo, simplemente vamos a mirar la tabla y buscamos la probabilidad asociada a 0,17. Esta probabilidad es de 0,5675. Eso significa que existe un 56,75% de probabilidad de que no alcancemos la meta de trabajo.

Ejercicio Estandarización: ¿Cuál es la probabilidad de que produzcan más de 31.000 tuercas?

TEOREMA CENTRAL DEL LÍMITE (20 minutos)

Arriba discutimos que estandarizar una variable normal muchas veces simplifica un problema. En vez de tener que realizar una integral compleja, simplemente hacemos una resta y una división. El teorema central del límite nos permite extender esta simplificación a diversos escenarios.

Cuando explicamos la distribución normal, calculamos la probabilidad asociada a no superar una cuota de 30.050 tuercas diarias que era de 0.57 aproximadamente. Supongamos ahora que, nosotros, como gerentes de operaciones, queremos saber cuál es la probabilidad de que dicha cuota no se alcance un máximo 6 veces al mes. ¿Cómo haríamos este cálculo?

Lo primero, es entender la variable que estamos usando. Ya no nos interesa la cantidad de tuercas. La variable de interés ahora es si se alcanza o no la meta. Es decir, es una variable dicotómica, que puede ser sí o no y sigue una distribución Bernoulli. Matemáticamente, la escribiríamos de la siguiente forma:

$$x = \begin{cases} 1 & \text{no se cumple la meta} \\ 0 & \text{se cumple la meta} \end{cases}$$

Si sumamos esa variable para todos los días del mes, vemos que solo contarían los días en los que no se cumple la meta, ya que cuando se cumple toma valor 0. Por ejemplo, si la meta no se cumple los 3 primeros días del mes, tendríamos que:

$$\sum_{i=1}^{30} x_i = 1 + 1 + 1 + 0 + 0 + 0 + \cdots + 0 = 3$$

Que es la sumatoria de x1, x2, hasta x30. Cada uno de estas x solo puede ser 1 o 0, como ya lo definimos más arriba. Por lo tanto, representará el número de veces que no se cumple la meta. Nosotros queremos saber la probabilidad de que esa suma sea, cuando máximo, 6. Por lo tanto, matemáticamente escribiremos:

$$\Pr\left(\sum_{i=1}^{30} x_i \leq 6\right) = ?$$

O sea, la probabilidad que la cantidad de veces que no alcance la cuota sea menor o igual a 6. Este cálculo resultaría muy complicado de hacer. Tendríamos que transformar esa sumatoria en una distribución binomial, y resolver un cálculo bastante extenso. Aquí es justamente donde el Teorema Central del Límite resulta de ayuda.

Ayudados por el Teorema Central del Límite, tomaremos otro camino mucho más sencillo para realizar el cálculo. El Teorema nos dice que cuando tenemos una muestra grande, la suma y el promedio de las variables se comportan como una distribución normal.

Apliquemos el teorema. Lo primero que haremos, será dividir dentro de la probabilidad por 30, para trabajar con el promedio:

$$\Pr\left(\frac{1}{30}\sum_{i=1}^{30} x_i \leq \frac{6}{30}\right) = \Pr(\bar{x} \leq 0{,}2)$$

Por lo tanto, tenemos que la sumatoria de los 30 x dividido por treinta debe ser menor o igual a 6 dividido por treinta. Es decir, la probabilidad que el promedio de x sea menor a 0.2. Lo bueno del teorema, es que podemos suponer que $\bar{x}$ es una variable normal y las variables normales las podemos estandarizar. Lo único que tenemos que saber, es cuál es la esperanza y la varianza de $\bar{x}$, que son:

$$E(\bar{x}) = E(x)$$

$$V(\bar{x}) = \frac{V(x)}{n}$$

Sabemos que el valor esperado de que no se cumpla la meta de 30.050 tuercas diarias es de 0.57. Además, la varianza se puede calcular como 0,57(1-0,57)=0.2451, ya que es una variable Bernoulli. Por lo tanto:

$$E(\bar{x}) = 0.57$$

$$V(\bar{x}) = \frac{0.2451}{30} = 0.008$$

Finalmente, con esos dos datos ya podemos estandarizar y resolver nuestro problema:

$$\Pr(\bar{x} \leq 0{,}2) = \Pr\left(\frac{\bar{x} - 0{,}57}{\sqrt{0{,}008}} \leq \frac{0{,}2 - 0{,}57}{\sqrt{0{,}008}}\right) = \Pr(Z \leq -4.14)$$

Este valor podríamos buscarlo directamente en una tabla y conocer la respuesta. Sin embargo, muchas tablas no muestran los valores positivos. La razón es que la distribución normal estándar es simétrica con respecto a cero, por lo tanto vamos a tener que:

$$\Pr(Z \leq -4{,}14) = \Pr(Z \geq 4{,}14)$$

Es decir, la probabilidad de que una variable sea menor a -4,14 es igual a la probabilidad de que un valor sea mayor a 4,14. También hay que considerar que las tablas muestran el valor acumulado a la izquierda, es decir $\Pr(Z \leq z)$. Entonces tenemos que hacer un último cambio:

$$\Pr(Z \geq 4{,}14) = 1 - \Pr(Z < 4{,}14)$$

Esto lo podemos hacer porque la probabilidad de que una variable sea menor a 4,14, más la probabilidad de que la misma variable sea mayor a 4,14 tiene que sumar uno. Esto viene dado por las propiedades de una función de distribución.

Ahora sí, podemos buscar el valor en una tabla:

$$\Pr(Z \geq 4{,}14) \approx 1 - 1 = 0$$

Es decir, la probabilidad de que fallen máximo 6 veces es casi cero. Esto significa que probablemente fallarán más veces al mes. Esta es una información muy útil para el gerente ya que podría, por ejemplo, decidir en base a esto colocar otra máquina.

Hemos visto un ejemplo de Teorema Central del Límite con una variable Bernoulli. Sin embargo, este teorema se puede usar con todas las funciones que hemos visto: Bernoulli, Poisson y Exponencial.