Unidad 2

Tratamiento estadístico de datos

Histogramas y distribución estadística

Consideremos una población de personas de una ciudad. Queremos analizar cómo se distribuyen las estaturas de la población. Para llevar adelante este estudio podemos medir la altura de todos los individuos de la *población*, o bien tomar una *muestra representativa* de la misma a partir de la cual inferiríamos las características de la población.

Esta clase de estudio es un típico problema de estadística. Si tomamos una muestra de tamaño N y para la misma medimos las alturas de cada individuo, este experimento dará N resultados: $x_1, x_2, x_3, \ldots, x_N$. Todos estos datos estarán comprendidos en un intervalo de alturas (x_{min}, x_{max}) entre la menor y mayor altura medidas.

Una manera útil de visualizar las características de este conjunto de datos consiste en dividir el intervalo (x_{min}, x_{max}) en m subintervalos delimitados por los puntos $(y_1, y_2, y_3, \ldots, y_m)$; a estos subintervalos los llamaremos el *rango de clases*. Seguidamente, contamos el número n_1 de individuos de la muestra cuyas alturas están en el primer intervalo $[y_1, y_2)$, el número n_j de los individuos de la muestra que están en el j-ésimo intervalo $[y_{j-1}, y_j)$, etc., hasta el m-ésimo subintervalo. Aquí hemos usado la notación usual de corchetes, […], para indicar un intervalo cerrado (incluye al extremo) y paréntesis comunes, (…), para denotar un intervalo abierto (excluye el extremo).

Con estos valores definimos la función de distribución f_j que se define para cada subintervalo como:

$$f_j = \frac{n_j}{\sum_j n_j} \tag{2.1}$$

Esta función de distribución está normalizada, es decir:

$$\sum_{j=1}^{m} f_j = 1 \tag{2.2}$$

El gráfico de f_j en función de x_j $[x_j = (y_{j-1} + y_j)/2]$ nos da una clara idea de cómo se distribuyen las alturas de los individuos de la muestra en estudio. Este tipo de gráfico se llama un *histograma* y la mayoría de las hojas de cálculo de programas comerciales

(Excel, Origin, etc.) tienen herramientas para realizar las operaciones descriptas y el gráfico resultante. En la Fig. 2.1 ilustramos dos histogramas típicos.

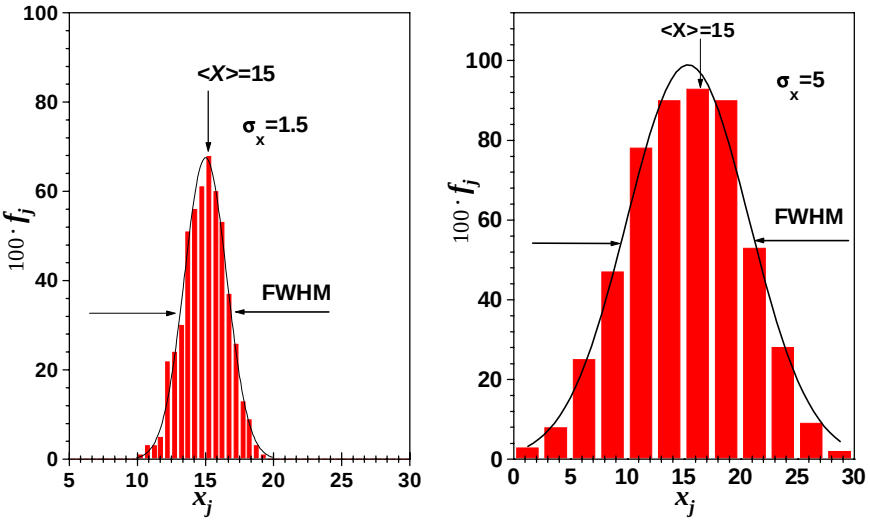

Figura 2.1. Histograma de dos muestras con igual valor medio pero con distintos grados de dispersión. En este ejemplo, los datos tienen una distribución gaussiana o normal, descripta por la curva de trazo continuo.

Tres parámetros importantes de una distribución son:

- El valor medio: $\bar{x} =< x >= \sum_{j=1}^{m} x_j \cdot f_j = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$ (2.3)

- La varianza: $Var(x) = \sigma_x{}^2 = \sum_{j=1}^{m} (x_j - \bar{x})^2 \cdot f_j$ (2.4)

- La desviación estándar: $\sigma_x = \sqrt{Var(x)}$ (2.5)

El valor medio (también llamado media y promedio) <***x***> da una idea de la localización del centro de masa (centroide) de la distribución. La desviación estándar σ_x es una medida de la dispersión de los datos alrededor del promedio. Cuando más concentrada esté la distribución de valores alrededor de <***x***>**,** menor será σ_x, y viceversa.

Una distribución de probabilidad muy común en diversos campos es la *distribución gaussiana o normal*, que tiene la forma de una campana como se ilustra en trazo continuo en la Fig. 2.1. La expresión matemática de esta distribución es:

$$f(x) = N(x; m, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{2.6}$$

La "campana de Gauss" está centrada en m y su ancho está determinado por la desviación estándar σ. Los puntos de inflexión de la curva están en $x-\sigma$ y $x+\sigma$. El área de esta curva entre estos dos puntos constituye el 68.3% del área total. El área entre $x-2\sigma$ y $x+2\sigma$ es del 96% del total. Es útil caracterizar para esta función el ancho a mitad de su altura, que está relacionado con σ a través de la expresión: FWHM = 2.35σ (FWHM, de "full width half maximum").

Cuando se desea comparar un histograma no normalizado con una curva normal, es necesario contar el número total de datos N_t, el valor medio de los mismos, $\bar{x}$, y la desviación estándar de los datos, σ_x. Para comparar el histograma con la curva normal debemos multiplicar la distribución dada por la Ec. (2.6) por un factor $N_t.\Delta x$, donde Δx es el ancho del rango de clases que suponemos idéntico para cada intervalo.

Aunque la distribución gaussiana ocurre naturalmente en muchos procesos, desde luego no es única y existen muchos tipos de distribuciones de ocurrencia común en la naturaleza.

📖 Los parámetros más usuales con los que puede caracterizarse la localización de una distribución asociada a un conjunto de N datos son:

a) la media
b) la mediana
c) la moda

La *media* o *promedio* de la distribución se define, según ya vimos, como $\bar{x} = \sum_i^N x_i / N$, y es la media aritmética de los valores observados.

La *moda* corresponde al valor de la variable donde está la máxima frecuencia, o sea, que en un histograma la moda corresponde al valor de la variable donde hay un pico o máximo. Si una distribución tiene dos máximos la denominamos distribución bimodal, si tiene tres máximos trimodal y así sucesivamente.

La *mediana* es el valor de la variable que separa los datos entre aquellos que definen el primer 50% de los valores de los de la segunda mitad. O sea que la mitad de los datos de la población o muestra están a derecha de la mediana y la otra mitad están a la izquierda de la misma.

Mientras que a la media la calculamos usando una fórmula, a la moda la evaluamos directamente del histograma.

Para estimar la mediana tenemos que observar la lista de datos ordenados de menor a mayor, y ubicar el valor central de la lista. Si el número de datos es impar, la mediana corresponde precisamente al valor central. Si el número *N* de datos es par, la mediana se estima como ½ ($X_{N/2} + X_{N/2+1}$). En una distribución dada, una línea vertical trazada desde la mediana divide a la distribución en dos partes de área equivalentes.

Media, moda y mediana no tienen, en general, porqué coincidir. Estos tres parámetros sí son iguales en el caso de distribuciones unimodales simétricas respecto del valor medio. Este es el caso de una distribución gaussiana. En el caso de una distribución asimétrica, las diferencias entre moda, media y mediana pueden ser sustanciales.

Es importante saber cuál parámetro de localización es más apropiado de usar o más representativo en una dada situación. Consideremos, para fijar ideas, la distribución del ingreso familiar en un país dado. La presencia de millonarios, aunque sean relativamente pocos, tiene un efecto sobre la media que contrarresta a muchos miembros de la población en el extremo inferior de la escala de salarios. De esta manera, la moda y la media difieren sustancialmente. En este caso tal vez la moda es un parámetro más representativo que la media. A menudo los datos estadísticos pueden ser interpretados de diversas maneras. El siguiente ejemplo ilustra las distintas interpretaciones que pueden extraerse de un conjunto de datos estadísticos.

✍ Una empresa analiza la necesidad de discutir los salarios. El cuadro de sueldos es el siguiente:

Director	$9000
Sub-director	$5000
2 Asesores	$2500
2 Encargados	$ 1350 c/u
Jefe de sección	$ 1200
6 Obreros	$600 c/u

La empresa argumenta que el salario medio es $2000. El delegado gremial sostiene que el sueldo representativo es de $600. Un político consultado asegura que el salario más representativo es $900. ¿Qué parámetros tuvo en cuenta para argumentar cada persona participante de la reunión? Calcule la moda, la mediana y la media de los ingresos para esta empresa.

Magnitud que se mide N veces

Cuando medimos una magnitud una única vez (N = 1), el mejor valor es simplemente el valor medido y su incertidumbre está dada por la incertidumbre nominal, σ_{nom}, que tiene en cuenta los errores del instrumento, del método y de las operaciones. Esto es consistente con la Ec. (1.3), que arroja como resultado $\Delta x = \sigma_{nom}$ cuando no disponemos del término estadístico σ_{est}.

En muchos casos prácticos estamos interesados en el estudio estadístico de la variación de las mediciones y realizamos N mediciones de la magnitud de interés. Por ejemplo podemos estar interesados en analizar la *repetibilidad* de un procedimiento, o en ver el efecto que tienen las *fluctuaciones* de un instrumento sobre las mediciones. Por otra parte, la realización de varias mediciones del mesurando minimiza la incidencia de los errores estadísticos. Dado el carácter aleatorio de estos tipos de errores es claro que, al promediar los resultados, el promedio estará menos afectado de las desviaciones estadísticas que lo que están los valores individuales. En todos estos casos es aplicable el tratamiento estadístico de datos que discutimos seguidamente. El procedimiento que se describe a continuación es un método para analizar estadísticamente las N mediciones y determinar las incertidumbres asociadas al promedio de las mismas. El procedimiento de repetición de mediciones no es aplicable para reducir los errores de carácter sistemático y mucho menos los espurios.

Supongamos que hemos hecho N mediciones de una misma magnitud con resultados $x_1, x_2, ..., x_j, ...x_N$. Estas N determinaciones pueden ser consideradas una *muestra* de todas las posibles mediciones que se podrían realizar (*población*). Bajo condiciones muy generales puede demostrarse que el *mejor estimador* de la magnitud x viene dado por el promedio de los valores:

$$\bar{x} = \frac{\sum_{j=1}^{N} x_j}{N} . \qquad (2.7)$$

Este resultado es llamado también el *mejor valor*, el *estimador* o el *valor más probable del mesurando*. Llamaremos a

$$\Delta x_j = x_j - \bar{x} \qquad j = 1, 2, ..., N$$

la desviación de cada medición respecto de $\bar{x}$. También definimos la desviación estándar o desviación cuadrática media de cada medición, S_x:

$$S_x = \sqrt{\frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{N-1}} \,, \tag{2.8}$$

que es un *estimador muestral* de la desviación estándar poblacional y da una idea global acerca de la dispersión de los puntos x_j alrededor del promedio $\bar{x}$. Si la distribución es ancha, S_x será grande, mientras que si está afinada alrededor del promedio $\bar{x}$, su valor será pequeño (ver Fig. 2.1). S_x tiene las mismas dimensiones físicas que $\bar{x}$, lo que hace posible compararla directamente con ésta a través del cociente $S_x / \bar{x}$. La calidad del proceso de medición será mayor cuanto menor sea $S_x / \bar{x}$, que en general es una constante del proceso de medición y no depende de *N*.

Si suponemos ahora que realizamos varias series de mediciones de *x*, y para cada una de estas series calculamos el valor medio $\bar{x}$, es de esperar que estos valores también se presenten distribuidos (puesto que variarán entre sí) pero con una menor dispersión que las mediciones individuales. Se puede probar que a medida que el número *N* de mediciones aumenta, la distribución de $\bar{x}$ será normal con una desviación estándar dada por[1,2,4,5]:

$$\sigma_{est} = \sigma_x = \sqrt{\frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{N.(N-1)}} = \frac{Sx}{\sqrt{N}} \,. \tag{2.9}$$

σ_x se llama la *desviación estándar del promedio* y en un experimento es una medida de la *incertidumbre estadística* asociada a $\bar{x}$ en el proceso de medir la magnitud *N* veces.

📖 Cuando el resultado de una medición se expresa como $(\bar{x} \pm \sigma_{est})$, esto es equivalente a decir que el valor de *x* está contenido en el intervalo $(\bar{x} - \Delta x,\ \bar{x} + \Delta x)$ con probabilidad $p_0 = 0.68$. Esto es equivalente a expresar:

$$P(\bar{x} - \Delta x < x < \bar{x} + \Delta x) = p_0 \,,$$

que se interpreta como "la probabilidad de que el *mejor estimador* de *x* esté comprendido entre $\bar{x} - \Delta x$ y $\bar{x} + \Delta x$ es igual a p_0." El valor de p_0 se conoce con

el nombre de *coeficiente de confianza* y el intervalo $(\bar{x} - \Delta x\,,\, \bar{x} + \Delta x)$ determina un *intervalo de confianza* para *x*.

Número óptimo de mediciones

Recordemos que S_x mide la dispersión de cada medición y que no depende de N sino de la calidad de las mediciones, mientras que $\sigma_{\bar{x}}$ sí depende de N y es menor cuanto más grande es N [Ec.(2.9)]. En principio parece tentador pensar que si medimos una magnitud un gran número de veces, podremos despreciar la contribución de la incertidumbre estadística en la Ec.(1.3). Ciertamente σ_{est} disminuye al aumentar N, pero desde un punto de vista físico, solo tiene sentido que disminuya hasta hacerse igual o del orden que σ_{nom}, que está determinado por el instrumental y el método de medición. Pensemos que si, por ejemplo, estamos midiendo una longitud con una regla graduada en milímetros, un aumento en el número de mediciones llevará a disminuir la incertidumbre de carácter estadístico, pero *nunca* con este instrumento podremos obtener con certeza cifras del orden de los micrones –por más que realicemos más y más mediciones.

La Ec. (1.3) indica que no es razonable esforzarse en disminuir σ_{est} mucho más que σ_{nom}. El balance óptimo se logra cuando $\sigma_{est} \approx \sigma_{nom}$. Esto nos da un criterio para decidir cual es el número óptimo de mediciones a realizar. Como suponemos que S_x es independiente de N, la idea es hacer un número pequeño de mediciones preliminares N_{prel} –digamos entre 5 y 10– y luego calcular S_x. Puesto que de un análisis previo de las características del instrumento y de los procedimientos ya conocemos σ_{nom}, podemos estimar el número óptimo de mediciones, N_{op}, como

$$N_{op} \approx 1 + \left(\frac{S_x}{\sigma_{nom}} \right)^2 , \tag{2.10}$$

que resulta de imponer la condición $\sigma_{\bar{x}} \approx \sigma_{nom}$ y usar la Ec.(2.9); el término unidad del segundo miembro nos asegura que siempre es necesario realizar al menos una medición. Si $N_{op} > N_{\text{prel}}$, se completan las mediciones para lograr N_{op} valores y se recalcula $\sigma_{\bar{x}}$. Si $N_{op} < N_{\text{prel}}$, no se realizan más mediciones que las preliminares y se usan todas ellas. Finalmente, en todos los casos la incertidumbre absoluta combinada Δx vendrá dada por la Ec. (1.3):

$$\Delta x = \sqrt{\sigma_{nom}^2 + \sigma_{\bar{x}}^2} \ . \tag{2.11}$$

Para la mayoría de los casos de interés práctico, si medimos 100 veces una magnitud *x*, aproximadamente 68 de ellas caerán en el intervalo $(\bar{x}-\sigma_x, \bar{x}+\sigma_x)$, 96 de ellas en el intervalo $(\bar{x}-2\sigma_x, \bar{x}+2\sigma_x)$ y 99 de ellas en el intervalo $(\bar{x}-3\sigma_x, \bar{x}+3\sigma_x)$. Estos resultados valen estrictamente para el caso en que los errores se distribuyan "normalmente", es decir, si el histograma formado con los resultados de las mediciones adopta la forma de una campana de Gauss.

Decálogo práctico

En resumen, los pasos a seguir para medir una magnitud física *X* son:

1. Se analizan posibles fuentes de errores sistemáticos y se trata de minimizarlos.
2. Se estima la incertidumbre nominal σ_{nom}
3. Se realizan unas 5 a 10 mediciones preliminares y se determina la desviación estándar de cada medición S_X (2.8).
4. Se determina el número óptimo de mediciones N_{op} (2.10).
5. Se completan las N_{op} mediciones de *X*.
6. Se calcula el promedio $\overline{X}$ y la incertidumbre estadística σ_x.
7. Se evalúa la incertidumbre absoluta de la medición combinando todas las incertidumbres involucradas (error efectivo (1.3)),
   $\Delta X = \sqrt{\sigma_x^2 + \sigma_{nom}^2}$.
8. Se expresa el resultado en la forma $X = \overline{X} \pm \Delta X$ con la *unidad correspondiente*, cuidando que el número de cifras significativas sea el correcto.
9. Es útil calcular e indicar la incertidumbre porcentual relativa de la medición ε_X=100* $\Delta X / \overline{X}$, lo que puede servir en comparaciones con resultados de otras experimentadores o por otros métodos.
10. Si se desea estudiar la distribución estadística de los resultados (por ejemplo si es normal o no), se compara el histograma de la distribución de los datos experimentales con la curva normal correspondiente, es decir con una distribución normal de media $\overline{X}$ y desviación estándar S_X.

Combinación de mediciones independientes

Una situación frecuente en ciencia es la determinación del mejor valor de una dada magnitud usando varios valores provenientes de mediciones independientes (obtenidos por diferentes autores, con diferentes técnicas e instrumentos, etc.). Cada una de estas mediciones independientes puede tener asociada distintas incertidumbres. Es decir, te-

nemos un conjunto de N mediciones, cada una caracterizada por un par (x_k, σ_k), con k = 1, 2, ..., N. Nuestro objetivo es obtener el mejor valor para la magnitud en discusión. Es claro que al combinar los distintos resultados para obtener el mejor valor, $<x>$, es preciso tener en cuenta las respectivas incertidumbres, de tal modo que aquellas mediciones más precisas contribuyan más ("que pesen más") en el resultado final. Es posible demostrar en este caso que el mejor valor $<x>$ viene dado por[1]:

$$< x >= \frac{\sum_{k=1}^{N} \frac{x_k}{\sigma_k^{\,2}}}{\sum_{k=1}^{N} \frac{1}{\sigma_k^{\,2}}}, \qquad (2.12)$$

con una incertidumbre absoluta $\Delta x \equiv \sigma_{<x>}$ dada por[1]

$$\frac{1}{\sigma_{<x>}^2} = \sum_{k=1}^{N} \frac{1}{\sigma_k^{\,2}}. \qquad (2.13)$$

📖 Un caso especial de interés, es cuando tenemos N determinaciones del mesurando, todas ellas con al misma incertidumbre σ. Como puede deducirse fácilmente de la Ec. (2.12) el promedio será:

$$< x >= \frac{\sum_{k=1}^{N} x_k}{N},$$

que, como es de esperar, coincide con la expresión (2.7). La incertidumbre asociada a este valor será, según la Ec. (2.13):

$$\sigma_{<x>} = \frac{\sigma}{\sqrt{N}},$$

que coincide con la Ec. (2.9). Además queda ilustrado el significado de σ como una medida de la dispersión asociada a cada medición individual y $\sigma_{<x>}$ como la dispersión asociada al mejor valor.

Discrepancia

Si una magnitud física se mide con dos o más métodos, o por distintos observadores, es posible –y muy probable– que los resultados no coincidan. Decimos entonces que existe una *discrepancia* en los resultados. El término *repetibilidad* se usa para describir la concordancia o no entre varias mediciones realizadas por el mismo observador con el mismo método. En cambio la *reproducibilidad* está asociada a la concordancia o no de mediciones realizadas por distintos observadores o distintos métodos.

Lo importante es saber si la discrepancia es significativa o no. Un criterio que se aplica frecuentemente es el siguiente. Si los resultados de las dos observaciones que se comparan son independientes (caso usual) y tienen como resultados:

$$\textit{Medición } 1: \qquad X_1 = \overline{X}_1 \pm \Delta X_1$$

$$\textit{Medición } 2: \qquad X_2 = \overline{X}_2 \pm \Delta X_2$$

definimos:

$$\Delta X^2 = \Delta X_1^2 + \Delta X_2^2$$

Decimos que con un límite de confianza del *p0* (=68% si los datos tiene distribución normal) las mediciones son distintas si:

$$\left|\overline{X}_1 - \overline{X}_2\right| \geq \Delta X\,,$$

y que con un límite de confianza del 96% las mediciones son distintas si:

$$\left|\overline{X}_1 - \overline{X}_2\right| \geq 2 \cdot \Delta X$$

Estos criterios pueden generalizarse para intervalos de confianza mayores en forma similar. También se aplican cuando se comparan valores obtenidos en el laboratorio con valores tabulados o publicados. Nótese la diferencia entre discrepancia e incertidumbre. La *discrepancia* está asociada a la falta de superposición de dos intervalos (incertidumbres) de dos resultados distintos.

Bibliografía

1. P. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, 2nd ed. (McGraw Hill, New York, 1993).

2. Stuardt L. Meyer, *Data analysis for scientists and engineers* (John Willey & Sons, Inc., New York, 1975).

3. *Statistics: Vocabulary and symbols,* International Organization of Standarization (ISO), Suiza; http://www.iso.ch/infoe/sitemap.htm.

4. Spiegel y Murray, *Estadística*, 2da ed. (McGraw Hill, Schaum, Madrid, 1995).

5. H. Cramér, *Teoría de probabilidades y aplicaciones* (Aguilar, Madrid, 1968); H. Cramér, *Mathematical method of statistics* (Princeton University Press, New Jersey, 1958).