TRATAMIENTOS ESTADÍSTICOS EN ENSAYOS DE APTITUD: APLICACIÓN DE LA MEDIANA PARA DETECCIÓN DE RESULTADOS ANÓMALOS

J. Laso Sánchez[1] y A. Peris García-Patrón[1]

[1] Gabinete de Servicios para la Calidad S.A.L., C/ Caridad 32, 28007 Madrid; e-mail: gscsal@gscsal.com

RESUMEN. El objetivo de este trabajo es analizar diversos sistemas de tratamientos estadísticos para la obtención de los valores de consenso en los ensayos de aptitud. En este artículo trataremos una nueva sistemática de tratamiento previo de los resultados de los participantes, alternativa a otros tests de eliminación de anómalos que utiliza la eficacia de la mediana como elemento de eliminación, sin la modificación de los resultados de los participantes. Este test ha sido desarrollado por GSC SAL, y aplicado en múltiples esquemas de intercomparación desde el año 2001.

1.- Introducción

La evaluación de la calidad de los resultados de los ensayos incluye una gran variedad de actividades entre las que se encuentra la realización de pruebas de precisión, en diferentes condiciones, o el uso de materiales de referencia, además de la intercomparación, cada una de las cuales aporta información diferente sobre las características del método o el mantenimiento de sus propiedades. Es una herramienta absolutamente potente para conseguir otros propósitos: controlar y comprobar nuestras incertidumbres y que éstas que declaramos son verdaderas, o incluso llegar a disponer de los datos necesarios para realizar una validación formal de nuestros métodos de ensayo.

En el caso de los ensayos de intercomparación, los datos aportados por los laboratorios participantes, a menudo, y, según corroboran las pruebas experimentales, aparecen distribuidos según modelos similares a distribuciones normales, a las que, con frecuencia, se añaden observaciones de carácter "anómalo" respecto del conjunto. El tratamiento previo de estos datos supone minimizar o eliminar el peso estadístico de estas observaciones anómalas, con un objetivo: establecer parámetros adecuados para evaluar los resultados de los laboratorios participantes.

2.- Necesidad de la aplicación de tests de detección de "anómalos"

Recordemos que, el objetivo final de la aplicación de los tests de detección de anómalos es la consecución de unos buenos y seguros "parámetros del ensayo de intercomparación o del ejercicio", que básicamente son los que, en la mayor parte de las ocasiones, se emplean para el cálculo de uno de los parámetros de valoración o clasificación de los resultados, cual es el estadístico "z-score", aunque también son de aplicación para la estimación de otros estadísticos de clasificación, como los En ó z´.

3.- Tests de detección de resultados anómalos

Sobre la detección de resultados estadísticamente anómalos, las normas que establecen sistemáticas sobre la existencia de valores atípicos y su tratamiento son fundamentalmente la ISO 5725, así como la ISO 35, en algún caso. La ISO 43 también plantea algunas propuestas sobre este tema

La orientación del comportamiento anómalo de los datos ha tenido, hasta el presente reciente, un doble objetivo: por un lado, el descubrimiento de datos anómalos en precisión (normalmente repetibilidad), respecto a la precisión de las medidas aportadas por el conjunto de laboratorios y, por otro lado, el descubrimiento de datos (medias u observaciones individuales) de los participantes que se apartan del valor más probable asignado por los laboratorios.

La detección de ambos "problemas", de precisión y de exactitud, ha impulsado el desarrollo de diferentes tests, muy conocidos en el campo de la intercomparación; algunos de ellos se han publicado en Normas Internacionales, como las mencionadas anteriormente.

En general, estos tests que identifican los resultados estadísticamente anómalos van encaminados, una vez detectados los mismos, a:

a) corregir los datos considerados como tales

b) eliminar del conjunto de resultados, aquellos que difieren estadísticamente del resto.

Como sea que, generalmente el comportamiento anómalo de los resultados es atribuible a errores humanos, mala aplicación de los métodos o de los requisitos de ensayo (por ejemplo, unidades de expresión o cálculos finales realizados), la corrección de éstos puede no ser la táctica más plausible, o, cuando menos, es discutible.

Una vez detectados (y, bien eliminados, bien corregidos) estos datos, es posible asignar al ejercicio, para el analito objeto de ensayo, los "parámetros del mismo", V_A y σ.

En la práctica de eliminación, han caído en desuso los tests de discriminación por precisión, como son los tests de Cochran.

Está demostrado, o, al menos bajo sospecha que, en un

ensayo de aptitud, la eliminación por repetibilidad puede considerarse "injusta", ya que, a menudo, los laboratorios realizan una "interpretación" de las condiciones de ensayo diferentes a las previstas: aún cuando los resultados son solicitados como más de una observación (duplicados, en general), no siempre el laboratorio informa de éstas, o, las observaciones informadas son idénticas o, tan parecidas que, hacen sospechar que la repetición no se haya realizado, en algún caso, a partir de la muestra inicial, sino sólo de una submuestra parte del proceso (por ejemplo, de un mismo extracto el laboratorio realiza dos valoraciones, dos inyecciones, etc…). Por otro lado, al tratarse de un ensayo de aptitud y, por tanto, al existir la posibilidad de que los laboratorios puedan aplicar métodos con diferentes precisiones conocidas, pueden condicionar la eliminación de algunos datos, respecto a otros, si éstos últimos datos son más abundantes y proceden de métodos más precisos y, por tanto más empleados. Un caso particularmente ilustrativo es la determinación de masa volúmica por métodos areométricos o por densimetría electrónica. Así mismo, la eliminación de un resultado anómalo por repetibilidad no tiene sobre el participante ninguna repercusión, que no sea la de la propia eliminación del conjunto de datos que valora los parámetros de consenso, ya que, finalmente, la evaluación del z-score se aplica sobre su resultado medio. Estos tests están evolucionando en los ensayos de aptitud hacia la práctica desaparición, aunque, la valoración de esta repetibilidad puede ser, a nuestro juicio, una herramienta útil que permita al laboratorio valorar si su media era o no adecuada y, por tanto, comprender la potencial causa de la obtención de un z-score no satisfactorio. GSC también ha desarrollado una sistemática de valoración informativa de la repetibilidad de un laboratorio respecto al conjunto.

Siguiendo con la estadística de detección de anómalos, en esta ocasión, de aquellos que se separan más del valor considerado como más probable o esperado, han sido muchos los que han sido empleados, por su sencillez e incluso su publicación didáctica en las propias normas de referencia (ejemplo de ellos es el test de Grubbs, publicado con ejemplos de aplicación en la propia ISO 5725). Entre ellos aunque, actualmente también en cierto desuso, por su falta de capacidad de identificación del resultado anómalo cuando éste se encuentra enmascarado por agrupación con otro o más resultados potencialmente anómalos, se encuentran por ejemplo:

-Test Q de Dixon, procedente de una prueba de contraste que evalúa el cumplimiento o no del estadístico Q, calculado, frente al establecido en tablas:

$$Q = \frac{V_{sospechoso} - V_{más.cercano}}{V_{máximo} - V_{mínimo}} \qquad (1)$$

-Test de Grubbs, prueba de contraste que emplea el estadístico G, como cálculo, comparándolo con tablas:

$$G = \frac{\left|V_{sospechoso} - V_{medio}\right|}{s} \qquad (2)$$

-O su variante Grubbs doble, mediante la estimación de varianzas de todos los resultados, y tras la eliminación de las dos más desfavorables:

$$G = S^2_{p-1,\ p} \Big/ S^2_0 \qquad (3)$$

Dixon ha sido el test empleado en algunos circuitos hasta hace poco tiempo.

El problema de la aplicación de los tests mencionados es que, en general, la eliminación no es del todo eficaz. La eficacia de identificación de anómalos y su eliminación recurrente ha sido demostrada como limitada, en estos casos, ya que su aplicación supone:

a) una población suficiente de datos

b) que éstos siguen una distribución próxima a normal

c) la no aparición de datos anómalos "múltiples" o agrupados.

Por ello, y, para vencer la imposibilidad de cumplir, en muchas ocasiones, las anteriores circunstancias, existe una tendencia actual a la aplicación de procedimientos de detección de anómalos basados en estadística robusta.

4.- Estadísticas robustas en la detección de anómalos

En este caso, criterios más modernos se apoyan en estadísticas robustas basadas en propiedades de la mediana que no se ven tan afectados por el tipo de población existente.

La aplicación de la estadística robusta parece marcar el actual panorama de la detección de anómalos, con los que el evaluador decide sobre su eliminación o transformación.

La Norma ISO 13528, en lo que se refiere al tratamiento de los datos aportados por los participantes se caracteriza por:

-Establecer las sistemáticas posibles de asignación de valores al valor central incluyendo la utilización de estadística robusta, y al valor de variabilidad, fundamentándose generalmente en la utilización de la σ objetivo.

-Establecer, así mismo, la conveniencia de la comparación de la σ realmente obtenida con respecto a la σ objetivo considerando que no se debe superar esta en un factor crítico de 1,2.

Una característica fundamental de la norma es que utiliza todos los valores obtenidos por los participantes, sin descartar ninguno pero modificando aquellos que considera atípicos.

El sistema del algoritmo A, establecido en la mencionada Norma, se basa en la realización de un proceso recursivo, hasta la convergencia de los datos obtenidos. Obtiene un valor central como media y una desviación estándar

robusta.

Si x_i es el valor del laboratorio i de total de p laboratorios.

x* = mediana (x_i)

s* = 1,483 • mediana /x_i – x*/

Se calcula φ = 1,5 • s*

Se sustituyen los valores iniciales x_i según la siguiente regla:

x_i^* = x* - φ si x_i < x* - φ

x* + φ si x_i > x* + φ

x_i, en resto de casos

De este modo los datos anómalos, se sustituyen para realizar los cálculos por el valor extremo, con lo que:

Se calculan los nuevos x* y s* como:

$$x^{*n} = \sum x_i^* / p \quad (4)$$

$$s^{*n} = 1{,}134\sqrt{\sum\left(x_i^* - x^{*n}\right) / \left(p - 1\right)} \quad (5)$$

Se repite el proceso hasta convergencia.

5.- Estadística robusta de eliminación de anómalos

Entendiendo que, la estadística robusta y, el empleo de la mediana como elemento fundamental era el horizonte más cercano y seguro para la detección de anómalos, Gabinete de Servicios para la Calidad ha desarrollado y aplicado, desde el año 2001 una sistemática propia.

Las suposiciones que subyacen en esta sistemática son:

-La población de los laboratorios sigue una distribución gaussiana.

-Existen laboratorios con resultados anómalos que modifican la distribución.

-Se deben eliminar los valores anómalos para calcular los parámetros reales de la población.

-Los valores anómalos se sitúan en los extremos.

-Los valores centrales permiten estimar los datos reales de la población minimizando la influencia de los anómalos.

La sistemática se desarrolla en los siguientes pasos:

1. Obtención de la mediana de los resultados aportados por los laboratorios participantes, en adelante M_e.

M_e=*Mediana* (x_i)

2. Cálculo, para cada participante *i*, del conjunto de participantes *p*, de la diferencia en valor absoluto, entre el valor obtenido V_{Li} y la Mediana del conjunto M_e:

$d_i = /V_{Li} - M_e/$.

3. Obtención de la mediana de estas diferencias, $M_{edi.}$:

$M_{edi.}$=*Mediana* (d_i)

Ese valor debería corresponder al punto donde se sitúa el 50% de la población.

4. Cálculo de la dispersión máxima admisible, estimada en función del número de laboratorios participantes, con un intervalo de probabilidad del 50 %, según la ecuación:

$$s_{máx} = \frac{|Me_{di}|}{t_{(0,5;n-1)}} \quad (6)$$

Siento *t*, la t de student, de dos colas, para un α=*0.50* y los *n-1* grados de libertad, equivalentes a los *n-1* datos empleados en la estimación de la mediana. Así, hacemos depender la *s*, y posteriormente, el intervalo de aceptación, del número de participantes *n*.

5. La $s_{máx}$, permitirá establecer un intervalo:

$$Me_{di} \pm 1{,}96 s_{máx} \quad (7)$$

tal que, los resultados que se encuentren fuera de él, se considerarán estadísticamente anómalos, es decir, fuera de la población con esperanza estadística del 95%, y serán eliminados.

Nota: se pueden utilizar otros intervalos, de 99% para mejorar convergencia.

Se excluye, pues, en virtud del resultado de la diferencia encontrada entre cada uno de los valores individuales y la mediana resultante, aceptada inicialmente, como la mejor estimación del valor central.

Este test se aplicará de forma recurrente, si se estima necesario, de manera que, finalmente, después de haber procedido a la eliminación de anómalos, con los resultados no excluidos se procederá a:

-Calcular la media de los mismos, que será el valor considerado como Valor de Consenso o Valor asignado V_A, empleado en la estimación del z-score.

-Calcular la σ, como desviación estándar de los datos no excluidos, y que, en general, será el empleado en la valoración del z-score.

6.- Evaluación de los parámetros del ejercicio. Estimación de la z-score

El sistema de eliminación presentado, incorpora características particulares, entre las que destacan:

-Estimación de intervalos usando el estadístico "mediana", menos sujeto al comportamiento más o menos normal de la población, especialmente conflictivo cuando los datos son reducidos en número.

-Adecuación de los intervalos de aceptación, en función del número de participantes (t-student).

-No manipulación o corrección de los datos de los laboratorios a valores límite que no han sido aportados por los participantes y que no pueden, en muchos casos asegurar. Recordemos que la corrección de datos propuesta en el Algoritmo A, implica asignar a los participantes con resultados inaceptables, los valores límites, originando valores extremos a ambos lados de la población que, afectan al cálculo de una desviación estándar robusta, ficticia.

Con ello, es posible encontrar parámetros de ejercicio con las siguientes características:

-Valor asignado, calculado, siempre que es posible, a través del valor de consenso de los laboratorios. El valor asignado será la media de los datos no eliminados por ser estadísticamente anómalos que, de responder a un comportamiento "cuasinormal" de la población de los datos aceptados, debe coincidir estadísticamente con la mediana.

-Establecimiento del valor σ, que, denotará, el grado de confianza o de seguridad, en el valor asignado empleado en el cálculo de la z-score.

-Este valor de σ, calculado como la desviación estándar experimental puede ser usado, previa valoración, para el cálculo de la z-score, aunque otros esquemas son posibles.

-El evaluador debe describir el origen de cada uno de los parámetros del ejercicio, para que, en caso necesario, el participante pueda realizar su propia valoración, si lo considera oportuno.

7.- Ejemplos de tratamiento

En primera instancia, compararemos el comportamiento de tres de los tests mencionados en el artículo, Test de Grubbs, Algoritmo A y Mediana Robusta de GSC.

En este supuesto se obtuvieron los siguientes resultados en una intercomparación de un metal pesado en alimento (resultados expresados en ppb,s):

Tabla 1. Resultados de intercomparación de metal pesado en alimento

Participante	Media	Participante	Media
1	180,05	11	198
2	*350*	12	199
3	*322,9*	13	224
4	*126,5*	14	222,95
5	244,99	15	234,65
6	225	16	288,25
7	220,8	17	*210*
9	205,2	18	222,15
10	181	19	241,05

Los datos anotados en cursiva representas los eliminados por el test de GSC, según se especifica en la tabla resumen, en la que se incluyen los resultados iniciales, y los del tratamiento estadístico de los tests mencionados:

Tabla 2. Resultados

RESULTADOS DE LOS TESTS				
Parámetro	Datos Iniciales	Resultados Grubbs sencillo/doble	Algoritmo A	Sistema GSC
Media	228	228	217	215
Mediana	223	223	223	222
S	51,7	51,7	20,4	20,3
N	18	18	18	14

El problema de muchos de los tests de eliminación se confirma con lo encontrado en este ejemplo: la imposibilidad de eliminación de resultados claramente anómalos.

Los resultados del Algoritmo A y GSC son equivalentes, habiendo aplicado una única iteración (más iteraciones no producen eliminación).

Por otro lado y, para valorar la adecuación de la sistemática presentada, se ha realizado el tratamiento de los datos presentados en los ejemplos del Anexo III del protocolo IUPAC mediante el test de GSC, para confirmar los resultados obtenidos con respecto al sistema propuesto.

7.1 Ejemplo 1 IUPAC

Distribución unimodal y simétrica (propiedad % masa).

Los resultados del tratamiento realizado por el sistema GSC fueron los siguientes:

Tabla 3. Resultados del tratamiento por sistema GSC

DATOS SISTEMA G.S.C.		
	Iniciales	Iteración 1
Media	53,103	53,307
Mediana	53,297	53,31
S	1,962	0,5036
N	68	60
Mediana diferencias	0,3805	0,32
Steórica	0,561	0,471

Se ha utilizado un intervalo 3σ para la definición de límites de eliminación, debido a la simetría de la distribución.

En comparación con los resultados obtenidos al aplicar el Algoritmo A:

Tabla 4. Comparación de resultados

DATOS COMPARATIVOS		
	Algoritmo A	Sistema GSC
Media	53,24	53,306
S	0,64	0,5036
N	68	60
Mediana	53,30	53,31

En algoritmo A, Media y S son la media y desviación típica robustas.

Los resultados obtenidos son equivalentes.

7.2 Ejemplo 2 IUPAC

Distribución unimodal asimétrica (propiedad en ppbs). La representación en histograma de los datos, muestra una población unimodal y asimétrica:

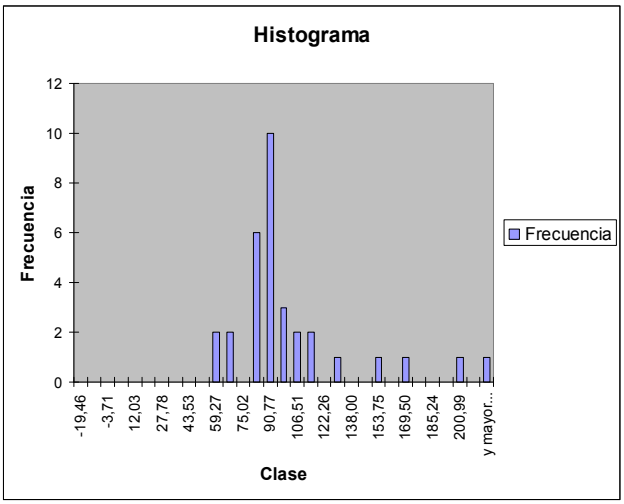

Fig. 1. Distribución unimodal

Los resultados de la aplicación de la estadística robusta de GSC, lleva a los siguientes resultados, después de aplicar el test hasta convergencia:

Tabla 5. Resultados de aplicación de estadística robusta GSC

DATOS sistema G.S.C.				
	Datos Iniciales	Iteración 1	Iteración 2	Iteración 3
Media	98,64	85,92	86,63	85,53
Mediana	89	84,74	85	84,74
S	39,37	16,30	11,73	10,59
Mediana diferencias	10,85	6,5	6	6
Steórica	15,89	9,51	8,76	8,76
N	32	28	25	24

Se ha utilizado intervalo 3σ debido al tipo de distribución.

La comparación con los resultados obtenidos tras la aplicación de otros tests, se muestra a continuación:

Tabla 6. Comparación de resultados

DATOS COMPARATIVOS		
	Algoritmo A	Sistema GSC
Media	91,45	85,53
S	23,64	10,59
N	32	24
Mediana	89	84,74

En el ejemplo del Anexo III, se indica que la estadística robusta no funciona satisfactoriamente, por lo que se aplica la detección mediante utilización de densidad Kernel y valoración de la moda 85,2.

La σ prevista por Horwitz es 19,7.

Puede comprobarse que, el Sistema GSC obtiene resultados excelentes de acuerdo a la decisión final del Anexo III, desde la primera iteración, con S inferior a las del resto de esquemas, incluso a la teórica de Horwitz, además de una buena coincidencia de media y mediana, coincidente así mismo con la moda de los datos iniciales, mejorando los resultados del Algoritmo A.

7.3 Ejemplo 3 IUPAC

Distribución bimodal (propiedad en ppm), según se muestra en el siguiente histograma de los datos:

Histograma 2.

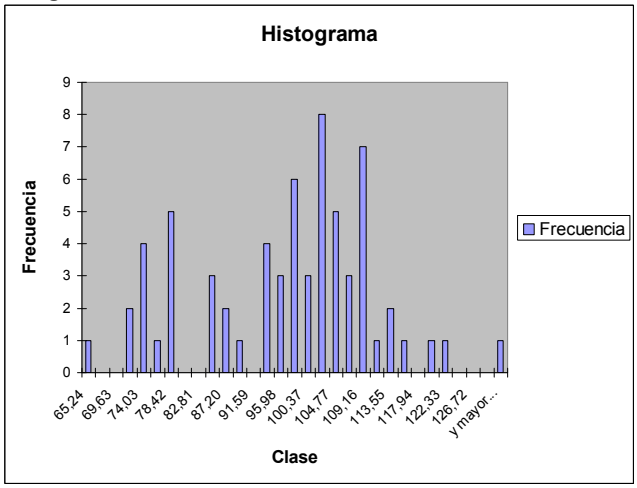

Fig. 2. Distribución bimodal

Se trata de la ejecución de un primer ejercicio tras un cambio de sistemática, que altera la expresión de resultados.

La σ de Horwitz es 7,71 ppm.

Los datos obtenidos tras la aplicación del Sistema de GSC hasta convergencia fueron:

Tabla 7. Datos obtenidos tras aplicación sistema GSC

DATOS SISTEMA G.S.C.				
	Iniciales	Serie 1	Serie 2	Serie 3
Media	95,69	98,18	100,58	101,25
Mediana	98,91	101	101	101,33
S	14,52	10,97	7,12	6,47
Mediana diferencias	8,09	6	5	5,33
S teórica	11,92	8,83	7,35	7,83
N	65	57	47	45

Se ha utilizado intervalo 2σ debido a la bimodalidad.

Tabla 8. Algoritmo A/GSC

	Algoritmo A	**GSC**
Media	95,78	101,25
S	14,63	6,47
N	65	45
Mediana	98,89	101,33

El en Anexo III (IUPAC), se indica que se estudian mediante densidad Kernel, eliminando la moda más baja, lo que conduce a un resultado asignado de 101,5.

En este caso, una vez más, el sistema GSC lleva a resultados equivalentes a la decisión tomada, mejorando el Algoritmo A, sin tener que llevarse a cabo consideraciones especiales.

8.- Conclusiones

La estadística de GSC, que utiliza la mediana como elemento de detección de los datos anómalos, se presenta en este artículo como alternativa a otras estadísticas robustas y tests de eliminación, a los que los mejora ya que:

1. Permite detectar los datos "estadísticamente diferentes", para poder eliminarlos, cuando otros tests alternativos no eran capaces, debido, principalmente a la agrupación de datos conjuntamente anómalos.
2. No modifica los valores de los laboratorios con resultados detectados como "diferentes", asignando o cambiando el dato que el laboratorio proporcionó como resultado, como algunos algoritmos defienden, creando una población de datos diferentes a los originales, aportados por los participantes.
3. Consigue resultados adecuados cuando la población es de reducido tamaño, y tiene en cuenta este tamaño, para la estimación del intervalo de exclusión.
4. La exclusión de datos en el tratamiento previo, no tiene como finalidad valorar la participación del resultado excluido, sino la obtención de mejores parámetros de ejercicio, V_A y σ, con los que se valorará el estadístico de clasificación, z-score o similar. Es posible que resultados excluidos no obtengan z no satisfactoria.
5. Es necesaria la evaluación de la bondad de los parámetros obtenidos y su seguridad, para proporcionar a los laboratorios una puntuación relativa al resto, adecuada y justa.
6. La intercomparación debe proporcionar una información, en primera instancia, diferente a la obtenida cuando se analiza una muestra con valor conocido, como puede ser un material de referencia. La intercomparación limita al laboratorio en :

a) el momento en que se realiza el control (que es decidido por el organizador, generalmente).

b) el plazo para realizarlo (que suele ser limitado en el tiempo, a pocos días).

c) el coste, en muchos casos superior al de la adquisición de materiales de referencia.

d) el natural retardo en la obtención de la valoración, lo que conlleva una dificultad adicional en la toma inmediata de decisiones e implantación de acciones correctivas, cuando es necesario (con un material de referencia, la evaluación es, en general, inmediata y, la toma de medidas, por consiguiente, también).

7. Los laboratorios tienen la posibilidad de valorar siempre, al margen de la evaluación del proveedor (con la mencionada normalización), y con la información obtenida en el ensayo, los resultados de su participación, con sus propios criterios (que no serán otros que los de procedencia o como quiera que hayan documentado, en muchos casos, la definición de las características de sus métodos), como se procedería con la evaluación del resultado obtenido del análisis de un material con valor asignado, material de referencia o similar, que es lo que, en la práctica muchos circuitos están aplicando.

Referencias

UNE 66543-1IN (ISO Guide 43-1) *"Ensayos de aptitud por intercomparación de laboratorios. Parte 1: Desarrollo y aplicación de programas de ensayos de aptitud"*

ISO 13528:2005 *"Statistical methods for use in proficiency testing by interlaboratory comparisons"*.

Protocolo IUPAC Technical Report *"The international harmonized protocol for the proficiency testing of analytical chemistry laboratories". Pure Appl.Chem., Vol 78, nº 1, pp.145-196, 2006. M. Thompson, S. L.R. Ellison y R. Wood.*