Física Experimental 1

Tratamiento Estadístico de Datos.

1. Conceptos básicos - Estimadores

Cuando medimos una magnitud x una única vez, el valor obtenido es el resultado de la medida, y su incertidumbre está dada por la incertidumbre nominal Δx, que tiene en cuenta la exactitud del instrumento, incertidumbres del método y de las operaciones. En muchas situaciones nos interesa el estudio estadístico de una magnitud y realizamos para ello n mediciones de la misma. Cada medida x es la realización de una variable aleatoria X con cierta densidad de probabilidad. Esta variable aleatoria tiene valor esperado $\mu = E(x)$ y varianza $\sigma^2 = var(x) = E[(x - E(x))^2]$. Estos valores no dependen de las medidas específicas que se realizan y no es posible conocerlos experimentalmente ya que para ello sería necesario analizar a toda la población. Entonces se utilizan estimadores estadísticos de los parámetros que se pueden calcular a partir de las n medidas realizadas. Supongamos que hemos hecho n medidas de una misma magnitud x con los resultados $x_1, x_2, ..., x_n$. Estas n determinaciones son una muestra de todas las posibles mediciones que se podrían efectuar, es decir una *muestra de la población*. Para saber como se distribuyen estas medidas, uno de los valores que es posible calcular es el promedio de los datos que se define como:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

Si tomamos otra muestra de tamaño n de la misma población, el promedio obtenido seguramente será diferente. Sin embargo se espera que los resultados sean similares (por ejemplo, si medimos una longitud tomando dos muestras de n valores no es muy probable que en una ocasión el promedio sea 10 cm y en otra 40 cm).

Se define también el desvío de los datos, (S_n), que describe la desviación de las medidas obtenidas respecto al promedio. El procedimiento para calcularla se explica a continuación:

El desvío de cada medida respecto a $\bar{x}$ es.

$$\Delta x_j = x_j - \bar{x}; j = 1, 2, ...n. \tag{2}$$

La suma de los desvíos $\sum \Delta x_j$ es cero, para cualquier conjunto de x_j. Dada la definición del promedio, la contribución a la sumatoria de los valores por encima del promedio $\bar{x}$ (contribución positiva), compensan la de aquellos que están por debajo (contribución negativa). Un valor cuantitativo del desvío está dado por $\sum \Delta x_j^2$, esto es, la suma de las *desviaciones cuadráticas*. Sin embargo, aún teniendo pequeños desvíos Δx_j, una suma de muchos de ellos resultaría en un número grande. Para independizarse del número de medidas se promedian esos valores, definiendo S como:

$$S^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1} = \frac{\sum_{j=1}^{n}(\Delta x_j)^2}{n-1} \tag{3}$$

Entonces se define el desvío S_n de una muestra de medidas como:

$$S_n = \sqrt{S^2} = \sqrt{\frac{\sum_{j=1}^{n}(\Delta x_j)^2}{n-1}} \tag{4}$$

Si su valor es grande, la probabilidad de hallar valores alejados del promedio es mayor, y si es pequeño los valores están más concentrados alrededor del promedio. El desvío S_n tiene las mismas dimensiones físicas que $\bar{x}$.

La ley de los grandes números establece que si el número de observaciones de una población aumenta, el promedio de la muestra $\bar{x}$ tiende al valor esperado μ. Entonces el promedio de una muestra puede ser usado como un estimador del valor esperado. De igual forma, S_n puede ser usado como estimador de la desviación standard σ.

2. Histograma

Cuando la variable de interés puede tomar muchos valores diferentes, una forma simple de representar gráficamente los resultados es mediante un histograma. Supongamos que tomamos N medidas de una magnitud x. Los resultados obtenidos estarán comprendidos en un intervalo entre el mínimo y el máximo valor $[x_{min}, x_{max}]$. Un histograma consiste en un gráfico, donde en el eje horizontal se colocan los valores obtenidos para la variable, agrupando valores próximos. Para ello se divide el intervalo de datos obtenidos en varios sub-intervalos o clases $[x_{min} = a_1, a_2), [a_2, a_3), ..., [a_{N-1}, a_N = x_{max}]$. Luego se cuenta el número de datos que caen en cada uno de los intervalos. Esas cantidades se definen como frecuencia absoluta, n_j.

Entonces se realiza un gráfico de barras del número de datos obtenidos en cada clase, tal como muestra el ejemplo de la Figura 1.

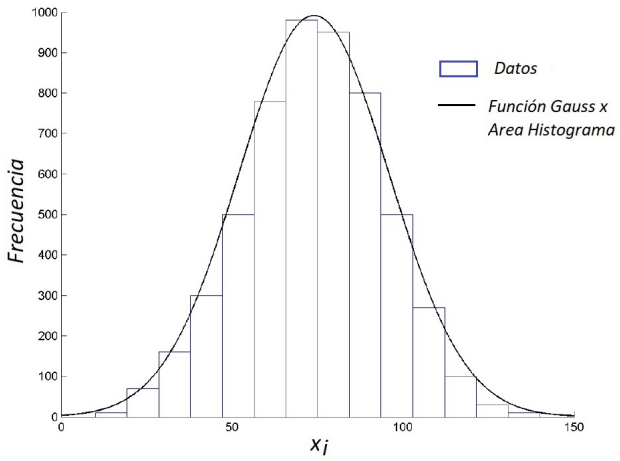


Figura 1: Histograma de barras típico (no normalizado) de una serie de medidas. La altura de las barras indica su frecuencia absoluta. En este ejemplo, los datos tienen una distribución gaussiana o normal, descrita por la curva de trazo continuo, superpuesta al histograma.

A partir de la frecuencia absoluta definimos la función de probabilidad puntual f_j como:

$$f_j = \frac{n_j}{\sum_j (n_j)} \tag{5}$$

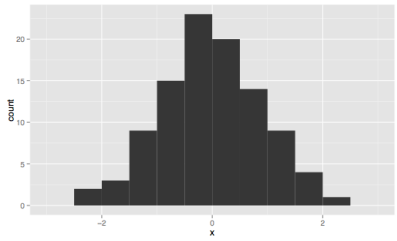
Esta función está normalizada, o sea:

$$\sum_j f_j = 1 \tag{6}$$

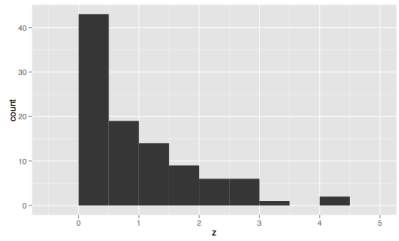
Si el número de medidas es muy grande y la longitud de los intervalos tiende a cero, obtenemos una función continua (llamada densidad de probabilidad): a cada punto (en el eje correspondiente a las clases) le corresponde un valor (su frecuencia).

En este punto es importante hacer énfasis en que los histogramas dan una idea aproximada de la distribución que siguen los datos y **NO** dan fehacientemente la información de qué distribución

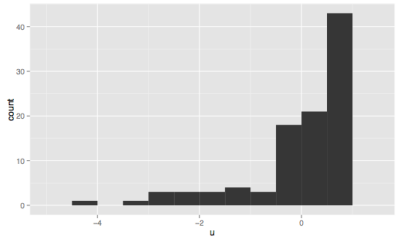
siguen dichos datos[1]. Por lo tanto, la manera de describir el comportamiento de un histograma es con términos como *unimodal*, *bimodal*, *multimodal*, *simétrico*, *skew-izquierda*, *skew-derecha*. Las Figura 2 muestra ejemplos de los diferentes tipos de histogramas mencionados.



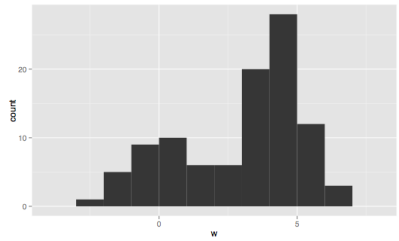
(a) Ejemplo de un histograma *unimodal* y simétrico. Notar que se observa un pico central con la mayor frecuencia de datos, mientras que a los costados se encuentra el resto de los datos.



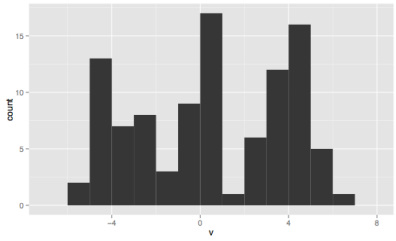
(b) Ejemplo de un histograma *skew-derecha*. Los datos tienen una distribución centrada en un costado, mientras hay una proporción de los mismos que se separan hacia la derecha.



(c) Ejemplo de un histograma *skew-izquierda*. Los datos tienen una distribución centrada en un costado, mientras hay una proporción de los mismos que se separan hacia la izquierda.



(d) Ejemplo de un histograma *bimodal*. Los datos presentan una distribución distribuida alrededor de dos modos (picos).



(e) Ejemplo de un histograma *multimodal*. Los datos tienen una distribución caracterizada por más de dos modos.

Figura 2: Ejemplos de diferentes tipos de histograma. Fuente: Wikipedia.

3. Estimadores estadísticos

Como ya mencionamos, al hacer un análisis estadístico, queremos obtener información de una población a partir de los datos de una muestra usando estimadores. Estos estimadores se pueden calcular a partir de una muestra de la población.

De esta forma el teorema central del límite (y la ley de los grandes números) establece que si el número de observaciones de una población aumenta, el promedio de la muestra x_n tiende a la media poblacional μ. Entonces podemos usar el promedio de una muestra como un estimador de la media poblacional. De la misma forma, si se define σ como la desviación estandard de la población, S_n se puede usar como un estimador de σ.

[1] Para ello es necesario hacer un análisis estadístico de mayor nivel.

4. Distribución normal o gaussiana.

Una de las funciones de densidad de probabilidad más importante es la llamada normal o gaussiana. Esta curva se describe completamente dados el valor esperado μ y de la desviación standard σ de una población. Tiene forma de campana y corresponde con la función:

$$f(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

Como no es posible conocer los valores de μ y σ, se utilizan su estimadores $\bar{x}$ y S_n por lo que la función que podemos graficar a partir una serie de datos experimentales (de una muestra) es:

$$f_n(x) = \frac{1}{\sqrt{(2\pi S_n^2)}} e^{-\frac{(x-\bar{x})^2}{2S_n^2}} \tag{8}$$

Esta función de densidad es muy importante ya que es la que describe correctamente una gran cantidad de datos de situaciones reales, como las alturas de una población, el nivel de ruido en telecomunicaciones, la vida media de una lámpara, etc. Además describe adecuadamente datos con resultados aleatorios, como son los que se obtienen cuando se realizan muchas medidas de una magnitud en iguales condiciones experimentales.

La campana gaussiana realizada a partir de una única serie de medidas es simétrica respecto a $\bar{x}$ y su ancho está determinado por el parámetro S_n. Los puntos de inflexión de la curva se encuentran en $\bar{x}-S_n$ y $\bar{x}+S_n$. El área comprendida entre estos puntos constituye aproximadamente el 68.3 % del área total. El área entre $\bar{x}-2S_n$ y $\bar{x}+2S_n$ es el 95.4 % del total. Para obtener los resultados correspondientes a las áreas comprendidas en los diferentes intervalos se recurre a técnicas de integración numérica, ya que la distribución normal no tiene una primitiva expresable en términos de funciones elementales. Los resultados de estas integrales se encuentran tabulados.

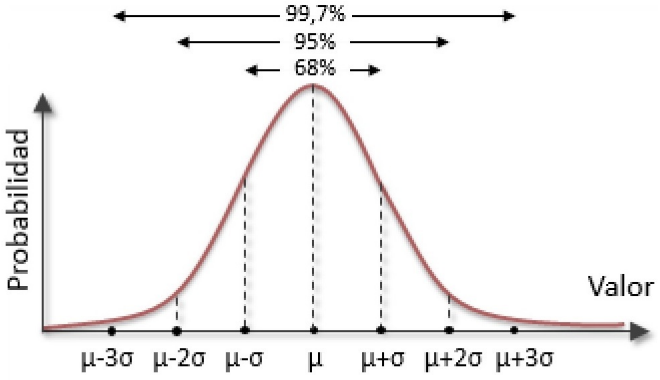


Figura 3: Distribución gaussiana, representando los porcentajes del gráfico bajo la curva en diferentes intervalos característicos

Una observación relevante es que para que las medidas obtenidas sigan una distribución gaussiana, la medición debe realizarse, cada vez, bajo idénticas condiciones.

Es importante destacar que los parámetros μ y la varianza (así como σ) no dependen del número de medidas. Si ahora nos planteamos realizar N series de n medidas de la magnitud x tendremos asociado un promedio $\bar{x_n}$ y un desvío standard S_{nN} para cada serie. Se puede demostrar que el desvío standard del promedio es $S_n/\sqrt{n}$ o sea, los promedios de cada serie se distribuyen en torno a $\bar{x}$ con un desvío $S_n(\bar{x}_n) = S_n/\sqrt{n}$. Esta relación permite predecir la fluctuación del promedio de una serie de medidas, sin tener que realizar medidas en mas series de datos.

Finalmente, es necesario utilizar algún criterio para definir un intervalo de incertidubre, de forma que el resultado de la medida quede expresado como $x = \bar{x} \pm \Delta x$.

1. El desvío del promedio $S_n/\sqrt{n}$, es una medida apropiada de la incertidumbre. Esto corresponde a tomar un intervalo de confianza del 68 %. Por lo tanto, tendremos que la incertidumbre es $S_n/\sqrt{n}$ con una probabilidad del 68 %. En este curso recomendamos utilizar $S_n/\sqrt{n}$ como medida para la incertidumbre estadística del promedio, por lo que el resultado se puede expresar como:

$$x = \bar{x} \pm S_n/\sqrt{n} \tag{9}$$

   Podemos observar que el desvío del promedio puede teóricamente ser disminuído arbitrariamente (tomando n suficientemente grande). Sin embargo en la práctica esto no es así. Aumentar el número de medidas implica seguramente un proceso de medición temporalmente más extenso, por lo que puede no ser razonable suponer que las medidas fueron tomadas bajo idénticas condiciones. Además tenemos un límite impuesto por la apreciación de los instrumentos, que determina una cota inferior que limita el valor mínimo que puede tomar la incertidumbre.

2. Dado que luego de realizar una única serie de medidas, el 68 % se concentra entre $\bar{x} \pm S_n$ otra opción es definir el intervalo de incertidumbre en relación al valor de S_n. Una elección posible es:

$$x = \bar{x} \pm S_n \tag{10}$$

5. Área bajo la gaussiana y descarte de medidas

El área bajo la curva gaussiana en un intervalo representa la probabilidad de encontrar una medida en dicho intervalo. Dado que el área bajo la gaussiana entre $\bar{x} - 2S_n$ y $\bar{x} + 2S_n$ es el 95,4 % del total y entre $\bar{x} - 3S_n$ y $\bar{x} + 3S_n$ es de 99,7 %, se deduce que la probabilidad de encontrar medidas por fuera de este último intervalo es extremadamente baja. Al analizar el histograma, debemos observar si aparecen datos atípicos. Estos datos son medidas que se encuentren en intervalos donde la probabilidad sea muy baja, lo que nos permite definir un *criterio de descarte*. Un criterio comúnmente usado es que se pueden descartar todas las medidas que se aparten de $\bar{x}$ más de $2S_n$ o $2,5S_n$.

Entonces, el procedimiento para obtener el resultado final luego de realizar una serie de n medidas, de una misma magnitud física en iguales condiciones experimentales sería:

- Calcular $\bar{x}$, $S_n(x)$ y $S_n(\bar{x})$.
- Realizar el histograma.
- Verificar si es necesario descartar medidas. En ese caso, se continúa trabajando con una nueva tabla de datos cuya longitud será n_1.
- Para la nueva serie de datos, se calcula nuevamente $\bar{x}_1$, S_{n1} y $S_{n1}(\bar{x})$.
- Se grafica el nuevo histograma y se superpone la correspondiente función gaussiana (con su respectiva normalización).
- Expresar el resultado final con su incertidumbre asociada.

6. Propagación de incertidumbres.

Para el cálculo de propagación de incertidumbres cuando se trabaja con magnitudes estimadas, es posible realizar el mismo análisis que en la repartido 1. En este caso, una estimación del valor del mensurando $\bar{y}$ al que nos referiremos como $\bar{y}$, se obtiene a partir de las cantidades estimadas $\bar{x_1}, \bar{x_2}, .., \bar{x_N}$ de tal modo que la estimación de $\bar{y}$, que es el resultado de la medida, está dado por:

$$\bar{\bar{y}} = f(\bar{x_1}, ...\bar{x_N}) \tag{11}$$

Los valores $\bar{x_1}, \bar{x_2}, .., \bar{x_N}$ tienen asociadas incertidubres $u(\bar{x_i})$. Por lo tanto el valor de y estará comprendido dentro de cierto intervalo de incertidumbre que habrá que determinar. La incertidumbre estándar combinada del resultado de la medida y, que designaremos $u_c(\bar{y})$ se calcula usando la ecuación:

$$u_c^2(y) = \sum_{j=1}^{n} (\frac{\partial f}{\partial x_i}\Big|_{\bar{x_i}})^2 u^2(\bar{x}_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (\frac{\partial f}{\partial x_i}\Big|_{\bar{x_i}})(\frac{\partial f}{\partial x_j}\Big|_{\bar{x_j}}) u^2(\bar{x}_i, \bar{x}_j) \tag{12}$$

En la ecuación 12, las derivadas parciales $(\frac{\partial f}{\partial x_i}\Big|_{\bar{x_i}})$, son llamadas coeficientes de sensibilidad, $u(\bar{x}_i)$ es la incertidumbre estándar asociada a $\bar{x}_i$, y $u(\bar{x}_i, \bar{x}_j)$ es la covarianza estimada asociada a $(\bar{x}_i)$ y $(\bar{x}_j)$ la cual se anula si las variables son estadísticamente independientes. Esta última hipótesis se cumplirá en la mayoría de las situaciones experimentales con las que trabajemos en este curso. Por lo tanto la ecuación 12 se reduce a:

$$u_c^2(\bar{y}) = \sum_{j=1}^{n} (\frac{\partial f}{\partial x_i}\Big|_{\bar{x_i}})^2 u^2(\bar{x}_i) \tag{13}$$