La colección de datos y la distribución gaussiana

El cálculo diferencial e integral está íntimamente ligado con el análisis de datos y las distribuciones de probabilidad.

En particular, en los laboratorios de la Facultad de Química de la UNAM, se acostumbra hacer mediciones de variables cuyos valores se distribuyen de acuerdo con la función gaussiana, lo cual permite un abordaje a través del cálculo integral y su asociación con los métodos y conceptos de la estadística descriptiva.

La colección de datos

La figura 1 muestra el resultado de las marcas que deja el lanzamiento de un balín sobre una hoja de papel puesta en el suelo. Los lanzamientos del balín se hacen con la intención de acertar en el centro de papel. Conviene numerar a las franjas: 1, 2, 3..., etcétera, para indicar que se trata de los posibles valores que puede tener una variable aleatoria; de hecho, podría ser cualquier serie de valores consecutivos.

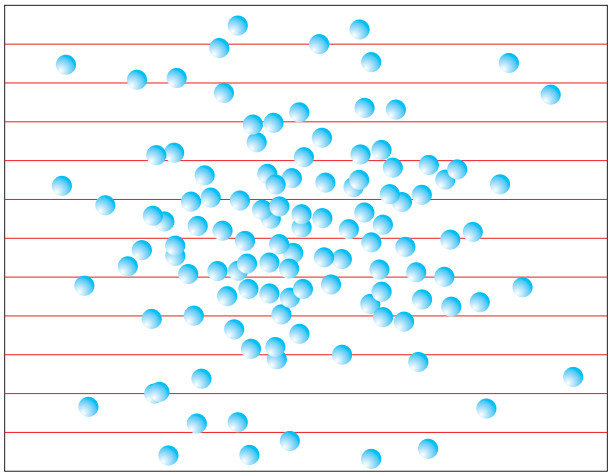

Figura 1. Marcas dejadas por el lanzamiento de un balín.

Los datos

Aquí se muestra un conjunto de datos generados aleatoriamente y que equivale al lanzamiento del balín.

0.06, 1.04, 1.21, 1.88, 2.06, 2.10, 2.14, 2.27, 2.47, 2.60, 2.70, 2.89, 3.07, 3.09, 3.30, 3.31, 3.50, 3.71, 3.78, 3.86, 3.93, 4.06, 4.15, 4.19, 4.22, 4.38, 4.40, 4.45, 4.55, 4.57, 4.66, 4.77, 4.86, 4.86, 4.87, 4.91, 4.98, 5.05, 5.08, 5.08, 5.13, 5.19, 5.19, 5.22, 5.37, 5.40, 5.43, 5.49, 5.55, 5.57, 5.59, 5.59, 5.66, 5.68, 5.69, 5.71, 5.72, 5.75, 5.90, 5.91, 5.91, 6.03, 6.13, 6.21, 6.29, 6.29, 6.42, 6.44, 6.50, 6.57, 6.60, 6.68, 6.72, 6.78, 6.83, 6.84, 6.85, 6.86, 7.03, 7.11, 7.26, 7.31, 7.61, 7.63, 7.64, 7.67, 7.72, 7.78, 7.83, 7.85, 7.86, 7.88, 7.89, 7.97, 7.99, 8.04, 8.10, 8.19, 8.23, 8.23, 8.28, 8.39, 8.39, 8.51, 8.57, 8.57, 8.62, 8.62, 8.66, 8.66, 8.69, 8.76, 9.10, 9.19, 9.24, 9.28, 9.35, 9.50, 9.60, 9.69, 9.76, 9.76, 9.86, 10.11, 10.12, 10.58, 10.65, 10.83, 11.13, 11.21, 11.30, 11.43, 11.48, 11.99, 12.01, 12.44, 13.59

Los datos se encuentran ordenados para facilitar la construcción del histograma correspondiente.

En la figura 2 se aprecia el histograma que muestra la distribución de datos presentados anteriormente.

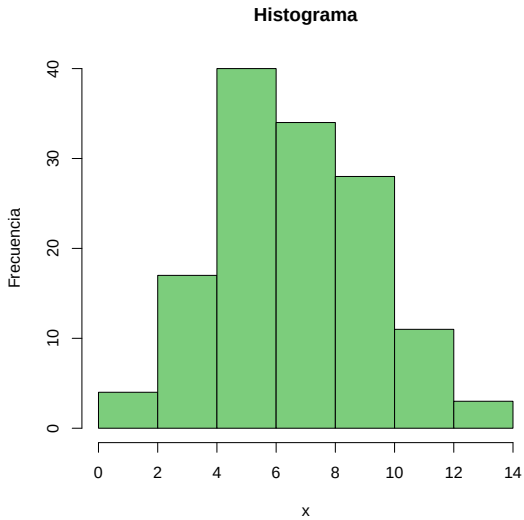

Figura 2. Histograma de los datos.

En algunas ocasiones es necesario cambiar el tamaño de la partición (o el número de intervalos de clase), como se muestra en la figura 3.

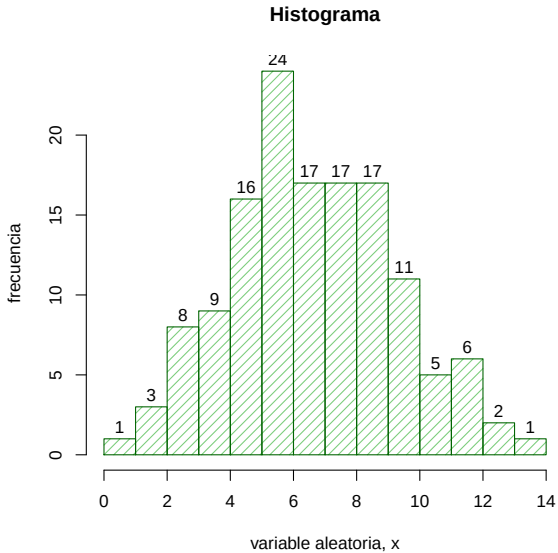

Figura 3. Histogramas de los datos.

A grandes rasgos, se puede ver que la distribución de frecuencias es más o menos gaussiana, como se ve en la figura 4

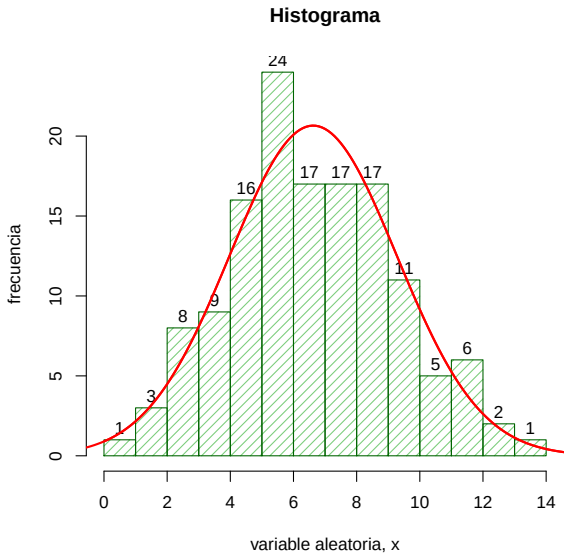

Figura 4. Histograma de los datos y la gaussiana asociada.

El probelma por resolver es encontrar los parámetros que permitan la construcción de la función gaussiana correspondiente al conjunto de datos:

$$f(x;\mu,\sigma) = \frac{A}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{1}$$

donde A es el área bajo la curva, μ es la media aritmética y σ es la desviación típica de la muestra.

Ahora determinemos cada uno de los parámetros mencionados.

En la figura 4 se hizo una partición de tal manera que la anchura de cada barra del histograma es unitaria, y en su parte superior se muestra el número de datos (frecuencia, que se conoce en el momento de construir el histograma) que pertenecen a cada barra.

Dado que el área de cada barra se puede calcular multiplicando la anchura, 1, de la barra por la frecuencia, f, correspondiente, es posible encontrar el área bajo el "histograma" (el área bajo la curva) si se toman en cuenta a todas las barras.
Así que

$$A = \sum_{i=1}^{N} \text{anchura}_i \times f_i \tag{2}$$

donde N es el número de barras en el histograma, ¡que es igual al tamaño de la partición!

$$A = 1 \cdot (1+3+8+9+16+24+17+17+17+11+5+6+2+1)$$

$$A = 137.$$

La media aritmética se determina como de costumbre

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

donde n es el número de datos registrados, por lo que

$$\mu = 6.624.$$

Finalmente se calcula la desviación típica de la muestra

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}} \tag{4}$$

así que

$$\sigma = 2.646.$$

Con esto ya es posible construir la función gaussiana asociada al histograma:

$$f(x;6.624,2.646) = \frac{137}{2.646\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-6.624}{2.646}\right)^2}. \tag{5}$$

Así, con la función gaussiana encontrada, es posible hacer una análisis de la misma para asociarla con algunos conceptos de la estadistica.

De acuerdo con los cursos de cálculo diferencial e integral, se pueden determinar los puntos críticos de una función aplicando la primera y segunda derivada de la función de interés.

El *máximo* local de la función $f(x;6.624,2.646)$ se encuentra calculando

$$\frac{d}{dx} f(x;6.624,2.646) = 0$$

y, resolviendo para x se tiene que el máximo corresponde a

$$x_{máx} = \mu = 6.624.$$

De la aplicación de la segunda derivada y su igualación con cero se encuentran los *puntos de inflexión*:

$$\frac{d^2}{dx^2} f(x; 6.624, 2.646) = 0$$

por lo que

$$x_{inflexión} = \mu \pm \sigma,$$

¡se tienen dos puntos de inflexión simétricos alrededor de la media, μ!: $a = \mu - \sigma$ y $b = \mu + \sigma$.

Pues bien, ahora hagamos uso del cálculo integral para estudiar el comportamiento de la gaussiana.

Calculemos primero el área bajo la curva, considerando solamente los valores mínimo y máximo del conjunto de datos, tomando en cuenta que $f(x; \mu, \sigma) > 0 \ \forall x \in \mathbb{R}$:

$$\int_{x_{mín}}^{x_{máx}} f(x; 6.624, 2.646)\, dx = 135.5243$$

¡ah!, casi igual que el valor de A(=137), ¿qué sucede si se cambian un poco los límites de integración?

$$\int_{x_{mín}-6}^{x_{máx}+6} f(x;6.624,2.646)\,dx = 136.9998$$

¡casi igual que A!... o prácticamente igual que A.

Sea

$$I_1 = \int_{x_{mín}-6}^{x_{máx}+6} f(x;6.624,2.646)\,dx = 136.9998$$

y cambiemos los límites de integración nuevamente...

$$I_2 = \int_{\mu-\sigma}^{\mu+\sigma} f(x;6.624,2.646)\,dx = 93.52846.$$

¿Qué fracción es I_2 de I_1?...

$$\frac{I_2}{I_1} = \frac{93.52846}{136.9998} = 0.6826905,$$

¡representa aproximadamente el 68 % del área total bajo la curva!

¿Qué sucede si se cambian nuevamente los límites de integración?

$$I_3 = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x; 6.624, 2.646)\, dx = 130.7665.$$

¿Qué fracción es I_3 de I_1?...

$$\frac{I_3}{I_1} = \frac{130.7665}{136.9998} = 0.9545014,$$

¡representa un poco más que el 95 % del área total bajo la curva!

¿Cuál debe ser el factor multiplicativo de σ en los límites de integración para que la fracción del área bajo la curva sea del 95 %? Cambiemos los límites de integración

$$I_4 = \int_{\mu-1.96001\sigma}^{\mu+1.96001\sigma} f(x; 6.624, 2.646)\, dx = 130.1507.$$

¿Qué fracción es I_4 de I_1?...

$$\frac{I_4}{I_1} = \frac{130.1507}{136.9998} = 0.9500065,$$

¡representa prácticamente el 95 % del área total bajo la curva!

Podríamos llamarle *factor de cobertura*, k, al factor multiplicativo, de tal manera que proporcione la fracción deseada del área bajo la curva.

La figura 5 muestra los resultados.

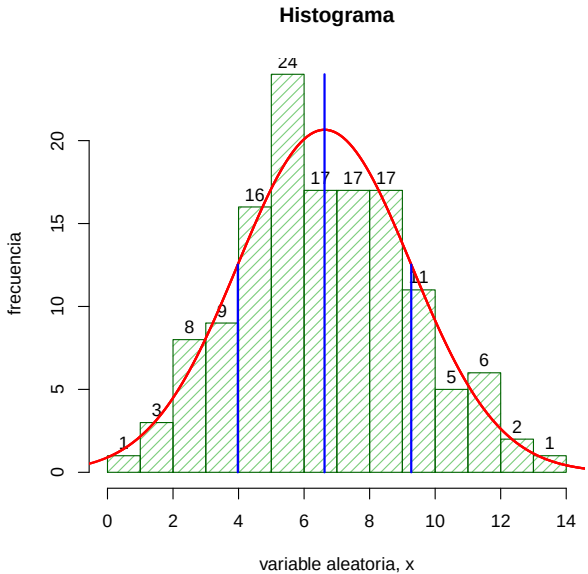

Figura 5. El histograma, la gaussiana asociada y los parámetros correspondientes.

O bien

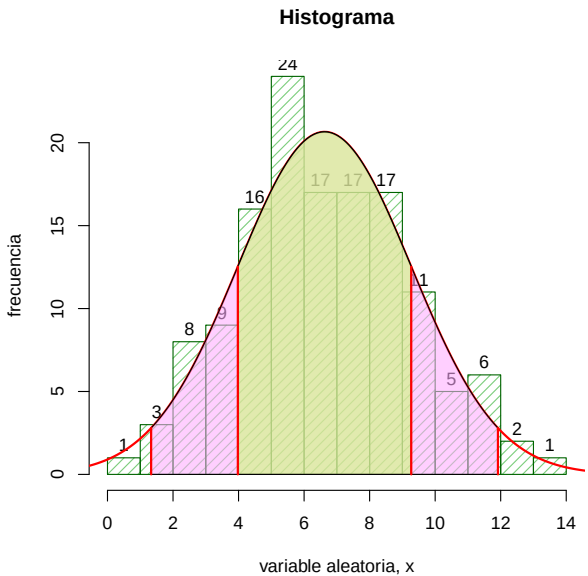

Figura 6. El histograma, la gaussiana asociada y las áreas bajo la curva.

El intervalo de confianza

¿Qué relación tiene todo esto con los *intervalos de confianza*?

Veamos un ejemplo. Se usa una máquina para llenar recipientes con 250 g de margarina. Dado que existe la posibilidad de que el llenado sea algo diferente, se considera que se trata de una variable aleatoria X. Se supone que la variación está distribuida normalmente alrededor del promedio de 250 g, con una desviación típica de 2.5 g. Para determinar si la máquina está calibrada adecuadamente, se toma una muestra al azar de $n = 25$ recipientes de margarinta y se mide la masa.

Las masas medidas $X_1, X_2, ..., X_{25}$ son una muestra aleatoria de X. Si se considera al estimador de la media, etnonces la muestra es $x_1, x_2, ...x_{25}$, así que

$$\bar{x} = \frac{1}{25} \sum_{i=1}^{25} x_i = 250.2 \text{ g}.$$

Así, existe un intervalo alrededor del valro observado de 250.2 g de la media de la muestra dentro del cual lso datos observados no serían considerados particularmente inusuales. Tal intevalo es el *intevalo de confianza* para la media.

¿Cómo se calcula dicho intervalo?

Dado que se ha supuesto que se tiene una distribución normal, entonces

$$\frac{\sigma}{\sqrt{n}} = 0.5 \text{ g}$$

y estandarizando se obtiene

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{0.5}.$$

Así, es posible encontrar números $-z$ y z, independientes de la media, entre los cuales se encuentra Z con una probabilidad dada por $1 - \alpha$, una medida de cuánta confianza se desea.

Sea $1 - \alpha$ =0.95, po rejemplo. Entonces se tiene

$$P(-z \leq Z \leq z) = 1 - \alpha = 0.95$$

Así,

$$\Phi(z) = P(Z \leq z) = 1 - \frac{a}{2} = 0.975,$$

$$z = \Phi^{-1}(\phi(z)) = \Phi^{-1}(0.975) = 1.96,$$

entonces

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) = P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$

$$0.95 = P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

En otras palabras, el extremo inferior del 95% del intervalo de confianza es

$$Límite\ inferior = \bar{X} - 1.96\frac{\sigma}{\sqrt{n}},$$

y el extremo superior del 95% del intervalo de confianza es

$$Límite\ superior = \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$

Así,

$$(\bar{x} - 0.98;\ \bar{x} + 0.98) = (249.22;\ 251.18).$$

Por lo tanto se debe decir "μ está en el intervalo de confianza, con un nivel de confianza de $100(1-\alpha)\%$."

La figura 7 muestra al histograma, la gaussiana y los intervalos de confianza al 80% y 90%.

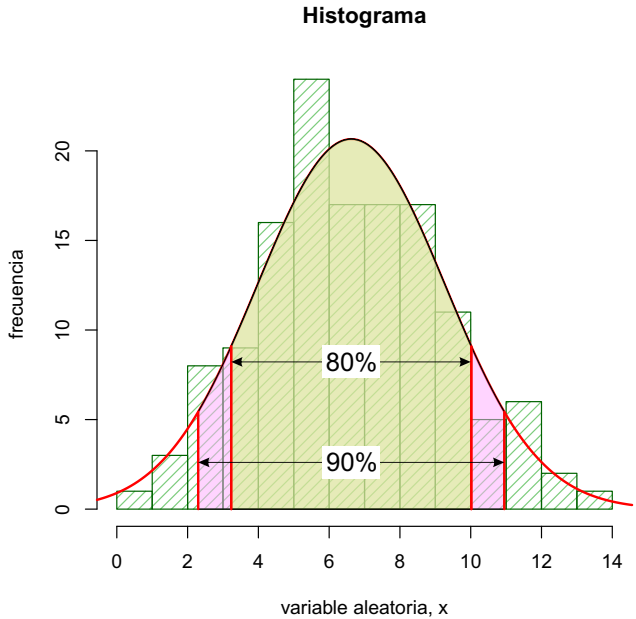

Figura 7. El histograma y los intervalos de confianza al 80% y 90%..

F I N