2.- GENERALIDADES

04
Conceptos principales de Estadística

1 Conceptos generales

Al fabricar un producto a partir de una determinada fórmula o consigna en condiciones ideales obtendríamos el mismo resultado para una determinada característica si el funcionamiento de toda la cadena productiva (recursos materiales y humanos) actuase sin cambios significativos para los fines prácticos. Pero, en realidad, al activar un proceso industrial se producen desviaciones de la fórmula original debido a oscilaciones inevitables (de mayor o menor amplitud) en el funcionamiento de los equipos y en el comportamiento de los operadores. Esta variabilidad inevitable en la práctica, aunque siempre reducible a valores menores, es la base de la necesidad de emplear la herramienta estadística para asegurarnos de que, cuando comprobamos mediante muestras las características de la producción, los diferentes resultados obtenidos se deben al azar en el campo de variabilidad esperable y no a que estamos produciendo o recibiendo un producto de características distintas e inaceptables.

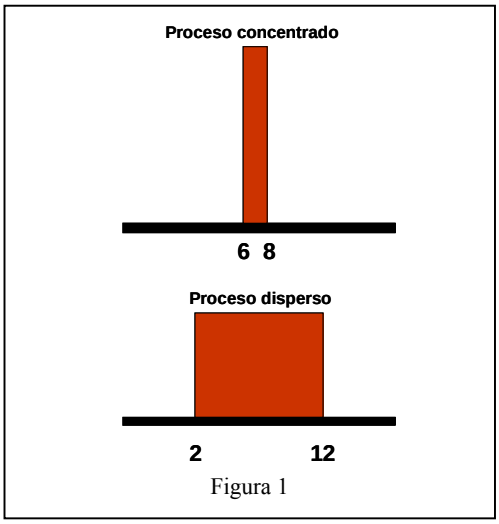


Figura 1

Hoy en día se entiende, en gran medida, como calidad industrial a la capacidad, primero, de mantener bajo control estadístico los procesos y, segundo, de saber actuar sobre las causas de variabilidad para reducirla todo lo posible como fuente de éxito empresarial.

El hormigón como todo proceso industrial sufre la variabilidad de sus características. Y si se concreta en la resistencia o la durabilidad, sólo actuando sobre las fuentes (humedad de los áridos, precisión de las balanzas, previsión de los errores humanos, etc.) es posible reducir la dispersión a valores que le permitan a la industria mantener bajo control su costes y a los compradores obtener un producto satisfactorio. Cuando un industrial del hormigón mantiene bajo control su producción está en condiciones de concertar planes de

control de recepción seguros cuya incertidumbre residual sea debidamente compartida por los intereses en juego. Este objetivo hace necesario que las herramientas estadísticas que se utilicen sean eficaces y transparentes para todas las partes.

En la figura 1 se representa un proceso concentrado como aquel que apuntando a obtener un producto con un determinada característica de valor 7 genera valores entre 6 y 8; y un proceso disperso, como aquel que apuntando con su fórmula de fabricación igualmente a 7 genera (por falta de control) valores entre 2 y 12. En ambos casos se representa el fenómeno con un rectángulo más o menos estrecho. El hacerlo así quiere decir, en términos estadísticos, que la probabilidad elemental en el entorno de todos los valores posibles (en cada uno de los dos procesos) es uniforme. Sin embargo, esta uniformidad de la densidad de probabilidad raramente sucede en los procesos industriales

Esto, que es verdad para los valores de las caras de un dado, no sucede en la práctica industrial. Los valores de resistencia de un hormigón fabricado con una misma dosificación nominal, no sólo se dispersan en torno al valor teórico, sino que lo hacen con probabilidades elementales distintas y paulatinamente menores a medida que se alejan de dicho valor teórico[1]. Decir probabilidades es equivalente (simplificando) a decir frecuencia, o mejor decir frecuencia relativa. O sea, en los procesos industriales los valores que representan una característica aparecen con más frecuencia entorno al valor medio que en zonas más alejadas. Esta distinta probabilidad es representada con curvas llamadas funciones de densidad, entre las que la más conocida por acercarse mejor a la realidad (y ser más conocida), es la curva Normal.

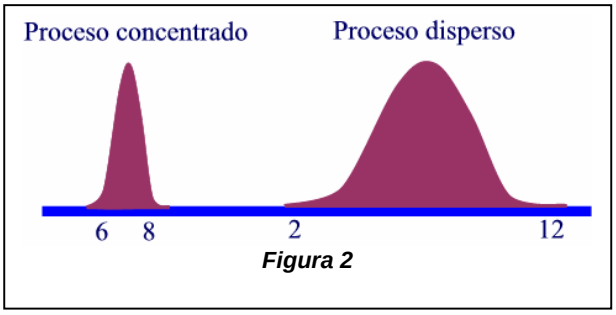


Figura 2

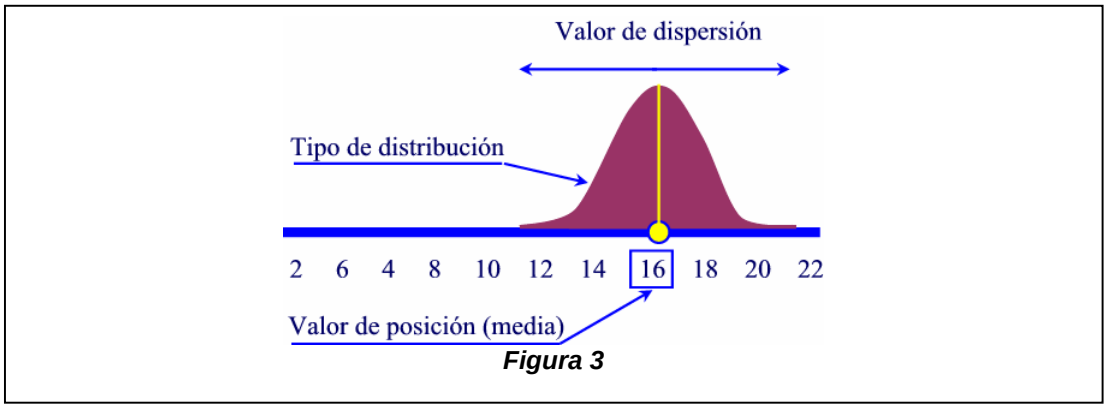


Figura 3

[1] El valor teórico de resistencia de un hormigón fabricado con una determinada dosificación coincide en gran medida con el valor medio de todos los valores aparecidos como consecuencia de la variaciones que, en la dosificación inicial, introducen las condiciones de los componentes y de las instalaciones industriales

Hoy en día se sabe, gracias a la acumulación de datos y la potencia de proceso de los ordenadores que este modelo necesita algunas correcciones. Pero para los fines de este capítulo la aceptaremos sin crítica. Así pues, debido al modo en que se distribuyen los valores la representación de un proceso concentrado y uno disperso se parece más a la figura 2

De lo dicho en el párrafo anterior se puede resumir que de una producción industrial concreta cabe esperar una serie de valores que quedan incluidos en un determinado intervalo. La información fundamental para mantener bajo control una producción está formada por:

1) la forma de la distribución de los valores
2) un valor de posición que fija al conjunto de valores esperables en un determinado punto del espectro de todos los valores posible de esa característica y
3) un valor de dispersión que permite conocer la extensión del intervalo dentro del cual están todos los valores físicamente posibles para ese proceso. En estadística decir todos es decir valores en torno al 99 % para dar cabida a la aparición de atípicos por algún fallo puntual del sistema productivo.

2 Datos

Los datos son el resultado de las observaciones. Las observaciones se producen sobre la naturaleza, los procesos industriales y sus resultados o sobre los fenómenos sociales. Es decir sobre todo aquello que tiene interés para el conocimiento y la acción humanas. En esta monografía se utilizarán, en general, ejemplos relativos al hormigón, tanto de métodos de aplicación general, como de métodos específicamente pensados para su aplicación a los procesos industriales de la construcción de estructuras de hormigón y su control. Las observaciones de estos fenómenos nos proporcionan valores que expresan las dimensiones de las magnitudes de un proceso. Una observación es siempre una medición con un resultado cuantitativo. Ya sea un mero recuento o la aplicación de un sofisticado instrumento. Los procesos industriales definidos observados (medidos) proporcionan masas de datos dentro de intervalos predecibles cuando están bajo control y masas de datos erráticas cuando están fuera de control. Las instalaciones industriales bajo control, salvo valores aberrantes, proporcionan masas de datos, que no sólo se mantienen en un intervalo determinado, sino que se distribuyen con regularidad determinada que suele responder a modelos con desviaciones aceptables, lo que permite su tratamiento matemático para la extracción de conclusiones prácticas.

Una serie de datos o conjunto de datos se puede expresar en la forma x_i, $i=1,2,...n$, en la que x_i representa a todos los datos del conjunto. También se puede expresar en la forma $\{x_1, x_2, ..., x_n\}$. En ambos caso la letra n señala el número total de datos se repitan, o no, los valores. El subíndice indica una veces el orden temporal en que aparecen los datos y, otras, sólo es un elemento diferenciador. Es decir, el subíndice 1 indica al primer dato pero no, necesariamente, al menor de ellos. Igualmente, el subíndice n indica al último dato, pero no, necesariamente, al mayor de ellos. Cuando nos referimos a la serie de valores prescindiendo de las repeticiones de los mismos se la representa por x_α, $\alpha=1,2,...,\nu$. Siendo ν el número de valores sin repetición y α el subíndice que diferencia a un valor de otro . Es decir, en una serie en la que hay más datos que valores n será mayor que ν.

2.1 Tipos de datos

Los datos son los resultados de las observaciones sobre un fenómeno que se quiere estudiar. Hay dos grandes categorías de datos:

a) datos experimentales (*experimental data*)
b) datos no experimentales (*observational data*)

Los datos experimentales, propios de la ciencia aplicada, son generados de forma consciente y programada por el experimentador. El diseño de experimentos o el control de calidad son ejemplos de la obtención de datos experimentales. En un caso para resolver cuestiones técnicas sobre el comportamiento de un material o un elemento y, en el otro, para decidir la aceptación o rechazo de un material o elemento .

Los datos no experimentales, propios de las ciencias sociales, surgen sin la colaboración activa del observador en su generación, pero su recogida es planificada por el estadístico al cargo. La Econometría es un ejemplo de ciencia que utiliza datos no experimentales.

Como se ve, el criterio utilizado para distinguir los tipos de datos es la actitud activa o pasiva del que los utiliza (experimentador o analista) para su generación. Esta es, como todas, una separación artificial que resulta útil pero que no siempre tiene la precisión que aparenta.

Los datos manejados en el control de calidad de los productos industriales son experimentales porque son generados mediante experiencias de laboratorio sobre las muestras extraídas del flujo de la producción de forma programada.

Una abstracción útil para aplicar el lenguaje matemático al análisis de los datos disponibles es la de asignar la idea de variable a la característica del fenómeno en estudio. Desde este punto de vista, los datos son concreciones de las diferentes variables observadas. Hay dos tipos de variables:

a) Variables cuantitativas o métricas
b) Variables cualitativas o no métricas

Las variables cuantitativas toman valores numéricos. Un ejemplo de variable cuantitativa es la resistencia a compresión del hormigón. En efecto, la variable f_c (resistencia del hormigón) puede tomar muchos valores numéricos para un determinado hormigón.

La variables cualitativas especifican atributos o cualidades del observable. En general estas variables no toman valores numéricos propiamente dicho, pero, muy a menudo, son codificadas con números. Un ejemplo de variable cualitativa es el color de un hormigón arquitectónico. En este caso, el hormigón sólo puede tener o no el color especificado mediante una comparación con un patrón que se expresa como una cualidad y no como un número.

Las variables cuantitativas pueden ser de dos tipos:

a) Continuas
b) Discretas

Las **variables continuas** pueden tener valores infinitamente próximos entre sí.

EJEMPLO . La variable *«Resistencia del hormigón»* puede alcanzar cualquier valor: 23 - 23,4 – 23,423 – 23,4236 – 23,42362 MPa, en una serie infinita cuya única limitación es la resolución del instrumento utilizado para la determinación de la resistencia.

Las **variables discretas** sólo pueden tomar determinados valores de una serie.

EJEMPLO La variable *«Número de la cara superior de un dado»* sólo puede tomar los valores 1,2,3,4,5 y 6.

Los criterios de clasificación de las variables en discretas y continuas muestra, también, cierta ambigüedad si se tiene en cuenta que los valores experimentales no puede aproximarse infinitamente entre sí debido a la resolución limitada de los instrumentos de medida. Así una resistencia de 23,4 MPa puede ser seguida de otra de 23,5 MPa con un tipo determinado de aparatos, pero con otros de mayor resolución podríamos obtener valores de 23,46 y de 23,47. Pero, en ambos casos, podemos observar que entre los valores más próximos que permiten los instrumentos *no existen* otros valores observables. La restricción comentada hace que, en la práctica, todas las variables que manejamos en el control de calidad puedan ser consideradas como discretas, aunque sean esencialmente discretas. Sin embargo, el tratamiento matemático de las distribuciones de las variables aleatorias discretas es más complicado que el de las continuas y, por eso, a la mayoría de las distribuciones de variables **conceptualmente** continuas se las trata como tal en los cálculos estadísticos. La resistencia del hormigón es un ejemplo claro de variables cuya distribución es conceptualmente continua y cuyo tratamiento estadístico responde a esta naturaleza dada el número relativamente alto de cifras significativas. Por otro lado, la consistencia del hormigón es un ejemplo de variable conceptualmente continua que, sin embargo, se trata como una variable discreta dada la baja resolución del procedimiento de medida y al redondeo normativo. Estas decisiones no suponen ningún error teórico ni menor rigor matemático. Lo que se puede decir igualmente cuando una variable conceptualmente discreta produce los datos de tal modo que es posible su tratamiento como variable continua.

Si se introduce el tiempo en el análisis los datos pueden ser clasificados en tres categorías:

a) De corte transversal
b) De corte longitudinal (series temporales)
c) Datos de panel

Los datos de corte transversal son resultados de observaciones realizadas en un instante o período de tiempo muy corto. Son datos de este tipo los que se utilizan, en general, para juzgar un lote de hormigón hormigonado en una mañana en una obra determinada o en la producción de una central de hormigonado durante, pongamos, una semana.

EJEMPLO En una obra con estructura de hormigón se ha llevado a cabo una sesión de hormigonado que ha empleado 36 cubas durante dos días. La resistencia característica especificada del hormigón es de 35 MPa. El responsable del control de recepción del lote

correspondiente de hormigón ha mandado extraer muestras de seis camiones. De cada muestra se han elaborado dos probetas cilíndricas para determinar la resistencia de cada amasada para la edad de 28 días conforme a las normas UNE correspondientes. Transcurridos los días especificados se procedió al ensayo de las seis parejas de probetas con los siguientes resultados en MPa:

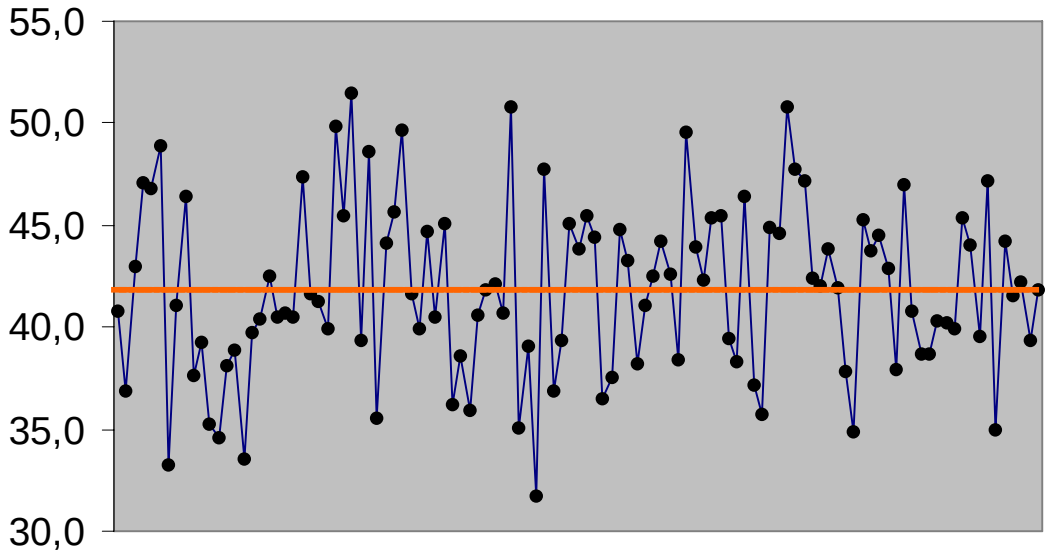
32,4 – 31,6
29,5 – 30,5
43,2 – 41,8
36,8 – 38,3
39,5 – 37,1
41,2 – 44,1

Estos datos de corte transversal permiten proceder a determinar si las muestras son aceptables mediante el recorrido relativo de cada pareja y la resistencia estimada del lote mediante el segundo estimador de la EHE.

Los datos de corte longitudinal se refieren a las observaciones de una variable realizadas en sucesivos instantes en un período de tiempo relativamente largo. Un ejemplo son los datos obtenidos en el control de producción

Los datos de panel combinan datos de corte transversal con datos de corte longitudinal. La variable a medir es única y se realiza sobre los mismos individuos en los mismos instantes de tiempo. Un ejemplo los constituye las series temporales de la resistencia del hormigón de las distintas centrales de hormigón de un grupo empresarial.

EJEMPLO En una central de hormigón preparado se lleva a cabo el control de producción mediante la extracción de una muestras de cada 100 m^3. Con estas muestras se elaboran probetas cilíndricas para la determinación de la resistencia a 7 días de edad. Dado que la planta produce 600 m^3 al día se cuentan con 6 resultados cada día de media. Al cabo de un año se han obtenido 1050 resultados cuya expresión gráfica es la siguiente para un tramo de un mes determinado:



La tendencia del control de calidad moderno es la de considerar la influencia del tiempo explícitamente en el análisis de los datos. En la actualidad se hace de un modo implícito y poco claro.

2.2Tipos de análisis de datos

Hay dos tipos de análisis de datos:

a) Exploratorio, que es aquel que aspira a responder a preguntas de carácter general, tales como qué estructura tienen los datos o si hay datos anómalos
b) Confirmatorio, que es aquel que responde a preguntas más concretas, tales como la existencia, o no, de subgrupos con diferencias relevantes; si existen relaciones de dependencia entre variables; si se da alguna tendencia en los datos de corte longitudinal o se es posible estimar alguna variable.

Hay dos perspectivas para su aplicación:

a) Descriptiva, que es aquella que busca enunciados o afirmaciones sobre los datos que se tienen. Generalmente se cuenta con todos los datos posibles o, bien se circunscribe el análisis al conjunto de los datos disponibles. En general se aplica este enfoque a las poblaciones finitas y el resultado de su aplicación es el conocimiento del conjunto a través de sus parámetros y la interpretación correspondiente. También se aplica a las muestras y a series en general.
b) Inductiva o inferencial, que es aquella que trata de obtener conclusiones de carácter general a partir de los datos con que se cuenta. En general, se cuenta con un número relativamente bajo de datos respecto de los posibles.

Los datos, por sí mismos no proporcionan información útil si no son tratados estadísticamente. La realidad es oscilante y, en general, contar con un solo dato es no tener prácticamente ninguna información, porque cualquier proceso industrial produce, a igualdad de intención (consigna o dosificación), series de datos. En Construcción es necesario realizar mediciones de las características temporales, geométricas, físicas o químicas de los productos y procesos. Estas operaciones proporcionan conjuntos de datos que es necesario analizar cuidadosamente para que resulten útiles.

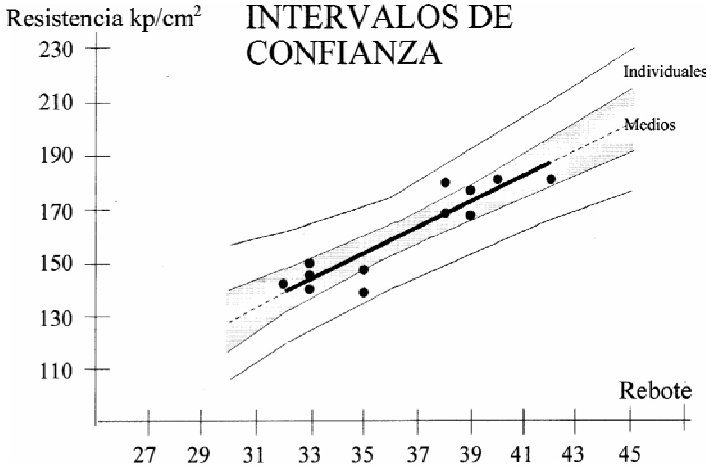
La relación entre el enfoque descriptivo y el enfoque inductivo tiene alguna ambigüedad. Así, por ejemplo, la media aritmética cuando se aplica a los valores de una población finita es un parámetro de la misma perfectamente definido y fijo. Pero cuando se aplica a una muestra es un valor oscilante que cambia meramente con la extracción de una nueva muestra. Por eso, en este caso, hablamos de una estimación o aproximación al parámetro poblacional. La nomenclatura convencional utiliza letras griegas para los parámetros poblacionales y letras latinas para los estimadores o estadísticos.

Uno de los objetivos más importantes del análisis de datos, desde el punto de vista inductivo, es mejorar el conocimiento de los parámetros poblacionales, ya que éstos gobiernan el comportamiento presente y futuro de los procesos productivos sometidos a control. El empleo de estimadores de los parámetros poblacionales es por razones práctico – económicas.

El planteamiento multivariante del análisis de datos es interesante porque pone de manifiesto dos tipos de problemas de interés: las dependencia y la interdependencia de las variables. En el caso del ejemplo 1.1.5, se puede considerar por razones teóricas que la variable X_4 es una variable dependiente mientras que las demás se pueden considerar independientes. El análisis de dependencia trataría de establecer un modelo explicativo de una respecto de las otras. Un ejemplo, aunque no el único, de este tipo de análisis es el modelo de regresión.

EJEMPLO Dados los valores de resistencia de un hormigón y los del índice de rebote del mismo, es posible un análisis de dependencia mediante el establecimiento de la línea de regresión correspondiente:

Índice de rebote	Resistencia
32	141
33	138
33	149
33	151
35	140
35	142
38	171
38	185
39	169
39	172
40	180
42	179



De este modo es posible estimar valores de la resistencia del hormigón a partir de los correspondientes índices de rebote.

2.3 Frecuencia de valor y frecuencia de clase

Más arriba se indicaba la diferencia entre valor y dato. Varios datos pueden tener el mismo valor. Por ejemplo, el número tres (valor) puede darse varías veces en la tirada de un dado. Cada una de las veces es un dato. Llamamos frecuencia al número de veces que se da el valor.

El concepto de frecuencia es de aplicación directa en los siguientes tipos de variables:

- Categóricas (ordinales y nominales). Por ejemplo, en la clase de resistencia (variable ordinal) de una serie de obras se pueden dar 4 datos de hormigones C20 y 3 de C35. En el tipo de cemento utilizado en una provincia se pueden dar 43 cementos CEM I
- Métricas discretas. Por ejemplo, en relación con la consistencia del hormigón en una obra, el valor 6 cm se puede dar en 22 hormigones y el valor 8 cm en 12.

Como se verá más adelante, en las variables continuas (Vg. la resistencia del hormigón) el grado de *presencia* se mide más eficazmente con el concepto de **densidad** de frecuencia**.** La razón es que, en puridad, en una variables continua *la probabilidad de obtener valores repetidos es nula.* Si se dan repeticiones es por la resolución limitada de los instrumentos de medida o por que llevamos a cabo algún tipo de redondeo premeditadamente.

Sin embargo, cuando se cuenta con muchos datos procedentes de una variable continua es posible utilizar también el concepto de frecuencia. Si se procede a la agrupación de datos mediante la división del dominio de definición de la variable *X* (ámbito en el que puede variar) en intervalos o **clases**, de modo que la frecuencia asociadas a una clase sería igual al número de resultados individuales contenidos en la misma.

Los intervalos en que se divide el dominio de la variable se identifican por sus extremos. El número total de intervalos creados se designa ν. Habrá pues ν intervalos para llevar a cabo la agrupación y se utilizará el índice genérico α (que toma valores desde 1 a ν) para designar a un intervalo genérico, concretamente el α*-ésimo.* Los extremos de los intervalos se designan L_0, L_1, L_2, ...,L_ν, pues si hay ν intervalos tiene que haber ν+1 extremos. Por el modo de designar a los extremos, el intervalo α*-ésimo* es el que tiene por extremo izquierdo $L_{\alpha-1}$ y por extremo derecho L_α al superior. En una variable discreta o categórica, ν, representa al número de valores (no confundir con el número de datos). Para asimilar mejor la variable continua a las categóricas o discretas en la utilización del análisis de frecuencia, cada intervalo de clase es representado por un único valor x_α que suele coincidir con el centro entre los extremos del intervalo y al que se denomina **marca de clase**:

$$x_\alpha = \frac{L_{\alpha-1} + L_\alpha}{2}; \quad (\alpha = 1,2,...,\nu)$$

La amplitud h_α del intervalo viene dada por:

$$h_\alpha = L_\alpha - L_{\alpha-1}; \quad (\alpha = 1,2,...,\nu)$$

Las marcas de clase siguen una secuencia ordenada tal que $x_\alpha \leq x_{\alpha+1}$. Puede darse la circunstancia de que una marca de clase cualquiera x_α no pertenezca a la serie original de los datos. Si no se conoce el dominio de definición de la variable es necesario calcular el **recorrido** o **rango** de la serie de datos:

$$r_X = máx\{x_i\} - mín\{x_i\},$$

para definir con su ayuda el número de intervalos y su amplitud.

Los intervalos son abiertos por el extremo inferior y cerrados por el extremo superior. De este modo si el valor de un dato coincide con el extremo inferior o superior de un intervalo de clase hay un criterio para su inclusión en una u otra.

De este modo la serie original de datos,

$$\{x_i, i = 1,2,...,n\},$$

Se convierte en la serie de datos agrupados,

$$\{x_\alpha; n_\alpha; \alpha = 1,2,...,\nu\}.$$

El proceso completo para aplicar el análisis frecuencial a la serie de datos es el siguiente:

1. Establecimiento del dominio de definición de la variable. En caso de que no se conozca el dominio de definición, cálculo del rango de la serie de datos
2. Fijación del número de intervalos de clase ν
3. Fijación de las amplitudes de los intervalos de clase
4. División del dominio de definición de la variable en ν intervalos de clase

Hoy en día el agrupamiento de datos no debe emplearse para la determinación de las medidas de centralización o de dispersión, dado que los ordenadores facilitan la labor. El interés de la agrupación reside en la representación gráfica que supone y en su carácter esquemático para los informes.

La frecuencia aplicada hasta ahora es la denominada absoluta n_α. Cuando este valor se divide por el número total de datos observados se obtiene la frecuencia relativa f_α de las series originales de las variables discretas y categóricas y de los datos agrupados de las variables continuas:

$$f_\alpha = \frac{n_\alpha}{n}$$

Donde n_α es la frecuencia absoluta del valor en el caso de variables categóricas o discretas y de la marca de clase en el caso de datos agrupados de una variable continua.

Dado que

$$\sum_{\alpha=1}^{\nu} n_\alpha = n\,,$$

donde ν es el número de valores en el caso de variables categóricas o discretas y de intervalos de clase en el caso de datos agrupados se ha de cumplir que:

$$\sum_{\alpha=1}^{n} f_a = 1$$

La representación en forma de gráfico de barras es aplicable tanto a los datos agrupados por clases como a los datos de un serie original de una variable discreta.

Sin embargo, cuando se trata de una variable continua es más frecuente hacer la representación mediante un histograma de frecuencias. La diferencia estriba en que en el histograma se representan, en abscisas, los extremos de los intervalos de clase y en las ordenadas el valor de la frecuencia absoluta n_α dividida por la amplitud del intervalo h_α , es decir:

$$d_\alpha = \frac{n_\alpha}{h_\alpha}$$

Este método gráfico tiene la ventaja de que consigue una representación de los datos que se aproxima a la representación continua del modelo de distribución de la variable, como se verá más adelante con el concepto de densidad de probabilidad.

Naturalmente, en el histograma la suma de las áreas de las rectángulos es igual a *n*, es decir, el número total de datos. Una alternativa es el histograma de frecuencias relativas. En éste, la altura de los rectángulos es

$$\delta_\alpha = \frac{f_\alpha}{h_\alpha}$$

y mide la densidad de frecuencia relativa. En este histograma la suma de las áreas los rectángulos es, obviamente, igual a la unidad. Esta es una mejor aproximación al concepto de densidad anunciado. En efecto, si aumenta el número de intervalos o clases y su amplitud cada vez es menor, la línea envolvente del histograma de frecuencias relativas toma la forma de una curva continua para la que la altura en un punto determinado es el valor de δ_α para un intervalo de clase infinitamente pequeño.

Otro tratamiento de interés de las frecuencias, tanto absolutas como relativas, son los valores acumulados de las mismas. De este modo se obtiene para cada valor la suma de las frecuencias, absolutas o relativas, del valor dado y todos los que le preceden.
La frecuencia absoluta acumulada, N_α, se obtiene sumando sucesivamente las frecuencias de cada intervalo o valor. La frecuencia relativa acumulada, F_α, es el resultado de dividir la frecuencia acumulada por el número total de datos. Es decir,

$$N_\alpha = n_1 + n_2 + \ldots + n_{\alpha-1} + n_\alpha$$

$$F_\alpha = \frac{N_\alpha}{n}$$

$$F_\alpha = f_1 + f_2 + ... + f_{\alpha-1} + f_\alpha$$

3 Medidas de centralización o de posición

La medidas de posición de una serie de datos son valores sintéticos que intentan reducir toda la serie a uno solo representativo del conjunto de datos en el dominio de definición de la variable. Las más importantes son:

- La media aritmética
- La mediana
- La moda

Tienen la función de situar o posicionar a la masa de datos en el eje de las unidades de la variable de que se trate. Unos lo hacen con el valor de número de unidades medio (la media aritmética) , otros con el valor que divide a la masa de datos en dos partes de igual frecuencia (la mediana) y otros con el valor que más veces se repite entre los observados (la moda).

3.1 La media aritmética $\bar{x}$

La media aritmética tiene la siguiente expresión

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ .$$

Es decir, es la suma de todos los resultados divididos por el número total de los mismos *n.* Si los datos se presentan agrupados, ya sea por repeticiones del mismo valor o porque se trate de datos agrupados por clases, la fórmula es la siguiente:

$$\bar{x} = \frac{1}{n}\sum_{\alpha=1}^{\nu} n_\alpha x_\alpha = \sum_{\alpha=1}^{\nu} f_\alpha x_\alpha$$

En las variables discretas no agrupadas por clases, las dos fórmulas proporcionan el mismo valor para la media. Pero no ocurre igual con las variables agrupadas en clases (discretas o continuas). En este caso es preferible calcular la media aritmética con los datos originales si están disponibles.

3.2 La mediana $\tilde{x}$

La mediana es el valor cuya frecuencia relativa acumulada es 0,50. Una vez ordenados los datos conforme a la serie $\left\{x_1' \le x_2' \le ... \le x_n'\right\}$ (las comillas la distinguen de la serie original, done los subíndice indican el orden de aparición de los datos y no el orden de menor a mayor como aquí), cuando *n* es impar la mediana coincide con el valor $x'_{(n+1)/2}$, es decir, el valor central de los datos ordenados de menor a mayor. Cuando *n* es par la mediana coincide por convenio con el valor $\left(x'_{n/2} + x'_{n/2+1}\right)/2$, es decir, la semisuma de los dos valores centrales.

Para el caso de datos agrupados por clases hay que identificar, previamente, el intervalo mediano, que es aquel cuyo extremo inferior α-1 tiene una frecuencia relativa acumulada $F_{\alpha-1}$ menor o igual a 0,50 y su extremo superior α tiene una frecuencia relativa acumulada F_{α} mayor que 0,50. El procedimiento consiste en recorrer la serie ordenada de frecuencias relativas acumuladas (última columna en el ejemplo 1.1.11) hasta encontrar el intervalo que cuenta con una frecuencia relativa acumulada en su extremo final, F_{α} , igual o mayor a 0,5. Si es igual, ese es el valor de la mediana. Si es mayor la mediana es:

$$\widetilde{x} = m \approx L_{\alpha-1} + \frac{0{,}5 - F_{\alpha-1}}{F_{\alpha} - F_{\alpha-1}} h_{\alpha}$$

Como se ve la fórmula proporciona la mediana como el valor de la variable que coincide con la frecuencia relativa acumulada de 0,5 dentro del intervalo mediano. Para ello se suma al valor del extremo inferior del intervalo la parte de h_a (amplitud del intervalo) que va desde el extremo inferior $L_{\alpha-1}$ hasta el valor que coincide con la probabilidad 0,5. Suma que proporciona la mediana.

3.3 La moda m_0

La moda de una serie de datos es el valor de la variable que más veces se repite. Si se trata de una variable discreta o categórica coincide con el valor de mayor frecuencia de valor. Si la variable es continua, la moda de la serie original $\{x_1, x_2, ..., x_i, ..., x_n\}$ no existe conceptualmente, pues en ésta no se repetiría ningún valor, si se contase con un instrumento de resolución infinita. Como esto no es posible por razones prácticas, los instrumentos miden valores cuya repetición es más o menos improbable pero no imposible. En un cierto modo, la resolución del instrumento es un intervalo de clase, dado que no informa de diferencias más pequeñas de las que es capaz de mostrar en el indicador.

En estos casos, se puede calcular una moda para la serie de datos mediante una agrupación de datos, como se viene haciendo en este apartado. En ese caso la moda viene dada por la aplicación al intervalo de mayor densidad de frecuencia absoluta o relativa de la fórmula

$$m_0 \cong L_{\alpha-1} + \frac{d_{\alpha+1}}{d_{\alpha-1} + d_{\alpha+1}} . h_{\alpha}$$

siendo:

$d_{\gamma} = d_{\alpha} = \frac{n_{\alpha}}{h_{\alpha}}; \quad \alpha = 1,2,..., \nu$ la densidad de frecuencia de los correspondientes intervalos de clase. Es decir, la frecuencia de clase en el intervalo dividida por la amplitud del mismo.

Como se puede comprobar la moda se calcula sumando al extremo inferior del intervalo de clase de más densidad de frecuencia una cantidad en unidades de la variable que se trate que la sitúe más cerca del intervalo vecino (anterior o posterior) que más densidad de frecuencia tenga.

4 Medidas de dispersión

Una vez establecida la posición del conjunto de datos es de interés medir la mayor o menor dispersión con que se presenta. Para ello, la estadística ha desarrollado métodos diversos de estructura más o menos compleja.

En general, la dispersión se mide respecto de las medidas de posición. La media suele ser el valor generalmente utilizado. La medida de dispersión más elemental de un dato es la diferencia, $\hat{x}_i = x_i - \bar{x}$, llamada desviación del dato *i-ésimo* respecto de la media aritmética. Existen *n* diferencias o desviaciones de la media. La suma vale:

$$\sum_{i=1}^{n} \hat{x} = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

Este resultado muestra que las desviaciones no son independientes y la media aritmética de la serie de desviaciones es cero. En consecuencia, dicha medida, no sirve como medida de dispersión.

Por eso, para cuantificar la variabilidad de una serie de datos se han definido otras medidas. Las principales son:

- la varianza
- la desviación típica
- el recorrido o rango
- la media de valores absolutos de las desviaciones
- el rango intercuartílico

4.1 La varianza y la desviación típica

La varianza de una serie de datos, s_n^2, tiene la siguiente expresión:

$$s_n^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 \ .$$

Y para datos agrupados:

$$s_n^2 = \frac{1}{n} \sum_{\alpha=1}^{\nu} n_\alpha (x_\alpha - \bar{x})^2 \quad = \sum_{\alpha=1}^{\nu} f_\alpha (x_\alpha - \bar{x})^2 \ .$$

Esta medida cobra sentido porque al elevar al cuadrado la desviación de cada dato respecto de la media se obtienen cantidades siempre positivas y la suma de las desviaciones cuadráticas es mayor que cero mientras existan datos diferentes de la media.

La desviación típica tiene la expresión

$$s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y para valores agrupados:

$$s_n = \sqrt{\sum_{\alpha=1}^{\nu} f_\alpha (x_\alpha - \bar{x})^2}\ .$$

La cuasivarianza ofrece una medida de dispersión similar a la varianza, solo que utilizando como denominador *n-1* en vez de *n*. La expresión es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

y para datos agrupados:

$$s_n^2 = \frac{1}{n-1}\sum_{\alpha=1}^{\nu} n_\alpha (x_\alpha - \bar{x})^2$$

La cuasidesviación típica es la raíz cuadrada de la cuasivarianza_

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Y para valores agrupados:

$$s = \sqrt{\frac{1}{n-1}\sum_{\alpha=1}^{\nu} n_\alpha (x_\alpha - \bar{x})^2}\ .$$

La cuasivarianza y la cuasidesviación típica tienen aplicación cuando se trata de operar sobre una muestra en vez de sobre una población. Es decir cuando se tiene unos pocos datos y no todos, o casi todos, los datos. La ventaja estriba en que para un determinado tamaño de muestra la media aritmética de muchos valores de s^2 está centrada sobre la varianza de la población. Como se ha dicho más arriba, en última instancia, el propósito de la estadística es conocer los parámetros de una población. Cuando se cuenta con pocos datos es necesario emplear las fórmulas que con más exactitud se aproximan al valor buscado. Éste es el caso de la cuasivarianza como fórmula para aproximarse (estimar) el parámetro varianza de la población. Estos aspectos se tratarán con más detalle en el apartado de estimación del hormigón.

Pero la diferencia entre la varianza y la cuasivarianza es, sin perjuicio de las razones para su empleo en datos procedentes de muestras, de carácter matemático. En efecto, la suma de *n* cuadrados de las desviaciones, $\sum_{i=1}^{n}(x_i - \bar{x})^2$, sólo tiene *n-1* sumandos

independientes, pues cualquier desviación respecto de la media, por ejemplo la primera, $x_1 - \bar{x}$, es una combinación lineal de las restantes. En efecto:

$$x_1 - \bar{x} = \hat{x}_1 = \sum_{i=1}^{n}(x_i - \bar{x}) - \sum_{i=2}^{n}(x_i - \bar{x}) = 0 - \sum_{i=2}^{n}(x_i - \bar{x})$$

$$\hat{x}_1 = -\sum_{i=2}^{n}(x_i - \bar{x}) = -\sum_{i=2}^{n}\hat{x}_i$$

Por lo que la suma cuadrática $\sum_{i=1}^{n}(x_i - \bar{x})^2$ en una función de *n-1* desviaciones independientes. Es decir, si la serie de datos original $\{x_1, x_2, ..., x_n\}$ tiene *n* grados de libertad la expresión de s_n^2 o de s^2 sólo tiene *n-1*, pues uno se ha consumido al calcular la media aritmética que interviene en la definición de la varianza.

Ahora se puede comprender mejor la cuasivarianza que representa un promedio más natural de las desviaciones cuadráticas que la varianza, al dividir la suma de aquellas por los grados de libertad que posee. En todo caso la discusión sobre que medida es mejor para evaluar la dispersión de una serie de datos es de escasa importancia, dado que la relación entre ellas es $s^2 = \frac{n}{n-1}s_n^2$. Es decir muy pequeña cuando *n* es grande. Todo lo dicho es de aplicación a la relación entre la desviación típica y la cuasidesviación típica cuya relación es $s = \sqrt{\frac{n}{n-1}}s_n$.

4.2 El recorrido o rango

El recorrido o rango es otra medida de la dispersión o variabilidad de una serie de datos $\{x_1, x_2, ..., x_n\}$ procedentes de las observaciones efectuadas a una variable. Se define

$$r = x_n' - x_1',$$

siendo x_n' el mayor valor de la serie y x_1' el menor valor de la serie $\{x_1' \le x_2' \le ... \le x_n'\}$ construida a partir de la serie original. En la primera los subíndices numéricos indican el orden de aparición del dato y en la segunda el orden de menor a mayor valor del dato.

4.3 La media de los valores absolutos de las desviaciones

Como se pudo comprobar en A.1.1.6.1 las media de las desviaciones $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$ es igual a cero por lo que no sirve como medida de dispersión. Sin embargo, si se utilizan las

desviaciones en valor absoluto su media pasa a ser una medida de la dispersión de los datos. Su expresión

$$\bar{d}_{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|.$$

Y para datos agrupados:

$$\bar{d}_{\bar{x}} = \sum_{\alpha=1}^{\nu} f_\alpha \left|x_\alpha - \bar{x}\right|.$$

4.4 El rango intercuartílico

El rango intercuartílico es, también, una medida de la dispersión o variabilidad de los datos. Los cuartiles dividen a la serie de los datos en cuatro partes de frecuencia 0,25 cada una.

- el primer cuartil $\tilde{Q}_1 = \tilde{Q}(0{,}25)$ deja a su izquierda el 25 % de los datos;
- el segundo cuartil $\tilde{Q}_2 = \tilde{Q}(0{,}50)$ deja a su izquierda el 50 % de los datos
- el tercer cuartil $\tilde{Q}_3 = \tilde{Q}(0{,}75)$ deja a su izquierda el 75 % de los datos

El rango intercuartílico se define como:

$$r_{ic} = \tilde{Q}(0{,}75) - \tilde{Q}(0{,}25).$$

Los cuartiles son un caso particular de los cuantiles. Un cuantil es un valor $\tilde{Q}(p)$ que deja a su izquierda (considerándose también incluidos alguna parte del dato que coincida con el cuantil) una fracción *p* de los valores de la serie observada. Para calcular un cuantil hay que seguir el siguiente proceso:

1. Se ordenan los datos, obteniéndose la serie ordenada $\left\{x_1' \le x_2' \le \ldots \le x_n'\right\}$
2. Se calcula el valor *p(n+1)*
3. Se determina la parte entera de *p(n+1)* que se designa $k = Int\left[p(n+1)\right]$
4. Se determina la parte fraccionaria de *p(n+1)* que se designa $\omega = Frac\left[p(n+1)\right] = p(n+1) - k$
5. Se determina $\tilde{Q}(p) = (1-\omega)x_k' + \omega x_{k+1}'$

Como se ve, el cuantil resulta ser una media ponderada de dos datos consecutivos de la serie ordenada, los de índice de orden *k* y *k+1*, respectivamente. La mediana es un caso particular de cuantil. En efecto es el cuantil para *p=0,5.* Lo que equivale a ser el segundo cuartil. La fórmula del punto 5 es coherente con la definición de mediana dada en el apartado 1.1.6.2.

Aunque el rango intercuartílico es una medida de dispersión, los cuartiles mismos son medidas de posición no centrada, con la excepción del segundo cuartil que coincide con la mediana, como se ha dicho. De hecho tienen la doble naturaleza de medidas de

posición y de dispersión, pues un cuartil esta posicionando la masa de datos pero está también informando parcialmente de su variabilidad.

Para caracterizar una serie de datos se recomienda usar un grupo de cinco medidas, de las que tres se basan en cuartiles. Son las siguientes: $x_1', \widetilde{Q}_1, \widetilde{Q}_2, \widetilde{Q}_3, x_n'$. Es decir, los valores extremos de la serie ordenada y los tres cuartiles, el segundo de los cuales coincide con la mediana.
Estas cinco medidas proporcionan una información valiosa pues aportan una medida de posición robusta como es la mediana; una medida de la dispersión, mediante el rango intercuartílico $\underset{\sim}{Q}^3 - \underset{\sim}{Q}^1$, una medida de la asimetría mediante la diferencia cuartil $(\widetilde{Q}_1 + \widetilde{Q}_2 - 2\widetilde{Q}_2)$, que es cero cuando la distribución de los datos es simétrica respecto de la mediana, y una medida de la curtosis de la distribución de los datos mediante las diferencias $\widetilde{Q}_1 - x_1'$ y $x_n' - \widetilde{Q}_3$, que indican el grado de concentración de datos en los extremos (colas) de la distribución de los datos. Los conceptos de distribución, asimetría, colas y curtosis serán desarrollados con mayor precisión más adelante.

Esa información de carácter sintético mejora si se expresa gráficamente conjuntamente con dos límites adicionales situados a 1,5 veces el rango intercuartílico, $\underset{\sim}{Q}^3 - \underset{\sim}{Q}^1$, del primer cuartil a la izquierda y del tercer cuartil a la derecha. Surge entonces el diagrama de caja y bigotes, en el que la caja (rectángulo central) está partida por la mediana y limitada por el primer y tercer cuartil y los bigotes (segmentos horizontales que se prolongan hasta los valores extremos no aberrantes, es decir hasta los valores extremos incluidos en los límites definidos más arriba. De este modo hay un criterio para identificar valores atípicos cuando los valores extremos sobrepasan los límites señalados. A veces esto extremos se trasladan hasta una distancia de 3 veces el rango intercuartílico.

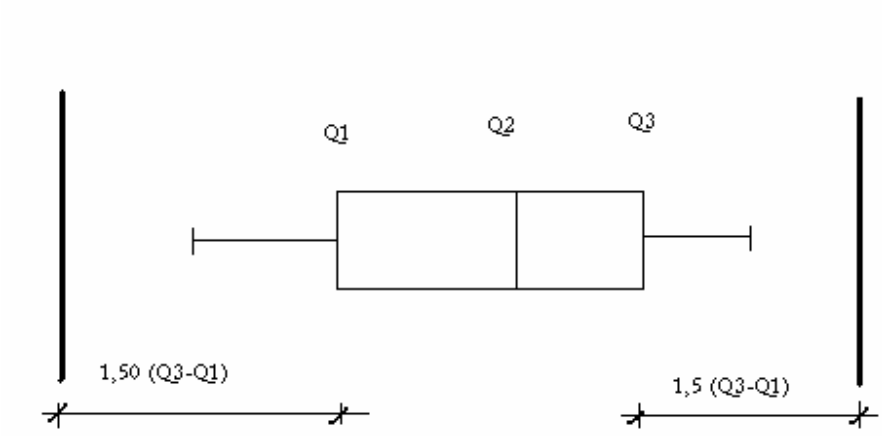


Diagrama de Caja y Bigotes

5 Medidas mixtas

Existen medidas mixtas formadas por la combinación de medidas de dispersión y medidas de posición. La principal es el coeficiente de variación cuyo carácter mixto procede de la inclusión de una medida de dispersión (la desviación típica) y de una medida de posición (la media aritmética). Aunque si hay que decidirse, se puede considerar antes una medida

de dispersión debido a que, en realidad mide la dispersión relativa a la media, lo que permite comparar las dispersiones de series de datos de diferente media.

Hay dos versiones del coeficiente de variación:

$$v = \frac{s}{\bar{x}}$$

basada en la cuasidesviación típica, y

$$v_n = \frac{s_n}{\bar{x}}$$

basada en la desviación típica. En la estadística inductiva o inferencial se emplea más *v* y en la estadística descriptiva v_n.

El recorrido relativo es la relación entre el recorrido de una serie de datos y su media aritmética.

$$w = \frac{r}{\bar{x}}$$

Dado que el recorrido es una medida de variabilidad basada en la serie de datos ordenados, tal vez sería más adecuado utilizar *m* (la mediana) en lugar de $\bar{x}$ en la expresión precedente, dado que la mediana es, también, una medida de posición (como la media aritmética) pero surgida de la serie ordenada (como el recorrido). La EHE, sin embargo ha optado por el valor que incluye a la media, lo que, en la práctica, no tiene gran incidencia puesto que el tamaño de la muestra utilizada generalmente es *n=2,* por lo que la media y la mediana coinciden. Marginalmente se puede considerar otra medida mixta como es la relación entre el rango intercuartílico y la media o la mediana.