UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

Curso Métodos y Modelos

Profesora:
Karen Ballesteros-González PhD.

Modelos Basados en Datos

a. **Pre-procesamiento de datos (limpieza, transformación, detección de valores atípicos)**

b. **Análisis Exploratorio de Datos**
   i. Métodos gráficos
   ii. Estimadores Muestrales

c. **Modelos de Probabilidad**
   i. Distribuciones de variables Discretas
   ii. Distribuciones de variables Continuas
   iii. Verificación de ajuste de modelos de probabilidad
   iv. Gráficas Q-Q plots
   v. Pruebas de hipótesis

Objetivo:
Introducir a los estudiantes en las etapas preliminares del análisis de datos, incluyendo el **preprocesamiento** y el **análisis exploratorio**, como base fundamental para construir modelos predictivos o explicativos

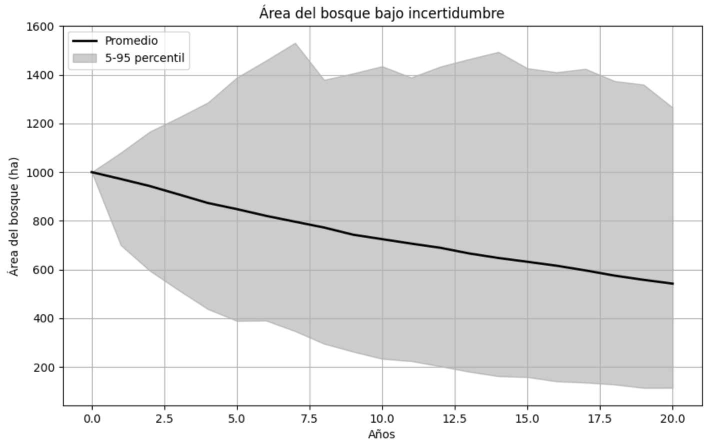
Modelos Estocásticos – **Método de Monte Carlo**

Ejercicio: Modelo Estocástico de Reforestación y Captura de Carbono

Este ejercicio simula el crecimiento del área de un bosque reforestado durante 20 años, considerando la posibilidad de eventos aleatorios que afectan dicho crecimiento.

Descripción del modelo
Cada año, el área del bosque puede crecer o reducirse dependiendo de un evento aleatorio:

- **Incendio** (probabilidad = 0.2): reduce el área en un 30%.
- **Plaga** (probabilidad = 0.1): reduce el área en un 15%.
- **Protección** (probabilidad = 0.3): aumenta la tasa de crecimiento a 8%.
- **Sin evento** (probabilidad = 0.4): el crecimiento se mantiene en 5%.



La ecuación básica de crecimiento de bosque es:

$$A_t = A_{t-1} \times (1 + r_t)$$

Donde:
A_t: área del bosque en el año (t)
r_t: tasa de crecimiento o decrecimiento, según el evento ocurrido en el año (t)

Ejercicio Monte Carlo – Sostenibilidad del Agua en una Cuenca

Objetivo
Simular cómo evoluciona el **almacenamiento de agua** de una cuenca durante 20 años, considerando incertidumbre en la precipitación, demanda y políticas de conservación.

Contexto del problema
Una cuenca tiene una capacidad promedio de recarga hídrica anual estimada a partir de la precipitación y la escorrentía. Sin embargo, la precipitación es variable de año a año. Además, el consumo de agua también varía según el crecimiento poblacional y hábitos de consumo.

Balance del Hídrico está dado por:

$$\Delta S_t = O_t - D_t$$

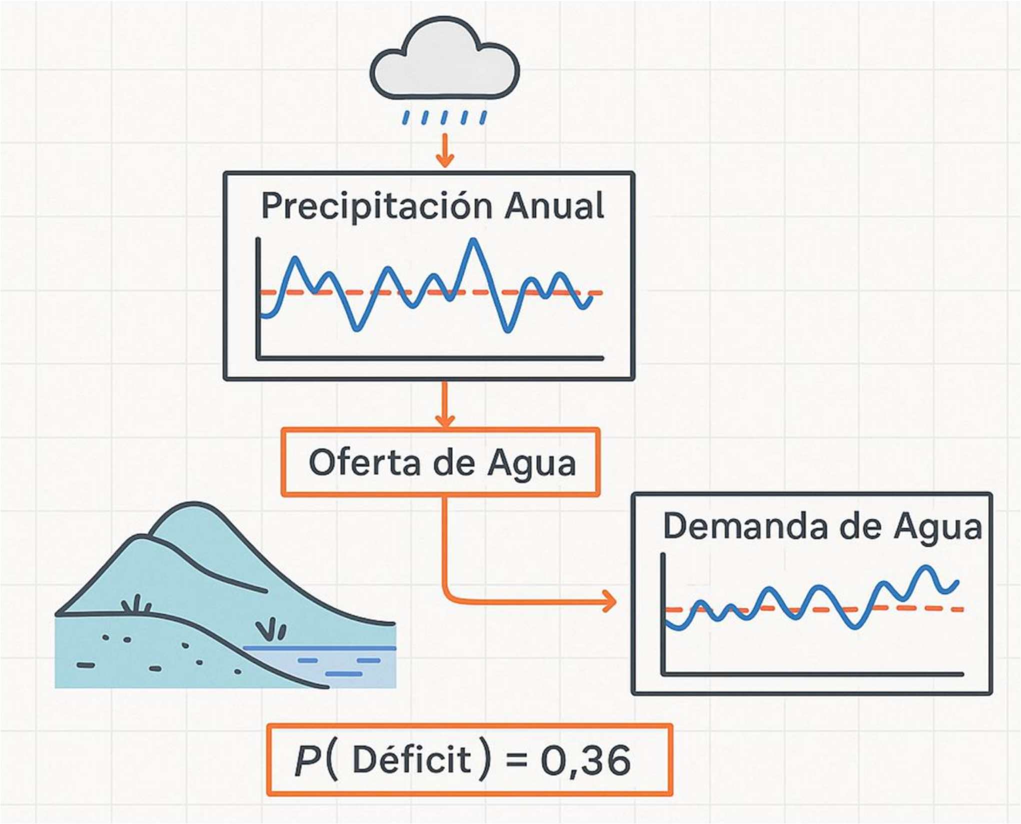
Capacidad de almacenamiento:

$S_{t+1} = \max(S_t + \Delta S_t, 0)$ - *Se usa para evitar que se retornen valores negativos.*

Modelos Estocásticos – **Método de Monte Carlo**

Ejercicio Monte Carlo – Sostenibilidad del Agua en una Cuenca

- **Precipitación anual** (P): Distribución normal con media 1200 mm y desviación estándar 200 mm.
- **Eficiencia de escorrentía** (e): 0.3 (30% de la precipitación se convierte en agua utilizable).
- **Demanda poblacional anual de agua** (D): Distribución normal con media 300 millones de m^3 y desviación estándar 50 millones de m^3.
- **Superficie de la cuenca** (A): 500 km^2.



Oferta de agua:

$$O = P \cdot e \cdot A.$$

$$para\ que\ Oferta\ este\ en\ m^3 (\frac{10^6}{1000})$$

Evaluación del balance:

$$Déficit\ si\colon\ D > O$$

Capacidad de almacenamiento:

$\boldsymbol{S_{t+1} = max(S_t + \Delta S_t, 0)}$ **- *Se usa para evitar que se retornen valores negativos.***

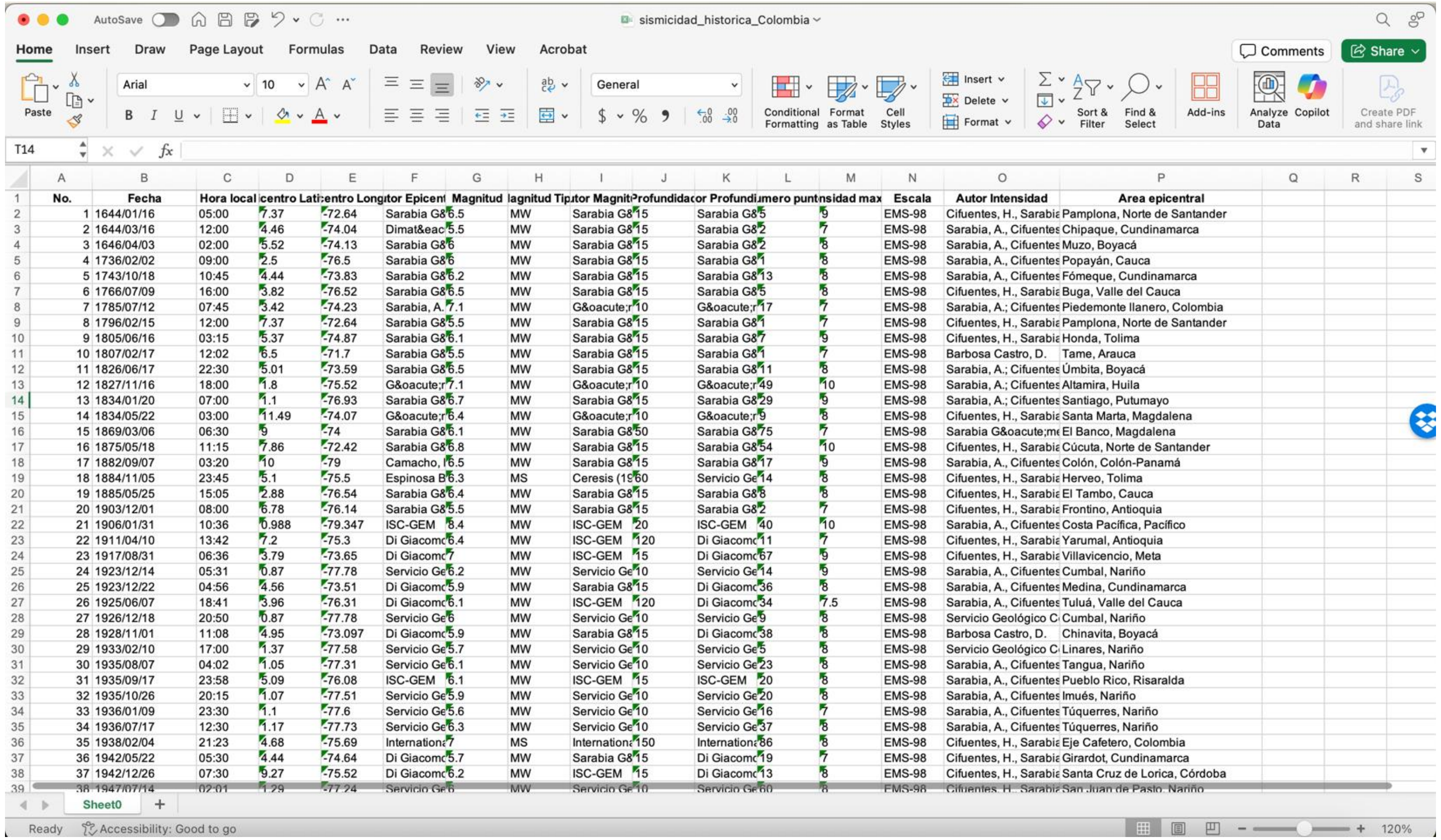
Ejemplo:

- Comienza el año con 100 millones de m^3 en el embalse.
- Hay un déficit este año de 150 millones de m^3

Entonces:

- $S_{t+1} = \max(100 - 150, 0) = \max(-50, 0) = 0$
- **Resultado realista**: el sistema se queda **sin agua**, pero no entra en negativo.

Ejercicio: Cadenas de Markov y Probabilidad de Sismos



No.	Fecha	Hora local	centro Lati	centro Long	tor Epicent	Magnitud	lagnitud Tip	tor Magnit	Profundidad	or Profundi	umero punt	nsidad max	Escala	Autor Intensidad	Area epicentral
1	1644/01/16	05:00	7.37	-72.64	Sarabia G&	6.5	MW	Sarabia G&	15	Sarabia G&	5		9	EMS-98	Cifuentes, H., Sarabia	Pamplona, Norte de Santander
2	1644/03/16	12:00	4.46	-74.04	Dimat&eac	5.5	MW	Sarabia G&	15	Sarabia G&	2	7	EMS-98	Sarabia, A., Cifuentes	Chipaque, Cundinamarca
3	1646/04/03	02:00	5.52	-74.13	Sarabia G&	6	MW	Sarabia G&	15	Sarabia G&	2	8	EMS-98	Sarabia, A., Cifuentes	Muzo, Boyacá
4	1736/02/02	09:00	2.5	-76.5	Sarabia G&	6	MW	Sarabia G&	15	Sarabia G&	1	8	EMS-98	Sarabia, A., Cifuentes	Popayán, Cauca
5	1743/10/18	10:45	4.44	-73.83	Sarabia G&	6.2	MW	Sarabia G&	15	Sarabia G&	13	8	EMS-98	Sarabia, A., Cifuentes	Fómeque, Cundinamarca
6	1766/07/09	16:00	3.82	-76.52	Sarabia G&	6.5	MW	Sarabia G&	15	Sarabia G&	5	8	EMS-98	Cifuentes, H., Sarabia	Buga, Valle del Cauca
7	1785/07/12	07:45	3.42	-74.23	Sarabia, A.	7.1	MW	Gór	10	Gór	17	7	EMS-98	Sarabia, A.; Cifuentes	Piedemonte llanero, Colombia
8	1796/02/15	12:00	7.37	-72.64	Sarabia G&	5.5	MW	Sarabia G&	15	Sarabia G&	1	7	EMS-98	Cifuentes, H., Sarabia	Pamplona, Norte de Santander
9	1805/06/16	03:15	5.37	-74.87	Sarabia G&	6.1	MW	Sarabia G&	15	Sarabia G&	7	9	EMS-98	Cifuentes, H., Sarabia	Honda, Tolima
10	1807/02/17	12:02	6.5	-71.7	Sarabia G&	5.5	MW	Sarabia G&	15	Sarabia G&	1	7	EMS-98	Barbosa Castro, D.	Tame, Arauca
11	1826/06/17	22:30	5.01	-73.59	Sarabia G&	6.5	MW	Sarabia G&	15	Sarabia G&	11	8	EMS-98	Sarabia, A.; Cifuentes	Úmbita, Boyacá
12	1827/11/16	18:00	1.8	-75.52	Gór	7.1	MW	Gór	10	Gór	49	10	EMS-98	Sarabia, A.; Cifuentes	Altamira, Huila
13	1834/01/20	07:00	1.1	-76.93	Sarabia G&	6.7	MW	Sarabia G&	15	Sarabia G&	29	9	EMS-98	Sarabia, A.; Cifuentes	Santiago, Putumayo
14	1834/05/22	03:00	11.49	-74.07	Gór	6.4	MW	Gór	10	Gór	9	8	EMS-98	Cifuentes, H., Sarabia	Santa Marta, Magdalena
15	1869/03/06	06:30	9	-74	Sarabia G&	6.1	MW	Sarabia G&	50	Sarabia G&	75	7	EMS-98	Sarabia Góme	El Banco, Magdalena
16	1875/05/18	11:15	7.86	-72.42	Sarabia G&	6.8	MW	Sarabia G&	15	Sarabia G&	54	10	EMS-98	Cifuentes, H., Sarabia	Cúcuta, Norte de Santander
17	1882/09/07	03:20	10	-79	Camacho, I	6.5	MW	Sarabia G&	15	Sarabia G&	17	9	EMS-98	Sarabia, A., Cifuentes	Colón, Colón-Panamá
18	1884/11/05	23:45	5.1	-75.5	Espinosa B	6.3	MS	Ceresis (19	60	Servicio Ge	14	8	EMS-98	Cifuentes, H., Sarabia	Herveo, Tolima
19	1885/05/25	15:05	2.88	-76.54	Sarabia G&	6.4	MW	Sarabia G&	15	Sarabia G&	8	8	EMS-98	Cifuentes, H., Sarabia	El Tambo, Cauca
20	1903/12/01	08:00	6.78	-76.14	Sarabia G&	5.5	MW	Sarabia G&	15	Sarabia G&	2	7	EMS-98	Cifuentes, H., Sarabia	Frontino, Antioquia
21	1906/01/31	10:36	0.988	-79.347	ISC-GEM	8.4	MW	ISC-GEM	20	ISC-GEM	40	10	EMS-98	Sarabia, A., Cifuentes	Costa Pacífica, Pacífico
22	1911/04/10	13:42	7.2	-75.3	Di Giacomo	6.4	MW	ISC-GEM	120	Di Giacomo	11	7	EMS-98	Cifuentes, H., Sarabia	Yarumal, Antioquia
23	1917/08/31	06:36	3.79	-73.65	Di Giacomo	7	MW	ISC-GEM	15	Di Giacomo	67	9	EMS-98	Cifuentes, H., Sarabia	Villavicencio, Meta
24	1923/12/14	05:31	0.87	-77.78	Servicio Ge	6.2	MW	Servicio Ge	10	Servicio Ge	14	9	EMS-98	Sarabia, A., Cifuentes	Cumbal, Nariño
25	1923/12/22	04:56	4.56	-73.51	Di Giacomo	5.9	MW	Sarabia G&	15	Di Giacomo	36	8	EMS-98	Sarabia, A., Cifuentes	Medina, Cundinamarca
26	1925/06/07	18:41	3.96	-76.31	Di Giacomo	6.1	MW	ISC-GEM	120	Di Giacomo	34	7.5	EMS-98	Sarabia, A., Cifuentes	Tuluá, Valle del Cauca
27	1926/12/18	20:50	0.87	-77.78	Servicio Ge	6	MW	Servicio Ge	10	Servicio Ge	9	8	EMS-98	Servicio Geológico C	Cumbal, Nariño
28	1928/11/01	11:08	4.95	-73.097	Di Giacomo	5.9	MW	Sarabia G&	15	Di Giacomo	38	8	EMS-98	Barbosa Castro, D.	Chinavita, Boyacá
29	1933/02/10	17:00	1.37	-77.58	Servicio Ge	5.7	MW	Servicio Ge	10	Servicio Ge	5	8	EMS-98	Servicio Geológico C	Linares, Nariño
30	1935/08/07	04:02	1.05	-77.31	Servicio Ge	6.1	MW	Servicio Ge	10	Servicio Ge	23	8	EMS-98	Sarabia, A., Cifuentes	Tangua, Nariño
31	1935/09/17	23:58	5.09	-76.08	ISC-GEM	6.1	MW	ISC-GEM	15	ISC-GEM	20	8	EMS-98	Sarabia, A., Cifuentes	Pueblo Rico, Risaralda
32	1935/10/26	20:15	1.07	-77.51	Servicio Ge	5.9	MW	Servicio Ge	10	Servicio Ge	20	8	EMS-98	Sarabia, A., Cifuentes	Imués, Nariño
33	1936/01/09	23:30	1.1	-77.6	Servicio Ge	5.6	MW	Servicio Ge	10	Servicio Ge	16	7	EMS-98	Sarabia, A., Cifuentes	Túquerres, Nariño
34	1936/07/17	12:30	1.17	-77.73	Servicio Ge	6.3	MW	Servicio Ge	10	Servicio Ge	37	8	EMS-98	Sarabia, A., Cifuentes	Túquerres, Nariño
35	1938/02/04	21:23	4.68	-75.69	Internationa	7	MS	Internationa	150	Internationa	86	8	EMS-98	Cifuentes, H., Sarabia	Eje Cafetero, Colombia
36	1942/05/22	05:30	4.44	-74.64	Di Giacomo	5.7	MW	Sarabia G&	15	Di Giacomo	19	7	EMS-98	Cifuentes, H., Sarabia	Girardot, Cundinamarca
37	1942/12/26	07:30	9.27	-75.52	Di Giacomo	6.2	MW	ISC-GEM	15	Di Giacomo	13	8	EMS-98	Cifuentes, H., Sarabia	Santa Cruz de Lorica, Córdoba
38	1947/07/14	02:01	1.29	-77.24	Servicio Ge	6	MW	Servicio Ge	10	Servicio Ge	60	8	EMS-98	Cifuentes, H., Sarabia	San Juan de Pasto, Nariño

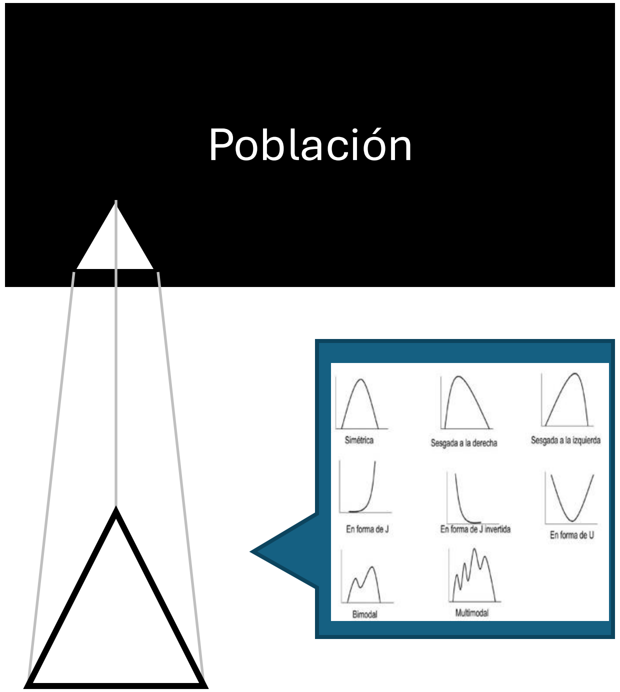
https://sish.sgc.gov.co/visor/sesionServlet?metodo=irAEpicentrosTodos&idDepartamento=&idMunicipio=&cuadranteXMin=&cuadranteXMax=&cuadranteYMin=&cuadranteYMax=

Distribuciones de Probabilidad

Decidir que distribución de probabilidad usar en el modelo

Levantamiento de datos

Data Collection



Muestra - Observaciones

Generar observaciones al azar por medio de distribuciones de probabilidad.

Sin observaciones reales:

Datos historicos

Datos observados

Datos de Entrada del Modelo

Tendencias asumidas

Variables Aleatorias [pseudo-aelatorios]

Def.

Una función (o regla) que asigna un número real (cualquier numero entre -∞ y ∞) dentro del espacio muestral.

e.g.

Un trabajador esta examinando el proceso en una estación de verificación que las piezas estén debidamente procesadas (buena=1; mala=0)

Calidad de Piezas S={1,0}

Si el 95% de las piezas son buenas P(x)=95% y

El 5% de las piezas son malas P(x)=5%

Variables aleatorias

Discretas

- Un operario de call center debe registrar las llamadas que recibe entre 11a.m y 12m.

  {0,1,2,3,4,5,....}

- Personas que llegan a un restaurante.
  - 1 persona = 20%
  - 2 personas = 50%
  - 3 personas = 10%
  - 4 o más = 20%

Continuas

- El tiempo de funcionamiento de una maquina antes de ser nuevamente reparada

  Cualquier número no negativo entre [0, ∞]

- El tiempo de atención de un paciente puede estar entre 2.5 y 4.5 minutos.

  Cualquier valor entre un intervalo [2.5,4.5]minutos

Variables aleatorias

Discretas

Típicamente representan elementos contables.

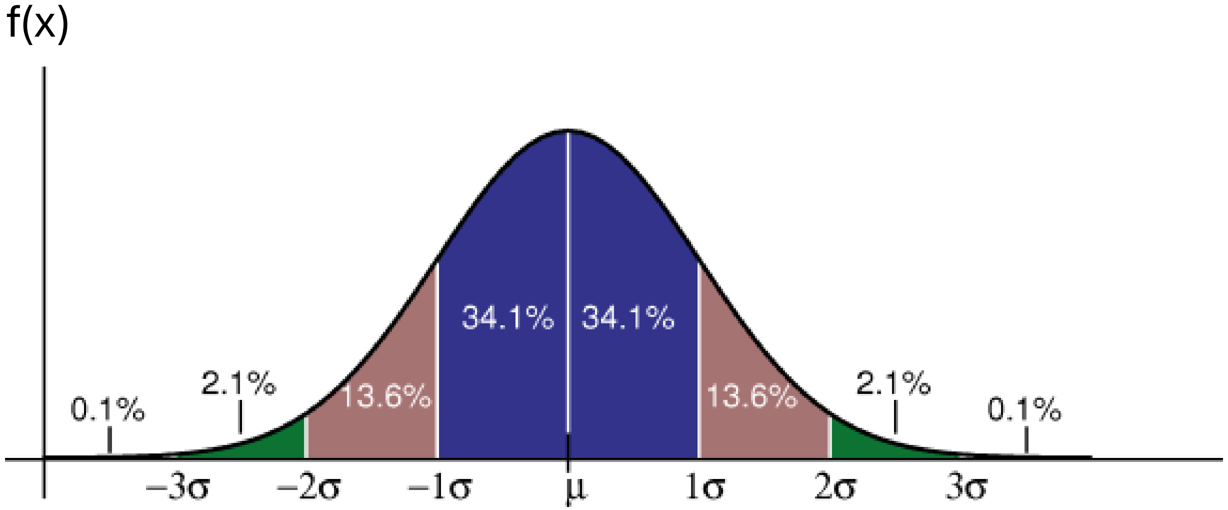
1. El número de partes inspeccionadas en un puesto de trabajo.
2. Número de personas que llegan a un hotel en un fin de semana.
3. Número de partes que están en una banda transportadora a un tiempo determinado.
4. Número de maquinas ocupadas en un tiempo determinado de una estación de torneado de tres tornos.

Continuas

Típicamente representan intervalos de tiempo.

1. El tiempo requerido para reparar una maquina.
2. El tiempo de ciclo necesario para que la siguiente parte a procesar llegue.
3. El intervalo de tiempo para que un nuevo cliente llegue a la fila.
4. El tiempo requerido para que un operario cargue una estiva sobre una estantería de almacenamiento.

Distribuciones de población



$$\mu = \frac{\sum x_i}{n}$$

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{(n-1)}$$

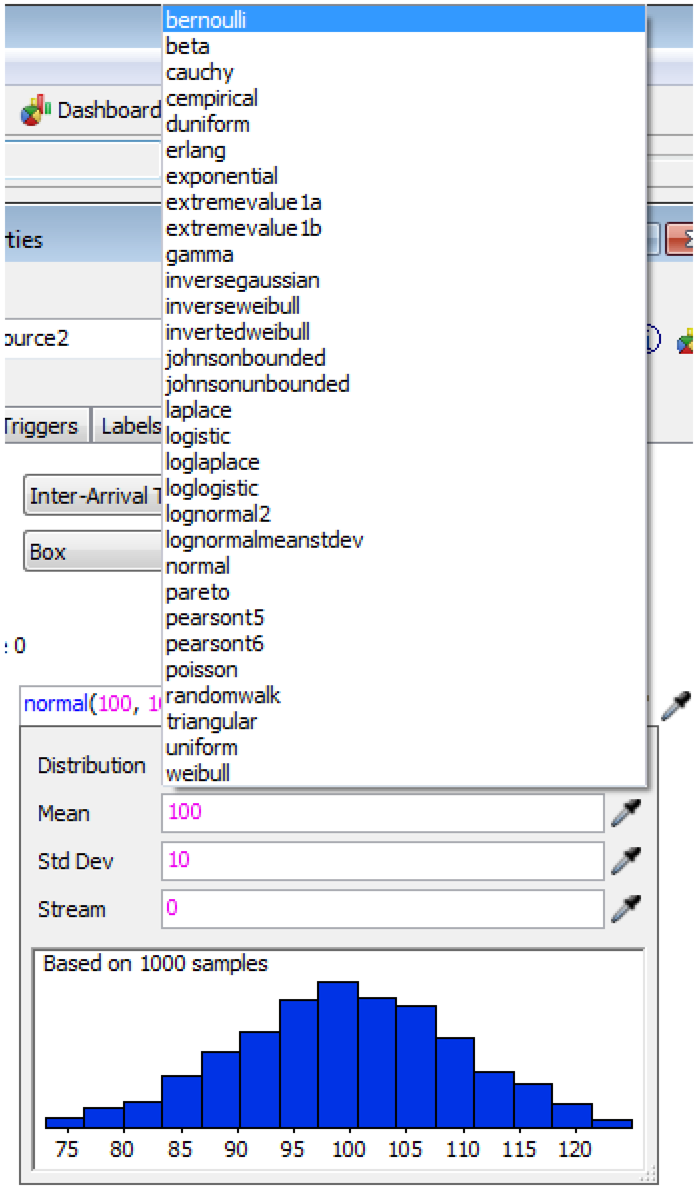
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{(n-1)}}$$

Promedio = Tendencia central de los datos

Varianza = medida de dispersión de los datos.

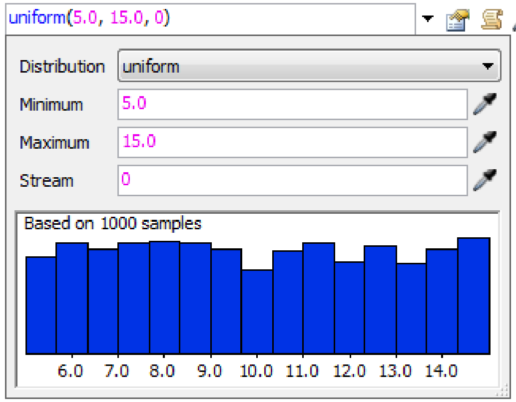
Desviación estándar = medida de dispersión de los datos.

Todas las medidas tienen las mismas unidades.



Tipos de distribuciones

Distribución Uniforme

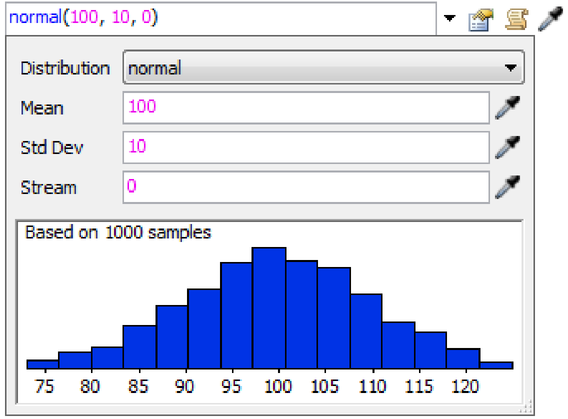


- Datos que siguen un patrón lineal.
- Parámetros que son los extremos del intervalo [min,max].
- Datos aleatorios: cualquier número entre el rango.

Parámetros	[min, max]
Rango	[a,b]
Media	$\frac{a+b}{2}$
Varianza	$\frac{(b-a)^2}{12}$

```
> n = 1000
> duniform = runif(n,10,20)
> hist(uniform, probability = TRUE)
> hist(uniform)
```

Distribución Normal

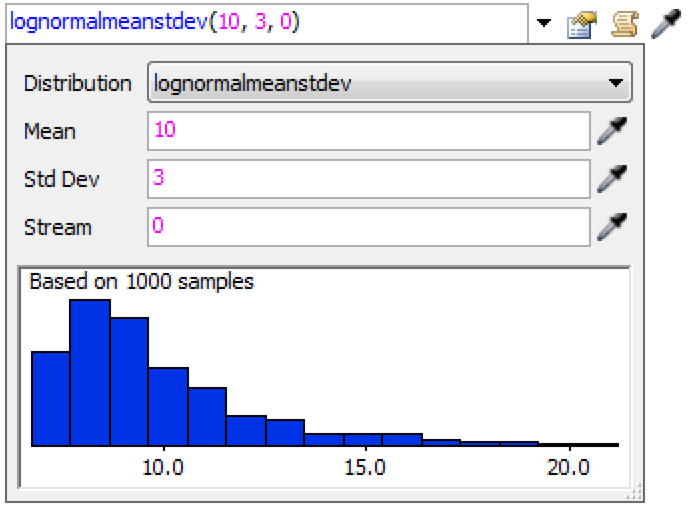


- Distribución de dos colas asimétricas.

Parámetros	μ, σ^2
Rango	[-∞, ∞]
Media	$\frac{\sum x_i}{n}$
Varianza	$\frac{\sum(x_i - \mu)^2}{(n-1)}$

```
> n = 1000
> normal = rnorm(n,8,2)
> hist(norm, probability = TRUE)
> hist(norm)
```

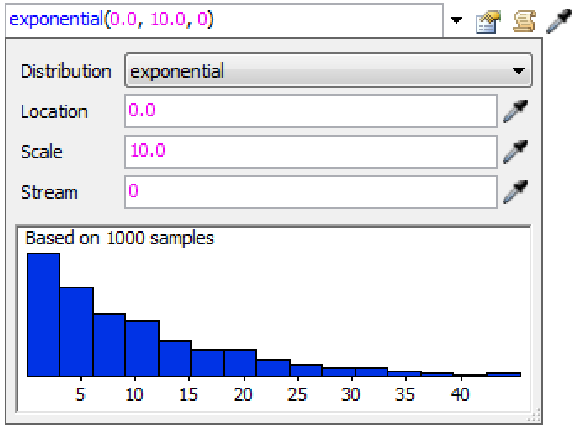
Distribución Lognormal



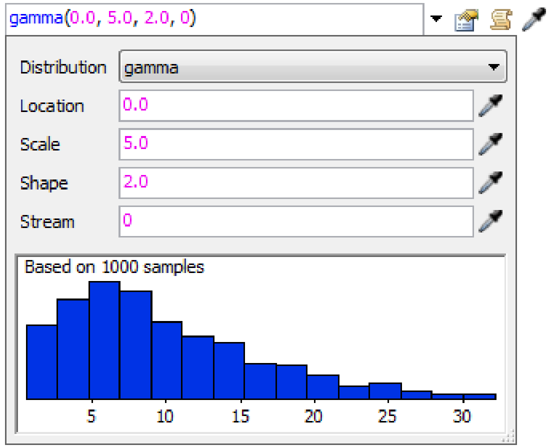
Parámetros	μ, σ^2
Rango	[0, ∞]
Media	$e^{\mu+\sigma^2/2}$
Varianza	$e^{\mu+\sigma^2/2}\left(e^{\sigma^2}-1\right)$

```
> n1 = 100
> lognormal = rlnorm(n,10,3)
> hist(beta, probability = TRUE)
> hist(beta)
```

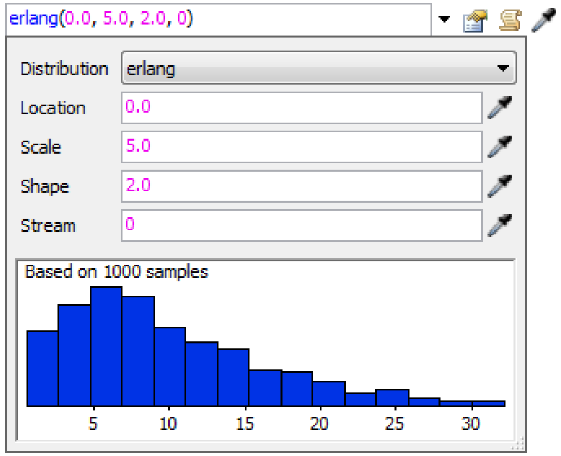
Distribución Exponencial



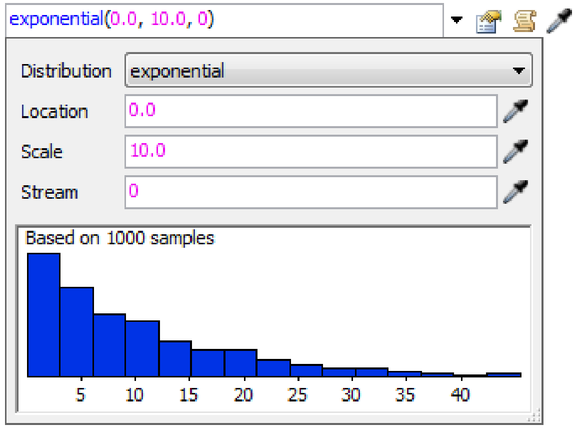
Distribución Gamma



Distribución Erlang



Distribución Exponencial

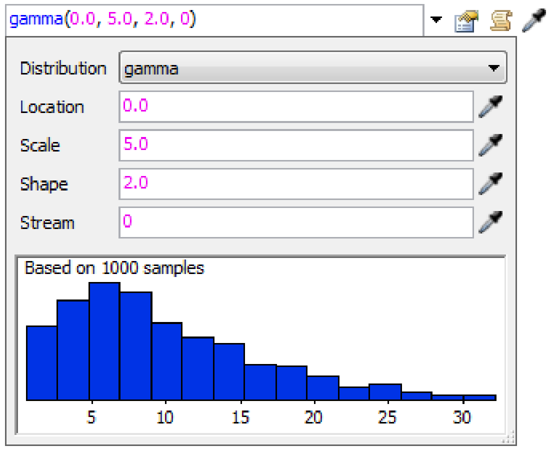


- Distribución que está representada únicamente por un parámetro el cuál representa tanto la media como la desviación estándar.
- Nunca negativa

Parámetros	β
Rango	[0, ∞]
Media	β
Varianza	β^2

```
> n = 1000
> expo = rexp(n,1)
> hist(expo, probability = TRUE)
> hist(expo)
```

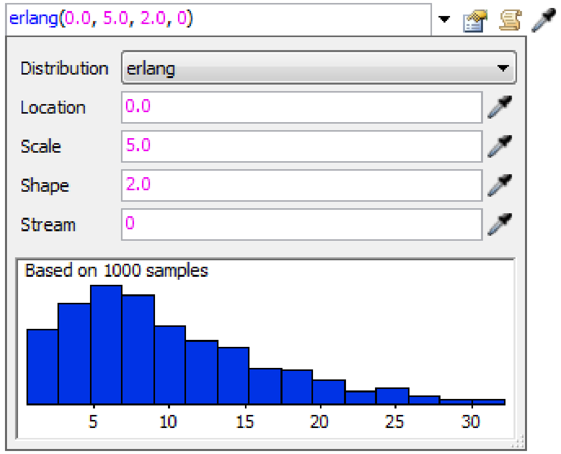
Distribución Gamma



Parámetros	α , β
Rango	[0, ∞]
Media	αβ
Varianza	$\alpha\beta^{2}$

```
> n = 1000
> gamma = rgamma(n,5,2)
> hist(gamma, probability = TRUE)
> hist(gamma)
```

Distribución Erlang

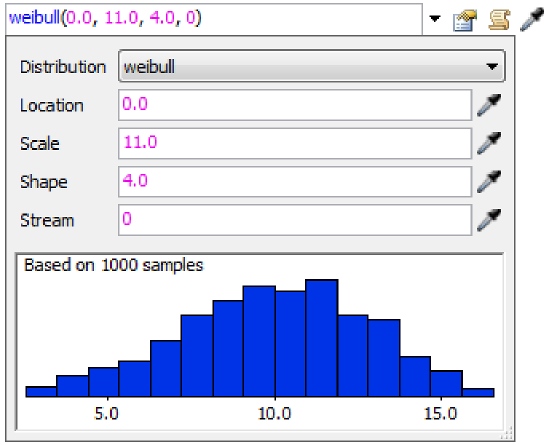


- Esta distribución es un caso específico de la distribución Gamma.
- k: forma
- λ: tasa

Parámetros	k , λ
Rango	[0, ∞]
Media	$\frac{k}{\lambda}$
Varianza	$\frac{k}{\lambda^2}$

```
> n = 1000
> Punto adicional en el taller - quien encuentre la
función para trabajar con este tipo de distribución
en R.
```

Distribución Weibull

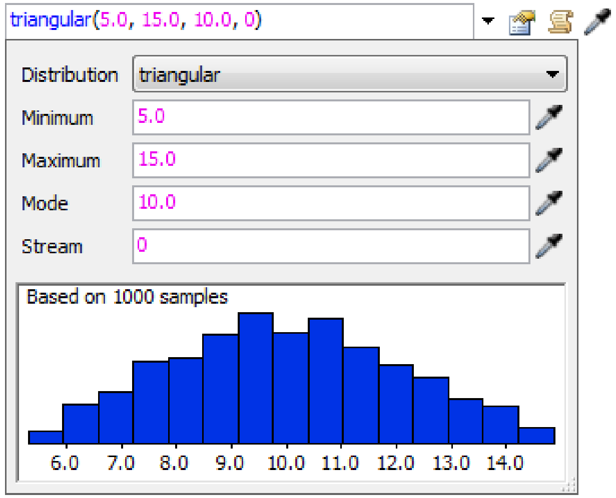


Parámetros	α , β
Rango	[0, ∞]
Media	$\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$
Varianza	$\frac{\beta^2}{\alpha}\left\{2\Gamma\left(\frac{2}{\alpha}\right)-\frac{1}{\alpha}\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^2\right\}$

```
> n = 1000
> weibull = rweibull(n,11,4)
> hist(weibull, probability = TRUE)
> hist(weibull)
```

Distribución Triangular

- Distribución usada comúnmente para representar variables aleatorias continuas en la ausencia de observaciones.

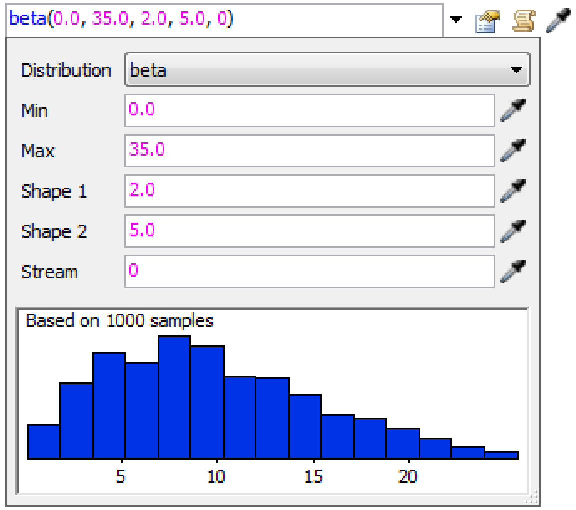


Parámetros	a , b , c
Rango	[a, b]
Media	$\dfrac{a+b+c}{3}$
Varianza	$\dfrac{a^2+b^2+c^2-ab-ac-bc}{18}$

- a: mínimo
- b: máximo
- c: moda

```
> library(triangle)
> n = 1000
> triangular = rtriangle(n,5,15,10)
> hist(triangular, probability = TRUE)
> hist(triangular)
```

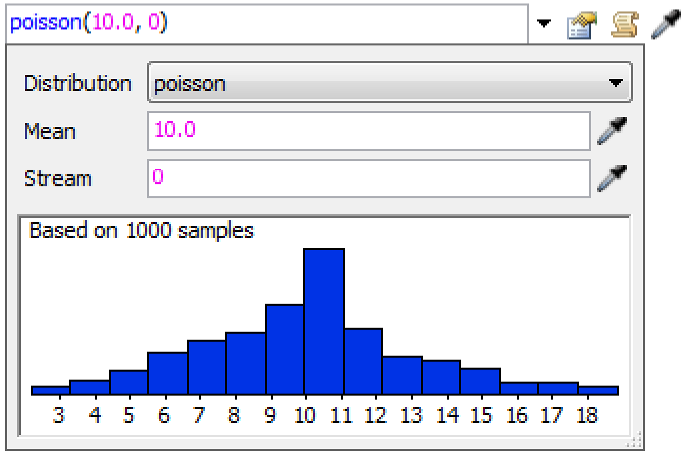
Distribución Beta



Parámetros	α_1 , α_2
Rango	[0, 1]
Media	$\dfrac{\alpha_1}{\alpha_1 + \alpha_2}$
Varianza	$\dfrac{\alpha_1\alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}$

```
> n = 1000
> beta = rbeta(n,35,2,5)
> hist(beta, probability = TRUE)
> hist(beta)
```

Distribución Poisson



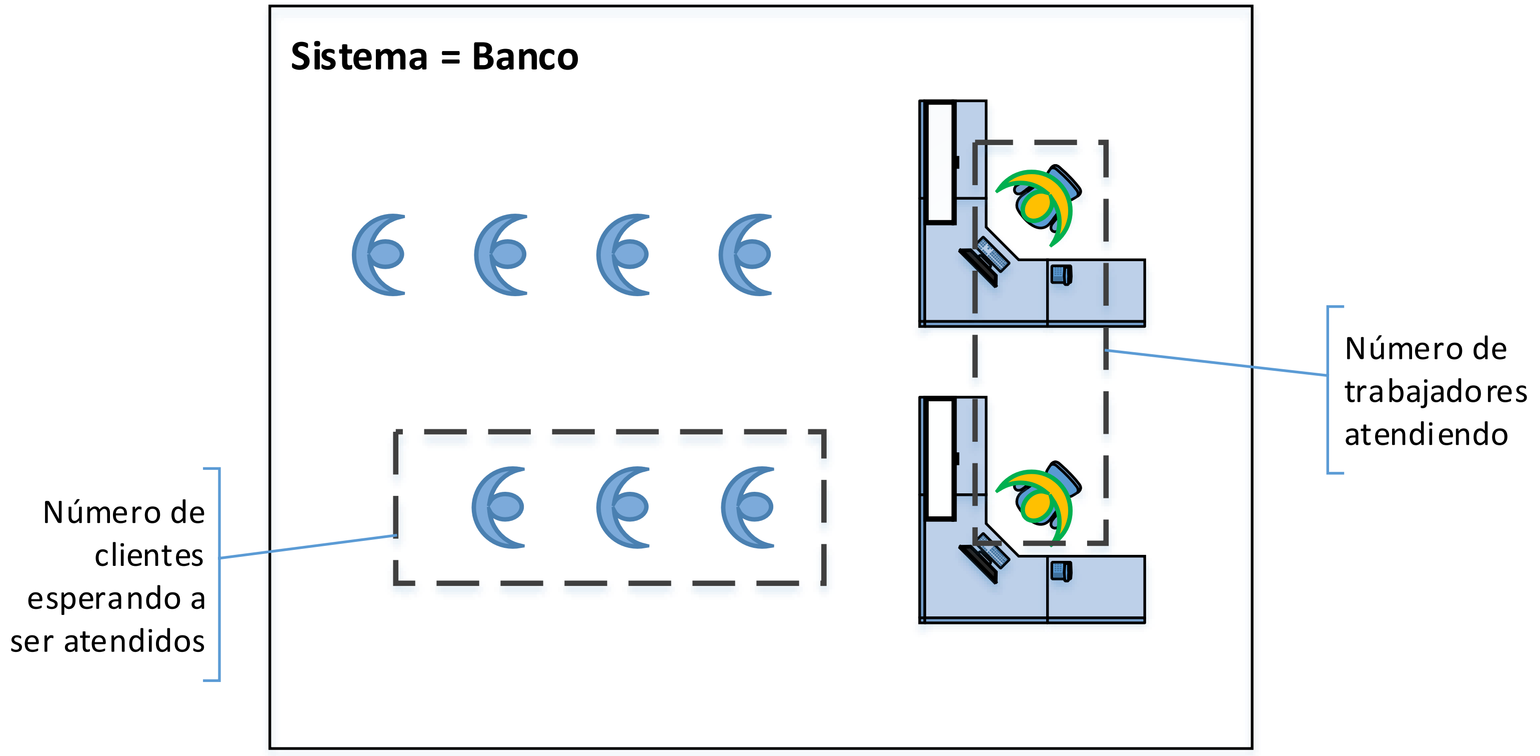
- Numero de eventos que ocurren en un intervalo de tiempo cuando los eventos ocurren a ritmo constante.

Parámetros	λ
Rango	[0, 1,…]
Media	λ
Varianza	λ

```
> n = 1000
> poissonn = rpois(n,10)
> hist(poissonn, probability = TRUE)
> hist(poissonn)
```

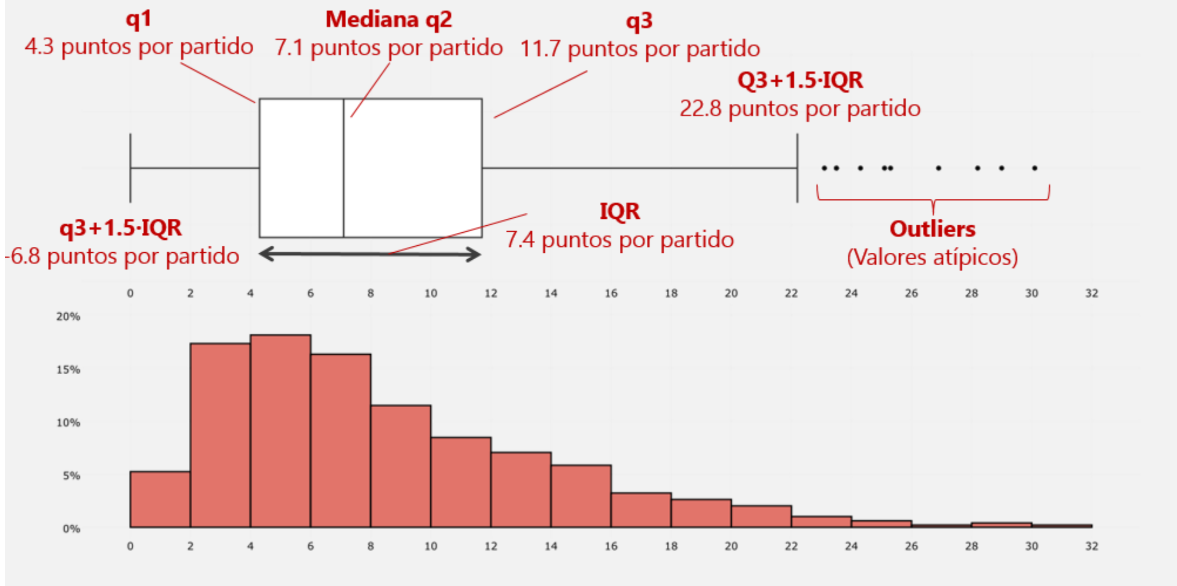
Ejercicio limpieza y análisis de datos

Sistema = colección de elementos (e.g. Operarios y maquinas) que interactúan para desarrollar acciones lógicas.



Propiedades básicas y los box-plots

Cuartil, Rango intercuartílico




```python
# Calcular valores límite para detectar outliers con el
# método del rango intercuartílico (IQR)

Q1 = df1.quantile(0.25)
print('Q1', Q1)

Q3 = df1.quantile(0.75)
IQR = Q3 - Q1
print('IQR', IQR)

# Limites
limite_inferior = Q1 - 1.5 * IQR
print(limite_inferior)
limite_superior = Q3 + 1.5 * IQR

# Mostrar outliers
outliers = df1[(df1 < limite_inferior) | (df1 > limite_superior)]
print("Valores atípicos (outliers):")
print(outliers.sort_values())
print('Minimo:', outliers.min())
```

```
Q1 7.266666666666667
IQR 11.95
-10.658333333333331
Valores atípicos (outliers):
6753     37.966667
305      38.516667
6173     38.583333
5007     39.850000
7560     40.383333
6506     41.116667
3619     43.250000
4510     45.700000
5175     47.333333
721      47.933333
267      48.383333
3400     49.483333
4938     49.750000
8354     55.066667
2557     57.733333
4129     61.500000
4533     62.316667
2277     79.516667
2354    837.733333
Name: Minutos, dtype: float64
Minimo: 37.96666666666667
```

```
(np.float64(14.168746533555188), np.float64(7.616037875882752))
```

	Minutos
count	686.000000
mean	15.246088
std	33.110063
min	0.000000
25%	7.266667
50%	12.700000
75%	19.216667
max	837.733333

	Minutos
count	601.000000
mean	14.168747
std	7.622382
min	0.566667
25%	8.533333
50%	13.250000
75%	18.816667
max	34.216667

Prueba de Kolmogorov–Smirnov (KS)

La prueba de Kolmogorov–Smirnov (KS) es una prueba estadística no paramétrica que se utiliza para comparar distribuciones. Tiene dos usos principales:

1. Prueba de ajuste (1 muestra)

Se utiliza para verificar si un conjunto de datos sigue una **distribución teórica específica**, como la Normal, Exponencial, etc.

Ejemplo:
¿Los datos de temperatura siguen una distribución normal?

2. Prueba de comparación (2 muestras)

Se usa para comparar **dos conjuntos de datos** y verificar si provienen de la **misma distribución**.

Ejemplo:
¿Las temperaturas medidas en dos estaciones diferentes siguen la misma distribución?

Ventajas:

- No requiere suposiciones sobre la forma de la distribución (no paramétrica).
- Se puede aplicar con pocos datos.