UNIVERSIDAD POLITECNICA DE VALENCIA

# *Gráficos y parámetros de posición, dispersión y forma de estadística descriptiva*

# *Aspectos prácticos*


| | |
|---|---|
| **Apellidos, nombre** | Prats Montalbán, José Manuel (jopramon@eio.upv.es) |
| **Departamento** | Estadística, Investigación Operativa Aplicadas y Calidad |
| **Centro** | ETS Ingenieros Industriales |

UNIVERSIDAD POLITECNICA DE VALENCIA

# 1 Resumen de las ideas clave

El presente artículo docente pretende mostrar al alumno de Ingeniería Química la conveniencia de realizar una análisis previo de cualquier conjunto de datos disponibles, antes de tomar cualquier tipo de decisión a partir de los mismos; utilizando los gráficos y parámetros más adecuados. Para ello, se toma como base el análisis descriptivo de variables continuas.

# 2 Introducción

En la actualidad, cualquier proceso productivo genera cantidades ingentes de datos relacionados con diferentes variables de calidad (*e.g.* octanaje final de un lote de gasolina, pureza y distribución de una mezcla farmacéutica, tono de una pieza cerámica, porcentaje de impurezas a la salida de un proceso de reciclaje de PET, densidad final de un polímero, etc.), así como variables propias generadas durante el propio proceso de producción (*e.g.* temperatura, presión, tipo y cantidad de materia prima utilizada, proveedor, tensiones e intensidades, etc.).

Por otro lado, estas variables son recogidas a lo largo de diferentes puntos del proceso, con diferentes frecuencias de muestreo y por sensores de distinto tipo; que en muchas ocasiones son volcados a bases de datos de diferente naturaleza. Ello puede dar lugar a diferentes tipos de fallos e incongruencias en las bases de datos que pueden a su vez originar errores de distinto tipo relacionados con fallos en los sensores (generando así datos denominados faltantes), mezclando datos medidos a diferentes escalas, etc.

Toda esta casuística debe prevenir al ingeniero de utilizar estos datos sin más, asumiendo que son del todo fiables desde un principio. Así, unas de las primeras preguntas que debe hacerse una vez dispone de una base de datos, es ¿cómo resumir todos los datos en unos pocos parámetros y gráficos que proporcionen información útil? ¿Son todos los datos válidos? ¿A qué tipo de distribuciones se pueden aproximar los mismos?

# 3 Objetivos

El valor añadido del presente objeto de aprendizaje se encuentra en la descripción de los aspectos prácticos relativos a la utilización de los parámetros y gráficos presentados. Por ello, la manera en que estos parámetros y gráficos se calculan no forma parte de los objetivos ni del contenido del mismo; si bien dichos procedimientos se encuentran detallados en cualquier libro de texto de Estadística básica. De esta manera, se plantean como objetivos los siguientes:

- Hacer reflexionar al alumno acerca de la necesidad de estudiar la naturaleza de los datos antes de cualquier análisis de los mismos
- Reflexionar acerca de los principales parámetros que ayudan a caracterizar la distribución de una variable
- Presentar una serie básica de gráficos para sintetizar la información proporcionada por una variable
- Utilizar cada uno de los parámetros y gráficos de manera adecuada, en función de los datos disponibles, tipo de distribución y fin perseguido (i.e. datos anómalos, normalidad, asimetría, etc.)

# 4 Parámetros de posición, dispersión y forma

Al analizar cualquier serie de datos, una de las primeras actuaciones que se realiza sobre las variables es la compresión de las mismas en un resumen de parámetros que son de utilidad a la hora de obtener información, caracterizarlas y ayudar a establecer hipótesis sobre su distribución. Aparecen así los parámetros de posición, de dispersión y de forma.

## 4.1 Parámetros de Posición

Los parámetros de posición tienen como función ubicar la distribución a lo largo de los valores de la misma. En función de la existencia o no de fenómenos tales como asimetría, datos anómalos, coexistencia de subpoblaciones, etc., diferentes parámetros pueden ser de mayor utilidad. Los principales parámetros de posición son:

- **Media (aritmética)**: es el parámetro de posición más empleado, ya que resume el conjunto de los datos en un único valor que da una idea de la posición global de los mismos. También se puede ver como el centro de gravedad de la distribución, por lo que se ve desplazada en caso de asimetrías respecto del valor intermedio; viéndose afectada por la existencia de datos anómalos, tanto más conforme disminuye el tamaño de muestra.
- **Mediana**: es el valor que deja el 50% de los datos a cada lado, no viéndose afectado por la presencia de datos anómalos (asumiendo un mismo tamaño de muestra). Es una medida alternativa a la media, por lo que la obtención de ambos datos y su comparación puede ser de utilidad, sobre todo en aquellos casos en los que el tamaño de muestra no es grande.
- Moda: es el valor que más se repite de la distribución. En el caso de distribuciones que agrupan a más de una subpoblación, podemos encontrarnos ante el fenómeno de multimodalidad.
- Cuartiles: valores que dejan el 25% (C1), 50% (C2, mediana) y 75% (C3) a la izquierda de una distribución.

## 4.2 Parámetros de Dispersión

Sin embargo, la posición por sí sola no nos da una idea global de los valores que puede tomar la distribución. Imaginemos que el proceso de reciclaje de PET de dos proveedores distintos proporcionad un contenido medio de impurezas en cada saca producida es del 0,3%. ¿Es esta información suficiente a la hora de comparar y seleccionar a uno de los dos proveedores?

A la vista de los gráficos presentados en la Figura 1, parece claro que de manera intuitiva, incluso cumpliendo con la tolerancia máxima admisible de impurezas en una saca, cualquiera elegiría al proveedor 1. Para ello, además de la posición, tenemos en cuenta la dispersión de los datos, eligiendo aquél que presenta la menor de los dos proveedores.

Por todo ello, parece bastante claro que a la hora de caracterizar una distribución (y poder tomar decisiones, no de manera intuitiva, sino en base a evidencias estadísticas) es necesario conocer, no sólo los parámetros de posición, sino también los de dispersión.

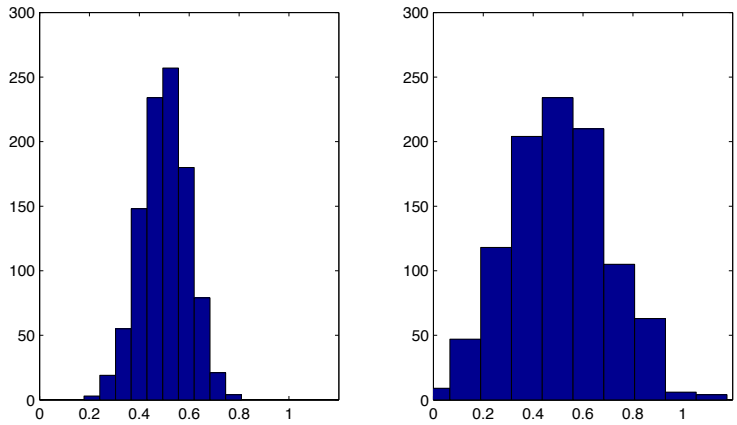


**Figura 1.** Distribución del porcentaje de impurezas en las sacas de PET producidas por el proveedor 1 (izquierda) y el proveedor 2 (derecha)

Los principales parámetros de dispersión son:

- Rango. Es la diferencia entre el valor máximo y el valor mínimo del conjunto de datos. Resulta útil en el caso de muestras pequeñas, pero tiene el peligro de verse afectado por la presencia de datos anómalos, ya que por definición es la diferencia entre los datos extremos de la muestra.
- **Varianza. Desviación típica.** La dispersión de los datos respecto de su centro de gravedad se puede resumir a partir del valor promedio del cuadrado de las deviaciones de cada dato respecto de la media. Es la medida de dispersión más importante ya que, en general para el tipo de datos que el ingeniero utiliza, la media suele ser un buen parámetro de posición. La desviación típica es simplemente la raíz cuadrada positiva de la varianza, que se calcula para tener una medida de dispersión en la misma dimensión que la variable medida.
- **Intervalo intercuartílico.** En aquellos casos en que existe una clara asimetría de los datos, la media deja de ser un buen indicador de posición y por tanto la varianza tampoco. En estos casos, el parámetro adecuado a utilizar es el intervalo intercuartílico, el cual se define como la diferencia entre el tercer y primer cuartil, que proporciona una idea de la dispersión del conjunto central de los datos (50% central), dejando fuera las zonas que se ven más afectadas por fenómenos tales como la asimetría o los datos anómalos.

## 4.3 Parámetros de Forma

En entornos industriales, al igual que en muchos otros, una de las distribuciones más comunes es la distribución Normal, caracterizada por una forma de campana, simétrica y con frecuencias decrecientes conforme nos alejamos de la media. Dado que existe una gran metodología desarrollada alrededor de esta distribución, resulta conveniente estudiar la posibilidad de aproximar la distribución de una serie de datos a una Normal. Por ello, además de los parámetros de posición y dispersión estudiados, se presentan dos nuevos parámetros de forma

- **Coeficiente de Asimetría.** El coeficiente de asimetría mide cuánto más de alejados respecto de la media se encuentran, en promedio, una serie de datos. Para ello, se calcula el promedio del cubo de las desviaciones de los datos respecto de la media muestral, de manera análoga al cálculo de la varianza. Valores positivos de asimetría indican una tendencia a tener valores más alejados por la derecha que por la izquierda, mientras que una asimetría negativa indica lo contrario. Con el fin de establecer hasta cuándo una distribución tiene valores de asimetría tolerables, un criterio es utilizar el intervalo [-2, 2] para el valor del coeficiente de asimetría estandarizada, el cual se calcula como el cociente del coeficiente de asimetría respecto del cubo de la desviación típica.
- **Coeficiente de Curtosis.** Por último, una manera de estudiar la posible existencia de datos anómalos, o bien de datos censurados (e.g. en un proceso de producción donde se superan las tolerancias del cliente, una manera de cumplir es eliminar toda la producción fuera de las mismas, ver Fig 2); es el cálculo de coeficiente de curtosis. Se calcula como el promedio de la potencia cuarta de las desviaciones de los datos respecto de la media muestral, si bien ciertos programas comerciales como *Statgraphics* restan a dicho valor la constante 3 con el fin de tener el valor final centrado en "0". De esta manera, valores positivos de curtosis vienen asociados a distribuciones leptocúrticas, en las cuales valores alejados de la media aparecen con una frecuencia mayor de lo esperada (indicativos por tanto de datos anómalos); mientras que valores negativos de curtosis se dan cuando valores alejados de la media aparecen con una frecuencia menor de la esperada (indicativos por tanto de datos censurados). Al igual que en el caso de la asimetría, el intervalo

[-2, 2] se utiliza para determinar has cuándo es razonable asumir que no hay curtosis en una distribución. Dicho intervalo se calcula para el valor del coeficiente de asimetría estandarizada, el cual se define como el cociente del coeficiente de curtosis respecto de la potencia cuarta de la desviación típica menos 3 (ya que en dicho caso centramos alrededor del cero).

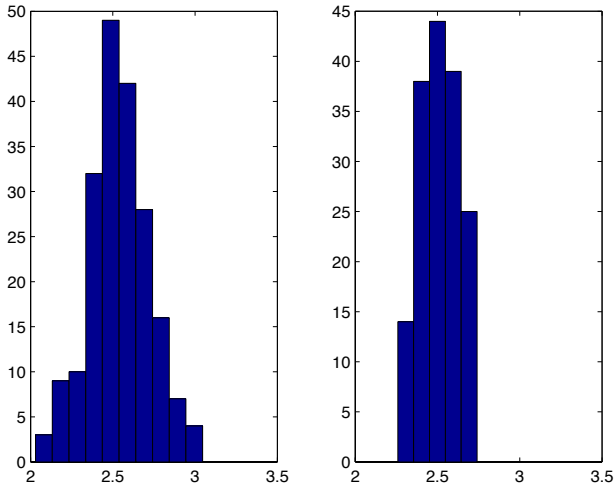


**Fig. 2.** Distribución no censurada (izquierda) y censurada (derecha) para valores menores de 2,25 y mayores de 2,75, de una distribución normal de media 2,5 y desviación típica de 0,2. En el primer caso, la curtosis estandarizada (y centrada) presenta un valor de 0,26, mientras que en el segundo caso (derecha), dicho valor es de -0,84

# 5 Gráficos

Aunque el resumen de una serie de datos en unos pocos parámetros de posición, dispersión y forma puede resultar muy útil, la utilización únicamente de éstos con el fin de establecer la caracterización de una distribución no es recomendable. Al contrario, la visualización de los datos resulta de gran ayuda a la hora de determinar qué tipo de distribución pueden seguir, así como la existencia de asimetrías, anomalías y/o censura en los datos. A continuación se presentan los gráficos más utilizados.

## 5.1 Histograma

El histograma es la representación gráfica de la tabla de frecuencias. Representa la forma de la distribución "vista desde enfrente", i.e. es el alzado de la distribución. Ayuda a caracterizar muy bien la distribución, ayudando a detectar:

a) Frecuencia anómala de un valor
b) Medidas inconsistentes (diagrama en forma de peine)
c) Mezclas de poblaciones distintas
d) Asimetría
e) Datos artificialmente modificados (censuras)
f) Datos anómalos

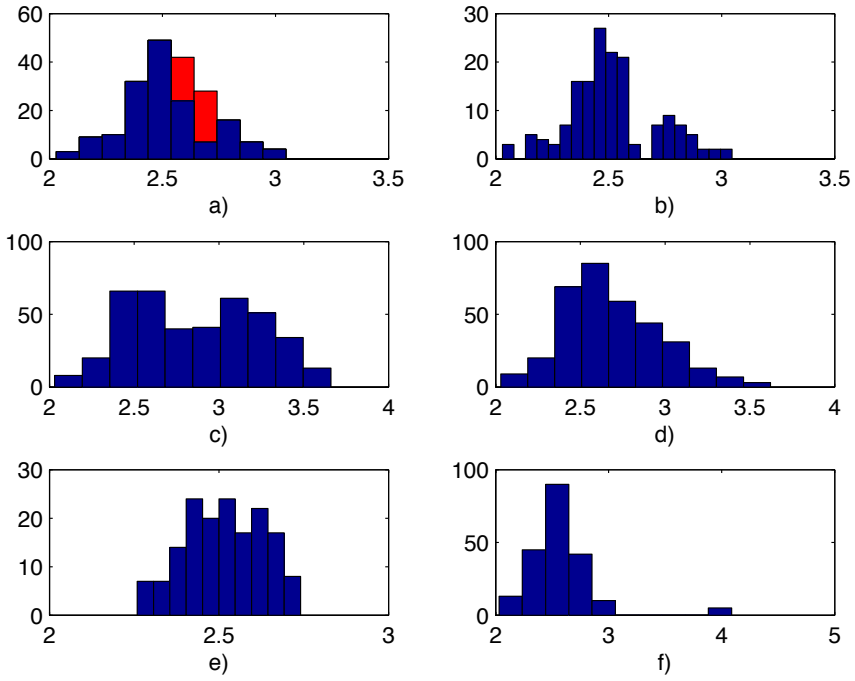


**Figura 3.** Representación de los tipos de anomalías que ayudan a dectectar los histogramas.

Sin embargo, el histograma presenta una serie de inconvenientes tales como:

- Necesidad de una gran cantidad de datos (mínimo recomendable de 50)
- El número de intervalos y el rango de representación utilizados puede variar la forma de la distribución

Por ello, su utilización no está exenta de subjetividad, siendo por tanto recomendable cierta experiencia para una adecuada utiliación. Las representaciónes gráficas de las distribuciones de las Figuras 1 y 2 son simplemente los histogramas de las mismas.

## 5.2 Diagrama Caja-Bigote (Box & Whisker Plot)

Se trata de una representación gráfica robusta de una serie de datos, a partir de los valores de los cuartiles de los mismos. Se compone de una caja (Box) dibujada entre el

primer y tercer cuartil, dentro de la cual se sitúan tanto la mediana como la media; y de dos bigotes (Whisker, uno a cada lado de la caja) cuya dimensión máxima es 1,5 veces el intervalo intercuartílico. Los datos que quedan más allá de estos bigotes se consideran datos anómalos. El diagrama Caja-Bigote se puede entender como la planta de una distribución, i.e. la visualización cenital (desde arriba) de la misma, con indicaciones de las posiciones de los cuartiles y media, así como de la dispersión de los datos en función de dichos cuartiles. De este modo, la caja alberga el 50% central de los datos, y los bigotes proporcionan una idea de hasta dónde deberían llegar os datos pertenecientes a dicha distribución.

Así, el diagrama Caja-Bigote permite una rápida detección de posibles datos anómalos y, por comparación de la anchura de ambos lados de la caja (respecto de la mediana) y de longitud de los bigotes, de posible presencia de asimetrías, siempre que esas comparaciones sean coherentes: e.g. si la caja es más ancha a la derecha de la mediana, y el bigote derecho de mayor longitud, ello puede ser indicativo de asimietría positiva (Fig. 4).

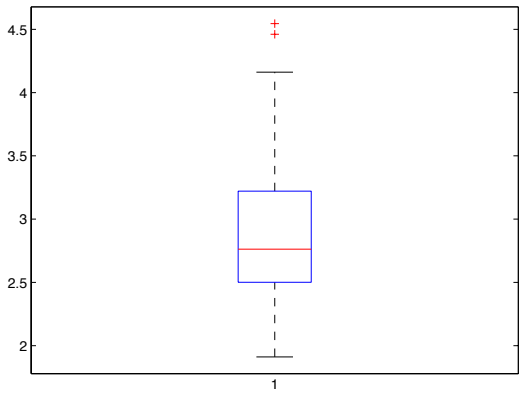


**Figura 4.** Diagrama Caja-Bigote de una serie de puntos. El hecho de que la caja y el bigote derechos (superiores) sean mayores que los izquierdos (Inferiores), así como la presencia de dos datos fuera de límites, si bien no muy alejados de la distribución, es indicativo de la posible existencia de asimetría positiva.

## 5.3 Papel Probabilístico Normal

El último gráfico que se presenta en este artículo docente es el papel probabilístico normal. Se trata de una representación de la serie ordenada de datos, y escalada de forma que, cuando éstos siguen aproximadamente una distribución normal, se alinean a lo largo de una recta. Permite detectar datos anómalos, cuando éstos se alejan de la recta asociada al conjunto de datos (Fig. 5 a), asimetrías, cuando los datos

presentan una curvatura a derechas (positiva, Fig. 5 b) o a izquierdas (negativa), e incluso mezcla de poblaciones (Fig. 5 c).

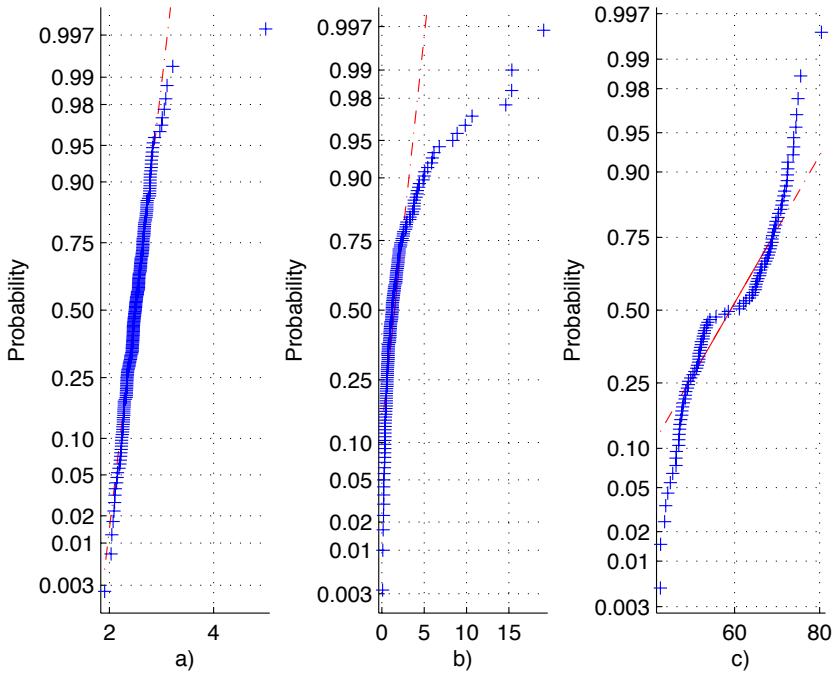


**Figura 5.** Papel probabilístico normal para series de datos asociadas a los casos de: datos anómalos (a), asimetría positiva (b) y mezcla de poblaciones (c).

# 6 Cierre

El presente objeto de aprendizaje ha presentado los aspectos prácticos de la utilización de los parámetros y gráficos que ayudan a caracterizar una distribución. Las cuestiones teóricas relativos a los mismos se encuentran en libro de texto tales como los recomendados en la Bibliografía.

# 7 Bibliografía

[1] Peña, D. (2001). *Fundamentos de Estadística*. (Ed.) Alianza Editorial, S.A. Madrid. ISBN: 84-206-8696-4.

[2] Romero, R y Zúnica, L.R. (1993). Estadística (Proyecto de Innovación Educativa). SPUPV-93.637.