QUIMIOMETRÍA

DISCIPLINA QUÍMICA QUE UTILIZA MATEMÁTICAS, ESTADÍSTICA Y LÓGICA FORMAL PARA (a) DISEÑAR O SELECCIONAR PROCEDIMIENTOS EXPERIMENTALES OPTIMIZADOS, (b) PROPORCIONAR INFOMACIÓN QUÍMICA RELEVANTE ANALIZANDO DATOS QUÍMICOS, Y (c) OBTENER CONOCIMIENTOS SOBRE SISTEMAS QUÍMICOS (MASSART)

CIENCIA QUE RELACIONA MEDIDAS HECHAS SOBRE UN SISTEMA O PROCESO QUÍMICO, CON EL ESTADO DEL SISTEMA MEDIANTE LA APLICACIÓN DEMÉTODOS ESTADÍSTICOS O MATEMÁTICOS (International Chemometrics Society)

EL ARCO DEL CONOCIMIENTO

CONOCIMIENTO

Creatividad Inteligencia química

DEDUCCIÓN HIPÓTESIS INFORMACIÓN INDUCCIÓN

(SÍNTESIS) (ANÁLISIS)

Diseño Datos

EXPERIMENTOS

QUIMIOMETRÍA Y CALIDAD

No siempre está implicada en la obtención de conocimiento

Muchas técnicas se utilizan para mejorar procesos y/o productos y para controlar la calidad

Hoy día se trata de igualar CHEMOMETRICS Y QUALIMETRICS

Las herramientas quimiométricas son fundamentales en la espiral de la calidad y en las políticas de CALIDAD y de QUALITY ASSURANCE

Relación con el entorno

PROBLEMA PROBLEMA PROCESO

ECONÓMICO ANALÍTICO ANALÍTICO

SOCIAL

Etapas generales para el planteamiento y resolución de un problema analítico

Si

No

Obtención de información Regresión, Métodos multivariantes

ERRORES EN QUÍMICA ANALÍTICA

1 Introducción

- **Poblaciones y muestras**
- **Variables**
- **Histogramas y distribuciones**

    - **Estadística descriptiva**
    - **Promedio y medidas de centralización**
    - **Medidas de dispersión**
    - **Medidas de la forma de la distribución**
        - ♦ **Medida de la calidad**
        - ♦ **Calidad y errores**
        - ♦ **Errores sistemáticos y aleatorios**
            - ◊ **Precisión y bias de las medidas**
            - ◊ **Otros tipos de error**
            - ◊ **Propagación de errores**
            - ◊ **Distribución normal o gaussiana**
            - ◊ **Propiedades de la distribución normal**
            - ◊ **Distribución normal estandarizada**
            - ◊ **Tablas de distribución normal estandarizada**
            - ◊ **Funciones EXCEL**
                - · **Teorema del límite central y distribución de medias muestrales**
                - · **Enunciado**
                - · **Intervalos de confianza de la media**
                - · **Muestras pequeñas y distribución t**
                - · **Funciones EXCEL**
                - · **Pruebas de normalidad**
                    - **Otras distribuciones**
                    - **Distribución binomial**
                    - **Distribución de Poisson**
                    - **Distribución o de Pearson**
                    - **Distribución t de Student**
                    - **Distribución F de Fischer**

10 Bibliografía

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:

Ejemplo de histograma

Resultados en mg.l−1 de 60 determinaciones del contenido en calcio del agua del Canal de Castilla

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 límites 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.

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 (Véase Figura adjunta).

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 21 y 28 mg/l es del 23,33 % y la de encontrar valores de hasta 28 mg/l es del 38,33 %. 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) $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

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

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

Media ponderada

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

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 Q1, Q2 y Q3. La mediana coincide con Q2

Deciles: Valores D1 a D9 que dividen los datos en 10 partes iguales

Percentiles: Valores P1 a P99 que dividen los datos en 100 partes iguales

Cálculo de cuartiles

Una vez calculada la mediana que coincide con Q2, 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 Q1, Q2 y Q3 Si n es impar, la mediana Q2 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 Q1, Q2 y Q3 son números enteros los datos que ocupen las posiciones Q1, Q2 y Q3 definen los cuartiles. Si se obtiene números decimales, Q1, Q2 y Q3, se calculan por interpolación lineal. En el ejemplo del contenido en calcio del agua del Canal de Castilla, Q1 = 24,51; Q2 = 32,69 y Q3 = 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: Q3 − Q1

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 Q1, Q2 (la mediana) y Q3 También se incluye un intervalo que engloba a los puntos más extremos comprendidos dentro de 1,5(Q3−Q1). Véase el ejemplo de boxplot. Lo ideal es una caja simétrica en torno a la mediana y de bigotes idénticos.

Desviación típica relativa:

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:

(j=1,...k)

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

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

Ejemplo de boxplot

Sean los datos siguientes (ya ordenados)

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).

Medidas de la forma de la distribución

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

a3 > 0. Sesgo positivo. La distribución tiene cola hacia la derecha. Media >Mediana

a3= 0 : La distribución es simétrica. Media = Mediana

a3 < 0. Sesgo negativo. La distribución tiene cola hacia la izquierda. Media > Mediana

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

a4−3 > 0: La curva es apuntada (leptocúrtica)

a4−3 = 0: La curva es normal

a4−3 < 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**	
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, xi, diferirá del valor verdadero , y la diferencia es el error:

Si se hacen más medidas, es decir se analiza una muestra de una población, se obtendrá la media, , que estima , o la media de la población de medidas. Si la muestra es lo bastante grande, entonces .

La primera parte es el error aleatorio, y la segunda parte o el error sistemático. **Los errores sistemáticos conducen a inexactitud (inaccuracy) y bias, mientras que los errores aleatorios conducen a la imprecisión**.

La exactitud (accuracy) de la media es definida por **ISO** como: *"cercanía entre el valor verdadero y el valor medio obtenido aplicando el procedimiento experimental un número elevado de veces. Cuanto menor sea la parte sistemática de los errores experimentales que afectan a los resultados, más exacto (accurate) es el procedimiento"*. Por su parte, un documento IUPAC define **bias** como: *"una medida de la exactitud (accuracy) o inexactitud (inaccuracy) de la media"*, y como *"la diferencia entre la media de la población y el valor verdadero teniendo en cuenta el signo"*. Es decir: **la exactitud es el concepto y el bias su medida**. Además hay dos tipos de bias: bias del laboratorio y bias del método.

La **precisión** se define por **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"*.

En Química Analítica se usa como medida de la precisión la desviación típica, o la desviación típica relativa. Debemos recordar que s estima , y que cuando el número n de medidas replicadas es bastante grande, se puede cosiderar que s= . El valor de n varía, y según la IUPAC basta que n"10, pero en muchos libros de estad´sitica se dice que n"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)** si las condiciones son más adversas (o heterogéneas): diferente analista, o día....

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 obliga a trabajar con la ayuda de otros laboratorios (Véase validación de métodos). En esencia, un laboratorio j trabajando en solitario únicamente puede evaluar la repetitividad y no puede diferenciar entre bias del método y bias del laboratorio, pues solo tiene acceso a la media de sus determinaciones que estima j.y a sus resultados individuales

Repetitividad

bias (método + laboratorio)

o de forma gráfica

Si se une a otros laboratorios, podrá diferenciar entre los bias del método y los bias de su laboratorio y podrá diferenciar entre

reproducibilidad y repetitividad. En este caso además de la media de sus determinaciones j, tiene acceso a la media de las medias de los diferentes laboratorios, , que estimará mucho mejor

Reproducibilidad

Repetitividad bias bias

laboratorio método

o de forma gráfica:

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:

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 y es el eror sistemático que afecta a y, para una ecuación de adición y/o sustracción se cumple:

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

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

Si la relación es multiplicativa

$y = axz/t$

$y/y = x/x + z/z - t/t$

Es decir los errores sistemáticos pueden compensarse entre sí.

(para otro tipo de relaciones véase el Resumen de fórmulas de propagación de errores)

Resumen de fórmulas de propagación de errores

Función	**Error aleatorio**	**Error sistemático**
Combinaciones lineales		
Expresiones multiplicativas		
Potencias		
Función de una variable independiente	**ó**	
Función de varias variables independientes		

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 ves 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:

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

La **distribución de probabilidad normal acumulativa** es

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

siendo la media de la población y su desviación típica

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

Los parámetros que describen una distribución normal son y : **N(,)**, y su efecto sobre la función de densidad de probabilidad es:

Propiedades de la distribución normal

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

Distribución normal tipificada o estandarizada

Tanto como 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−)/** , con lo que se obtiene una nueva distribución **N(0,1)**. Sus f(x) y F(x) son respectivamente:

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

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

Por tanto si se tiene una variable x N(,), para un valor dado x1, si se calcula z1 = (x1 −)/ , se cumple que:

$P(X''x1) = \int_{-\infty}^{x_1} f(x)\,dx$

y $P(Z''z1) = \int_{-\infty}^{z_1} f(z)\,dz$

son idénticas

x1 0 z1

Tablas para la distribución estandarizada

La transformación z permite que en vez de hacer infinitos cálculos y gráficas para las infinitas combinaciones de y , 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 a z

DISTR.NORMAL.ESTAND.INV(p) devuelve el valor de z tal que la probabilidad desde

a z valga p

Ejemplo

Se sabe que los matraces aforados suministrados por una cierta compañía tienen una distribución normal, con una media de 100,00 ml y una desviación típica de 0,25 ml

- ¿Qué proporción de matraces estará por encima de 100,33 ml?
- ¿Qué proporción de matraces tendrá un contenido inferior a 99,55 ml?
- ¿Qué proporción de matraces estará comprendida entre 99,77 y 100,20 ml?
- Si se desea que sólo un 5% de matraces esté por encima del volumen especificado ¿cuál debería ser éste?

TEOREMA DEL LÍMITE CENTRAL Y DISTRIBUCIÓN DE MEDIAS MUESTRALES

Enunciado

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

yi = x1 + x2 + ... + xn

con medias i y varianzas i2, para grandes valores de n, la distribución de y es **aproximadamente normal** con una media i y varianza i2.

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 y varianza se toman todas las posibles muestras de tamaño n, la distribución de las medias muestrales tendrá una media y una varianza /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 ± 1,96 .

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

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

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

Muestras pequeñas y la distribución t

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

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

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

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

siendo 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 = 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 = = 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 = = 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)

- 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 = = 0,05
- Si lo que se desea es de dos colas, hay que entrar por la columna encabezada por (o se sustituye en la función) p = /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)

- 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 = = 0,05
- Si lo que se desea es de una cola, hay que entrar por la columna (o se sustituye en la función) p = 2 = 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 , 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. La frecuencia absoluta de cada uno de los valores es la unidad. La frecuencia acumulativa se obtiene sumando uno a cada frecuencia absoluta anterior, y el porcentaje de frecuencia relativa correspondiente es:

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

Si la distribución es normal, el % de frecuencia relativa acumulada coincide con la probabilidad de que un número aparezca y esa probabilidad se puede convertir en valores z. Es decir, p.e. el valor con una frecuencia relativa acumulada del 7,7 % es equivalente al valor que delimita la **cola inferior** de una distribución normal con una cola del 7,7%, utilizando una tabla de z (tal como la 3,1; 3,2 ó 3.3). Se obtiene una tabla de valores ordenados (ranked) de z, que se denomina rankits.

La representación de los valores de la medida original (x) frente a los de z, debe dar una línea recta

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

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)

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).

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) =

P($\overline{A}$
) = 1 − P(A) = 1−

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.

$$P(2i) = C_{10}^2 \ 0{,}02^{\,2} (1 - 0{,}02)^{10-2} = \frac{10!}{2!(10-2)!} 0{,}02^{\,2i} (0{,}98)^8 = 0{,}0153 \quad = 1{,}5$$

La media y la varianza de una distribución binomial son = n y = n (1−). (En el ejemplo = y =)

La distribución binomial se aplica a:

- Muestreo sin reemplazamiento de una población muy grande comparada con el tamaño de la muestra
- Muestreo con reemplazamiento de una población finita
- Muestreo de procesos continuos de fabricación con población grande

Cuando n y n(1−) 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 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: =
=

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

Si > 10, la distribución de Poisson tiende a la normalidad

Distribución o de Pearson

Si un conjunto de variables independientes z1, z2, ..., zn 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 =n grados de libertad

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

es la función gamma

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

La función 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: = y = 2

Habitualmente las variables independientes iniciales no están normalizadas. Si tenemos n variables independientes x1, x2, ..., xn con distribución normal N(,) podemos obtener una nueva variable independiente

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

que estará distribuida como con = n−1 grados de libertad.

La distribución se utiliza en el **test** 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 otra variable aleatoria independiente de la anterior, con una distribución . Se puede definir una nueva variable t como

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

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

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

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

Distribución F de Fischer

1) Sean x e y dos variables aleatorias independientes que siguen distribuciones con y 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

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

La media y la varianza son:

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

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

2) Sea x1, x2,...xn una muestra aleatoria de variables aleatorias independientes con distribución N(x, x2) e y1, y2,...yn una muestra aleatoria de variables aleatorias independientes con distribución N(y, y2)

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 con x =nx−1 y y = ny−1 grados de libertad respectivamente. La variable F es en este caso es

$$F = \frac{\chi_x^2 / \nu_x}{\chi_y^2 / \nu_y} = \frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2}$$

con una distribución F con nx−1 y ny−1 grados de libertad

Si las varianzas de ambas poblaciones son iguales:

$$F = \frac{s_x^2}{s_y^2}$$

Si este cociente es mucho menor que la unidad, la hipótesis anterior será falsa, ya que si x2 y y2 son idénticas, sus estimadores deberían ser muy parecidos, por lo que F debería ser muy próximo a 1.

El valor de la variable F está tabulado para diferentes probabilidades y diferentes grados de libertad.

Se utiliza para comparar la precisión, es decir las varianzas de dos procedimientos o de dos medias.

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

PRECISIÓN Y VERACIDAD. EVALUACIÓN DE DATOS ANALÍTICOS

PRECISIÓN Y VERACIDAD. EVALUACIÓN DE DATOS ANALÍTICOS

INTRODUCCIÓN

Supongamos una situación en la que se está determinando el contenido en carbonato cálcico de una muestra de un material de referencia certificado que contiene 50,00 mg de CaCO3.

- Se llevan a cabo 4 determinaciones (n=4). La media $\overline{x}$ = 49,30 y se sabe que la vale 0,82 (estimada previamente)
- Se hacen 6 réplicas (n=6), tales que x

= 49,10 y s = 0,80

Deseamos conocer si las medias obtenidas por nosotros son significativamente diferentes del valor verdadero o teórico esperado. De esta manera podremos determinar si nuestras medidas llevan asociados un componente de error sistemático (bias) además del error aleatorio inevitable.

(Recordemos que xi − = (xi − $\bar{x}$) +($\bar{x}$ −) o bien xi − = (xi −) +(−))

Para ello se lleva a cabo un test de hipótesis o prueba de significación, que trata de encontrar si hay diferencias estadísticamente significativas entre (estimado por $\bar{x}$) y (lo cual no significa necesariamente que esas diferencias sean relevantes desde el punto de vista químico)

HIPÓTESIS NULA Y ALTERNATIVA

La **hipótesis nula** es que no hay diferencias entre y , lo que se escribe:

H0: =

Para el caso en que H0 no sea cierta necesitamos una **hipótesis alternativa**

H1: " . (También puede plantearse > o <)

PLANTEAMIENTO DE LAS PRUEBAS

Con un nivel de confianza del 95%, es decir con un **nivel de significación = 0,05**, veamos los dos ejemplos planteados:

- Puesto que $\bar{x}$ es un estimador de , sabemos que $\mu = \bar{x} \quad 1{,}96\,\sigma / \sqrt{n}$, luego = 49,30 ± 1,96.0,82/ $\sqrt{4}$ = 49,30 ± 0,80, es decir 48,50 " " 50,10. Como ese intervalo incluye a = 50,00, podemos concluir que = , y por tanto **aceptamos la hipótesis nula**. Esto no significa que hayamos probado que es cierta, sino solo que **no tenemos evidencias para rechazarla**.
- En este caso no se conoce por lo que debe utilizarse su estimador s, y ya que el número de datos es pequeño, debemos utilizar $\mu = x \quad t_{\alpha/2,(n-1)}\,(s / \sqrt{n})$ luego = 49,10 ± 2,57.0,80/ $\sqrt{6}$ = 49,10 ± 0,84, es decir 48,26 " " 49,94. Como ese intervalo no incluye a = 50,00, podemos **rechazar la hipótesis nula**, y en su lugar **aceptaremos la hipótesis alternativa** y concluiremos que " .

Las etapas del test o prueba de significación tal como lo hemos realizado son

- Plantear H0 y H1
- Elegir el nivel de significación
- Calcular el intervalo de confianza alrededor de $\bar{x}$ a un nivel de 100− %= 95%
- Investigar si está dentro de ese intervalo
- Si le respuesta es positiva, se acepta H0. Si es negativa se rechaza H0 (y se acepta H1)

Comparación con valores críticos

Hay otra forma de hacer el test, que parece diferente pero que es exactamente la misma.

- Sabemos que el 95% de todas las medias muestrales cae dentro del intervalo $\mu \quad 1{,}96\,\sigma / \sqrt{n}$. Si suponemos que H0 es cierta, = , también caerán dentro de $\mu_0 \quad 1{,}96\,\sigma / \sqrt{n}$. En la Figura aparecen los intervalos de confianza alrededor de en unidades originales y en unidades z. Sabemos que el 95% de todas las medias compatibles con **H0:** = están dentro del intervalo comprendido entre z=−1,96 y z=+1,96, por lo que si expresamos la distancia de $\bar{x}$ a en unidades z, cuando **el valor absoluto de ese z experimental (zexp) sea menor que 1,96** $\bar{x}$ **estará dentro del intervalo, y la H0 se aceptará**. Si el valor absoluto de zexp es mayor que 1,96, entonces $\bar{x}$ estará fuera del intervalo, por lo que H0 se rechazará (y se aceptará H1).

  Por consiguiente 1,96 es un **valor crítico**, ya que representa la máxima distancia en unidades z a que puede estar situada una media $\bar{x}$ del valor esperado , para poder ser considerada como perteneciente a la población de con un 95% de confianza.

  En este caso el valor experimental de |z|, calculado a partir de $|z| = \dfrac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}$, vale , y por tanto está más cerca que 1,96, o lo que es igual |zexp|<zcrit por lo que la H0 se mantiene.

- En este segundo ejemplo los pasos son exactamente los mismos pero en unidades t, por lo que el valor crítico de t se determina en la tabla correspondiente y vale 2,57.

  El valor experimental de |t| se calcula mediante $|t| = \dfrac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$

y vale , que está más lejos que 2,57 por lo que se encuentra fuera del intervalo de confianza del 95% y la **H0 puede rechazarse, aceptándose la H1**.

En esta otra modalidad las **etapas del test** son:

- Plantear H0 y H1
- Elegir el nivel de significación
- Determinar el valor crítico del estadístico correspondiente
- Calcular el valor experimental del estadístico
- Comparar ambos valores
- Si |valorexp| < valorcrit se acepta H0. En caso contrario se rechaza H0(y se acepta H1)
-

Evidentemente, ambas pruebas son la misma y **deben originar el mismo resultado.**

En ocasiones, se calcula un **nivel de significación a posteriori**, que se denomina **p** para diferenciarlo del nivel de significación a priori. Para ello se calcula el valor de p para el cual se obtiene un valor de z o de t igual al |valor|exp. Para el Ejemplo 1, p= 0,089 (Función de EXCEL (1−DISTR.NORM.ESTAND(1,70))*2). Para el Ejemplo 2, p=0,040 (Función de EXCEL DISTR.T(2,76;5;2)). **El valor de p representa el nivel de significación hasta el que se puede mantener la H0**. Siempre que **p > la H0 se mantiene**, y si p < la H0 se rechaza.

Ejemplo 1 Ejemplo 2

ERRORES DE TIPO I Y II

Error de tipo I o error _

Es la **probabilidad de rechazar la H0 cuando es cierta**. Es decir es la probabilidad de decir que $\bar{x}$
no pertenece a la población de , cuando sí que pertenece a ella. **El error de tipo I coincide con el nivel de significación a priori** .

Así para un nivel = 0,05 (un nivel de confianza del 95 %), cuando examinemos muchas medias determinadas por nosotros, el 5% de las mismas se rechazará a pesar de pertenecer a la población de .

Error de tipo II o _

Sea la situación del ejemplo número 1 y supongamos que la H0 es cierta, es decir que = =50,00. Los valores de $\bar{x}$
que deberíamos obtener deberían estar (con un 95% de probabilidad) dentro del intervalo 50,00 ± 1,96.0,82/$\sqrt{4}$
= 50,00 ± 0,80, es decir 49,20 " " 50,80. Cualquier valor de $\bar{x}$
que obtengamos y que esté fuera de ese intervalo hará que rechacemos H0 y aceptemos H1. Si obtenemos $\bar{x}$
= 48,80 rechazaremos H0.

Supongamos que el método que utilizamos tiene un **bias de −1,20, desconocido para nosotros**, de manera que la media de la población de las determinaciones es en realidad = 50,00−1,20 = 48,80. Como **no** lo sabemos, si encontramos un valor de $\bar{x}$ mayor de 49,2 aceptaremos H0, cuando de hecho eso no es correcto ya que hay un bias. Hemos cometido un **error de tipo II o error .** que consiste en **aceptar la H0 cuando es falsa** (o de rechazar H1 cuando es cierta). El bias no habría sido detectado.

Por tanto P(I) = y P(II) = . Sabiendo que hay un bias de −1,20 y por tanto que = 48,80, para calcular nos debemos preguntar por la probabilidad de encontrar un valor mayor de 49,20 para la distribución centrada en . El y la fracción de medidas con z > 0,976 es 0,165 (EXCEL =1−DISTR.NORM.ESTAND(0,976)) o sea del 16,5 %. Ese es el valor de .

Ambos tipos de error están relacionados entre sí: Una disminución de origina un aumento correspondiente de . Por ejemplo, si = 0,01, cuando no haya bias el intervalo será ahora 50,00 ± 2,57.0,82/$\sqrt{4}$ = 50,00 ± 1,05, es decir 48,95 " " 51,05. Cuando hay un bias de −1,20 el valor La fracción de medidas con z > 0,366 es 0,357 (EXCEL =1−DISTR.NORM.ESTAND(0,366)) o sea del 35,7 %. Por tanto, se ha producido un aumento de .

Por eso se deben tener muy buenas razones para que sea inferior a 0,05.

Para un dado, cuanto mayor es el bias o diferencia entre y , menor es . Por el contrario, cuanto menor es el bias, más grande se hace . Esto es lógico ya que cuanto más pequeño es un bias, más difícil es su detección.

Cuanto mayor es n, menores son y , ya que las distribuciones de medias muestrales se hacen más estrechas al disminuir la desviación típica que es $\sigma / \sqrt{n}$

.

Se denomina **potencia o poder de una prueba a la probabilidad de rechazar correctamente H0 cuando es falsa**. Puesto que es la probabilidad de aceptar H0 en esas condiciones, la potencia o poder de un test es **1 − .**

En resumen:

Se acepta H0 No hay error. P=1 −

CIERTA

Se rechaza H0 Hay error de tipo I. P(I) =

H0

Se acepta H0 Hay error de tipo II. P(II) =

FALSA

Se rechaza H0 No hay error. P=1 −

Ejemplo: Un fabricante de gamusinos certifica que su contenido medio en luciferina es de 0,300 g con una desviación típica de 0,020 g

- Se analizan 20 gamusinos y se encuentra un contenido medio de 0,285 g. Evalúe la certeza de la afirmación del fabricante al 95% y al 99%
- Determine la probabilidad de cometer un error del tipo I en ambos casos
- Si se demuestra que nuestro método analítico tiene un bias de manera que la media de nuestras medias muestras es 0,290 g, determine la probabilidad de error de tipo II y la potencia del test para tamaños de muestra n=9 y n=27, con un nivel de significación del 5%.

Importancia del tamaño de muestra

El intervalo de confianza de cualquier media puede disminuirse simplemente aumentando el tamaño de la muestra.

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

Así aumentando cuatro veces el tamaño, el intervalo se reduce a la mitad, y esto es manipulable.

Hasta ahora hemos visto como se puede calcular el valor una vez conocidos todos los demás valores: , n, , y el bias | − |. Otra posibilidad más interesante es el poder calcular el **tamaño n de la muestra** que debe ser analizado para llevar a cabo nuestra prueba con la **suficiente confianza**, es decir para poder detectar una diferencia (bias) = | − | mediante un procedimiento que tiene una precisión (o s), de manera que haya no más de un % de probabilidades de decidir que hay diferencia (bias) cuando no la hay, y al mismo tiempo una probabilidad de no más que % de no detectar la diferencia (bias) cuando sí que existe.

Ya que la distancia | − | en unidades z es z /2+z , al transformarla en las unidades originales, resulta al reordenar:

$$n \quad \left[\left(z_{\alpha/2} + z_{\beta}\right)\sigma/\delta\right]^2$$

También pueden utilizarse gráficos publicados por la ISO.

TEST DE UNA O DOS COLAS

En el contexto de los ejemplos anteriores, tan malo era que el resultado del análisis fuera mayor (>) o menor (<) que el

valor certificado. Por ello la hipótesis alternativa era siempre " .
El test se llama de **dos colas pues se exploran ambos lados** de la distribución.

Pero hay ocasiones en que estamos solo interesados en saber si el valor hallado es mayor o menor que la referencia. Por ejemplo al comprobar la calidad de una materia prima, o al estudiar la estabilidad de un compuesto. El test es de **una cola pues solo se explora un lado** de una distribución.

En resumen:

H0: = Siempre

H1: " Test de dos colas

> Test de una cola

< Test de una cola

Una cola

Una cola

Dos colas

Prueba , para ver si una distribución es normal

Se aplica al ejemplo de la Lección 2: Resultados del calcio de 60 muestras de agua del Canal de Castilla. Se utiliza la salida del histograma

H0: La muestra está extraída de una población Normal

H1: La muestra está extraída de una población no−normal

- Se calculan la media y la desviación típica
- Se calcula el valor z del límite superior de cada clase
- Se calcula la Frecuencia relativa acumulativa esperada para cada z (Tabla 3.3 o EXCEL =DISTR.NORM.ESTAND(z))
- Se calculan las Frecuencias relativas esperadas por diferencia entre cada dos acumulativas
- Se calcula la Frecuencia esperada (Ei) multiplicando lo anterior por el nº de datos
- Se ve la Frecuencia observada (Oi) a partir de los datos del histograma en cada clase
- Se calculan los (Oi−Ei)2/Ei para cada clase y se suman: SUMA
- Se compara esa suma con el , crítico con el nivel de significación requerido (= 0,05) y nº de g.d.l. = k −3 siendo k el nº de clases utilizadas.

(Tabla, o en EXCEL PRUEBA.CHI.INV(0,05;k−3))

- Si SUMA < , Crítico, se mantiene H0

  Si SUMA > , Crítico, se rechaza H0 y se acepta H1

Clase	*Frecuencia*	*% acumulado*
10,708929	1	1,67%
17,8696342	4	8,33%
25,0303394	11	26,67%
32,1910447	12	46,67%
39,3517499	14	70,00%
46,5124551	9	85,00%
53,6731603	5	93,33%
y mayor...	4	100,00%

Resultados del Ejemplo

Inferior	*Superior*	*Frec (O)*	*z superior*	*Frec. Relat. acu espe*	*Frec. Relat. espe*	*Frec. es (E)*
7,16070522	10,708929	1	−1,88198267	0,02991912	0,02991912	1,79514
10,708929	17,8696342	4	−1,28987047	0,09854788	0,06862876	4,11772
17,8696342	25,0303394	11	−0,69775828	0,24266411	0,14411623	8,64697
25,0303394	32,1910447	12	−0,10564608	0,45793153	0,21526741	12,9160
32,1910447	39,3517499	14	0,48646612	0,68668164	0,22875011	13,7250
39,3517499	46,5124551	9	1,07857832	0,85961209	0,17293045	10,3758
46,5124551	53,6731603	5	1,67069051	0,95260861	0,09299652	5,57979
53,6731603	60,8338655	4	2,26280271	0,98817611	0,0355675	2,13404
Media	33,4686749					SUMA
desv	12,0934939					, Crítico
n	60					
						Normal

PRUEBAS DE SIGNIFICACION ESTADISTICA (TESTS DE HIPOTESIS)

PRUEBAS PARA EVALUAR LA EXACTITUD (DETECCION DE ERR SISTEMATICO)		
Comparación de una media experimental con un valor conocido		
H0: $\bar{x} = \mu_o$ H1: $\bar{x} \quad \mu_o$ (test de 2 colas)		
conocida	$z = \frac{\lvert \bar{x} - \mu_o \rvert}{\sigma / \sqrt{n}}$	$z_{crit\ 1-\frac{\alpha}{2}}$

desconocida n>30	$z = \frac{\lvert \bar{x} - \mu_o \rvert}{s/\sqrt{n}}$	$z_{crit\ 1-\frac{\alpha}{2}}$	
desconocida n<30	$t = \frac{\lvert \bar{x} - \mu_o \rvert}{s/\sqrt{n}}$	$t_{crit\ 1-\frac{\alpha}{2},\nu}$	$\nu = n-1$
Comparación de dos medias			
H0: $\bar{x}_1 = \bar{x}_2$ (ó $\bar{x}_1 - \bar{x}_2 = 0$) H1: $\bar{x}_1 \quad \bar{x}_2$ (ó $\bar{x}_1 - \bar{x}_2 \quad 0$) (test de 2 colas)			
, conocidas	$z = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$z_{crit\ 1-\frac{\alpha}{2}}$	
, desconocidas n1,n2>30	$z = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$z_{crit\ 1-\frac{\alpha}{2}}$	
, desconocidas s1,s2 comparables n1,n2<30	$t = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	$t_{crit\ 1-\frac{\alpha}{2},\nu}$	$\nu = (n_1 - 1) + (n_2 - 1) = n_1 + n_2$ $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$
, desconocidas s1,s2 no comparables n1,n2<30	$t = \frac{\lvert \bar{x}_1 - \bar{x}_2 \rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$t_{crit\ 1-\frac{\alpha}{2},\nu}$	$\frac{1}{\nu} = \frac{1}{(n_1 - 1)} \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} + \frac{1}{(n_2}$
		Test de Cochran: y	

Comparación de pares de valores (medidas pareadas)		
$d_i = x_{1i} - x_{2i}$ H0: $\mu_d = 0$ H1: $\mu_d \quad 0$ (test de 2 colas)		
n>30	$z = \frac{\lvert \bar{d} - \mu_d \rvert}{s_d/\sqrt{n}} = \frac{\lvert \bar{d} \rvert}{s_d/\sqrt{n}}$	$z_{crit\ 1-\frac{\alpha}{2}}$

n<30	$t = \frac{\lvert \bar{d} - \mu_d \rvert}{s_d/\sqrt{n}} = \frac{\lvert \bar{d} \rvert}{s_d/\sqrt{n}}$	$t_{crit\ 1-\frac{\alpha}{2},\nu}$	**=n−1**

Comparación de pares de valores cuando s depende de C y ésta varía en un amplio rango

Comparación de dos métodos por regresión lineal

$x_{2i} = a + bx_{1i}$

H0: $a = 0$
, $b = 1$

H1: $a \quad 0$
$b \quad 1$

(test de 2 colas)

$t = \frac{\lvert a - 0 \rvert}{s_a}$	$t_{crit\ 1-\frac{\alpha}{2},\nu}$	=n−2

PRUEBAS PARA EVALUAR LA PRECISION

Comparación de dos varianzas: Prueba F de Fischer

H0: $\sigma_1^2 = \sigma_2^2$

H1: $\sigma_1^2 > \sigma_2^2$

(test de 1 cola)

$s_1^2 \quad s_2^2$			=n1−1 =n2−1

PRUEBAS DE RECHAZO DE VALORES DISCREPANTES (OUTLIERS)

Prueba Q de Dixon

H0: el valor sospechoso " N(,) H1: el valor sospechoso " N(,) (test de 2 colas)

$Q = \frac{\lvert x_{scspechoso} - x_{pr\ ximo} \rvert}{x_{mayor} - x_{menor}}$	$Q_{crit\ 1-\frac{\alpha}{2},\nu}$	=n

TESTS DE NORMALIDAD

Prueba chi−cuadrado ( )

H0: X " N(,) H1: X " N(,) (test de 2 colas)

$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$	$\chi^2_{crit(1-\alpha,\nu)}$	$\nu = k - 1 - h$ k, número de clases

		h, número de parámetros de la distribución normal (2: y)
Calcular la media y la desviación estándar de los datos experimentales Dividir los datos experimentales en k clases, siendo k"n/5 Calcular las frecuencias Oi observadas en cada clase (Oi=n). Transformar los valores límite superiores (LS) de cada clase en valores z: Calcular la frecuencia acumulada para cada valor de z, y la frecuencia relativa de cada clase. Calcular las frecuencias esperadas, Ei: $E_i = nf_i$, siendo fi la frecuencia relativa de cada clase		

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

EVALUACIÓN DE FUENTES DE VARIACIÓN DE DATOS. ANOVA

- **Introducción**
- **Análisis de varianza de una vía**

- Fuentes de varianza y significación
- La tabla del ANOVA
- Modelo de efectos fijos y efectos aleatorios
- Suposiciones implícitas

- **ANOVA de dos vías y multivía**
- Fuentes de varianza y significación
- La tabla del ANOVA
- Interacción y su estimación

4 **Bibliografía**

EVALUACIÓN DE FUENTES DE VARIACIÓN DE LOS DATOS. ANOVA

INTRODUCCIÓN

Anteriormente se han discutido test de hipótesis para comparación de medias. A veces se comparan más de dos

medias, como se puede apreciar en los datos de la tabla que muestra los datos resultantes de la determinación de Cu en suelos de una misma muestra mediante 7 procedimientos de mineralización.

			MÉTODO				
	1	**2**	**3**	**4**	**5**	**6**	**7**
	5,59	5,67	5,75	4,74	5,52	5,52	5,43
	5,59	5,67	5,47	4,45	5,47	5,62	5,52
	5,37	5,55	5,43	4,65	5,66	5,47	5,43
	5,54	5,57	5,45	4,94	5,52	5,18	5,43
	5,37	5,43	5,24	4,95	5,62	5,43	5,52
	5,42	5,57	5,47	5,06	5,76	5,33	5,52
Media	5,48	5,57	5,47	4,80	5,59	5,43	5,48
Desviación	0,11	0,093	0,16	0,23	0,11	0,15	0,05

En vez de comparar las medias de cada columna dos a dos, podemos plantear una cuestión más general: El factor que hace que las columnas difieran, ¿tiene algún efecto sobre las medias de esas columnas?. Dicho de otra manera: ¿todos los procedimientos de mineralización originan el mismo resultado?.

Si no fuera el caso, la varianza total de la tabla dependería exclusivamente de la precisión de los procedimientos, por lo que cada elemento de la tabla sería

$$x_{ij} = \mu + e_{ij}$$

siendo el valor verdadero, eij errores aleatorios de media cero y varianza e2 = .

Si los procedimientos de mineralización tienen efecto podemos escribir:

$$x_{ij} = \mu + a_j + e_{ij} \quad ; \qquad a_{ij} = 0$$

siendo aj el efecto del pretratamiento j sobre la media global. El término aj introduce una varianza adicional en los datos de manera que ésta será mayor de e2.

Hay otras ocasiones similares, p.e. puede tratarse de los resultados de un estudio interlaboratorios, en los que se reparte una misma muestra homogénea a 7 laboratorios diferentes y se les pide que la analicen 6 veces mediante el mismo procedimiento. En este caso si hay algún efecto adicional deberá ser debido a que la muestra no es homogénea (o a que los laboratorios dan resultados diferentes).

		LABORATORIO					
	1	**2**	**3**	**4**	**5**	**6**	**7**
	5,59	5,67	5,75	4,74	5,52	5,52	5,43
	5,59	5,67	5,47	4,45	5,47	5,62	5,52
	5,37	5,55	5,43	4,65	5,66	5,47	5,43
	5,54	5,57	5,45	4,94	5,52	5,18	5,43
	5,37	5,43	5,24	4,95	5,62	5,43	5,52
	5,42	5,57	5,47	5,06	5,76	5,33	5,52
Media	5,48	5,57	5,47	4,80	5,59	5,43	5,48
Desviación	0,11	0,093	0,16	0,23	0,11	0,15	0,05

En este caso se lleva a cabo un ANOVA de una vía (hay un solo factor) a siete niveles.

ANÁLISIS DE VARIANZA DE UNA VÍA

Fuentes de varianza y significación

Pongamos la tabla de forma más general

			MUESTRA			
	1	**2**	**....**	**j**	**....**	**k**
	x11	x21	....	x1j	....	x1k
	x21	x22	....	x2j	....	x2k
	....	....	....	....	....	....
	xi1	xi2	....	xij	....	xik
	....	....	....	....	....	....
	xn11	xn22	....	xnjj	....	xnkk
Media	$\overline{x}_1$	$\overline{x}_2$	....	$\overline{x}_j$	....	$\overline{x}_k$
Varianza	s_1^2	s_2^2	....	s_j^2	....	s_k^2

Supongamos que el lote es **homogéneo** y que la **única fuente** de variación son las **incertidumbres** de las medidas. En ese caso la varianza podría ser estimada a partir de la primera columna:

$$s_1^2 = \sum_{i=1}^{n1} \left(x_{i1} - \overline{x}_1\right)^2 /(n_1 - 1)$$

o bien a partir de cualquiera de las k columnas. Si todos los datos proceden de la misma población de media y , todas las columnas deberían dar el mismo resultado:

$$\sigma_1^2 = \sigma_2^2 = ... = \sigma_j^2 = ... = \sigma_k^2 = \sigma^2$$

La varianza puede calcularse como una **varianza promediada (pooled) de las varianzas de las k columnas:**

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + \ldots + (n_k - 1)s_k^2}{n_1 + \ldots + n_k - k} = \sum_{j=1}^{k} (n_j - 1)s_j^2 \Big/ \sum_{j=1}^{k} (n_j - 1)$$

que será una estimación de mejor que cualquiera de las s_j^2 individuales. Por ahora consideraremos que todas las columnas tienen el mismo número de datos n1 = n2 =... = nk.

Otra posibilidad para estimar es a partir de la **varianza de las medias de las columnas** $s_{\bar{x}}^2$:

$$s_{\bar{x}}^2 = \sum_{j=1}^{k} (\bar{x}_j - \bar{\bar{x}})^2 / (k - 1)$$

siendo $\bar{\bar{x}}$ la gran media, que es también la media de las k medias columnares $\bar{x}_j$. Evidentemente, $s_{\bar{x}}^2$ estima $\sigma_{\bar{x}}^2$, y como hay nj datos en cada columna se cumplirá que $\sigma_{\bar{x}}^2$ = /nj (Teorema del Límite Central) o bien = nj $\sigma_{\bar{x}}^2$, luego es estimada mediante nj $s_{\bar{x}}^2$:

$$n_j s_{\bar{x}}^2 = n_j \sum_{j=1}^{k} (\bar{x}_j - \bar{\bar{x}})^2 / (k - 1)$$

Las dos estimaciones de serán iguales si el material es homogéneo.

Si el material es heterogéneo, s_p^2 no resulta afectada, ya que sus componentes se determinan dentro de cada columna: **varianza dentro de las columnas (within column).**

La varianza de las medias columnares $s_{\bar{x}}^2$ es la **varianza entre columnas (between−column).** Si el material es heterogéneo, no estima únicamente /nj sino que debe añadirse un **componente adicional** σ_a^2 que estima la varianza adicional debida a la heterogeneidad: $\sigma_{\bar{x}}^2 = \sigma^2 / n_j + \sigma_a^2$, por tanto nj$s_{\bar{x}}^2$

estima + nj σ_a^2

.

Todo esto nos permite formular una hipótesis y comprobarla.

1) Si el material es homogéneo s_p^2
y nj$s_{\bar{x}}^2$
estiman :

H0 : σ_p^2
= nj$\sigma_{\bar{x}}^2$
o H0 : σ_a^2
= 0

2) Si el material es heterogéneo nj$s_{\bar{x}}^2$
estima una varianza más grande que s_p^2

H1 : σ_p^2
< nj$\sigma_{\bar{x}}^2$
o H1 : σ_a^2
> 0

Estas varianzas pueden ser comparadas con una prueba F de una cola.

La tabla del ANOVA

Una forma de comprender mejor los cálculos es considerar el ANOVA como la **separación de la varianza total en sus componentes**. La varianza total es:

$$s_T^2 = \sum_{j=1}^{k} \sum_{i=1}^{nk} (x_{ij} - \bar{x})^2 / (n-1) \quad ; \quad n = \sum_{j=1}^{k} n_j$$

Por razones de facilidad de cálculo trabajaremos primero con las sumas de cuadrados SS

$$SS_T = \sum_{j} \sum_{i} (x_{ij} - \bar{x})^2$$

Como se cumple que

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

al elevar al cuadrado

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_j)^2 + (\bar{x}_j - \bar{x})^2 + 2(x_{ij} - \bar{x}_j)(\bar{x}_j - \bar{x})$$

Al sumar primero sobre las filas (i) y luego sobre las columnas (j) el último término se anula y el resultado es:

$$SS_T = \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 + \sum_j n_j(\bar{x}_j - \bar{x})^2$$

o lo que es igual

SST = SSR + SSA

SSR se llama **suma residual de cuadrados** o suma de cuadrados residuales y coincide con la SS **dentro** de las columnas. **SSA es la suma de cuadrados debida al efecto** del factor estudiado (la heterogeneidad entre muestras) y coincide con la SS entre columnas. A veces se llama SSTRATAMIENTO debido al origen agronómico del ANOVA.

Los **grados de libertad** de SST son (n−1) ya que se gasta uno en estimar $\bar{x}$
. Se reparten entre los de SSR y SSA. Para SSA se gastan (k−1) y para SSR quedan (n−1)−(k−1) = (n−k). Con ellos y las SS se calculan las medias cuadradas o cuadrados medios:

MSA = SSA/(k−1)

MSR = SSR/(n−k)

Las MS son estimaciones de las varianzas, así MSR es una estimación de , mientras que MSA es una estimación de
+ nj σ_a^2
(o de $\sigma^2 + \sigma_a^2 \frac{n - (\sum n_j^2 / n)}{k - 1}$
cuando los k valores nj no son iguales)

El **test de hipótesis** consiste en calcular
$F = \frac{MS_A}{MS_R} = \frac{SS_A/(k-1)}{SS_R/(n-k)}$
que se compara con el F tabulado de una cola con k−1 y n−k grados de libertad.

Los resultados se presentan en forma de tabla

Fuente	**g.d.l**	**SS**	**MS**	**F**
Entre columnas (A)	k−1	SSA	SSA/(k−1)	MSA/MSR
Dentro de columnas (residual)	n−k	SSR	SSR/(n−k)	
Total	n−1	SST		
Fcrit (0,05,k−1,n−k)=.....				

Para el ejemplo

Fuente	**g.d.l**	**SS**	**MS**	**F**
Entre columnas (A)	6	2,6834	0,4472	23,1529
Dentro de columnas (residual)	35	0,6761	0,0193	
Total	41	3,3595		
Fcrit (0,05,6,35)= 2,38				

Modelos de efectos fijos y efectos aleatorios

Hasta ahora hemos supuesto que los efectos son aleatorios. La diferencia no afecta ni a los experimentos reales ni al test F, pero los objetivos del ANOVA son diferentes.

Puesto que $x_{ij} = \mu + a_j + e_{ij}$
, cada resultado se descompone en varios componentes uno de los cuales es el efecto del factor. Hay dos posibilidades:

1) **Modelos de efectos fijos (fixed effect models) o ANOVA modelo I**

El efecto del factor hace desviarse de forma fija la media de cada grupo j de la gran media. Es el caso del ejemplo de la Tabla 1 en la que se estudia el efecto del pretratamiento, de manera que cada resultado consiste por un lado en +aj (a media + el efecto del pretratamiento) y por otro de los errores aleatorios o residuales ei.

En rigor MSA estima $\sigma^2 + \frac{\sum_j n_j a_j^2}{k-1}$
y la hipótesis a comprobar es:

H0 : a1 = a2 =....=ak = 0

H1 : aj " 0 para al menos un j

Esto no tiene consecuencias para los cálculos. En este modelo se rechaza H0, si al menos una columna tiene un valor medio diferente del testo, es decir es significativamente diferente de las otras (al menos un método de pretratamiento es diferente de los otros). En ocasiones se **desea saber qué columna es la diferente y cuál es la cuantía de la diferencia**. El método más simple por su sencillez es el de las diferencia menos significativa (LSD), y es parecido a las

pruebas t de comparación de medias. Para ello podría calcularse un valor de t, para cada pareja de columnas:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{MS_R\left[(1/n_1) + (1/n_2)\right]}} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{MS_R\left[2 / n_j\right]}} \quad \text{si} \quad n_1 = n_2$$

que se compararía con el t crítico.

La prueba se hace en la práctica chequeando cada

par $\left|\bar{x}_1 - \bar{x}_2\right|$

frente a LSD, calculado como

donde tcrit se calcula al nivel de significación que se desea y con un número de g.d.l. igual al que se utilizó para calcular MSR.

2) Modelo de efectos aleatorios (random effect models) ANOVA modelo II

Es el que se ha utilizado hasta ahora. No estamos interesados en un efecto específico debido a una cierta columna, sino en un efecto general sobre todas las columnas y que dicho efecto esté normalmente distribuido.

Las estimaciones son las indicadas anteriormente, así MSA estima $\sigma^2 + \sigma_a^2 \frac{n - (\quad n_j^2 / n)}{k - 1}$

o para iguales nj : + nj $\boldsymbol{\sigma}_a^2$

.

Puesto que el efecto es aleatorio no tiene sentido conocer que media columnar es significativamente diferente de las otras, pero si el efecto existe hay que conocer su **cuantía**. Cuando todos los nj son idénticos, MSA estimaba + nj $\boldsymbol{\sigma}_a^2$

, mientras MSR estima . Por tanto la varianza debida a la heterogeneidad de la muestra es:

$$s_a^2 = (MS_A - MS_R) / n_j$$

La diferencia entre ambos modelos no es siempre evidente: En el ejemplo de intercomparación podríamos estar interesados en comparar si todos los laboratorios trabajan bien (proficiency testing) en cuyo caso se trataría de un modelo de efectos fijos. Por otro lado, podemos suponer que todos los laboratorios trabajan bien, por lo que si el material es homogéneo la varianza dentro de las columnas describiría la repetitividad, la varianza global la repetibilidad y la varianza entre columnas el componente debido a la varianza entre laboratorios, En este caso es un modelo de efectos

aleatorios. **La diferencia es solo de tipo filosófico**, ya que los cálculos se hacen de la misma manera

Suposiciones implícitas

LA MSR se estima a partir de una varianza promediada, por tanto se supone que todas las columnas tienen varianzas homogéneas: **homocedasticiad**. A veces eso no es cierto.

Una forma de comprobarla es hacer una inspección visual de los datos antes del ANOVA. Un auxiliar muy potente son los boxplots, pero existen otras pruebas para comprobar que las varianzas son similares.

Si se comprueba la heterocedasticidad cabe varias opciones: Eliminar las columnas con varianza muy grande o transformando las variables.

ANOVA DE DOS VÍAS Y MULTI−VÍA

Fuentes de varianza y significación

En ocasiones se desean tener en cuenta **dos o más factores**. Por ejemplo, podemos estar estudiando el efecto que tiene sobre una la determinación de Cu en suelos varios procedimientos de mineralización de la muestra y simultáneamente varias formas de desecarla, por ejemplo al aire o en estufa. Los datos podrían ser:

		MÉTODO	
	1	2	3
	5,59	5,67	5,75
Seca al	5,59	5,67	5,47
Aire	5,37	5,55	5,43
	5,54	5,57	5,45
	5,37	5,43	5,24
	5,42	5,57	5,47
	4,74	5,52	5,52
Seca en	4,45	5,47	5,62
estufa	4,65	5,66	5,47
	4,94	5,52	5,18
	4,95	5,62	5,43
	5,06	5,76	5,33

De manera más general, la tabla se puede poner:

			FACTOR B				
FACTOR A	1	2	....	j	....	k	**Medias factor A**
1	x11	x21	....	x1j	....	x1k	$\bar{x}_{1.}$
2	x21	x22	....	x2j	....	x2k	$\bar{x}_{2.}$
....	....	....	....	....	....	....	....
h	xh1	xh2	....	xhj	....	xhk	$\bar{x}_{h.}$
...	....	....	....	....	....	....	
l	xl1	xl2	....	xij	....	xlk	$\bar{x}_{l.}$
Medias							**Gran media**
Factor B	$\bar{x}_{.1}$	$\bar{x}_{.2}$	....	$\bar{x}_{.j}$	....	$\bar{x}_{.k}$	$\bar{\bar{x}}$

Tablas de este estilo se llaman **tablas de dos vías (two−way)** y su análisis **ANOVA de dos vías (two−way ANOVA)** ya que los datos están sujetos a una doble clasificación. Cada intersección se denomina **celda** y pude haber uno o más datos en ella. Si hay **replicación** cada celda contienen más de un dato (en el ejemplo, cada celda contiene 6 datos).

La tabla del ANOVA

La tabla sigue un modelo lineal:

$$x_{hj} = \mu + a_h + b_j + e_{hj}$$

es decir que cada valor viene afectado por el efecto del factor a, el efecto del factor b y queda un residual que debería ser aleatorio.

Supongamos primero que no hay replicación es decir que solo hay un resultado en cada celda, como en la tabla general.

La gran media de los datos es:

$$\bar{\bar{x}} = \sum_h \sum_j x_{hj} / lk \quad ; \quad h = 1 \text{ a } l \; ; \; j = 1 \text{ a } k$$

Hay l niveles del factor A y la media de cada uno de estos niveles está dada por $\bar{x}_{1.}$
,$\bar{x}_{2.}$
,....$\bar{x}_{h.}$
,...$\bar{x}_{l.}$
, tal que

$$\bar{x}_{h.} = \sum_{j} x_{hj} / k$$

Similarmente hay k niveles del factor B y la media de cada nivel viene dada por

$$\bar{x}_{.j} = \sum_{h} x_{hj}$$

La SST se obtiene de forma similar a como antes y puede dividirse en componentes debidos a los diferentes factores y a los residuales

$$SS_T = \sum_{h} \sum_{j} (x_{hj} - \bar{x})^2 = SS_A + SS_B + SS_R$$

Los diferentes componentes, g.d.l. y MS son:

Para el factor A

$$SS_A = \sum_{h} \sum_{j} (\bar{x}_{h.} - \bar{x})^2 = k \sum_{h} (\bar{x}_{h.} - \bar{x})^2$$

; g.d.l. = l −1 ; MSA = SSA/(l − 1)

Para el factor B

$$SS_B = \sum_{h} \sum_{j} (\bar{x}_{.j} - \bar{x})^2 = l \sum_{j} (\bar{x}_{.j} - \bar{x})^2$$

; g.d.l. = k −1 ; MSB = SSB/(k − 1)

y para los residuales

SSR = SST − SSA − SSB ; g.d.l. = (kl − 1) − (k − 1) − (l − 1) ; MSR = SSR/g.d.l.

Puede demostrarse que:

$$SS_R = \sum_{h} \sum_{j} (x_{hj} - \bar{x}_{h.} - \bar{x}_{.j} + \bar{x})^2$$

; g.d.l. = (k −1)(l − 1) ; MSR = SSR/(k −1)(l − 1)

Se puede distinguir también aquí entre modelos de efectos fijos y aleatorios, y existe también un modelo mixto en el cual un factor es de efectos fijo y otro aleatorio.

Cuando se hacen réplicas, en cada celda existe más de un valor. En nuestro ejemplo numérico hay 3x2 celdas y cada una contiene además 6 réplicas. Aunque no es preciso que todas las celdas contengan el mismo número de réplicas, debería evitarse el que haya grandes diferencias. A partir de los valores anteriores se construye una tabla del ANOVA que es similar a la del ANOVA de una vía y que resulta ser:

Fuente	g.d.l	SS	MS	F
Efectos globales	(l −1)+(k −1)	SSA+SSB		
Factor A	l −1	SSA	SSA/(l−1)	MSA/MSR
Factor B	k−1	SSB	SSb/(k−1)	MSB/MSR
Residual	t−(k−1)−(l−1)=r	SSR	SSR/r	
Total	njkl−1=t	SST		

Se hacen pruebas F de 1 cola con los correspondientes g.d.l. de numerador y denominador para comprobar si los efectos de os factores A y B son estadísticamente significativos

En el caso del ejemplo

Fuente	g.d.l	SS	MS	F
Efectos globales	3	1,7501		
Factor A	1	0,5041	0,5041	10,34
Factor B	2	1,2460	0,6230	12,78
Residual	32	1,5601	0,0488	
Total	35	3,3102		

Los F críticos son: Factor A (1 y 32 g.d.l.) 4,15; Factor B (2 y 32 g.d.l.) 3,30. Luego ambos factores tienen efectos significativamente estadísticos sobre los resultados.

Interacción

En ocasiones el **efecto de uno de los factores depende del nivel del otro factor**. Esto es la interacción. La forma de considerar esto es añadir otro término adicional como fuente de varianza, o sea aumentar el modelo lineal subyacente:

$$x_{hij} = \mu + a_h + b_j + (ab)_{hj} + e_{hij}$$

El número de g.d.l. necesarios es el producto de (k −1)(l −1), y ese es el número de g.d.l. que queda para los residuales si no se hacen réplicas (ver más arriba), es decir que no hay g.d.l. para los residuales cuando no hay replicación. Por tanto, si se desea estudiar la interacción debe hacerse réplicas. En nuestro ejemplo concreto sí que la había, luego puede calcularse (No entraremos en detalles). La Tabla del ANOVA que resulta es:

Fuente	g.d.l	SS	MS	F
Efectos globales	(l −1)+(k −1)	SSA+SSB		
Factor A	l −1	SSA	SSA/(l−1)	MSA/MSR
Factor B	k−1	SSB	SSb/(k−1)	MSB/MSR

Interacción	(l −1)(k −1) = inter	SSinte	SSinte/inter	MSinter/MSR
Residual	t−(k−1)−(l−1)−(l−1)(k −1)=r	SSR	SSR/r	
Total	njkl−1=t	SST		

En nuestro caso concreto:

Fuente	**g.d.l**	**SS**	**MS**	**F**
Efectos globales	3	1,7501		
Factor A	1	0,5041	0,5041	22,81
Factor B	2	1,2460	0,6230	28,19
Interacción	2	0,8962	0,4481	20,27
Residual	30	0,6639	0,0221	
Total	35	3,3102		

Luego todos los efectos, incluida la interacción son significativos.

En el caso de que la interacción se estime y no sea significativa, puede ser incorporada en el residual. Para ello simplemente se suman las SS y los g.d.l. correspondientes.

Existen otros tipos más complicados, como el **MANOVA o ANOVA multivía**, así como los **ANOVA encajados (nested).** Estos últimos son de interés en los ensayos de validación interlaboratorios.

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

Química Analítica Avanzada Curso 2001/2002

Quimiometría (Lección 1)

Química Analítica Avanzada Curso 2001/2002

Identificación y confirmación de las necesidades de información del problema económico−social

Problema resuelto

Definición de la información analítica requerida

Planificación de la estrategia requerida

Fin

Comprobar

Monitorización de resultados

Acciones correctoras

ð

0

ð

ð

j

x

ij

Bias

del

método

Error aleatorio

(

Reproducibilidad

)

Bias

del

laboratorio

Errores aleatorios

del laboratorio

x

F(x)

x

F(x)

f(z)

f(x)

−z

z

z

ð

z

ð

/2

z

ð

ð

0

x

z

ð

z

ð

/2

z

ð

ð

0

ð

0