Ajuste de una distribución a un conjunto de datos observados.

Cuando se dispone de un conjunto de datos reales (observados) debemos realizar una prueba de "bondad de ajuste" para verificar si los mismos proceden de alguna distribución de probabilidad teórica conocida. Si tal ajuste no es posible, podemos construir una distribución empírica. En general, para realizar el ajuste de datos reales con una distribución teórica se realizan los siguientes pasos:

1. *Dibujar un histograma de los datos empíricos* (abcisa: valor de la variable aleatoria; ordenada: frecuencia en % con que aparece ese valor en todas las observaciones).

2. *Identificar (estimar) los valores de los parámetros de cada distribución potencial para la realización del ajuste con el patrón representado por los datos empíricos*. Una vez seleccionada tentativamente una distribución teórica que se ajusta a las observaciones, es necesario estimar los parámetros de la misma. En ciertos casos, como en la *Poisson* o en la *Normal*, es fácil. En otros, como la *Beta*, la *Weibull* o la *Gamma*, se requiere de más trabajo estadístico para hacer una buena estimación de parámetros.

3. *Realizar una prueba de bondad de ajuste*. Se trata de una comparación del histograma con la distribución seleccionada (prueba *ji-cuadrada o chi-cuadrada*, o prueba de *Kolmogorov-Smirnov*).

Para seleccionar una distribución a ajustar a los datos observados hay que tener presente lo siguiente:

1. Características de cada función de distribución en particular.
2. Exactitud con la cual la función de distribución puede representar el conjunto de datos empíricos dados.
3. Facilidad con la cual se puede hacer el ajuste.
4. Eficiencia computacional cuando se generan tales variables.

Construcción de una distribución empírica.[1]

Sean $X_1, X_2, \ldots, X_n$ valores aleatorios obtenidos (de la evaluación de algún modelo matemático o datos recolectados). Deseamos agrupar la información en varios intervalos sucesivos y determinar la distribución acumulada correspondiente.

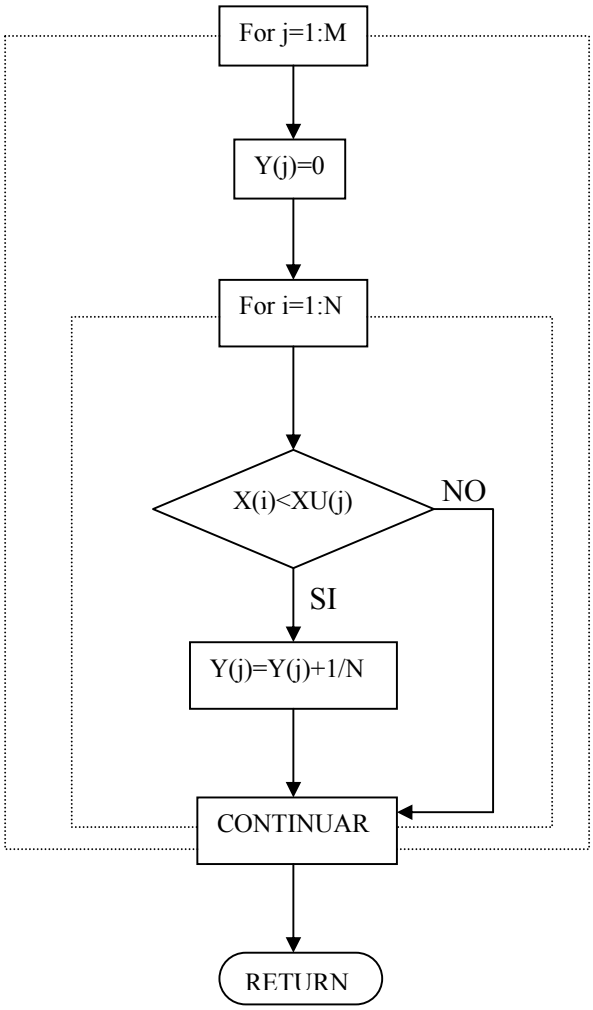
Se determinan cuantos valores caen dentro de cada intervalo. Se compara cada X_i con $XU_1, XU_2, \ldots, XU_j, \ldots$ hasta encontrar $X_i < XU_j$.

[1] *Byron S. Gottfried,* "Elements of Stochastic Process Simulation", Prentice Hall Inc., Englewood Cliffs, New Jersey, 1984.

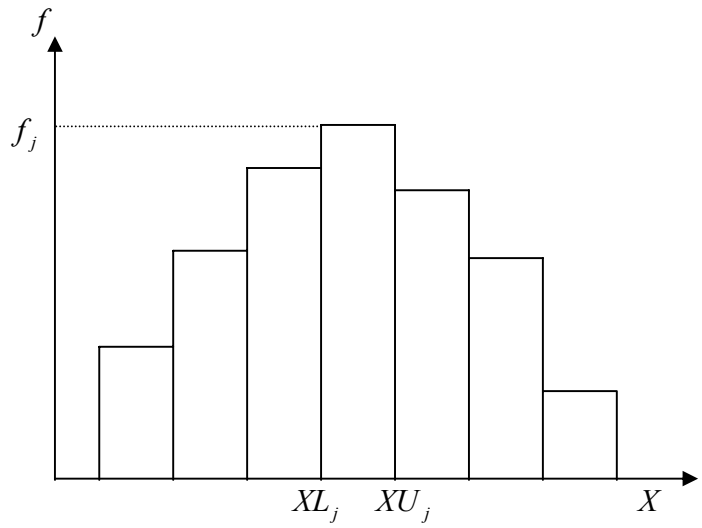
$$a < XU_1 < XU_2 < \cdots < XU_m = b \text{ (cotas superiores)}$$

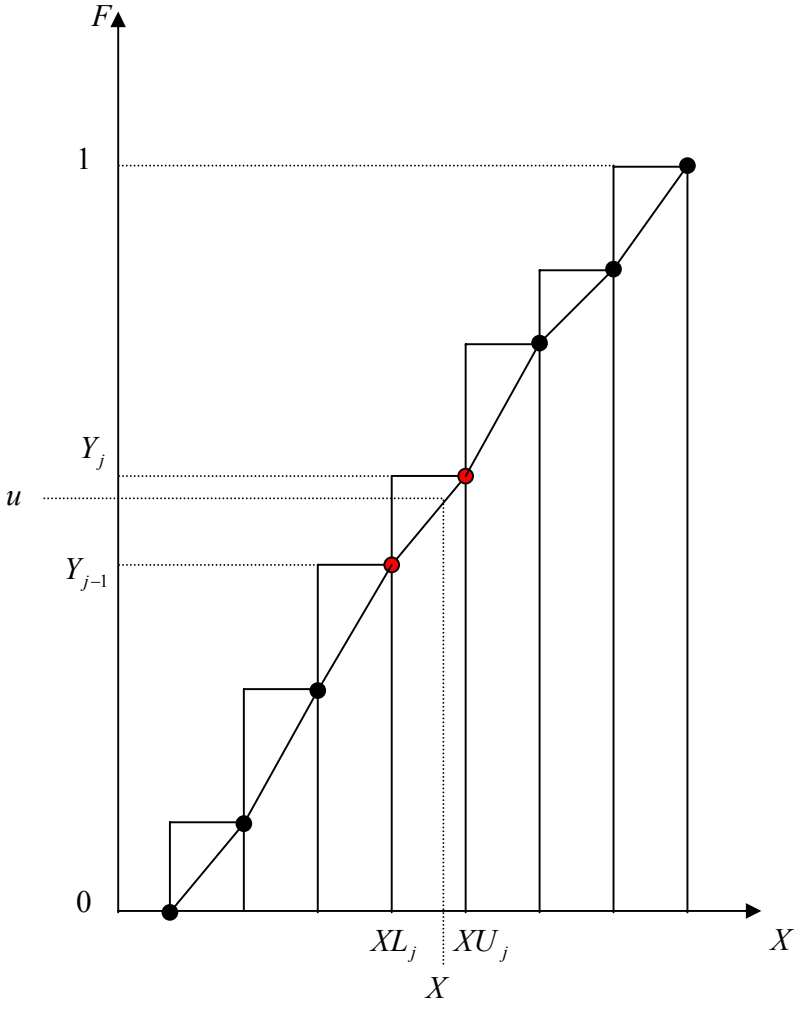
$$a = XL_1 < XL_2 < \cdots < XL_m < b \text{ (cotas inferiores)}$$

$$\left.\begin{aligned} XL_j &= a + \left(\frac{b-a}{m}\right)(j-1) \\ XU_j &= a + \left(\frac{b-a}{m}\right) j \end{aligned}\right\} \quad j = 1,2,\ldots,m \text{ (}a\text{, }b\text{, }m\text{ conocidos)}$$



Generación de variables a partir de una distribución empírica.





$Y_1 = f_1 \quad Prob(X)$ en $XL_j < X < XU_j$

$Y_2 = f_1 + f_2$

...

$Y_j = f_1 + f_2 + \cdots + f_j$

...

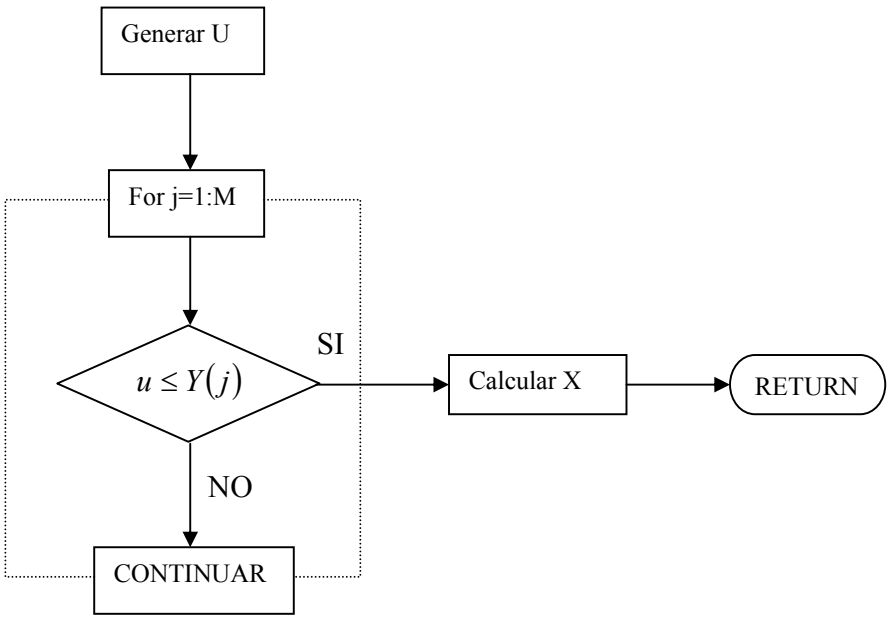
$Y_m = f_1 + f_2 + \cdots + f_m = 1$

Para usar el método de la transformada inversa debemos pasar una curva continua a través de la distribución acumulada. En la figura se utilizó la forma más sencilla, usando segmentos de recta. Entonces, se genera un valor *u* (entre 0 y 1) y el valor correspondiente de *X* se obtiene así:

$$y - y_0 = m(x - x_0)$$

$$(u - Y_{j-1}) = \left(\frac{Y_j - Y_{j-1}}{XU_j - XL_j}\right)(X - XL_j)$$

$$\therefore X = XL_j + \left(\frac{u - Y_{j-1}}{Y_j - Y_{j-1}}\right)(XU_j - XL_j)$$



Tests para ajuste de una distribución.[2]

En general, tales tests proceden de acuerdo a la siguiente secuencia de pasos:

1. Determinar donde caerían tales puntos si ellos hubiesen pertenecido a la distribución dada.
2. Comparar los puntos actuales (observados) con los anticipados (1) a fin de calcular una medida de "error" (depende del test) o diferencia entre el valor actual y el esperado.
3. Comparar esta medida de error (test estadístico) con una tabla de valores críticos. Si el error observado es menor que el valor crítico, no se tendrá una razón para no utilizar tal distribución teórica como modelo para el fenómeno que generó los datos. De otro modo, el error es más grande que el atribuido a una variación aleatoria y tal distribución no podrá ser usada.

Test Ji-cuadrada para distribuciones discretas.

Procede mediante el conteo del número de ocurrencias en cada intervalo. El conteo actual se compara con aquel esperado en cada intervalo. Si la diferencia es muy grande la distribución es rechazada.

Ejemplo:

Se quiere determinar si los datos que se muestran a continuación corresponden a una distribución donde los diez dígitos, 0,1,2,...,9, tienen la misma probabilidad de ocurrir (uniformidad).

3 6 2 1 5
8 2 6 7 1
1 6 5 4 6... (últimos dígitos de los primeros 100 teléfonos de la guía telefónica)

Se quiere probar la hipótesis que los dígitos individuales mostrados anteriormente poseen la distribución, $P(X = k) = \frac{1}{10}; \quad k = 0,1,\ldots,9$.

Dígito	0	1	2	3	4	5	6	7	8	9	*Total*
Frecuencia esperada	10	10	10	10	10	10	10	10	10	10	100
Frecuencia actual	4	15	9	9	9	12	10	7	10	15	100

Se consideran las dos hipótesis siguientes:

[2] *Arne Thesen, Laurel E. Travis,* "Simulation for Decision Making", West Publishing Company, 1992.

Ho: *nuestro conjunto de datos es una serie de observaciones independientes provenientes de una distribución específica.*

Ha: *Nuestro conjunto de datos no es una serie de observaciones independientes provenientes de una distribución específica.*

El test consiste en calcular: $Ji^2 = \sum_{i=1}^{K} \frac{[c_i - E(c_i)]^2}{E(c_i)}$

donde:

- c_i = número de datos observados.
- $E(c_i)$ = número esperado de observaciones en el *i*-ésimo intervalo (asumiendo ***Ho*** correcta).
- K = número de intervalos.

Dígito i	*Observaciones* c_i	*Obs. Esperadas* $E(c_i)$	*Errores* $c_i - E(c_i)$	*Ji-cuadrada* $\frac{[c_i - E(c_i)]^2}{E(c_i)}$
0	4	10	-6	3.6
1	15	10	5	2.5
2	9	10	-1	0.1
3	9	10	-1	0.1
4	9	10	-1	0.1
5	12	10	2	0.4
6	10	10	0	0.0
7	7	10	-3	0.9
8	10	10	0	0.0
9	15	10	5	2.5
Total	100	100	0	$Ji^2 = 10.2$

nivel de significación (α) = *P*(rechazar ***Ho/Ho*** es cierta) = *P*($d > d_c$***/Ho*** es cierta)

Error tipo I = *P*(rechazar ***Ho/Ho*** es cierta)
Error tipo II = *P*(no rechazar ***Ho/Ho*** es falsa)

$(1-\alpha)$ = *P*(aceptar ***Ho/Ho*** es cierta) = *P*($d \le d_c$ (aceptación)***/Ho*** es cierta)

Si fijamos el nivel de significación en, $\alpha = 0.10 \Rightarrow 1-\alpha = 0.90$.
Ji-cuadrada se distribuye con $\nu = K - 1 = 9$ grados de libertad.
$\therefore Ji^2_{9;0.90} = 14.7 = d_c$ (valor tabulado)

Similarmente, $\alpha = 0.05 \Rightarrow 1-\alpha = 0.95 \quad \therefore Ji^2_{9;0.95} = 16.9$

Bosquejo del test:

1: Establecer la hipótesis:
Ho: *El conjunto de datos es una serie de observaciones independientes provenientes de la distribución a ser evaluada.*

2: Especificar el nivel de significancia:
α = *riesgo de rechazar erróneamente una hipótesis verdadera* (***Ho***).
$\alpha = 10\%$ ó $\alpha = 5\%$.

3: Establecer el número de intervalos a ser usados:
En el ejemplo se usaron 10. El número de observaciones esperadas en cada intervalo debe ser al menos 5. Si son menores de 5,se utilizarán menos intervalos o se tomarán más observaciones.

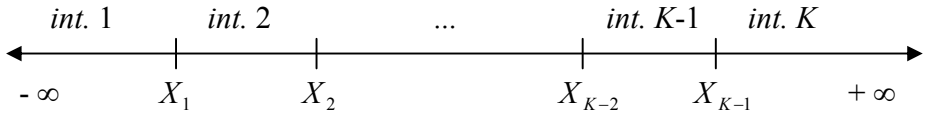
4: Contar frecuencias y calcular errores (tabla anterior).

5: Calcular el valor de la *Ji-cuadradra* (tabla anterior).

6: Encontrar el valor crítico en una tabla estadística (rechazar ***Ho*** si el valor de la *Ji-cuadrada* es mayor que el valor crítico).

Test Ji-cuadrada para distribuciones continuas.

Modifiquemos el test anterior para ajustar distribuciones continuas y distribuciones cuyos parámetros han sido estimados a partir del conjunto de datos. Se separarán las observaciones *K* intervalos especificando *K*-1 puntos finales X_i donde $i = 1,2,...,K\text{-}1$.



$$Prob\{X \text{ esté en el intervalo } i\} = \begin{cases} F(X_1); & i = 1 \\ F(X_i) - F(X_{i-1}); & 1 < i \le K-1 \\ 1 - F(X_{K-1}); & i = K \end{cases}$$

donde *F* es la función de distribución acumulada. El ancho de los intervalos se seleccionará de manera que el número esperado de observaciones (bajo ***Ho***) en cada intervalo sea mayor o igual a 5. El ¿cómo? se determina el ancho en la práctica, depende de la distribución a ajustar.

Ejemplo:

Se tiene el siguiente conjunto de datos. Deseamos evaluar el ajuste de los datos a una exponencial (***Ho***).

3.20	3.87	2.51	2.64	2.97
3.09	1.56	3.12	3.04	2.37
1.93	3.70	3.62	3.95	1.71
1.75	1.02	3.17	1.58	3.88
2.39	2.81	3.81	2.80	1.35
2.10	1.49	1.50	3.10	1.12

$$\hat{\lambda} = \frac{1}{\bar{x}} \Rightarrow \hat{\lambda} = 0.39 \text{ (estimador máximo verosímil)}$$

Como tenemos 30 datos, usaremos 6 intervalos de modo que el número esperado de puntos en cada uno sea 5. Deseamos los valores de la variable para los cuales tengamos igual probabilidad de caer en cada uno de los 6 intervalos. En consecuencia, el chance de caer será de 1/6. Establecemos X_1, X_2, X_3, X_4, X_5 tales que,

$$F(X_1) = \frac{1}{6}; \quad F(X_2) = \frac{2}{6}; \quad F(X_3) = \frac{3}{6}; \quad F(X_4) = \frac{4}{6}; \quad F(X_5) = \frac{5}{6}.$$

$$F(X) = 1 - e^{-\lambda X} = \frac{1}{6} \Rightarrow e^{-\lambda X} = 1 - \frac{1}{6} \Rightarrow X = -\frac{1}{\lambda}\ln\left(1 - \frac{1}{6}\right)$$

$$\therefore X_i = -\frac{1}{\hat{\lambda}}\ln\left(1 - \frac{i}{K}\right) \quad i = 1,2,\ldots,K-1$$

$$X_1 = -\frac{1}{0.39}\ln\left(1 - \frac{1}{6}\right) = 0.47$$

$$X_2 = -\frac{1}{0.39}\ln\left(1 - \frac{2}{6}\right) = 1.04$$

$$X_3 = 1.78 \quad X_4 = 2.82 \quad X_5 = 4.59$$

El número de grados de libertad se calcula a través de la fórmula, $\nu = K - m - 1$, donde m es igual al número de parámetros que han sido estimados a partir de los datos. Luego, $\nu = 6 - 1 - 1 = 4$ (solo se estimó un parámetro, $\hat{\lambda}$).

intervalo	c_i	$E(c_i)$	$c_i - E(c_i)$	$\frac{[c_i - E(c_i)]^2}{E(c_i)}$
0.00→0.47	0	5	-5	5.0
0.48→1.04	1	5	-4	3.2
1.05→1.78	8	5	3	1.8
1.79→2.82	8	5	3	1.8
2.83→4.59	13	5	8	12.8
4.60→∞	0	5	-5	5.0
Total	30	30	0	$Ji^2 = 29.6$

Si $\alpha = 0.05 \Rightarrow Ji^2 = 29.6 > Ji^2_{4;0.95} = 9.49 = d_c$, debemos rechazar ***Ho***.

Tratemos ahora de ajustar una normal al mismo conjunto de datos.

$$\hat{\mu} = \bar{x} = 2.57 \text{ (estimador máximo verosímil)}$$

$\hat{\sigma}^2 = s_{un}^2 = 0.819$ (estimador máximo verosímil) ó $\hat{\sigma}^2 = s^2 = 0.90$ (estimador insesgado)

donde, $s_{un}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (unadjusted) y $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Usando la tabla normal estándar los cinco puntos correctos son: -0.97, -0.43, 0, 0.43, 0.97

$$X_1 = \hat{\mu} + (-0.97)\hat{\sigma} = 1.70$$
$$X_2 = \hat{\mu} + (-0.43)\hat{\sigma} = 2.18$$
$$X_3 = \hat{\mu} + (0)\hat{\sigma} = 2.57$$
$$X_4 = \hat{\mu} + (0.43)\hat{\sigma} = 2.96$$
$$X_5 = \hat{\mu} + (0.97)\hat{\sigma} = 3.44$$

intervalo	c_i	$E(c_i)$	$c_i - E(c_i)$	$\frac{[c_i - E(c_i)]^2}{E(c_i)}$
-∞→1.70	7	5	2	0.8
1.71→2.18	4	5	-1	0.2
2.19→2.57	3	5	-2	0.8
2.58→2.96	3	5	-2	0.8
2.97→3.44	7	5	2	0.8
3.45→∞	6	5	1	0.2
Total	30	30	0	$Ji^2 = 3.6$

$$\nu = K - m - 1 = 6 - 2 - 1 = 3$$

Si $\alpha = 0.05 \Rightarrow Ji^2 = 3.6 < Ji^2_{3;0.95} = 7.81 = d_c$, luego se acepta ***Ho***.