Manual práctico de *Quimiometría*

1ª Edición (2011)

Germán Tortosa Muñoz
Departamento de Microbiología del Suelo y Sistemas Simbióticos
Estación Experimental del Zaidín (EEZ-CSIC)
http://www.compostandociencia.com

ÍNDICE

Capítulo 1. Introducción

La Química Analítica tiene en la Estadística una de sus herramientas fundamentales. Esta imprescindible relación ha dado lugar en los últimos años al desarrollo de la ***Quimiometría***, una disciplina que aplica las técnicas matemáticas de la estadística a los problemas analíticos de la identificación y cuantificación de las sustancias químicas, siendo habitual el uso de la quimiometría en cualquier análisis químico. En la actualidad, esta disciplina ha ganado importancia debido sobre todo por cantidad de información que obtenemos a través de los equipos instrumentales (los cuales generan una gran cantidad de datos numéricos) y por el incremento en la capacidad de cálculo de los ordenadores actuales.

La Estadística describe el comportamiento aleatorio de las variables analíticas que usamos en el laboratorio. Así, se puede usar para deducir las leyes de la ***probabilidad*** que rigen dichos comportamientos, con el fin de hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones. Por lo tanto, podemos clasificar a la ***estadística*** como ***descriptiva***, la cual nos dará solo información detallada de un conjunto de datos, e ***inferencial,*** cuando el objetivo del estudio se centra en derivar las conclusiones obtenidas de nuestro estudio a un conjunto de datos más amplio, es decir, hacer predicciones de los comportamientos de las variables analíticas.

Así, el siguiente manual describe conceptos básicos de probabilidad, de estadística descriptiva e inferencial, siempre desde un punto de vista práctico y aplicado al análisis químico. Fruto de ese enfoque práctico, se ofrecen alternativas para el desarrollo de los estudios estadísticos a través de numerosos recursos gratuitos disponibles actualmente en internet. Es importante constar que para una mayor profundización en los conceptos teóricos aquí comentados, se recomienda consultar los textos didácticos referenciados que se han seguido para elaborar este texto, así como las diversas fuentes de información comentadas en el último capítulo de este manual.

Este trabajo está dedicado al Dr. Ignacio F. López García (Universidad de Murcia) por sus enseñanzas universitarias en quimiometría, al Dr. Félix Belzunce Torregrosa (Universidad de Murcia) por trasmitirme sus conocimientos matemáticos de estadística y a la Dra. Diana Marco (Universidad Nacional de Córdoba), por sus importantes comentarios en la aplicación de la estadística en ecología.

Nota: Muchos de los ejemplos comentados en este Manual están disponibles en internet a través de Applets desarrollados en Javascrip, por lo que se recomienda instalar un software para ello (http://java.com/es/).

Capítulo 2. Conceptos básicos y Probabilidad.

Antes de profundizar en técnicas estadísticas, es conveniente aclarar algunos conceptos básicos de probabilidad, los cuales nos ayudarán a entender mejor los principios matemáticos usados en quimiometría.

El primero de ellos es el ***Fenómeno aleatorio***, el cual se puede definir como aquel en que los resultados son inciertos, imprevisibles o impredecibles (ejemplo: la medición de la concentración de nitratos en agua de un rio de cauce natural). El caso contrario sería el ***Fenómeno determinista,*** en el cual si podemos saber los resultados al estar descritos por modelos matemáticos (ejemplo: el tiempo que tardará un coche en llegar a su destino a una velocidad constante).

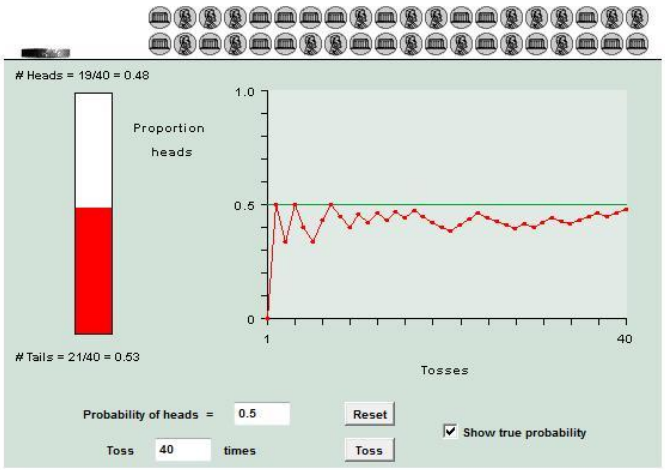
Otro concepto importante es el de ***Población,*** que conjunto global del sistema que queremos estudiar (ejemplo: una encuesta sobre intención de voto de un país de 40 millones de habitantes). Normalmente es imposible optar a su totalidad, por lo en estos casos se suele coger una muestra representativa de la misma. El subconjunto de la población al cual si tenemos acceso para estudiar el comportamiento de la misma se denomina ***Muestra poblacional***. Esta debe ser representativa de la población y cogida de la forma más imparcial posible (totalmente al azar), a la cual se le estudia una ***Variable***, que se define como aquella propiedad que es observable y medible (ejemplo: la masa de una persona después de una cena medida en un peso).

Veamos una representación gráfica de estos conceptos:

2.1. El concepto de ***Probabilidad***.

Para entender este concepto realizaremos un sencillo experimento que se basará en el lanzamiento de una moneda repetidas veces. Como sabemos, los resultados posibles de cada lanzamiento se resumen en dos: cara o cruz. El lanzamiento de la moneda en un ***Fenómeno aleatorio*** ya que no sabemos con exactitud el resultado que saldrá cada vez que tiramos la moneda.

Imaginemos que lanzamos una moneda al aire 40 veces. Con los resultados obtenidos calculamos la ***Frecuencia relativa***, decir, el cociente entre el número de veces que se obtiene un resultado deseado (las veces que ha salido cara o cruz) con respecto al total de veces que se realiza. Finalmente, el resultado lo representamos gráficamente y obtenemos lo siguiente:



Sacado de http://bcs.whfreeman.com/ips4e/cat_010/applets/Probability.html

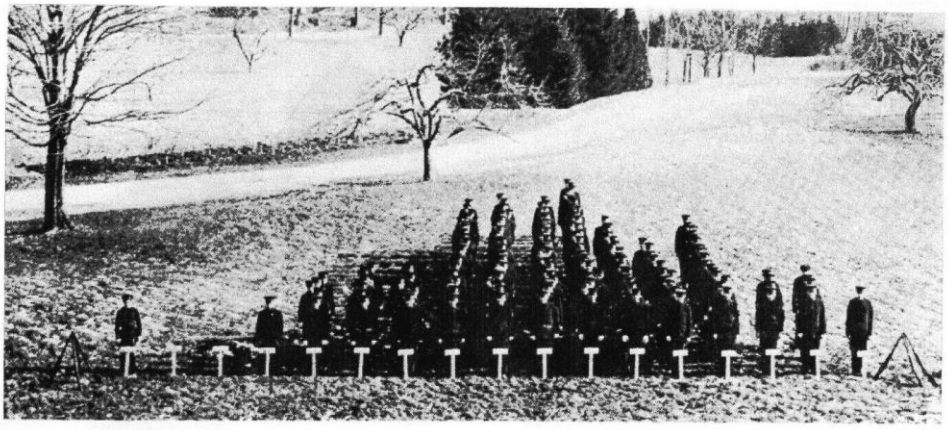
Como podemos observar, la frecuencia relativa tiende a 0,5 conforme el número de tiradas es mayor, o lo que es lo mismo, la probabilidad de tirar una moneda y sea cara o cruz es del 50%.

Nota: si no tienes paciencia de tirar monedas, en la siguiente web puedes obtener los resultados de este experimento (http://www.ematematicas.net/simulacionmoneda.php).
También puedes verlo en esta web (está en inglés):
http://bcs.whfreeman.com/ips4e/cat_010/applets/Probability.html

Así, podemos definir la ***probabilidad*** del suceso (en nuestro caso el lanzamiento de monedas) como el valor al cual tiende la frecuencia relativa en un experimento. Por lo tanto, con la probabilidad podemos conocer el comportamiento que rigen los fenómenos aleatorios que estudiamos y estimar su resultado.

2.2. Distribución de datos aleatorios.

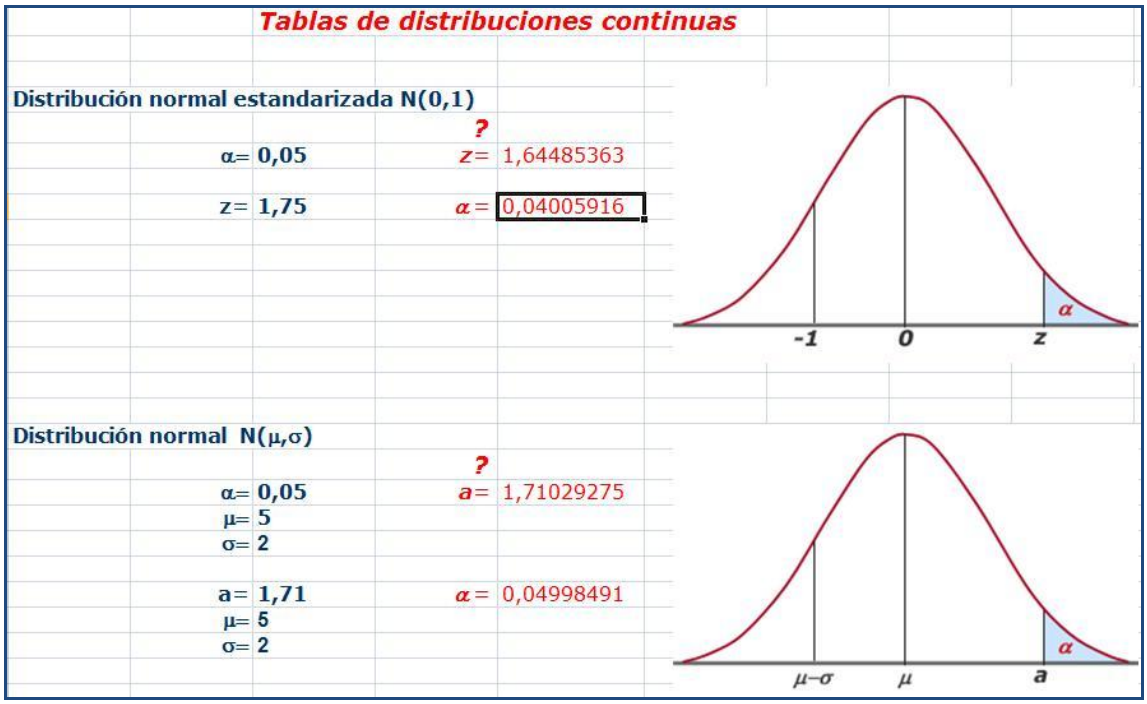
Al estudiar un fenómenos aleatorio, concretamente una variable de una muestra de una población (ejemplo: altura de un destacamento de soldados de un ejército), los resultados podemos representar en forma de ***histogramas***, es decir, representando la frecuencia de los resultados obtenidos.

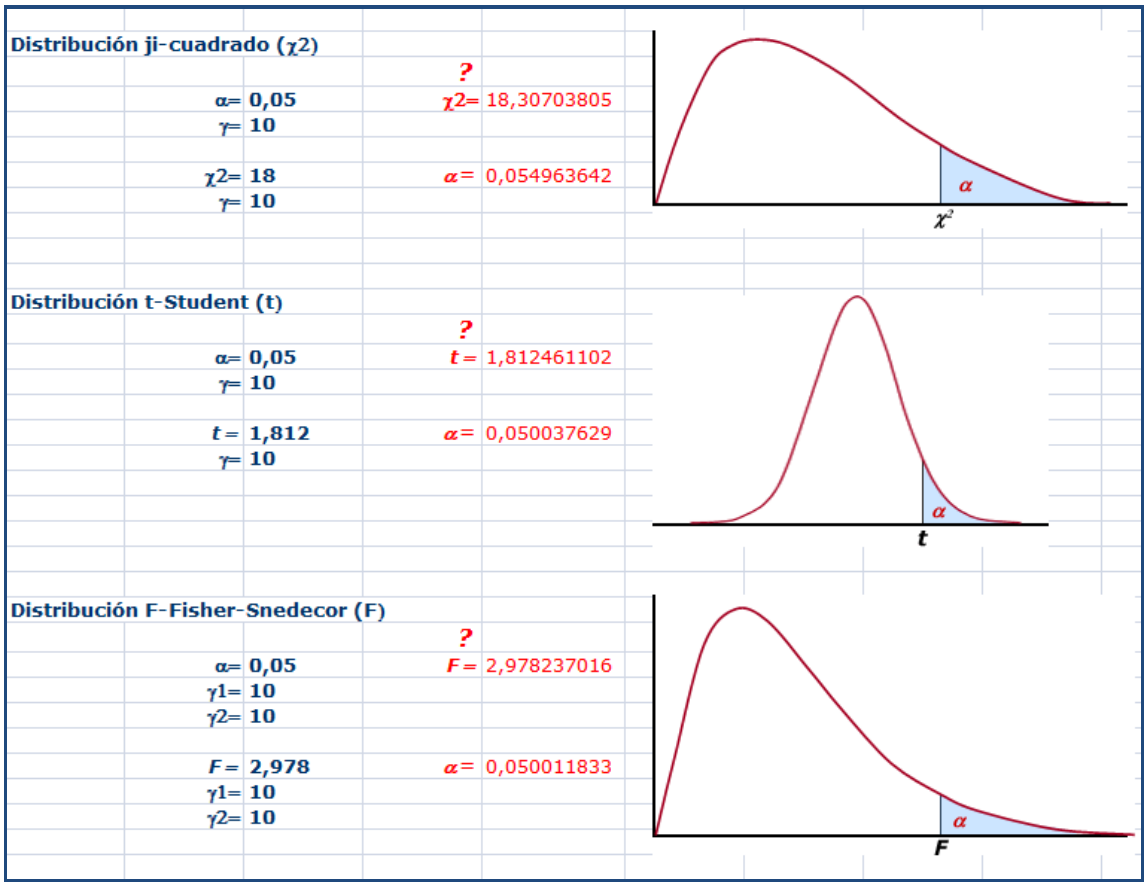


Número de individuos	1	0	0	1	5	7	7	22	25	26	27	17	11	17	4	4	1
Estatura en pulgadas	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74

Nota: no se conoce la fuente exacta de esta imagen.

En este caso vemos que la distribución de los datos aleatorios obtenidos puede presentar comportamientos definidos. En este caso, la altura de los soldados presenta una distribución de campana de Laplace-Gauss y se denomina ***distribución normalizada***. Existen muchos ejemplos de distribuciones, como las distribuciones binomiales, Chi-cuadrado, F de Fisher-Snedecor, etc.





Ejemplos sacados de http://sebbm.bq.ub.es/BioROM/contenido/UIB/bioinfo/index.htm

)

2.3. El caso de la distribución normalizada

Como comentan Miller y Miller (2009), la distribución normalizada es común en los análisis cuantitativos que requieren muestras repetidas. De hecho, aproximadamente el 90% de los métodos estadísticos se basan en que los datos aleatorios se rigen por una distribución normalizada, siendo esta distribución es muy importante en análisis químico. La distribución normalizada se caracteriza por ser simétrica con respecto a un valor central denominado μ, siendo la ecuación matemática que la describe la siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R},$$

donde σ es un parámetro que nos da información de la dispersión de los datos, es decir, sobre la anchura de la campana de Gauss.

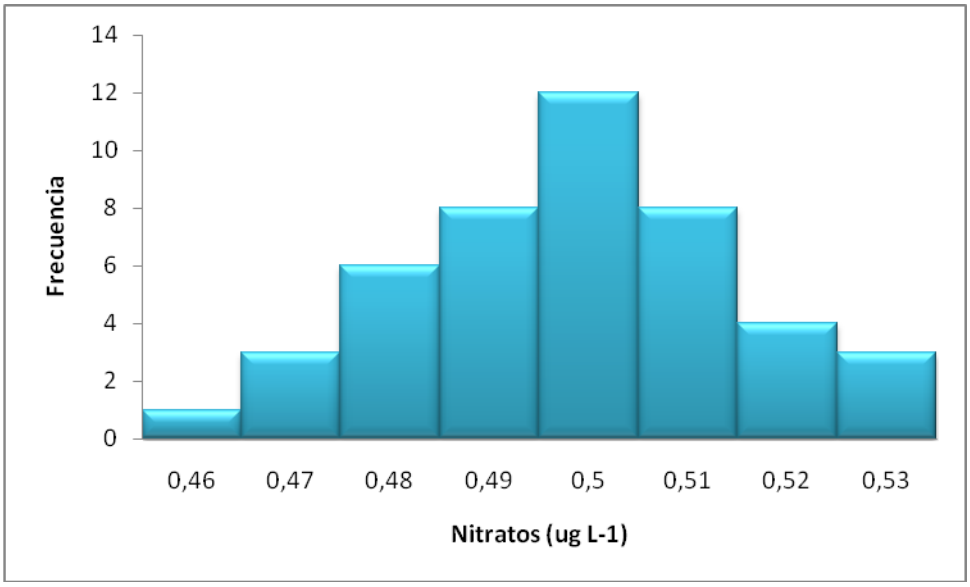
Nota: con el fin de calcular la probabilidad, a esta ecuación se le aplica una transformación matemática para estandarizarla, es decir, para que el área de la campana sea 1 (o el 100%).

Veamos un ejemplo de química analítica, un análisis cuantitativo de la concentración de nitrato en una muestra de agua:

Resultados de la concentración de nitratos en agua (µg L^{-1})								
0,51	0,50	0,50	0,50	0,50	0,49	0,52	0,50	0,47
0,51	0,52	0,53	0,48	0,49	0,50	0,52	0,49	0,50
0,49	0,48	0,46	0,49	0,49	0,48	0,49	0,51	0,47
0,50	0,51	0,51	0,48	0,48	0,47	0,50	0,49	0,48
0,50	0,50	0,50	0,53	0,53	0,52	0,50	0,51	0,51

Ejemplo sacado de Miller y Miller (2009).

Estos son los resultados obtenidos de una misma muestra de agua en un laboratorio de análisis. Si representamos en forma de histograma observamos que se rigen mediante una distribución normal o Gausiana.



Como hemos comentado antes, la curva normalizada o campana de Gauss es una curva simétrica respecto a un valor central µ (que si no existe error sistemático en el equipo de medida coincide con la media aritmética de las muestras) y con el parámetro σ como medida de la dispersión de los datos (siendo esta la desviación estándar). Cuanto mayor sea esta, la campana de Gauss será más grande como puede verse en la siguiente figura.

Nota: Los conceptos de media y desviación estándar se verán más adelante y están relacionados con la exactitud y precisión respectivamente.

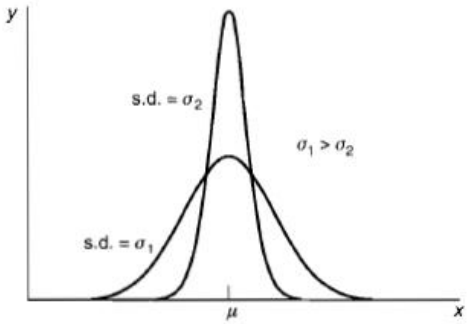


Figura 2.3. Distribuciones normales con la misma media pero con diferentes valores de la desviación estándar (d.e.).

Ejemplo sacado de Miller y Miller (2009).

Una de las características más importantes de esta distribución la podemos observar en la siguiente Figura:

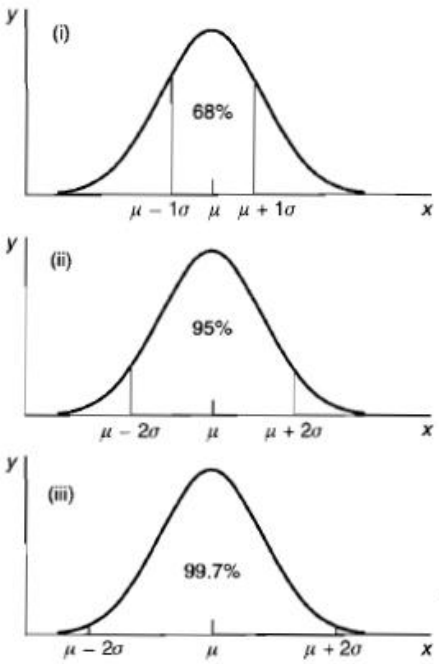


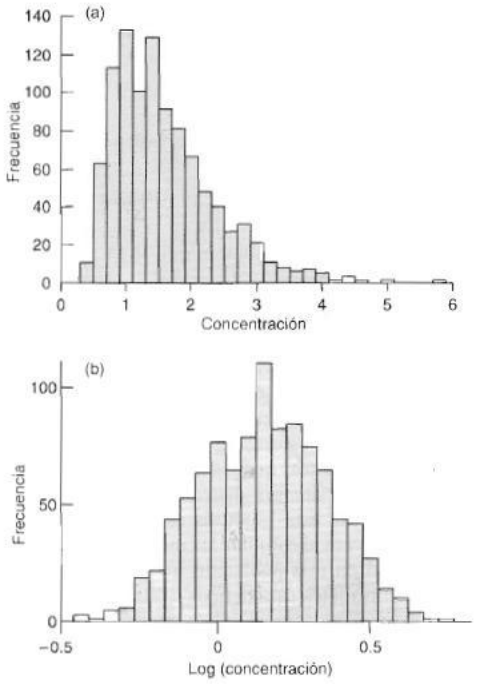
Figura 2.4. Propiedades de la distribución normal: (i) aproximadamente el 68% de los valores caen dentro de $\pm 1\sigma$ de la media; (ii) cerca del 95% de los valores se ubican dentro de $\pm 2\sigma$ de la media; (iii) aproximadamente el 99.7% de los valores se encuentran dentro de $\pm 3\sigma$ de la media.

Ejemplo sacado de Miller y Miller (2009).

Al ser simétrica con respecto al valor medio μ, podemos estimar la cantidad de muestras que están en la campana de Gauss con ayuda del parámetro σ, el cual nos da información de la dispersión de datos, Así, y tal y como se puede leer en la leyenda de

esta Figura, podemos saber que el 68% de los datos que se distribuyen normalmente están comprendido en el intervalo $\mu \pm \sigma$, el 95% en el intervalo $\mu \pm 2\sigma$ y finalmente, el 99,7% en el intervalo $\mu \pm 3\sigma$.

Como ya hemos comentado, este tipo de distribución es muy importante siendo referencia para muchas pruebas estadísticas. Cuando nuestros datos no siguen esta distribución, una alternativa es transformarlos matemáticamente mediante el cálculo del logaritmo, tal y como podemos ver en el siguiente ejemplo.



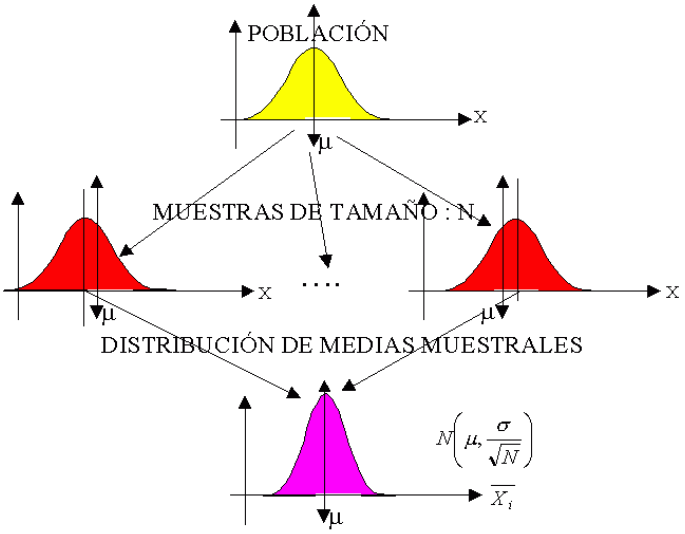
Ejemplo sacado de Miller y Miller (2009). Se muestra la concentración de anticuerpos de inmunoglobulina M en suero de varones y su transformación logarítmica para conseguir la normalización de los datos.

2.4. Muestras representativas y Teorema Central del límite.

Como ya hemos comentado anteriormente, cuando analizamos un fenómeno aleatorio es imposible acceder a la totalidad de la población por lo que accedemos a una muestra de la misma. Dicha muestra debe ser lo suficientemente representativa como para poder extrapolar las conclusiones obtenidas con este subconjunto a la totalidad de la población.

¿Cómo podemos saber si una muestra es representativa o no?, ¿qué criterios debemos seguir para obtener una muestra representativa?

Veamos el ***Teorema Central del Límite***. Este nos dice que cuando analizamos una población de datos mediante varios subgrupos de muestras representativas (ejemplo: queremos saber la población media de los hombres de un país y se cogen varios grupos de individuos correspondientes a las principales ciudades del mismo), cada una de ellas puede obtener una distribución distinta entre ellas (Los hombres de las ciudades del norte son más bajos, y los del sur son más variables, etc.). Si aumentamos el número de ciudades que estudiamos (como por ejemplo a 30), la distribución de las medias de las alturas de todas las ciudades tenderá a una distribución normalizada. Este ejemplo lo podemos visualizar en la siguiente figura:



Ejemplo sacado de http://terra.es/personal2/jpb00000/testimaciondelamedia.htm

Nota: para una visualización más explicativa, consultar http://terra.es/personal2/jpb00000/ttcentrallimite.htm

La principal conclusión práctica de este ejemplo es que el número de muestras debe ser siempre mayor o igual a 30 para que el tamaño de la muestra poblacional sea lo suficientemente representativo de la totalidad de la población.

Nota: Existen fórmulas para calcular exactamente este número dependiendo de varios criterios como el nivel de confianza que queremos.

Aunque las teorías estadísticas nos indiquen el tamaño muestral necesario para nuestros experimentos, muchas veces esos valores están en contraposición con la viabilidad técnica y económica de quien hace los estudios. Realizar un experimento con una carga analítica de más de 30 muestras por cada tratamiento puede ser inviable en muchos casos.

La solución a este problema no es fácil, y se tiene que llegar a una relación de compromiso teniendo en cuenta la siguiente premisa: cuanto mayor sea el número de repeticiones, más potencia tendrá nuestro estudio estadístico, y por tanto, las conclusiones que saquemos.

Capítulo 3. Estadística descriptiva.

Como comentábamos en el Capítulo 1, al aplicar la estadística podremos obtener información de un conjunto de datos utilizando su totalidad (***estadística descriptiva***) o a partir de una parte de este conjunto de datos (***inferencia estadística***). Es decir, la primera de ellas se utiliza para un análisis descriptivo del conjunto de datos y la otra para poder hacer predicciones de las características de una población de datos, de los cuales solo podemos acceder a un subconjunto.

Algunas de las preguntas que podemos resolver usando la inferencia estadística serían las siguientes:

- *Ante un fenómeno aleatorio, ¿Cuál es el modelo probabilística que describe dicho fenómeno?, y conocido dicho modelo, ¿cuáles son los parámetros que lo caracterizan?*

Para resolver este tipo de preguntas, primero tenemos que profundizar en la estadística descriptiva, ya que la inferencial se basa en esta última (como verenos en el siguiente capítulo).

3.1. El concepto de *Estadístico.*

Usando la terminología matemática de la estadística, podemos definir como ***estadístico*** a cualquier transformación matemática de una muestra aleatoria simple.

Para ilustrar este concepto, podemos ver las definiciones de los siguientes estadísticos:

Media aritmética:

$$\overline{X} = \frac{\sum X_i}{n}$$

Desviación estándar (DE):

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Varianza:

$$s^2(x) = (\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \overline{x}^2 \qquad \text{Varianza} = \text{DE}^2$$

Coeficiente de variación (CV):

$$\text{Coeficiente de variación} = \text{Desviación estándar relativa} = 100\frac{DE}{\overline{X}}$$

Existen muchos más ejemplos de estadísticos como la moda, la mediana, la media geométrica, etc., aunque los anteriores son los fundamentales.

Nota: Para tener una idea de otros conceptos, visitar
http://es.wikipedia.org/wiki/Parametro_estadistico
Para realizar este tipo de cálculos, pinchar aquí:
http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Descriptive.htm
http://graphpad.com/quickcalcs/CImean1.cfm

Veámoslo con el ejemplo anterior de la concentración de nitratos:

Resultados de la concentración de nitratos en agua (µg L^{-1})								
0,51	0,50	0,50	0,50	0,50	0,49	0,52	0,50	0,47
0,51	0,52	0,53	0,48	0,49	0,50	0,52	0,49	0,50
0,49	0,48	0,46	0,49	0,49	0,48	0,49	0,51	0,47
0,50	0,51	0,51	0,48	0,48	0,47	0,50	0,49	0,48
0,50	0,50	0,50	0,53	0,53	0,52	0,50	0,51	0,51

Ejemplo sacado de Miller y Miller (2009).

Media: 0,50
Desviación estándar: 0,02
Varianza: 0,0003
Coeficiente de variación (%): 3,36%

3.2. El sentido de un valor analítico.

En un laboratorio de análisis, el analista puede enfrentarse a dos tipos de preguntas al trabajar con una muestra desconocida. Una es de carácter cualitativo (¿qué compuestos existen en esta muestra?) y otra cuantitativa (¿qué concentración tienen estos compuestos?). El segundo tipo de preguntas suele ser la más demandada e importante y lleva implícita un resultado numérico.

Este resultado (el resultado de la medida cuantitativa) suele llevar asociado un error. Como dicen Miller y Miller (2009), *"no existen resultados cuantitativos de interés si no van acompañados de alguna estimación de los errores inherentes a los mismos"*.

Por lo tanto, los errores analíticos se pueden clasificar en tres grandes grupos:

- ***Groseros o accidentales***
- ***Aleatorios***
- ***Sistemáticos***

El primero de ellos es fácilmente identificable (ejemplo: cuando se rompe algún instrumento, se pierde alguna disolución, etc.). Los otros dos, los aleatorios y sistemáticos, se diferencian en que los últimos muestran cierta tendencia o comportamiento repetido (ejemplos: una pipeta que siempre alícuota de menos, un analista que siempre pesa por exceso, etc.) y los aleatorios no.

Además de estos, existen otros conceptos importantes que un analista tiene que tener en cuenta:

- ***Precisión***: describe los errores aleatorios. Hace mención a la dispersión de los datos con respecto al valor real. En el caso de las distribuciones normalizadas de datos, haría referencia a la anchura de la campana gaussiana siendo más ancha cuanto más dispersos están los datos y viceversa.

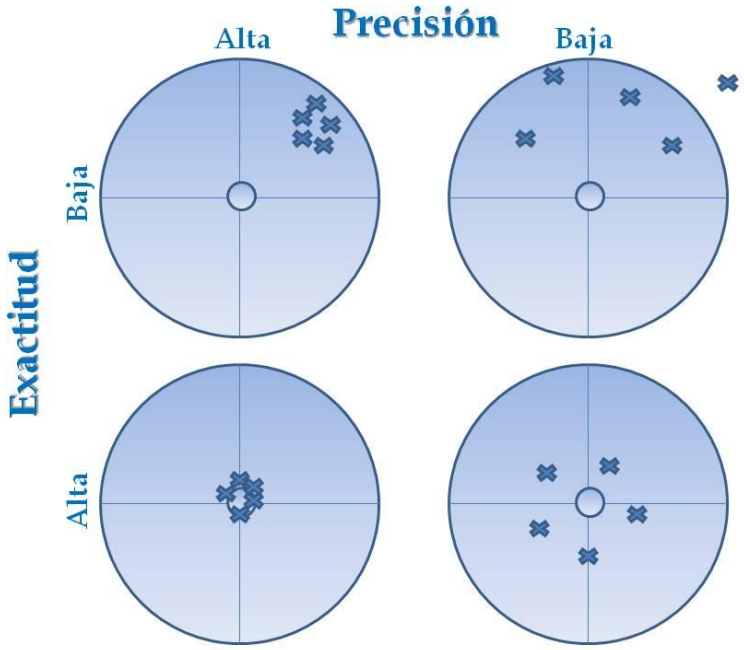
- ***Exactitud***: es la proximidad al verdadero valor de una medida individual o un valor promedio (ejemplo: una pipeta mide teóricamente 50 ml y experimentalmente como media mide 49,9 ml, por lo que sería bastante exacta). Está afectada por los errores aleatorios y sistemáticos.

- ***Reproducibilidad***: es la capacidad de obtener los mismos resultados en un análisis independientemente de las condiciones usadas (laboratorio, analista, fecha, material, etc.).

- ***Repetibilidad***: capacidad de obtener los mismos resultados en un análisis en las mismas condiciones usadas (laboratorio, analista, fecha, material, etc.).

- ***Incertidumbre***: intervalo dentro del cual es razonablemente verosímil que se encuentre el verdadero valor de la magnitud. Hace referencia a la expresión de un resultado analítico con su error inherente (aleatorio y sistemático).

A continuación se muestra una figura aclaratoria de la relación entre precisión y exactitud de un análisis cuantitativo:



Nota: el verdadero valor está situado en el centro de cada circunferencia y las cruces corresponden a cada repetición del análisis

3.3. **Incertidumbre de un valor analítico.**

En ausencia de error sistemático, la media aritmética debería coincidir con el verdadero valor que se espera encontrar (si tenemos en cuenta que los datos siguen una distribución normalizada, la media coincidirá con μ y la desviación estándar σ). Aún así, el error aleatorio inherente en la medida hará improbable que la media de la muestra sea exactamente igual al valor verdadero. Por lo tanto es recomendable hablar de un intervalo de valores que sea probable encontrar el valor verdadero que buscamos.

La amplitud de este intervalo depende de dos factores:

- La precisión de las medidas, que a su vez depende de la desviación estándar.
- El número de medidas que se realice.

Puede concluirse que cuantas más medidas se hagan, más fiable será la estimación de que la media aritmética ya que el intervalo de confianza será menor.

¿Cuál es el intervalo dentro del cual se puede suponer de forma razonable que se encuentra el valor verdadero de la medida? Para eso primero debemos calcular el error estándar de la media, que se define de la siguiente manera:

Error estándar de la media: Desviación estándar/√n,
donde n es el número de medidas.

A continuación, se define el ***Intervalo de Confianza de la Media***, que es el intervalo de valores dentro del cual podemos afirmar con cierta probabilidad que el valor verdadero se encuentra. Dependerá de la certeza que queramos: cuanto mayor sea, mayor la certeza de acertar.

Intervalo de confianza:

$$\overline{x} \pm zs/\sqrt{n}$$

donde s es la desviación estándar,
Z depende del grado de confianza requerido:
95%, z = 1,96
99%, z = 2,58
99,7%, z = 2,97,
y n es el número de muestras

Cuando el tamaño de muestra es más pequeño, se utiliza esta ecuación:

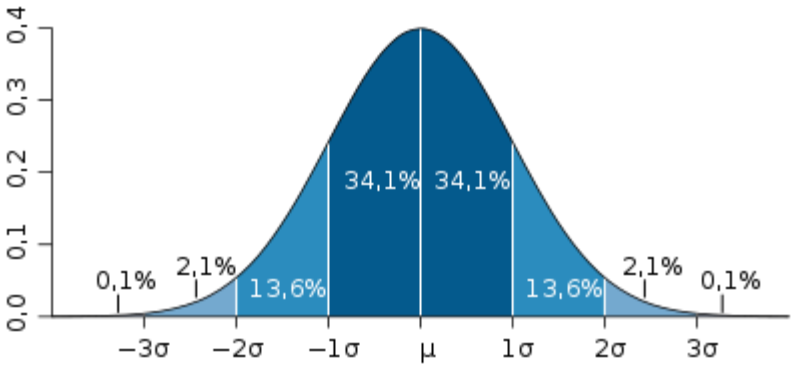
$$\overline{x} \pm t_{n-1} s/\sqrt{n}$$

donde t_{n-1} es un valor tabulado que dependerá del valor de confianza
son los grados de libertad,
donde s es la desviación estándar,
y n es el número de muestras

Grados de libertad	Valores de t para intervalos de confianza de 95 %	99%
2	4,30	9,92
5	2,57	4,03
10	2,23	3,17
20	2,09	2,85
50	2,01	2,68
100	1,98	2,63

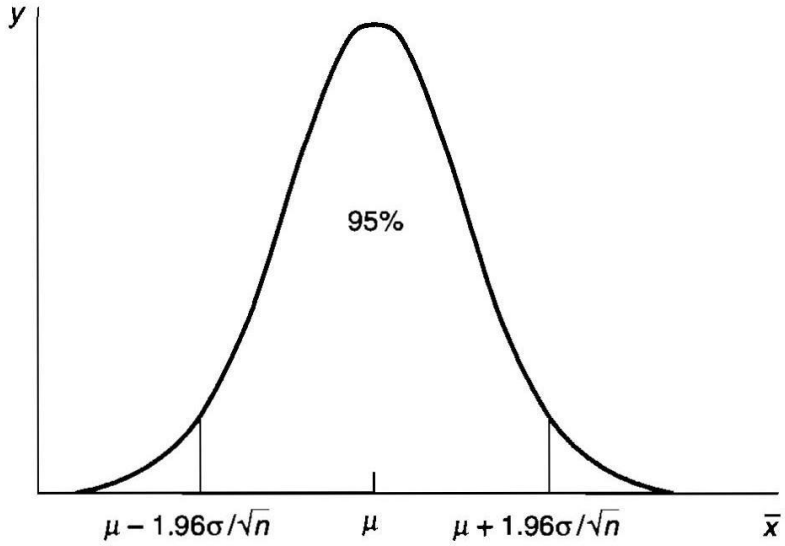
Tabla sacada de Miller y Miller (2009)

En una distribución normalizada, los datos se distribuyen de la siguiente manera alrededor de μ, el valor promedio de la población, y σ, (la desviación estándar).



Sacado de http://es.wikipedia.org/wiki/Distribucion_normal

Utilizando los valores comentados anteriormente, el 95% de los datos de este ejemplo estarían dentro del intervalo μ ± 1,96σ/√n.



Ejemplo sacado de Miller y Miller (2009)

3.4. **Presentación de resultados y propagación de errores**

Lo más común para presentar los resultados de un análisis es utilizar la media y la desviación estándar como estimación de la exactitud de la cantidad medida y precisión respectivamente. Menos frecuente es usar el error estándar de la media en vez de la desviación estándar o incluso, presentar el intervalo de confianza, ya que no existe unanimidad, tal y como observa en algunos trabajos científicos:

Table 1
Main characteristics of the "alperujo" samples (dry weight)

Parameters	Mean	Range	CV (%)
Moisture (% fresh weight)	64.0	55.6–74.5	7.6
pH[a]	5.32	4.86–6.45	6.6
EC[a] (dS m^{-1})	3.42	0.88–4.76	33.9
Ash (g kg^{-1})	67.4	24.0–151.1	42.5
TOC (g kg^{-1})	519.8	495.0–539.2	2.8
C/N ratio	47.8	28.2–72.9	22.1
TN (g kg^{-1})	11.4	7.0–18.4	24.5
P (g kg^{-1})	1.2	0.7–2.2	29.7
K (g kg^{-1})	19.8	7.7–29.7	34.2
Ca (g kg^{-1})	4.5	1.7–9.2	57.3
Mg (g kg^{-1})	1.7	0.7–3.8	58.7
Na (g kg^{-1})	0.8	0.5–1.6	36.6
Fe (mg kg^{-1})	614	78–1462	74.9
Cu (mg kg^{-1})	17	12–29	28.8
Mn (mg kg^{-1})	16	5–39	70.2
Zn (mg kg^{-1})	21	10–37	36.3

CV: coefficient of variation.
[a] water extract 1:10.

Table 3 – Contents, expressed in micrograms of biophenol per gram of alperujo, of the target biophenols in the polar fraction as obtained using the SAME and conventional extraction methods

Biophenol	Proposed method	Conventional method
Hydroxytyrosol	890.93 (1.75)	213.48 (1.04)
Verbascoside	103.32 (2.94)	67.55 (2.61)
Luteolin-7-glucoside	17.60 (3.27)	10.76 (2.15)
Apigenin-7-glucoside	7.82 (2.69)	4.46 (2.83)
Oleuropein	21.30 (2.95)	9.24 (3.68)
Luteolin	72.25 (3.06)	35.16 (2.10)
Apigenin	41.75 (1.48)	18.23 (2.49)

Errors, in brackets, are expressed as percent relative standard deviation (n = 3 replicates).

Table 1 Chemical, biochemical, microbiological and physical characteristics of the soil

pH (H_2O)	8.5±0.0[a]
EC (1:5, μs cm^{-1})	225±2
Texture	Loam
Total organic C (g kg^{-1})	10.3±0.3
Total carbohydrates (μg g^{-1})	552±20
Water-soluble C (μg g^{-1})	100±1
Water-soluble carbohydrates (μg g^{-1})	8±0
Total N (g kg^{-1})	0.95±0.02
Available P (μg g^{-1})	7±0
Extractable K (μg g^{-1})	222±4
Microbial biomass C (μg g^{-1})	396±11
Dehydrogenase (μg INTF g^{-1})	51±1
Urease (μmol NH_3 g^{-1} h^{-1})	0.31±0.03
Protease-BAA (μmol NH_3 g^{-1} h^{-1})	0.60±0.04
Phosphatase (μmol PNP g^{-1} h^{-1})	0.28±0.02
β-Glucosidase (μmol PNP g^{-1} h^{-1})	0.46±0.01
Aggregate stability (%)	11.5±0.4

[a]Mean±standard error (N=6)

Nota: el intervalo de confianza tiene sentido cuando tenemos muchas repeticiones. Cuando son pocas, (n ≤ 10) no tiene mucho sentido

A la hora de expresar los resultados de un análisis, las cifras significativas de un resultado (decimales que suelen llevar) no pueden exceder a la precisión usada, es decir, no tiene sentido dar un resultado de 0,0234234 g cuando la balanza que usamos tiene de precisión 0,001 g.

Además, para eso debemos tener en cuenta que los errores se propagan con las operaciones aritméticas que hagamos a los descriptivos (o combinación de cantidades observables). Para eso es necesario conocer la precisión de cada observación.

Veamos cómo podemos calcular la propagación de errores:
Nota: información y ejemplos asacados de http://www.uv.es/zuniga/tefg.htm

Propagación de errores en sumas y diferencias:

Si queremos saber el error de una variable (q) calculada de forma indirecta mediante la suma o resta de dos mediciones previas (x e y) y conocido su error:

$$q = x \pm y \quad \Rightarrow \quad \delta q \approx \delta x + \delta y$$

Veámoslo con un ejemplo: En un experimento se introducen dos líquidos en un matraz y se quiere hallar la masa total del líquido. Se conocen:

M1 = Masa del matraz 1 + contenido = 540 ± 10 g
m1 = Masa del matraz 1 = 72 ± 1 g
M2 = Masa del matraz 2 + contenido = 940 ± 20 g
m2 = Masa del matraz 2 = 97 ± 1 g

La masa de líquido será:
$M = M1 - m1 + M2 - m2 = 1311$ g

Su error:
$\delta M = \delta M1 + \delta m1 + \delta M2 + \delta m2 = 32$ g

El resultado se expresará:
$M = 1310 \pm 30$ g

Propagación de errores en productos y cocientes:

$$q = xy \quad \Rightarrow \quad \frac{\delta q}{|q|} \approx \frac{\delta x}{|x|} + \frac{\delta y}{|y|}$$

$$q = \frac{x}{y} \quad \Rightarrow \quad \frac{\delta q}{|q|} \approx \frac{\delta x}{|x|} + \frac{\delta y}{|y|}$$

Para medir la altura de un árbol, L, se mide la longitud de su sombra, L_1, la altura de un objeto de referencia, L_2, y la longitud de su sombra, L_3. Por semejanza:

$$L = L_1 (L_2 / L_3)$$

Realizadas las medidas resultan:
$L_1 = 200 \pm 2$ cm, $L_2 = 100{,}0 \pm 0{,}4$ cm, $L_3 = 10{,}3 \pm 0{,}2$ cm

Por tanto
$L = 200 \times (100/10) = 2000$ cm

Su error será

$$\frac{\delta L}{|L|} \approx \frac{\delta L_1}{|L_1|} + \frac{\delta L_2}{|L_2|} + \frac{\delta L_3}{|L_3|} = \frac{2}{200} + \frac{0.4}{100} + \frac{0.2}{10.3} =$$

$$= (1 + 0{,}4 + 2)\% = 3{,}4\%$$

$$\delta L = (3{,}4/100) \times 2000 = 68$$

$$L = 2000 \pm 70 \text{ cm}$$

Propagación de errores en producto por una constante y una potencia:

$$\delta q = |A| \delta x \qquad \frac{\delta q}{|q|} = |n| \frac{\delta x}{|x|}$$

Para calcular la propagación de errores, consultar la siguiente web:
http://graphpad.com/quickcalcs/ErrorProp1.cfm?Format=SD

Capítulo 4. Estadística inferencial.

Como ya hemos visto, la estadística la podemos utilizar de forma **descriptiva** (nos dará toda la información posible sobre un grupo de datos) o de forma **inferencial** (estudiando el comportamiento de los datos de una muestra, podremos obtener conclusiones y predicciones sobre la población).

Cuando estudiamos un parámetro o variable de la muestra, existen tres formas de estimación:

Puntual
Por intervalos
Por contraste de hipótesis

La primera se correspondería con el valor obtenido de diversas mediciones expresado como la media aritmética y su error asociado.

La segunda se correspondería con el Intervalo de confianza, un rango de valores entre los que se encuentra el valor verdadero del parámetro estudiado afirmándolo con una determinada probabilidad.

4.1. Contraste de hipótesis.

Los dos ejemplos anteriores ya los hemos visto en el capítulo anterior. La tercera opción corresponde a una herramienta fundamental en la inferencia estadística, el ***Contraste de Hipótesis***. Se basa en estudiar la probabilidad de formular una afirmación (o hipótesis) sobre un caso concreto y que estemos en lo cierto.

Nota: Este tipo de estimaciones se usan en los casos de distribución normalizada.

Veámoslo con el siguiente ejemplo: queremos estudiar si hay un problema de contaminación y nos preguntamos ¿es mayor de 50 mg L^{-1} la concentración de nitratos en un rio que pasa cerca de una granja agrícola? Si es mayor, habrá contaminación y si es menor, no.

Para estudiar este caso, hemos ido a río y hemos recogido muestras de agua para analizarlas en el laboratorio. Tenemos dos posibles hipótesis antagónicas a contrastar:

- **La hipótesis nula (H_0),** correspondería a la afirmación que queremos contrastar (ejemplo: la concentración de nitratos es menor o igual de 50 mg L^{-1}).
- **La hipótesis alternativa (H_1)** o la contraria a H_0 (ejemplo: la concentración es mayor de 50 mg L^{-1}).

A obtener los resultados de los análisis, vemos las siguientes posibilidades:

Nuestra elección	Realidad	
	H_0 es cierta	**H_1 es cierta**
Escogemos H_0 como cierta	Decisión correcta de Tipo A ($p = 1-\alpha$)	**Error de tipo II ($p= \beta$)**
Escogemos H_1 como cierta	**Error de tipo I ($p = \alpha$)**	Decisión correcta de Tipo B ($p= 1-\beta$)

Como podemos observar, tenemos cuatro posibilidades. Dos de ellas en las que no cometemos error en la formulación de nuestra hipótesis.

- Decisión correcta de Tipo A: Suponemos que los nitratos son menores de 50 mg L^{-1} y lo confirmamos con los análisis. Podemos afirmar que NO HAY contaminación.
- Decisión correcta de Tipo B: Suponemos que los nitratos son mayores de 50 mg L^{-1} y lo confirmamos con los análisis. Podemos afirmar que HAY contaminación.

Las otras dos nos informan del error que hemos cometido en nuestra hipótesis:

- Error de Tipo I (o de tipo α): Suponemos que la concentración de nitratos era mayor de 50 mg L^{-1} y no lo era.
- Error de Tipo II (de tipo β): Suponemos que la concentración de nitratos era menor de 50 mg L^{-1} y no lo era.

Nota: Para hacer un contraste de hipótesis, debemos definir muy bien cual es la hipótesis nula (H_0) y alternativa (H_1), ya que todas las conclusiones se harán en base a esto.

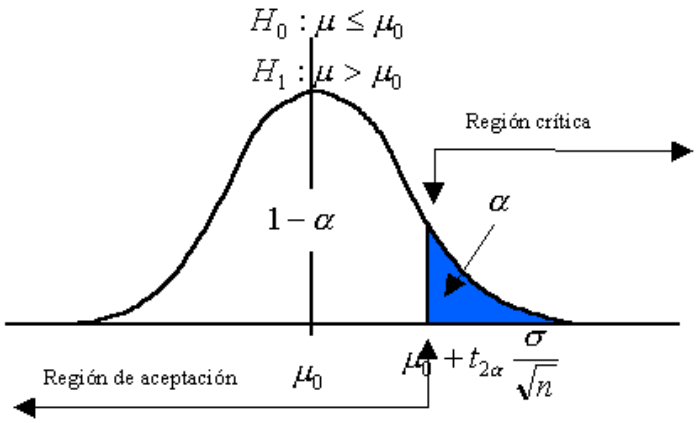
A la hora de descartar una u otra hipótesis, no podremos afirmar exactamente cual es la decisión correcta debido a que solo tenemos acceso a una parte de la población (una muestra de ella). Por lo tanto, es lógico que hablemos en términos probabilísticos, ya que así podremos evaluar el error que cometemos al equivocarnos.

Así, en términos de probabilidad tendríamos lo siguiente:

Probabilidad de escoger H_0 siendo H_0 cierta = $1-\alpha$
Probabilidad de escoger H_0 siendo H_1 cierta = β
Probabilidad de escoger H_1 siendo H_0 cierta = α
Probabilidad de escoger H_1 siendo H_1 cierta = $1-\beta$

Nota: Es usual utilizar el Error de tipo I, es decir, el error que comentemos al afirmar que la hipótesis nula no es la correcta ya que es más fácil de controlar.

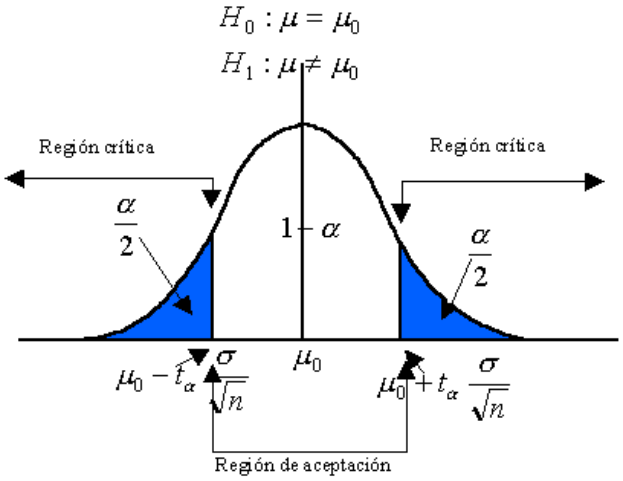
Veámoslo gráficamente:



Ejemplo de gráfica de una cola sacado de
http://www.terra.es/personal2/jpb00000/ttesthipotesis.htm

μ se corresponde con el valor de la concentración de nitratos de las muestras de agua y μ_0 con 50 mg L^{-1}. Aquí vemos que el nivel de significancia o probabilidad de cometer el error de tipo I es α.

Nota: si hubiésemos definido las hipótesis como H_0 fuese la concentración de nitratos en agua 50 mg L^{-1} y H_1, una concentración distinta a esta (independientemente que fuese mayor o menor), tendríamos una gráfica de "dos colas".

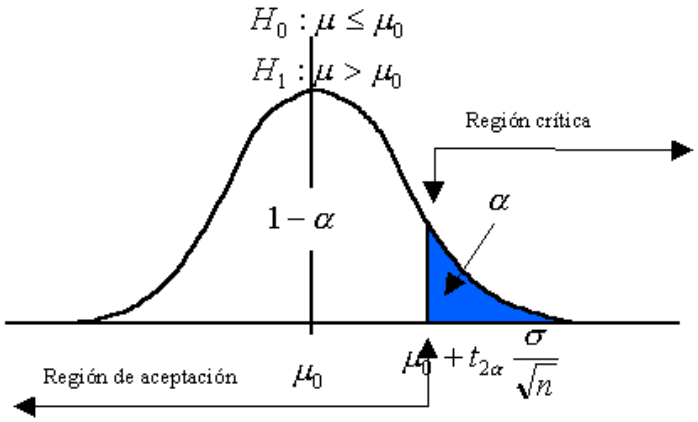


Ejemplo de gráfica de dos colas sacado de
http://www.terra.es/personal2/jpb00000/ttesthipotesis.htm

4.2. Estadístico de contraste y concepto de *p*-valor.

Una vez vistas todas las posibilidades, ¿cuál es el criterio que debemos seguir para poder afirmar u descartar una hipótesis? Para eso debemos calcular el ***estadístico de contraste***, que a su vez nos dará el ***p-valor***.

Este es un parámetro que se calcula teniendo en cuenta que los datos cumplen la distribución normalizada y se relaciona con el área de la curva la cual se correspondería con la región de rechazo, es decir, que si nuestra muestra está en esa zona podríamos descartar la hipótesis nula (H_0).



La región de rechazo sería la sombreada en azul en un ejemplo de gráfica de una cola sacado de http://www.terra.es/personal2/jpb00000/ttesthipotesis.htm

Si ese valor fuese menor que el nivel de significación (α) que hemos prefijado, podríamos rechazar la hipótesis nula.

Nota: los valores más usados para a son 0,05, 0,01 y 0,001, es decir, que la probabilidad de acertar en nuestra afirmación sería del 95%, 99% y 99,9% respectivamente.

Para el caso de que conozcamos la desviación estándar (σ), el estadístico de contraste tendría esta expresión:

$$Z^* = \frac{\overline{X} - \mu_{H_0}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_{H_0}}{\sigma / \sqrt{n}}$$

Si no lo conociéramos, sería esta:

$$t^* = \frac{\overline{x} - \mu}{s / \sqrt{n}} \approx t - Student(n-1)$$

En este último caso, para obtener el *p*-valor, deberíamos consultar las tablas para la distribución estandarizada de la t-Student, en la cual deberemos saber los grados de libertad (n-1), siendo n el número de muestras.

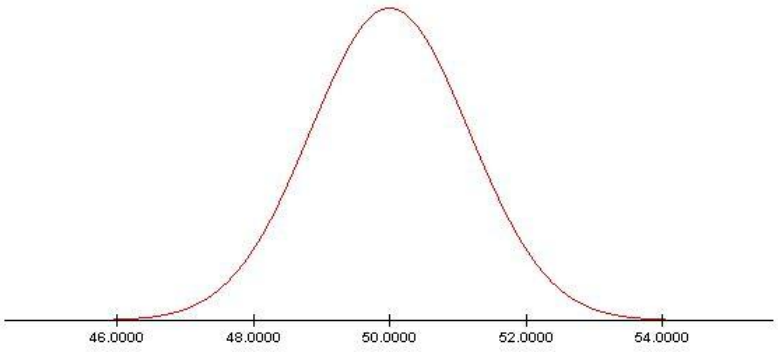
Nota: En cualquier manual de estadística podemos encontrar estas tablas y cualquier software de estadística las lleva incorporadas en sus análisis dándote directamente el p-valor. Si no, se pueden consultar en internet aquí:
http://es.wikibooks.org/wiki/Tablas_estadisticas/Distribucion_t_de_Student

En resumen, descartaremos H_0 si *p*-valor ≤ α, siendo α = 0,05, (también puede ser 0,01 y 0,001).

Veámoslo con nuestro ejemplo de la concentración de nitratos anteriormente descrito en el apartado de contraste de hipótesis.

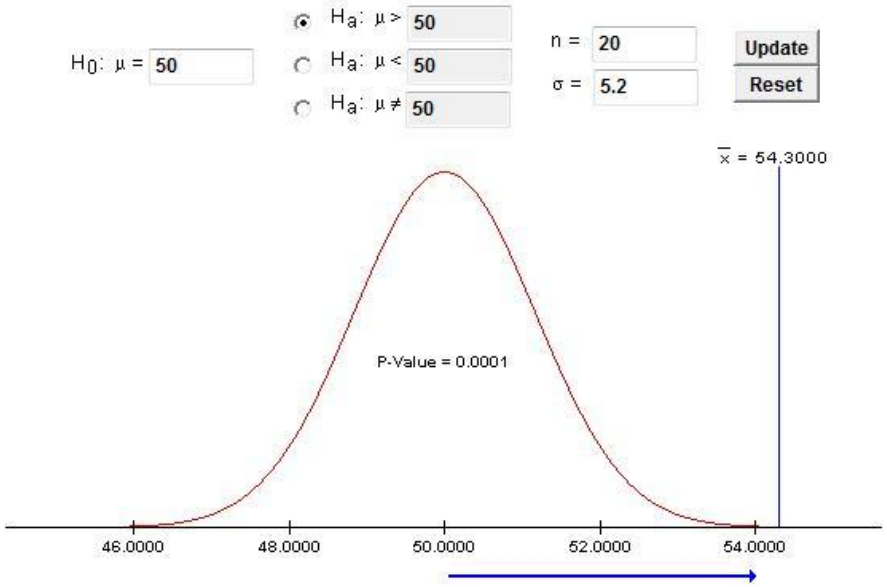
Nota: usaremos la aplicación gratuita disponible en
http://bcs.whfreeman.com/ips4e/cat_010/applets/pvalue_ips.html.

Imaginemos que hemos analizado las muestras de agua del río que pasa por al lado de una granja. Hemos obtenido un valor de 54,3 ± 5,2 mg L^{-1} al medir 20 muestras. Vemos que los resultados siguen una distribución normalizada, tal y como vemos en el siguiente gráfico:



Queremos saber si hay contaminación en estas aguas. ¿El valor obtenido de nuestras muestras es distinto significativamente de 50 mg $L^{-1,}$ valor que según el cual la legislación te dice que tienes contaminación por nitratos? Al hacer el contraste de hipótesis, definimos H_0 y H_1:

H_0: El valor de nuestros análisis es igual a 50 mg L^{-1}
H_1: El valor de nuestros análisis es mayor a 50 mg L^{-1}



Como podemos observar, nos ha salido un p-valor = 0.0001 (p-valor ≤ 0,05). Por lo tanto, podemos rechazar la hipótesis nula y confirmar que la concentración de nitratos obtenida en las muestras analizadas es mayor que 50 significativamente (con un 95% de probabilidad).

4.3. Tipos de variables y clasificación de los tests estadísticos.

Como ya se ha comentado, para el estudio estadístico de las poblaciones utilizamos las variables, las cuales definen alguna propiedad de las muestras elegidas que podemos medir experimentalmente. Es importante conocer las distintas formas que pueden tener las variables, ya que su naturaleza condicionará el tipo de test estadístico que aplicaremos.

Existen varias clasificaciones para las variables siendo la más importante la siguiente:

- ***Cualitativas***. Son aquellas que expresan una propiedad de las muestras que no se puede expresar numéricamente (ejemplo: el color de los ojos). A su vez, podemos clasificarlas en dos grupos:
    - Ordinal o cuasicuantitativa, en la cual puede adoptar varios regidos por un cierto orden (ejemplo: leve, moderado o grave).
    - Nominal, en la cual no existe ningún tipo de orden (ejemplo: tipo de medio de locomoción, coche, motocicleta, etc.).
- ***Cuantitativas o numéricas,*** las cuales se pueden expresar mediante un valor numérico (ejemplo: la concentración de nitratos de 50 mg L^{-1}). A su vez, se pueden clasificar en dos grupos:

- Discretas, en las cuales solo pueden adoptar valores enteros (ejemplo: número de hijos, días, etc.).

- Continuas, en las cuales pueden adoptar todos los valores posibles reales (ejemplo: la altura de un jugador de baloncesto es de 2,15 metros, hace -2ºC en la calle, etc.).

También es importante la siguiente clasificación:

- ***Variable independiente.*** Es aquella que no depende de ningún factor concreto y que a su vez, puede provocar una modificación en otras variables.
- ***Variable dependiente.*** Es aquella que puede ser modificada por otra variable (independiente). Un ejemplo es aquella propiedad que se mide en una muestra a lo largo del tiempo.

Para aclarar estos conceptos, vemos el siguiente ejemplo. Imaginemos que ponemos un experimento en el cual vamos a medir el contenido en nitrógeno de una determinada planta al añadirle distintas concentraciones de nitrato. Para eso, montamos varios grupos de macetas (tendremos 10 repeticiones en cada caso) a las que regaremos con 0, 10, 20 y 30 mg L^{-1} de nitrato. Observamos que al añadir más concentración de nitrato, la planta crece más y tiene un mayor contenido en nitrógeno. Este último será la variable dependiente (la que vamos a medir en nuestro experimento) y la concentración de nitratos será la independiente (modifica la anterior y es la que nosotros manipulamos).

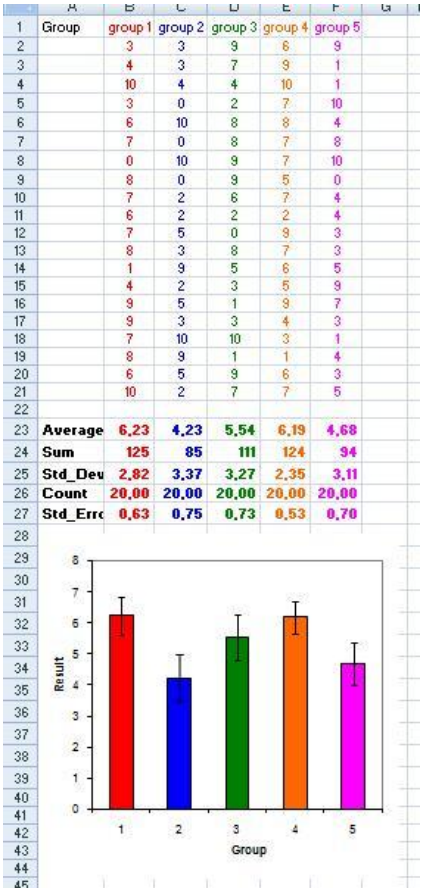
En este manual nos centraremos en las técnicas estadísticas de las variables cuantitativas continuas, que son las más comunes en un laboratorio de análisis. Para este tipo de variables, existen dos grandes grupos de técnicas estadísticas que dependerán fundamentalmente de la distribución de sus probabilidades (como vimos en el Capítulo 1). Así, encontramos estos dos grandes grupos:

- ***Técnicas paramétricas.*** Se utilizan cuando las variables siguen una distribución normalizada y se basan en la media y la varianza.
- ***Técnicas no paramétricas o "robustas".*** Se utilizan cuando no siguen un tipo de distribución conocida. Están basadas en el empleo de la mediana.

Nota: en general, estos tests presentan la misma filosofía. Calculan estadísticos de contraste que nos permitirán obtener un p-valor, el cual utilizaremos para afirmar la validez de nuestra hipótesis nula H_0 planteadas al compararlo con el valor de significancia a (0,05, 0,01 y 0,001 según convengamos).

A continuación, en el siguiente apartado veremos algunos tests estadísticos básicos que nos podremos encontrar en cualquier análisis de muestras de interés biológico. Estas se centrarán en tres grandes grupos: análisis para una muestra poblacional, análisis para varias muestras poblacionales y análisis de regresión y correlación.

Capítulo 5. Tests estadísticos básicos de interés en análisis químico.



	Group	group 1	group 2	group 3	group 4	group 5
2		3	3	9	6	9
3		4	3	7	9	1
4		10	4	4	10	1
5		3	0	2	7	10
6		6	10	8	8	4
7		7	0	8	7	8
8		0	10	9	7	10
9		8	0	9	5	0
10		7	2	6	7	4
11		6	2	2	2	4
12		7	5	0	9	3
13		8	3	8	7	3
14		1	9	5	6	5
15		4	2	3	5	9
16		9	5	1	9	7
17		9	3	3	4	3
18		7	10	10	3	1
19		8	9	1	1	4
20		6	5	9	6	3
21		10	2	7	7	5
22						
23	Average	6,23	4,23	5,54	6,19	4,68
24	Sum	125	85	111	124	94
25	Std_Dev	2,82	3,37	3,27	2,35	3,11
26	Count	20,00	20,00	20,00	20,00	20,00
27	Std_Error	0,63	0,75	0,73	0,53	0,70

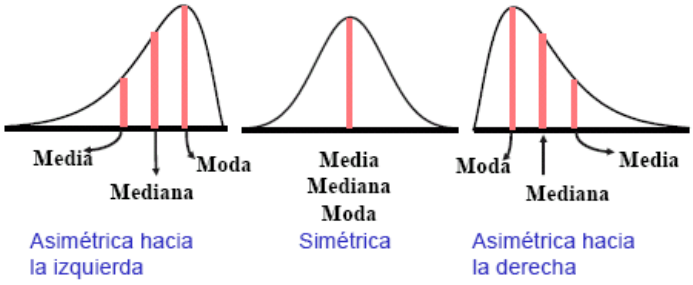
En este capítulo veremos una descripción breve de algunos de los ejemplos más comunes de tests que un analista necesitar realizar de forma rutinaria.

Nota: Para profundizar en los fundamentos teóricos de dichas pruebas, se recomienda la visita a la excelente web de estadística http://statpages.org. En ella, se encuentran una amplia selección de tests estadísticos que se pueden realizar online.

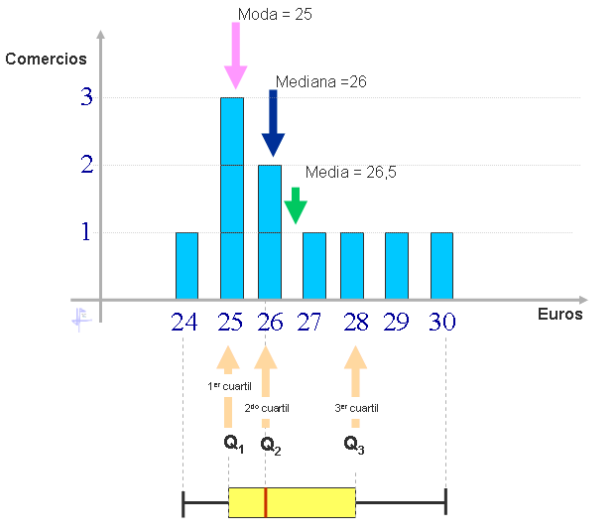
5.1. Análisis para una muestra poblacional.

5.1.1. Análisis descriptivos

Son fundamentales en cualquier medida analítica, dándote bastante información sobre la naturaleza de las muestras y suelen incluir los estadísticos básicos que hemos comentado en el Capítulo 2 (*media, desviación estándar, error relativo u coeficiente de variación y la varianza*), aunque hay bastantes más. Algunos interesantes son los siguientes:



Sacado de
http://www.tuveras.com/estadistica/estadistica02.htm



Moda: Es el valor más frecuente.
Mediana: Al ordenar los datos, es el valor que ocupa la posición central.
Cuartiles: Son los tres valores que dividen el conjunto de datos en cuatro partes iguales.

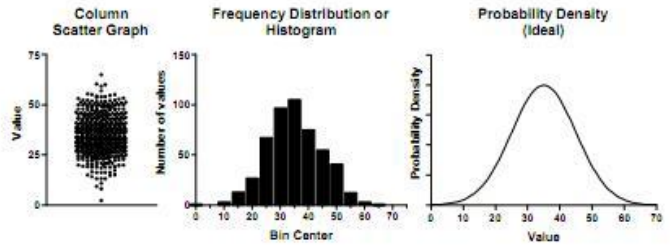
Son muy útiles para representar gráficamente y cualquier hoja de cálculo lleva incorporada la posibilidad de realizar este tipo de cálculos.

En las siguientes páginas se pueden calcular:

- http://graphpad.com/quickcalcs/CImean1.cfm
- http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Descriptive.htm
- http://www.openepi.com/OE2.3/Menu/OpenEpiMenu.htm

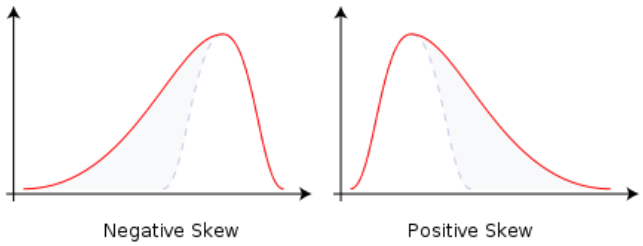
5.1.2. Test de Normalidad

Evaluar la distribución de los datos aleatorios en una muestra poblacional es el primer paso que debemos seguir antes de elegir el test estadístico a aplicar.

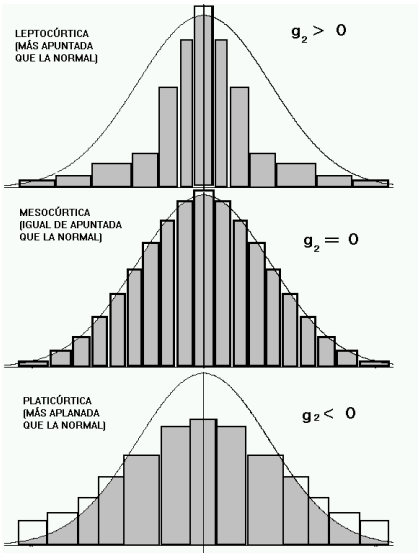


Ejemplo sacado de http://www.graphpad.com/articles/AnalyzingData.pdf

Existen varios tests para evaluar la distribución como la prueba de Kolmogórov-Smirnov, el test de Shapiro–Wilk o la prueba de Anderson-Darling. Estas pruebas evalúan si los datos están normalmente distribuidos y se basan en cálculos de parámetros tales como el Skewness y el Curtosis, ambos relacionados con la forma de la campana de Gauss de la distribución normalizada.



Sacado de http://en.wikipedia.org/wiki/Skewness



Sacado de http://www.uv.es/ceaces/base/descriptiva/curtosis.htm

Existen numerosas webs donde se pueden realizar este tipo de tests, cuyos resultados suelen ser de este tipo:

Evidencia fuerte en contra de la normalidad
Evidencia suficiente en contra de la normalidad
Evidencia subjetiva en contra de la normalidad
Poca evidencia en contra de la normalidad
Ninguna evidencia en contra de la normalidad
Evidencia fuerte en contra de la normalidad

En este enlace se puede realizar un test de normalidad de datos aleatorios:
http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Normality.htm

Nota: cuando tengamos pocas repeticiones, este tipo de tests saldrán siempre normalizados. Si los datos están normalizados, deberemos usar los tests paramétricos y si no, los no paramétricos.

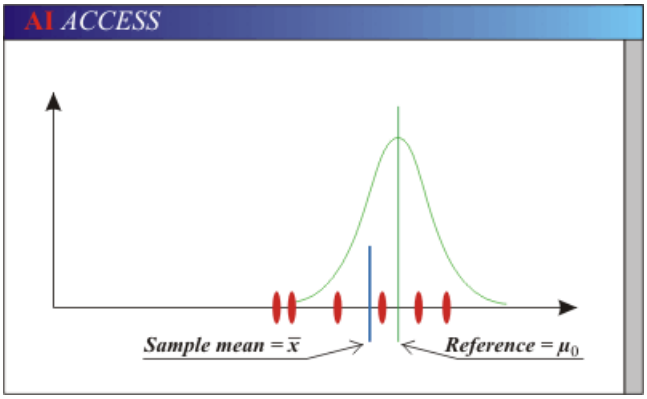
Contrastes de la media de una población con un valor de referencia.

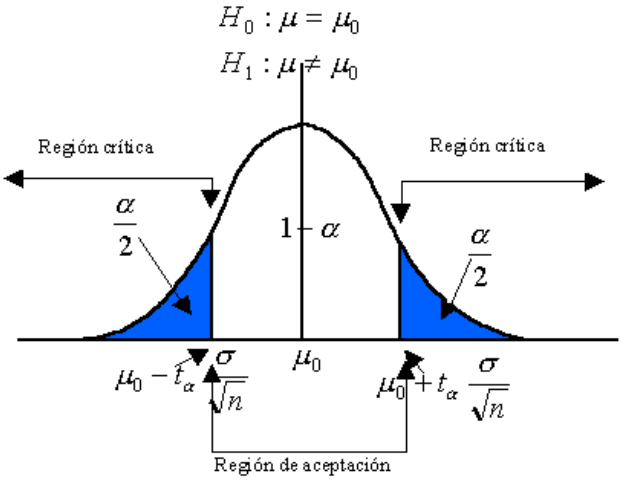
Prueba de la t de Student (paramétrico)

Este test se basa en el contraste de hipótesis y cálculo del estadístico t, tal y como comentamos en el Capítulo 3. La condición fundamental es que los datos sigan una distribución normalizada. Al final obtendremos un valor de *p* con el que podremos

saber si la media de nuestros datos es diferente significativamente o no a un valor de referencia que nosotros queremos comparar (hipótesis nula).

p >0,05, no significativo, (NS)
p entre 0,01 y 0,05, significativo, (*)
p entre 0,001 y 0,01, muy significativo, (**)
p < 0,001, extremadamente significativo, (***)





Sacado de http://www.aiaccess.net/English/Glossaries/GlosMod/e_gm_t_test.htm

En estas páginas se pueden realizar este tets:

http://www.graphpad.com/quickcalcs/OneSampleT1.cfm?Format=SD
http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/MeanTest.htm
http://www1.assumption.edu/users/avadum/applets/applets.html

5.1.3. Prueba de los signos (no paramétrico)

Este test es la versión no paramétrica de la t de Student. Se basa obtener la diferencia de cada uno de los valores experimentales con respecto al valor referencia teniendo solo en cuenta el signo (negativo si es menor y positivo si es mayor). Estos se distribuirán siguiendo la ley binomial, con la cual calcularemos un p-valor que contrastaremos con nuestro valor de significancia

http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html

5.2. Análisis para dos muestras poblacionales.

Cuando empezamos a analizar y a comparara dos muestras poblacionales, toma mucha importancia la naturaleza de las mismas, es decir, si son independientes (no pareadas) o dependientes (pareadas) unas de otras.

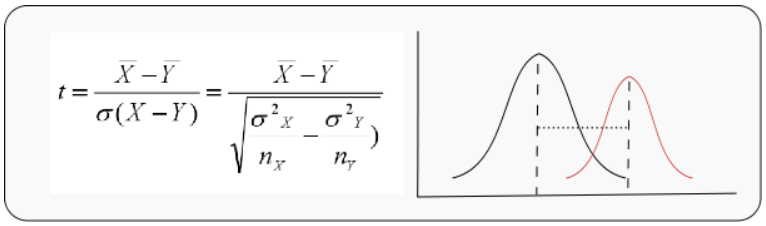
5.2.1 Prueba de la t de Student (paramétrico)

Al igual que hemos visto anteriormente, se hará un contraste de hipótesis y se calculará el valor del estadístico t, para obtener finalmente, un valor de *p* con el que decidiremos si rechazamos o no la hipótesis nula. Para utilizar esta prueba tenemos que tener en cuenta que las dos muestras deben seguir una distribución normalizada

En las siguientes páginas se pueden realizar este test con la posibilidad de elegir si las muestras son independientes o dependientes:

http://graphpad.com/quickcalcs/ttest1.cfm
http://faculty.vassar.edu/lowry/tu_esp.html

Nota: Para profundizar sobre esta prueba, consultar esta web:
http://www.fisterra.com/mbe/investiga/t_student/t_student.asp#dependientes



Sacado de http://personales.upv.es/jcanizar/modulo_3/diferenciales_4.html

5.2.2. Prueba de U de Mann-Whitney para muestras independientes (no paramétrico)

Este test es la versión no paramétrica de la t de Student para muestras independientes. Las alternativas paramétricas son menos robustas que las paramétricas ya que se basan en la mediana (valor que está situado en el centro al ordenar los datos). Análogo a la t de Student, este test se basa en el cálculo del estadístico U:

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - \Sigma R_1$$

$$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - \Sigma R_2$$

El cual contrastaremos con los valores de significancia (0,05, 0,01 y 0,001) para así saber si existe diferencia estadísticamente significativa entre ambas poblaciones de muestras.

Para realizar esta prueba, se pueden consultar las siguientes webs:

http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Ustat.htm
http://faculty.vassar.edu/lowry/utest.html

Nota: Pata más información, consultar esta web:
http://members.fortunecity.com/bucker4/estadistica/pruebaumw2mi.htm

5.2.3. Prueba de Wilcoxon para muestras dependientes (no paramétrico)

Se basa en calcular el estadístico W que contrastaremos con el *p*-valor.

En las siguientes webs se puede realizar esta prueba:

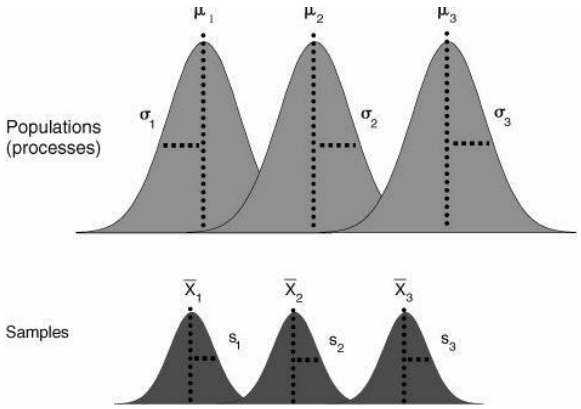
http://faculty.vassar.edu/lowry/wilcoxon.html
http://www.fon.hum.uva.nl/Service/Statistics/Signed_Rank_Test.html

5.3. Análisis para más de dos muestras poblacionales.

5.3.1. Análisis de la varianza (ANOVA, paramétrico)

Esta prueba estadística es de las más utilizadas para poder comparar más de dos muestras poblacionales, las cuales deben cumplir los siguientes requisitos:

- Que las variables sean independientes
- Que tengan una distribución normalizada
- Que sus varianzas no difieran significativamente.



Sacado de http://www.bexcellence.org/Anova.html

Se basa en el contraste de las medias de las muestras y su varianza. Para saber más sobre los cálculos aritméticos que incluyen, consultar las siguientes páginas:

http://www.seh-lelha.org/anova.htm

http://e-stadistica.bio.ucm.es/cont_mod_1.html#Anova

Finalmente, se calculará un estadístico de contraste F y que dará un valor de p que compararemos según nuestro nivel de significación (0,05, 0,01 y 0,001). Este test nos dirá si las muestras poblacionales son distintas significativamente pero no entre si, es decir, por parejas de muestras poblacionales. Para eso se realizan los tests *"post-hoc"*, como los tests de Duncan, Tukey o Fisher (mínima diferencia significativa o LSD), todos ellos basados en la t de Student.

Para calcular este test, se puede consultar las siguientes webs:

http://www.amstat.org/publications/jse/v18n2/ANOVAExercise.xls
http://www.physics.csbsju.edu/stats/anova.html
http://e-stadistica.bio.ucm.es/mod_anova/anova_applet.html
http://faculty.vassar.edu/lowry/ank3.html

Para las pruebas *"post-hoc"*:
http://graphpad.com/quickcalcs/posttest1.cfm

5.3.2. Análisis de Kruskal-Wallis (no paramétrico)

Es similar al ANOVA pero cuando se cuentan con datos que nos siguen una distribución normalizada.

En esta página se puede realizar dicha prueba:

http://department.obg.cuhk.edu.hk/researchsupport/KruskallWallis.asp

5.4. Análisis de la correlación.

Hasta ahora, hemos comparado grupos de muestras poblacionales entre si con la intención de si cumplían o no determinadas características. Los ejemplos vistos se centraban en discernir si esas muestras diferían significativamente de un valor prefijado o incluso, si diferían entre dos o más grupos de muestras poblacionales.

Una vez descritos los procedimientos estadísticos más básicos para poder evaluar estas cuestiones, el siguiente paso en el análisis inferencial nos lleva a estudiar la relación entre dichas muestras poblacionales, para observar y describirlas matemáticamente con el fin de poder hacer predicciones. El estudio de estas relaciones nos lo da correlación entre variables.

En este manual solo veremos un tipo de correlación ya que es la más importante para un analista, la regresión lineal

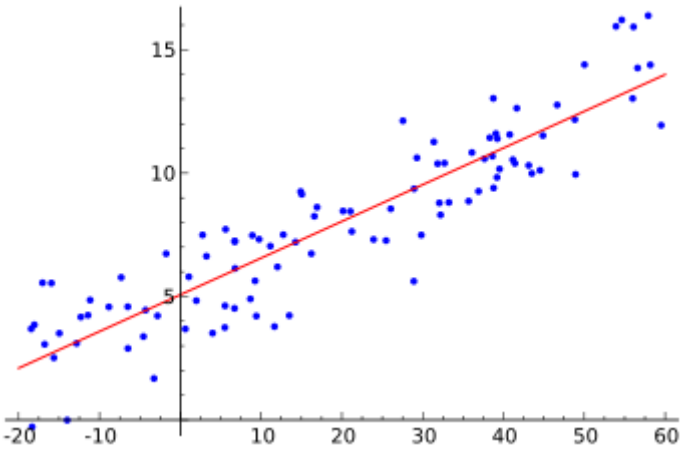
5.4.1 Regresión lineal

Imaginemos que tenemos datos de dos variables, una dependiente y otra independiente, y al contrastar dichas variables, la relación matemática que las relaciona tiene forma de una recta de este tipo:

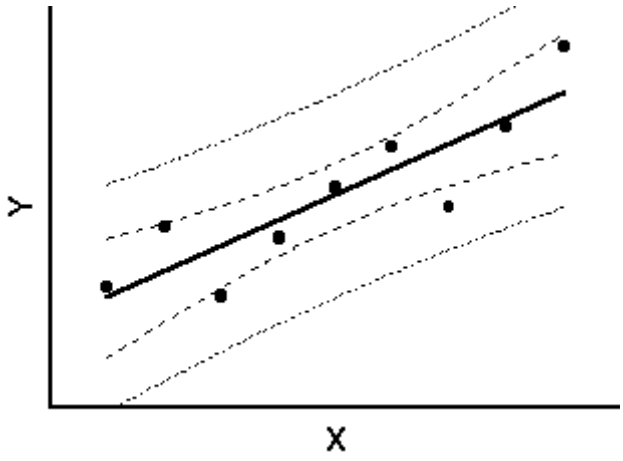
$$Y = a + bX,$$

donde a sería la ordenada en el origen y b la pendiente.

Nota: La recta suele ser de primer orden aunque puede ser mayor



Ejemplo sacado de http://es.wikipedia.org/wiki/Regresion_lineal



Ejemplo sacado de http://graphpad.com/curvefit/linear_regression.htm

El método para calcular experimentalmente la ecuación de la recta se realiza mediante el ***Método de Mínimos Cuadrados***. Para saber más sobre la aritmética de este método, consular la siguiente página web:

http://www.uv.es/jbosch/PDF/RectaMinimosCuadrados.pdf

Aparte de calcular los coeficientes de la recta (a, ordenada en el origen o el valor donde corta con el eje de las X, y b, que nos da información sobre la pendiente de la recta), también calcula un coeficiente que nos informa de la bondad de la regresión. Este se denomina R ó su cuadrado R^2, siendo mejor el ajuste cuanto más cerca sea de la unidad (lo ideal es 0,999 o mejor).

Para hacer la regresión lineal, consultar estos enlaces:
http://department.obg.cuhk.edu.hk/researchsupport/regressionFit.asp
http://www.sc.ehu.es/sbweb/fisica/cursoJava/numerico/regresion1/regresion1.htm
http://faculty.vassar.edu/lowry/corr_stats.html
http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Regression.htm

5.5. Cuadro resumen.

A continuación, se muestra un esquema de actuación de lo comentado anteriormente:

Capítulo 6. Bibliografía recomendada y recursos disponibles en internet.

A continuación recopilamos algunos libros importantes para profundizar en los fundamentos y aplicaciones de la quimiometría, así como recursos disponibles en internet interesantes y que han ayudado a la elaboración de este manual. También se recomiendan algunas herramientas de software libre interesantes para el estudio estadístico.

Libros de texto fundamentales

- ***Estadística y Quimiometría para Química Analítica***. Miller y Miller. ISBN: 84-205-3514-1
- ***Quimiometría***. Carlos Mongay Fernández. ISBN: 9788437059235
- ***Quimiometría***. Guillermo Ramis Ramos y Mª Cecilia García Álvarez-Coque. ISBN: 8477389047

- Edición digital:

  - ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*** Second Edition. February 2009. Trevor Hastie, Robert Tibshirani and Jerome Friedman. Disponible en: (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)
  - ***Bioestadística: métodos y aplicaciones.*** Autores: Francisca Ríus Díaz, Francisco Javier Barón Lopez, Elisa Sánchez Font y Luis Parras Guijosa. Disponible en: http://www.bioestadistica.uma.es/libro/

Webs:

- **Web Pages that Perform Statistical Calculations! (http://statpages.org)**

- STAT-ATTIC STATistics Applets for Teaching Topics in Introductory Courses. (http://sapphire.indstate.edu/~stat-attic/index.php)
- Apuntes sobre Bioestadística, Universidad de Málaga (http://www.bioestadistica.uma.es/baron/apuntes/).
- Aula Virtual de Bioestadística. http://e-stadistica.bio.ucm.es/index.html
- Dr Arsham´s Statistic site (http://home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm)
- Material de clase de la asignatura Estadística 2 (http://augusta.uao.edu.co/moodle/course/view.php?id=284)
- Estadística aplicada a la Bioinformática (http://sebbm.bq.ub.es/BioROM/contenido/UIB/bioinfo/index.htm)
- Bioestadística, Universidad de Granada (http://www.ugr.es/~bioestad/)
- Department of Obstetrics and Gynaecology. The Chinese University of Hong Kong. (http://department.obg.cuhk.edu.hk/researchsupport/statstesthome.asp)

Applets y software gratuito:

- Recopilación muy completa sobre software estadístico libre y de pago (http://statpages.org/javasta2.html)
- Recopilación muy completa sobre Applets (aplicaciones en Java) para hacer cálculos estadísticos online (http://statpages.org/index.html)