Pruebas de bondad del ajuste

Jhon Jairo Padilla Aguilar, PhD.

Necesidad

- Una vez estimados los parámetros de la distribución a partir de los datos experimentales x1, . . . , xn, cabe preguntarse en qué medida los datos experimentales responden a la distribución ajustada.
- Se trata de responder a la pregunta: ¿cabría haber obtenido los datos experimentales muestreando la distribución teórica o, por el contrario, existen grandes discrepancias entre la distribución ajustada y los datos experimentales?.
- Para intentar responder a esta pregunta se usan técnicas gráficas y tests estadísticos

Precaución

- Hay que tener en cuenta que la mayoría de los tests no son demasiado potentes: cuando se dispone de pocas observaciones son poco sensibles a discrepancias entre los datos y la distribución ajustada y, sin embargo, cuando se dispone de muchas observaciones una pequeña discrepancia puede hacer que se rechace el ajuste.

Pruebas estadísticas comunes

- Dos pruebas estadísticas muy utilizadas son la prueba de chi-cuadrado y la de Kolmogorov-Smirnov
- La decisión acerca de cuál de estos dos tests usar depende de la naturaleza de la distribución y del tamaño de la muestra

Comparativo entre Pruebas

El test de Kolmogorov-Smirnov

- Es sólo válido para distribuciones continuas
- Es aplicable a cualquier tamaño de muestra.

El test chi-cuadrado

- Es aplicable tanto a distribuciones discretas como continuas.
- Requiere la clasificación de los datos de la muestra en clases.
- Dado que se recomienda que haya al menos 3 observaciones por clase y conviene tener un número razonablemente grande de clases, el test chi-cuadrado no es aplicable a muestras pequeñas.

Técnicas gráficas

- Las técnicas gráficas son una herramienta muy potente de análisis. Comúnmente se usan el mismo tipo de representaciones que para seleccionar la familia de distribuciones, es decir, histogramas y gráficas Q-Q.

Gráfico Cuantil-Cuantil (Gráfico Q-Q)

- Existen técnicas para reducir el problema de la comparación de funciones distribución de probabilidad acumulada a decidir cuál, de entre varias gráficas, se asemeja más a una recta.
- Una de estas técnicas, denominada *gráfica cuantil-cuantil* o *gráfica Q-Q*, está basada en la comparación de los cuantiles o puntos críticos de las distribuciones continuas.

Definición de Cuantil

- El cuantil q (con $0 < q < 1$) de una distribución es un número x_q que satisface:

$$F_X(x_q) = q$$

- Representando F_X^{-1} a la función inversa de la probabilidad acumulada, una definición equivalente del cuantil es:

$$x_q = F_X^{-1}(q)$$

Definición del gráfico Q-Q

La técnica gráfica Q-Q se basa en las dos propiedades siguientes:

- Si las variables aleatorias X e Y están igualmente distribuidas, entonces sus densidades de probabilidad y sus cuantiles son iguales: $y_q = x_q$. La representación gráfica de los puntos (x_q, y_q), para $q \in (0, 1)$, es una recta con pendiente unidad que pasa por el origen.
- Si las variables aleatorias X e Y pertenecen a la misma familia de distribuciones y sus densidades acumuladas difieren en el valor los parámetros de posición (γ) y escala (β), es decir $F_Y(y) = F_X\left(\frac{y-\gamma}{\beta}\right)$
- entonces la relación entre los cuantiles de las distribuciones es: $y_q = \gamma + \beta \cdot x_q$
- Consiguientemente, la representación gráfica de los puntos (xq, yq), para q $\in$ (0, 1), es una recta con pendiente que no pasa por el origen.

Ventaja de la prueba Q-Q

- A efectos de la aplicación de la técnica, sólo es relevante el parámetro de forma de la distribución teórica. Los parámetros de escala y posición son irrelevantes, con lo cual pueden escogerse de la forma que resulte más sencilla.

Construcción del gráfico Q-Q

- Los dos pasos a seguir para comparar la distribución empírica de un conjunto de datos experimentales, x_1, ..., x_n, con la distribución F_X son los siguientes:
- En primer lugar hay que construir una función de probabilidad acumulada empírica, a partir de los datos experimentales. Para ello es preciso ordenar crecientemente los datos experimentales: x(1) ≤ x(2) ≤ · · · ≤ x(n), donde x(i) representa el dato que ocupa la i-´esima posición. El valor de la distribución empírica en cada uno de los datos experimentales es igual al número de datos menor o igual que éste:

$$\tilde{F}\left(x_{(i)}\right) = \frac{i}{n} \text{ para } i : 1, \ldots, n$$

Construcción del gráfico Q-Q

- Esta definición presenta la desventaja de que la probabilidad acumulada vale 1 para el valor experimental mayor. Una forma de evitar este inconveniente es modificar ligeramente la definición (Law & Kelton 2000): $\tilde{F}\left(x_{(i)}\right) = \frac{i - 0.5}{n}$ para $i : 1, \ldots, n$

- A continuación hay que construir la gráfica Q-Q. Puesto que $x_q = x(i)$ es el cuantil $q = (i-0.5)/n$ de la distribución empírica, la gráfica consiste en la representación de los puntos:

$$\left(x_{(i)}, F_X^{-1}\left(\frac{i - 0.5}{n}\right)\right) \text{ para } i : 1, \ldots, n$$

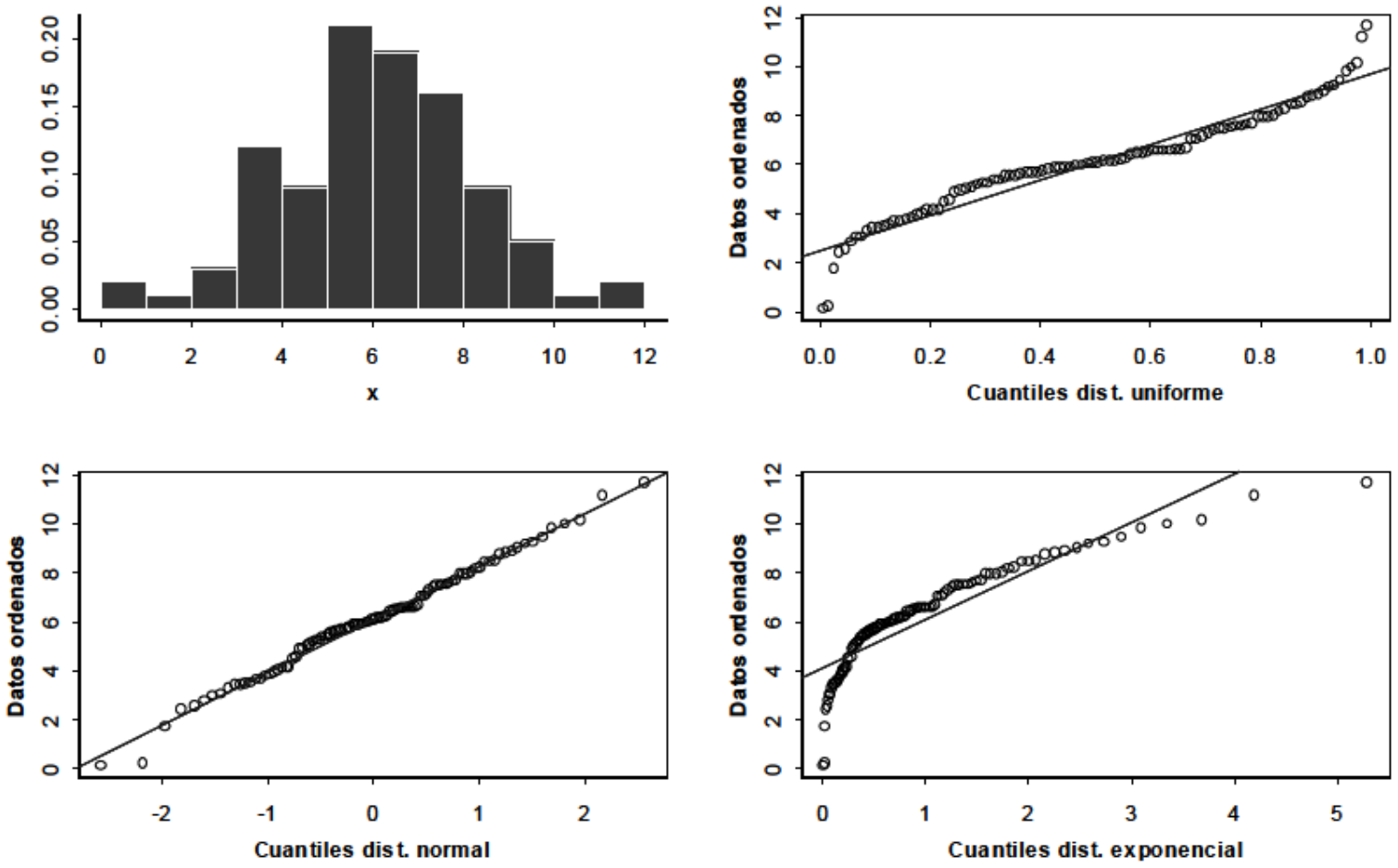
Análisis del gráfico Q-Q

- Si estos puntos se ajustan razonablemente bien a una recta (con independencia de su pendiente o de si pasa o no por el origen), entonces está justificada la hipótesis de que la familia y el factor de forma de la distribución de la que están muestreados los datos experimentales y de F_X coinciden.

Ejemplo 1

- Se ha recogido una muestra de 100 datos experimentales.
- En primer lugar se dibuja el histograma de los datos. Se observa que el histograma es aproximadamente simétrico (esta observación puede contrastarse estimando el sesgo a partir de los datos experimentales). Por ello, tiene sentido realizar comparaciones con la distribución normal.
- Como puede observarse de la segunda gráfica Q-Q, la comparación con la distribución normal es bastante satisfactoria.
- Sobre las gráficas Q-Q se ha dibujado la recta ajustada en cada caso. Como referencia del comportamiento de la gráfica Q-Q cuando la distribución empírica y teórica tienen una forma considerablemente diferente, se han comparado los datos experimentales con la distribución uniforme y exponencial (primera y tercera gráfica Q-Q respectivamente).

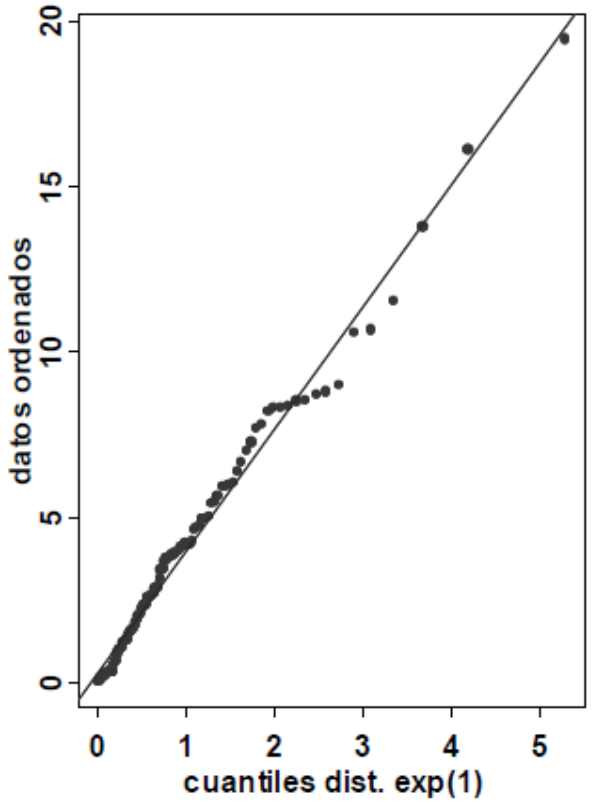
Gráfico Q-Q (Cuantiles)

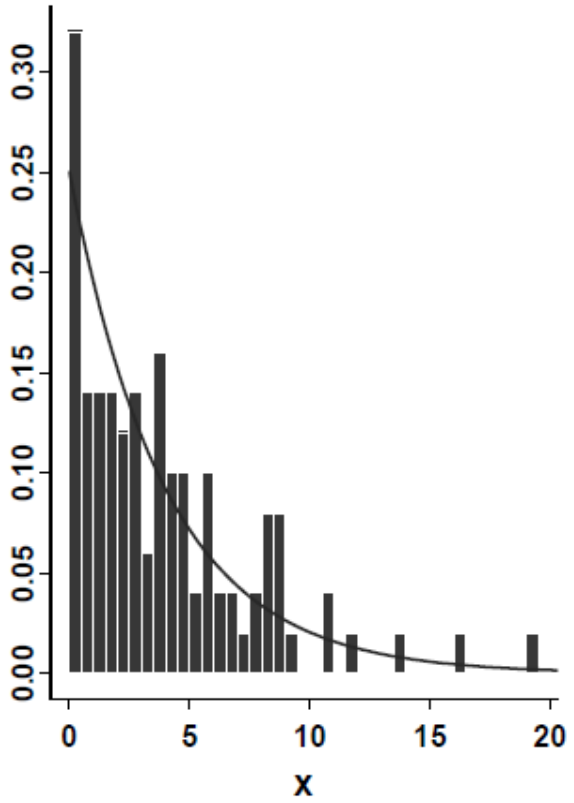


Ejemplo 2

- Se han tomado 100 muestras de un proceso de llegada (tiempos entre llegadas sucesivas). Apoyándose en razonamientos teóricos y en la gráfica Q-Q , se ha realizado la hipótesis de que los datos están distribuidos exponencialmente. El estimador del parámetro de la distribución se calcula de los datos experimentales: media= 4.
- A continuación se compara la distribución ajustada con los datos experimentales. Para ello, se representa el histograma de los datos experimentales escalado como una densidad de probabilidad (la suma del área de las barras vale uno) y la densidad de probabilidad de la distribución ajustada: expo (4).
- Este gráfico se muestra en la parte derecha de la Figura.

Gráfico Q-Q (Cuantiles)





Test de Kolmogorov-Smirnov

- El test chi-cuadrado se basa en observaciones realizadas sobre la función de probabilidad o densidad.
- El test de Kolgomorov-Smirnov (K-S) se basa en la función de distribución o acumulativa (CDF).

Principios del Test K-S

- A partir de una secuencia de números (x_1, . . . , x_n), las observaciones se obtienen a partir de la función CDF discreta producida por esta serie:

$$F_n(x) = \frac{\sharp\{x_i \leq x\}}{n}$$

- Es decir, esta función toma en x el número de elementos de la serie que son menores o iguales a x.

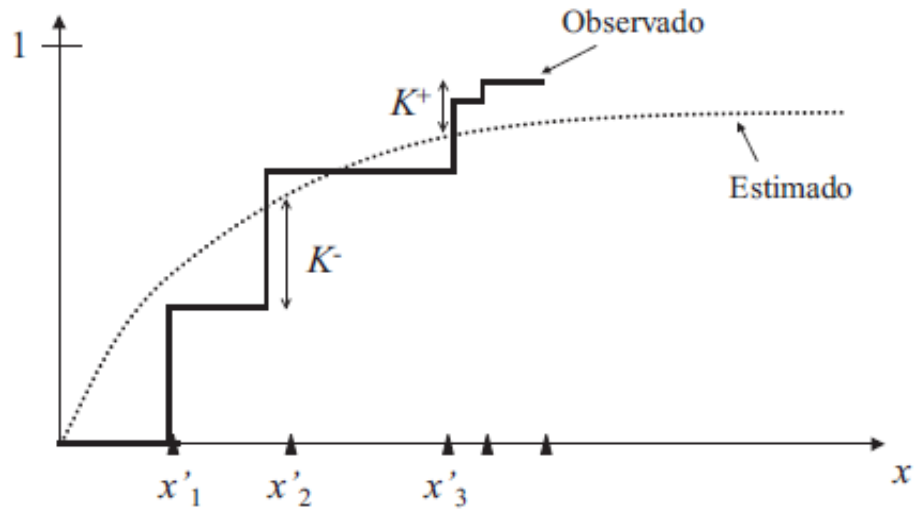
Principios del Test K-S

- Esta función se compara con otra función estimada $F_e(x)$. Concretamente, mediremos la desviación máxima entre ambas funciones:

$$K^+ = \sqrt{n}\, max_x[F_n(x) - F_e(x)]$$

$$K^- = \sqrt{n}\, max_x[F_e(x) - F_n(x)]$$

Principios del Test K-S



- El significado de K+ y K– se puede ver en la gráfica.
- El valor de K+ expresa la máxima diferencia entre ambas funciones cuando la CDF de las observaciones supera a la CDF estimada.
- K- expresa la máxima diferencia entre ambas funciones cuando la CDF estimada supera a la CDF de las observaciones.

Principios del Test K-S

- Al igual que en el test chi-cuadrado, en este test **las desviaciones calculadas siguen una distribución aleatoria**, denominada K, con *n* grados de libertad. Así, dependiendo del grado de confianza deseado, la desviación máxima debería ser menor que el valor $K_{[1-\alpha;n]}$

Procedimiento de la prueba K-S

Por ejemplo, si queremos comprobar si un generador de números aleatorios es bueno, compararemos la serie generada de longitud n con la distribución uniforme U(0, 1), del siguiente modo:

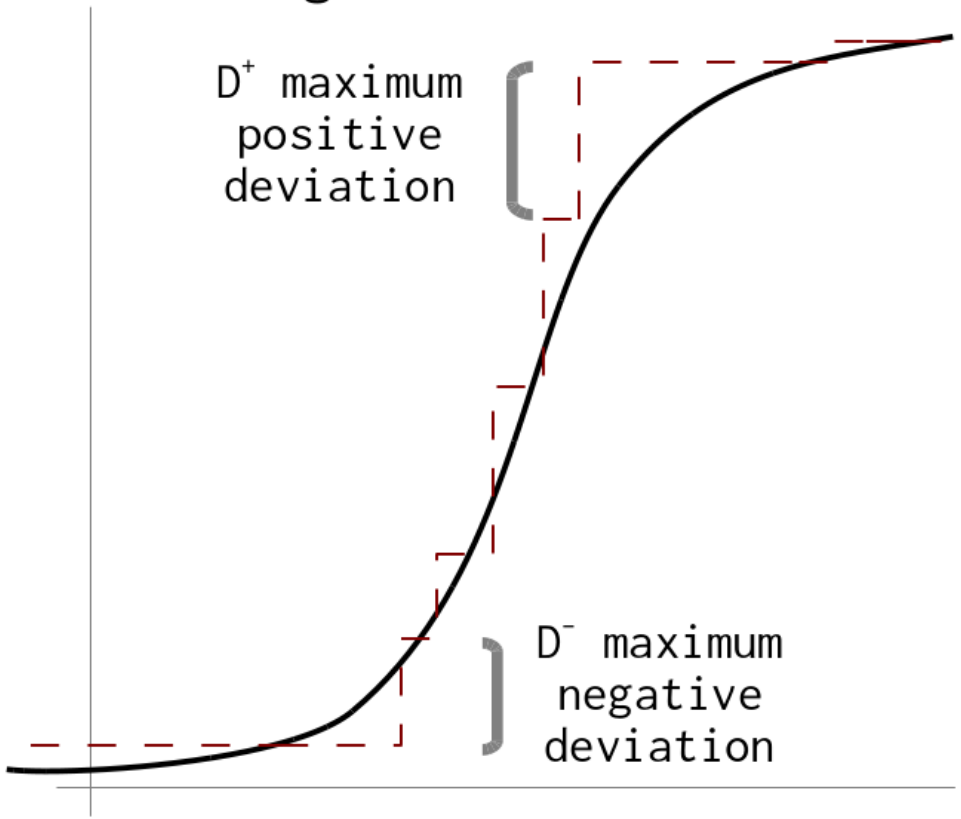
- Ordenamos la serie de menor a mayor ($x'_1, \ldots, x'_n$).
- Calculamos las desviaciones máximas como sigue:

$$K^+ = \sqrt{n}\, max_j[\frac{j}{n} - x'_j]$$

$$K^- = \sqrt{n}\, max_j[x'_j - \frac{j-1}{n}]$$

- Finalmente, si el valor de K+ o K− es mayor que $K[1-\alpha;n]$, entonces descartaremos el generador.

Schematic Representation of Kolmogorov-Smirnov Test



— Theoretical CDF
– - Empirical CDF

For *n* empirical observations:

$K^+ = \sqrt{n}\, D^+$

$K^- = \sqrt{n}\, D^-$

Example: Kolmogorov-Smirnov Test

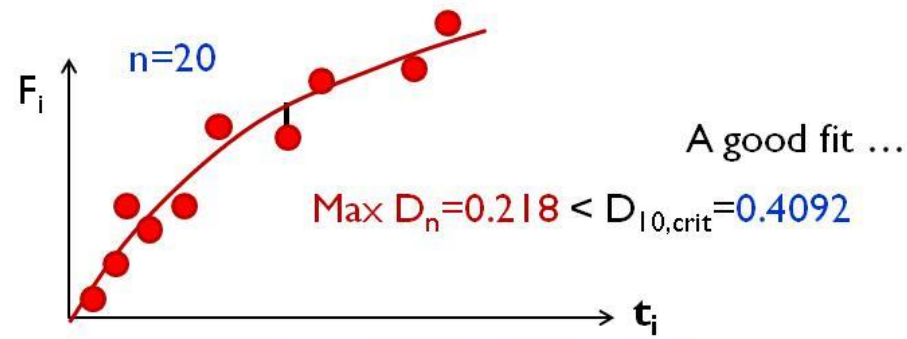
Compute … $D_n = \max\left|F_{obs}(t_i) - F_{theory}(t_i)\right|$

Sample size

If $D_n > D_n^{crit}$, fit is poor …

n	$D_{crit}(n)$
5	0.5633
10	0.4092
20	0.2941
50	0.1884

t_i	3	20	40	52	53	54	85	318	429	553
$F_i=(i-.3)/(n+.4)$										



A good fit …

Max D_n=0.218 < $D_{10,crit}$=0.4092

```
[h,p] = kstest(x,'CDF',test_cdf,'Alpha',0.01)
```