ERRORES EN QUÍMICA ANALÍTICA

INTRODUCCIÓN

Poblaciones y muestras

Dentro del contexto de un laboratorio, la **población** consiste en todas las posibles determinaciones que puedan llevarse a cabo, mientras que la **muestra** es solo una pequeña parte, es decir las determinaciones que realmente se llevan a cabo.

Las poblaciones pueden ser muy grandes e incluso infinitas. Aunque algunas pruebas estadísticas hacen distinciones entre poblaciones finitas o infinitas, casi siempre se puede considerar una población como consistente en un número infinito de individuos, objetos o determinaciones, de la cual se toma una muestra finita (y de tamaño mucho más reducido). A partir del estudio de la muestra, se extraen conclusiones acerca de toda la población.

Hay un problema de **terminología** en lo que respecta al término "muestra". La IUPAC propone que en química se emplee la palabra *muestra* solo cuando ésta última sea una porción de un material seleccionado a partir de una cantidad más grande de material, lo cual es consistente con la terminología estadística. Esto uso implica también la existencia de un error de muestreo, cuando la muestra no refleja exactamente el contenido de la cantidad más grande dela que procede. Si los errores de muestreo son despreciables, por ejemplo cuando se parte de un líquido y se toma una pequeña porción, la IUPAC sugiere el uso de ***porción de prueba*** *(test portion), alícuota o espécimen*.

Variables

Las variables pueden definirse como propiedades respecto de las que los elementos individuales de una muestra difieren de alguna forma mensurable. Pueden medirse en diferentes escalas:

Escala nominal: los objetos se describen con palabras. Las variables medidas de esta forma se denominan cualitativas, categóricas o atributos.

Escala ordinal: las variables se ordenan según una gradación, p.e. de muy malo a muy bueno. Las variables medidas así se denominan variables ordenadas (ranked)

Escalas cuantitativas o de medida. En ellas cada valor se expresa con un número. Si el cero no tiene sentido y es arbitrario (p.e. º C), la escala se llama de intervalo. Si el cero tiene significado (p.e. ºK, o cualquier concentración química) la escala se denomina a veces de razón (ratio), si bien este nombre no suele emplearse.

Otra clasificación divide las variables en **continuas** y **discretas**. Estas últimas suelen ser el resultado de conteo (nº de bacterias, nº de defectos de un material) y sus únicos valores son entonces números enteros. Es peligroso confundir las variables discretas con las medidas con una escala ordinal. Dependiendo del número de variables, la estadística se denomina univariada o multivariada.

Histogramas y distribuciones

Cuando se dispone de muchos datos y se quieren describir, resulta útil agruparlos en clases y visualizar su distribución con un histograma. El rango es el intervalo entre el mayor y el menor, y el número de clases debe fijarse de antemano., pero un punto de partida es utilizar:

$$nº_{\text{clases}} = \sqrt{nº \text{ de datos}}$$

Véase el ejemplo de Histograma realizado con EXCEL. Obsérvese la forma peculiar que tiene la macro 'Histograma' de 'Análisis de datos' a la hora de presentar las clases del histograma, lo cual conduce a un error (bug) a la hora de representar la gráfica.

Según la **ISO**:

Clase es cada uno de los intervalos consecutivos en los cuales se divide el intervalo total de variación.

Límites de clase son los valores que definen los límites inferior y superior de una clase.

El **punto medio** (mid-point) es la media aritmética de los lím ites y se denomina a veces **marca** (class mark) de la clase (la ISO no lo recomienda)

El intervalo de clase es la diferencia entre los mismos.

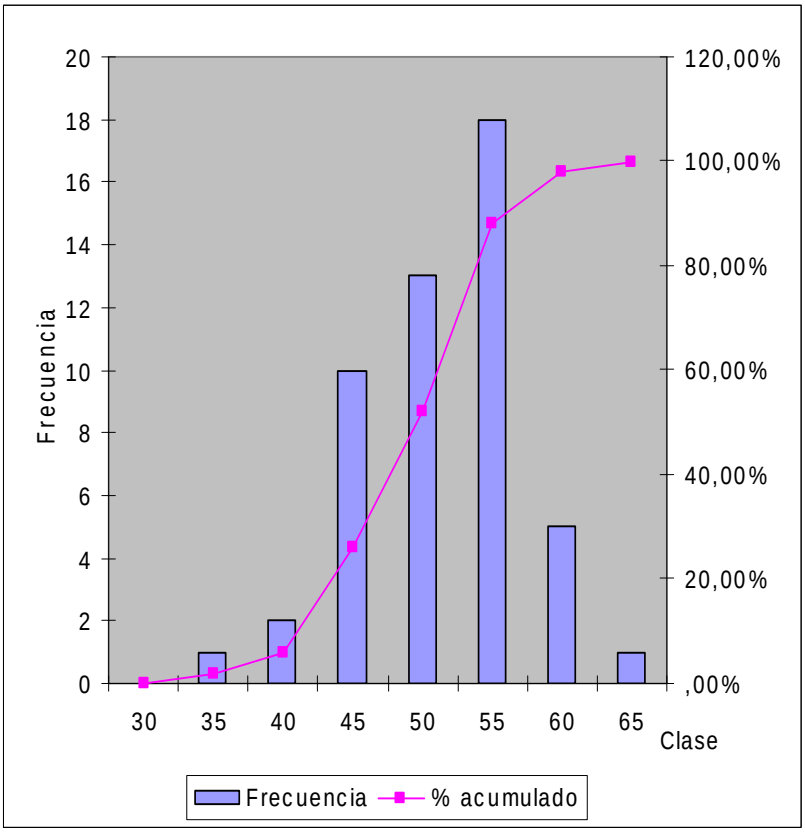
Ejemplo de histograma

Resultados procedentes de la determinación del contenido en calcio de aguas del canal de la reina Juana II

Datos
53,9
49,0
41,9
60,7
53,6
58,1
50,2
56,2
54,7
49,3
56,4
47,7
43,5
39,2
55,8
54,8
52,2
50,7
53,2
44,1
47,7
52,5
45,9
52,2
43,9
44,2
42,2
33,4
48,5
53,1
48,9
39,5
45,6
49,1
54,4
54,1
52,7
43,0
40,5
47,5
51,1
49,9
54,7
47,7
56,4
53,7
48,6
44,6
51,2
42,5

Histograma

	Clase	Frecuencia	% acumulado
	30	0	,00%
30	35	1	2,00%
35	40	2	6,00%
40	45	10	26,00%
45	50	13	52,00%
50	55	18	88,00%
55	60	5	98,00%
60	65	1	100,00%
65	y mayor...	0	100,00%



Si se cuenta el número de individuos en cada clase y se divide por el número total de individuos, se obtiene la **frecuencia relativa** de cada clase, y la tabla de estos valores se denominan **distribución de frecuencias relativas**. Suele representarse en función del punto medio de cada clase.

Si se suman todas las frecuencias hasta una determinada clase, se obtienen las **frecuencias acumuladas o acumulativas**. Se representa en función del punto medio de cada clase y dicha representación diagrama se suele denominar diagrama de **frecuencias (relativas) acumuladas o acumulativas**.

Todas estas distribuciones son **discretas**, ya que las frecuencias se dan para clases discretas o valores discretos de la variable (punto medio de cada clase). Si la variable puede tomar valores continuos, se obtienen distribuciones continuas. Si los datos de la tabla son verdaderamente representativos de la población, las frecuencias nos dan la probabilidad de encontrar ciertos valores en la población. Así en la tabla del contenido en calcio del agua del Canal de Castilla, la probabilidad de encontrar un contenido entre 40 y 45 mg/l es del 20,00 % y la de encontrar valores de hasta 45 mg/l es del 26,00 %. De esa manera las representaciones gráficas anteriores pueden considerarse como la **distribución de probabilidades (o función de densidad de probabilidad)** y **distribución de probabilidad acumulativa**. Hay una sutil diferencia entre la distribución de frecuencias y la distribución de probabilidades: la distribución de frecuencias describe los datos de la **muestra** estudiada. La distribución de probabilidad describe la **población**, es decir la distribución que se obtendría para un número infinito de datos.

ESTADÍSTICA DESCRIPTIVA

Se precisan: El número de observaciones n
Un parámetro para la tendencia central
Un parámetro para la dispersión

Promedio y medidas de centralización

Media (muestral) $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Media poblacional $\mu = \frac{\sum_{i=1}^{n} x_i}{N} = E(x)$

Media de datos agrupados $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$

Media ponderada $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Mediana: Valor central de una serie de datos
Ejemplo: Concentración de calcio en el Canal de Castilla

32,1	30,8	31,9	35,0	29,8	33,6	31,6	30,5	32,6	33,1

Ordenados:

29,8	30,5	30,8	31,6	31,9	32,1	32,6	33,1	33,6	35,0

Si n es impar, la mediana es el dato que aparece en el lugar (n+1)/2. Si n es par, la mediana es la media de los datos que aparecen en los lugares n/2 y (n+2)/2. En ste caso n= 10, luego la mediana es la media de los resultados que aparecen en los lugares 5 y 6, es decir (31,9+32,1)/2 = 32,0
La mediana es más **robusta** o insensible que la media. En la serie anterior tenemos que
Media: 32,1

Mediana: = 32,0
Si cambiamos 35,0 por 40,8
Media: 32,7 pero la mediana no cambia

Moda: Valor que se presenta con mayor frecuencia

Cuartiles: Al ordenar los datos de menor a mayor, la mediana divide a un conjunto en dos partes iguales. Cada una de esas dos partes puede dividirse en otras dos y se generan cuatro partes o cuartiles divididos por Q_1, Q_2 y Q_3. La mediana coincide con Q_2
Deciles: Valores D_1 a D_9 que dividen los datos en 10 partes iguales
Percentiles: Valores P_1 a P_{99} que dividen los datos en 100 partes iguales

Cálculo de cuartiles
Una vez calculada la mediana que coincide con Q_2, los valores de Q1 y Q3 se calculan igual pero en la primera y segunda mitad de los datos respectivamente (incluyendo la mediana)

Se ordenan los datos de mayor a menor y se determinan los números de los datos que corresponden a Q_1, Q_2 y Q_3 Si n es impar, la mediana Q_2 es el dato que aparece en el lugar (n+1)/2. Si n es par, la mediana es la media de los datos que aparecen en los lugares n/2 y (n+2)/2

Otra manera de calcular los números de los datos correspondientes a los cuartiles es a partir de n:

$$Q_1 = \frac{n+1}{4} = 0{,}25(n+1)$$

$$Q_2 = \frac{2(n+1)}{4} = 0{,}50(n+1)$$

$$Q_3 = \frac{3(n+1)}{4} = 0{,}75(n+1)$$

Si Q_1, Q_2 y Q_3 son números enteros los datos que ocupen las posiciones Q_1, Q_2 y Q_3 definen los cuartiles. Si se obtiene números decimales, Q_1, Q_2 y Q_3, se calculan por interpolación lineal. En el ejemplo del contenido en calcio del agua del Canal de Castilla, Q_1 = 24,51; Q_2 = 32,69 y Q_3 = 40,60.

Medidas de dispersión

Rango: $R = x_{max} - x_{min}$

Varianza muestral: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Varianza poblacional: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$

Desviación típica muestral: $s = \sqrt{s^2}$

Desviación típica poblacional: $\sigma = \sqrt{\sigma^2}$

Recorrido intercuartil: $Q_3 - Q_1$
Cuanto menor es, más agrupados están los datos

Boxplot (box and whiskers plot) Gráfico de caja y bigotes:

Es la representación gráfica de una caja de ancho indiferente pero cuya largura corresponde a los valores Q_1, Q_2 (la mediana) y Q_3 También se incluye un intervalo que engloba a los puntos más extremos comprendidos dentro de $1,5(Q_3-Q_1)$. Véase el ejemplo de boxplot. Lo ideal es una caja simétrica en torno a la mediana y de bigotes idénticos.

Ejemplo de boxplot

Sean los datos siguientes (ya ordenados)

Resultado	Orden
9,00	1
10,06	2
13,52	3
14,06	4
15,65	5
16,22	6
16,57	7
16,60	8
16,83	9
16,85	10
17,78	11
18,13	12
18,34	13
19,05	14
19,07	15
19,10	16
19,16	17
19,28	18
19,28	19
19,47	20
19,55	21
19,80	22
20,24	23
20,25	24
20,39	25

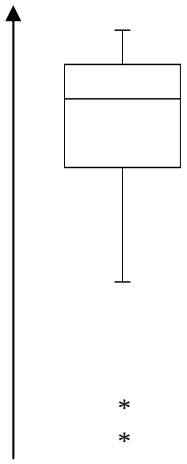
Al ser n impar, la mediana es el dato que aparece en (25+1)/2=13 lugar: 18,34

Q1 se calcula como la mediana de los primeros 13 puntos, o sea es el 7º valor: 16,57

Q3 se calcula en los últimos 13 puntos, o sea el el 19º valor: 19,28

El intervalo intercuartil es Q3-Q1= 19,28-16,57=2,71, y 1,5 veces ese valor vale 4,065, por lo que los límites extremos (bigotes) se extienden en principio hasta Q1-1,5(Q3-Q1)= 12,50 y hasta Q3+1,5(Q3-Q1)= 23,34. Todos los valores inferiores a 12,50 y superiores a 23,34 son outliers.

Se dibuja una caja cuyos límites corresponden a Q1 y Q3 (16,57 y 19,28), con la mediana (18,34) representada por una barra horizontal. Desde los extremos de la caja se dibujan unas líneas que van hasta el punto más remoto que no es un outlier. Por la parte inferior llega hasta 13,52, y por la parte superior llega hasta el último punto 20,39. Estos puntos más remotos se representan con un pequeña línea. Los outliers se dibujan como asteriscos (9,00 y 10,06).



Desviación típica relativa: $s_r = s/\bar{x}$; $s_r(\%) = s/\bar{x} * 100$

A veces se llama coeficiente de variación pero la IUPAC lo desaconseja

Desviación típica promediada (pooled)

Cuando se obtienen conjuntos de datos en diferentes momentos o en diferentes (pero similares) muestras, y se desea obtener la varianza (desviación típica) de los datos agrupados, se emplea la varianza (desviación típica) promediada (pooled), de acuerdo a:

$$s^2_{pooled} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{(n_1-1)+(n_2-1)+\ldots+(n_k-1)} = \frac{\sum (g.d.l)_j s_j^2}{\sum (g.d.l.)_j} \quad (j=1,\ldots k)$$

En el caso en que se trate de medidas replicadas apareadas, todos los $n_j=2$

$$s^2_d = s^2_{pooled} = \frac{\sum d_j^2}{2k}$$

siendo d_j las diferencias entre cada par de valores y k el número de parejas

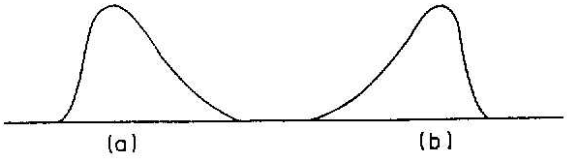
Medidas de la forma de la distribución

Coeficiente de sesgo $a_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$

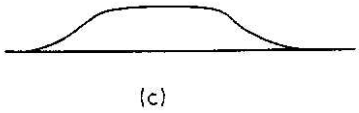
$a_3 > 0$. Sesgo positivo. La distribución tiene cola hacia la derecha. Media >Mediana
$a_3 = 0$: La distribución es simétrica. Media = Mediana
$a_3 < 0$. Sesgo negativo. La distribución tiene cola hacia la izquierda. Media > Mediana

Coeficiente de curtosis $a_4 - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$

a_4-3 > 0: La curva es apuntada (leptocúrtica)
a_4-3 = 0: La curva es normal
a_4-3 < 0: La curva es achatada (platicúrtica)



Sesgo positivo Sesgo negativo



Kurtosis negativa

MEDIDA DE LA CALIDAD

Calidad y errores

Quality Assurance (Aseguramiento de la Calidad) es un sistema de actividades cuyo propósito es proporcionar al productor o usuario de un producto o servicio la seguridad de que éste cumple con unas estándares o normas de calidad definidas con un nivel de confianza dado. Un proceso debe originar un producto con ciertas características dentro de ciertos márgenes de error. La calidad de una medida se obtiene si el resultado dado se aproxima al resultado correcto, es decir no está sujeto a un error mayor del que se considera aceptable.

Errores sistemáticos y aleatorios

Supongamos que el valor correcto para una determinación es 10,0

	Determinación							
Serie	**1**	**2**	**3**	**4**	**5**	**6**	**7**	$\bar{x}$
A	10,1	9,9	10,1	10,0	9,9	10,2	9,8	10,0
B	10,3	9,7	9,6	9,8	10,1	10,5	10,2	10,0
C	11,3	10,7	11,0	10,9	11,1	11,3	10,7	11,0
D	10,0	9,8	10,1	9,9	10,2	10,3	12,7	10,4
E	9,7	9,8	9,9	10,0	10,1	10,2	10,3	10,0

Se observa en la tabla anterior que A y B dan el valor correcto pero que sus resultados individuales están dispersos alrededor de esa media. Se dice que los resultados individuales están sujetos a **errores aleatorios** (random errors). En A la situación es mejor que en B, y se dice que la precisión de A es mejor que la de B.

En el caso C, todos los resultados son claramente mayores que el valor correcto: hay un **error sistemático**. Este error sistemático está acompañado también de un error aleatorio pues los resultados están dispersos: El **error sistemático siempre va acompañado del error aleatorio**, y gran parte de las pruebas estadísticas se dirigen a diferenciar entre ambos tipos de error. Los casos D y E se verán más adelante.

PRECISIÓN Y BIAS DE LAS MEDIDAS

El objetivo de una medida química es encontrar el verdadero valor de un parámetro químico. La **ISO** define el **valor verdadero** como: *"Valor que caracteriza una cantidad perfectamente definida en las condiciones que existen en el momento en que tal cantidad es observada. Es un valor ideal que solo puede obtenerse cuando todas las fuentes de error son eliminadas y la población es infinita*".

Hay dos causas por las que un resultado analítico difiera del valor verdadero: existencia de error aleatorio o de error sistemático. Si se obtiene un único resultado analítico, x_i, diferirá del valor verdadero μ_0, y la diferencia es el error: $e_i = x_i - \mu_0$

Si se hacen más medidas, es decir se analiza una muestra de una población, se obtendrá la media, $\bar{x}$, que estima μ, o la media de la población de medidas. Si la muestra es lo bastante grande, entonces $\bar{x} = \mu$.

$$e_i = (x_i - \bar{x}) + (\bar{x} - \mu_0) = (x_i - \mu) + (\mu - \mu_0)$$

La primera parte $(x_i - \bar{x})$ ó $(x_i - \mu)$ **es el error aleatorio, y la segunda parte** $(\bar{x} - \mu_0)$ **ó** $(\mu - \mu_0)$ **el error sistemático**.

Precisión

La **precisión** es una medida del componente aleatorio. Para la **ISO**, precisión es *"la cercanía entre resultados de pruebas independientes obtenidos bajos condiciones estipuladas" ("closeness of agreeement between independent test results obtained under stipulated conditions")*. Para la **IUPAC**: *"Cercanía entre los resultados obtenidos aplicando el procedimiento experimental varias veces bajo las condiciones prescritas. Cuanto menor es la parte aleatoria de los errores experimentales que afectan a los resultados, más preciso es el procedimiento".*

La precisión se cuantifica mediante la **desviación típica** que según la **ISO** es "la medida cuantitativa de la precisión" (o mediante la desviación típica relativa). Debemos recordar que s estima σ, y que cuando el número n de medidas replicadas es bastante grande, se puede cosiderar que $s=\sigma$. El valor de n varía, y según la IUPAC basta que $n \geq 10$, pero en muchos libros de estad´sitica se dice que $n \geq 25$ o 30. Dependiendo de las condiciones experimentales, hay dos tipos de precisión: **repetitividad (repeatability)** si las condiciones de medida son homogéneas: mismo método, mismo analista, mismo día, etc. y la **reproducibilidad (reproducibility)** que es la precisión obtenida en las condiciones más adversas (o heterogéneas). Asimismo se define la **precisión intermedia (intermediate precision) o reproducibilidad dentro del laboratorio (within-laboratory reproducibility)** que es la medida de la precisión dentro de un mismo laboratorio debida a la variación entre diferentes días (T), analistas o equipamiento (E).

Exactitud, veracidad y bias

En el caso de los errores sistemáticos, la terminología es confusa y ambigua ya que aparecen los términos **exactitud (accuracy)** y **veracidad (trueness)** que se usan indiscriminadamente dependiendo del organismo. Para la **ISO**, **veracidad** es *"la concordancia entre el valor medio de una gran serie de medidas y el valor aceptado como referencia" ("closeness of agreement between the average result obtained from a large series of test results and the acepted reference value")*, mientras que la exactitud es *"la cercanía entre un resultado*

(test result), que puede ser la media de una serie de medidas) y el valor aceptado como referencia" ("the closeness of agreement between test result and the acepted reference value"). La diferencia entre ambos conceptos es que la ISO aplica la veracidad a un gran número de series de medidas, mientras que la exactitud se refiere a un único resultado (o a la media de una única serie). La IUPAC utiliza para exactitud una definición similar, pero no reconoce el término veracidad.

La **ISO** cuantifica la veracidad mediante el **bias**[1], que es *"la diferencia entre la esperanza de un resultado y un valor aceptado como referencia" ("the difference between the expectation of the test results and an acepted reference value").* Es decir que para la **ISO**, bias y veracidad son esencialmente la misma cosa: veracidad es el concepto y el bias su medida. Por su parte, la **IUPAC** da la misma definición de bias pero no reconoce el término veracidad, sino solamente la exactitud (de la media) que se refiere a la "concordancia entre el valor verdadero y la media de la población obtenida aplicando el procedimiento experimental un gran número de veces". Veamos un ejemplo para clarificar (¿) la cuestión:

Supongamos una magnitud cuyo valor verdadero se conoce: $\mu_0 = 100$. Se hace un único análisis y se obtiene 93: para la ISO la exactitud es -7. Si se hacen varios análisis (p.e. 5) como en un estudio de repetitividad, y su media es 88, para la ISO la exactitud sería –12. Si se tienen varias medias y su media fuera 91, la ISO diría que la veracidad es –9, mientras que la IUPAC diría ahora que la exactitud es –9. Ambas organizaciones dirían ahora que el bias es –12.

La definición más común de **bias** es: $\Delta = \mu - \mu_0$, siendo μ la media de la población de resultados experimentales, y μ_0 el valor verdadero. μ no es conocido, sino solo estimado a partir de $\bar{x}$, por lo que el bías debería de ser definido como $D = \bar{x} - \mu_0$. El bias tiene dos componentes: bias del método (error inherente al método) y bias del laboratorio (introducido por un laboratorio concreto)

La diferenciación entre los bias del laboratorio y los del método, así como la evaluación de la repetitividad y de la reproducibilidad, no es tan simple como parece a primera vista y puede obligar a trabajar con la ayuda de otros laboratorios (Véase validación de métodos). Para detectar los bias es preciso disponer de un Material de Referencia Certificado que proporcione un valor verdadero que sirva de referencia al determinado por el o los laboratorios.

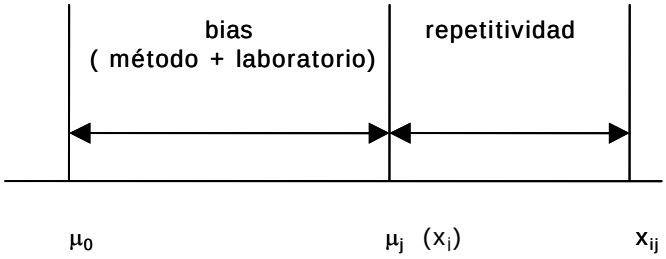
Si un laboratorio trabaja en solitario solo tiene acceso a la media de sus determinaciones $\bar{x}_j$ (que estima μ_j) por lo que el **bias** que puede detectar incluye tanto al **bias del método** como al componente sistemático que dicho laboratorio introduce **(bias del laboratorio).** Evidentemente, si el laboratorio utiliza un método normalizado (standard) libre de bias del método, cualquier bias que se detecte será achacable únicamente al componente del laboratorio. En cuanto a la precisión solo podrá evaluar la **repetitividad** (obtenida en condiciones homogéneas) y la **precisión intermedia** (obtenida al variar factores como el tiempo (T), el operador (O) o el equipamiento (E)).

Repetitividad

$$e_{ij} = (x_{ij} - \mu_j) + (\mu_j - \mu_0)$$

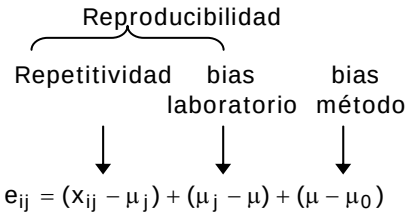
bias (método + laboratorio)

o de forma gráfica

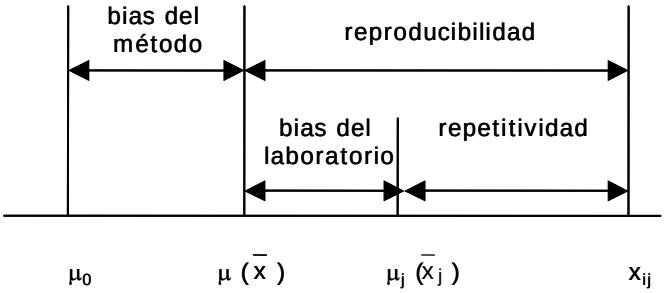


[1] Este término inglés no tiene equivalente en castellano y suele traducirse por sesgo

Si **se une a otros laboratorios** en un estudio inter-laboratorio, además de la media de sus determinaciones, μ_j, se tendrá acceso a la media de las medias de los diferentes laboratorios, μ, y se podrá **diferenciar por tanto entre los bias** del método y el bias que introduce el laboratorio. Seguirá teniendo acceso a su r**epetitividad**, pero ahora se podrá evaluar además la **reproducibilidad**.



o de forma gráfica:



Obsérvese que dependiendo de la situación, el bias del laboratorio puede ser considerado como parte del error sistemático (un laboratorio individual) o del aleatorio (muchos laboratorios tomados en conjunto). Para evitar ambigüedades sobre la naturaleza sistemática o aleatoria de los errores, la ISO define el término **incertidumbre (uncertainty):** Un estimador añadido a un resultado que caracteriza el rango de valores dentro de los cuales se afirma que está el valor verdadero. Este rango viene afectado por diversos componentes, tanto aleatorios como sistemáticos.

OTROS TIPOS DE ERROR

Errores espurios, groseros o inaceptables (gross errors)

Conducen a valores aberrantes (outliers). Se obtienen por un fallo que no es sistemático ni aleatorio, p.e. en la tabla el resultado de 12,7 en la serie D claramente superior al resto. Estos outliers falsifican las estimaciones estadísticas y deben ser detectados y eliminados (evidentemente hay que encontrar la razón de su aparición)

Deriva (drift)

Aparece en el caso de procesos que no están bajo control estadístico (Véase Quimiometría y Control de Calidad): su media y su desviación típica no son constantes. P.e. la situación de la serie E de la tabla que muestra un aumento constante de los valores encontrados.

Ruido de la línea base (baseline noise).

Las medidas en química suelen obtenerse por diferencia entre una señal obtenida cuando se mide el analito y una señal obtenida para un **blanco**. Hay varios tipos de blancos, pero por ahora consideraremos como blanco al que está compuesto del mismo material que la muestra pero sin analito. El ruido del blanco se superpone a la medida de la muestra y en ocasiones es indistinguible, p.e. cuando se opera por debajo del **límite de detección**.

PROPAGACIÓN DE ERRORES

Si el resultado final se obtiene a partir de varias medidas **independientes** entre sí, o cuando está influido por varias fuentes de error **independientes** (p.e. muestreo y medida), los errores se propagan y pueden acumularse o compensarse.

Los errores **aleatorios siempre se acumulan** de acuerdo a:

$$\sigma_y^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2 + \left(\frac{\partial f}{\partial t}\right)^2 \sigma_t^2 + ...$$

cuando y = f(x, z, t...)

Se demuestra que las **varianzas son aditivas cuando la función es adición o sustracción, y que cuando es multiplicación o división lo que son aditivos son los cuadrados de las desviaciones típicas relativas**. Debe hacerse hincapié en que esto solo se cumple si las variables son independientes, lo cual a veces no es el caso.

Los errores **sistemáticos se propagan con su signo**. P.e si Δy es el eror sistemático que afecta a y, para una ecuación de adición y/o sustracción se cumple:

$$y = a + b\,x + c\,z - d\,t$$
$$\Delta y = b\,\Delta x + c\,\Delta z - d\,\Delta t$$

Si la relación es multiplicativa

$$y = axz/t$$
$$\Delta y/y = \Delta x/x + \Delta z/z - \Delta t/t$$

Es decir los errores sistemáticos pueden compensarse entre sí.

Para otros tipos de relaciones véase la Tabla resumen que se presenta a continuación

Resumen de fórmulas de propagación de errores

Función	Error aleatorio	Error sistemático
Combinaciones lineales $y = k + k_a a + k_b b + \ldots$	$s_y = \sqrt{(k_a s_a)^2 + (k_b s_b)^2 + \ldots}$	$\Delta y = k_a \Delta a + k_b \Delta b + \ldots$
Expresiones multiplicativas $y = k\,ab/c$	$\frac{s_y}{y} = \sqrt{\left(\frac{s_a}{a}\right)^2 + \left(\frac{s_b}{b}\right)^2 + \left(\frac{s_c}{c}\right)^2}$	$\frac{\Delta y}{y} = \frac{\Delta a}{a} + \frac{\Delta b}{b} - \frac{\Delta c}{c}$
Potencias $y = a^k$	$\frac{s_y}{y} = \left\lvert k \frac{s_a}{a} \right\rvert$	$\frac{\Delta y}{y} = k \frac{\Delta a}{a}$
Función de una variable independiente $y = f(a)$	$s_y = \left\lvert s_a \frac{dy}{da} \right\rvert$ ó $s_y^2 = s_a^2 \left(\frac{dy}{da}\right)^2$	$\Delta y = \Delta a \frac{dy}{da}$
Función de varias variables independientes $y = f(a, b, \ldots)$	$s_y^2 = s_a^2 \left(\frac{\partial y}{\partial a}\right)_b^2 + s_b^2 \left(\frac{\partial y}{\partial b}\right)_a^2 + \ldots$	$\Delta y = \Delta a \left(\frac{\partial y}{\partial a}\right)_b + \Delta b \left(\frac{\partial y}{\partial b}\right)_a + \ldots$

Estas ecuaciones han adquirido gran importancia en **metrología**, porque permiten describir las fuentes individuales de error y combinarlas en lo que se denomina **incertidumbre** (uncertainty). Esta magnitud se define como un intervalo en el cual debe estar incluido el valor

verdadero. Si la incertidumbre se expresa como desviación típica, se denomina incertidumbre típica. Cuando hay varias fuentes de error debe utilizarse una incertidumbre típica combinada, obtenida mediante las leyes de propagación de errores.

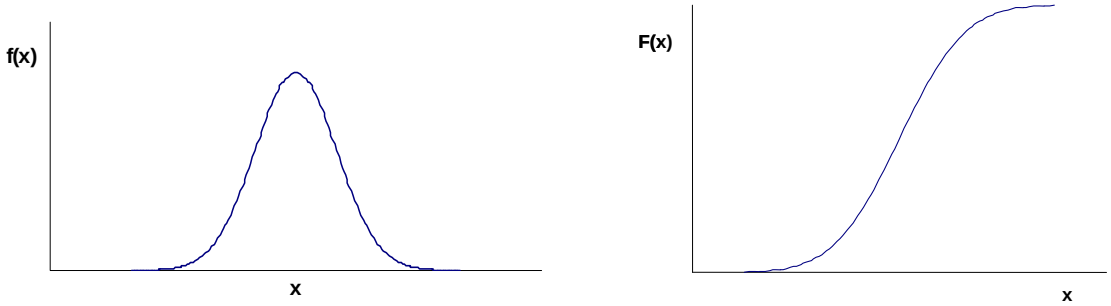
DISTRIBUCION NORMAL O GAUSSIANA

Cuando se tienen infinitos datos, en vez de una diagrama de distribución de frecuencias (histograma), se obtiene una distribución continua de probabilidades (o **función de densidad de probabilidad**), que para distribución normal o gaussiana tiene la fórmula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

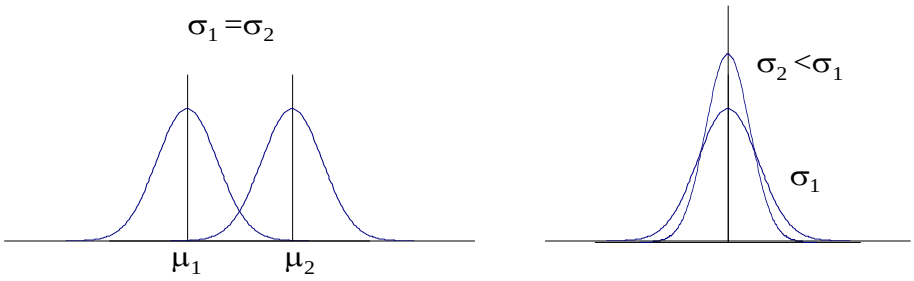
La **distribución de probabilidad normal acumulativa** es

$$F(x_0) = \int_{-\infty}^{x_0} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] dx$$



siendo μ la media de la población y σ su desviación típica $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{N}}$

Los parámetros que describen una distribución normal son μ y σ^2: **$N(\mu,\sigma^2)$**, y su efecto sobre la función de densidad de probabilidad es:



<u>Propiedades de la distribución normal</u>

- Simétrica respecto a la media
- Media, mediana y moda coinciden
- Un aumento (disminución) de σ origina un ensanchamiento (estrechamiento), es decir una mayor (menor) dispersión
- Un aumento (disminución) de μ origina un desplazamiento
- El rango $\mu \pm \sigma$ incluye al 68,26% de los datos
- El rango $\mu \pm 2\sigma$ incluye al 95,44% de los datos
- El rango $\mu \pm 3\sigma$ incluye al 99,74% de los datos

-

Distribución normal tipificada o estandarizada

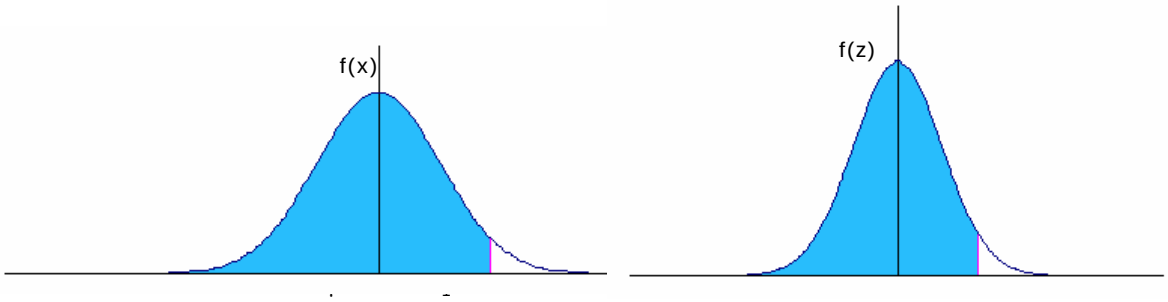
Tanto μ como σ tienen su escala y unidades, por lo que las formas y valores de f(x) y F(x) serían infinitas. Para evitar este efecto se hace una transformación por **escalado o autoescalado**. Se calcula **$z = (x-\mu)/\sigma$**, con lo que se obtiene una nueva distribución **N(0,1)**. Sus f(x) y F(x) son respectivamente:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(z)^2}{2}\right]$$

$$\Phi(z_0) = \int_{-\infty}^{z_0} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(z)^2}{2}\right] dz$$

Por tanto si se tiene una variable x $N(\mu,\sigma^2)$, para un valor dado x_1, si se calcula $z_1 = (x_1 - \mu)/\sigma$, se cumple que:

$P(X \le x_1) = \int_{-\infty}^{x_1} f(x)dx$ y $P(Z \le z_1) = \int_{-\infty}^{z_1} f(z)dz$ son idénticas



Tablas para la distribución estandarizada

La transformación z permite que en vez de hacer infinitos cálculos y gráficas para las infinitas combinaciones de μ y σ, solo se precisen cálculos y gráficas para la variable z, que se obtiene por transformación de cualquier variable x. Por tanto solo se necesita **una tabla, que se presenta de varias formas .**

Las tablas pueden ser de **una o dos colas**. Las tablas de dos colas dan qué parte del área cae dentro o fuera de un intervalo (+z, -z). La Tabla 3.1 da qué valor de z corresponde con una probabilidad p dada dividida en dos colas.
Las Tablas 3.2 y 3.3 dan el valor de p correspondiente a un z dado. La Tabla 3.3 es acumulativa (es igual a la3.2 con todos los valores de p incrementados en 0,500)

Funciones EXCEL

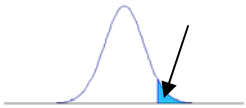
DISTR.NORMAL.ESTAND(z) devuelve la probabilidad desde $-\infty$ a z
DISTR.NORMAL.ESTAND.INV(p) devuelve el valor de z tal que la probabilidad desde $-\infty$ a z valga p

Ejemplo
Durante nuestro proceso de fabricación de ampollas para luciferina, hemos comprobado que el volumen de dichas ampollas tiene una distribución normal, con una media de 4,98 mL y una desviación típica de 0,13 ml. Debido a que vamos a exportar una partida de 500 ampollas a Bechuanalandia deseamos predecir:

1) ¿Cuántos de ellos estarán por encima de 5,20 ml?

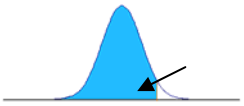
$$x = 5{,}20 \text{ ml} \Rightarrow z = \frac{5{,}20 - 4{,}98}{0{,}13} = 1{,}690$$

P(x>5,20) = P(z>1,690)
Necesitamos
Conocer



1,690

La tabla 3.3
nos da



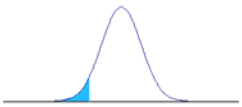
1,690

La tabla 3.3 indica para z = 1,690, P = 0,954 luego la proporción buscada es P = 1- 0,954 = 0,046, por lo que el número de ampollas que tendrá un volumen superior a 5,20 ml es 500*0,046 = 22,8 ≈ 23 ampollas
(EXCEL da DIST.NORMAL.ESTAND(1,690)= 0,954)

2) ¿Cuántos de ellos contendrán menos de 4,80 ml?

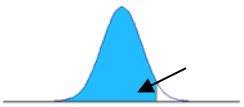
$$x = 4{,}80 \text{ ml} \Rightarrow z = \frac{4{,}80 - 4{,}98}{0{,}13} = -1{,}385$$

P(x<4,80) = P(z<-1,385)
Necesitamos
Conocer



--1,385

La tabla 3.
nos da (pero es simétrica)



1,385

De la tabla 3.3 para z = 1,385 , P = 0,917, luego la proporción buscada es P = 1- 0,917 = 0,083, por lo que habrá 500*0,083 = 41,5 ≈ 42 ampollas
(EXCEL da directamente el valor buscado DIST.NORMAL.ESTAND(-1,385)= 0,083)

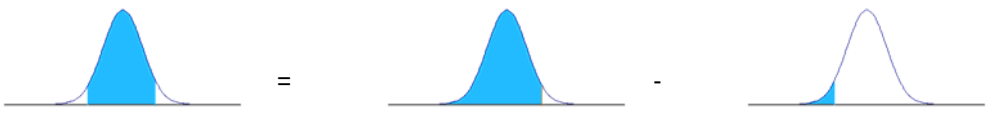
3) ¿Qué proporción de matraces estará comprendida entre 4,95 y 5,05 ml?

$$x_1 = 4{,}95 \text{ ml} \Rightarrow z_1 = \frac{4{,}95 - 4{,}98}{0{,}13} = -0{,}231$$

$$x_2 = 5{,}05 \text{ ml} \Rightarrow z_2 = \frac{5{,}05 - 4{,}98}{0{,}13} = 0{,}538$$

P(4,95 <X<5,05) = P(-0,231<Z<0,538)

Necesitamos



-0,231 0,538 0,538 -0,231

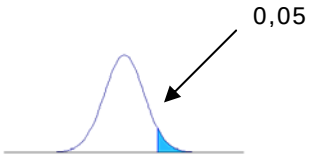
La primera área sale directamente de la Tabla 3.3 para z = 0,538 y vale P = 0,705
(**EXCE**L DISTR.NORMAL.ESTAND(0,538)=0,705)

La segunda área sale como en apartado 1) de la misma tabla 3.3. Para z= 0,231 sale P =0,591 pero teniendo en cuenta que es simétrica, el valor que nos interesa es P = 1 - 0,591 = 0,4,09, por lo que el resultado final es **P = 0,705 - 0,409 = 0,296**. El número de ampollas según las especificaciones será 500*0,296 = 148.
(EXCEL da directamente el resultado DISTR.NORMAL.ESTAND(-0,231)=0,409 para la segunda área)

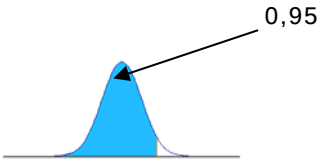
4) Si se desea que sólo un 5% de los matraces esté por encima del volumen especificado ¿cuál debería ser éste?

Es una aproximación inversa, ya que conocemos P = 0,05 y necesitamos saber z

Necesitamos conocer:



La Tabla 3.3 nos da



De la Tabla 3,3 para P = 0,95 sale z = 1,65
(**EXCE**L DISTR.NORMAL.ESTAND.INV(0,95) = 1,65)

El valor de x, sale invirtiendo la transformación z:

$$x_1 = 4{,}98 + 1{,}65 * 0{,}13 = 5{,}19\,\text{ml}$$

TEOREMA DEL LÍMITE CENTRAL Y DISTRIBUCIÓN DE MEDIAS MUESTRALES

Enunciado

Si se tiene un suma de n variables aleatorias independientes x_i, cuyas distribuciones no son necesariamente normales

$$y_i = x_1 + x_2 + ... + x_n$$

con medias μ_i y varianzas σ_i^2, para grandes valores de n, la distribución de y es **aproximadamente normal** con una media $\Sigma\mu_i$ y varianza $\Sigma\sigma_i^2$.
Este teorema tiene gran importancia pues explica porqué las distribuciones de errores aleatorios tienden a la normalidad, puesto que el error global suele ser una combinación lineal de componentes independientes.

Si de una población con media μ y varianza σ^2 se toman todas las posibles muestras de tamaño n, la distribución de las medias muestrales tendrá una media μ y una varianza σ^2/n. La distribución de las medias será **normal** si la población original es normal. Será **aproximadamente normal** para cualquier tipo de distribución de la población original, tanto más cuanto más grande sea n.

El valor crítico suele ser n>30, pero para poblaciones no normales, simétricas y unimodales, se obtienen medias distribuidas normalmente con muestras de tamaño 4-5.

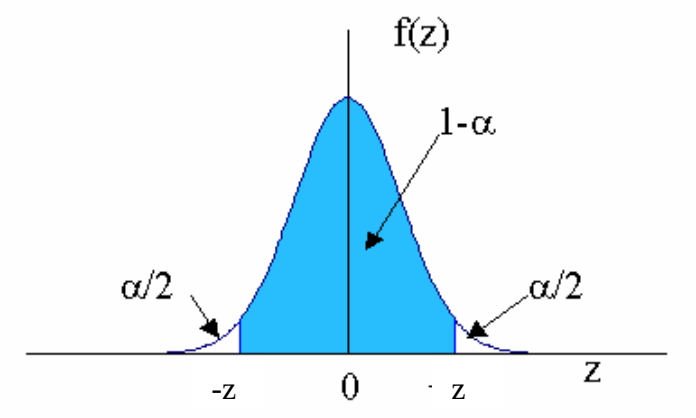
Intervalos de confianza de la media

Para una distribución normal, el 95% de los datos cae dentro de los límites de z=-1,96 a z=+1,96 (Tabla 3.1). Esto puede ser refraseado para indicar que el 95% de los datos caen dentro del intervalo $\mu \pm 1{,}96\sigma$.

Lo anterior puede aplicarse a la distribución de las medias muestrales, luego podemos decir que el 95% de los datos estará dentro del intervalo $\mu \pm 1{,}96\sigma/\sqrt{n}$. Es decir:

$$\mu = \bar{x} \pm 1{,}96\sigma/\sqrt{n}\ .$$

En general, $\mu = \bar{x} \pm z_{\alpha/2}\,(\sigma/\sqrt{n})$ con 100-α% de confianza, donde α se deriva de una tabla de z de dos colas.



Muestras pequeñas y la distribución t

En las ecuaciones anteriores se utiliza **σ**. En muchas ocasiones únicamente se conoce su **estimación s**. Si n≥30 (ó 25 según investigadores), puede hacerse la sustitución:

$$\mu = \bar{x} \pm z_{\alpha/2}\,(s/\sqrt{n}) \quad (n \geq 30)$$

Si **n<30**,s es un estimador incierto de σ, y **se sustituye z por t**

$$\mu = \bar{x} \pm t_{\alpha/2,(n-1)}\,(s/\sqrt{n}) \quad (n < 30)$$

siendo α el nivel de significación, y n-1 el número de grados de libertad con el que debe buscarse t en la correspondiente tabla de t de dos colas. Habitualmente, se emplea un nivel de significación α = 0,05 (nivel de confianza del 95 %).

Si la tabla t que utilizamos es de dos colas, no habrá problema y deberá utilizarse la columna encabezada por p = α = 0,05. Si la tabla de la que disponemos es de una cola, para poder emplearla en este caso, habrá que buscar en la columna encabezada por p = α = 0,025

Funciones EXCEL

DISTR.T.INV(p;g.d.l) devuelve el valor de t con el nº de g.d.l. elegido y el nivel de probabilidad p, en una **tabla de dos colas**
Para que sirva para **una cola** hay que ponerla como DISTR.T.INV(2p;g.d.l)

NOTA IMPORTANTE SOBRE TABLAS DE UNA Y DOS COLAS
Si se dispone de una **tabla de una cola o de la función EXCEL** DISTR.NORMAL.ESTAND.INV(p)

a) Si lo que se desea buscar es de una cola no hay problema: se entra por la columna encabezada por (o se sustituye en la función) p = α = 0,05
b) Si lo que se desea es de dos colas, hay que entrar por la columna encabezada por (o se sustituye en la función) p = α/2 = 0,05/2= 0,025

Si se dispone de una **tabla de dos colas o de la función EXCEL** DISTR.T.INV(p;g.d.l)

a) Si lo que se desea buscar es de dos colas no hay problema: se entra por la columna encabezada por (o se sustituye en la función) p = α = 0,05
b) Si lo que se desea es de una cola, hay que entrar por la columna (o se sustituye en la función) p = 2α = 2*0,05 = 0,10

Pruebas de normalidad

No todas las distribuciones son normales. El **conocer si una determinada muestra o población sigue una distribución normal es importante**, ya que permite detectar un efecto que no es explicable mediante errores de medida aleatorios. Aunque la comprobación puede hacerse mediante una prueba χ^2, es útil disponer de métodos gráficos que permitan visualizar de forma rápida la normalidad o no de la distribución. Uno de ellos se denomina **rankit** y es el más recomendado por la ISO.

Supongamos una serie de medidas:

2,286 ; 2,327 ; 2,388; 3,172 ; 3,158 ; 2,751 ; 2,222; 2,367 ; 2,247 ; 2,512 ; 2,104; 2,707

En primer lugar se ordenan los datos:
2,104; 2,222; 2,247; 2,286; 2,327; 2,367; 2,388; 2,512 ; 2,707; 2,751; 3,158; 3,172

Puesto que tenemos 12 (n) valores, podemos diferenciar esta muestra en 12+1 (n+1) intervalos, todos los cuales tienen una frecuencia absoluta unidad. La frecuencia acumulativa se obtiene sumando uno a cada frecuencia absoluta anterior y coincide con el número de orden de cada uno de los datos (de ahí la denominación del método, ya que en inglés rank significa orden). Cuando hay varios valores iguales, todos ellos tienen la misma Frecuencia acumulativa (mismo rango) que es la media de las que les corresponderían.

El porcentaje de frecuencia relativa correspondiente es:

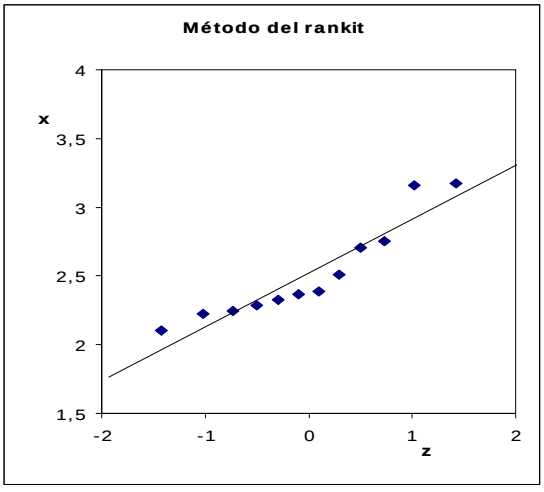
$$\%\ \text{frec. acum.} = (100 \times \text{fre. acum}) / (n+1)$$

Si la distribución fuera normal, el porcentaje de frecuencia relativa acumulativa de un determinado valor coincidiría con la probabilidad de que dicho valor apareciera y esa probabilidad se puede convertir en valores z. Así un valor experimental con una frecuencia relativa acumulada del 23,1 % debería ser equivalente a un valor de z que delimitara la **cola inferior** de una distribución normal con una cola de P = 0,231. Ese valor se obtiene con la adecuada tabla de z o bien con EXCEL (Función =DISTR.NORM.ESTAND.INV(xx)) y vale –0,740. Esos valores z (que también aparecen ordenados) se denominan rankits.

Puesto que existe una relación lineal entre x y z, la representación de los valores de la medida original (x) frente a los de z, debe dar una **línea recta**. (Deduzca como podrían estimarse los valores de μ y σ a partir de dicha representación).

Medidas (x)	Frec. Acum.	% Frec. Acum.	Z
2.104	1	7.7	-1.43
2.222	2	15.4	-1.02
2.247	3	23.1	-0.74
2.286	4	30.8	-0.50
2.327	5	38.5	-0.28
2.367	6	46.1	-0.10
2.388	7	53.8	0.10
2.512	8	61.5	0.28
2.707	9	69.2	0.50
2.751	10	76.9	0.74
3.158	11	84.6	1.02
3.172	12	92.3	1.43

El procedimiento es rápido pero en ocasiones no se sabe si la línea es recta o no. El método permite **detectar muy fácilmente la presencia de outliers** (espurios).



En la práctica se utiliza **papel de probabilidad normal**. En él hay un eje lineal en el que se representan los valores de x, y un eje no lineal en el que se representan los correspondientes porcentajes de frecuencia acumulada (no es preciso calcular los valores z)

Ejercicio: Estudiar si los datos que a continuación se exponen siguen una distribución normal. Utilice el método del rankit y represente los datos también en papel de probabilidad normal.

Valores	Valores ordenados	Frecuencia absoluta	Frecuencia acumulativa	Valor de z
0,0819	0,0797			
0,0806	0,0805			
0,0814	0,0806			
0,0805	0,0808			
0,0821	0,0811			
0,0826	0,0811			
0,821	0,0814			
0,0826	0,0819			
0,0948	0,0821			
0,0811	0,0821			
0,0831	0,0823			
0,1095	0,0823			
0,0824	0,0824			
0,0824	0,0824			
0,0811	0,0826			
0,0844	0,0826			
0,0797	0,0831			
0,0823	0,0844			
0,0808	0,0948			
0,0823	0,1095			

OTRAS DISTRIBUCIONES

Distribución Binomial

Una población binomial es aquella cuyos elementos pertenecen a **dos categorías mutuamente excluyentes**. Por ejemplo, un producto puede ser defectuoso o no, un paciente puede estar enfermo o sano....

Sea una serie de medidas o experimentos independientes cuyo resultado puede ser A o $\overline{A}$ (No A). La probabilidad de que ocurra A o $\overline{A}$ será:

$$P(A) = \Pi$$

$$P(\overline{A}) = 1 - P(A) = 1 - \Pi$$

Si se realizan n experimentos, la probabilidad de que A ocurra i veces será:

$$P(i) = C_n^i \Pi^i (1-\Pi)^{n-i} = \frac{n!}{i!(n-i)!} \Pi^i (1-\Pi)^{n-i}$$

Ejemplo: La probabilidad de encontrar un elemento defectuosos en un lote de productos es 0,02. Calcule la probabilidad de que aparezcan 2 elementos defectuosos en una muestra aleatoria de 10 elementos.

$$P(2i) = C_{10}^2\, 0{,}02^2 (1-0{,}02)^{10-2} = \frac{10!}{2!(10-2)!} 0{,}02^{2i} (0{,}98)^8 = 0{,}0153 = 1{,}53\%$$

La media y la varianza de una distribución binomial son $\mu = n\Pi$ y $\sigma^2 = n\Pi(1-\Pi)$. (En el ejemplo $\mu =$ y $\sigma^2 =$)

La distribución binomial se aplica a:
1) Muestreo sin reemplazamiento de una población muy grande comparada con el tamaño de la muestra
2) Muestreo con reemplazamiento de una población finita
3) Muestreo de procesos continuos de fabricación con población grande

Cuando $n\Pi$ y $n(1-\Pi)$ son grandes (>5) la distribución binomial tiende a la normalidad.

Distribución De Poisson

Describe **variables discretas relacionadas con sucesos discretos en un intervalo continuo, tal como tiempo, espacio o volumen**. Se supone que los sucesos ocurren de forma aleatoria e independiente. Por ejemplo: el número de caramelos defectuosos de una fábrica, el número de averías mensuales en la maquinaria, el número de cuentas en la desintegración de un radioisótopo...

La probabilidad de que ocurra un número i de sucesos viene dada por

$$P(i) = \frac{e^{-\lambda}\lambda^i}{i!}$$

siendo λ la media, o número de sucesos que tiene lugar en un período dado de tiempo, espacio o volumen.

La media y la varianza en una distribución de Poisson coinciden: $\mu = \sigma^2 = \lambda$

La distribución de Poisson puede considerarse como una distribución binomial en la cual n tiende a infinito, Π tiende a cero y $n\Pi$ tiende a λ.

Si $\lambda > 10$, la distribución de Poisson tiende a la normalidad

Distribución χ^2 o de Pearson

Si un conjunto de variables independientes $z_1, z_2, \ldots, z_n$ están distribuidas según una distribución normal unidad N(0,1), la variable $\chi^2 = \sum_1^n z_i^2$ tiene una función de densidad de probabilidad con ν=n grados de libertad

$$f(\chi^2) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(\chi^2)^{\nu/2-1}e^{-\chi^2/2} \qquad ;\nu \geq 0; \quad 0 \leq \chi^2 \leq \infty; \nu = n-1$$

Γ es la función gamma

La función es asimétrica con una cola hacia la derecha cuando ν es pequeño. Según aumenta el número de grados de libertad, la distribución χ^2 tiende a la normalidad.

La función χ^2 está tabulada para diversos grados de libertad, por lo que en la práctica no es preciso recordar ni calcular la función de densidad de probabilidad

La media y la varianza son: $\mu = \nu$ y $\sigma^2 = 2\nu$

Habitualmente las variables independientes iniciales no están normalizadas. Si tenemos n variables independientes $x_1, x_2, \ldots, x_n$ con distribución normal $N(\mu,\sigma^2)$ podemos obtener una nueva variable independiente

$$\sum_{i=1}^{n}\frac{\left(x_i - \bar{x}\right)^2}{\sigma^2} = (n-1)\frac{s^2}{\sigma^2}$$

que estará distribuida como χ^2 con $\nu = n-1$ grados de libertad.

La distribución χ^2 se utiliza en el **test χ^2** para comprobar si la distribución de los datos de una muestra de tamaño n se ajusta a una cierta distribución teórica, habitualmente la normal.

Distribución t de Student

Se utiliza para **describir muestreos con pocos elementos (n<30)**, es decir para describir la distribución de una muestra en vez de la distribución de la población.

Sea z una variable aleatoria N(0,1) y χ^2 otra variable aleatoria independiente de la anterior, con una distribución χ^2_ν . Se puede definir una nueva variable t como

$$t = \frac{z}{\sqrt{\chi^2/\nu}}$$

La distribución t de Student con ν grados de libertad se describe mediante la función de densidad de probabilidad:

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\nu+1/2)}{\Gamma(\nu/2)} \frac{1}{\left(1+t^2/\nu\right)^{\nu+1/2}}$$

Esta función no es preciso recordarla ya que está tabulada para diferentes grados de libertad.

La distribución t con ν grados de libertad es simétrica en torno a cero, y su varianza es $\sigma^2 = \nu/(\nu-2)$. Se parece a la normal, pero es algo más apuntada y tiene la base más estrecha. **Tiende a la normalidad cuando ν tiende a infinito, y en la práctica coincide con la normal tipificada z cuando $\nu > 30$.**

Se utiliza siempre que no se conocen las verdaderas media y varianza de una distribución para:

1) Estimar intervalos de confianza
2) Evaluar la veracidad de un resultado (ausencia de bias)
3) Comparar medias obtenidas por métodos diferentes

Distribución F de Fischer

1) Sean x e y dos variables aleatorias independientes que siguen distribuciones χ^2 con ν_1 y ν_2 grados de libertad respectivamente. Se puede definir la variable F

$$F = \frac{x/\nu_1}{y/\nu_2} = \frac{\chi_x^2/\nu_1}{\chi_y^2/\nu_2}$$

con una distribución F cuya función de densidad de probabilidad f(F) es

$$f(F;\nu_1,\nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} F^{\frac{\nu_1-2}{2}} (\nu_2+\nu_1)^{-\frac{(\nu_1+\nu_2)}{2}} \quad ; \quad 0 < F < \infty$$

La media y la varianza son:

$$E(F) = \frac{\nu_2}{\nu_2-2} \quad ; \quad \nu_2 > 2$$

$$Var(F) = \frac{\nu_2^2(2\nu_2+2\nu_1-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)} \quad ; \quad \nu_2 > 4$$

2) Sea $x_1, x_2, \ldots x_n$ una muestra aleatoria de variables aleatorias independientes con distribución $N(\mu_x, \sigma_x^2)$ e $y_1, y_2, \ldots y_n$ una muestra aleatoria de variables aleatorias independientes con distribución $N(\mu_y, \sigma_y^2)$

Las variables $\chi_x^2 = \frac{(n_x-1)s_x^2}{\sigma_x^2}$ y $\chi_y^2 = \frac{(n_y-1)s_x^2}{\sigma_y^2}$ siguen una distribución χ^2 con $\nu_x = n_x-1$ y $\nu_y = n_y-1$ grados de libertad respectivamente. La variable F es en este caso es

$$\frac{\qquad}{\qquad} \qquad \frac{\dfrac{2}{x}}{\dfrac{2}{y}} \qquad \frac{2}{y}$$

BIBLIOGRAFÍA

Massart D.L., Vandegintse B.G.M., Buydens L.M.C., De Jong S., Lewi P.J. and Smeyers-Verbeke J., *Handbook of Chemometrics and Qualimetrics*. Elsevier, Amsterdam, 1997

Sharaff M.A., Illman D.L. and Kowalski B.R. *Chemometrics*, Wiley-Interscience, New York, 1986

Eurachem/CITAC, *Quantifying uncertainty in analytical measurement*, 2nd Ed., 2000
http://www.eurachem.ul.pt/guides/QUAM2000-1.pdf