Nivelatorio: Estadística para la Ciencia de datos

Presentación 2: Examinar las principales distribuciones de probabilidad que se emplean para describir datos

Diego Antonio Bohórquez Ordóñez
dabohorquez@Icesi.edu.co

Maestría en Ciencia de Datos

Agenda

1. ¿Qué es la probabilidad?
2. Algunas reglas de la probabilidad
3. Teorema de Bayes
4. ¿Qué son las distribuciones de probabilidad y por qué son útiles?
5. Distribuciones de variables discretas (Binomial, Uniforme, Poisson y Binomial Negativa)
6. Distribuciones de variables continuas (Normal)
7. Práctica en Python

¿Qué es la probabilidad?

¿Qué es la probabilidad?

¿Qué es: “Valor entre 0 y 1 que describe la posibilidad relativa de que ocurra un evento”. (Lind, Marchal & Wathen, 2015)*

Ejemplo: Probabilidad del resultado en los 90 minutos para la final del Mundial Qatar 2022

33%

VS

31%

36%

*Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). Estadística aplicada a los negocios y la economía. McGraw-Hill, Ed 16.

¿Qué es la probabilidad?

Experimento: Es un proceso que induce a que ocurra un resultado.
Ejemplo: Lanzamiento de un dado.

Resultado: Es la consecuencia de un experimento
Ejemplo: Se observa 1, 2, 3, 4, 5 o 6.

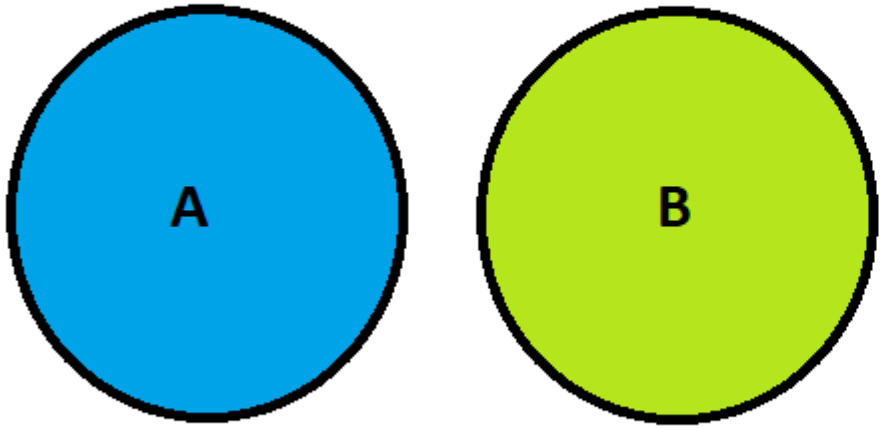
Evento: Conjunto de uno o más resultados de un experimento
Ejemplo: Se observa un número par, un número mayor a 2, etc.

¿Cómo se ajusta el ejemplo de la final del mundial a estas definiciones?

Algunas reglas de la probabilidad

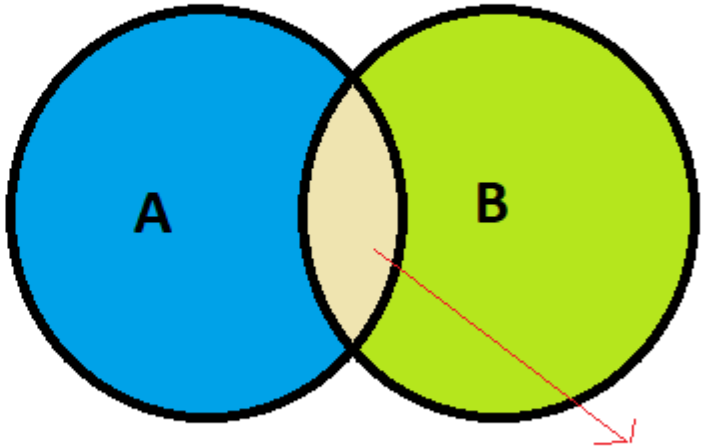
Regla adición:

Mutuamente exclyentes:



$$P(A \text{ o } B) = P(A) + P(B)$$

No son mutuamente excluyentes:



$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$$

Ejercicios:

- En un partido de fútbol las casas de apuestas pronostican que la probabilidad de que el equipo A gane es de 48%, que el equipo B gane es 22% y que empaten es de 30%. ¿Cuál es la probabilidad de que no haya empate?

- En el mismo partido, A tiene 69% de probabilidad de anotar al menos 1 gol, B tiene 50% de probabilidad de anotar al menos 1 gol y la probabilidad de que ambos anoten es 35%. ¿Cuál es la probabilidad de que A o B anoten al menos un gol?

Regla multiplicación:

Definiciones iniciales:

Probabilidad conjunta: Probabilidad de que dos o más eventos sucedan al mismo tiempo.

Probabilidad condicional: Probabilidad de que un evento ocurra dado que otro ha acontecido.

Algunas reglas de la probabilidad

Regla multiplicación:

- Eventos independientes: P(A y B) = P(A)*P(B)

Ejemplo: ¿Cuál es la probabilidad de que llueva en Cali y en Londres mañana?
P(llueva en Cali y llueva en Londres) = P(llueva en Cali)xP(llueva en Londres)

- Eventos no independientes: P(A y B) = P(A)*P(B|A)

Ejemplo: ¿Cuál es la probabilidad de que un adulto mayor escuche reggaetón?
P(adulto mayor y escucha reggaetón) = P(adulto mayor) x P(escucha reggaetón|adulto mayor)

Tablas de contingencia:

- ¿Qué es?: es una tabla utilizada para clasificar las observaciones según dos o más características.

Películas vistas (A)	género (B)		Total
	H	M	
0	20	40	**60**
1	40	30	**70**
>= 2	10	10	**20**
Total	**70**	**80**	**150**

Calcular: P(no ver películas), P(ser hombre), P(ser hombre dado que no ve películas), P(no ve películas y es hombre)

Calcular: P(no ver películas)

Películas vistas (A)	Total
0	**60**
1	**70**
>= 2	**20**
Total	**150**

Calcular: P(ser hombre)

Películas vistas (A)	género (B)		Total
	H	M	
Total	**70**	**80**	**150**

P(ser hombre dado que no ve películas)

Películas vistas (A)	género (B)		Total
	H	M	
0	20	40	**60**

P(ser hombre dado que no ve películas)

Películas vistas (A)	género (B)		Total
	H	M	
>= 2	10	10	**20**

P(No ve películas dado que es mujer)

Películas vistas (A)	género (B)		Total
	H	M	
0	20	40	**60**
1	40	30	**70**
>= 2	10	10	**20**
Total	**70**	**80**	**150**

P(no ve películas y es hombre)

Películas vistas (A)	género (B)		Total
	H	M	
0	20	40	**60**
1	40	30	**70**
>= 2	10	10	**20**
Total	**70**	**80**	**150**

Teorema de Bayes

Teorema de bayes

- El teorema de Bayes es utilizado para calcular la probabilidad de un suceso, teniendo información de antemano sobre ese suceso*.

- En otras palabras, permite calcular la probabilidad de un suceso, conociendo información que condiciona su probabilidad.

- ¿Cómo así?**:

Existe una situación en la que conocemos la probabilidad de ocurrencia de un suceso (ser diabético). PROBABILIDAD A PRIORI

Esta probabilidad puede verse modificada por la ocurrencia de otro suceso (digamos ser obeso), ya que la probabilidad de ser obeso difiere entre quienes son diabéticos y no diabéticos.

Conociendo que alguien es obeso, el teorema de Bayes nos indica como condiciona esta información la probabilidad de ser diabético. PROBABILIDAD A POSTERIORI

*Tomado de: https://economipedia.com/definiciones/teorema-de-bayes.html

** Tomado de: https://www.ugr.es/~jsalinas/bayes.htm

Teorema de bayes

- Formulación de la regla:

$$P(A|B) = \frac{P(A\ y\ B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B)}$$

- La importancia de este teorema es que permite relacionar P(A|B) con P(B|A).

Teorema de bayes

- Ejemplo:

Una empresa tiene una fábrica en Estados Unidos que dispone de tres máquinas A, B y C, que producen envases para botellas de agua. Se sabe que la máquina A produce un 40% de la cantidad total, la máquina B un 30% , y la máquina C un 30%. También se sabe que cada máquina produce envases defectuosos. De tal manera que la máquina A produce un 2% de envases defectuosos sobre el total de su producción, la máquina B un 3%, y la máquina C un 5%.

¿Cuál es la probabilidad de que un envase sea defectuoso si se fabrica en EEUU? ¿Cuál es la probabilidad de que haya sido fabricado por la máquina A?

Teorema de bayes

- Información que conocemos:

P(A)=40%

P(B)=30%

P(C)=30%

P(D|A)=2%

P(D|B)=3%

P(D|C)=5%

- ¿Qué están preguntando?

P(D)= ¿?

P(A|D)= ¿?

Primera pregunta:

P(D)= ¿?

P(D)=P(A)*P(D|A) +P(B)*P(D|B)+P(C)*P(D|C)

P(D)=3,2%

Segunda pregunta:

P(A|D)= ¿?

$$P(A|D) = \frac{P(D|A) * P(A)}{P(D)} = \frac{2\% * 40\%}{3{,}2\%}$$

P(A|D)=25%

Teorema de bayes

- Usos en el machine learning: Clasificadores de Naive Bayes, LDA y QDA

- Ejemplo clásico: Correo spam

$$P(S|x1, x2, x3, \dots, xn) = \frac{P(S\ y\ x1, x2, x3, \dots, xn)}{P(x1, x2, x3, \dots, xn)} = \frac{P(x1, x2, x3, \dots, xn|S) * P(S)}{P(x1, x2, x3, \dots, xn)}$$

¿Qué son las distribuciones de probabilidad y por qué son útiles?

¿Qué son las distribuciones de probabilidad y por qué son útiles?

¿Qué es: "Es un listado que muestra los posibles resultados de un experimento y la probabilidad de que cada uno ocurra". (Lind, Marchal & Wathen, 2015)*

Ejemplo: Tirar un dado.

Resultado posible	Probabilidad
1	16,67%
2	16,67%
3	16,67%
4	16,67%
5	16,67%
6	16,67%

*Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). Estadística aplicada a los negocios y la economía. McGraw-Hill, Ed 16.

¿Qué son las distribuciones de probabilidad y por qué son útiles?

¿Por qué es útil estudiar distribuciones de probabilidad?:

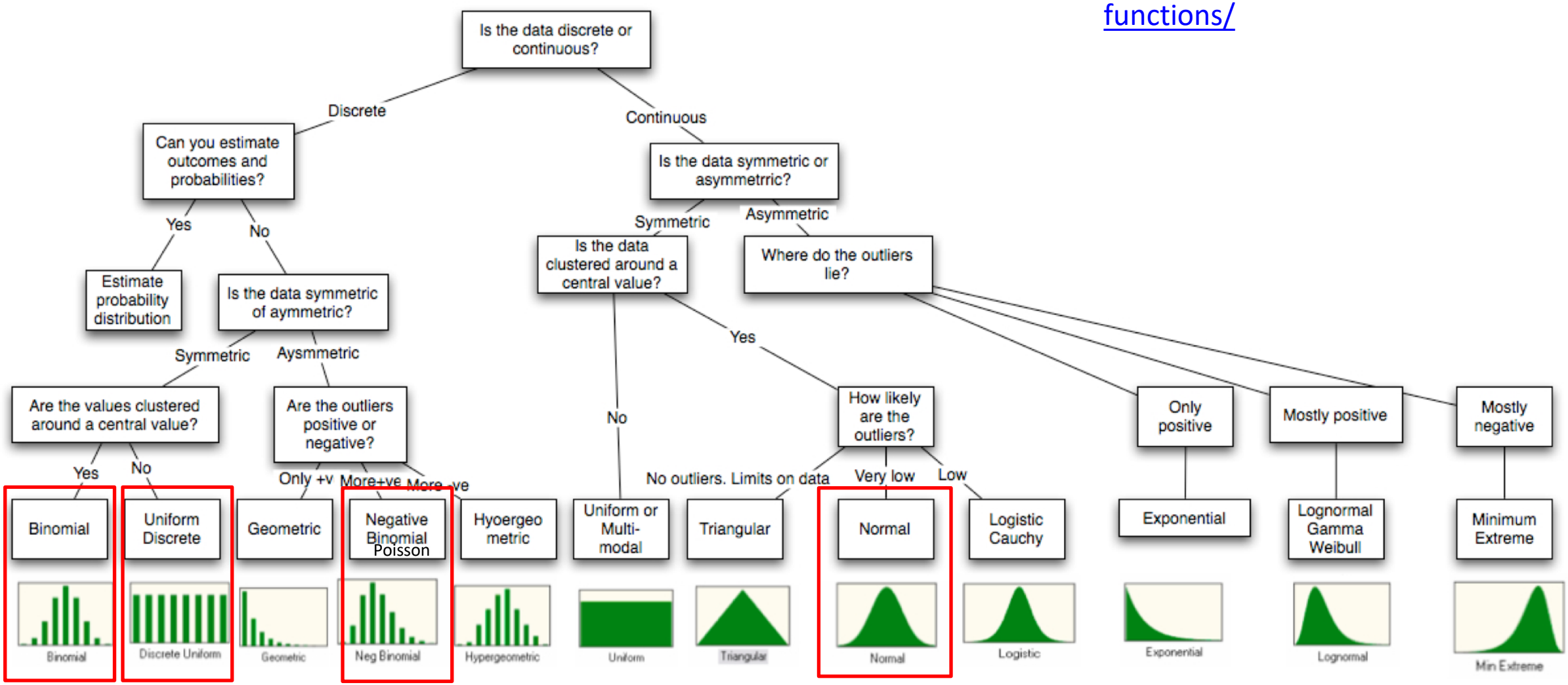
- Diferentes distribuciones de probabilidad han sido identificadas con regularidad en la naturaleza (Normal, Poisson, Binomial, etc.), tal que, luego de haber sido estudiadas, fue posible establecer un conjunto de propiedades y reglas empíricas.

- Dichas propiedades y reglas empíricas son de gran aplicabilidad para caracterizar distribuciones y en el establecimiento de escenarios sobre el comportamiento futuro.

¿Qué son las distribuciones de probabilidad y por qué son útiles?

Figure 6A.15: Distributional Choices

https://www.abstractclasses.in/2017/10/08/table-probability-distribution-functions/



¿Qué son las distribuciones de probabilidad y por qué son útiles?

- De acuerdo con las características que se observan gráficamente es posible intuir aproximadamente qué distribución podría ajustarse a nuestros datos.

- Sin embargo, es necesario llevar a cabo pruebas estadísticas formales que veremos en siguientes clases.

- Por ahora lo importante es saber como “lucen” gráficamente las distribuciones de probabilidades más relevantes y conocer sus propiedades y reglas empíricas.

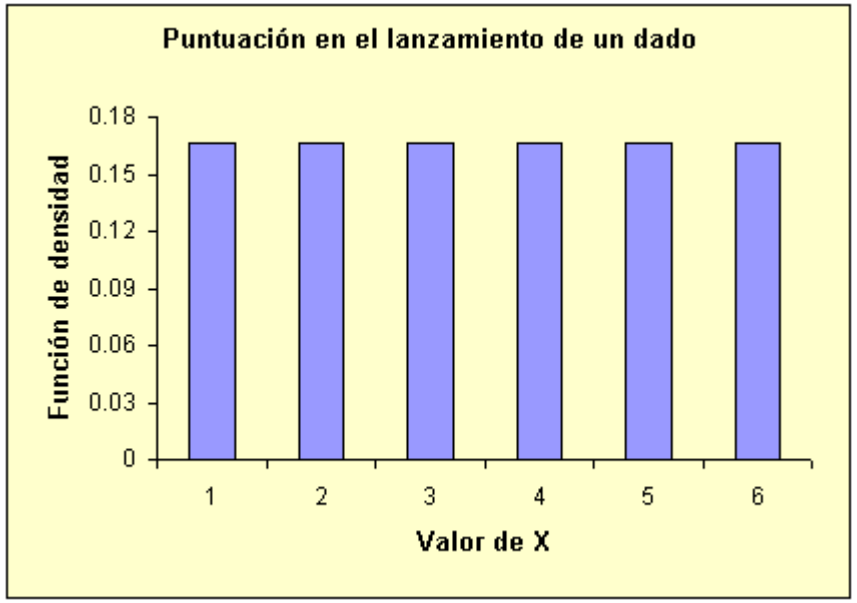
Distribuciones de variables discretas

Distribución Uniforme:

- Cada uno de los resultados tiene una misma probabilidad de ocurrencia.
- Nuestro ejemplo anterior del dado sigue una distribución uniforme.
- También está el caso de la moneda (50%-50%)

Su media será: $\mu = \frac{n+1}{2}$ y su varianza: $\sigma^2 = \frac{n^2-1}{12}$

Resultado posible	Probabilidad
1	16,67%
2	16,67%
3	16,67%
4	16,67%
5	16,67%
6	16,67%



Su media será: $\mu = n\pi$ y su varianza: $\sigma^2 = n\pi(1-\pi)$

Distribución Binomial:

- El resultado de cada ensayo de un experimento se clasifica en una de dos categorías mutuamente excluyentes: éxito o fracaso.
- La variable aleatoria permite contar el número de éxitos en una cantidad fija de ensayos.
- La probabilidad de éxito y fracaso es la misma en cada ensayo.
- Los ensayos son independientes, lo cual significa que el resultado de un ensayo no influye en el resultado del otro.
- El interés está en la ocurrencia de un evento.

Distribución Binomial:

- Ejemplo: El porcentaje de estudiantes mujeres en Icesi es 55%. ¿Cuál es la probabilidad de que 8 estudiantes en un grupo de 15 sean mujeres?

- Debemos definir a qué le llamaremos “éxito”: Mujer.

- Aplicamos la fórmula de la distribución binomial (R nos ayuda en el cálculo):

$$P(x) = \binom{n}{x}\pi^{x}(1-\pi)^{n-x}$$

Distribución Binomial:

- Ejemplo: El porcentaje de estudiantes mujeres en Icesi es 55%. ¿Cuál es la probabilidad de que 8 estudiantes en un grupo de 15 sean mujeres?

- $\pi = 55\%$
- $x = 8$
- $n = 15$

$$P(x = 8) = 20{,}1\%$$

Distribución Binomial:

- Ejemplo: ¿Cuál es la probabilidad de que sean 5 mujeres?

- $\pi = 55\%$
- $x = 5$
- $n = 15$

$$P(x = 5) = 5{,}1\%$$

Distribución Binomial:

- Ejemplo: ¿Cuál es la probabilidad de que todas sean mujeres?

- $\pi = 55\%$
- $x = 15$
- $n = 15$

$$P(x = 15) = 0{,}01\%$$

Distribución Binomial:

- Ejemplo: Tabulemos todos los casos
- ¿Por qué 8 es lo más probable?
- La probabilidad de ser mujer en Icesi es de 55%, por tanto, **en promedio**, en una muestra de 15 estudiantes uno esperaría:

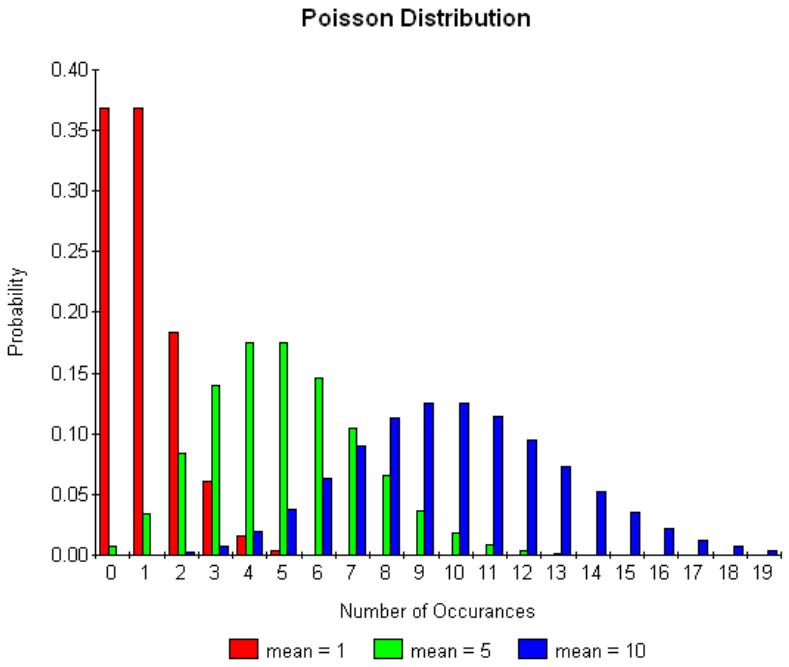
$$15 * 55\% = 8{,}25\ mujeres$$

Número de Mujeres	Probabilidad
0	0,0006%
5	5,14%
6	10,48%
7	16,47%
8	20,13%
9	19,14%
10	14,04%
11	7,79%
15	0,01%

Distribución de Poisson:

Su media será: $\mu = \lambda$ y su varianza: $\sigma^2 = \lambda$

- Describe el número de veces que se presenta un evento durante un intervalo.
- La media es igual que la varianza.
- Cada intervalo es independiente.
- El sesgo positivo se reduce a medida que la media aumenta.



Distribución de Poisson:

- La representación de que una variable sigue dicha distribución es: $X \sim Poisson(\lambda)$

- Matemáticamente, la distribución de Poisson es:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Donde λ es la media de la distribución.

- ¿Cómo se calcula la media?: Como vimos en la clase anterior, es el promedio.

Ejemplo: En una oficina se recibieron 35 llamadas en 5 días. Para modelar el número de llamadas por día con la distribución de Poisson, ¿cuál es la media de llamadas por día?

$$\lambda = \frac{35\ llamadas}{5\ días} = 7\ {llamadas}/{día}$$

¿Cómo usar la distribución de Poisson?:

- Sigamos con el ejemplo: La media de las llamadas por día es 7. ¿Cuál es la probabilidad de que no reciba llamadas (x=0)?

$$P(x = 0) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{7^0 e^{-7}}{0!} = 0{,}091\%$$

¿Cuál es la probabilidad de que reciba menos de 6 llamadas (x<6)?

$$P(x \leq 5) = P(x = 0)+P(x = 1)+P(x = 2)+P(x = 3) + P(x = 4) + P(x = 5) = 30{,}1\%$$

¿Cuál es la probabilidad de que reciba más de 10 llamadas (x>10)?

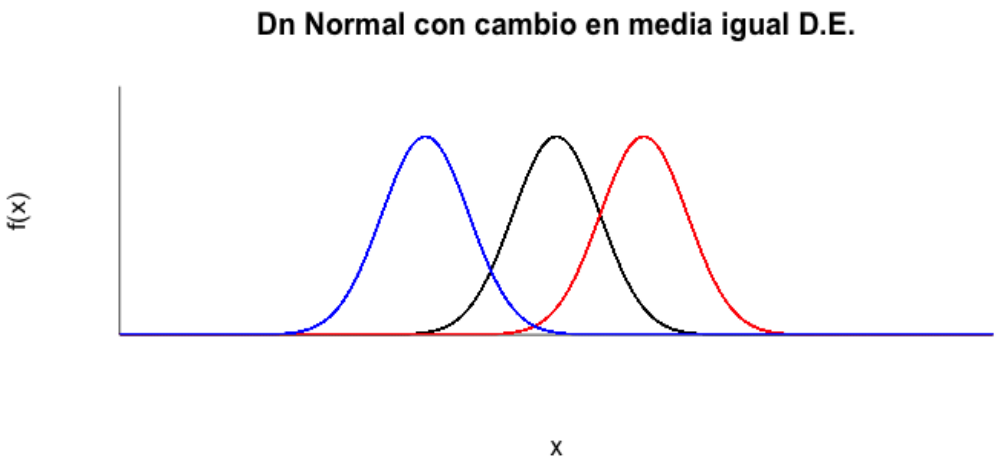
$$P(x > 10) = 1 - [P(x = 0) + P(x = 1) + \cdots + P(x = 9) + P(x = 10)] = 9{,}85\%$$

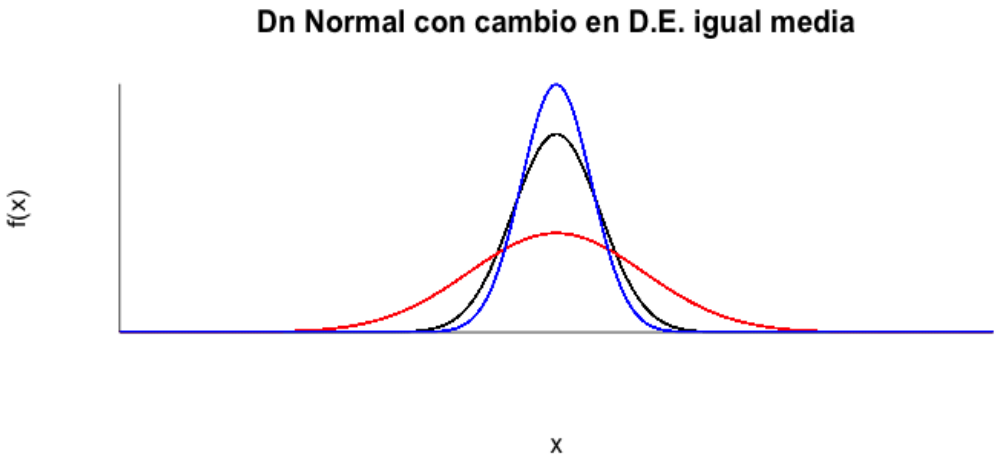
¿Cuál es la probabilidad de que reciba entre 6 y 10 llamadas (5<x<11)?

Distribuciones de variables continuas

Distribución Normal:

- Es una de las más usadas en la estadística, debido a que numerosos fenómenos siguen dicho "comportamiento" o se aproximan bien a este, y por la utilidad de sus propiedades al ser insumo para un gran número de pruebas estadísticas.
- Tiene forma de campana y es simétrica.
- La media y la desviación estándar son las que caracterizan la distribución.





Distribución Normal:

- La representación de que una variable sigue dicha distribución es: $X \sim N(\mu, \sigma^2)$

- Matemáticamente, la distribución Normal es:

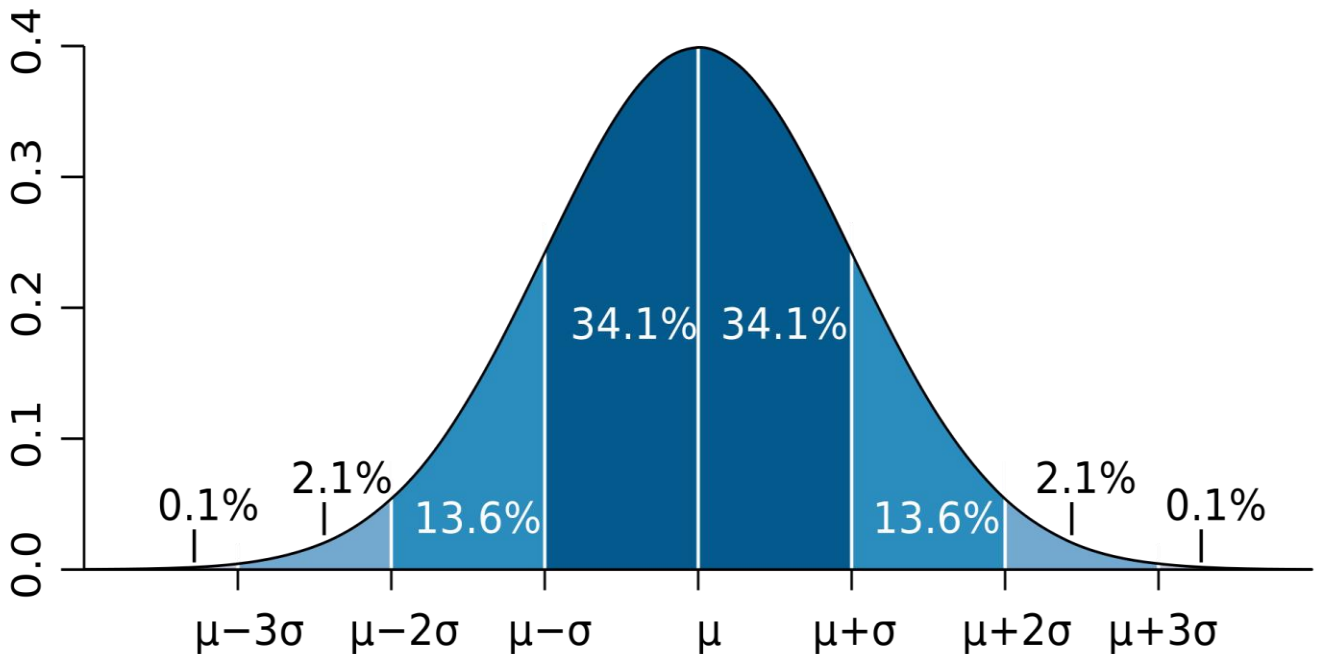
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-[\frac{(x-\mu)^2}{2\sigma^2}]}$$

- Así como hicimos con Poisson, podemos calcular la probabilidad de que x tome un valor, no muy usado en variables continuas, o la probabilidad de que x esté en un intervalo de valores, lo que sí tiene más sentido para este tipo de variables. (R nos ayuda con eso)

Distribución Normal:

- Regla empírica:

1. Aproximadamente el 68,2% de las observaciones está entre $\mu \pm \sigma$
2. Aproximadamente el 95,4% de las observaciones está entre $\mu \pm 2\sigma$
3. Aproximadamente el 99,7% de las observaciones está entre $\mu \pm 3\sigma$



Ejemplo: Se toma el tiempo de espera de cada cliente en la sucursal más concurrida. Se concluye que esta variable sigue una distribución normal, con una media de 20 minutos y una desviación estándar de 2 minutos. Use la regla empírica, ¿qué puede afirmar sobre el tiempo de espera?

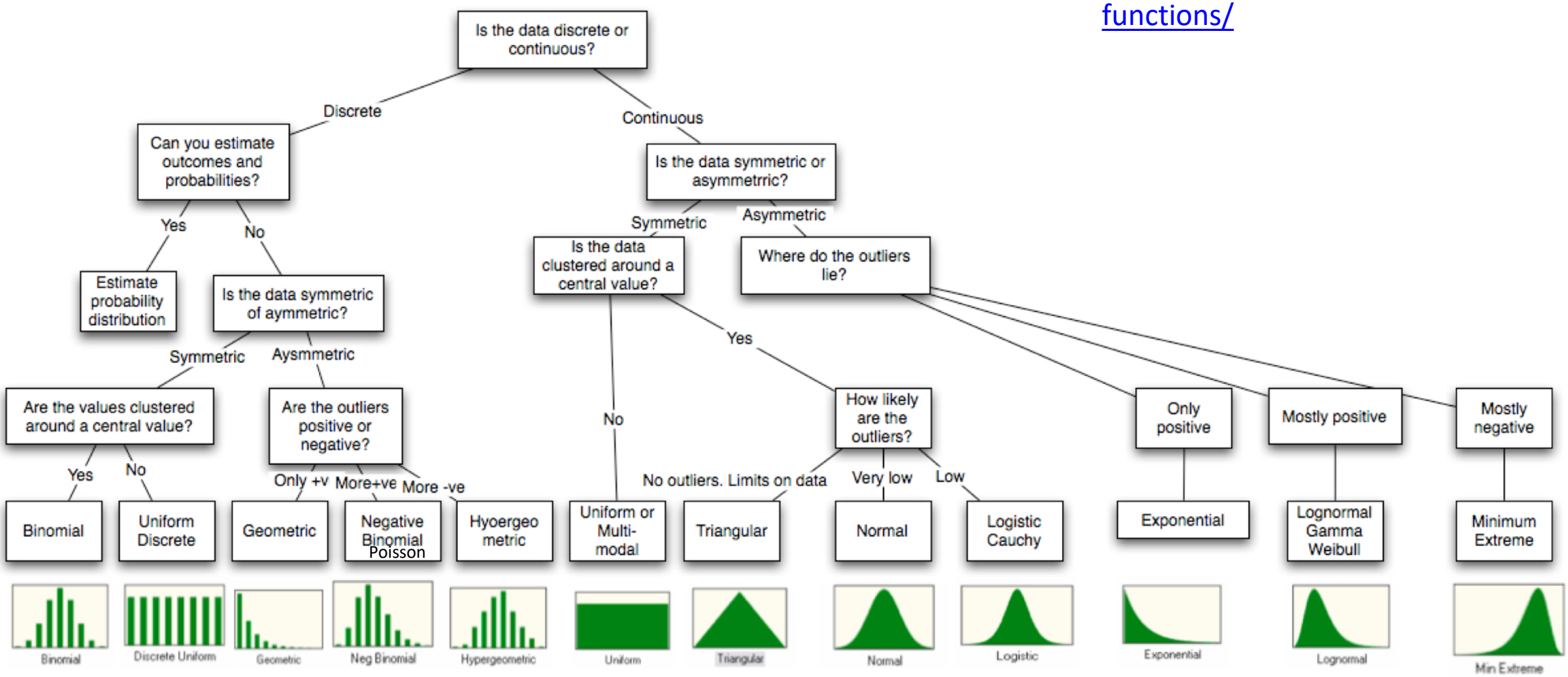
- Regla empírica:

1. Aproximadamente el 68,2% de las observaciones está entre _____.
2. Aproximadamente el 95,4% de las observaciones está entre _____.
3. Aproximadamente el 99,7% de las observaciones está entre _____.

En resumen…

Figure 6A.15: Distributional Choices

https://www.abstractclasses.in/2017/10/08/table-probability-distribution-functions/



1. Se calcula que 0.5 % de quienes se comunican al departamento de servicio al cliente, escuchará un tono de línea ocupada. ¿Cuál es la probabilidad de que de las 1200 personas que se comunicaron hoy, por lo menos 5 hayan escuchado un tono de línea ocupada? ¿Cuál es la probabilidad de que más de 10 personas hayan escuchado el tono de línea ocupada? Construir una tabla que contenga la probabilidad para x=0,1,2,3,4,5,6,7,8,9,10. Construir una gráfica de barras.

2. Se ha encontrado que, en promedio, en 5 días al mes nos enfrentamos a problemas de liquidez. ¿Cuál es la probabilidad de que en el próximo mes ocurran problemas de liquidez en 12 días? ¿y entre 4 y 9 días? Construir una tabla que contenga la probabilidad para x=0,1,2,3,4,5,6,7,8,9,10. Construir una gráfica de barras.

3. En los resultados del último Saber 11 la media del puntaje global fue 260, con una desviación estándar de 45. Teniendo en cuenta que el Saber 11 es una prueba estandarizada, ¿cuál es la probabilidad de que un estudiante haya obtenido 200 o menos? ¿a partir de qué puntaje se encuentra el 10% de estudiantes con mayor puntaje? ¿cuáles son los valores de la regla empírica?

4. El equipo de fútbol de la ciudad juega el 70% de sus partidos de noche y el 30% de día. Cuando juega de noche, gana el 50% de las veces, mientras que si juega de día gana el 90% de las veces. Hoy me levanté y vi que el equipo ayer ganó. ¿Cuál es la probabilidad de que el partido se haya jugado de noche?

5. En los últimos 20 partidos el América ha anotado en promedio 1.2 goles de local, mientras que el Cali ha anotado en promedio 0.9 goles de visitante, en el mismo número de partidos. Suponiendo independencia en los goles que anotan ambos equipos, ¿qué probabilidad tiene América de no anotar goles? ¿ y el Cali? ¿Qué probabilidad tiene el América de anotar 1 o 2 goles? ¿ y el Cali? Construir una gráfica de barras para cada equipo desde 0 a 5 goles. ¿Qué probabilidad hay de que el partido termine 0-0, 1-0, 1-1, 2-1, 1-2?