Métodos de Estadística Descriptiva Univariante

En el trabajo experimental hay ocasiones en las que es preciso realizar un análisis estadístico descriptivo de una variable aleatoria X. El resultado de dicho análisis consiste en la obtención de una serie de medidas, y a las que se conoce como medidas de centralización, dispersión y forma:

- Medidas de centralización: Su finalidad es informar de la tendencia de los datos.

  - ✓ Tamaño muestral (n): No es en sí una medida de centralización aunque suele incluir en este grupo. Es el número de individuos, elementos u objetos que componen la muestra. Una muestra grande es la que consta de más de 30 individuos (n>30), siendo pequeña cuando dicho número es inferior a 30 (n<30).

  - ✓ Media aritmética ($\bar{x}$): Es el valor promedio de las observaciones. Es una de las medidas de centralización más importantes.

  - ✓ Media geométrica (m_g): Es un valor medio que es utilizado con porcentajes, tasas etc.

  - ✓ Media armónica (m_a): Es un valor medio muy robusto a valores extremos con utilidad en el cálculo del promedio de velocidades, tiempos, rendimientos etc.

  - ✓ Media cuadrática (m_c): Es un valor medio que es utilizado en el cálculo del promedio de corrientes alternas, ondas, gases, etc. eliminando el efecto del signo cuando algunas observaciones toman valores negativos. Se conoce también como RMS o valor cuadrático medio.

  - ✓ Mediana (Me): Si ordenamos los valores de las observaciones de menor a mayor la mediana es aquella observación que deja al 50% de las observaciones por encima y al otro 50% restante por debajo. Se conoce con otros nombres como "cuartil dos", "percentil 50".

- ✓ Moda (Mo): Es el valor más frecuente en las observaciones.

- ✓ Cuantiles: Se trata de la aplicación de lo que se conoce como teoría elemental de rangos, introducida por el estadístico Kendall en 1940. Si ordenamos los valores de las observaciones de menor (MIN) a mayor (MAX), tendremos los siguientes valores o cuantiles.

  Si a continuación dividimos la muestra en cuatro partes iguales entonces Q1 (primer cuartil) será el valor que deja el 25% y el 75% de las observaciones a izquierda y derecha, Q2 (segundo cuartil) es la mediana (Me), y Q3 (tercer cuartil) que es aquel valor que dejará el 75% y el 25% de las observaciones a izquierda y derecha. Ahora bien, si dividiéramos la muestra en 100 partes iguales entonces tendremos percentiles. Por ejemplo, P25, P50 y P75 se corresponden con Q1, Q2 y Q3 respectivamente.

$$\overline{x} = \frac{\sum x_i}{n}$$

$$m_g = \sqrt[n]{x_1 . x_2 . \ldots . x_n}$$

$$m_a = \frac{n}{\sum \frac{1}{x_i}}$$

$$m_c = \sqrt{\frac{\sum x_i^2}{n}}$$

- Medidas de dispersión: Nos informan de la variabilidad o dispersión de los datos, y por tanto del grado de representatividad de las medidas de centralización. A menor variabilidad, más representativa será la medida de centralización obtenida.

- ✓ Rango (r): Se define como la diferencia entre el valor máximo y el mínimo de las observaciones, es decir Max(x)-min(x). Se utiliza en experimentos de control de calidad.

- ✓ Recorrido intercuartílico (IQR): Es una medida de variabilidad definida a partir de la teoría de cuantiles, siendo igual a Q3-Q1.

- ✓ Desviación media (D_m): Es un buen indicador de dispersión definido como un promedio de las diferencias en valor absoluto de cada observación con respecto a la media aritmética.

- ✓ Varianza (s^2): Es la medida de dispersión más importante en el trabajo experimental, también denominada como error cuadrático medio. Se define como un promedio de las diferencias al cuadrado de cada observación con respecto a la media aritmética. No confundir con la cuasivarianza, una versión de estimador insesgado de la varianza (su esperanza matemática es la varianza poblacional, $E[s^2] = \sigma^2$) :

$$s^2 = s^2 . \frac{n}{n-1}$$

- ✓ Desviación típica (s): Es la raíz cuadrada de la varianza (s^2). De esta forma las unidades se expresan linealmente y no al cuadrado tal y como ocurre con la varianza.

- ✓ Error estándar: Puesto que la media muestral es una variable aleatoria, ya que su valor fluctúa de muestra a muestra, el error estándar es la desviación típica de la media aritmética. Se utiliza en inferencia estadística, aunque a veces su valor se incluya en estadística descriptiva.

- ✓ Coeficiente de variación de Pearson (CV): Se define como el cociente entre la desviación típica y la media muestral. Suele multiplicarse por 100 y permite comparar la variabilidad entre dos variables cualesquiera.

$$D_m = \frac{\sum \left| x_i - \bar{x} \right|}{n}$$

$$s^2 = \frac{\sum \left(x_i - \bar{x} \right)^2}{n}$$

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x} \right)^2}{n}}$$

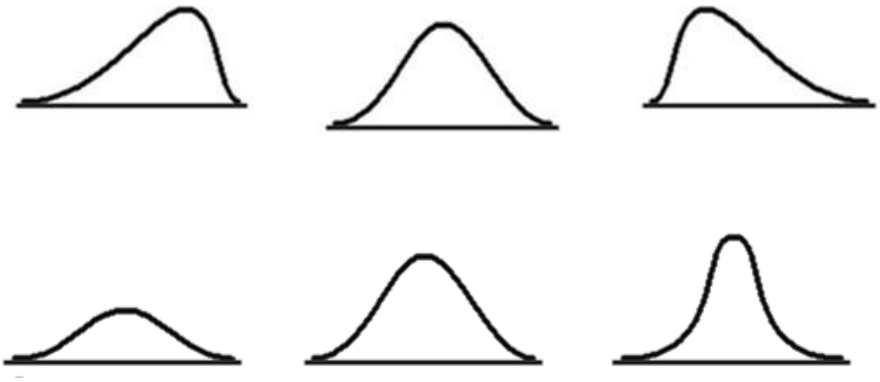
$$CV = \frac{s}{\bar{x}}.100$$

- Medidas de forma: Su finalidad es medir la forma de la distribución de frecuencias de los datos. La forma de la distribución se cuantifica con dos medidas:

  - ✓ Asimetría (g_1 o "skewness"): Si la distribución es simétrica entonces g_1 será próximo a cero. En caso contrario, cuando $g_1>0$ entonces un número considerable de datos se concentran en los valores altos de la distribución, estando ésta sesgada a la derecha. Si $g_1<0$ entonces la distribución estará sesgada hacia los valores bajos, es decir hacia la izquierda.

  - ✓ Curtosis (g_2 o "kurtosis"): Es una medida del grado de apuntamiento de la distribución de los datos. Si g_2 es cero la distribución es mesocúrtica, si $g_2>0$ entonces se llama leptocúrtica, y si $g_2<0$ entonces la distribución se denomina platicúrtica. En una distribución leptocúrtica la

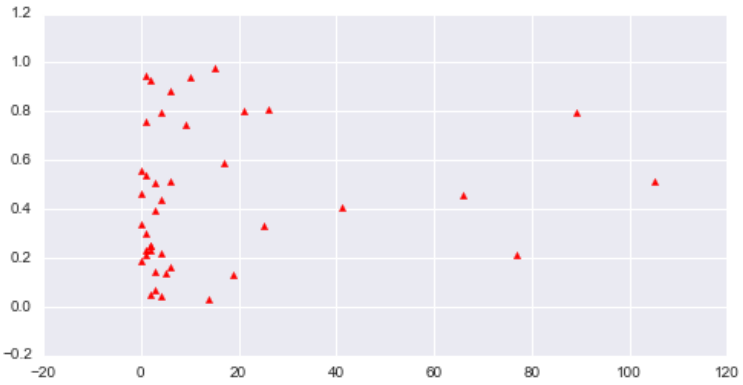
varianza s^2 es muy baja, mientras que en una distribución platicúrtica la varianza s^2 es muy alta. Por consiguiente, la variabilidad tiene un valor medio únicamente cuando la distribución es mesocúrtica.



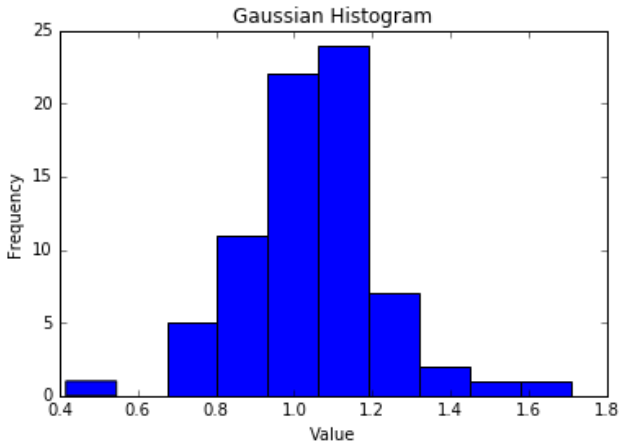
Forma de una distribución. (Superior): Distribuciones sesgada a la izquierda ($g_1<0$), insesgada ($g_1=0$) y sesgada a la derecha ($g_1>0$). (Inferior): Distribuciones platicúrtica ($g_2<0$), mesocúrtica ($g_2=0$) y leptocúrtica ($g_2>0$).

Además de las medidas de centralización, dispersión y forma la estadística descriptiva utiliza métodos gráficos. Los gráficos más utilizados en el trabajo experimental son:

- Diagrama de dispersión ("Scatterplot"): Representación de las observaciones en la que el eje X o de abcisas representa los valores de los datos u observaciones, y el eje Y u ordenadas es un valor ficticio que se obtiene simulando la "agitación" o "jitter" del eje X, permitiendo así una visualización más cómoda de las observaciones.

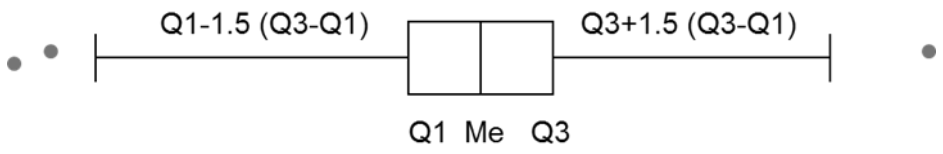


- Histograma: Es uno de los gráficos más populares en estadística descriptiva. Representa la distribución de frecuencias en variables cuantitativas que sean continuas. No confundir con los diagramas de barras utilizados con variables cuantitativas discretas.



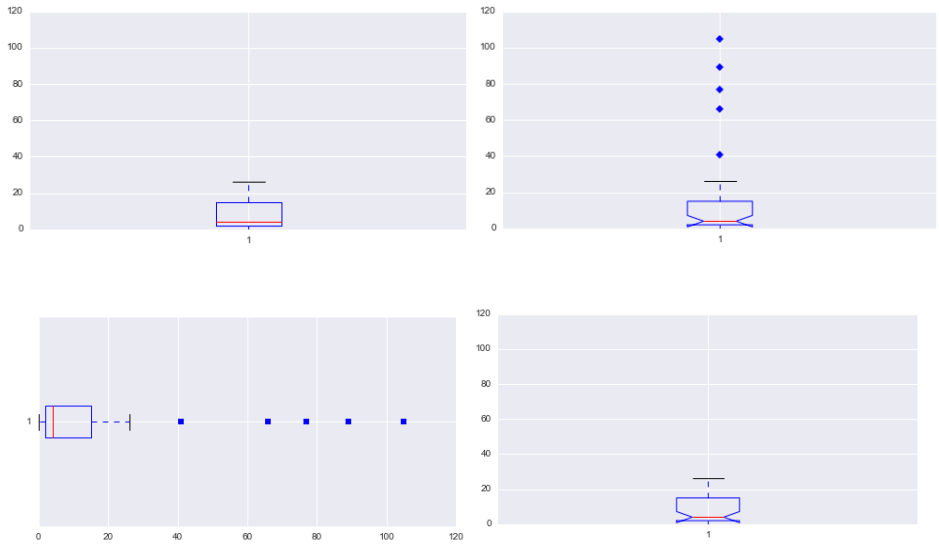
En un histograma las barras son rectángulos que representan intervalos y tienen una altura proporcional a la frecuencia de cada intervalo, una vez que los datos son agrupados y organizados en una tabla de clases y frecuencias.

- Gráfico de caja y bigotes ("Box and whisker plot"): Es un gráfico actualmente muy utilizado en el trabajo experimental. Fue introducido por el estadístico Tukey en 1977 popularizándose con el uso extendido de los ordenadores.
  Permite analizar (a) la simetría de la distribución, (b) identificar valores atípicos o extremos de los datos u observaciones ("outliers") y (c) comparar entre sí lotes de datos.



Los valores extremos mínimo y máximo de los "bigotes" se calculan con las expresiones que se muestran a izquierda y derecha respectivamente. El tamaño de la "caja" representa los valores de los cuartiles primero, segundo o mediana

y tercero. En ocasiones una pequeña cruz u otra marca dentro de la "caja" indica la media aritmética. Los puntos más alla de los bigotes son los valores atípicos o "outliers".



- Gráfico de probabilidad normal: Permite evaluar la normalidad de la variable aleatoria objeto de estudio, comparando los datos con la distribución normal. Es decir este gráfico evalúa si los datos proceden o no de una distribución normal o Gauss de forma empírica, a partir del ajuste de los puntos que representan a los datos experimentales a una línea recta que representaría a la distribución normal.