

NC - ISO 13528 Metodos Estadisticos Utilizados EN LOS EAA

Quimica (Universidad de Concepción)



Escanea para abrir en Studocu

NORMA CUBANA

NC

ISO 13528: 2017
(Publicada por la ISO en 2015)

MÉTODOS ESTADÍSTICOS UTILIZADOS EN LOS ENSAYOS DE APTITUD POR COMPARACIÓN ENTRE LABORATORIOS (ISO 13528: 2015, IDT)

Statistical methods for use in proficiency testing by interlaboratory comparison

ICS: 03.120.30

1. Edición Marzo 2017
REPRODUCCIÓN PROHIBIDA

Oficina Nacional de Normalización Calle E No. 261 Vedado, Ciudad de La Habana.
Cuba. Teléfono: 830-0835 Fax: (537) 836-8048 Correo electrónico: nc@ncnorma.cu



Cuban National Bureau of Standards

Prefacio

La Oficina Nacional de Normalización (NC) es el Órgano Nacional de Normalización de la República de Cuba y representa al país ante las organizaciones internacionales y regionales de normalización.

La elaboración de las Normas Cubanas y otros documentos relacionados se realiza generalmente a través de los Comités Técnicos de Normalización. Su aprobación es competencia de la Oficina Nacional de Normalización y se basa en las evidencias del consenso.

Esta Norma Cubana:

- Ha sido elaborada por el **Comité Técnico de Normalización NC/CTN 46 de Métodos de Análisis y Toma de Muestras**, en el que están representadas las siguientes entidades:
 - Centro de Investigaciones Pesqueras (CIP). MINAL.
 - Corporación CUBARON S.A. – Dirección de Calidad. MINAL.
 - Dirección de Normalización (DN) – ONN. CITMA.
 - Instituto Nacional de Higiene, Epidemiología y Microbiología (INHEM). MINSAP.
 - Laboratorio Central de Calidad (LACCAL). CIDCI – MINCIN.
 - Laboratorio de Supervisión de la Calidad CUBACONTROL, OSDE CAUDAL S.A. - CM.
 - Laboratorio Nacional de Higiene de los Alimentos (LNHA), ULCSA. MINAG.
 - Oficina Nacional de Inspección Estatal (ONIE). MINAL.
- Es una adopción idéntica por el método de traducción de la Norma Internacional ISO 13528: 2015 *Statistical methods for use in proficiency testing by interlaboratory comparison*. Se ha mantenido el formato original de la norma ISO adoptada.
- Contiene los Anexos A, B y C normativos; y los Anexos D y E informativos.

© NC, 2017

Todos los derechos reservados. A menos que se especifique, ninguna parte de esta publicación podrá ser reproducida o utilizada en alguna forma o por medios electrónicos o mecánicos, incluyendo las fotocopias, fotografías y microfilmes, sin el permiso escrito previo de:

Oficina Nacional de Normalización (NC)

Calle E No. 261, Vedado, Ciudad de La Habana, Habana 4, Cuba.

Impreso en Cuba.

Prefacio de la Norma Internacional

ISO (Organización Internacional de Normalización) es una federación mundial de cuerpos nacionales de normalización (organismos miembros de ISO). El trabajo de preparación de las normas internacionales normalmente se realiza a través de los comités técnicos de ISO. Cada organismo miembro interesado en una materia para la cual se haya establecido un comité técnico, tiene el derecho de estar representado en dicho comité. Las organizaciones internacionales, gubernamentales y no gubernamentales, en coordinación con ISO, también participan en el trabajo. ISO colabora estrechamente con la Comisión de Electrotécnica Internacional (IEC) en todas las materias de normalización electrotécnica.

Los procedimientos utilizados para desarrollar este documento y los destinados a su posterior mantenimiento se describen en las Directivas ISO / IEC, Parte 1. En particular, debe tenerse en cuenta los diferentes criterios de aprobación necesarios para los diferentes tipos de documentos ISO. Este documento fue elaborado de acuerdo con las normas editoriales de las Directivas ISO / IEC, Parte 2 (ver www.iso.org/directives).

Llama la atención la posibilidad de que algunos de los elementos de este documento puedan estar sujetos a derechos de patente. ISO no se hace responsable por la identificación de cualquiera o todos los derechos de patente. Los detalles de cualquier derecho de patente identificados durante el desarrollo del documento estarán en la introducción y / o en la lista de las declaraciones de patentes ISO recibidas (ver www.iso.org/patents).

Cualquier nombre comercial usado en el presente documento es información dada para la conveniencia de los usuarios y no constituye un aval.

Para obtener una explicación sobre el significado de los términos específicos de la ISO y las expresiones relacionadas con la evaluación de la conformidad, así como información sobre el cumplimiento de la norma ISO a los principios de la WTO en las Barreras Técnicas del Comercio (TBT) consulte el siguiente URL: [Foreword - Supplementary information](#)

El comité responsable de este documento es el ISO / TC 69, Aplicaciones de métodos estadísticos, Subcomité SC 6, Métodos de medición y resultados.

Esta segunda edición de la Norma ISO 13528 anula y reemplaza la primera edición (ISO 13528: 2005), de la que constituye una revisión técnica.

Esta segunda edición ofrece cambios para traer el documento en armonía con la norma ISO / IEC 17043: 2010, la cual reemplaza a la Guía ISO 43 1: 1997. Esta sigue una estructura revisada, para describir mejor el proceso del diseño, el análisis y el reporte de los esquemas de ensayos de aptitud. También elimina algunos procedimientos que ya no se consideran apropiados, y agrega o modifica algunos procedimientos que no son ampliamente considerados como apropiados y añade o revisa algunas otras secciones para ser coherente con la norma ISO / IEC 17043 y proporcionar claridad y corregir errores menores. Nuevas secciones han sido añadidas para los datos cualitativos y métodos estadísticos robustos.

Índice

PREFACIO.....	2
PREFACIO DE LA NORMA INTERNACIONAL.....	3
0 INTRODUCCIÓN	8
1. ALCANCE.....	11
2. REFERENCIAS NORMATIVAS.....	11
3. TÉRMINOS Y DEFINICIONES.....	12
4 PRINCIPIOS GENERALES.....	15
4.1 REQUISITOS GENERALES PARA LOS MÉTODOS ESTADÍSTICOS	15
4.2 MODELO BÁSICO	15
4.3 ENFOQUES GENERALES PARA LA EVALUACIÓN DEL DESEMPEÑO.....	16
5. DIRETRICES PARA EL DISEÑO ESTADÍSTICO DE LOS ESQUEMAS DE ENSAYOS DE APTITUD.....	16
5.1 INTRODUCCIÓN AL DISEÑO ESTADÍSTICO DE LOS ESQUEMAS DE ENSAYOS DE APTITUD.....	16
5. 2 BASES DE UN DISEÑO ESTADÍSTICO.....	17
5.3 CONSIDERACIONES PARA LA DISTRIBUCIÓN ESTADÍSTICA DE LOS RESULTADOS ..	18
5.4 CONSIDERACIONES PARA UN PEQUEÑO NÚMERO DE PARTICIPANTES.....	19
5.5 PAUTAS PARA ELEGIR EL FORMATO DE PRESENTACIÓN DE INFORMES.....	20
6 DIRETRICES PARA LA REVISIÓN INICIAL DE LOS ELEMENTOS DE ENSAYOS DE APTITUD Y LOS RESULTADOS.....	21
6.1 HOMOGENEIDAD Y ESTABILIDAD DE LOS ELEMENTOS DE ENSAYO DE APTITUD	21
6.2 CONSIDERACIONES PARA DIFERENTES MÉTODOS DE MEDICIÓN.....	23

6.3 ELIMINACIÓN DE ERRORES GROSEROS	24
6.4 REVISIÓN VISUAL DE LOS DATOS.....	24
6.5 MÉTODOS ESTADÍSTICOS ROBUSTOS	24
6.6 TÉCNICAS DE VALORES ATÍPICOS PARA RESULTADOS INDIVIDUALES	25
7 DETERMINACIÓN DEL VALOR ASIGNADO Y SU INCERTIDUMBRE TÍPICA	26
7.1 ELECCIÓN DEL MÉTODO DE DETERMINACIÓN DEL VALOR ASIGNADO	26
7.2 DETERMINACIÓN DE LA INCERTIDUMBRE DEL VALOR ASIGNADO.....	27
7.3 FORMULACIÓN	28
7.4 MATERIAL DE REFERENCIA CERTIFICADO.....	29
7.5 RESULTADOS DE UN LABORATORIO	29
7.6 VALOR DE CONSENSO DE LOS LABORATORIOS ESPECIALIZADOS.....	31
7.8 COMPARACIÓN DEL VALOR ASIGNADO CON UN VALOR DE REFERENCIA INDEPENDIENTE	33
8 DETERMINACIÓN DE LOS CRITERIOS PARA LA EVALUACIÓN DEL DESEMPEÑO	34
8.1 MÉTODOS PARA LA DETERMINACIÓN DE LOS CRITERIOS DE EVALUACIÓN	34
8.2 POR LA PERCEPCIÓN DE LOS EXPERTOS	35
8.3 POR LA EXPERIENCIA DE LAS RONDAS ANTERIORES DE ENSAYOS DE APTITUD	35
8.4 MEDIANTE EL USO DE UN MODELO GENERAL	36
8.5 USO DE LAS DESVIACIONES TÍPICAS DE REPETIBILIDAD Y REPRODUCIBILIDAD DE UN ESTUDIO EN COLABORACIÓN DE UN MÉTODO DE MEDICIÓN PREVIO DE PRECISIÓN	37
8.6 A PARTIR DE LOS DATOS OBTENIDOS EN LA MISMA RONDA DE UN PROGRAMA DE ENSAYOS DE APTITUD	37
8.7 SEGUIMIENTO DEL ACUERDO ENTRE LABORATORIOS	39

9 EL CÁLCULO DE LAS ESTADÍSTICAS DE RENDIMIENTO.....	39
9.1 CONSIDERACIONES GENERALES PARA LA DETERMINACIÓN DEL RENDIMIENTO	39
9.2 LIMITACIÓN DE LA INCERTIDUMBRE DEL VALOR ASIGNADO	40
9.3 LAS ESTIMACIONES DE DESVIACIÓN (ERROR DE MEDICIÓN).....	41
9.4 Z SCORE	42
9.5 Z'-SCORE	43
9.6 ZETA SCORES (Z).....	44
9.7 E _N SCORES	46
9.8 EVALUACIÓN DE INCERTIDUMBRES DE PARTICIPANTES EN ENSAYOS	47
9.9 PUNTUACIONES COMBINADAS DE EFICIENCIA.....	48
10. MÉTODOS GRÁFICOS PARA DESCRIBIR LAS PUNTUACIONES DE EFICIENCIA	49
10.1 APLICACIÓN DE MÉTODOS GRÁFICOS.....	49
10.2 HISTOGRAMAS DE RESULTADOS O PUNTUACIONES DE EFICIENCIA.....	49
10.3 DENSIDAD KERNEL PLOTEADA.....	50
10.4 PLOTEOS DE BARRAS DE PUNTUACIONES DE EFICIENCIA ESTANDARIZADOS....	52
10.5 PLOTEOS DE YOUNDEN	52
10.6 PLOTEOS DE REPETIBILIDAD DE DESVIACIONES TÍPICAS.....	53
10.7 MUESTRAS DIVIDIDAS	54
10.8 MÉTODOS GRÁFICOS PARA COMBINAR LAS PUNTUACIONES DE EFICIENCIA A TRAVÉS DE PRUEBAS DE UN ESQUEMA DE ENSAYOS DE APTITUD	55
11. DISEÑO Y ANÁLISIS DE ESQUEMAS DE ENSAYOS DE EFICIENCIA CUALITATIVA (INCLUIDAS LAS PROPIEDADES ORDINALES Y NOMINALES)	56

11.1 TIPOS DE DATOS CUALITATIVOS	56
11.2 DISEÑO ESTADÍSTICO.....	57
11.3 VALORES ASIGNADOS PARA ESQUEMAS DE ENSAYOS DE APTITUD CUALITATIVOS	58
11.4 EVALUACIÓN DE DESEMPEÑO PARA PROGRAMAS DE ENSAYOS DE APTITUD CUALITATIVOS Y DE PUNTUACIÓN	59
ANEXO A.....	62
(NORMATIVO).....	62
SÍMBOLOS.....	62
ANEXO B.....	64
(NORMATIVO).....	64
LA HOMOGENEIDAD Y LA ESTABILIDAD DE LOS ELEMENTOS DE PRUEBA DE APTITUD	64
ANEXO C.....	73
(NORMATIVO).....	73
ANÁLISIS ROBUSTO.....	73
ANEXO D.....	87
(INFORMATIVO).....	87
ORIENTACIONES ADICIONALES SOBRE PROCEDIMIENTOS ESTADÍSTICOS	87
ANEXO E.....	93
(INFORMATIVO).....	93
EJEMPLOS ILUSTRATIVOS.....	93
BIBLIOGRAFÍA	122

0 introducción**0.1 El propósito de los ensayos de aptitud**

Los ensayos de aptitud involucran el uso de comparaciones interlaboratorios para determinar el desempeño de los participantes (que pueden ser laboratorios, cuerpos de inspección, o individuos) para pruebas específicas o mediciones, y para monitorear su desempeño continuo. Existen una serie de propósitos característicos de los ensayos de aptitud, como se describe en la Introducción para la norma ISO / IEC 17043: 2010. Estos incluyen la evaluación del desempeño de los laboratorios, la identificación de problemas en los laboratorios, establecimiento de la efectividad y la comparación de los métodos de ensayo o medición, la provisión de confianza adicional para los clientes de laboratorio, la validación de las incertidumbres requeridas y la educación de los laboratorios participantes. El diseño estadístico y las técnicas analíticas aplicadas deben ser apropiados para el propósito declarado (s).

0.2 Razón para la puntuación en los esquemas de ensayos de aptitud

Una variedad de estrategias de calificación está disponible y en uso de las pruebas de aptitud. Aunque los detalles de los cálculos difieren, la mayoría de los esquemas de ensayos de aptitud comparan la desviación del participante de un valor asignado con un criterio numérico que se utiliza para decidir si la desviación representa una preocupación o no. Las estrategias utilizadas para el valor asignado y para la elección de un criterio para la evaluación de las desviaciones de los participantes son por lo tanto críticas. En particular, es importante considerar si el valor asignado y el criterio para la evaluación de las desviaciones deben ser independientes de los resultados de los participantes, o deberían ser derivados de resultados presentados. En esta norma, se proporcionan ambas estrategias. Sin embargo, se llama la atención a la discusión que se encuentra en las secciones 7 y 8 de las ventajas y desventajas de la selección de los valores asignados o criterios para la evaluación de las desviaciones que no se derivan de los resultados de participantes. Se verá que, en general, que la elección de los valores asignados y criterios de evaluación independiente de los resultados de participantes ofrece ventajas. Este es particularmente el caso para el criterio usado para evaluar desviaciones del valor asignado - tales como la desviación típica para la evaluación de la aptitud o un consentimiento para el error de medición - para lo cual una elección coherente basada en la idoneidad para un uso final particular de los resultados de la medición, es especialmente útil.

0.3 ISO 13528 e ISO/IEC 17043

ISO 13528 proporciona ayuda para la implementación de ISO / IEC 17043 particularmente, sobre los requisitos para el diseño estadístico, la validación ensayos de aptitud, la revisión de los resultados y la presentación de los resúmenes estadísticos. El Anexo B de la norma ISO / IEC 17043: 2010 describe brevemente los métodos estadísticos generales que son usados en los esquemas de ensayos de aptitud. Esta Norma Internacional pretende ser complementaria a la norma ISO / IEC 17043, proporcionando una guía detallada que no está incluida en este documento sobre los métodos estadísticos particulares para los ensayos de aptitud.

La definición de ensayo de aptitud en ISO / IEC 17043 se repite en la norma ISO 13528, con las notas que describen diferentes tipos de ensayos de aptitud y la variedad de diseños que se pueden utilizar. Esta Norma no puede cubrir específicamente todos los propósitos, diseños, matrices y mensurandos. Las técnicas presentadas en ISO 13528 pretenden ser ampliamente aplicables, especialmente para los esquemas de ensayos de aptitud de reciente creación. Se espera que las técnicas estadísticas usadas para un esquema de ensayos de aptitud en particular se desarrollen como el esquema madura y las puntuaciones, los criterios de evaluación y las técnicas gráficas se perfeccionan para un mejor servicio a las necesidades de un grupo específico de participantes, los organismos de acreditación y las autoridades reguladoras.

ISO 13528 incorpora una guía publicada para los ensayos de aptitud de los laboratorios de análisis químicos [32] pero, además, incluye amplia gama de procedimientos para permitir el uso con métodos de medición válidos y de identificaciones cualitativas. Esta revisión de la norma ISO 13528: 2005 contiene la mayoría de los métodos estadísticos y la guía de la primera edición, ampliada por los documentos referidos anteriormente y un alcance ampliado de la norma ISO / IEC 17043. ISO / IEC 17043 incluye ensayos de aptitud para individuos y cuerpos de inspección y el Anexo B, que incluye consideraciones para resultados cualitativos. Esta norma incluye técnicas estadísticas que son consistentes con otras normas internacionales, particularmente aquellas de TC69 SC6, considerablemente la serie de normas ISO 5725 sobre Exactitud: veracidad y precisión. Las técnicas también pretenden reflejar otras normas internacionales, donde es adecuado, y se pretende ser consistentes con la norma ISO / IEC Guide 98-3 (GUM) y la Guía ISO / IEC 99 (VIM).

0.4 Experticia estadística

ISO / IEC 17043: 2010 con el objetivo de ser competente requiere, que el proveedor de ensayos de aptitud tenga acceso a la especialización estadística y se debe autorizar a un personal específico para llevar a cabo el análisis estadístico. Ni la norma ISO / IEC 17043, ni esta Norma Internacional pueden especificar más allá de lo que la experiencia necesaria lo puede hacer. Para algunas aplicaciones un grado avanzado en las estadísticas es útil, pero por lo general las necesidades de conocimientos se puede cumplir por las personas con experiencia técnica en otras áreas, que están familiarizados con los conceptos estadísticos básicos y tienen experiencia o capacitación en las técnicas comunes aplicables al análisis de los datos de los esquemas de ensayos de aptitud. Si un individuo está a cargo del diseño estadístico y / o análisis, es muy importante la experiencia que esta persona tenga en la comparaciones inter laboratorios, incluso si esa persona tiene un nivel avanzado de conocimiento estadístico. Los entrenamientos estadísticos avanzados convencionales a menudo no incluye ejercicios con comparaciones entre laboratorios, y las causas únicas de error de medición que se presentan en los ensayos de aptitud puede parecer oscuras. La guía que aparece en esta norma no puede proveer toda la experiencia necesaria para considerar todas las aplicaciones, y no puede sustituir la experiencia adquirida por el trabajo con las comparaciones entre laboratorios.

0.5 Programas computacionales

Los programas informáticos que se necesita para el análisis estadístico de los datos de ensayos de aptitud pueden variar mucho, desde una simple hoja de cálculo aritmético para esquemas de ensayos de aptitud pequeños utilizando valores de referencia conocidos hasta software estadísticos sofisticados utilizados para los métodos estadísticos que dependen de cálculos iterativos u otros métodos numéricos avanzados. La mayor parte de las técnicas en esta norma internacional se puede ejecutar mediante aplicaciones de hojas de cálculos convencionales,

quizás con rutinas personalizadas para un esquema o un análisis en particular; algunas técnicas pueden requerir de aplicaciones informáticas que están libremente disponibles (en el momento de publicación de esta Norma). En todos los casos, los usuarios deben verificar la exactitud de sus cálculos, especialmente cuando rutinas especiales han sido introducidas por el usuario. Sin embargo, aun cuando las técnicas en esta Norma Internacional son adecuadas y correctamente implementadas por aplicaciones informáticas adecuadas, ellas no pueden aplicarse sin la atención de una persona con experiencia técnica y estadística que sea suficiente para identificar e investigar anomalías que pueden ocurrir en cualquier ronda del ensayo de aptitud.

MÉTODOS ESTADÍSTICOS UTILIZADOS EN ENSAYOS DE APTITUD POR COMPARACIÓN ENTRE LABORATORIOS

1. Alcance

Esta Norma Cubana proporciona una descripción detallada de los métodos estadísticos a ser usados por proveedores de ensayos de aptitud en el diseño de los esquemas de ensayos de aptitud y el análisis de los datos obtenidos en estos esquemas. Esta norma proporciona recomendaciones sobre la interpretación de los datos de ensayos de aptitud para los participantes en dichos esquemas y por los organismos de acreditación.

Los procedimientos de la presente norma pueden ser aplicados para demostrar que los resultados de las mediciones obtenidas por los laboratorios, cuerpos de inspección e individuos cumplen criterios especificados para un desempeño aceptable.

Esta norma es aplicable a ensayos de aptitud tanto para resultados de mediciones cuantitativas como para observaciones cualitativas del artículo de ensayo.

NOTA Los procedimientos de esta Norma también pueden ser aplicables a la evaluación de la opinión de expertos, donde las opiniones o juicios se presentan en una forma que puede compararse objetivamente con un valor de referencia independiente o un consenso estadístico. Por ejemplo, al clasificar los elementos de ensayo de aptitud en categorías conocidas por la inspección - o en determinar mediante comprobación de si surgen o no elementos de ensayo de aptitud, desde la misma fuente original - y los resultados de la clasificación son comparados objetivamente, las disposiciones de esta norma que se refieren a propiedades nominales (cuantitativas) se pueden aplicar.

2. Referencias normativas

Los siguientes documentos, en su totalidad o en parte, están normativamente referidos en este documento y son indispensables para su aplicación. Para las referencias con fecha, sólo se aplica la edición citada. Para las referencias sin fecha se aplica la última edición del documento de referencia (incluyendo cualquier modificación).

- Guía ISO 30, Materiales de referencia - Términos seleccionados y definiciones
- ISO 3534-1, Estadística - Vocabulario y símbolos - Parte 1: Términos estadísticos generales y términos usados en probabilidad
- ISO 3534-2, Estadística - Vocabulario y símbolos- Parte 2 : Aplicaciones Estadísticas
- ISO 5725-1, Exactitud (veracidad y precisión) de los métodos de medición y resultados - Parte 1: Principios generales y definiciones
- ISO / IEC 17043, Evaluación de la conformidad - Requisitos generales para los ensayos de aptitud
- ISO / IEC Guía 99, Vocabulario Internacional de Metrología - Conceptos básicos y generales y términos asociados (VIM)

3. Términos y definiciones

Para los propósitos de este documento, los términos y definiciones dados en la Norma ISO 3534-1, ISO 3534-2, ISO 5725-1, ISO / IEC 17043, ISO / IEC Guía 99, la Guía ISO 30, y los siguientes se aplican. En el caso de las diferencias entre estas referencias sobre el uso de términos, se aplican las definiciones de ISO 3534 partes 1-2. Los símbolos matemáticos se enumeran en el anexo A.

3.1 Comparación inter laboratorios

Organización, ejecución y evaluación de mediciones o ensayos sobre los mismos elementos o elementos similares por dos o más laboratorios de acuerdo con condiciones predeterminadas.

3.2 ENSAYOS DE APTITUD

Evaluación del desempeño de los participantes contra criterios pre establecidos por medio de comparaciones entre laboratorios

NOTA 1: A los efectos de esta norma, el término "ensayos de aptitud" se toma en su sentido más amplio e incluye, pero no se limita a:

- Esquema cuantitativo - donde el objetivo es cuantificar uno o más medidos para cada elemento de ensayo de aptitud;
- Esquema cualitativo - donde el objetivo es identificar o describir una o más características cualitativas del elemento de ensayo de aptitud;
- Esquema secuencial - donde uno o más elementos de ensayo de aptitud son distribuidos de forma secuencial para la prueba o medición y devueltos al proveedor de ensayos de aptitud a intervalos;
- Esquema simultáneo - donde los elementos de ensayos de aptitud son distribuidos para ensayar o medir simultáneamente dentro de un período de tiempo definido;
- Ejercicio de ocasión única - donde los artículos de ensayos de aptitud se proporcionan en una sola ocasión;
- Esquema Continuo - donde los elementos de ensayos de aptitud se proporcionan a intervalos regulares;
- Muestreo - donde se toman las muestras para su posterior análisis y el propósito del esquema de ensayos de aptitud incluye la evaluación de la ejecución del muestreo; y
- Interpretación de datos — donde unos conjuntos de datos u otra información están disponibles y la información se procesa para proporcionar una interpretación (u otro resultado).

3.3 Valor asignado

Valor atribuido a una propiedad particular de un elemento de ensayo de aptitud

3.4 Desviación típica para la evaluación de la competencia

Medida de dispersión usada en la evaluación de los resultados de ensayos de aptitud

NOTA 1: Esto se puede interpretar como la desviación típica de la población de los resultados provenientes de una población hipotética de laboratorios que funcionan exactamente en acuerdo con requisitos.

NOTA 2: La desviación típica para la evaluación de la competencia sólo se aplica a la relación y los resultados de intervalo de escala.

NOTA 3: No todos los esquemas de ensayos de aptitud evalúan el desempeño basado en la dispersión de los resultados.

[FUENTE: ISO / IEC 17043: 2010, modificada - En la definición ", en base a la información disponible" se ha eliminado. NOTA 1 se ha añadido, y las Notas 2 y 3 han sido ligeramente editadas.]

3.5 Error de medición

Valor cuantitativo medido menos un valor cuantitativo de referencia.

[FUENTE: ISO / IEC Guide99: 2007, Modificada - se han suprimido las Notas]

3.6 Error máximo permisible

Valor extremo de error de medición, con respecto a un valor cuantitativo de referencia conocido, permitido por las especificaciones o regulaciones para una medición dada, instrumento de medición, o sistema de medición.

[FUENTE: ISO / IEC Guide99: 2007., Modificada - se han suprimido las Notas]

3.7 Valor z

Medida estandarizada de desempeño, calculada utilizando el resultado del participante, el valor asignado y la desviación típica para la evaluación de la competencia.

NOTA 1: Una variación común en la puntuación z, a veces se denotada z' (comúnmente pronunciado z -prima), está formada por la combinación de la incertidumbre del valor asignado con la desviación típica para la evaluación de la competencia antes de calcular la puntuación z.

3.8 Valor zeta

Medida estandarizada de desempeño, calculada utilizando el resultado del participante, el valor asignado y las incertidumbres típicas combinadas para el resultado y el valor asignado.

3.9 Proporción de la puntuación límite permitido

Medida estandarizada de desempeño, calculada usando el resultado del participante, el valor asignado y el criterio de error de medición en un ensayo de aptitud.

NOTA 1: Para resultados individuales, el desempeño puede ser expresado como la desviación del valor asignado (D o $D\%$).

3.10 Señal de acción

Indicación de una necesidad de acción derivada de un resultado del ensayo de aptitud.

Ejemplo: Una puntuación z superior a 2 se toma convencionalmente como una indicación de la necesidad de investigar las posibles causas; una puntuación z superior a 3 se toma convencionalmente como una señal de acción que indica la necesidad de una acción correctiva.

3.11 Valor de consenso

Valor derivado de una colección de resultados en una comparación inter laboratorios.

NOTA 1: La frase "valor de consenso" se utiliza típicamente para describir los estimados de localización y dispersión derivada de resultados de los participantes en una ronda de ensayos de aptitud, pero también puede ser utilizado para referirse a valores derivados de los resultados de un subconjunto especificado de tales resultados o, por ejemplo, a partir de un número de laboratorios expertos.

3.12 Valor atípico (Outlier)

Miembro de un conjunto de valores que es incongruente con los demás miembros de ese conjunto.

NOTA 1: Un valor atípico puede surgir por cambio de la población esperada, proceder de una población diferente, o ser el resultado de un registro incorrecto u otra equivocación.

NOTA 2: Muchos esquemas utilizan el término valor atípico para designar un resultado que genera una señal de acción. Este no es el uso previsto de este término. Mientras que los valores atípicos usualmente generan señales de acción, es posible tener señales de acción a partir de resultados que no son valores atípicos.

[FUENTE: ISO 5725-1: 1994, modificada - Las Notas se han añadido.]

3.13 Participante

Laboratorio, organización o individuo que recibe los elementos de un ensayo de aptitud y somete los resultados a revisión por parte del proveedor de ensayos de aptitud.

3.14 Elemento de ensayo de aptitud

Muestra, producto, artefacto, material de referencia, pieza de un equipo, patrón de medición, conjunto de datos u otra información utilizada para evaluar el desempeño de un participante en ensayos de aptitud.

NOTA 1: En la mayoría de los casos, los elementos del ensayo de aptitud cumplen con la definición de "material de referencia" (3.17) de la Guía ISO 30.

3.15 Proveedor de ensayos de aptitud

Organización que asume la responsabilidad de todas las tareas en el desarrollo y operación de un esquema de ensayos de aptitud.

3.16 Esquema de ensayos de aptitud

Ensayos de aptitud diseñado y operado en una o más rondas de un área específica de ensayo, medición, calibración o inspección.

NOTA 1: Un esquema de ensayos de aptitud podría cubrir un tipo particular de ensayo, calibración, inspección o un número de ensayos, calibraciones o inspecciones en elementos de ensayo de aptitud.

3.17 Material de referencia (RM)

Material, suficientemente homogéneo y estable con respecto a una o más propiedades especificadas, las cuales han sido establecidas para ser aptos para el uso previsto en un proceso de medición.

NOTA 1: RM es un término genérico.

NOTA 2: Las propiedades pueden ser cuantitativas o cualitativas, por ejemplo, identidad de las sustancias o especies.

NOTA 3: Los usos pueden incluir la calibración de un sistema de medición, la evaluación de un procedimiento de medición, valores asignados para otros materiales, y control de calidad.

[FUENTE: Guía ISO 30: 2015, modificada -NOTA 4 se ha eliminado.]

3.18 Material de referencia certificado (CRM)

Material de referencia (RM) caracterizado por un procedimiento metrológicamente válido para una o más propiedades especificadas, acompañados por un certificado RM que proporciona el valor de

la propiedad especificada, su incertidumbre asociada, y una declaración de la trazabilidad metrológica.

NOTA 1: El concepto de valor incluye una propiedad nominal o un atributo cualitativo tales como la identidad o secuencia. Las incertidumbres para tales atributos pueden ser expresadas como probabilidades o niveles de confianza
[FUENTE: Guía ISO 30: 2015, modificada -Notas 2, 3 y 4 se han suprimido.]

4 Principios generales

4.1 Requisitos generales para los métodos estadísticos

4.1.1 Los métodos estadísticos utilizados deberán ser adecuados para los propósitos y estadísticamente válidos. Cualquier hipótesis estadística en la que se basan los métodos o diseños deberán declararse en el diseño o en una descripción escrita del esquema de ensayos de aptitud, y estas hipótesis deberán ser demostradas razonablemente.

NOTA Un método estadísticamente válido tiene una base teórica sólida, ha conocido el rendimiento en las condiciones de uso previstas y se basa en supuestos o condiciones que se pueda demostrar que se aplican a los datos suficientemente bien para el propósito que nos ocupa.

4.1.2 El diseño estadístico y las técnicas de análisis de datos deberán ser consistentes con los objetivos establecidos para el esquema de ensayos de aptitud.

4.1.3 El proveedor de ensayos de aptitud debe proporcionar a los participantes una descripción de los métodos de cálculo utilizados, una explicación de la interpretación general de los resultados, y una declaración de las limitaciones relativas a la interpretación. Esta deberá estar disponible ya sea en cada informe para cada ronda del esquema de ensayos de aptitud o en un resumen independiente de los procedimientos que están disponibles para los participantes.

4.1.4 El proveedor de ensayos de aptitud debe asegurarse de que todo el software esté validado adecuadamente.

4.2 Modelo básico

4.2.1 Para obtener resultados cuantitativos en los esquemas de ensayos de aptitud donde un solo resultado se reporta para un elemento de ensayo de aptitud dado, el modelo básico se da en la ecuación (1).

$$x_i = \mu + \varepsilon_i \quad (1)$$

Dónde

x_i = resultado ensayo de aptitud del participante i

μ = valor verdadero para el mensurando

ε_i = error de medición para el participante i, distribuidos de acuerdo con un modelo pertinente

NOTA 1 Modelos comunes para ε incluyen: la distribución normal $\varepsilon_i \sim N(0, \sigma^2)$ con media 0 y varianza constante o diferente para cada laboratorio; o más comúnmente, una distribución "normal contaminada de valores atípicos" que consiste en una mezcla de una distribución normal con una distribución más amplia que representa la población de resultados erróneos.

NOTA 2 La base de la evaluación de desempeño con puntuaciones z y σ_{pt} es que en una población "idealizada" de laboratorios competentes, la desviación típica inter laboratorios sería opt o menos.

NOTA 3 Este modelo difiere del modelo básico en ISO 5725, en que este no incluye el término de sesgo del laboratorio Bi. Esto es porque el sesgo del laboratorio y los términos de errores residuales no pueden distinguirse cuando solo se reporta una observación. Cuando se consideran los resultados de un participante de varias rondas o artículos de ensayo, sin embargo, puede ser útil incluir un término separado para el sesgo de laboratorio.

4.2.2 Para resultados ordinales o cualitativos, otros modelos pueden ser apropiados, o podría haber un modelo no estadístico.

4.3 Enfoques generales para la evaluación del desempeño

4.3.1 Existen tres enfoques generales diferentes para evaluar el desempeño en un esquema de ensayos de aptitud. Estos enfoques se utilizan para satisfacer diferentes propósitos de un esquema de ensayos de aptitud. Los enfoques se listan a continuación:

- a) desempeño evaluado por comparación con criterios derivados externamente;
- b) desempeño evaluado por comparación con otros participantes;
- c) desempeño evaluado por comparación con la incertidumbre de medición requerida.

4.3.2 Los enfoques generales se pueden aplicar de forma diferente para la determinación del valor asignado y para la determinación de los criterios para la evaluación del desempeño; por ejemplo, cuando el valor asignado es la media robusta de resultados de participantes y la evaluación del desempeño se deriva σ_{pl} o δ_E , donde δ_E es una asignación predefinida para el error de medición y $\sigma_{pl} = \delta_E/3$; del mismo modo, en algunas situaciones el valor asignado puede ser un valor de referencia, pero σ_{pt} puede ser una desviación típica robusta de resultados de los participantes. En el enfoque c) el uso de la incertidumbre de la medición, el valor asignado es típicamente un valor de referencia apropiado.

5. Directrices para el diseño estadístico de los esquemas de ensayos de aptitud

5.1 Introducción al diseño estadístico de los esquemas de ensayos de aptitud

El ensayo de aptitud se ocupa de la evaluación del desempeño de los participantes y, como tal, no aborda específicamente el sesgo o la precisión (aunque éstos pueden ser evaluados con diseños específicos). El desempeño de los participantes se evalúa a través de una evaluación estadística sus resultados después de las mediciones o las interpretaciones que ellos hacen de los elementos del ensayo de aptitud. El desempeño se expresa a menudo en forma de puntuaciones de desempeño que permiten una interpretación coherente a través de una gama de mensurandos y pueden permitir resultados para diferentes mensurandos para ser comparados sobre una misma base. Las puntuaciones de desempeño se derivan típicamente mediante la comparación de la diferencia entre un resultado reportado por un participante y un valor asignado con una desviación permisible o con un estimado de la incertidumbre de la medición de la diferencia. El examen de las puntuaciones de desempeño sobre múltiples rondas de un esquema de ensayos de aptitud puede proporcionar información sobre si los laboratorios individuales muestran evidencia de efectos sistemáticos constantes ("sesgo") o mala precisión a largo plazo.

Las siguientes secciones 5-10 dan orientación sobre el diseño de esquemas cuantitativos de ensayos de aptitud y sobre el tratamiento estadístico de los resultados, incluyendo el cálculo e

interpretación de varias puntuaciones de rendimiento. Consideraciones para los esquemas de ensayos de aptitud cualitativos (incluyendo esquemas ordinales) se dan en la Sección 11.

5.2 Bases de un diseño estadístico

5.2.1 De acuerdo con la norma ISO / IEC 17043, 4.4.4.1, el diseño estadístico "será desarrollado para cumplir los objetivos del esquema de ensayos de aptitud, en base a la naturaleza de los datos (cuantitativos o cualitativos incluyendo ordinales y categóricos), supuestos estadísticos, la naturaleza de los errores, y el número de resultados esperados". Por lo tanto los esquemas de ensayos de aptitud con objetivos diferentes y con diferentes fuentes de error podrían tener diseños diferentes.

Consideraciones de diseño para objetivos comunes se enumeran a continuación. Otros objetivos son posibles.

Ejemplo 1: Para un esquema de ensayos de aptitud para comparar un resultado de un participante contra un valor de referencia predeterminado y dentro de límites que son especificados antes de que comience la ronda, el diseño requerirá un método para la obtención y un valor de referencia externamente definido, un método de fijación de límites, y un método de puntuación;

Ejemplo 2: Para un esquema de ensayos de aptitud para comparar el resultado de un participante con resultados combinados de un grupo en la misma ronda, y límites que se especifican antes de que comience la ronda, el diseño necesitará considerar cómo el valor asignado se determinará a partir de los resultados combinados así como los métodos para la fijación de límites y puntuaciones;

Ejemplo 3: Para un esquema de ensayos de aptitud para comparar el resultado de un participante con resultados combinados de un grupo en la misma ronda, y los límites determinados por la variabilidad de los resultados de los participantes, el diseño necesitará considerar el cálculo de un valor asignado y una medida de dispersión apropiada, así como el método de puntuación;

Ejemplo 4 Para un esquema de ensayos de aptitud para comparar el resultado de un participante con el valor asignado, utilizando la propia incertidumbre de la medición de los participantes, el diseño necesitará considerar cómo se han de obtener el valor asignado y su incertidumbre y cómo las incertidumbres de medición del participante son utilizadas en la puntuación.

Ejemplo 5 Para un esquema de ensayos de aptitud con el objetivo de comparar el desempeño de los diferentes métodos de medición, el diseño deberá tener en cuenta los resúmenes estadísticos y los procedimientos correspondientes para calcularlos.

5.2.2 Hay varios tipos de datos utilizados en los ensayos de aptitud, incluyendo cuantitativa, nominales (categóricos) y ordinales. Entre las variables cuantitativas, algunos resultados podrían estar en una escala de intervalo; o una relativa, o escala de proporción. Para algunas mediciones en una escala cuantitativa, sólo un conjunto discreto y discontinuo de valores se puede realizar (por ejemplo, diluciones secuenciales); Sin embargo, en muchos casos, estos resultados pueden ser tratados mediante técnicas que son aplicables a las variables cuantitativas continuas.

NOTA 1 Para los valores cuantitativos, una escala de intervalo es una escala en la que los intervalos (diferencias) son significativos, pero las proporciones no lo son, tales como la escala de temperatura Celsius. Una escala de proporción es una escala en la que los intervalos y proporciones son significativos, tales como la escala de temperatura Kelvin, o más común las unidades para la longitud.

NOTA 2 Para los valores cualitativos, una escala categórica tiene distintos valores para los que el orden no es significativo, tales como los nombres de las especies bacterianas. Los valores en una escala ordinal

tienen un orden significativo, pero las diferencias no son significativas; por ejemplo, una escala como "grande, mediana, pequeña" se puede ordenar, pero las diferencias entre los valores está indefinido otra cosa que en términos del número de valores intermedios.

5.2.3 Esquemas de ensayos de aptitud pueden ser utilizados para otros fines, además de lo anterior, como se discutió en la sección 0.1 y en la norma ISO / IEC 17043. El diseño deberá ser adecuado para todos los fines establecidos para el esquema de ensayos de aptitud en particular.

5.3 Consideraciones para la distribución estadística de los resultados

5.3.1 ISO / IEC 17043: 2010, 4.4.4.2, exige que las técnicas de análisis estadístico sean consecuentes con los supuestos estadísticos de los datos. Las técnicas de análisis más comunes para los ensayos de aptitud suponen que un conjunto de resultados de participantes competentes tendrá aproximadamente una distribución normal, o al menos unimodal y razonablemente simétrico (después de la transformación, si es necesario). Una suposición común adicional es que la distribución de los resultados de mediciones competentemente determinadas, es mixto (o "contaminado") con resultados provenientes de una población de valores erróneos los cuales pueden generar valores atípicos. Usualmente, la interpretación de puntuación se basa en el supuesto de normalidad, pero sólo por el fundamento de la distribución asumida para participantes competentes.

5.3.1.1 Usualmente, no es necesario verificar que los resultados se distribuyen normalmente, pero es importante verificar la simetría aproximada, al menos visualmente. Si la simetría no se puede verificar entonces el proveedor de ensayos de aptitud debe utilizar técnicas robustas para la asimetría (ver Anexo C).

5.3.1.2 Cuando la distribución esperada para el esquema de ensayos de aptitud no es suficientemente simétrica (teniendo en cuenta la contaminación por los valores atípicos), el proveedor de ensayos de aptitud debe seleccionar métodos de análisis de datos que tengan en cuenta la asimetría esperada y que sean resistentes a los valores atípicos, y métodos de puntuación que también tengan debidamente en cuenta la distribución esperada de los resultados de los participantes competentes. Esto puede incluir:

- transformación para proporcionar simetría aproximada;
- Los métodos de estimación que sean resistentes a la asimetría;
- Métodos de estimación que incorporan supuestos de distribución adecuados (por ejemplo, máxima probabilidad ajustada con los supuestos de distribución adecuados y, si es necesario, el rechazo de valores atípicos).

Ejemplo 1 Los resultados basados en la dilución, como para los conteos microbiológicos cuantitativos o para técnicas de inmunoensayo, a menudo se distribuyen de acuerdo con la distribución normal logarítmica, y así una transformación logarítmica pueden ser apropiada como el primer paso en el análisis.

Ejemplo 2 El Conteo de pequeñas cantidades de partículas pueden ser distribuido de acuerdo a una distribución de Poisson, y por lo tanto los criterios para la evaluación del desempeño se puede determinar usando una tabla de probabilidades de Poisson, basada en el conteo promedio para el grupo de participantes.

5.3.1.3 En algunas zonas de calibración, los resultados de los participantes puede seguir distribuciones estadísticas que se describen en el procedimiento de medición (por ejemplo, exponenciales, o una forma de onda); estas distribuciones definidas deben ser consideradas en cualquier protocolo de evaluación.

5.3.2 De acuerdo con la norma ISO / IEC 17043: 2010, 4.4.4.2, el proveedor de ensayos de aptitud deberá declarar la base para cualquier supuesto estadístico y demostrar que los supuestos son razonables. Esta demostración puede basarse en, por ejemplo, los datos observados, los resultados provenientes de rondas anteriores de esquema de ensayos de aptitud, o la literatura técnica.

NOTA La demostración de la razonabilidad de un supuesto de distribución es menos rigurosa que la demostración de la validez de esa suposición.

5.4 Consideraciones para un pequeño número de participantes

5.4.1 El diseño estadístico para un esquema de ensayos de aptitud debe considerar el número mínimo de participantes que se necesitan para cumplir con los objetivos del diseño, y la declaración de los enfoques alternativos que se utilizarán si el número mínimo no se logra (ISO / IEC 17043: 2010, 4.4.4.3b)). Los métodos estadísticos que son apropiados para un gran número de participantes pueden no ser apropiados con un número limitado de participantes. La preocupación es que las estadísticas determinadas a partir de un pequeño número de resultados de participantes pueden no ser lo suficientemente confiable, y un participante se pudiera evaluar contra un grupo de comparación inapropiado.

NOTA El Informe Técnico IUPAC/CITAC: Selección y uso de los esquemas de ensayos de aptitud para un número limitado de participantes [24] proporciona una guía útil para los esquemas de ensayos de aptitud donde existen pocos participantes. En resumen, el informe de la IUPAC / CITAC recomienda que el valor asignado debe basarse en mediciones independientes fiables; por ejemplo, mediante el uso de un material de referencia certificado, asignación independiente mediante una calibración o instituto nacional de metrología, o por preparación gravimétrica. El informe señala además que la desviación típica para la evaluación de la competencia no puede estar basada en la dispersión observada entre los resultados de participantes para una sola ronda de un esquema de ensayos de aptitud.

5.4.2 El número mínimo de participantes necesarios para los diversos métodos estadísticos dependerá de una variedad de situaciones:

- Los métodos estadísticos utilizados, por ejemplo, el método robusto en particular o estrategia de eliminación de valores atípicos se ha seleccionado;
- La experiencia de los participantes con el esquema de ensayos de aptitud en particular;
- La experiencia del proveedor de ensayos de aptitud con la matriz, mensurando, métodos, y el grupo de participantes;
- Si la intención es determinar el valor asignado o la desviación típica (o ambos).

Más orientación sobre técnicas para el manejo de un pequeño número de participantes se proporciona en el anexo D.1.

5.5 Pautas para elegir el formato de presentación de informes

5.5.1 Es un requisito de la norma ISO / IEC 17043: 2010, 4.6.1.2, que los proveedores de ensayos de aptitud instruyan a los participantes para llevar a cabo las mediciones y reportar resultados sobre los elementos de la ensayo de aptitud en la misma forma que para la mayoría de las mediciones realizadas de forma rutinaria, excepto en circunstancias especiales.

Este requisito puede, en algunos casos, hacer difícil la obtención de una evaluación exacta de la precisión y la veracidad de los participantes, o la competencia con un procedimiento de medición.

El proveedor de ensayos de aptitud debe adoptar un formato de informe consistente para el esquema de ensayos de aptitud, pero debe, en lo posible, utilizar las unidades familiares para la mayoría de los participantes y elegir un formato de informe que minimice la transcripción y otros errores. Esto puede incluir la advertencia automatizada de las unidades inapropiadas cuando se conocen los participantes para reportar rutinariamente en unidades distintas de las requeridas por el esquema.

NOTA 1: Para algunos esquemas de ensayos de aptitud, un objetivo es evaluar la habilidad de un participante para seguir un método estándar, lo que podría incluir el uso de una unidad particular de medida o el número de dígitos significativos.

NOTA 2: Errores de transcripción en la comparación de los resultados por el proveedor de ensayos de aptitud pueden ser sustancialmente reducidos o eliminados por el uso de sistemas electrónicos de reporte que permitan a los participantes a entrar directamente sus propios datos.

5.5.2 Si un esquema de ensayos de aptitud requiere replicar mediciones sobre elementos de ensayo de aptitud, el participante debe exigir reportar todos los valores replicados. Esto puede ocurrir, por ejemplo, si un objetivo es evaluar la precisión de un participante sobre elementos de ensayo de aptitud de réplica conocida, o cuando un procedimiento de medición requiere el reporte de informes por separado de múltiples observaciones. En estas situaciones, el proveedor de ensayos de aptitud también puede necesitar pedir el valor de la media a los participantes (u otra estimación de la ubicación) y la incertidumbre para ayudar el análisis de datos por el proveedor de ensayos de aptitud.

5.5.3 Donde el reporte práctico convencional es para reportar resultados como "menos" o "mayor que" un límite (como un nivel de calibración o un límite de cuantificación) y donde se requieren resultados numéricos para la puntuación, el proveedor de ensayos de aptitud debe determinar cómo los resultados serán procesados.

5.5.3.1 El proveedor de ensayos debe de cualquier modo adoptar procedimientos de tratamiento de datos y de puntuación validados que se adapten a los datos censurados (ver Anexo E.1), o requerir que los participantes reporten el valor medido del resultado, ya sea en lugar de, o además de, el valor convencionalmente reportado.

NOTA 1: Una opción de procedimiento de puntuación podría ser la de no registrar tales datos.

NOTA 2: Exigir a los participantes reportar valores numéricos fuera del rango normalmente informado (por ejemplo, por debajo del límite de cuantificación del participante) permitirá el uso de métodos estadísticos que requieren valores numéricos, pero pueden dar lugar a resultados que no reflejan el servicio de rutina de los participantes a los clientes.

5.5.3.2 Cuando se utiliza el consenso estadístico, tal vez no sea posible evaluar el desempeño si el número de valores censurados es lo suficientemente grande, un método robusto se ve afectado por la censura. En circunstancias donde el número de resultados censurados es suficiente para afectar a un método robusto, entonces, los resultados deben ser evaluados utilizando métodos estadísticos que permiten la estimación imparcial, en presencia de datos censurados^[21], o los resultados no deben ser evaluados. En caso de duda sobre el efecto del procedimiento elegido, el proveedor de ensayos de aptitud debe calcular un resumen estadístico y las evaluaciones de desempeño con cada uno de los procedimientos estadísticos alternativos considerados potencialmente aplicables en las circunstancias, e investigar la importancia de cualquier diferencia (s).

5.5.3.3 Donde los resultados censurados como 'menor que' de lo que espera en los informes o han sido observados, el diseño de ensayos de aptitud debe incluir disposiciones para la puntuación y / u otra acción sobre los valores censurados reportados por los participantes, y los participantes deben ser notificados de estas disposiciones.

NOTA Anexo E.1: tiene un ejemplo de algunos de los enfoques para el análisis de datos censurados. Este ejemplo muestra consenso estadístico robusto; con tres diferentes enfoques; con los valores censurados eliminados; con los valores censurados retenidos pero el '<' signo eliminado, y con los resultados reemplazados por medio del valor límite.

5.5.4 Usualmente, el número de dígitos significativos a reportar será determinado por el diseño del esquema de ensayos de aptitud.

5.5.4.1 Cuando se especifique el número de dígitos significativos para ser reportados, el error de redondeo debe ser insignificante en comparación con la variación esperada entre los participantes.

NOTA: En algunas situaciones, la presentación correcta del informe es parte de la determinación de la competencia de los participantes, y el número de dígitos significativos y decimales puede variar.

5.5.4.2 Cuando el número de dígitos reportado bajo condiciones de medición rutinarias tiene un efecto adverso apreciable sobre el tratamiento de datos por el proveedor de ensayos de aptitud (por ejemplo, donde los procedimientos de medición requieren que se informe un pequeño número de dígitos significativos), el proveedor de ensayos de aptitud puede especificar el número de dígitos a ser informado.

Ejemplo: Un procedimiento de medición podría especificar la presentación de informes a 0,1 g, dando lugar a una gran proporción (> 50%) de resultados idénticos y, hacer que se tenga que calcular medias robustas y desviaciones típicas. El proveedor de ensayos de aptitud podría entonces exigir a los participantes que informen a dos o tres cifras decimales para obtener estimaciones suficientemente fiables de la ubicación y la variación.

5.5.4.3 Si se permite que los diferentes participantes reporten resultados utilizando diferentes números de dígitos significativos, el proveedor de ensayos de aptitud debe tener esto en consideración cuando genera cualquier estadística de consenso (como el valor asignado y la desviación típica para la evaluación del desempeño).

6 Directrices para la revisión inicial de los elementos de ensayos de aptitud y los resultados

6.1 Homogeneidad y estabilidad de los elementos de ensayo de aptitud

6.1.1 El proveedor de ensayos de aptitud debe asegurarse de que los lotes de elementos de ensayo de aptitud sean lo suficientemente homogéneos y estables para los propósitos del esquema de ensayos de aptitud. El proveedor deberá asegurar la homogeneidad y la estabilidad

mediante el uso de criterios que aseguren que la falta de homogeneidad y la inestabilidad de los elementos de ensayo de aptitud no afectan negativamente la evaluación del desempeño. Para la evaluación de la homogeneidad y la estabilidad se deben utilizar uno o más de los siguientes enfoques:

- a) estudios experimentales que se describen en el Anexo B o métodos experimentales alternativos que ofrecen una confianza equivalente o mayor de homogeneidad y estabilidad;
- b) experiencia con el comportamiento de los elementos de ensayo de aptitud muy similares en rondas anteriores del esquema de ensayos de aptitud, verifique si es necesario para la ronda actual;
- c) evaluación de los datos de los participantes en la ronda actual del esquema de ensayos de aptitud para evidenciar la coherencia con las rondas anteriores, para evidenciar el cambio con el tiempo de reporte o el método de producción, o cualquier dispersión inesperada atribuible a la falta de homogeneidad o inestabilidad.

NOTA 1: Estos enfoques se pueden adoptar sobre la base de un caso por caso, usando técnicas estadísticas apropiadas y justificación técnica. El enfoque a menudo cambiará durante la vida útil de un esquema de ensayos de aptitud, por ejemplo, como la experiencia acumulada reduce el requisito inicial para el estudio experimental.

NOTA 2: Contando con la experiencia (como en b anterior) es solamente razonable, siempre y cuando:

1. El procedimiento para producir lotes del elemento (s) de ensayo de aptitud no cambia en ninguna manera que pueda tener un impacto sobre la homogeneidad;
2. Los materiales utilizados en la producción del elemento de ensayo de aptitud (s) no cambian de ninguna manera que pueda tener impacto en la homogeneidad;
3. No existe un "fallo" en la homogeneidad identificado ya sea a través de pruebas de homogeneidad o respuestas de un participante; y,
4. Los requisitos de homogeneidad para el material sean revisados periódicamente, teniendo en cuenta el uso previsto del material en el momento de la revisión, para asegurar que la homogeneidad alcanzada por el proceso de producción sigue siendo apta para el fin previsto.

Ejemplo Si rondas anteriores de un esquema de ensayos de aptitud utilizaron elementos de ensayo de aptitud que se ensayaron y demostraron ser suficientemente homogéneos y estables, y con los mismos participantes como en las rondas anteriores, entonces si una desviación típica inter laboratorios en la ronda actual no es mayor que el desviación típica en las rondas anteriores, existen evidencias de homogeneidad y estabilidad en la ronda actual.

6.1.2 Para esquemas de ensayos de aptitud de calibración donde el mismo objeto es utilizado por múltiples participantes, el proveedor de ensayos de aptitud debe asegurar la estabilidad durante la ronda, o debe tener procedimientos para identificar y cuantificar la inestabilidad a través de la progresión de una ronda del esquema de ensayos de aptitud. Esto debería incluir la consideración de las tendencias para un elemento de ensayo de aptitud en particular y mensurando, tal como la deriva. Donde sea apropiado, el aseguramiento de la estabilidad debe considerar los efectos de múltiples envíos de un mismo objeto.

6.1.3 A todos los mensurandos (o propiedades) normalmente se les deben comprobar la homogeneidad y la estabilidad. Sin embargo, cuando el comportamiento de un subconjunto de propiedades puede ser mostrado para proporcionar una buena indicación de la estabilidad y / o la homogeneidad para todas las propiedades reportadas sobre una ronda, la evaluación descrita en la sección 6.1.1 se puede limitar a ese subconjunto de propiedades. Los mensurandos que se comprueban deben ser sensibles a fuentes de falta de homogeneidad o inestabilidad en el proceso del elemento de ensayos de aptitud. Algunos casos importantes son:

- a) cuando la medición es una proporción, una característica que es una pequeña proporción puede ser más difícil para homogeneizar y así es más sensibles en una comprobación de la homogeneidad;
- b) si un elemento de ensayo de aptitud se calienta durante el proceso, entonces seleccione un mensurando que sea sensible al calentamiento designial;
- c) si una propiedad medida puede verse afectada por la sedimentación, la precipitación, u otros efectos dependientes de la duración del proceso de preparación de los elementos del ensayo de aptitud, entonces, esta propiedad debe ser chequeada a través del proceso completo.

Ejemplo: En un esquema de ensayos de aptitud para el contenido de metales tóxicos en los suelos, la medida del contenido de metal se ve afectada principalmente por el contenido de humedad. Un chequeo del contenido de humedad puede entonces ser considerado suficiente para garantizar una adecuada estabilidad de los metales tóxicos.

NOTA Un ejemplo de los controles de homogeneidad y estabilidad se proporciona en el anexo E.2, utilizando métodos estadísticos recomendados en el Anexo B.

6.2 Consideraciones para diferentes métodos de medición

6.2.1 Cuando se espera que todos los participantes reporten un valor para el mismo mensurando, el valor asignado normalmente debe ser el mismo para todos los participantes. Sin embargo, cuando a los participantes se les permite elegir su propio método de medición, es posible que un solo valor asignado para cada analito o propiedad pueda no ser apropiado para todos los participantes. Esto puede ocurrir, por ejemplo, cuando los diferentes métodos de medición producen resultados que no son comparables. En este caso, el proveedor de ensayos de aptitud puede utilizar un valor asignado diferente para cada método de medición.

Ejemplos:

a) ensayos médicos, donde diferentes métodos de medición aprobados se conoce que responden de manera diferente sobre el mismo material de ensayo y se utilizan diferentes rangos de referencia para el diagnóstico;

b) mensurandos operacionalmente definidos, como los metales tóxicos lixiviados en suelos, para que los cuales diferentes métodos típica están disponibles y no se espera que sean directamente comparables, pero donde el esquema de ensayos de aptitud especifica el mensurando sin referencia a un método de ensayo específico.

6.2.2 La necesidad de diferentes valores asignados para los subgrupos de los participantes deben ser considerados en el diseño del esquema de ensayos de aptitud (por ejemplo, para hacer provisión para la presentación de informes de métodos específicos) y también deben ser considerados en la revisión de datos de cada ronda.

6.3 Eliminación de errores groseros

6.3.1 ISO / IEC 17043: 2010, B.2.5 y el Protocolo Armonizado IUPAC recomienda eliminar los errores groseros obvios de un conjunto de datos en una etapa temprana del análisis, antes del uso de cualquier procedimiento robusto o cualquier ensayo para identificar valores estadísticos atípicos. Generalmente, estos resultados deberían tratarse por separado (por ejemplo contactando con el participante). Puede ser posible corregir algunos errores groseros, pero esto sólo debe hacerse de acuerdo con una política y un procedimiento aprobado.

NOTA errores groseros obvios, como informar los resultados en las unidades incorrectas o cambiar los resultados de diferentes elementos de ensayo de aptitud, se producen en la mayoría de las rondas de ensayos de aptitud, y estos resultados sólo perjudican el funcionamiento de los métodos estadísticos posteriores.

6.3.2 Si hay alguna duda sobre si un resultado es un error grosero, este debe mantenerse en el conjunto de datos y se somete a un tratamiento posterior, tal como se describe en las secciones de 6.4 a 6.6.

6.4 Revisión visual de los datos

6.4.1 Como primer paso en cualquier análisis de datos el proveedor debe organizar la revisión visual de los datos, realizada por una persona que tiene una experiencia técnica y estadística adecuada. Esta verificación es para confirmar la distribución esperada de los resultados, y para identificar anomalías o fuentes no previstas de variabilidad. Por ejemplo, una distribución bimodal podría ser evidencia de una población mixta de resultados causados por diferentes métodos, muestras contaminadas o instrucciones mal redactadas. En esta situación, la preocupación debe ser resuelta antes de proceder con el análisis o evaluación.

NOTA 1: Un histograma es un procedimiento de revisión útil y ampliamente disponible, para buscar una distribución que es unimodal y simétrica, y para identificar valores atípicos inusuales (sección 10.2). Sin embargo, los intervalos utilizados para combinar los resultados en un histograma son sensibles a los números de los resultados y puntos de corte, y por lo tanto puede ser difícil de crear. Un gráfico de densidad kernel suele ser más útil para identificar posibles bimodalidades o falta de simetría (sección 10.3).

NOTA 2: Otras técnicas de revisión pueden ser útiles, por ejemplo, un gráfico de distribución acumulativa o un diagrama de tallo y hoja. Algunos métodos gráficos para la revisión de datos se ilustran en los anexos E.3 y E.4.

6.4.2 Cuando no es posible llevar a cabo revisión visual de todos los conjuntos de datos de interés, se procederá a un procedimiento para advertir de la variabilidad inesperada en un conjunto de datos; por ejemplo, mediante la revisión de la incertidumbre del valor asignado en comparación con los criterios de evaluación, o por comparación con las rondas anteriores del programa de ensayos de aptitud.

6.4.3 Cuando no es posible llevar a cabo revisión visual de todos los conjuntos de datos de interés, se procederá a advertir la variabilidad inesperada en un conjunto de datos; por ejemplo, mediante la revisión de la incertidumbre del valor asignado en comparación con los criterios de evaluación, o por comparación con las rondas anteriores del esquema de ensayos de aptitud.

6.5 Métodos estadísticos robustos

6.5.1 Los métodos estadísticos robustos pueden ser usados para describir la parte central de un conjunto de resultados distribuido normalmente, pero sin requerir la identificación de valores específicos como valores atípicos y excluirlos de los análisis posteriores. Muchas técnicas robustas utilizadas se basan (en el primer paso) en la mediana y el rango central del 50% de los resultados – estas son medida del centro y de la dispersión de los datos, similar a la media y la desviación típica. En general, los métodos robustos deben ser utilizados en preferencia a los métodos que eliminan los resultados etiquetados como valores atípicos.

NOTA: Las estrategias que aplican estadística clásica, tales como la desviación típica después de eliminar valores extremos por lo general conducen a menores estimaciones de dispersión para los datos cercanos a la distribución normal; las estadísticas robustas se ajustan generalmente para dar estimaciones imparciales de dispersión.

6.5.2 La mediana, la desviación absoluta mediana escalada (MAD_e), y normalizado IQR ($nIQR$) son permitidos como estimadores simples. El Algoritmo A transforma los datos originales por un proceso llamado “*winsorización*” para proporcionar estimadores alternativos de media y desviación típica para los datos casi normales y es más útil donde la proporción esperada de valores atípicos está por debajo de 20%. Los métodos Qn y Q (descritos en el anexo C) para la estimación de la desviación típica son particularmente útiles para situaciones en las que una proporción grande (> 20%) de resultados pueden discrepar, o cuando los datos no pueden ser revisadas de forma fiable por los expertos. Otros métodos descritos en el Anexo C también proporcionan un buen rendimiento cuando la proporción esperada de los valores extremos es más del 20% (véase el anexo D).

NOTA: La mediana, el rango inter-cuartílico y la desviación absoluta mediana escalada tienen varianza mayor que la media y la desviación típica cuando se aplica a los datos aproximadamente una distribuidos de forma normal. Estimadores robustos más sofisticados proporcionan un mejor desempeño para los datos de aproximadamente una distribución normal, conservando gran parte de la resistencia a los resultados de valores atípicos que es ofrecido por la mediana y el rango inter-cuartílico.

6.5.3 La elección de los métodos estadísticos es la responsabilidad del proveedor de ensayos de aptitud. La media y la desviación típica robusta se puede utilizar para diversos fines, de las cuales la evaluación del desempeño es sólo uno. La media robusta y la desviación típica también se pueden utilizar como resumen estadístico para diferentes grupos de participantes o para métodos específicos.

NOTA Los detalles para procedimientos robustos se proporcionan en el Anexo C. Anexos E.3 y E.4 tienen ejemplos detallados que ilustran el uso de una variedad de técnicas estadísticas robustas presentadas en el Anexo C.

6.6 Técnicas de valores atípicos para resultados individuales

6.6.1 Las pruebas de valores atípicos pueden utilizarse para apoyar el examen visual para anomalías o, junto con el rechazo de los valores atípicos, para proporcionar un grado de resistencia a los valores extremos en el cálculo de resúmenes estadísticos. Cuando se utilicen técnicas de detección de valores atípicos, los supuestos subyacentes de la prueba deben ser demostrados para aplicar suficientemente para el propósito del programa de ensayos de aptitud; En particular, muchas pruebas para valores atípicos asumen la normalidad subyacente.

NOTA: ISO 16269 4 [10] y la ISO 5725 2 [1] proporcionan varios procedimientos de identificación de valores atípicos que son aplicables a los datos inter laboratorios.

6.6.2 Las estrategias de rechazo de valores atípicos, que se basan en el rechazo de los valores atípicos detectados por una prueba de valor atípico con un alto nivel de confianza, seguido por la aplicación de estadísticas simples, como la media y la desviación típica, son permitidas donde los métodos robustos no son aplicables (véase 6.5.1). Cuando se utilizan las estrategias de rechazo de atípicos, el proveedor de ensayos de aptitud debe:

- a) documentar las pruebas y el nivel de confianza requerido para el rechazo;
- b) establecer límites para el porcentaje de datos rechazados por las pruebas de valores atípicos sucesivas, si son usadas;
- c) demostrar que las estimaciones de ubicación resultantes y (si es apropiado) escalada en su caso tienen un desempeño suficiente (incluida la eficiencia y sesgo) para los propósitos del programa de ensayos de aptitud.

NOTA: ISO 5725-2 proporciona recomendaciones para el nivel de confianza adecuado para el rechazo de valores atípicos en estudios entre laboratorios para la determinación de la precisión de los métodos de ensayo. En particular, la norma ISO 5725-2 recomienda el rechazo sólo en el nivel del 99% a menos que haya otra razón de peso para rechazar un resultado particular.

6.6.3 Cuando el rechazo atípico es parte de un procedimiento de manejo de datos, y el resultado es eliminado como un valor atípico, el desempeño del participante aún debe ser evaluado de acuerdo a los criterios utilizados para todos los participantes en el programa de ensayos de aptitud.

NOTA 1: Los valores atípicos entre los valores reportados son a menudo identificados por el empleo de la prueba de Grubbs para valores atípicos, como se indica en la norma ISO 5725-2. La evaluación en este procedimiento es aplicada usando la desviación típica de todos los participantes, incluidos los valores atípicos potenciales. Por lo tanto este procedimiento se debe aplicar cuando el desempeño de los participantes es consistente con las expectativas de rondas anteriores y hay un pequeño número de valores atípicos (uno o dos valores atípicos en cada lado de la media). Las tablas convencionales para el procedimiento Grubbs asumen una única solicitud para un posible valor atípico (o 2) en una ubicación definida, no a ilimitada solicitud secuencial. Si las tablas de Grubbs se aplican de forma secuencial, el error tipo I probabilidades para el ensayo puede no aplicarse.

NOTA 2: Cuando los resultados replicados son devueltos o elementos de la ensayo de aptitud idénticos se incluyen en una ronda de un programa de ensayos de aptitud, es común el uso de la prueba de Cochran para valores atípicos de repetibilidad, también se describen en la norma ISO 5725-2.

NOTA 3: Los valores atípicos también se pueden identificar mediante técnicas robustas o no paramétricas; por ejemplo, si una media robusta y la desviación típica son calculadas los valores de desviación de la media robusta por más de 3 veces la desviación típica robusta podría ser identificado como valores atípicos.

7 Determinación del valor asignado y su incertidumbre típica

7.1 Elección del método de determinación del valor asignado

7.1.1 Cinco maneras de determinar el valor asignado Xpt se describen en las secciones 7.3 a 7.7. La elección entre estos métodos es la responsabilidad del proveedor de ensayos de aptitud.

NOTA: Las Secciones 7.3-7.6 son muy similares a los enfoques utilizados para determinar los valores de las propiedades de los materiales de referencia certificados descritos en la Guía ISO 35 [13].

7.1.2 Métodos alternativos para la determinación del valor asignado y su incertidumbre pueden ser utilizados siempre que tengan una base estadística sólida y que el método utilizado se describa en el plan documentado para el programa de ensayos de aptitud, y describen completamente a los participantes. Independientemente del método utilizado para determinar el valor asignado, siempre es adecuado verificar la validez del valor asignado para una ronda de programa de ensayos de aptitud. Esto se discute en la sección 7.8.

7.1.3 Métodos para la determinación de los valores cualitativos asignados se discuten en la sección 11.3.

7.1.4 El método para determinar el valor asignado y su incertidumbre asociada se indicará en cada informe a los participantes o claramente descrito en un esquema de protocolo disponible para todos los participantes.

7.2 Determinación de la incertidumbre del valor asignado

7.2.1 La Guía para la expresión de la incertidumbre en la medición (Guía ISO / IEC 98 3^[14]) da orientación sobre la evaluación de las incertidumbres de medición. La Guía ISO 35 proporciona orientación sobre la incertidumbre del valor asignado para el valor de las propiedades certificadas, que se puede aplicar para muchos diseños de ensayos de aptitud.

7.2.2 Un modelo general para el valor asignado y su incertidumbre se describe en las ecuaciones (2) y (3):

El modelo para el valor asignado puede ser expresado a través de la ecuación:

$$x_{pt} = x_{char} + \delta_{hom} + \delta_{trans} + \delta_{stab} \quad (2)$$

donde:

x_{pt} indica el valor asignado;

x_{char} el valor de la propiedad obtenida a partir de la caracterización (determinación del valor asignado);

δ_{hom} indica un término de error debido a la diferencia entre los elementos de ensayo de aptitud;

δ_{trans} indica un término de error debido a la inestabilidad en las condiciones de transporte;

δ_{stab} indica un término de error debido a la inestabilidad durante el período de ensayos de aptitud.

El modelo asociado a la incertidumbre del valor asignado puede ser expresado a través de la ecuación:

$$u(x_{pt}) = \sqrt{u_{char}^2 + u_{hom}^2 + u_{trans}^2 + u_{stab}^2} \quad (3)$$

donde:

$u(x_{pt})$ indica la incertidumbre típica del valor asignado;

u_{char} indica la incertidumbre típica debido a la caracterización;

u_{hom} indica la incertidumbre típica debido a las diferencias entre los elementos de ensayo de aptitud;

u_{trans} indica la incertidumbre típica debido a la inestabilidad causada por el transporte de los elementos de ensayo de aptitud;

u_{stab} indica la incertidumbre típica debido a la inestabilidad durante el período de ensayos de aptitud.

NOTA 1: La covarianza entre las fuentes de incertidumbre, o fuentes insignificantes, puede dar lugar a un modelo diferente para aplicaciones específicas. Cualquiera de los componentes de la incertidumbre puede ser cero o insignificante, en algunas situaciones.

NOTA 2: Cuando σ_{pt} se calcula como la desviación típica de los resultados de los participantes, los componentes de la incertidumbre debido a la falta de homogeneidad, el transporte, y la inestabilidad reflejan en gran parte la variabilidad de los resultados de los participantes. En este caso la incertidumbre de caracterización, como se describe en las secciones 7.3-7.7, es suficiente.

NOTA 3: El proveedor de ensayos de aptitud espera normalmente para asegurar que los cambios relacionados con la inestabilidad o incurridos en el transporte sean insignificantes en comparación con la desviación típica para la evaluación de aptitud; es decir, para garantizar que δ_{trans} y δ_{stab} sean insignificantes. En caso de que se cumpla este requisito, u_{stab} y u_{trans} pueden ser puestos a cero.

7.2.3 No puede haber sesgo en el valor asignado que no se contabiliza en la expresión anterior. Esto deberá, en lo posible, ser considerado en el diseño para el programa de ensayos de aptitud. Si hay un ajuste de sesgo en el valor asignado, la incertidumbre de este ajuste se incluirá en la evaluación de la incertidumbre del valor asignado.

7.3 Formulación

7.3.1 El elemento de ensayo de aptitud se puede preparar por materiales con diferentes niveles conocidos de una propiedad en proporciones especificadas de mezcla, o mediante la adición de una proporción específica de una sustancia a un material de base.

7.3.1.1 El valor asignado x_{pt} se deriva mediante el cálculo de las masas de propiedades usadas. Este enfoque es especialmente valioso cuando los elementos de ensayo de aptitud individuales se preparan de esta manera, y es la proporción de las propiedades que se va a determinar.

7.3.1.2 Se debe de tener cuidado razonable para asegurarse de que:

a) el material de base es efectivamente libre de la constituyente añadido, o que la proporción del componente añadido en el material de base se conoce con precisión;

b) los componentes se mezclan de forma homogénea (cuando así lo requieran);

c) se identifican todas las fuentes significativas de error (por ejemplo, no es siempre se dio cuenta de que el vidrio absorbe compuestos de mercurio, de modo que la concentración de una solución acuosa de un compuesto de mercurio puede ser alterada por su contenedor);

d) no hay ninguna interacción adversa entre los componentes y la matriz;

e) el comportamiento de los elementos de prueba de aptitud que contienen materiales añadidos es similar a las muestras de los clientes que son probados con regularidad. Por ejemplo, los materiales puros añadidos a una matriz natural se extraen a menudo del material con mayor facilidad que la misma sustancia de origen natural.

7.3.1.3 Cuando la formulación proporciona objetos de la prueba de aptitud en el que la adición está más mal adherida que en las muestras analizadas de manera rutinaria, o en una forma diferente, puede ser preferible utilizar otro método para preparar los objetos de la prueba de aptitud.

7.3.1.4 Determinar el valor asignado por la formulación es un caso de enfoque general para la caracterización de materiales de referencia certificados descritos por la Guía ISO 35, donde un único laboratorio determina de un método de medición primaria un valor asignado. Otros usos de un método principal por un solo laboratorio pueden ser utilizados para determinar el valor asignado para pruebas de competencia (véase la sección 7.5).

7.3.2 Cuando el valor asignado se calcula a partir de la formulación del producto de ensayo de aptitud, la incertidumbre típica para la caracterización (u_{char}) se calcula mediante la combinación de incertidumbre utilizando un modelo apropiado. Por ejemplo, en ensayos de aptitud para mediciones químicas de las incertidumbres por lo general se les asocia con las mediciones gravimétricas y volumétricas y la pureza de los materiales utilizados en la formulación. La incertidumbre típica del valor asignado ($u(x_{pt})$) se calcula entonces según la ecuación (3).

7.4 Material de referencia certificado

7.4.1 Cuando un elemento de prueba de aptitud es un material de referencia certificado (CRM), el valor de la propiedad certificada x_{CRM} es usado como el valor asignado x_{pt} . Las limitaciones de este enfoque son las siguientes:

- Puede ser costoso para proporcionar a cada participante una unidad de un material de referencia certificado;
- Los CRMs son a menudo procesados muy fuertemente para garantizar la estabilidad a largo plazo, lo que puede comprometer la comutabilidad de los elementos de la prueba de aptitud.
- Un CRM puede ser conocido por los participantes por lo que es importante ocultar la identidad del producto de ensayo de aptitud.

7.4.2 Cuando se utiliza un material de referencia certificado como el elemento de la prueba de aptitud, la incertidumbre típica del valor asignado se deriva de la información sobre la incertidumbre del valor del inmueble indicado en el certificado. La información del certificado debe incluir los componentes de la ecuación (3), y tienen un uso previsto apropiado para el propósito del programa de ensayos de aptitud.

7.5 Resultados de un laboratorio

7.5.1 Un valor asignado puede ser determinado por un único laboratorio usando un método de referencia, tal como, por ejemplo, un método primario. El método de referencia utilizado deberá

describirse completamente y comprendido, y con una declaración de incertidumbre y trazabilidad metrológica documentada que es apropiada para el programa de ensayos de aptitud. El método de referencia debe ser comutable para todos los métodos de medición utilizados por los participantes.

7.5.1.1 El valor asignado debe ser el promedio de un estudio diseñado utilizando la opción de más de un dominio de prueba o condiciones de medición, y un número suficiente de mediciones repetidas.

7.5.1.2 La incertidumbre de caracterización es la estimación apropiada de la incertidumbre para el método de referencia y las condiciones de estudio diseñada.

7.5.2 El valor asignado x_{pt} del objeto de ensayo de aptitud puede ser derivado por un solo laboratorio que utiliza un método de medición adecuado, a partir de una calibración frente a los valores de referencia de un material de referencia certificado. Este enfoque supone que el CRM es comutable para todos los métodos de medición utilizados por los participantes.

7.5.2.1 Esta determinación requiere una serie de pruebas que se llevarán a cabo, en un laboratorio, a elementos de ensayo de aptitud y al CRM, utilizando el mismo método de medición, y en condiciones de repetibilidad. Cuando:

x_{CRM} es el valor asignado por el CRM

x_{pt} es el valor asignado para el elemento de prueba de aptitud

d_i es la diferencia entre los resultados promedio para el elemento de prueba de aptitud y las muestras i^{th} del CRM

\bar{d} es el promedio de las diferencias d_i

Entonces,

$$x_{pt} = x_{CRM} + \bar{d} \quad (4)$$

NOTA: x_{CRM} y \bar{d} son independientes, salvo en raras ocasiones, cuando el laboratorio experto también produce el CRM.

7.5.2.2 La incertidumbre típica de caracterización se deriva de la incertidumbre de la medida utilizada para la asignación del valor. Este enfoque permite que el valor asignado a establecerse sea de una manera metrológicamente trazable al valor certificado del CRM, con una incertidumbre típica que se puede calcular a partir de la ecuación (5).

$$U_{char} = \sqrt{u_{CRM}^2 + u_d^2} \quad (5)$$

El ejemplo en el Anexo E5 ilustra cómo la incertidumbre requerida se puede calcular en un caso sencillo cuando el valor asignado de un elemento de prueba de competencia se establece por comparación directa con un único CRM.

7.5.3 Cuando se asigna un valor de referencia antes del comienzo de una partida de esquema secuencial de ensayo de aptitud, y luego el valor de referencia se comprueba posteriormente

utilizando el mismo sistema de medición, la diferencia entre los valores deberá ser inferior a dos veces la incertidumbre de esa diferencia (es decir, los resultados deberán ser compatibles vista metrológica). En tales casos, el proveedor de ensayos de aptitud puede optar por utilizar un promedio de las mediciones como el valor asignado, con la incertidumbre apropiada. Si los resultados no son compatibles vista metrológica, el proveedor de ensayos de aptitud debe investigar la razón de la diferencia, y tomar las medidas apropiadas, incluyendo el uso de métodos alternativos para determinar el valor asignado y su incertidumbre o el abandono de la ronda.

7.6 Valor de consenso de los laboratorios especializados

7.6.1 Los valores asignados se pueden determinar mediante un estudio de comparación entre laboratorios con laboratorios especializados, tal como se describe en la Guía ISO 35 para su uso en comparaciones entre laboratorios para caracterizar un CRM. Se preparan primero los objetos de la prueba de aptitud y se distribuyen a los participantes. A continuación, algunos de estos objetos de la prueba de aptitud, se seleccionan al azar y son analizados por un grupo de expertos usando un protocolo que especifica el número de elementos y repeticiones de ensayos de aptitud y demás condiciones pertinentes. Se requiere que cada laboratorio experto para proporcionar una incertidumbre típica con sus resultados.

7.6.2 Cuando los laboratorios expertos informan un solo resultado y no son requeridos por el protocolo de medición para proporcionar suficiente información sobre la incertidumbre de los resultados, o donde la evidencia de los resultados reportados o en otro lugar sugiere que las incertidumbres informadas no son suficientemente fiables, el valor de consenso debe normalmente obtenerse por los métodos de la sección 7.7, aplicados al conjunto de los resultados de laboratorio de los expertos. Cuando los laboratorios expertos informan de más de un resultado cada uno (por ejemplo, incluyendo repeticiones), el proveedor de ensayos de aptitud debe establecer un método alternativo para la determinación del valor asignado y la incertidumbre asociada que es estadísticamente válida (ver 4.1.1) y permite la posibilidad de valores atípicos u otras desviaciones de la distribución esperada de los resultados.

7.6.3 Cuando los laboratorios expertos informan incertidumbres con los resultados, la estimación de un valor por consenso de los resultados es un problema complejo y una amplia variedad de enfoques se ha sugerido, incluyendo, por ejemplo, promedios ponderados, promedios no ponderados, los procedimientos que la asignación de maquillaje por más de dispersión y procedimientos que permitan posibles resultados erróneos o periféricas y estimaciones de la incertidumbre [16]. El proveedor de ensayos de aptitud debe establecer en consecuencia un procedimiento para la estimación que:

a) debe incluir el control de validez de las estimaciones de las incertidumbres declaradas, por ejemplo, mediante la comprobación plena de las incertidumbres informadas y la dispersión observada de los resultados;

b) debe utilizar un procedimiento de ponderación apropiada para la escala y la fiabilidad de las incertidumbres informadas, que se podrán incluir en la misma ponderación si las incertidumbres reportadas son similares o malas o de desconocido fiabilidad (véase 7.6.2);

c) debe tener en cuenta la posibilidad de que informó incertidumbres podrían no tener plenamente en cuenta la dispersión observada ("sobre la dispersión"), por ejemplo mediante la inclusión de un término adicional para tener en cuenta sobre la dispersión;

- d) debe tener en cuenta la posibilidad de valores inesperados para el resultado comunicado o la incertidumbre;
- e) deberían tener una base teórica sólida;
- f) deberá tener un rendimiento demostrado (por ejemplo, en los datos de prueba o en simulaciones) suficiente para los efectos del plan de ensayos de aptitud.

7.7 Valor de consenso de resultados de participantes

7.7.1 Con este enfoque, el x_{pt} valor asignado para el elemento de prueba de capacidad utilizada en una ronda de un programa de ensayos de aptitud es la estimación de localización (por ejemplo, significan, media robusta, la mediana o la aritmética) formado a partir de los resultados reportados por los participantes en la ronda, calculado mediante un procedimiento adecuado de acuerdo con el diseño, tal como se describe en el anexo C. Las técnicas descritas en las secciones 6.2-6.6 se deben utilizar para confirmar que existe un acuerdo suficiente, antes de combinar los resultados.

7.7.1.1 En algunas situaciones, el proveedor de ensayos de aptitud puede desechar utilizar un subconjunto de determinaciones fiables por participantes para algunos criterios predefinidos, como el estado de acreditación o sobre la base de los resultados anteriores. Las técnicas de esta sección se aplican a esas situaciones, incluyendo consideraciones para el tamaño del grupo.

7.7.1.2 Otros métodos de cálculo pueden ser utilizados en lugar de los del Anexo C, siempre que tengan una base estadística sólida y el informe establezca el método a utilizar.

7.7.1.3 Las ventajas de este enfoque son las siguientes:

- a) no son necesarias las mediciones adicionales para obtener el valor asignado;
- b) el enfoque puede ser particularmente útil con un mensurando normalizado, definido operacionalmente, ya que a menudo no existe un método más fiable para obtener resultados equivalentes.

7.7.1.4 Las limitaciones de este enfoque las siguientes:

- a) no haya suficiente acuerdo entre los participantes;
- b) el valor de consenso puede incluir sesgo desconocido debido a la utilización general de la metodología defectuosa y este sesgo no se reflejará en la incertidumbre típica del valor asignado;
- c) el valor de consenso puede ser sesgado debido al efecto de sesgo en los métodos que se utilizan para determinar el valor asignado.
- d) Puede ser difícil determinar la trazabilidad metrológica del valor de consenso. Mientras que el resultado siempre es trazable a los resultados de los laboratorios individuales, una declaración clara de trazabilidad más allá de eso sólo se puede hacer cuando el proveedor de ensayos de aptitud tiene la información completa acerca de los estándares de calibración utilizados y el control

de otras condiciones del método relevantes por todos los participantes contribuyendo al valor de consenso.

7.7.2 La incertidumbre típica del valor asignado dependerá del procedimiento utilizado. Si se necesita un enfoque totalmente general, el proveedor de ensayos de aptitud debe considerar el uso de técnicas de remuestreo ("bootstrapping") para estimar un error típico para el valor asignado. Referencias [17,18] dar detalles de las técnicas de bootstrapping.

NOTA Un ejemplo utilizando una técnica de arranque está previsto en el anexo E.

7.7.3 Cuando el valor asignado se deriva como un robusto promedio calculado usando los procedimientos en el anexo C.2, C.3, o C.5, la incertidumbre típica de la x_{pt} valor asignado puede estimarse como:

$$u(x_{pt}) = 1,25 \times \frac{s^*}{\sqrt{p}} \quad (6)$$

Donde s^* es la desviación típica robusta de los resultados. (Aquí un "resultado" de un participante es el promedio de todas sus mediciones en el objeto de ensayo de aptitud.)

NOTA 1: En este modelo, en el que el valor asignado y desviación típica robusta se determinan a partir de los resultados de los participantes, la incertidumbre del valor asignado puede ser asumida para incluir los efectos de la incertidumbre debido a la falta de homogeneidad, el transporte, y la inestabilidad.

NOTA 2: El factor 1,25 se basa en la desviación típica de la media, o la eficiencia de la mediana como una estimación de la media, en un gran conjunto de resultados procedentes de una distribución normal. Se aprecia que la eficiencia de los métodos más sofisticados robusto puede ser mucho mayor que la de la mediana, lo que justifica un factor de corrección menor que 1,25. Sin embargo, este factor ha sido recomendado ya que los resultados de ensayos de aptitud por lo general no son estrictamente una distribución normal, y contienen proporciones desconocidas de los resultados de diferentes distribuciones ('Resultados contaminados'). El factor de 1,25 se considera que es una (alta) estimación conservadora, para tener en cuenta una posible contaminación. Proveedores de ensayos de aptitud pueden ser capaces de justificar el uso de un factor más pequeño, o una ecuación diferencial, dependiendo de la experiencia y el procedimiento robusto utilizado.

NOTA 3 Un ejemplo del uso de un valor asignado de resultados de participantes figura en el Anexo E.3.

7.8 Comparación del valor asignado con un valor de referencia independiente

7.8.1 Cuando se usan los métodos descritos en 7.7 para establecer el valor asignado (x_{pt}), y cuando se dispone de una estimación fiable independiente (x_{ref} de NOTADA), por ejemplo, a partir del conocimiento de la preparación o de un valor de referencia, el valor de consenso x_{pt} debe compararse con referencia externa x_{ref} .

Cuando se usan los métodos descritos en el apartado 7.3 a 7.6 para establecer el valor asignado, las medias robustas x^* derivadas de los resultados de la ronda deben compararse con el valor asignado después de cada ronda de un programa de ensayos de aptitud.

La diferencia se calcula como: $x_{diff} = (x_{ref} - x_{pt})$ (o $(x^* - x_{pt})$) y la incertidumbre típica de la diferencia se calcula como:

$$u_{diff} = \sqrt{u^2(x_{ref}) + u^2(x_{pt})} \quad (7)$$

Donde:

$u(x_{ref})$ es la incertidumbre del valor de referencia para la comparación;
 $u(x_{pt})$ y es la incertidumbre del valor asignado.

NOTA: Un ejemplo de una comparación de un valor de referencia con un valor de consenso está incluido en anexo E.7.

7.8.2 Si la diferencia es más del doble de su incertidumbre típica, la razón debe ser investigada. Las posibles razones pueden ser:

- Sesgo en el método de medición de referencia;
- Un sesgo común en los resultados de los participantes;
- Incapacidad para apreciar las limitaciones del método de medición cuando se usa el método de formulación descrito en 7.3;
- Sesgo en los resultados de los "expertos" cuando se utilizan los enfoques en las secciones 7.5 o 7.6;
- El valor de comparación y el valor asignado no tienen su origen en la misma referencia metroológica.

7.8.3 En función de la razón de la diferencia, el proveedor de ensayos de aptitud debe decidir si se debe evaluar los resultados o no, y (para continuar con los programas de ensayos de aptitud). Cuando la diferencia es suficientemente grande como para afectar a la evaluación del rendimiento o para sugerir importante sesgo en los métodos de medición utilizados por los participantes, la diferencia debe tenerse en cuenta en el informe para la ronda. En tales casos, la diferencia se debe considerar en el diseño de futuros programas de ensayos de aptitud.

8 Determinación de los criterios para la evaluación del desempeño

8.1 Métodos para la determinación de los criterios de evaluación

8.1.1 El enfoque básico para todos los efectos es comparar el resultado en un elemento de prueba de competencia (x) con un valor asignado (x_{pt}). Para la evaluación, la diferencia se compara con una estimación para el error de medición. Esta comparación se hace comúnmente a través de una estadística de rendimiento estandarizados (por ejemplo, z , z' , ζ , E_n), como se explica en las secciones 9.4-9.7. Esto también se puede hacer mediante la comparación de la diferencia con un criterio definido (D o $D\%$ en comparación con δ_E) como se discute en 9.3. Un enfoque alternativo a la evaluación es comparar la diferencia con la reclamación de un participante por la incertidumbre de su resultado combinado con la incertidumbre del valor asignado (E_n y ζ).

8.1.2 Si un requisito reglamentario o una aptitud para el propósito meta se da como una desviación típica puede ser utilizado directamente como σ_{pt} . Si el requisito o meta es un error máximo admisible de medición, este criterio puede ser dividido por el límite de acción para obtener σ_{pt} . Un error permisible máximo prescrito puede ser utilizado directamente como δ_E para su uso con D o $D\%$. Las ventajas de este enfoque para los esquemas continuos son:

- a) las puntuaciones de rendimiento tienen una interpretación consistente en términos de aptitud para el uso de una fase a la siguiente;
- b) las puntuaciones de rendimiento no están sujetos a la variación esperada en la estimación de la dispersión de los resultados reportados.

Ejemplo: Si se especifica un criterio normativo como un error máximo permisible y 3,0 es un límite de acción para la evaluación con una puntuación z score, entonces el criterio especificado se divide por 3,0 para determinar σ_{pt} .

8.1.3 Cuando el criterio para la evaluación de los resultados se basa en estadísticas de consenso de la actual ronda o rondas anteriores del esquema de ensayos de aptitud, a continuación, una estimación robusta de la desviación típica de los resultados de los participantes es la estadística preferida. Cuando se utiliza este método por lo general es más conveniente utilizar una puntuación de rendimiento, tales como la puntuación z score y para ajustar la desviación típica para la evaluación de aptitud (σ_{pt}) a la estimación calculada de la desviación típica.

8.2 Por la percepción de los expertos

8.2.1 El error máximo permisible o la desviación típica para la evaluación de aptitud puede ser fijado en un valor que se corresponda con el nivel de rendimiento de una autoridad reguladora, el órgano de acreditación, o los expertos técnicos del proveedor de ensayos de aptitud creen que esto es razonable para los participantes.

8.2.2 Al especificar un error máximo admisible se puede transformar en una desviación típica para la evaluación de aptitud dividiendo el límite por el número de múltiplos de la σ_{pt} que se utilizan para definir una señal de acción (o resultado inaceptable). Del mismo modo, un σ_{pt} especificado se puede transformar en δ_E .

8.3 Por la experiencia de las rondas anteriores de ensayos de aptitud

8.3.1 La desviación típica para la evaluación de competencia (σ_{pt}), y el error máximo admisible (δ_E), puede ser determinada por la experiencia con las anteriores rondas de ensayos de aptitud para el mismo mensurando con valores de propiedades comparables, y donde los participantes utilizan procedimientos de medición compatibles. Este es un enfoque útil cuando no hay acuerdo entre los expertos acerca de la aptitud para el propósito. Las ventajas de este enfoque son las siguientes:

- Las evaluaciones se basan en las expectativas razonables de rendimiento;
- Los criterios de evaluación no varían de ronda a ronda del esquema de ensayos de aptitud debido a la variación aleatoria o cambios en la población participante;
- Los criterios de evaluación no van a variar entre los diferentes proveedores de ensayos de aptitud, cuando hay dos o más proveedores de ensayos de aptitud aprobados para un área de ensayo o calibración.

8.3.2 La revisión de las rondas previas de un programa de ensayos de aptitud debe incluir la consideración de rendimiento que se puede lograr por los participantes competentes, y no se ve afectado por los nuevos participantes o la variación aleatoria debido a, por ejemplo, tamaño de los

grupos más pequeños o de otros factores propios de un particular, redondo. Las determinaciones se pueden hacer subjetivamente mediante el examen de las rondas previas de la coherencia, u objetivamente con promedios o con un modelo de regresión que se ajusta por el valor de la magnitud a medir. La ecuación de regresión podría ser una línea recta, o podría ser curvo [31]. Las desviaciones típicas y las desviaciones típicas relativas deben ser consideradas, con la selección sobre la base de que es más coherente en toda la gama adecuada de los niveles mensurando. Un error permisible máximo apropiado también se puede obtener de esta manera.

8.3.3 Cuando el criterio para la evaluación de los resultados se basa en estadísticas de consenso de rondas anteriores de un esquema de ensayos de aptitud, deben ser utilizados estimaciones robustas de la desviación típica.

NOTA 1: Algoritmo S (Anexo C.4) proporciona una robusta desviación típica combinada que es aplicable cuando todas las rondas anteriores de un programa de ensayos de aptitud bajo consideración tienen la misma desviación típica esperada o (si se utilizan desviaciones relativas a la evaluación) la misma relación desviación típica.

NOTA 2: Un ejemplo de derivar un valor de la experiencia de las anteriores rondas de ensayos de aptitud se proporciona en el anexo E.8.

8.4 Mediante el uso de un modelo general

8.4.1 El valor de la desviación típica para la evaluación de aptitud se puede derivar de un modelo general para la reproducibilidad del método de medición. Este método tiene la ventaja de la objetividad y la consistencia a través de valores de medida, así como basándose empíricamente. Dependiendo del modelo utilizado, para un propósito, este enfoque podría ser considerado como un caso especial de un criterio de aptitud.

8.4.2 Cualquier desviación típica esperada elegida por un modelo general debe ser razonable. Si proporciones muy grandes o muy pequeños de los participantes se les asigna las señales de advertencia o de acción, el proveedor de ensayos de aptitud debe asegurarse de que esto es consistente con el propósito del programa de ensayos de aptitud.

8.4.3 Una estimación específica teniendo las características específicas del problema de la medida en la consideración general, es preferible un enfoque genérico. En consecuencia, antes de usar un modelo general, la posibilidad de utilizar los enfoques descritos en 8.2, 8.3 y 8.5 se debe explorar. Ejemplo curva de Horwitz.

Un modelo general común para aplicaciones químicas fue descrito por Horwitz^[22] y modificado por Thompson^[31]. Este enfoque da un modelo general para la reproducibilidad de los métodos de análisis que pueden utilizarse para derivar la siguiente expresión para la desviación típica de la reproducibilidad:

$$\sigma_R = \begin{cases} 0,22c & \text{cuando } c < 1,2 \times 10^{-7} \\ 0,02c^{0,8495} & \text{cuando } 1,2 \times 10^{-7} \leq c \leq 0,138 \\ 0,01c^{0,5} & \text{cuando } c > 0,138 \end{cases} \quad (8)$$

Donde c es la fracción en masa de las especies químicas determinadas cuando $0 \leq c \leq 1$

NOTA 1: El modelo de Horwitz es empírico, basado en observaciones de los ensayos de colaboración de muchos parámetros a lo largo de un período de tiempo prolongado. Los valores σ_R son los límites superiores previstos de la variabilidad entre laboratorios cuando el ensayo colectivo no tenía ningún problema significativo. Los valores σ_R por lo tanto, podrían no ser criterios adecuados para determinar la competencia en un programa de ensayos de aptitud.

NOTA 2: Un ejemplo de derivar un valor a partir del modelo Horwitz modificada se proporciona en el anexo E.9.

8.5 Uso de las desviaciones típicas de repetibilidad y reproducibilidad de un estudio en colaboración de un método de medición previo de precisión.

8.5.1 Cuando el método de medición para ser utilizado en el esquema de ensayos de aptitud está estandarizado, y la información sobre la repetibilidad (σ_r) y reproducibilidad (σ_R) del método está disponible, la desviación típica para la evaluación de aptitud (σ_{pt}) puede calcularse utilizando esta información, como:

$$\sigma_{pt} = \sqrt{\sigma_R^2 - \sigma_r^2(1 - 1/m)} \quad (9)$$

Donde m es el número de mediciones repetidas de cada participante en una ronda del programa de ensayos de aptitud

NOTA: Esta ecuación se deriva de un modelo básico de efectos aleatorios de la norma ISO 5725-2.

8.5.2 Cuando las desviaciones típicas de repetibilidad y reproducibilidad dependen del valor medio de los resultados de la prueba, las relaciones funcionales deben ser derivadas por los métodos descritos en la norma ISO 5725-2. Estas relaciones deben entonces ser utilizadas para calcular los valores de la norma de la repetibilidad y la reproducibilidad de desviaciones apropiadas para el valor asignado que se va a utilizar en el programa de ensayos de aptitud.

8.5.3 Para que las técnicas anteriores sean válidas, el estudio de colaboración debe haber sido realizado de acuerdo con los requisitos de la norma ISO 5725-2 o un procedimiento equivalente.

NOTA Un ejemplo se presenta en el anexo E.10.

8.6 A partir de los datos obtenidos en la misma ronda de un programa de ensayos de aptitud

8.6.1 Con este enfoque, la desviación típica para la evaluación de competencia, σ_{pt} , se calcula a partir de los resultados de los participantes en la misma ronda del programa de ensayos de aptitud. Cuando se utiliza este método por lo general es más conveniente utilizar una puntuación de rendimiento, tales como la puntuación z score. Una estimación robusta de la desviación típica de los resultados reportados por todos los participantes, calculados utilizando una técnica que figuran en el Anexo C, normalmente se debe utilizar para calcular σ_{pt} . En general, la evaluación con D o D% y el uso de δ_E no son apropiados en estas situaciones, sin embargo P_A todavía se puede utilizar como una puntuación estandarizada, para la comparación entre los valores de medida (sección 9.3.6).

8.6.2 El uso de resultados de participantes puede dar lugar a criterios de evaluación de desempeño que no son apropiados. El proveedor de ensayos de aptitud debe garantizar que la σ_{pt} utilizada para las evaluaciones de rendimiento es adecuada para el propósito.

8.6.2.1 El proveedor de ensayos de aptitud debe poner un límite en el valor más bajo de σ_{pt} que será utilizado, en el caso de que la desviación típica robusta sea muy pequeña. Este límite debe ser elegido de modo que cuando el error de medición es apto para el uso previsto más desafiante, la puntuación de rentabilidad no será z <3,0.

Ejemplo: En un esquema de ensayos de aptitud para la tela, una magnitud a medir es el número de hilos por centímetro. La desviación típica robusta puede ser pequeña en algunas rondas (<1 hilo por cm.). Y los errores de menos de 4 hilos / cm se consideran insignificantes. El proveedor de ensayos de aptitud determina que la desviación típica robusta se usa como σ_{pt} , a menos que sea inferior a 1,3 hilos / cm, en cuyo caso se utiliza $\sigma_{pt} = 1,3$.

8.6.2.2 El proveedor de ensayos de aptitud debe poner un límite a la σ_{pt} más grande que el que se utilizará, o en los resultados de las mediciones que se pueden evaluar como (sin señal) "aceptable", en el caso de que la desviación típica robusta es muy grande. Este límite debe ser elegido de modo que los resultados que no son aptos para el propósito recibirán una señal de acción.

8.6.2.3 En algunos casos, el proveedor de ensayos de aptitud puede poner límites superiores o inferiores en el intervalo de los resultados que se puede evaluar como (sin señal de aviso o acción) "aceptable", cuando los intervalos simétricos son resultados que no serían aptos para el propósito.

Ejemplo: Para un programa de ensayos de aptitud reguladora de agua no potable, la normativa específica de los resultados debe de estar dentro de la 3 σ_{pt} robusta media de resultados de los participantes. Sin embargo, debido a que en algunos casos, el rango de resultados aceptables podría incluir 0 g / L, cualquier resultado de menos de 10% de un valor formulado deberá generar una señal de acción (o "inaceptable"). Un elemento de ensayos de aptitud se formula con 4,0 g / L de una sustancia regulada. Significa que el participante robusta es de 3,2 g / L y σ_{pt} es 1,1 mg / L. Por lo tanto, es posible que un participante que presente un resultado de 0,0 mg / L está dentro de 3 σ_{pt} , pero cualquier resultado inferior a 0,4 mg / L se evaluará como "inaceptable".

8.6.3 Las principales ventajas de este enfoque son la sencillez y la aceptación convencional debido a uso con éxito en muchas situaciones. Este puede ser el único enfoque factible.

8.6.4 Hay varias desventajas con este enfoque:

a) El valor de σ_{pt} puede variar sustancialmente de ronda a ronda de un programa de ensayos de aptitud, por lo que es difícil para un participante de utilizar los valores de la puntuación z para buscar tendencias que persisten en varias rondas.

b) Las desviaciones típicas pueden ser poco fiables cuando el número de participantes en el programa de ensayos de aptitud es pequeño o cuando los resultados de diferentes métodos se combinan. Por ejemplo, si P = 20, la desviación típica para los datos distribuidos normalmente puede variar alrededor de ± 30% de su valor verdadero de una ronda de esquema de ensayo de aptitud a la siguiente.

c) El uso de medidas de dispersión derivadas de los datos puede conducir a una proporción aproximadamente constante de resultados aparentemente aceptables. Generalmente el pobre

rendimiento no será detectado por la inspección de los resultados, y en general un buen rendimiento dará lugar a buenos participantes que recibieron las puntuaciones pobres.

d) No existe una interpretación útil en términos de idoneidad para cualquier uso final de los resultados.

NOTA: Ejemplos de la utilización de datos de los participantes se proporcionan en el ejemplo completo en el Anexo E.3.

8.7 Seguimiento del acuerdo entre laboratorios

8.7.1 Como comprobación sobre el desempeño de los participantes, y para evaluar el beneficio del programa de ensayos de aptitud, el proveedor debe aplicar un procedimiento para supervisar el acuerdo entre laboratorios, para realizar un seguimiento de los cambios en el rendimiento y asegurar la razonabilidad de procedimientos estadísticos.

8.7.2 Los resultados obtenidos en cada ronda de un programa de ensayos de aptitud deben ser utilizados para calcular las estimaciones de las desviaciones típicas de la reproducibilidad del método de medición (y repetibilidad, si está disponible), utilizando los métodos robustos que se describen en el Anexo C. Estas estimaciones deben ser representadas en los gráficos de forma secuencial o como una serie de tiempo, junto con los valores de las desviaciones típicas de repetibilidad y reproducibilidad obtenidos en los experimentos de precisión de la norma ISO 5725-2 (si está disponible), y / o σ_{pt} , si las técnicas en las secciones 8.2 a 8.4 se utilizan.

8.7.3 Estos gráficos a continuación, deben ser examinados por el proveedor de ensayos de aptitud. Si los gráficos muestran que los valores de precisión obtenidos en una determinada ronda de ensayos de aptitud son mayores en un factor de dos o más de los valores esperados a partir de datos o experiencia previa, entonces el proveedor de ensayos de aptitud debe investigar según los acuerdos en rondas anteriores. Del mismo modo, una tendencia hacia mejores o peores valores de precisión debe dar lugar a una investigación de las causas más probables.

9 El cálculo de las estadísticas de rendimiento

9.1 Consideraciones generales para la determinación del rendimiento

9.1.1 Estadísticas utilizadas para la determinación del rendimiento deberán ser compatibles con el objetivo (s) para el programa de ensayos de aptitud.

NOTA: Son más útiles si las estadísticas y su derivación son entendidas por los participantes y otras partes interesadas.

9.1.2 Los puntajes de desempeño deben ser revisados fácilmente a través de los niveles mensurando y diferentes rondas de un programa de ensayos de aptitud.

9.1.3 La participante deben ser revisados y determinados a ser coherente con los supuestos utilizados en el diseño del programa de ensayos de aptitud, para tener en cuenta las estadísticas de rendimiento significativas. Por ejemplo, que no hay evidencia de deterioro del producto de ensayo de aptitud, o de una mezcla de las poblaciones de los participantes, o de violaciones graves de los supuestos estadísticos sobre la naturaleza de los datos.

9.1.4 En general, no es apropiado utilizar métodos de evaluación que clasifican intencionadamente una proporción fija de resultados como generador de una señal de acción.

9.2 Limitación de la incertidumbre del valor asignado

9.2.1 Si la incertidumbre típica $u(x_{pt})$ del valor asignado es grande en comparación con el criterio de evaluación de rendimiento, entonces hay un riesgo de que algunos participantes recibirán de acción y de advertencia señales debido a la inexactitud en la determinación del valor asignado, no debido a cualquier causa del participante. Por esta razón, la incertidumbre típica del valor asignado será determinada y se notificará a los participantes (véase la norma ISO / IEC 17043: 2010, 4.4.5 y 4.8.2).

Si se cumple el siguiente criterio, entonces la incertidumbre del valor asignado puede ser considerada insignificante y no es preciso incluirla en la interpretación de los resultados de la ronda de ensayos de aptitud.

$$u(x_{pt}) < 0,3\sigma_{pt} \quad \text{o} \quad u(x_{pt}) < 0,1\delta_E \quad (10)$$

NOTA: $0,3\sigma_{pt}$ es equivalente a $0,1\delta_E$ cuando $|z| \geq 3,0$ genera una señal de acción.

9.2.2 Si no se cumple este criterio, entonces el proveedor de ensayos de aptitud debe tener en cuenta lo siguiente, asegurando cualquier acción tomada sigue siendo consistente con la política de evaluación de desempeño acordada para el programa de ensayos de aptitud.

a) Seleccionar un método para determinar el valor asignado de tal manera que su incertidumbre cumple con el criterio de la ecuación (10).

b) Usar la incertidumbre del valor asignado en la interpretación de los resultados del régimen de ensayos de aptitud (ver secciones 9.5 sobre la z' score, o 9.6 sobre ζ scores, o 9.7 sobre E_n scores).

c) Si el valor asignado se deriva de resultados de participantes, y surge la gran incertidumbre de las diferencias entre subpoblaciones identificables, se reportan los valores separados e incertidumbres para cada sub-población (por ejemplo, los participantes utilizando diferentes métodos de valoración).

NOTA: El protocolo armonizado IUPAC [32] describe un procedimiento específico para la detección de bimodalidad, sobre la base de una inspección de un gráfico de densidad kernel con un ancho de banda especificado.

d) Informar a los participantes que la incertidumbre del valor asignado no es insignificante, y las evaluaciones podrían verse afectadas.

Si ningún inciso del a) al d) se aplica, a continuación, los participantes serán informados de que ningún valor asignado fiable puede ser determinado y que no hay puntuaciones de rendimiento que puedan ser proporcionada.

NOTA: Las técnicas presentadas en esta sección se demuestran en los anexos E.3 y E.4.

9.3 Las estimaciones de desviación (error de medición)

9.3.1 Permita que x_i represente el resultado (o la media de las réplicas) reportadas por un participante i para la medición de una propiedad del elemento de ensayo de aptitud en la primera ronda de un programa de ensayos de aptitud. A continuación, una simple medida de rendimiento del participante puede ser calculado como la diferencia entre el resultado de X_i y el x_{pt} valor asignado:

$$D_i = x_i - x_{pt} \quad (11)$$

D_i se puede interpretar como el error de medición para ese resultado, en la medida en la que el valor asignado puede ser considerado un valor de cantidad convencional o de referencia.

La diferencia D_i puede ser expresada en las mismas unidades que el valor asignado o como una diferencia porcentual, calculada como:

$$D_i \% = 100(x_i - x_{pt}) / x_{pt} \% \quad (12)$$

9.3.2 La diferencia D o $D\%$ es por lo general en comparación con un criterio δ_E basado en la aptitud para el propósito con la experiencia de las rondas anteriores de un programa de ensayos de aptitud; el criterio se observa aquí como δ_E , un ajuste por error de medición. Si $-\delta_E < D < \delta_E$ entonces el rendimiento se considera que es "aceptable" (o "sin señal"). (El mismo criterio se aplica para $D\%$, dependiendo de la expresión de δ_E)

9.3.3 δ_E está estrechamente relacionado con σ_{pt} tal como se utiliza para las puntuaciones z score (ver 9.4), cuando σ_{pt} está determinada por la aptitud para el propósito o expectativas de rondas anteriores. La relación viene determinada por el criterio de evaluación para las puntuaciones z score. Por ejemplo, si $z \geq 3$ crea una señal de acción, entonces $\delta_E = 3\sigma_{pt}$, o equivalentemente $\sigma_{pt} = \delta_E/3$. Diversas expresiones de δ_E son convencionales en los ensayos de aptitud para aplicaciones médicas y en las especificaciones de rendimiento para los métodos y productos de medición.

9.3.4 La ventaja de D como una estadística rendimiento y δ_E como un criterio de rendimiento es que los participantes tengan una comprensión intuitiva de estas estadísticas, ya que están directamente ligados a errores de medición y son comunes como criterios para determinar la aptitud para el uso. La ventaja de $D\%$ es que la comprensión es intuitiva, que ha sido estandarizada para el nivel de magnitud a medir, y está relacionado con las causas más comunes de errores (por ejemplo, calibración incorrecta o sesgo en la dilución).

9.3.5 Las desventajas son que no es convencional para las pruebas de competencia en muchos países o áreas de medición; D no es estándar, para permitir la exploración sencilla de informes para las señales de acción en los programas de ensayos de aptitud con múltiples analitos o la aptitud para el propósito de criterios puede variar según el nivel de la magnitud a medir.

NOTA: El uso de D y $D\%$ asume generalmente la simetría de la distribución de resultados de participantes en el sentido de que el rango aceptable es $-\delta_E < D < \delta_E$.

9.3.6 Para efectos de comparación entre los niveles mensurando, donde la aptitud para el propósito de criterios puede variar; o para la combinación en las vueltas oa través de valores de medida, D y $D\%$ se pueden transformar en una puntuación de rendimiento estandarizado que muestra las diferencias relativas a los criterios de funcionamiento de los valores de medida. Para ello, se calcula el "Porcentaje de desviación permitida" (P_A) para cada resultado de la siguiente manera:

$$P_{Ai} = (D_i / \delta_E) \times 100 \% \quad (13)$$

Por lo tanto $P_A \geq 100\%$ o $P_A \leq -100\%$ indica una señal de acción (o "funcionamiento inaceptable").

NOTA 1: Las puntuaciones P_A se pueden comparar entre diferentes niveles y rondas de un programa de ensayos de aptitud, o rastreados en los gráficos. Estas puntuaciones de rendimiento son similares en el uso e interpretación de las puntuaciones z score que tienen un criterio de evaluación común como $z \leq -3$ o $z \geq 3$ para señales de acción.

NOTA 2: Variaciones de esta estadística se utilizan comúnmente, en particular en aplicaciones médicas, en donde por lo general hay una mayor frecuencia de las pruebas de competencia y un gran número de analitos.

NOTA 3: Puede ser apropiado utilizar el valor absoluto de P_A para reflejar coherentemente aceptable (o inaceptable) en relación con el valor asignado.

9.4 z score

9.4.1 La z score para una x_i en la prueba de aptitud se calcula como:

$$z_i = \frac{(x_i - x_{pt})}{\sigma_{pt}} \quad (14)$$

donde:

x_{pt} es el valor asignado; y

σ_{pt} es la desviación típica para la evaluación de competencia

9.4.2 La interpretación convencional de la z score es la siguiente (véase la norma ISO / IEC 17043: 2010, B.4.1.1):

- Un resultado que de $|z| \leq 2,0$ se considera que es aceptable.
- Un resultado que de $2,0 < |z| < 3,0$ se considera para dar una señal de advertencia.
- Un resultado que de $|z| \geq 3,0$ se considera que es inaceptable (o señal de acción).

NOTA1: En algunas aplicaciones, proveedores de ensayos de aptitud utilizan 2,0 como una señal de acción para el z-score.

NOTA 2: La elección del criterio σ_{pt} deberá ser tomadas a fin de permitir la interpretación anterior, que es ampliamente utilizado para la evaluación de competencia y también es muy similar a los límites del gráfico de control familiares.

NOTA 3: La justificación para el uso de los límites de 2,0 y 3,0 para la z score es el siguiente. Las mediciones que se llevan a cabo correctamente se asumen para generar resultados que pueden ser descritos (después de la transformación, si es necesario) por una distribución normal con media σ_{pt} y desviación típica x_{pt} . Las z score serán distribuidas normalmente con una media de cero y una desviación típica de 1,0. En estas circunstancias, se esperaría que sólo alrededor del 0,3% de las puntuaciones cayeran fuera del rango de $-3,0 \leq z \leq 3,0$ y solo alrededor del 5% caerían fuera del rango de $-2,0 \leq z \leq 2,0$. Debido a que la probabilidad de que no respondan $z \pm 3,0$ es tan bajo, es poco probable que las señales de acción ocurrirán por casualidad cuando no existe ningún problema real, por lo que es probable que exista una causa identificable de una anomalía cuando se da una señal de acción.

NOTA 4: La hipótesis en la que se basa esta interpretación se aplica sólo a una distribución hipotética de los laboratorios competentes y no en ninguna suposición sobre la distribución de los resultados observados. Su hipótesis necesita ser realizada sobre los mismos resultados observados.

NOTA 5: Si la verdadera variabilidad entre laboratorios es más pequeño que σ_{pt} , se reducen las probabilidades de errores de clasificación.

NOTA 6: Cuando la desviación típica para la evaluación de competencia se fija mediante cualquiera de los métodos descritos en 8.2 o 8.4, puede diferir sustancialmente de la (amplia) desviación típica de los resultados, y las proporciones de los resultados que quedan fuera $\pm 2,0$ y $\pm 3,0$ pueden diferir considerablemente de 5% y 0,3%, respectivamente.

9.4.3 El proveedor de ensayos de aptitud debe determinar el redondeo adecuado para los z-scores reportados, basados en el número de dígitos significativos para el resultado, y por el valor asignado y la desviación típica para las pruebas de competencia. Las normas de redondeo se incluirán en la información disponible para los participantes.

NOTA Es raramente útil disponer de más de dos dígitos después del punto decimal para el z score.

9.4.4 Cuando se utiliza la desviación típica de los resultados de participantes como los programas de ensayos de aptitud σ_{pt} e involucran a un gran número de participantes, el proveedor de ensayos de aptitud podrá comprobar la normalidad de la distribución, el uso de los resultados reales o z score. En el otro extremo, cuando sólo hay un pequeño número de participantes, puede que no haya señal de acción dada. En este caso, los métodos gráficos que combinan las puntuaciones de rendimiento en varias rondas pueden proporcionar indicaciones más útiles de la actuación de los participantes que los resultados de rondas individuales.

9.5 z'-score

9.5.1 Cuando existe preocupación por la incertidumbre de un valor asignado $u(x_{pt})$, por ejemplo cuando $u(x_{pt}) > 0,3\sigma_{pt}$, la incertidumbre puede ser tomada en cuenta por la ampliación del denominador de la puntuación de rendimiento. Esta estadística se llama z'score y se calcula de la siguiente manera (con la NOTAción de la sección 9.4):

$$z'_i = \frac{x_i - x_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} \quad (15)$$

NOTA Cuando x_{pt} y/o σ_{pt} se calculan a partir de resultados de participantes, la puntuación de los resultados se correlacionan con los resultados individuales de los participantes, ya que los resultados individuales

tienen un impacto tanto en una robusta media y la desviación típica. La correlación de un participante individual depende de la ponderación dada a ese participante en la estadística combinada. Por esta razón, las puntuaciones de rendimiento, incluyendo la incertidumbre del valor asignado sin incluir un ajuste por correlación representan subestimaciones de las puntuaciones que resultarían si se incluyera la covarianza. Por ejemplo, cuando $u(x_{pt})=0,3\sigma_{pt}$ entonces hay una subestimación de aproximadamente el 10% del z'-score. Por lo tanto la ecuación (15) se puede utilizar cuando x_{pt} y/o σ_{pt} esta determina a partir de resultados de participantes.

9.5.2 D y D% sus resultados también pueden ser modificados para tener en cuenta la incertidumbre del valor asignado a la siguiente fórmula para expandir δ_E y δ'_E

$$\delta'_E = \sqrt{\delta_E^2 + U^2(x_{pt})} \quad (16)$$

donde $U(x_{pt})$ es la incertidumbre expandida del valor asignado x_{pt} con un factor de cobertura $k = 2$ calculado.

9.5.3 Los z'-scores pueden interpretarse de la misma manera que los z-scores (ver 9.4) y utilizando los mismos valores críticos de 2,0 y 3,0, dependiendo del diseño para el programa de ensayos de aptitud. Del mismo modo, D y D% sus resultados a continuación se compararían con δ'_E (véase 9.3).

9.5.4 La comparación de las fórmulas del z score y z' score de 9.4 y 9.5 muestra que el z' score aNOTAdo para una ronda de un programa de ensayos de aptitud siempre serán más pequeños que los z scores correspondientes a un factor constante de

$$\frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}}$$

Cuando se cumple la directriz para limitar la incertidumbre del valor asignado en el apartado 9.2.1, este factor caerá en el rango:

$$0,96 < \frac{\sigma_{pt}}{\sqrt{\sigma_{pt}^2 + u^2(x_{pt})}} < 1,00$$

Por lo tanto, en este caso, la NOTA de z' score será casi idénticas a las z score, y se puede concluir que la incertidumbre del valor asignado es insignificante para la evaluación del desempeño.

Cuando no se cumple la directriz en el apartado 9.2.1 de la incertidumbre del valor asignado, la diferencia en la magnitud del z' score y el z score puede ser tal que en algunos puntajes z excede los valores críticos de 2,0 o 3,0 por lo que dan "señales de aviso" o "señales" de acción, mientras que el z' score correspondiente no supera estos valores críticos y por lo tanto no da señales.

En general, para estas situaciones en las que el valor asignado y/o σ_{pt} no se determina a partir de resultados de participantes, z' score puede ser preferible porque cuando el criterio 9.2.1 se cumple la diferencia entre Z y Z' será insignificante.

9.6 Zeta scores (ζ)

9.6.1 Las Zeta pueden ser útiles cuando un objetivo para el programa de ensayos de aptitud sea evaluar la capacidad de un participante de tener resultados cerca del valor asignado dentro de su incertidumbre. Con la NOTación como en 9.4, las puntuaciones zeta se calculan como:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u^2(x_i) + u^2(x_{pt})}} \quad (17)$$

donde:

$u(x)$ es la estimación propia del participante de la incertidumbre típica de su resultado $[x]$, y

$u(x_{pt})$ es la incertidumbre típica de la x_{pt} y su valor asignado.

NOTA 1: Cuando el x_{pt} valor asignado se calcula como el valor de consenso a partir de los resultados de los participantes, entonces x_{pt} se correlaciona con los resultados individuales de los participantes. La correlación de un participante individual depende de la ponderación dada a ese participante en el valor asignado, y en menor medida, en la incertidumbre del valor asignado. Por esta razón, las puntuaciones de rendimiento, incluyendo la incertidumbre del valor asignado sin incluir un ajuste por correlación representan subestimaciones de las puntuaciones que resultarían si se incluyera la covarianza. La subestimación no es grave si la incertidumbre del valor asignado es pequeño; cuando se utilizan métodos robustos que es menos grave para los participantes más extrovertidos con más probabilidades de recibir las puntuaciones de rendimiento adversas. Por lo tanto, la ecuación (17) se puede utilizar con las estadísticas de consenso sin ajuste para su correlación.

NOTA 2: zeta difiere de las puntuaciones en (sección 9.7) mediante el uso de la incertidumbre típica $u(x)$ y $u(x_{pt})$, en lugar de la incertidumbre expandida $U(x)$ y $U(x_{pt})$. ζ scores por encima de 2 o por debajo de -2 pueden ser causada por métodos sistemáticamente sesgadas o por una mala estimación de la incertidumbre de la medición por el participante. Por lo tanto, los zeta scores proporcionan una evaluación rigurosa del resultado completo presentado por el participante.

9.6.2 Valiéndose de los resultados zeta permite la evaluación directa si los laboratorios si son capaces de ofrecer resultados correctos, es decir, resultados que concuerdan con x_{pt} dentro de sus incertidumbres de medición. ζ scores puede ser interpretada utilizando los mismos valores críticos de 2,0 y 3,0 en cuanto a las z scores, o con múltiplos del factor de cobertura de los participantes a utilizar para calcular la incertidumbre expandida. Sin embargo, una puntuación adversa de ζ puede indicar o bien una gran desviación de x_i de x_{pt} , una sub-estimación de la incertidumbre por parte del participante, o una combinación de ambos.

NOTA: Puede ser útil para el proveedor de ensayos de aptitud para dar información adicional acerca de la validez de las incertidumbres reportadas las directrices útiles para dicha evaluación que se sugieren en la sección 9.8.

9.6.3 Zeta score se puede utilizar en conjunción con la z score, como una ayuda para la mejora del rendimiento de los participantes, de la siguiente manera. Si un participante obtiene z score que superan varias veces el valor crítico de 3,0, puede resultar de utilidad para examinar su prueba de procedimiento paso a paso y obtener una evaluación de la incertidumbre para ese procedimiento.

La evaluación de la incertidumbre identificará los pasos en el procedimiento en el que se presentan las mayores incertidumbres, por lo que el participante puede ver donde utilizar esfuerzos para

lograr una mejoría. Zeta de las puntuaciones de los participantes también superan varias veces el valor crítico de 3,0, esto implica que la evaluación a los participantes y la incertidumbre no incluye todas las fuentes significativas de incertidumbre (es decir, que se están perdiendo algo importante). Por el contrario, si un participante obtiene repetidamente puntuaciones $z \geq 3$ pero zeta puntuajes <2, esto demuestra que el participante puede haber evaluado la incertidumbre de sus resultados con precisión pero que sus resultados no cumplen con el rendimiento esperado para el programa de ensayos de aptitud. Este puede ser el caso, por ejemplo, por un participante que utiliza un método de cribado en los procedimientos de medición, donde los otros participantes aplican métodos cuantitativos. No es necesaria ninguna acción si el participante considera que la incertidumbre de sus resultados es suficiente.

NOTA: Cuando una puntuación ζ se utiliza sola, se puede interpretar sólo como una prueba de si la incertidumbre del participante es consistente con la desviación observada en particular y no se puede interpretar como una indicación de la aptitud para el uso de los resultados de un participante particular. La determinación de la aptitud para el uso podría hacerse por separado (por ejemplo, por el participante o por un organismo de acreditación) mediante el examen de la desviación ($x-x_{pt}$) o las incertidumbres típicas combinadas en comparación con una incertidumbre de destino.

9.7 E_n scores

9.7.1 E_n scores pueden ser útiles cuando un objetivo para el programa de ensayos de aptitud es evaluar la capacidad de un participante a tener resultados cercanos al valor asignado dentro de su incertidumbre expandida. Esta estadística es convencional para las pruebas de competencia en la calibración, pero puede ser utilizado para otros tipos de ensayos de aptitud.

Esta estadística de rendimiento se calcula como:

$$(E_n)_i = \frac{x_i - x_{pt}}{\sqrt{U^2(x_i) + U^2(x_{pt})}} \quad (18)$$

Donde:

x_{pt} es el valor asignado determinado en un laboratorio de referencia;

$U(x_{pt})$ es la incertidumbre expandida de la x_{pt} valor asignado;

$U(x)$ es el resultado de la incertidumbre expandida x_i de un participante

NOTA: La combinación directa de incertidumbres expandidas no es compatible con el requisito de la norma ISO / IEC Guía 98 3 y no es equivalente al cálculo de una incertidumbre expandida combinado salvo tanto los factores de cobertura y los grados de libertad efectivos son idénticos para $U(x)$ y $U(x_{pt})$.

9.7.2 E_n scores deben interpretarse con precaución, ya que son las razones de dos medidas de desempeño separadas (pero relacionadas). El numerador es la desviación del resultado del valor asignado, y tiene una interpretación discutida en la sección 9.3. El denominador es una incertidumbre expandida combinada que no debe ser mayor que la desviación en el numerador, si el participante ha determinado $U(x)$ correctamente y si el proveedor de ensayos de aptitud ha determinado $U(x_{pt})$ correctamente. Por lo tanto, decenas de veces $E_n \geq 1,0$ o $E_n \leq -1,0$ podría indicar la necesidad de revisar las estimaciones de la incertidumbre, o para corregir un problema de medición; Del mismo modo $-1,0 < E_n < 1,0$ se debe tomar como un indicador de desempeño exitoso sólo si las incertidumbres son válidas y la desviación (x_i-x_{pt}) es la más pequeña que necesitan los participante.

NOTA: Si bien la interpretación de las puntuaciones E_n puede ser difícil, que no impide su uso. La incorporación de información sobre la incertidumbre en la interpretación de los resultados de ensayos de aptitud puede desempeñar un papel importante en la mejora de la comprensión de los participantes de la incertidumbre de medición y su evaluación.

9.8 Evaluación de incertidumbres de participantes en ensayos

9.8.1 Con el aumento de la aplicación de la norma ISO/IEC 17025 existe un mejor entendimiento de la incertidumbre de la medición. El uso de las evaluaciones de laboratorio de la incertidumbre en la evaluación del desempeño ha sido común en programas de ensayos de aptitud en diferentes áreas de calibración, como con las puntuaciones E_n , pero no ha sido una práctica común en las pruebas de aptitud para laboratorios de ensayo. Las puntuaciones ζ se describen en la sección 9.6, y las puntuaciones e_n en la sección 9.7, son opciones para la evaluación de los resultados en con respecto a la incertidumbre exigida.

9.8.2 Algunos proveedores de ensayos de aptitud han reconocido la utilidad de pedir a los laboratorios el informe de la incertidumbre de los resultados en las pruebas de aptitud. Esto puede ser útil, incluso cuando las incertidumbres no se utilizan en la puntuación. Hay varios propósitos para reunir dicha información:

- los organismos de acreditación pueden asegurar que los participantes están informando de incertidumbres que son coherentes con su ámbito de acreditación;
- Los participantes pueden revisar su incertidumbre en conjunto con los de otros participantes, para evaluar la coherencia (o no) y obtener así una oportunidad para determinar si su evaluación de la incertidumbre no incluye la de todos los componentes pertinentes, o está sobrevalorando la de algunos componentes;
- Los ensayos de aptitud pueden ser utilizados para confirmar las reclamaciones de incertidumbre, y esto es más fácil cuando la incertidumbre se informan junto con el resultado.

NOTA Un ejemplo del análisis de datos cuando las incertidumbres son reportadas se informa en el Anexo E.3.

9.8.3 Donde x_{pt} se determina mediante los procedimientos en las secciones 7.3-7.6 y $u(x_{pt})$ cumple el criterio en 9.2.1, de esta forma es improbable que un resultado de un participante tenga menor incertidumbre típica de este, por lo que $u(x_{pt})$ podría ser utilizada como un límite inferior para el screening, llamado u_{min} . Si el valor asignado se determina a partir de los resultados del participante (punto 7.7), entonces el proveedor de ensayos de aptitud debería determinar los límites de detección práctico u_{min} .

NOTA Si la $u(x_{pt})$ incluye la variabilidad debida a la falta de homogeneidad o inestabilidad, el participante $u(x)$ podría ser menor a u_{min} .

9.8.4 Es improbable también que para cualquier participante registrado, la incertidumbre típica sea mayor que 1,5 veces la desviación típica robusta de los participantes ($1,5 s^*$), con lo que podría utilizarse como un límite práctico para el screening de incertidumbres, llamado u_{max} .

NOTA El factor 1,5 es el límite superior de la variabilidad en las desviaciones típicas que se puede esperar de un consenso en la desviación típica con 10 o más resultados, basado en la raíz cuadrada de percentiles de la distribución F. Cualquier proveedor de ensayos de aptitud que adopte este procedimiento puede utilizar un multiplicador diferente.

9.8.5 Si la u_{min} o la u_{max} , u otros criterios, son utilizados para identificar las incertidumbres alejadas de lo esperado, las pruebas de eficiencia del proveedor deben explicar esto a los participantes y dejar claro que el informe de una incertidumbre, $u(x)$, puede ser válido incluso si es menor de u_{min} o mayor que u_{max} ; y cuando esto ocurre, los participantes y las partes interesadas deberían comprobar el resultado o la estimación de incertidumbre. Asimismo, la incertidumbre reportada puede ser más grande que u_{min} y menor u_{max} , y aun así no ser válida. Estos son indicadores solamente de carácter informativo.

9.8.6 Los proveedores de ensayos de aptitud puede también llamar la atención sobre las incertidumbres inusualmente altas o bajas basándose, por ejemplo:

- los cuartiles especificados para las incertidumbres reportadas (por ejemplo por debajo del 5^o percentil y por encima del 95^o percentil del estándar reportado o incertidumbre expandida);
- límites sobre la base de una supuesta distribución con escala basada en la dispersión de incertidumbres.
- una incertidumbre de medida requerida.

NOTA: Dado que las incertidumbres son poco probable que se distribuyan normalmente, es probable que sea necesaria la transformación cuando se utilizan límites que dependen aproximadamente o subyacentemente de la normalidad; por ejemplo, los límites basados en el rango de inter cuartiles tienen una interpretación probabilística únicamente cuando la distribución es aproximadamente normal.

9.9 Puntuaciones combinadas de eficiencia

9.9.1 Es común, en una sola serie de un programa de pruebas de eficiencia, que los resultados sean obtenidos por más de un elemento de prueba de aptitud o para más de un mesurando. En esta situación, los resultados de cada prueba de aptitud y para cada mensurando deben interpretarse como se describe en el apartado 9.3 a 9.7; es decir, los resultados de la prueba de aptitud para cada elemento y cada mensurando deben evaluarse por separado.

9.9.2 Hay aplicaciones, donde dos o más elementos de prueba de aptitud con niveles especialmente diseñados están incluidos en un sistema de ensayos de aptitud para medir otros aspectos de eficiencia como, por ejemplo, para investigar la repetibilidad, el error sistemático, o linealidad. Por ejemplo, dos elementos de prueba de aptitud similares pueden utilizarse en un sistema de ensayos de aptitud con la intención de tratarlos con el esquema de Youden, tal como se describe en 10.5. En tales casos, el proveedor de ensayos de aptitud deberá proporcionar a los participantes una descripción completa del diseño y los procedimientos estadísticos que se utilizan.

9.9.3 La métodos gráficos descritos en la sección 10 deberán utilizarse cuando los resultados son obtenidos por más de un elemento de prueba de aptitud o de varias magnitudes sometidas a medición, siempre que estén estrechamente relacionadas y/u obtenidos mediante el mismo método. Estos procedimientos combinan las puntuaciones de eficiencia de forma que no ocultan los valores altos de las puntuaciones individuales, y pueden revelar información adicional sobre el desempeño de los participantes, tales como la correlación entre los resultados de diferentes

magnitudes sometidas a medición - que no es observable en las tablas de las puntuaciones individuales.

9.9.4 En programas de ensayos de aptitud que involucran a un gran número de magnitudes sometidas a medición, un recuento o la proporción del número de acciones y señales de advertencia pueden ser utilizados para evaluar la eficiencia.

9.9.5 Las puntuaciones de eficiencia combinado o alejadas del valor central, sólo deben utilizarse con precaución, porque puede ser difícil de describir las supuestas estadísticas subyacentes de las puntuaciones. Aunque la combinación de puntuaciones de eficiencia para los resultados de la prueba de aptitud de distintos elementos en el mismo mesurando pueden tener distribuciones útiles para detectar sesgos persistentes, o sumadas las puntuaciones promedio en diferentes magnitudes sometidas a medición en el mismo o en diferentes elementos de prueba de aptitud puede disimular el sesgo en los resultados para una única magnitud sometida a medición. El método de cálculo, la interpretación, y las limitaciones de cualquier magnitud alejada del valor central o puntuaciones combinadas utilizadas deberán ser claras para los participantes.

10. Métodos gráficos para describir las puntuaciones de eficiencia

10.1 Aplicación de métodos gráficos

El proveedor de ensayos de aptitud deberá usar normalmente las puntuaciones de eficiencia obtenido en cada serie de un programa de pruebas de aptitud para preparar los gráficos tales como las descritas en el artículo 10.2 y 10.3. El uso de puntuaciones de eficiencia, tales como P_A , z , z' , ζ , o E_n puntajes en estos gráficos tiene la ventaja de que pueden trazarse con ejes normalizados, lo que simplifica su presentación e interpretación. Los gráficos deben estar a disposición de los participantes, permitiendo a cada participante ver dónde caen sus propios resultados en relación con los obtenidos por otros participantes. Los códigos de letras o número de códigos pueden ser utilizados para representar a los participantes para que cada participante sea capaz de identificar sus propios resultados, pero no es capaz de determinar los resultados obtenidos por otro. Los gráficos también pueden ser utilizados por el proveedor de ensayos de aptitud y cualquier organismo acreditador, para permitirles evaluar la eficacia global del programa de pruebas de aptitud y para ver si hay necesidad de revisar los criterios utilizados para evaluar el desempeño.

10.2 Histogramas de resultados o puntuaciones de eficiencia

10.2.1 El histograma es una herramienta estadística común, y es útil en dos puntos diferentes en el análisis de los resultados de las pruebas de aptitud. El gráfico es útil en la etapa de análisis preliminar, para verificar si las hipótesis estadísticas son razonables, o si hay una anomalía, como una distribución bimodal, una gran proporción de los valores atípicos, o asimetría inusual que no estaba prevista.

Los histogramas también pueden ser útiles en los informes de las pruebas de aptitud, para describir el esquema de puntuaciones de eficiencia, o para comparar los resultados, por ejemplo, métodos diferentes o diferentes elementos de prueba de aptitud. Los histogramas son particularmente útiles en los informes individuales de tamaño pequeño o moderados programas de ensayos de aptitud (menos de 100 participantes) para permitir a los participantes evaluar cómo se compara su eficiencia con la de otros participantes, por ejemplo, destacando un bloque dentro de una barra vertical para representar un resultado del participante o, en pequeños programas de ensayos de aptitud (menos de 50 participantes), utilizando caracteres ploteados individualizados

para cada participante.

10.2.2 Los histogramas se pueden preparar utilizando resultados de participantes reales o puntuaciones de eficiencia. Los resultados participantes tienen la ventaja de estar directamente relacionados con los datos enviados y pueden ser evaluados sin otro cálculo o transformación a partir de la puntuación de eficiencia del error de medición. Los histogramas basados en las puntuaciones de eficiencia tienen la ventaja de relacionar directamente a las evaluaciones del desempeño, y pueden ser fácilmente comparables con magnitudes sometidas a medición y series de un plan de ensayos de aptitud.

El rango y tamaño utilizado para un histograma debe determinarse para cada conjunto de datos, basado en la variabilidad y el número de resultados. A menudo es posible hacerlo sobre la base de la experiencia con las pruebas de aptitud, pero en la mayoría de las situaciones, los agrupamientos tendrán que ser ajustados después de la primera vista. Si se utilizan los resultados de eficiencia en el histograma, es útil tener una escala a partir de la desviación típica para la evaluación de la aptitud y puntos de corte para señales de alerta y acción.

10.2.3 Los intervalos de la escala y de ploteo deben elegirse de manera que se pueden detectar la bimodalidad (si está presente), sin crear falsas advertencias debido a la resolución de los resultados de la medición o un pequeño número de resultados.

NOTA 1: La apariencia de los histogramas es sensible al ancho escogido y a la ubicación de la banda (límites de ancho constante que depende en gran medida del punto de partida). Si el ancho de banda es demasiado pequeño, el gráfico mostrará muchos modos pequeños; los modos demasiado grandes y apreciables cerca del cuerpo principal pueden no ser lo suficientemente distintivos. La aparición de modos y la estrecha relación de las alturas de barras adyacentes pueden cambiar apreciablemente la posición inicial o ancho, especialmente cuando el conjunto de datos es pequeño y/o muestra algunos clústeres.

NOTA 2: Un ejemplo de un histograma ploteado es proporcionada en el Anexo E.3.

10.3 Densidad kernel ploteadas

10.3.1 Una densidad kernel ploteadas, a menudo abreviada como "Densidad de ploteo", proporciona una suave curva que describe la forma general de la distribución de un conjunto de datos. La idea subyacente en la estimación del kernel es que cada punto de datos se sustituye por una determinada distribución normal (normalmente), centrada sobre el punto y con una desviación típica σ_k ; σ_k se suele llamar el "ancho de banda". Estas distribuciones se suman y la distribución resultante, a escala para tener una unidad de superficie, da una estimación "densidad" que puede representarse como una curva suave.

10.3.2 Los siguientes pasos pueden seguirse para preparar una densidad kernel ploteadas. Se supone que un conjunto de datos compuesto de p valores x_1, x_2, \dots, x_p se incluirán en el ploteo. Estos son generalmente los resultados del participante, pero pueden ser las puntuaciones de eficiencia derivadas de los resultados.

i) Elegir un ancho de banda apropiado σ_k . Las dos opciones son particularmente útiles:

a) Para su inspección general, establezca $\sigma_k = 0,9 s^*/p^{0,2}$ donde s^* es la desviación típica robusta de los valores x_1, \dots, x_p calculados mediante los procedimientos que figuran en el anexo C.2 y C.3.

b) Al examinar el conjunto de datos para el modo Gross, importante en comparación con el criterio de evaluación del desempeño, establecer $\sigma_k = 0,75\sigma_{pt}$ si utiliza puntuaciones z o ζ , o $\sigma_k = 0,25\delta_E$ si se utiliza D o $D\%$.

NOTA 1 La opción a) de acuerdo con Silverman^[30] que recomienda s^* basado en el rango de intercuartílos (IQR). Otras reglas de selección de ancho de banda que proporcionan resultados similares incluyen la de Scott^[29], que sustituye al multiplicador de 0,9 por 1,06. Referencia^[29] que describe un método de selección de ancho de banda casi óptimo, pero mucho más complejo. En la práctica, las diferencias para la inspección visual son ligeras y la elección depende de la disponibilidad del software.

NOTA 2 La opción de la NOTA 2 b) anterior según la Guía IUPAC^[32].

- ii) Se establece un rango de trazado q_{min} a $q_{máx}$. de modo que $q_{min} \leq \min(x_1, \dots, x_p) - 3\sigma_k$ y $q_{máx} \geq \max(x_1, \dots, x_p) + 3\sigma_k$.
- iii) Elegir un número de puntos n_k para trazar la curva, $n_k = 200$ generalmente es suficiente, a menos que sean puntos extremos atípicos dentro del rango de los ploteados.
- iv) Calcular el trazado de las ubicaciones desde q_1 hasta q_{n_k} a partir de:

$$q_i = q_{min}(i - 1) \frac{(q_{n_k} - q_1)}{n_k - 1} \quad (19)$$

- v) Calcular las densidades desde n_k hasta h_{n_k} a partir de:

$$h_i = \frac{1}{p} \sum_{j=1}^p \varphi \left(\frac{x_j - q_i}{\sigma_k} \right) \text{ desde } i = 1 \text{ hasta } i = n_k \quad (20)$$

Donde $\varphi(\cdot)$ indica la densidad normal típica.

- vi) Ploteo de h_i con respecto a q_i .

NOTA 1 Puede ser útil agregar las ubicaciones de los puntos de datos individuales a la ploteados. Esto es más comúnmente realizado para representar las ubicaciones por debajo de la curva de densidad de trazados como marcadores verticales cortos (a veces denominados "rug"), pero también puede hacerse por trazar los puntos de datos en los puntos adecuados a lo largo de la curva de densidad calculada.

NOTA 2 El ploteo de densidad se realiza mejor por software. El anterior cálculo gradual del conjunto de datos puede hacerse en una hoja de cálculo de modestas dimensiones. Patentado con software estadístico disponible libremente a menudo incluye puntos de densidad ploteados, basados en ancho de banda predeterminado con opciones similares. La implementación de software avanzado de densidad ploteada puede utilizar este algoritmo de cálculo basado en métodos de densidad.

NOTA 3 Ejemplos ploteados de densidad de kernel figuran en los anexos E.3, E.4 y E.6.

10.3.3 La forma de la curva es, tomado como una indicación de la distribución a partir de la cual se extrajeron los datos. Los diferentes modos aparecen como picos separados. Los valores vecinales aparecen como picos separados bien separados del cuerpo principal de los datos.

NOTA 1 Un ploteo de densidad es sensible al ancho de banda σ_k elegido. Si el ancho de banda es demasiado pequeño, el gráfico mostrará muchos modos pequeños; y modos demasiados grandes, apreciables, cerca del cuerpo principal y puede no ser lo suficientemente específica.

NOTA 2 Semejante a los histogramas, la densidad de ploteo se utiliza mejor con conjuntos de datos moderados a grandes porque los conjuntos de datos pequeños (10 o menos) pueden incluir por casualidad, leves o aparentes modos atípicos, especialmente cuando la desviación típica robusta se utiliza como base para el ancho de banda.

10.4 Ploteos de barras de puntuaciones de eficiencia estandarizados

10.4.1 Los ploteos de barras son un método adecuado de presentar las puntuaciones de eficiencia para un número de características similares en un gráfico. Pueden revelar si hay alguna característica común en los puntajes de un participante, por ejemplo, si un participante obtiene varios puntajes altos de z indica generalmente una eficiencia deficiente, el participante puede tener sesgo positivo.

10.4.2 Para preparar un ploteo de barras, se recogen las puntuaciones de eficiencia estandarizados en un gráfico, como se muestra en la figura E.10, en el que las puntuaciones para cada participante se agrupan unidas. Otras puntuaciones de eficiencia estandarizados, como $D\%$ o P_A pueden ser trazadas con el mismo propósito.

10.4.3 Cuando se repliquen las determinaciones en una serie de un programa de pruebas de aptitud, los resultados pueden ser utilizados para calcular un gráfico de medidas de precisión; por ejemplo, k estadísticas como se describe en ISO 5725 - 2, o una escala de medida relacionados contra la desviación típica robusta promedio tal y como es definida en el algoritmo S (Anexo C.4).

NOTA Un ejemplo de un ploteo de barra con puntajes z se proporciona en el Anexo E.11.

10.5 Ploteos de Youden

10.5.1 Cuando dos elementos de prueba de aptitud similares han sido probados en una serie de un programa de pruebas de aptitud, los ploteos de Youden proporcionan un método gráfico muy informativo de estudiar los resultados. Puede ser útil para demostrar la correlación (o la independencia) de los resultados de la prueba de aptitud para distintos elementos, y para orientar las investigaciones, con razones, para la toma de decisiones.

10.5.2 El gráfico está construido ploteando los resultados de los participantes o los puntajes z , obtenidos en la prueba de aptitud de uno de los elementos contra los resultados del participante o z puntuaciones obtenidas en la otra prueba de aptitud de los elementos. Las líneas verticales y horizontales son típicamente dibujadas para crear cuatro cuadrantes de valores, para ayudar a la interpretación. Las líneas se dibujan en los valores asignados o en las medianas de las dos distribuciones de resultados o dibujando al 0, si son trazadas las puntuaciones z .

NOTA: Para la correcta interpretación de ploteos de Youden es importante que los dos elementos de prueba de aptitud tengan similares (o idénticos) niveles del mesurando; además de que la naturaleza de cualquier error sistemático de medición es el mismo en esa zona del intervalo de medición. Los ploteos de Youden puede ser útiles para los diferentes niveles de un mesurando, consistente en presencia de error sistemático, pero puede ser engañosos si hay un error de calibración que no sea constantemente positivo o negativo en todo el rango del nivel de medición.

10.5.3 Cuando se confecciona un ploteo de Youden, se interpreta como sigue:

- a) Inspeccionar el ploteo de los puntos verificando que estén bien separados del resto de los datos. Si un participante no está siguiendo el método de prueba correctamente, de modo que sus resultados estén sujetos a un error sistemático, un punto dado puede estar ahora en la parte inferior izquierda o superior derecha de los cuadrantes. Los puntos alejados de los demás en la esquina superior izquierda e inferior derecha de los cuadrantes representan a aquellos participantes cuya repetibilidad es más grande que la de la mayoría de los demás participantes, cuyos métodos de medición muestran diferente sensibilidad a la prueba de aptitud del elemento o, a veces, debido a los participantes que accidentalmente han intercambiado elementos de pruebas de aptitud.
- b) Inspeccionar el ploteo para ver si hay evidencia de una relación general entre los resultados de dos elementos en la prueba de aptitud (por ejemplo, si se encuentran aproximadamente a lo largo de una línea inclinada). Si hay evidencia de una relación y, a continuación, muestra que hay evidencia de sesgos de participantes que afecten a diferentes elementos de las pruebas de aptitud de forma similar. Si no existe aparente relación visual entre los resultados (por ejemplo, puntos que se distribuyen aproximadamente de forma uniforme en una región circular, generalmente con mayor densidad hacia el centro) que los errores de medición para los dos elementos de pruebas de aptitud que son en gran medida independientes. Esto puede comprobarse con un rango de correlación estadística, si el examen visual no es concluyente.
- c) Inspeccionar el ploteo para grupos cerrados de participantes, tanto a lo largo de la diagonal o en otros lugares. La presencia de grupos perfectos probablemente indica diferencias entre los distintos métodos.

NOTA 1 En estudios donde todos los participantes utilizan el mismo método de medición, o ploteos de resultados son a partir de un método único de medición, si los resultados se encuentran a lo largo de una línea, esto puede ser una evidencia de que el método de medición no ha sido adecuadamente especificado. La investigación del método de ensayo puede permitir la reproducibilidad que permita ser generalmente mejorado.

NOTA 2 Un ejemplo de un ploteo de Youden es proporcionado en el Anexo E.12.

10.6 Ploteos de repetibilidad de desviaciones típicas

10.6.1 Cuando las réplicas de las mediciones realizadas por los participantes se realizan en un esquema de pruebas de aptitud, los resultados pueden ser usados para obtener un gráfico a fin de identificar los participantes cuyo promedio y desviación típica son inusuales.

10.6.2 El gráfico está construido ploteando la desviación típica s_i para cada participante, contra el promedio correspondiente x_i para dicho participante. Alternativamente, el rango de resultados duplicados puede utilizarse en lugar de la desviación típica. Quedando

x^* = Promedio robusto de x_1, x_2, \dots, x_p , calculado por el algoritmo A

w^* = Promedio robusto combinado de s_1, s_2, \dots, s_p , calculado por el algoritmo S

Y asumiendo que los datos se distribuyen normalmente. Bajo la hipótesis nula de que no hay ninguna diferencia entre los participantes en los valores de la población media o dentro de desviaciones típicas de los participantes, la estadística

$$\left(\sqrt{m} \frac{x_i - x^*}{w^*} \right)^2 + \left(\sqrt{2(m-1)} \ln \left(\frac{s_i}{w^*} \right) \right)^2 \quad (21)$$

Tiene aproximadamente la distribución χ^2 con 2 grados de libertad. Por lo tanto una región crítica, con un nivel de significación de aproximadamente el 1 % puede ser dibujado sobre el gráfico poteando

$$s = w^* \exp \left[\pm \frac{1}{\sqrt{2(m-1)}} \sqrt{\chi_{2,0.99}^2 - \left(\sqrt{m} \frac{x - x^*}{w^*} \right)^2} \right] \quad (22)$$

Sobre la desviación típica del eje con respecto a las x en el promedio del eje de las abscisas para

$$x = x^* - w^* \sqrt{\frac{\chi_{2,0.99}^2}{m}} \quad \text{to} \quad x^* + w^* \sqrt{\frac{\chi_{2,0.99}^2}{m}} \quad (23)$$

NOTA: Este procedimiento se basa en la Técnica de círculo introducida por van Nuland [36]. El método descrito utiliza una simple aproximación normal de la distribución de la desviación típica que podría dar una región crítica negativa que contienen desviaciones típicas. El método dado aquí utiliza una aproximación de la distribución de la desviación típica que evita este problema, pero la región crítica ya no es un círculo como en el original. Además, se utilizarán los valores robustos para el punto central en lugar de promedios simples como en el método original.

10.6.3 El ploteo puede indicar a los participantes con sesgo la parcialidad que es inusualmente grandes, dada su repetibilidad. Si hay un gran número de réplicas, esta técnica también puede identificar a los participantes con la repetibilidad excepcionalmente pequeña. Sin embargo, debido a que usualmente hay un pequeño número de repeticiones, las interpretaciones son difíciles.

NOTA Un ejemplo de un gráfico de desviaciones típica de la repetibilidad se proporciona en el Anexo E.13.

10.7 Muestras divididas

10.7.1 La opción de dividir las muestras se utiliza cuando es necesario realizar una comparación detallada de dos participantes, o cuando no se dispone de pruebas de aptitud y es necesario efectuar alguna verificación externa. Son obtenidas muestras de varios materiales, representando un amplio rango de las propiedades de interés, cada muestra se divide en dos partes, y cada laboratorio obtiene cierto número (al menos dos) de réplicas de las determinaciones de una parte de cada muestra.

En ocasiones, más de dos participantes pueden estar involucrados, en cuyo caso debe ser tratada como una referencia, y las demás deben ser comparadas con ella utilizando las técnicas descritas aquí.

NOTA 1 Este tipo de estudio es común, pero a menudo con nombres diferentes, como "muestras pareadas" o "comparaciones bilaterales".

NOTA 2 Este diseño de división de muestras no debe confundirse con el diseño "nivel split" utilizado en la ISO 5725, que implica dos elementos de prueba con niveles ligeramente diferentes suministrados a todos los participantes.

10.7.2 Los datos a partir de un diseño de muestra dividida pueden ser utilizados para obtener los gráficos que muestran la variación entre mediciones repetidas para los dos participantes y las diferencias entre el promedio de sus resultados para cada elemento de prueba de aptitud. El ploteo bivariado utilizando toda la gama de concentraciones puede tener una escala donde sea difícil identificar diferencias importantes entre los participantes, además, el ploteo de las diferencias o de los diferentes porcentajes entre resultados de los dos participantes puede ser más útil. El análisis ulterior dependerá de las deducciones realizadas a partir de estos gráficos.

10.8 Métodos gráficos para combinar las puntuaciones de eficiencia a través de pruebas de un esquema de ensayos de aptitud

10.8.1 Cuando las puntuaciones de eficiencia estandarizadas se combinan a través de varias series de un esquema de ensayos de aptitud, las puntuaciones de eficiencia pueden considerar la elaboración de gráficos, tal como se describe en el apartado 10.8.2 o 10.8.3. El uso de estos gráficos, en el que las puntuaciones de eficiencia de varias series de un esquema de ensayos de aptitud son combinadas, puede permitir que sean determinadas las tendencias, y otras características de los resultados, que no son evidentes cuando se examinan por separado las puntuaciones de eficiencia para cada serie.

NOTA Con el uso de "running scores" o "puntuaciones acumulativas", en las que la eficiencia obtenida por un participante se combina a través de varias series de un plan de ensayos de aptitud, las puntuaciones de eficiencia deben mostrarse gráficamente. El participante puede tener un error que se muestre con el elemento utilizado en la prueba de aptitud en una sola serie, pero no en las demás; una puntuación en curso podría ocultar este error. Sin embargo, en algunas circunstancias (por ejemplo, con series frecuentes) "suavizados" de ocasionales puntuaciones alejadas del valor central, pueden ser útiles para demostrar la eficiencia primaria con más claridad.

10.8.2 Las cartas de control de Shewhart son un método efectivo para identificar los problemas que causan grandes valores erráticos de puntajes z. Véase ISO 7870 - 2^[6] para información sobre ploteo de cartas y reglas de Shewhart así como sus límites de acción.

10.8.2.1 Para preparar estas cartas, son ploteadas las puntuaciones estandarizadas, tales como puntuaciones z o P_A , para un participante se representan como puntos individuales, con acción y límites establecidos de advertencia que concuerden con el diseño para el esquema de ensayos de aptitud. Cuando son medidas varias características en cada serie, las puntuaciones de eficiencia pueden ser trazadas en el mismo gráfico, pero los puntos para las diferentes características deben representarse utilizando diferentes símbolos y/o colores. Cuando varios elementos se incluyen en la misma serie del programa de pruebas de aptitud las puntuaciones de eficiencia se pueden trazar junto con múltiples puntos en cada período de tiempo. Las líneas que unen las puntuaciones medias en cada momento también pueden ser añadidas al ploteo.

10.8.2.2 Reglas convencionales para interpretar las cartas de control de Shewhart. Cuando hay señal de puntos fuera de control, éstas se interpretan de la siguiente manera:

- un punto cae fuera de los límites de acción ($\pm 3,0$ para puntajes z, o 100 % de P_A);
- dos de tres puntos sucesivos caen fuera de los límites de advertencia ($\pm 2,0$ para puntuaciones z o 70 % de P_A);
- seis resultados consecutivos, ya sean positivos o negativos.

10.8.2.3 Cuando una carta de control de Shewhart proporciona una señal de control fuera de los límites, el participante debe investigar las posibles causas.

NOTA La desviación típica para la evaluación de la aptitud de σ_{pi} generalmente no es la desviación típica de las diferencias ($x_i - x_{pi}$), de modo que los niveles de probabilidad que suelen asociarse con la acción y los límites de advertencia de un gráfico de control de Shewhart pueden no ser aplicables.

10.8.3 Cuando el nivel de una propiedad de una serie de ensayos de aptitud varía de un esquema a otro, los indicadores de eficiencias estandarizados ploteados, tales como las puntuaciones z y P_A , con respecto al valor asignado mostrarán si el sesgo del participante cambia con el nivel. Cuando más de una prueba de aptitud para una propiedad está incluida en la misma serie las puntuaciones de eficiencia pueden representarse de forma independiente.

NOTA 1 Puede ser útil tener un símbolo o color diferente de trazado para los resultados de la presente serie de ensayos de aptitud, a fin de distinguir el punto(s) de las series anteriores.

NOTA 2 Se muestra un ejemplo de ploteo en el Anexo E.14, utilizando una puntuación P_A . Este ploteo podría emplear fácilmente z, con sólo un cambio en la escala vertical.

11. Diseño y análisis de esquemas de ensayos de eficiencia cualitativa (incluidas las propiedades ordinales y nominales)

11.1 Tipos de datos cualitativos

Una gran cantidad de ensayos de aptitud se realizan para las propiedades evaluadas o identificados en escalas cualitativas. Esto incluye lo siguiente:

- Programas de ensayos de aptitud que requieren la presentación de informes sobre una escala categórica (a veces llamados "nominal"), donde el valor de la propiedad no tiene la magnitud (como, por ejemplo, un tipo de sustancia u organismo);
- Programas de ensayos de aptitud para determinar la presencia o ausencia de una propiedad, ya sea determinado por criterios subjetivos o por la magnitud de la señal de un procedimiento de medición. Esto puede considerarse como un caso especial de una escala categórica u ordinal, con sólo dos valores (también llamados "dicotómico", o binario);
- Programas de ensayos de aptitud que requieren los resultados reportados en una escala ordinal, que pueden ser ordenados de acuerdo a su magnitud, pero para los cuales no existen relaciones aritméticas entre los diferentes resultados. Por ejemplo, "alto, medio y bajo" en forma de una escala ordinal.

Tales programas de ensayos de aptitud que requieren especial consideración para el diseño, la asignación de valor y etapas de evaluación de desempeño (puntuación) porque:

- los valores asignados son muy frecuentemente basados en la opinión de expertos; y
- el tratamiento estadístico diseñado para valores continuos y el recuento de los datos no es aplicable a los datos cualitativos. Por ejemplo, no tiene sentido tomar medias y desviaciones típicas de resultados de escala ordinal incluso cuando pueden ser colocados en un orden de prioridades.

En consecuencia los párrafos siguientes proporcionan orientación sobre el diseño, la asignación de valor y eficiencia para la evaluación cualitativa de programas de ensayos de aptitud.

NOTA Guía para datos ordinales que no se aplica a los resultados de las mediciones, basados en una escala cuantitativa con indicaciones discontinuas (como diluciones o títulos), véase la sección 5.2.2.

11.2 Diseño estadístico

11.2.1 Para programas de ensayos de aptitud en los que la opinión de los expertos es esencial tanto para la asignación de valor o para la evaluación de informes de participantes, normalmente será necesario reunir un panel de expertos debidamente calificados y dar tiempo para el debate a fin de lograr el consenso sobre la asignación apropiada. Donde hay una necesidad de confiar en los expertos individuales para puntuar o dar el valor asignado por el proveedor de ensayos de aptitud que debe además proporcionar el aseguramiento para la evaluación y el control de la consistencia de opinión entre los diferentes expertos.

Ejemplo En un esquema de ensayos de aptitud clínica que se basa en la microscopía para el diagnóstico, la opinión de los expertos se utiliza para evaluar los portaobjetos proporcionados a los participantes y el diagnóstico clínico adecuado de los elementos de prueba de aptitud. El proveedor de ensayos de aptitud puede elegir elementos de prueba de aptitud para estudios "ciegos" a los diferentes miembros del grupo de expertos para asegurar la consistencia del diagnóstico o realizar ejercicios periódicos para evaluar el consenso del panel.

11.2.2 Para esquemas de ensayos de aptitud de un simple reporte, un solo valor categórico u ordinal, los resultados de las pruebas de aptitud del proveedor deben considerar

- proporcionar dos o más elementos de prueba de aptitud por serie; o
- solicitar los resultados de una serie de observaciones replicadas en cada elemento de prueba de aptitud, con el número de repeticiones especificado previamente.

Cualquiera de estas estrategias permite contar de resultados para cada participante de forma que puede utilizarse tanto en el examen de datos o en la puntuación. El suministro de dos o más elementos de prueba de aptitud puede proporcionar información adicional sobre la naturaleza de los errores y también permitir ensayos de aptitud más sofisticados para una mejor evaluación de la eficiencia.

Ejemplo 1 En un programa de pruebas de aptitud destinado a informar sobre la presencia o ausencia de un contaminante, la provisión de elementos de prueba de aptitud que contiene un rango de niveles de contaminante permite examinar al proveedor de ensayos de aptitud el número de detecciones de éxito en cada nivel, en función del nivel de contaminante presente. Esto puede ser utilizado, por ejemplo, para proporcionar información a los participantes sobre la capacidad de detección de su método de prueba elegido, o para obtener una promedio de probabilidad de detección, lo cual a su vez puede permitir puntuaciones de eficiencia para ser asignadas a los participantes sobre la base de probabilidades estimadas de patrones particulares de respuesta.

Ejemplo 2 Las pruebas de eficiencia en comparaciones forenses a menudo requieren la coincidencia de elementos de prueba de aptitud si provienen de la misma fuente o fuentes distintas (por ejemplo, huellas dactilares, ADN, casquillos de bala, huellas, etc.). En muchos casos "Indeterminado" es una respuesta permitida. Un plan de ensayos de aptitud puede incluir varios elementos de las pruebas de aptitud de

distintas fuentes, y se pide a los participantes que su estado sea del "mismo origen", "diferente", o "indeterminado" para cada pareja. Este objetivo permite puntuaciones de números (o %) correctas o incorrectas, o el número (%) correcto coincide, o corregir los rechazos. Los criterios de desempeño pueden determinarse sobre la idoneidad para el uso, o el grado de dificultad de la tarea.

11.2.3 La homogeneidad debe ser demostrada con el examen de una muestra adecuada de los elementos de prueba de aptitud, todos los cuales deben demostrar el valor esperado de la propiedad. Para algunas propiedades cualitativas, por ejemplo, presencia o ausencia, es posible verificar la homogeneidad con mediciones cuantitativas; por ejemplo, un recuento microbiológico o un espectro de absorbancia por encima de un umbral. En estas situaciones, una prueba de homogeneidad convencional puede ser adecuada, o una demostración de todos los resultados que están por encima o por debajo de un valor límite.

11.3 Valores asignados para esquemas de ensayos de aptitud cualitativos

11.3.1 Los valores pueden ser asignados a los elementos de prueba de aptitud:

- a) por el juicio de expertos.
- b) por el uso de materiales de referencia como elementos de pruebas de aptitud;
- c) a partir del conocimiento del origen o la preparación del elemento de la prueba de aptitud (s);
- d) usando el modo o la mediana de los resultados del participante (la mediana es apropiada sólo para valores ordinales).

Cualquier otro método de asignación de valor que pueda ser demostrado que proporcionen resultados fiables también puede ser utilizado. En los siguientes párrafos se examina cada una de las estrategias mencionadas.

NOTA No es generalmente apropiado para proporcionar información cuantitativa sobre la incertidumbre del valor asignado en programas de ensayos de aptitud cualitativos. Cada uno de los párrafos 11.3.2 a 11.3.5 no obstante requieren el suministro de información básica relativa a la confiabilidad en el valor asignado de manera que los participantes puedan juzgar si un mal resultado pudiera ser razonablemente atribuible a un error en el valor asignado.

11.3.2 Los valores asignados por la opinión de expertos normalmente deberían basarse en un consenso de un grupo de expertos debidamente calificados. Cualquier desacuerdo significativo entre los participantes debe estar registrado en el informe de la serie. Si el grupo no puede llegar a un consenso para un determinado elemento de prueba de aptitud, el proveedor de ensayos de aptitud puede considerar un método alternativo de asignación de valor de aquellos enumerados en la Sección 11.3.1. Si esto no es apropiado, el elemento de la prueba de aptitud no debe ser utilizado para la evaluación del desempeño de los participantes.

NOTA: En algunos casos, es posible que un solo experto determine el valor asignado.

11.3.3 Cuando se proporciona un material de referencia a los participantes de un elemento en un ensayo de aptitud, el valor de referencia asociado, o valor certificado, normalmente deberá ser utilizado como el valor asignado a la serie. Toda la información facilitada con el material de referencia se refiere a la confiabilidad en el valor asignado que debe estar a disposición de los participantes en la serie siguiente.

NOTA: Las limitaciones de este enfoque son relacionadas en la Sección 7.4.1.

11.3.4 Cuando los elementos de las pruebas de aptitud son preparados a partir de una fuente conocida, el valor asignado puede ser determinado basándose en el origen del material. Las pruebas de eficiencia del proveedor deben mantener un registro del origen, el transporte y la manipulación del material(es) utilizado(s). Debido a que debe tenerse cuidado para evitar la contaminación que pueda derivar resultados incorrectos de los participantes. Las pruebas de origen y/o detalle de preparación deben estar a disposición de los participantes después de la serie previa solicitud o como parte del informe para la serie de ensayos de aptitud.

Ejemplo Prueba de elementos de aptitud de vino distribuido para un esquema de ensayo de aptitud de autenticidad directamente desde un productor adecuado en la denominada región de origen, o a través de un proveedor comercial capaz de ofrecer garantía de autenticidad.

11.3.4.1 Las pruebas confirmatorias o mediciones son recomendadas siempre y cuando sea posible, especialmente en los casos en que la contaminación puede comprometer la utilización como elemento de prueba de aptitud. Por ejemplo, una prueba de aptitud del elemento identificado como un simple microbio, especies vegetales o animales, normalmente debe ser ensayada para evaluar la respuesta a las pruebas de otras especies importantes. Dichas pruebas, deberían ser tan sensibles como sea posible para asegurar que está ausente la contaminación de las especies o que el nivel de contaminación está cuantificado.

11.3.4.2 El proveedor de ensayos de aptitud deberá proporcionar información sobre cualquier contaminación detectada o sobre las posibles dudas acerca del origen que puedan comprometer la utilización del elemento de la prueba de aptitud.

NOTA Mayor detalle sobre la caracterización de tales elementos de prueba de aptitud está más allá del alcance de esta norma internacional.

11.3.5 El modo (la observación más común) puede ser usado como el valor asignado a los resultados en una escala categórica u ordinal, mientras que la mediana puede ser utilizada como el valor asignado a los resultados en una escala ordinal. Cuando se utilizan estos estadígrafos, el informe de la serie de las pruebas de aptitud deberá incluir una declaración de la proporción de los resultados utilizados en la asignación de valores que coinciden con el valor asignado. Nunca es apropiado calcular medias o desviaciones típicas para los resultados de las pruebas de aptitud para las propiedades cualitativas, incluyendo valores ordinales. Esto es porque hay relación aritmética entre valores diferentes en cada escala.

11.3.6 Cuando los valores asignados están basados en mediciones (por ejemplo, presencia o ausencia), el valor asignado normalmente puede determinarse de manera definitiva; es decir, con bajos niveles de incertidumbre. Los cálculos estadísticos de incertidumbre pueden ser apropiados para los niveles del mensurando en niveles "indeterminado" o "equivoca".

11.4 Evaluación de desempeño para programas de ensayos de aptitud cualitativos y de puntuación

11.4.1 La evaluación del desempeño de los participantes en un esquema de ensayos de aptitud cualitativa depende en parte de la naturaleza del informe requerido. En algunos programas de ensayos de aptitud, donde se requiere una cantidad significativa de evaluaciones de los participantes, las conclusiones requieren un cuidadoso examen y redacción de los informes, los reportes de los participantes pueden pasar a los expertos para su evaluación y puede darse una

marca global. En el otro extremo, los participantes podrán ser juzgados únicamente si su resultado coincide exactamente con el valor asignado al elemento de la correspondiente prueba de aptitud. En consecuencia los párrafos siguientes proporcionan orientación sobre la evaluación del desempeño y de puntuación para una amplia gama de circunstancias.

11.4.2 La valoración por los expertos de los informes de los participantes requiere de uno o más expertos para revisar cada informe de participante para cada elemento de prueba de aptitud y asignar un valor de eficiencia o puntuación. En tal esquema de ensayos de aptitud, el proveedor de ensayos de aptitud debe garantizar que:

- el participante en particular no es conocido por el experto. En particular, el informe que pasa al (los) experto(s) no debe incluir ninguna información que pueda identificar razonablemente el participante;
- la revisión, identificación y evaluación de la eficiencia siga un conjunto de criterios previamente acordados, que sea tan objetiva como razonablemente sea posible.
- las disposiciones del apartado 11.3.2 con respecto a la coherencia entre los expertos se cumplen;
- En la medida de lo posible, se prevé la apelación de un participante en particular contra la opinión de los expertos y/o la revisión de las opiniones secundarias acerca de cualquier umbral de eficiencia importante.

11.4.3 Pueden utilizarse dos sistemas para marcar un único informe de resultado cualitativo basado en un valor asignado:

- i) Cada resultado es señalado como aceptable (o registrado como un éxito) si coincide exactamente con el valor asignado y es señalado como inaceptable, o dando una puntuación de eficiencia adversa, de lo contrario.

Ejemplo en un esquema para determinar la presencia o ausencia de un contaminante, los resultados correctos se califican como 1 y los resultados incorrectos como 0.

- ii) Los resultados que coincidan exactamente con el valor asignado se marcan como aceptable y se les da una puntuación correspondiente; los resultados que no coincidan exactamente con el valor asignado se les da una puntuación que depende de la naturaleza de la discrepancia. Los diseños para dicha calificación deben asignar puntuaciones inferiores a una mejor eficiencia, para que sean coherentes con otros tipos de puntuaciones de eficiencia (por ejemplo, la puntuaciones z, P_A , ζ , y E_n).

Ejemplo 1 En un esquema de ensayos de aptitud en patología clínica, el proveedor asigna una puntuación de "0" para una exacta y correcta identificación de una especie microbiológica, el punto "1" para obtener un resultado que es incorrecto pero que no cambiaría el tratamiento clínico (por ejemplo, identificación de relacionadas especies microbiológicas pero diferentes que requieren un tratamiento similar), y 3 puntos para una identificación que es incorrecta y conduciría a un tratamiento incorrecto para un paciente. Para este sistema de puntuación se suele requerir la opinión de expertos sobre la naturaleza de la disconformidad, tal vez obtenido antes de la puntuación.

Ejemplo 2 En un sistema de ensayos de aptitud para que seis posibles respuestas clasificadas en una escala ordinal son posibles, un resultado que coincide con el valor asignado que se le da un puntaje de 0 y la

puntuación se incrementa en 2 para cada diferencia en rango hasta la puntuación que aumenta hasta un máximo de 6 (para un resultado adyacente al valor asignado con una puntuación de 2).

Las puntuaciones de eficiencia individual para cada elemento de prueba de aptitud deben ser proporcionadas a los participantes. Donde las réplicas de las observaciones que se realizan como un resumen de puntuaciones de eficiencia para cada resultado pueden ser proporcionadas.

11.4.4 Donde son reportadas múltiples repeticiones para cada elemento de prueba de aptitud o donde varios elementos de prueba de aptitud se proporcionan a cada participante, el proveedor de ensayos de aptitud puede calcular y utilizar las puntuaciones de eficiencia combinando los resúmenes de puntuación en la evaluación de la eficiencia. Las puntuaciones de eficiencia combinada o los resúmenes pueden calcularse como, por ejemplo:

- la simple suma de las puntuaciones de eficiencia en todos los puntos de la prueba de aptitud.
- el recuento de cada nivel de eficiencia asignado;
- la proporción de resultados correctos,
- un estadígrafo basado en las diferencias entre los resultados y los valores asignados.

Ejemplo Un estadígrafo muy general utilizado a veces para datos cualitativos estadísticos es el coeficiente de Gower^[20]. Este puede combinar las variables cuantitativas y cualitativas sobre la base de una combinación de las puntuaciones para la similitud. Para datos binarios o categóricos el índice asigna una puntuación de 1 para que coincida exactamente con las categorías y 0 en caso contrario; para escalas ordinales se asigna una puntuación igual a 1 menos la diferencia en rango dividido por el número de filas disponibles, y para el intervalo o escala de relación datos que asigna una puntuación igual a 1 menos la diferencia absoluta dividido por el rango observado de todos los valores. Estas calificaciones, que son necesariamente de 0 a 1, se suman y la suma es dividida por el número de variables utilizadas. Una variante ponderada también puede ser utilizada.

Las puntuaciones de eficiencia combinadas pueden estar asociadas con un resumen de la evaluación de los resultados. Por ejemplo, particular (generalmente alto) proporción de puntuación correcta que puede ser considerada una eficiencia "aceptable", si es coherente con los objetivos de la serie de ensayos de aptitud.

11.4.5 Para proporcionar información sobre el desempeño a los participantes, o para proporcionar información resumida en un informe sobre una ronda, se pueden utilizar métodos gráficos.

NOTA Un ejemplo del análisis de datos ordinales se presenta en el Anexo E.15.

Anexo A (Normativo)

Símbolos

d	Diferencia entre un valor de medición para un elemento de prueba de aptitud y un valor asignado para un CRM
\bar{d}	Diferencia media entre los valores de medición y el valor asignado para un CRM
D	Diferencia participante del valor asignado (x_{xpt})
$D\%$	Diferencia participante del valor asignado expresado como un porcentaje de x_{pt}
δ_E	Criterio de error máximo permisible de las diferencias
δ_{hom}	Error debido a la diferencia entre los elementos de prueba de aptitud
δ_{stab}	Error debido a la inestabilidad durante el período de ensayos de aptitud
δ_{trans}	Error debido a la inestabilidad en condiciones de transporte
E_n	"Error, normalizado" resultado que incluye incertidumbres para el resultado de los participantes y el valor asignado
g	Número de elementos de prueba de aptitud probado en un control de homogeneidad
m	Número de mediciones repetidas realizadas por elemento de prueba de aptitud
p	Número de participantes que tomaron parte en una ronda de un programa de ensayos de aptitud
P_A	Proporción de error permitido ($D/\delta E$), se puede expresar como un porcentaje
s_r	Estimación de la desviación típica de la repetibilidad
s_R	Estimación de la desviación típica de la reproducibilidad
s_s	Estimación de la desviación típica entre la muestra
s^*	Estimación robusta de la desviación típica de los participantes
$s_{\bar{x}}$	La desviación típica de los promedios de la muestra
s_w	Desviación típica dentro de la muestra o dentro del laboratorio
σ_k	Desviación típica de ancho de banda utilizado para gráficos de densidad kernel
σ_L	Desviación típica entre laboratorios (o participante)
σ_{pt}	La desviación típica para la evaluación de la competencia
σ_r	Desviación típica de la repetibilidad
σ_R	Desviación típica de la reproducibilidad
u_{hom}	Incertidumbre típica debido a la diferencia entre los elementos de prueba de aptitud
u_{stab}	Incertidumbre típica debido a la inestabilidad durante el período de ensayos de aptitud

u_{trans}	Incertidumbre típica debido a la inestabilidad en condiciones de transporte
$u(x_i)$	Incertidumbre Típica de un resultado del participante i
$u(x_{pt})$	Incertidumbre típica del valor asignado
$u(x_{ref})$	Incertidumbre típica de un valor de referencia
$U(x_i)$	Incertidumbre expandida del resultado reportado por el participante i
$U(x_{pt})$	Incertidumbre expandida del valor asignado
$U(x_{ref})$	Incertidumbre expandida de un valor de referencia
w_t	Rango entre-test-porción
w^*	Estimación robusta de la repetibilidad del participante
x	Resultado de la medición (genérico)
x_{char}	Valor de la propiedad obtenida a partir de la determinación del valor asignado
x_{CRM}	Valor asignado para una propiedad en un Material de Referencia Certificado
x_i	Resultado de la medición del participante i
x_{pt}	Valor asignado
x_{ref}	Valor de referencia para un propósito declarado
x^*	Estimación robusta significativa del participante
\bar{x}	Media aritmética de un conjunto de resultados
z	Puntuación utilizado para la evaluación de la competencia
z'	Puntuación z score que incluye la incertidumbre del valor asignado
ζ	Zeta score – Modificado z score que incluye incertidumbres para el resultado de los participantes y el valor asignado

**Anexo B
(Normativo)****La homogeneidad y la estabilidad de los elementos de prueba de aptitud****B.1 procedimiento general para un control de homogeneidad**

B.1.1 Para llevar a cabo una evaluación de la homogeneidad para una mayor parte de la preparación de los elementos de la prueba de aptitud, siga el procedimiento que se indica a continuación:

Elija una propiedad (o propiedades) o mensurando(s) para evaluar con el control de la homogeneidad.

Elige un laboratorio para llevar a cabo la verificación de la homogeneidad y un método de medición a usar. El método debe tener una desviación típica de repetibilidad (s_r) suficientemente pequeño para que cualquier falta de homogeneidad significativa pueda ser detectada. La relación de la desviación típica de la repetibilidad del método de la desviación típica para la evaluación de la competencia debe ser inferior a 0,5, como se recomienda en el Protocolo Armonizado IUPAC (o 1/6 de δ_E). Se reconoce que esto no siempre es posible, por lo que en ese caso el proveedor de ensayos de aptitud debe utilizar más repeticiones.

Preparar y empaquetar los artículos de la prueba de aptitud para una ronda del programa de ensayos de aptitud, asegurando que existen suficientes elementos de prueba de aptitud para los participantes en el programa de ensayos de aptitud y para la comprobación de la homogeneidad. Seleccione un número g de los ítems de la prueba de aptitud en su forma envasada final utilizando un proceso de selección al azar adecuado, donde $g \geq 10$. El número de elementos de la prueba de competencia incluidas en el control de la homogeneidad puede reducirse si se dispone de datos adecuados de los controles de homogeneidad anteriores en elementos de prueba de aptitud similares preparados por los mismos procedimientos.

Preparar $m \geq 2$ porciones de ensayo de cada ítem de ensayos de aptitud mediante técnicas apropiadas para el ítem de ensayos de aptitud para reducir al mínimo las diferencias entre-test-porción.

Tomar las porciones de ensayo $g \times m$ en un orden aleatorio, obtener un resultado de medición en cada uno, completando todas las series de mediciones en condiciones de repetibilidad.

Calcular el promedio general \bar{x} , la desviación típica dentro de la muestra s_{m} y la desviación típica entre muestras s_s , como se muestra en B.3.

B.1.2 Cuando no es posible llevar a cabo mediciones duplicadas, por ejemplo, con pruebas destructivas, entonces la desviación típica de los resultados se puede utilizar como s_s . En esta situación es importante tener un método con una desviación típica de repetibilidad s_r suficientemente baja.

B.2 Criterios de evaluación para el control de homogeneidad

B.2.1 Los siguientes tres controles deben ser utilizados para asegurar que los datos de las pruebas de homogeneidad son válidos para el análisis:

a) Examinar los resultados de cada porción de ensayo con el fin de buscar una tendencia (o deriva) para la medición en el análisis; si hay una aparente tendencia, tomar las medidas correctivas apropiadas en relación con el método de medición, o tenga cuidado en la interpretación de los resultados.

b) Examinar los resultados de los promedios de los ítem de ensayos de aptitud por orden de producción; si hay una tendencia grave que causa que el elemento de la prueba de aptitud exceda el criterio en B.2.2 o si no previene el uso del elemento de prueba de aptitud, entonces (i) o bien asignar valores individuales a cada elemento de la prueba de aptitud; o (ii) descartar un subconjunto de los elementos de prueba de aptitud afectados significativamente y vuelva a probar el resto de homogeneidad suficiente; o (iii) si la tendencia afecta a todos los elementos de la prueba de aptitud, siga las disposiciones de B.2.4.

c) Comparar la diferencia entre repeticiones (o rango, si hay más de 2 repeticiones) y, si fuera necesario, prueba una diferencia estadísticamente significativa entre repeticiones, utilizando la prueba de Cochran (ISO 5725-2). Si la diferencia entre las réplicas es grande para cualquier par, revise una explicación técnica para la diferencia y si es apropiado, elimine el grupo periférico del análisis o, si $m > 2$ y la alta varianza es causada por un único valor atípico, quite el punto periférico.

NOTA Si $m > 2$ y se elimina una sola observación, el cálculo posterior de s_w y s_s tendrá que tomar en cuenta el desequilibrio resultante.

B.2.2 Compare la desviación típica entre la muestra s_s con la desviación típica para evaluación de aptitud σ_{pt} . Los ítems de ensayos de aptitud pueden ser considerados adecuadamente homogéneos si:

$$s_s \leq 0,3 \sigma_{pt} \quad (\text{B.1})$$

NOTA 1 La justificación para el factor de 0,3 es que cuando se cumple este criterio la desviación típica entre la muestra aporta menos del 10% de la varianza para la evaluación del desempeño, por lo que es poco probable que sea afectada la evaluación del desempeño.

NOTA 2 De manera equivalente, s_s se puede comparar con δ_E :

$$s_s \leq 0,1\delta_E \quad (\text{B.2})$$

B.2.3 Esto puede ser útil para ampliar el criterio a tener en cuenta para el error de muestreo real y repetibilidad en el control de la homogeneidad. En estos casos, tome las siguientes medidas:

a) Calcular $\sigma_{allow}^2 = (0,3\sigma_{pt})^2$

b) Calcular $c = F_1\sigma_{allow}^2 + F_2 s_w^2$, donde

s_w es la desviación típica dentro de la muestra, calculado en la sección B.3 y

F_1 y F_2 son de tablas estadísticas típicas, reproducido en la Tabla B.1, para el número de ítems de ensayos de aptitud seleccionados y con cada elemento de prueba por duplicado^[33].

Tabla B.1 - Factores F_1 y F_2 para uso en pruebas de homogeneidad suficiente

gm	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1,59	1,60	1,62	1,64	1,67	1,69	1,72	1,75	1,79	1,83	1,88	1,94	2,01	2,10
F_2	0,57	0,59	0,62	0,64	0,68	0,71	0,75	0,80	0,86	0,93	1,01	1,11	1,25	1,43

Cuando $m > 2$, F_2 en B.2.3 b) y en la Tabla B.1 se sustituye con $F_{2m} = (F_{g-1, g(m-1), 0.95-1})/m$ cuando $F_{g-1, g(m-1), 0.95-1}$ es el valor superado con una probabilidad de 0,05 por una variable aleatoria con una distribución F con $g-1$ y $g(m-1)$ grados de libertad.

NOTA Las dos constantes de la Tabla B.1 se derivan de cuadros estadísticos estándar de la siguiente manera:

$F_1 = \chi^2_{0.95(g-1)}$ donde $\chi^2_{0.95(g-1)}$ es el valor superado con probabilidad 0,05 por una variable aleatoria chi-cuadrado con $g-1$ grados de libertad, y

$F_2 = (F_{0.95(g-1);g}-1)/2$ donde $F_{0.95(g-1);g}$ es el valor superado con una probabilidad de 0,05 por una variable aleatoria con una distribución F con $g-1$ y g grados de libertad.

c) Si $s_s > \sqrt{c}$ entonces hay evidencia de que el lote de elementos de la prueba de aptitud no es suficientemente homogéneo.

B.2.4 Cuando σ_{pt} no se conoce de antemano, por ejemplo, cuando σ_{pt} es la desviación típica robusta de resultados de participantes, el proveedor de ensayos de aptitud debe seleccionar otros criterios para determinar la homogeneidad suficiente. Tales procedimientos pueden incluir:

a) comprobar si hay diferencias estadísticamente significativas entre los ítems de la prueba de aptitud, utilizando, por ejemplo, el análisis de varianza F de prueba en $\alpha = 0,05$;

b) utilizar la información de las rondas anteriores del programa de ensayos de aptitud para estimar σ_{pt}

c) utilizar los datos de un experimento de precisión (tal como, una desviación típica de la reproducibilidad como se describe en ISO 5725-2);

d) aceptar el riesgo de la distribución de elementos de la prueba de aptitud que no son suficientemente homogéneos, y comprobar el criterio después del consenso σ_{pt} se ha calculado.

B.2.5 Si no se cumplen los criterios de homogeneidad suficiente, el proveedor de ensayos de aptitud debe considerar la adopción de una de las siguientes acciones.

a) Incluir la desviación típica entre la muestra en la desviación típica para la evaluación de la competencia, mediante el cálculo de σ'_{pt} como en la ecuación (B.3). Tenga en cuenta esto necesita ser descrito completamente a los participantes.

$$\sigma'_{pt} = \sqrt{\sigma_{pt}^2 + s_s^2} \quad (\text{B.3})$$

b) Incluir s_s en la incertidumbre del valor asignado y utilizar z o δ_E' para evaluar el desempeño (ver 9.5);

c) Cuando σ_{pt} es la desviación típica robusta de los resultados participantes, entonces la falta de homogeneidad entre los elementos de prueba de aptitud se incluye en σ_{pt} y así el criterio de aceptabilidad de la homogeneidad puede ser relajada, con precaución.

Si ninguno de a) a c) se aplican, deseche el elemento de prueba de aptitud y repita la preparación después de corregir la causa de la falta de homogeneidad.

B.3 Fórmulas para comprobar la homogeneidad

La estimación de la desviación típica en la muestra s_w y la desviación típica entre muestras s_s se pueden calcular utilizando el análisis de varianza como se muestra a continuación. El método mostrado es para un número elegido g de elementos de pruebas de aptitud, medida en réplicas m veces.

Los datos de un chequeo de homogeneidad están representados por $x_{t,k}$

Donde

t representa el elemento de prueba de aptitud ($t = 1, 2, \dots, g$)
 k representa la porción de ensayo ($k = 1, 2, \dots, m$)

Definir el promedio del elemento de prueba de aptitud y la varianza como:

$$\bar{x}_t = \frac{1}{m} \sum_{k=1}^m x_{t,k}$$

$$s_t^2 = \frac{1}{m} \sum_{k=1}^m (x_{t,k} - \bar{x}_t)^2 \quad (B.4)$$

y la estimación de varianza entre-test-porción como:

$$w_t^2 = \frac{1}{(m-1)} \sum_{k=1}^m (x_{t,k} - \bar{x}_t)^2 \quad (B.5)$$

Calcula el promedio general:

$$\bar{\bar{x}} = \frac{1}{g} \sum_{t=1}^g \bar{x}_t \quad (B.6)$$

la estimación de los promedios de la varianza de la muestra:

$$s_x^2 = \frac{1}{(g-1)} \sum_{t=1}^g (\bar{x}_t - \bar{\bar{x}})^2 \quad (B.7)$$

y la varianza dentro de la muestra:

$$s_w^2 = \frac{1}{g} \sum_{t=1}^g s_t^2 \quad (B.8)$$

Estime la varianza combinada de s_s y s_w

$$s_{s,w}^2 = \frac{1}{(g-1)} \sum_{t=1}^g (\bar{x}_t - \bar{\bar{x}})^2 + \left(1 - \frac{1}{m}\right) s_w^2 = s_s^2 + s_w^2 \quad (B.9)$$

Por último, estimar la variación entre muestras como

$$s_s^2 = s_{s,w}^2 - s_w^2 = \frac{1}{(g-1)} \sum_{t=1}^g (\bar{x}_t - \bar{\bar{x}})^2 - \frac{1}{m} s_w^2 \quad (B.10)$$

NOTA En el caso de que $s_s^2 < 0$, entonces es apropiado utilizar $s_s = 0$.

Para un diseño común cuando m es 2, se pueden utilizar las siguientes fórmulas.

Definir los promedios de la muestra como:

$$\bar{x}_t = (x_{t,1} + x_{t,2}) / 2 \quad (B.11)$$

y el rango entre-test-porción como:

$$w_t = |x_{t,1} - x_{t,2}| \quad (B.12)$$

Calcular el promedio general:

$$\bar{\bar{x}} = \frac{1}{g} \sum_{t=1}^g \bar{x}_t \quad (B.13)$$

Calcular la desviación típica de las medias de la muestra:

$$s_x = \sqrt{\sum_{t=1}^g (\bar{x}_t - \bar{\bar{x}})^2 / (g-1)} \quad (B.14)$$

y la desviación típica dentro de la muestra:

$$s_w = \sqrt{\sum_{t=1}^g w_t^2 / (2g)} \quad (\text{B.15})$$

donde las sumas en las fórmulas B.13, B.14 y B.15 están sobre las muestras ($t = 1, 2, \dots, g$).

Por último, estimar la desviación típica entre la muestra como:

$$s_s = \max \left(0, \sqrt{s_x^2 - (s_w^2 / 2)} \right) \quad (\text{B.16})$$

NOTA 1 La estimación de la varianza entre muestras s_s^2 con frecuencia se vuelve negativo cuando s_s es relativamente menor que s_w . Esto se puede esperar cuando los elementos de prueba de aptitud son altamente homogéneos. En este caso, $s_s = 0$.

NOTA 2 En lugar de utilizar rangos, se podría utilizar las desviaciones típicas entre las porciones de ensayo tales como

$$s_t = w_t / \sqrt{2}$$

NOTA 3 Un ejemplo se proporciona en el anexo E.2

B.4 Procedimientos para el control de la estabilidad

B.4.1 Consideraciones generales para la comprobación de la estabilidad

Estas cláusulas dan orientaciones para el cumplimiento de los requisitos de estabilidad de la sección 6.1. Las disposiciones de la sección 6.1.3 con respecto a las propiedades que han de estudiarse se aplican a cualquier verificación experimental sobre la estabilidad en la duración de la ronda de pruebas de competencia y en la estabilidad durante el transporte.

B.4.1.1 Donde hay una garantía razonable de estudios previos experimentales, experiencia o conocimiento previo de que la inestabilidad es poco probable, controles experimentales de estabilidad se puede limitar a una comprobación por cambio significativo en el transcurso de la ronda de ensayos de aptitud, llevado a cabo durante y después de la misma ronda. En otras circunstancias, los estudios sobre los efectos del transporte y la estabilidad de la duración típica de una ronda de ensayos de aptitud pueden tomar la forma de estudios previstos antes de la distribución de los elementos de la prueba de aptitud, ya sea para cada ronda o durante los primeros estudios de planificación y factibilidad para establecer el transporte consecuente y condiciones de almacenamiento. Proveedores de ensayos de aptitud también pueden comprobar si hay evidencia de inestabilidad comprobando los resultados reportados para una tendencia con la fecha de la medición.

B.4.1.2

Las siguientes consideraciones se aplican a los controles de estabilidad:

- Todas las propiedades que se utilizan en el programa de ensayos de aptitud deben revisarse o verificarse de otro modo para la estabilidad. Esto se puede lograr con la experiencia previa y justificación técnica basada en el conocimiento de la matriz (o artefacto) y mesurando.

- Más de 2 elementos de prueba de aptitud deben probar si la variabilidad entre los ítems de la prueba de aptitud es grande; más muestras o más repeticiones se deben usar si la repetibilidad es sospechosa (por ejemplo, si s_w o $s_s > 0,5\sigma_{pt}$).

NOTA Guía ISO 35 proporciona estrategias para minimizar el efecto en los estudios de estabilidad de la variación a largo plazo en el proceso de medición, como estudios isocrónos o el uso de materiales de referencia estables.

B.4.2 Procedimiento para comprobar la estabilidad durante el curso de una ronda de ensayos de aptitud

B.4.2.1 Un modelo práctico para probar la estabilidad en los ensayos de aptitud es probar una pequeña muestra de los elementos de ensayo de aptitud en la conclusión de una ronda de ensayos de aptitud y compararlos con los elementos de ensayo de aptitud probados antes de la ronda, para asegurar que ningún cambio ocurrió a través del tiempo de la ronda. La verificación puede incluir una comprobación por cualquier efecto de las condiciones de transporte por exponer los elementos del ensayo de aptitud retenidos por la duración del estudio a condiciones que representan las condiciones de transporte. Para los estudios destinados únicamente para buscar efectos de transporte, la comparación es entre los ítems de la prueba de aptitud que se entregan con los elementos de prueba de aptitud que se conservan en condiciones controladas.

NOTA 1 proveedores de ensayos de aptitud pueden utilizar los resultados de la homogeneidad de prueba anteriores de la ronda de ensayos de aptitud en lugar de seleccionar y medir un conjunto separado de los elementos de prueba de aptitud.

NOTA 2 Este modelo se aplica igualmente a los programas de ensayos de aptitud en las pruebas y en la calibración

B.4.2.2 Si un proveedor de ensayos de aptitud incluye el envío de los elementos de la prueba de aptitud en la evaluación de la estabilidad en B.4.2.1, entonces los efectos del transporte se incluyen en la evaluación de la estabilidad. Si los efectos del transporte se comprueban por separado, entonces se debe utilizar el procedimiento descrito en la sección B.6.

B.4.2.3 Un procedimiento para una comprobación básica de la estabilidad utilizando mediciones antes y después de una ronda de ensayos de aptitud es el siguiente:

- Seleccionar un número $2g$ de los elementos de la prueba de aptitud al azar, donde $g \geq 2$.
- Seleccionar un solo laboratorio utilizando un único método de medición con buena precisión intermedia.
- Medir g elementos de ensayos de aptitud antes de la fecha prevista de distribución de los elementos de la prueba de aptitud a los participantes. Deben hacerse repeticiones de las mediciones en un orden totalmente aleatorio.
- Reservar el resto g de elementos de ensayos de aptitud en condiciones similares a las condiciones de almacenamiento previstas en los locales participantes.
- Tan pronto como sea razonablemente posible después de la fecha límite para emitir de resultados de los participantes, evaluar los elementos de prueba de aptitud restantes g , usando el

mismo laboratorio, método de medición y número de repeticiones como en a), con todas las repeticiones en un orden aleatorio.

f) Calcular los promedios \bar{y}_1 y \bar{y}_2 de los resultados para los dos grupos (antes y después), respectivamente.

B.4.2.4 Se puede utilizar las siguientes variaciones en el procedimiento en B.4.2.3:

a) El primer grupo de elementos de prueba de competencia g puede omitirse si otras medidas en el conjunto de los elementos de prueba de competencia están disponibles en el mismo laboratorio y método de ensayo. Por ejemplo, se pueden usar los datos de un control de homogeneidad previo.

b) Las condiciones que puedan acelerar el cambio se pueden utilizar para proporcionar una mayor garantía de estabilidad.

c) El segundo conjunto de elementos de prueba de aptitud, puede adicionalmente estar sujeto a las condiciones previstas en el transporte marítimo, con el fin de incluir una prueba del efecto de envío.

d) Cualquier otro diseño y condiciones que, junto con el criterio de comprobación de estabilidad elegido, proporciona igual o más confianza de estabilidad se pueden utilizar.

B.5 Criterios de evaluación para un control de la estabilidad.

B.5.1 Compara la media general de las medidas obtenidas en el registro previo a la distribución con el promedio general de los resultados obtenidos en la comprobación de la estabilidad. Los ítems de ensayos de aptitud pueden ser considerados adecuadamente estable si:

$$|\bar{y}_1 - \bar{y}_2| \leq 0,3 \sigma_{pt} \text{ o } \leq 0,1 \delta_E \quad (\text{B.17})$$

B.5.2 Si es probable que la precisión intermedia del método de medición (o la incertidumbre de medición del elemento) contribuya a la incapacidad de encontrar el criterio, entonces una de las siguientes opciones se debe tomar:

- a) utilizar un estudio de estabilidad isocrónico (consulte la Guía ISO 35);
- b) aumentar la incertidumbre del valor asignado para tener en cuenta la posible inestabilidad;
- c) ampliar el criterio de aceptación mediante la adición de la incertidumbre de la diferencia a σ_{pt} utilizando la siguiente fórmula:

$$|\bar{y}_1 - \bar{y}_2| \leq 0,3\sigma_{pt} + 2\sqrt{u^2(\bar{y}_1) + u^2(\bar{y}_2)} \quad (\text{B.18})$$

NOTA El factor de 2 en la ecuación (B.18) es un factor de cobertura de la incertidumbre expandida de la diferencia, proporcionando aproximadamente el 95% de confianza, y el cálculo de incertidumbre combinada ha asumido que \bar{y}_1 y \bar{y}_2 son independientes.

B.5.3 Si no se cumple el criterio en las ecuaciones (B.17) o (B.18), las siguientes opciones se deben considerar:

- Cuantificar el efecto de la inestabilidad y tenerlo en cuenta en la evaluación (por ejemplo, con z' score); o
- Examinar los procedimientos de preparación y almacenamiento de los elementos de ensayos de aptitud para ver si las mejoras son posibles; o
- No evaluar el desempeño de los participantes.

B.5.4 El criterio en B.5.1 o B.5.2 puede ser reemplazado por una prueba estadística apropiada para una diferencia entre los dos conjuntos de datos proporcionados que la prueba tiene debidamente en cuenta la réplica y proporciona una garantía de identificación de estabilidad al menos igual a la proporcionada por la ecuación (B.18).

NOTA Un test- t para una diferencia significativa del 95% del nivel de confianza, usando los medios para cada ítem de ensayos de aptitud, por lo general proporcionará una seguridad similar o mejor de detectar la inestabilidad de la ecuación (B.18), siempre que el número de unidades examinadas sean 3 o más.

B.6 Estabilidad en condiciones de transporte

B.6.1 El proveedor de ensayos de aptitud debe comprobar los efectos del transporte sobre los ítems de ensayos de aptitud, por lo menos en las primeras etapas del programa de ensayos de aptitud. Dicha verificación debe, en lo posible, comparar artículos de la prueba de aptitud retenidos en las instalaciones del proveedor de ensayos de aptitud con artículos de la prueba de aptitud sujetos a gastos de envío y devolución. Por ejemplo, también se pueden utilizar estudios basados en la exposición a condiciones razonablemente previsibles de transporte.

B.6.2 Los efectos conocidos de transporte deben ser considerados en la evaluación de desempeño. Cualquier aumento significativo en la incertidumbre debido al transporte debe ser incluido en la incertidumbre del valor asignado.

B.6.3 Si la comprobación de estabilidad de transporte consiste en la comparación de los resultados de dos grupos de artículos de la prueba de aptitud, un grupo está expuesto a las condiciones de transporte y un grupo que no lo es, el criterio para la estabilidad suficiente en el transporte es el mismo que en la sección B. 5.1 o B.5.2.

NOTA 1 Si el valor asignado y la desviación típica para la evaluación de la competencia se determinan a partir de resultados de participantes (por ejemplo, mediante métodos robustos), entonces el promedio y la desviación típica para la evaluación de aptitud reflejará cualquier sesgo y una mayor variabilidad (respectivamente) causado por las condiciones de transporte.

NOTA 2 Un ejemplo de una comprobación de estabilidad se muestra en el anexo E.2

**Anexo C
(Normativo)**
Análisis robusto
C.1 Análisis robusto: Introducción

Comparaciones entre laboratorios presentan desafíos únicos para el análisis de datos. Aunque la mayoría de las comparaciones entre laboratorios proporcionan datos unimodales y aproximadamente simétricas, la mayoría de los conjuntos de datos de ensayos de aptitud incluyen una proporción de resultados que son inesperadamente distante de la mayoría. Estos pueden surgir por una variedad de razones; por ejemplo, de los participantes con menos experiencia, a partir de los métodos de medición menos precisas, o quizás nuevas, o de los participantes que no entendían las instrucciones o que procesan los artículos de la prueba de aptitud de forma incorrecta. Tales resultados periféricos pueden ser muy variables y hacer técnicas estadísticas convencionales, incluyendo la media y la desviación típica, poco fiable.

Se recomienda (ver 6.5.1) que los proveedores de ensayos de aptitud utilizan técnicas estadísticas que son robustas a los valores atípicos. Se han propuesto muchas de tales técnicas en la literatura estadística, y muchos de los se han utilizado con éxito para ensayos de aptitud. La mayoría de técnicas robustas, además, confieren resistencia a las distribuciones asimétricas atípicas.

Este anexo describe varias técnicas que se han aplicado en los ensayos de aptitud y tienen diferentes capacidades con respecto a la solidez de las poblaciones contaminadas (por ejemplo, la eficiencia y el punto de ruptura), y diferentes sencillez de aplicación. Se presentan aquí con el fin de la simplicidad (más simple primero, más compleja pasado), que es aproximadamente inversamente relacionada con la eficiencia porque los estimadores más complejos tienden a ser desarrollados con el fin de mejorar la eficiencia.

NOTA 1 Anexo D proporciona más información sobre la eficiencia, punto de ruptura y la sensibilidad a los modos menores - tres indicadores importantes del desempeño de los diversos estimadores robustos.

NOTA 2 La robustez es una propiedad del algoritmo de estimación, no de las estimaciones que produce, por lo que no es estrictamente correcto llamar a los promedios y las desviaciones típicas calculadas por dicho algoritmo "robusto". Sin embargo, para evitar el uso de la terminología excesivamente engorrosa, los términos "promedio robusto" y "desviación típica robusta" deben entenderse en esta norma en el sentido de las estimaciones de la población significan o de la desviación típica de la población calculada usando un algoritmo robusto.

C.2 Estimadores atípicos resistentes simples para la población media y desviación típica
C.2.1 La mediana

La mediana es un estimador simple y altamente atípico resistente de la población significa para distribuciones simétricas. Para determinar la mediana, deNOTA $med(x)$:

- deNOTAr los elementos p de datos, ordenados en orden creciente, a través de:

$$x_{\{1\}}, x_{\{2\}}, \dots, x_{\{p\}}$$

- Calcular

$$med(x) = \begin{cases} x_{\{(p+1)/2\}} & p \text{ impar} \\ \frac{x_{\{\frac{p}{2}\}} + x_{\{1+p/2\}}}{2} & p \text{ par} \end{cases}$$

(C.1)

C.2.2 Desviación absoluta de la mediana escalada MADe

La desviación absoluta de la mediana escalada $MADe(x)$ proporciona una estimación de la desviación típica de la población de los datos distribuidos normalmente y es altamente resistente a los valores atípicos. Para calcular hecho (x):

- Calcular las diferencias absolutas di (para $i = 1$ a p) del

$$d_i = |x_i - med(x)| \quad (C.2)$$

- Calcular $MADe(x)$ de

$$MADe(x) = 1,483 med(d) \quad (C.3)$$

Si el 50% o más de los resultados de participantes son los mismos, entonces $med(x)$ será cero, y puede ser necesario utilizar el nIQR en la sección C.2.3, una desviación típica aritmética (después de la eliminación de valores atípicos), o el procedimiento de descrito en el apartado C.5.2.

C.2.3 Rango intercuartil Normalizado nIQR

Un estimador robusto de la desviación típica similar al hecho (x) y un poco más simple para obtener ha demostrado ser útil en muchos programas de ensayos de aptitud, y puede ser obtenida a partir de la diferencia entre el 75 por ciento (o tercero cuartil) y percentil 25 (o primero cuartil) de los resultados de participantes. Esta estadística es comúnmente llamado el "normalizado rango intercuartil" (o nIQR), y se calcula que en la fórmula (C.4):

$$nIQR(x) = 0,7413(Q_3(x) - Q_1(x)) \quad (C.4)$$

donde

$Q_1(x)$ deNOTA el 25 percentil de x_i ($i=1,2,\dots,p$)

$Q_3(x)$ deNOTA el 75 percentil de x_i ($i=1,2,\dots,p$)

Si los percentiles 75 y 25 son los mismos, la nIQR será cero (como voluntad $MADe(x)$) y un procedimiento alternativo, tal como una desviación típica de la aritmética (después de la eliminación de valores atípicos) o el procedimiento en C.5.2 se debe utilizar para calcular la desviación típica robusta.

NOTA 1 El nIQR sólo requiere la clasificación de los datos una vez frente a hecho, pero tiene un punto de 25% en carretera (véase el anexo D), mientras que hicimos tiene punto de 50% en carretera. Por lo tanto, hizo puede tolerar una apreciable proporción mayor de los valores extremos que nIQR.

NOTA 2 Tanto nIQR y los estimadores hizo mostrar sesgo negativo apreciable en $p < 30$ que puede afectar negativamente a las puntuaciones si estas estimaciones se utilizan en los resultados de puntuación de los participantes.

NOTA 3 diferentes paquetes estadísticos pueden utilizar diferentes algoritmos para calcular los cuartiles, y por lo tanto puede producir un poco diferente nIQR.

NOTA 4 Un ejemplo usando estimadores robustos simples se incluye en el Anexo E.3.

C.3 Análisis robusto: Algoritmo A

C.3.1 Algoritmo A con escala iterada

Este algoritmo produce estimaciones robustas de la media y la desviación típica de los datos a los que se aplica.

DeNOTAn los elementos p de datos, ordenados en orden creciente, a través de:

$$x_{(1)}, x_{(2)}, \dots, x_{(p)}$$

DeNOTAr la desviación típica promedio y robusta robusta de estos datos por x^* y s^* .

Calcular los valores iniciales para x^* y s^* como:

$$x^* = \text{mediana de } x_i \quad (i = 1, 2, \dots, p) \quad (\text{C.5})$$

$$s^* = 1,483 \text{ mediana de } |x_i - x^*| \text{ with } (i = 1, 2, \dots, p) \quad (\text{C.6})$$

NOTA 1 Algoritmos A y S dada en este anexo se reproducen de la norma ISO 5725-5, con una ligera además de algoritmo A para especificar un criterio de parada: ningún cambio en la 3^a cifras significativas de la media robusta y la desviación típica.

NOTA 2 En algunos casos más de la mitad de los resultados x_i será el idéntico (por ejemplo, número de hilos en la tela, o electrolitos en el suero). En estos casos el valor inicial de s^* será cero y el procedimiento robusto no realizará correctamente. En el caso de que la inicial $s^* = 0$, es aceptable sustituir la desviación típica de la muestra, tras la comprobación de valores atípicos brutos que podrían hacer que la desviación típica de la muestra excesivamente grande. Esta sustitución se realiza sólo para la s inicial *, y después de que el algoritmo iterativo puede proceder como se describe.

Actualización de los valores de x^* y s^* como sigue el resultado:

$$\delta = 1,5s^* \quad (\text{C.7})$$

Para cada x_i ($i = 1, 2, \dots, p$), calcular:

$$x_i^* = \begin{cases} x^* - \delta & \text{cuando } x_i < x^* - \delta \\ x^* + \delta & \text{cuando } x_i > x^* + \delta \\ x_i & \text{de lo contrario} \end{cases}$$

(C.8)

Calcular los nuevos valores de x^* y s^* de:

$$x^* = \sum_{i=1}^p x_i^* / p \quad (\text{C.9})$$

$$s^* = 1,134 \sqrt{\sum_{i=1}^p (x_i^* - x^*)^2 / (p-1)} \quad (\text{C.10})$$

donde se extiende la suma i.

Las estimaciones robustas x^* y s^* pueden ser derivados por un cálculo iterativo, es decir, mediante la actualización de los valores de x^* y s^* varias veces utilizando los datos modificados en las ecuaciones C.7 a C.10, hasta que el proceso converge. La convergencia puede suponer cuando no hay ningún cambio de una iteración a la siguiente en la tercera cifra significativa de la media y la desviación típica robusta (x^* , y s^*). Criterios de convergencia alternativos se pueden determinar de acuerdo con los requisitos de diseño y presentación de informes de resultados de las pruebas de aptitud.

NOTA Ejemplos de uso del Algoritmo A con escala iterada se proporcionan en el Anexo E.3 y E.4.

C.3.2 Variantes del algoritmo A

El algoritmo A con escala iterativa en la sección C.3.1 tiene un modesto desglose (aproximadamente el 25% para grandes conjuntos de datos [25]) y el punto de partida de s^* sugiere en C.3.1 para los conjuntos de datos donde hizo (x) es cero puede degradar seriamente resistencia atípica cuando hay valores extremos graves en el conjunto de datos. Las siguientes variaciones se deben considerar cuando se espera que la proporción de valores atípicos para ser más del 20% en cualquier conjunto de datos, o cuando el valor inicial de s^* se ve afectada negativamente por los valores atípicos extremos:

- i) Reemplazar $MADe$ con $\text{med}(|x_i - \bar{x}|)$ cuando $MADe=0$, utilizar un estimador alternativa como la que se describe en C.5.1 o la desviación típica de la aritmética (después de la eliminación de valores atípicos).
- ii) Cuando no se utiliza la desviación típica robusta en la puntuación, uso que se hace (modificado como i) anterior) y no actualizar s^* durante la iteración. Cuando se utiliza la desviación típica robusta en aNOTACIÓN, reemplace s^* con el estimador Q descrito en C.5 y no actualizar s^* durante la iteración.

NOTA Variante ii) mejora el punto de Algoritmo un desglose a 50% [25], lo que permite el algoritmo para hacer frente a una mayor proporción de valores atípicos.

C.4 análisis robusto: Algoritmo S

Este algoritmo se aplica a las desviaciones típicas (o rangos), que se calculan cuando los participantes que se someten a m réplicas de los resultados de un mensurando en un elemento de prueba de competencia, o en un estudio con elementos de prueba m de competencia idénticas. Se produce un valor robusto agrupado de las desviaciones típicas o rangos a los que se aplica. Se deNOTAn las desviaciones o rangos p típicos, ordenados en orden creciente, a través de:

$$w_{(1)}, w_{(2)}, \dots, w_{(p)} \quad (\text{C.11})$$

DeNOTAr el valor robusto combinado por w^* , y los grados de libertad asociados con cada w_i por v . (Cuando w_i es un rango, $v = 1$. Cuando w_i es la desviación típica de los resultados de las pruebas m, $v = m - 1$). Obtener los valores de ξ y η requerido por el algoritmo de la Tabla C.1. Calcula un valor inicial para w^* como:

$$w^* = \text{mediana de } w_i \quad (i = 1, 2, \dots, p) \quad \text{C.1.}$$

NOTA Si más de la mitad de la desviación w_i es cero, entonces la inicial w^* será cero y el procedimiento robusto no se realizará correctamente. Cuando la inicial w^* es cero, sustituya la aritmética agrupada desviación típica media (o promedio) después de eliminar cualquier valor atípico extremo que pueden influir en la media. Esta sustitución es sólo para la w inicial *, después de lo cual debe seguir el procedimiento que se describe.

Actualizar el valor de w^* como sigue el resultado:

$$\psi = \eta \times w^* \quad (\text{C.12})$$

Para cada w_i ($i = 1, 2, \dots, p$), calcular:

$$w_i^* = \begin{cases} \psi & \text{si } w_i > \psi \\ w_i & \text{de lo contrario} \end{cases} \quad (\text{C.13})$$

Calcular el nuevo valor de w^* a partir de:

$$w^* = \xi \sqrt{\sum_{i=1}^p (w_i^*)^2 / p} \quad (\text{C.14})$$

La estimación robusta w^* se calcula mediante un cálculo iterativo del valor de w^* varias veces, hasta que el proceso converge. La convergencia puede suponerse cuando no hay ningún cambio de una iteración a la siguiente en la tercera cifra significativa de la estimación robusta.

NOTA: El algoritmo S proporciona una estimación de la desviación típica de la población como fuente de desviación típica a partir de una única distribución normal (y por lo tanto proporciona una estimación de la repetibilidad de la desviación típica cuando se aplican los supuestos de la norma ISO 5725 2).

Tabla C.1 - Factores necesarios para el análisis robusto: Algoritmo S

Grados de libertad v	Factor Límite η	Factor de ajuste ξ
1	1,645	1,097
2	1,517	1,054
3	1,444	1,039
4	1,395	1,032
5	1,359	1,027
6	1,332	1,024
7	1,310	1,021
8	1,292	1,019
9	1,277	1,018
10	1,264	1,017

NOTA Los valores de ξ y η se derivan en el anexo B de la norma ISO 5725 5: 1998.

C.5 Estimadores intensivos computacionalmente robustos: La forma de Q y Hampel estimador

C.5.1 Justificación para estimadores computacionalmente intensivos

Los estimadores robustos de la población media y desviación típica descrito en las secciones C.2 y C.3 son útiles cuando los recursos computacionales son limitados, o cuando es necesario proporcionar explicaciones concisas de los procedimientos estadísticos. Estos procedimientos han demostrado ser útiles en una amplia variedad de situaciones, incluso para los programas de ensayos de aptitud en nuevas áreas de ensayo o calibración y en economías en las pruebas de competencia no ha sido previamente disponible. Sin embargo, estas técnicas pueden llegar a ser poco fiable cuando más del 20% de los resultados son valores extremos, o cuando hay distribuciones bimodales (o multimodal), y algunos pueden llegar a ser variable de manera inaceptable para un número menor de participantes. Además, nadie puede manejar datos replicados de los participantes. ISO / IEC 17043 requiere que estas situaciones se anticipen por diseño o sean detectados por la revisión competente antes de la evaluación del desempeño, pero hay ocasiones en las que esto puede no ser posible.

Además, algunas de las técnicas robustas que se describen en las secciones C.2 y C.3 carecen de términos de eficiencia estadística - si el número de participantes es inferior a 50, y la media robusta y / o desviación típica se utilizan para aNOTAr allí es un riesgo considerable para la clasificación errónea de los participantes debido a la utilización de métodos estadísticos ineficaces.

Técnicas robustas que combinan una buena eficiencia (es decir, comparativamente baja variabilidad) con la tolerancia de una alta proporción de los valores extremos tienden a ser más

complejos y requieren más recursos computacionales, pero las técnicas son referenciadas en la literatura disponible y las Normas Internacionales. Algunas de ellas ofrecen, además, útiles mejoras de rendimiento cuando la distribución subyacente de los datos es sesgada o cuando algunos resultados son citados como debajo de una detección o informar límite.

A continuación se describen algunos de alta eficiencia, los métodos de alta degradación para la estimación de la desviación típica y la ubicación (media) que son útiles para los datos con una mayor proporción de valores atípicos y que muestran la variabilidad menor que los estimadores simples. Uno de los estimadores descritos también se puede utilizar para estimar una desviación típica de la reproducibilidad cuando los participantes informan múltiples observaciones.

C.5.2 Determinación de una desviación típica robusta utilizando métodos Q y Qn

C.5.2.1 Qn [34] es una gran ruptura, el estimador de alta eficiencia de la desviación típica de la población que es imparcial para datos distribuidos normalmente (es decir, bajo el supuesto de que no hay valores atípicos). Qn utiliza un único resultado comunicado (incluyendo una media o mediana de repeticiones) para cada participante. El cálculo se basa en el uso de diferencias por pares dentro del conjunto de datos y por lo tanto no depende de una estimación de la media o la mediana de los datos. La implementación descrita aquí incluye correcciones para asegurar que la estimación es imparcial para todos los tamaños de conjuntos de datos prácticos.

Para calcular Qn para un conjunto de datos (x_1, x_2, \dots, x_p) con p reportados resultados:

i) Calcular el $p(p-1)/2$ diferencias absolutas

$$d_{ij} = |x_i - x_j| \text{ para } i = 1, 2, \dots, p-1 \text{ y } j = i+1, i+2, \dots, p \quad (\text{C.15})$$

ii) DeNOTAn las diferencias ordenadas por D_{ij}

$$d_{\{1\}}, d_{\{2\}} \dots d_{\{p(p-1)/2\}} \quad (\text{C.16})$$

iii) Calcular

$$k = \frac{h(h-1)}{2} \quad (\text{C.17})$$

es decir, k es el número de pares distintos elegidos entre objetos h, donde:

$$h = \begin{cases} p/2 & p \text{ par} \\ (p-1)/2 & p \text{ impar} \end{cases} \quad (\text{C.18})$$

iv) Calcular Qn como:

$$Q_n = 2,2219d_{(k)}b_p \quad (\text{C.19})$$

donde se selecciona b_p de la Tabla C.2 para un número p de puntos de datos particular, o, por $p > 12$, se calcula a partir:

$$b_p = \frac{1}{r_p + 1} \quad (\text{C.20})$$

Donde:

$$r_p = \begin{cases} \frac{1}{p} \left[1,6019 + \frac{1}{p} \left(-2,128 - \frac{5,172}{p} \right) \right] & p \text{ impar} \\ \frac{1}{p} \left[3,6756 + \frac{1}{p} \left(1,965 + \frac{1}{p} \left(6,987 - \frac{77}{p} \right) \right) \right] & p \text{ par} \end{cases} \quad (\text{C.21})$$

NOTA 1 El factor de 2,2219 es un factor de corrección para dar una estimación no sesgada de la desviación típica para la gran p. Los factores de corrección bp para pequeña p están en la tabla C.2 y el cálculo para rp para $p > 12$ están en lo dispuesto en la referencia [34] de la extensa simulación y análisis de regresión posterior.

NOTA 2 El algoritmo simple descrito anteriormente, requiere considerables recursos de computación para los conjuntos de datos más grandes, por ejemplo $p > 1000$. Una implementación rápida y la memoria eficiente capaz de manejar conjuntos de datos mucho más grandes ha sido publicado con código informático completo [34] para su uso con mayor conjuntos de datos; referencia [34] citó un rendimiento aceptable para p durante 8000 en el momento de su publicación.

Tabla C.2 - Factor de Corrección b_p para $2 \leq p \leq 12$

p	2	3	4	5	6	7	8	9	10	11	12
b_p	0,9937	0,9937	0,5132	0,8440	0,6122	0,8588	0,6699	0,8734	0,7201	0,8891	0,7574

C.5.2.2 El Q produce una alta degradación, estimación de alta eficiencia de la desviación típica de los resultados de ensayos de aptitud comunicados por diferentes laboratorios. El método Q no sólo es robusto frente a los resultados de la periferia, sino también contra una situación en la que muchos resultados de la prueba son iguales, por ejemplo, debido a los datos cuantitativos en una escala discontinua o debido a las distorsiones de redondeo. En tal situación, otros métodos-Q como pueden fallar debido a que muchas diferencias por pares son cero.

El método Q se puede utilizar para ensayos de aptitud, ambos con resultados individuales por participante (incluyendo una media o mediana de repeticiones) y para repeticiones. El uso directo de repeticiones en el cálculo mejora la eficiencia del método.

El cálculo se basa en el uso de diferencias por pares dentro del conjunto de datos y por lo tanto no depende de una estimación de la media o la mediana de los datos. El método se conoce como Q / Hampel cuando se utiliza junto con el algoritmo de paso finita para el estimador Hampel describe en C.5.3.3.

Denotemos los resultados de las mediciones reportadas, agrupados por laboratorio, por:

$$\underbrace{y_{11}, \dots, y_{1n_1}}_{Lab\ 1}, \underbrace{y_{21}, \dots, y_{2n_2}}_{Lab\ 2}, \dots, \underbrace{y_{p1}, \dots, y_{pn_p}}_{Lab\ p}$$

Calcular la función de distribución acumulativa de todas las diferencias entre absolutos-laboratorio

$$H_1(x) = \frac{2}{p(p-1)} \sum_{1 \leq i < j \leq p} \frac{1}{n_i n_j} \sum_{k=1}^{n_i} \sum_{m=1}^{n_j} I\{|y_{ik} - y_{jm}| \leq x\} \quad (C.22)$$

donde

$$I\{|y_{ik} - y_{jm}| \leq x\} = \begin{cases} 1 & \text{si } |y_{ik} - y_{jm}| \leq x \\ 0 & \text{de lo contrario} \end{cases}$$

de NOTA la función de indicador.

Denotemos los puntos de discontinuidad de $H_1(x)$ por:

x_1, \dots, x_r , where $x_1 < x_2 < \dots < x_r$.

Calcular para todos los puntos de discontinuidad x_1 positivo, ..., x_r :

$$G_1(x_i) = \begin{cases} 0,5 \cdot (H_1(x_i) + H_1(x_{i-1})) & \text{si } i \geq 2 \\ 0,5 \cdot H_1(x_1) & \text{si } i = 1; x_1 > 0 \end{cases} \quad (C.23)$$

y deja

$$G_1(0)=0$$

Calcular la función $G_1(x)$ para todo x fuera del intervalo $[0, x_r]$ por interpolación lineal entre los puntos de discontinuidad $0 \leq x_1 < x_2 < \dots < x_r$.

Calcular la desviación típica robusta s^* de resultados de pruebas de laboratorios diferentes

$$s^* = \frac{G_1^{-1}(0,25 + 0,75 \times H_1(0))}{\sqrt{2} \Phi^{-1}(0,625 + 0,375 \times H_1(0))} \quad (C.24)$$

Donde $H_1(0)$ se calcula como en la ecuación (C.22) y es igual a cero a menos que haya lazos exacta en el conjunto de datos, y donde $\Phi^{-1}(q)$ es el cuantil q -ésima de la distribución normal típica.

NOTA 1 Este algoritmo no depende de un valor medio; que se puede utilizar junto con un valor de resultados de participantes combinadas o un valor de referencia especificado.

NOTA 2 Otras variantes del método Q proporcionan estimaciones robustas de ambas desviaciones típicas de la repetibilidad y la reproducibilidad [25,34].

NOTA 3 La base teórica para el método Q, incluyendo el rendimiento asintótico y desglose muestra finita, se describen en las referencias [26] y [34].

NOTA 4 Si los datos subyacentes de los participantes representan resultados de las mediciones individuales obtenidos con un método de medición específica, la desviación típica robusta es una estimación de la desviación típica de la reproducibilidad como en la ecuación (C.21).

NOTA 5 La desviación típica de la reproducibilidad no es necesariamente la desviación típica más adecuada para su uso en pruebas de competencia, ya que es por lo general una estimación de la dispersión de los resultados individuales y no una estimación de la dispersión de medias o medianas de resultados replicados de cada participante. Sin embargo, la dispersión de medias o medianas de resultados replicados es sólo ligeramente por debajo de la dispersión de los resultados individuales de diferentes laboratorios, si la relación de la desviación típica de la reproducibilidad dividida por la desviación típica de la repetibilidad es mayor que 2. Si esta relación es inferior a 2, para la puntuación en pruebas de competencia puede considerarse para reemplazar la desviación típica de la reproducibilidad s_R por el valor

corregido $\sqrt{s_R^2 - \frac{m-1}{m} s_r^2}$, donde m de NOTA número de repeticiones y S_r^2 la varianza repetibilidad tal como se calcula en [35], o no usar las repeticiones, pero la media de repeticiones por participante para el método Q.

NOTA 6: La NOTA 5 se aplica sólo si el marcador se lleva a cabo sobre la base de medias o medianas de resultados replicados. Si las réplicas son ítems de la prueba de aptitud réplica ciegos, las puntuaciones se deben dar por cada réplica. En este caso la desviación típica de la reproducibilidad es la desviación típica más apropiada.

NOTA 7 Un ejemplo al que se ha aplicado el método Q se muestra en el Anexo E.3.

C.5.3 Determinación de una media robusta utilizando el estimador Hampel

C.5.3.1 La estimación Hampel es una estimación muy robusta y eficiente de la media global de los resultados reportados por diferentes laboratorios. Como no existe una fórmula explícita para la obtención de la estimación Hampel, en este párrafo dos algoritmos se proporcionan. El primero puede ser implementado con mayor facilidad, pero puede conducir a desviarse resultados en diferentes implementaciones. El segundo ofrece resultados únicos que sólo depende de la desviación típica subyacente.

C.5.3.2 El siguiente cálculo proporciona un esquema de **reponderación** iterativa para la obtención de la estimación Hampel de ubicación.

i) de NOTA los datos mientras que $x_1, x_2 \dots x_p$

ii) Establecer x^* a med (x) (sección C.2.1)

iii) Establecer s^* a una estimación sólida adecuada de desviación típica, hecha por ejemplo, Qn o s^* a partir del método Q.

iv) Para cada x_i punto de datos, calcule el q_i de

$$q_i = \left| \frac{x_i - x^*}{s^*} \right|$$

v) Calcular los pesos w_i desde

$$w_i = \begin{cases} 0 & |q| > 4,5 \\ (4,5 - q)/q & 3 < |q| \leq 4,5 \\ 1,5/q & 1,5 < |q| \leq 3,0 \\ 1 & |q| \leq 1,5 \end{cases}$$

vi) Recalcular x^* desde

$$x^* = \frac{\sum_{i=1}^p w_i x_i}{\sum_{i=1}^p w_i}$$

vii) Repita los pasos iv) a vi) hasta x^* converge. La convergencia puede ser asumida cuando el cambio en x^* de una iteración a la siguiente es inferior a $(0,01s^*/p^{1/2})$, que corresponde a aproximadamente el 1% del error típico en x^* . Otros criterios de convergencia, y más precisos pueden ser utilizados.

Esta implementación del estimador Hampel no está garantizado para tener una solución única o para dar como resultado la mejor solución porque una mala elección de la ubicación inicial x^* y / o s^* puede excluir una parte importante del conjunto de datos. El proveedor de ensayos de aptitud en consecuencia debe implementar medidas para verificar si hay la posibilidad de una solución pobre o proporcionar reglas claras para la elección de la ubicación. La regla más común es elegir la solución más cercana a la mediana. Revisión de los resultados para asegurar que ninguna gran proporción del conjunto de datos está fuera del intervalo $|q| > 4,5$ también puede ayudar a confirmar una solución viable.

NOTA 1 Esta implementación del estimador de Hampel tiene aproximadamente el 96% de eficiencia para los datos distribuidos normalmente.

NOTA 2 Un ejemplo utilizando esta implementación se da en el Anexo E.3

NOTA 3 Estimador de Hampel puede ser sintonizado para una mayor eficiencia o una mayor resistencia a los valores atípicos cambiando la función de peso. La forma general de la función de ponderación es

$$w_i = \begin{cases} 0 & |q| > c \\ a(c - q)/[q(c - b)] & b < |q| \leq c \\ a/q & a < |q| \leq b \\ 1 & |q| \leq a \end{cases}$$

Donde a, b y c son parámetros de ajuste. Para la puesta en práctica aquí, $a = 1,5$, $b = 3,0$ y $c = 4,5$. Una mayor eficiencia se obtiene mediante el aumento de la gama; mejor resistencia a los valores atípicos o modos menores se obtiene mediante la reducción de la gama.

C.5.3.3 El siguiente algoritmo paso finito produce la estimación Hampel de ubicación sin reponderación iterativa [25].

Calcula las medias aritméticas de cada laboratorio, y_1 ahora etiquetados, y_2, \dots, y_p .

Calcular la media robusta, x^* , mediante la resolución de la ecuación

$$\sum_{i=1}^p \Psi \left(\frac{y_i - x^*}{s^*} \right) = 0 \quad (\text{C.25})$$

Donde

$$\Psi(q) = \begin{cases} 0 & q \leq -4,5 \\ -4,5 - q & -4,5 < q \leq -3 \\ -1,5 & -3 < q \leq -1,5 \\ q & -1,5 < q \leq 1,5 \\ 1,5 & 1,5 < q \leq 3 \\ 4,5 - q & 3 < q \leq 4,5 \\ 0 & q > 4,5 \end{cases} \quad (\text{C.26})$$

y s^* es la desviación típica robusta según el método Q.

La solución exacta se puede obtener en un número finito de pasos, lo que no significa de manera iterativa, utilizando la propiedad de que Ψ en el argumento de x^* es parcialmente lineal, teniendo en cuenta que los nodos de interpolación en el lado izquierdo de la ecuación (C.25) (interpretado aquí como una función de x^*) son los siguientes:

Calcula todos los nodos de interpolación

- Para el primer valor y_1 :

$$d_1 = y_1 - 4,5 \cdot s^*, d_2 = y_1 - 3 \cdot s^*, d_3 = y_1 - 1,5 \cdot s^*, d_4 = y_1 + 1,5 \cdot s^*, d_5 = y_1 + 3 \cdot s^*, d_6 = y_1 + 4,5 \cdot s^*$$

- Para el segundo valor y_2 :

$$d_7 = y_2 - 4,5 \cdot s^*, d_8 = y_2 - 3 \cdot s^*, d_9 = y_2 - 1,5 \cdot s^*, d_{10} = y_2 + 1,5 \cdot s^*, d_{11} = y_2 + 3 \cdot s^*, d_{12} = y_2 + 4,5 \cdot s^*$$

- Y así sucesivamente para todos los valores y_3, \dots, y_p .

Ordena estos datos $d_1, d_2, d_3, \dots, d_{6,p}$ en orden ascendente,

$$d_{\{1\}}, d_{\{2\}}, d_{\{3\}}, \dots, d_{\{6 \cdot p\}}$$

Luego calcule para cada uno

$$m = 1, \dots, (6 \cdot p - 1)$$

$$p_m = \sum_{i=1}^p \Psi \left(\frac{y_i - d_{\{m\}}}{s^*} \right)$$

y comprobar si

- (i) $p_m = 0$. Si es así, $d_{\{m\}}$ es una solución de la ecuación (C.25).
- (ii) $p_{m+1} = 0$. Si es así, $d_{\{m+1\}}$ es una solución de la ecuación (C.25).
- (iii) $p_m \times p_{m+1} < 0$. Si es así, $x_m = d_{\{m\}} - \frac{p_m}{p_{m+1} - p_m}$ es una solución de la ecuación (C.25).

Se de NOTA S al conjunto de todas estas soluciones de la ecuación (C.25).

La solución más cercana a la mediana $x^* \in S$ se utiliza como parámetro de localización x^* , es decir,

$$\left| x^* - \text{med}(y_1, y_2, \dots, y_p) \right| = \min \left\{ |x - \text{med}(y_1, y_2, \dots, y_p)| ; x \in S \right\}$$

Pueden existir varias soluciones. Si hay dos soluciones más cercanas a la mediana, o si no existe una solución en absoluto, el propio medio se utiliza como parámetro de ubicación x^* .

NOTA 1 Esta implementación del estimador de Hampel tiene aproximadamente el 96% de eficiencia para los datos distribuidos normalmente.

NOTA 2 Si se utiliza este método de estimación, los resultados de laboratorio que difieren de la media en más de 4,5 veces la desviación típica de la reproducibilidad ya no tiene ningún efecto en el resultado del cálculo, es decir, que se tratan como valores atípicos.

C.5.4 El método Q / Hampel

El método conocido como Q / Hampel utiliza el método Q que se describe en C.5.3.2 para el cálculo de la desviación típica robusta s^* junto con el algoritmo de paso finito para el estimador Hampel descrito en C.5.3.3 para el cálculo de la ubicación del parámetro x^* .

Cuando los participantes informan múltiples observaciones, el método Q se describe en C.5.3.2 se utiliza para el cálculo de la desviación típica robusta reproducibilidad s_R . Para el cálculo de la desviación típica robusta repetibilidad sr un segundo algoritmo utilizando las diferencias por pares dentro de los laboratorios se aplica.

NOTA Una aplicación web para el método Q / Hampel está disponible [37].

C.6 Otras técnicas robustas

Los métodos descritos en el presente anexo no constituyen una colección completa de enfoques válidos, y ninguno se garantiza que sea óptima para todas las situaciones. Otros estimadores robustos pueden utilizarse a discreción del proveedor de ensayos de aptitud, sujeto a la demostración, por referencia a la eficacia conocida, punto de ruptura y cualquier otra propiedad adecuada, que cumplan con los requisitos particulares del programa de ensayos de aptitud.

**Anexo D
(Informativo)**

Orientaciones adicionales sobre Procedimientos estadísticos

D.1 Procedimientos para un pequeño número de participantes

D.1.1 Consideraciones generales

Muchos programas de ensayos de aptitud tienen pocos participantes, o tienen grupos de comparación con un número reducido de participantes, incluso si hay un gran número de participantes en el sistema. Esto puede ocurrir con frecuencia cuando los participantes se agrupan y se obtuvieron por el método, como se hace habitualmente en los ensayos de aptitud para laboratorios médicos, por ejemplo.

Cuando el número de participantes es pequeño, el valor asignado idealmente debe ser determinado usando un procedimiento metrológicamente válida, independiente de los participantes, tales como mediante la formulación o de un laboratorio de referencia. Criterios de evaluación de resultados también deben basarse en criterios externos, como la opinión de expertos o criterios basados en la aptitud para el propósito. En estas situaciones ideales, el rendimiento se evalúa mediante el valor asignado previamente determinado y criterio de rendimiento, por lo que las pruebas de competencia se pueden realizar con un solo participante. Este tipo de comparación entre laboratorios se puede llamar una comparación bilateral, o una auditoría de medición, y puede ser muy útil en muchas situaciones, por ejemplo, en la calibración. Cuando no se pueden cumplir estas condiciones ideales, ya sea el valor asignado o la dispersión, o ambos, pueden necesitar ser derivada de resultados de participantes. Si el número de participantes es demasiado pequeño para los procedimientos particulares usados la evaluación del desempeño puede llegar a ser poco fiable; por tanto, es importante considerar si un número mínimo de participantes se debe establecer para la evaluación del desempeño. Los siguientes párrafos presentan orientación para situaciones de números pequeños, cuando los criterios de evaluación del desempeño se determinan utilizando resultados de participantes.

D.1.2 Procedimientos para la identificación de los valores atípicos

Aunque las estadísticas robustas son muy recomendables para las poblaciones atípicas, no se recomienda a menudo para los conjuntos de datos muy pequeños (ver más abajo para excepciones). Las pruebas de valores atípicos, sin embargo, es posible que los conjuntos de datos muy pequeñas. Rechazo de valores atípicos seguido por, por ejemplo, el cálculo de la media o desviación típica, por tanto, puede ser preferible en el caso de regímenes o grupos muy pequeños. Diferentes pruebas de valores atípicos son aplicables a diferentes tamaños de conjuntos de datos. ISO 5725 proporciona 2 tablas para la prueba de Grubbs para un único valor atípico y para dos valores extremos simultáneos en la misma dirección. Grubbs y otras pruebas requieren el número de posibles valores atípicos a especificar por adelantado y puede fallar cuando hay múltiples valores atípicos, haciéndolos más útiles para $p > 10$ (dependiendo de la proporción de los valores extremos).

NOTA 1 Se debe tener cuidado al estimar la dispersión después del rechazo atípico como estimaciones de dispersión estarán sesgadas baja. El sesgo no suele ser grave si el rechazo se lleva a cabo sólo en el nivel del 99% de confianza o superior.

NOTA 2 estimadores robustos mayoría univariantes para localización y dispersión realizar aceptablemente para $p \geq 12$.

D.1.3 Procedimientos para estimaciones de ubicación

D.1.3.1 Los valores asignados derivados de pequeños conjuntos de datos de los participantes deben, en lo posible, cumplir con el criterio de incertidumbre del valor asignado dada en 9.2.1. Para una situación utilizando una media simple como el valor asignado y una desviación típica de los resultados como la desviación típica para la evaluación de la competencia, este criterio no se puede satisfacer de una distribución normal con $p \leq 12$, después de cualquier eliminación de valores atípicos. Para el uso de la mediana como el valor asignado (tomando la eficiencia, 0,64), el criterio no puede ser cubierto por $p \leq 18$. Otros estimadores robustos, tales como el algoritmo A (C.3), tienen la eficiencia intermedia y pueden cumplir con el criterio de $p > 12$ si se tienen en cuenta las disposiciones de 7.7.3 NOTA 2.

D.1.3.2 Hay establecidos limitaciones de tamaño sobre la aplicabilidad de algunos estimadores de ubicación de los datos. Se recomiendan unos estimadores robustos computacionalmente intensivos para la media para los pequeños conjuntos de datos; un límite inferior típico es $p \geq 15$, aunque los proveedores pueden ser capaces de demostrar un rendimiento aceptable para los supuestos específicos en los conjuntos de datos más pequeños. La mediana es aplicable a $p = 2$ (cuando es igual a la media), pero al $3 \leq p \leq 5$ la mediana ofrece pocas ventajas con respecto a la media menos que haya una inusualmente alto riesgo de malos resultados.

D.1.4 Procedimientos para las estimaciones de la dispersión

D.1.4.1 El uso de criterios de rendimiento basados en la dispersión de los resultados de participantes no se recomienda para pequeños conjuntos de datos debido a la muy alta variabilidad de las estimaciones de dispersión. Por ejemplo, para $p = 30$, se espera que las estimaciones de la desviación típica para los datos distribuidos normalmente para variar en aproximadamente un 25% a cada lado de su verdadero valor (basado en un nivel de confianza del 95%). Ningún otro estimador de mejora en esta para los datos distribuidos normalmente.

D.1.4.2 Cuando se requieren estimadores de dispersión para otros fines (por ejemplo, como las estadísticas de resumen o una estimación de la dispersión de sólidos estimadores de localización), o cuando el programa de ensayos de aptitud puede tolerar una alta variabilidad en las estimaciones de dispersión, las estimaciones de dispersión con la más alta disponible la eficiencia se debe seleccionar al manipular conjuntos de datos más pequeños.

NOTA 1 "más alta disponible" se entiende para tener en cuenta la disponibilidad del software y la experiencia adecuada.

NOTA 2 La estimación Q_n de desviación típica descrito en el apartado C.5 es considerablemente más eficiente que la $MADe$ o el nQR del anexo C.1.

NOTA 3 Las recomendaciones específicas se han hecho las estimaciones robustas de dispersión en pequeños conjuntos de datos [24] de la siguiente manera:

— $p=2$: usar $|x_1-x_2|/\sqrt{2}$;

— $p=3$ ubicaciones y escala desconocida: utilizar $MADe$ para proteger contra las estimaciones excesivamente altas de la desviación típica o la desviación absoluta media para proteger contra indebidamente pequeñas estimaciones de la desviación típica, por ejemplo, cuando el redondeo pueden dar dos valores idénticos.

— $p \geq 4$: Un M-estimación concreta de la desviación típica de una función de ponderación logarítmica fue recomendado por la referencia [27]; un contravalor es el algoritmo A sin iteración del lugar, utilizando la mediana como una estimación de localización.

NOTA 4 Para obtener una estimación de la desviación típica de la distancia absoluta a la mediana use:

$$s^* = \frac{1}{0,798 \times p} \sum_{i=1}^p |x_i - \text{med}(x)| \quad (\text{D.1})$$

D.2 Eficiencia y puntos de ruptura para procedimientos robustos

D.2.1 Diferentes estimadores estadísticos (por ejemplo, técnicas sólidas) pueden ser comparados en tres características principales:

Punto de ruptura - la proporción de valores en el conjunto de datos que se puede sustituir por valores arbitrariamente grandes sin la estimación también ser arbitrariamente grande.

Eficiencia - la relación de la varianza del estimador dividido por la varianza de un estimador de varianza mínimo para la distribución en cuestión.

Resistencia a modos menores - la capacidad de un estimador para resistir el sesgo causado por un grupo minoritario de resultados discrepantes (típicamente menos de 20% del conjunto de datos).

Estas características dependen en gran medida de la distribución subyacente de resultados para una población de participantes competentes, y la naturaleza de los resultados que son de los participantes incompetentes (o de los participantes que no siguieron las instrucciones o el método de medición). Los datos contaminantes pueden aparecer como valores atípicos, los resultados con mayor varianza, o resultados con una media diferente (por ejemplo, bimodal).

Los puntos de ruptura y las eficiencias para los diferentes estimadores serán diferentes para diferentes situaciones, y una revisión a fondo está más allá del alcance de este documento. Sin embargo las comparaciones simples se pueden hacer bajo el supuesto de una distribución normal para los resultados de los laboratorios competentes, con una media igual a x_{pt} y desviación típica igual a σ_{pt} .

D.2.2 Punto de ruptura

El punto de ruptura es la proporción de los valores del conjunto de datos que pueden ser valores atípicos y sin la estimación se produzcan efectos adversos. El punto de ruptura es una medida de la resistencia a los valores atípicos; alto punto de ruptura se asocia con resistencia a una alta proporción de los valores extremos. Puntos Desglose y resistencia a los modos de menor importancia para los estimadores en el Anexo C se presentan en la Tabla D.1. Cabe señalar que los procedimientos requeridos en las secciones 6.3 y 6.4 deben evitar el análisis de datos de los conjuntos de datos con una gran proporción de los valores extremos. Sin embargo, hay situaciones en opinión visual no es práctico.

Tabla D.1 - puntos Desglose de las estimaciones de la media y la desviación típica (proporción de valores atípicos que pueden llevar al fracaso del estimador)

Estimador estadístico	Parámetro de la población estimada	Puntos desglose	Resistencia a los modos menores
media de la muestra	media	0 %	Pobre
Desviación típica de la muestra	Desviación típica	0 %	Pobre
la mediana de la muestra	media	50 %	Bueno
<i>n</i> / <i>Q</i> R	Desviación típica	25 %	Moderado
<i>MADe</i>	Desviación típica	50 %	Moderado - Bueno
Algoritmo A	Media y desviación típica	25 %	Moderado
<i>Q</i> _n y <i>Q</i> /Hampel	Media y desviación típica	50 %	Moderado (Muy Bueno para modos menores más distantes de 6 * s)

NOTA La definición de punto de ruptura que se utiliza aquí es la proporción de un gran conjunto de datos distribuidos normalmente que se puede mover a + infinito sin la estimación también se mueve hasta el infinito. Por ejemplo, si un poco menos de 50% de un conjunto de datos se sustituye por + infinito, la mediana se mantendrá dentro de los datos finitos restantes.

En resumen, la media muestral y la desviación típica se puede romper con un único valor atípico. Los métodos robustos utilizando la mediana, *MADe*, y *Q*/Hampel pueden tolerar una proporción muy grande de valores atípicos. Algoritmo A con desviación típica iterada y *n*/*Q*R tiene un punto de ruptura del 25%. En cualquier situación, con una gran proporción de los valores atípicos (>20%), cualquier procedimiento convencional o robusto puede producir estimados no razonables de localización y dispersión, y se debe tener precaución en la interpretación de estos valores.

D.2.3 Eficiencia relativa

Todas las estimaciones tienen varianza de muestreo - es decir, las estimaciones pueden variar en cada ronda a ronda de un programa de ensayos de aptitud, incluso si todos los participantes son competentes y no hay valores atípicos o subgrupos de participantes con diferentes medias o desviaciones. Modificando los estimadores robustos se presentaron resultados que están excepcionalmente lejos de la media de la distribución presentada, basado en suposiciones teóricas, y así estos estimadores tienen una varianza más grande que los estimadores de varianza mínima, en el caso de que el conjunto de datos sea, de hecho, una distribución normal.

La media muestral y la desviación típica son los estimadores de la varianza mínimas de la población media y desviación típica, y por lo que tienen la eficiencia del 100%. Estimadores con menor eficiencia tienen una mayor varianza - es decir, que podrían variar más de ronda a ronda, incluso si no hay valores atípicos o diferentes subgrupos de participantes. Tabla D.2 ofrece eficiencias relativas de los estimadores presentados en el Anexo C.

Tabla D.2 - Eficiencia relativa de los estimadores robustos para la media y la desviación típica de la población, para conjuntos de datos con distribución normal con n = 50 o 500 participantes:

Estimador estadístico	Media, n=50	Media, n=500	SD, n=50	SD, n=500
media de la muestra y desviación típica	100 %	100 %	100 %	100 %
Mediana y nIQR	66 %	65 %	38 %	37 %
Mediana y MADe	66 %	65 %	37 %	37 %
Algoritmo A	97 %	97 %	74 %	73 %
Q _n y Q / Hampel	96 %	96 %	73 %	81 %

Estos resultados demuestran que no existe un método estadístico que es perfecto para todas las situaciones. La media muestral y la desviación típica son óptimas con una distribución normal, pero se descomponen en el caso de los valores atípicos. Métodos robustos simples, como media, MADe o nIQR realizan comparativamente poco para datos distribuidos normalmente pero pueden ser eficaces cuando los valores atípicos están presentes o el conjunto de datos es pequeña.

D.3 El uso de los datos de ensayos de aptitud para evaluar la reproducibilidad y repetibilidad de un método de medición

D.3.1 La Introducción a la norma ISO / IEC 17043 establece que la evaluación de las características de funcionamiento de un método generalmente no es un objetivo de los ensayos de aptitud. Sin embargo, es posible utilizar los resultados de los programas de ensayos de aptitud para verificar, y tal vez establecer la repetibilidad y reproducibilidad de un método de medición [15], cuando el programa de ensayos de aptitud cumple las siguientes condiciones:

- a) el nivel de competencia probando artículos son suficientemente homogéneas y estables;
- b) los participantes son capaces de un rendimiento satisfactorio compatible,
- c) la competencia de los participantes (o un subconjunto de los participantes) se ha demostrado antes de la ronda de ensayos de aptitud, y su competencia no está puesta en duda por los resultados de la ronda.

D.3.2 A fin de proporcionar datos suficientes para la evaluación de la repetibilidad y reproducibilidad de un método de prueba de un programa de ensayos de aptitud, se utilizará las siguientes condiciones de diseño:

- a) un número suficiente de participantes para satisfacer un estudio en colaboración han demostrado competencia con un método de medición en las rondas anteriores de un programa de

ensayos de aptitud, y se han comprometido a seguir el método de medición sin modificaciones;

- b) en caso de repetición se ha de evaluar, cada ronda de pruebas de competencia utilizado para la evaluación de la repetibilidad debe incluir al menos dos artículos de la prueba de aptitud o un requisito para observaciones repetidas;
- c) cuando sea posible, los participantes deben contar con réplicas ciegos identificados por separado en lugar de ser solicitado para realizar replicar mediciones sobre el mismo tema del ensayo de aptitud;
- d) elementos de prueba de aptitud utilizados en una o varias rondas del programa de ensayos de aptitud para abarcar la gama de niveles y tipos de muestras de rutina para la que está destinado el método de medición;
- e) procedimientos de análisis de datos que se utilizan para evaluar la repetibilidad y reproducibilidad deben ser coherentes con la norma ISO 5725 o el protocolo de estudio colaborativo en uso.

**Anexo E
(Informativo)
Ejemplos ilustrativos**

Estos ejemplos están destinados a ilustrar los procedimientos especificados en la presente Norma, por lo que el lector puede determinar que sus cálculos son correctos. Ejemplos específicos no deben ser considerados como recomendaciones para su uso en programas de ensayos de aptitud particulares.

E.1 Efecto de valores censurados (sección 5.5.3.3)

Tabla E.1 muestra 23 resultados para una ronda de un programa de ensayos de aptitud, de los cuales 5 resultados se indican como "menos que" una cierta cantidad. La media robusta (\bar{x}^*) y la desviación típica (s^*) Algoritmo A se muestran durante 3 cálculos diferentes, donde los '<' signos se desechan y los datos analizados como datos cuantitativos; los resultados con '<' valores se ignoran; y donde 0,5 veces el resultado se inserta como una estimación del resultado cuantitativo. En cada escenario los resultados que habrían sido fuera del límite de aceptación se indican con '#'. Esto supone que la evaluación sería (señal de acción) "inaceptable" para cualquier resultado en la parte cuantitativa se encuentra fuera del $\bar{x}^* \pm 3s^*$. El proveedor de ensayos de aptitud podría tener reglas alternativas para la evaluación de resultados con los signos "<" o ">".

Tabla E.1 - Muestra de datos con resultados truncados (<), y tres opciones para el alojamiento de los resultados

Participante	Resultado	'<' ignorado	'<' eliminado	0,5 '<' valor
A	<10	10	--	5
B	<10	10	--	5
C	12	12	12	12
D	19	19	19	19
E	<20	20	--	10
F	20	20	20	20
G	23	23	23	23
H	23	23	23	23
J	25	25	25	25
K	25	25	25	25
L	26	26	26	26
M	28	28	28	28
N	28	28	28	28
P	<30	30	--	15
Q	28	28	28	28

R	29	29	29	29
S	30	30	30	30
T	30	30	30	30
U	31	31	31	31
V	32	32	32	32
W	32	32	32	32
Y	45	45	45 #	45
Z	<50	50 #	--	25
Sumario				
Número de Resultados	23	23	18	23
\bar{x}^*		26,01	26,81	23,95
s^*		7,23	5,29	8,60

La elección de cómo manejar los "menos de" muestras tiene un efecto significativo sobre la media y la desviación típica robusta, y en la evaluación del desempeño. Se espera por el proveedor de ensayos de aptitud para determinar un método adecuado.

E.2 La homogeneidad y la prueba de estabilidad - Arsénico (As) en el chocolate (sección 6.1)

Artículos del ensayo de aptitud se preparan para su uso en una prueba de competencia internacional, y luego para su posterior utilización como material de referencia. 1000 viales se fabrican.

Prueba de Homogeneidad: 10 artículos de la prueba de aptitud se seleccionan mediante una selección aleatoria estratificada de elementos de pruebas de eficiencia de las diferentes partes del proceso de fabricación. 2 porciones de ensayo se extraen de cada botella y se ensayaron en un orden aleatorio, en condiciones de repetibilidad. Los datos se presentan en la Tabla E.2 continuación. El procedimiento en el anexo B.3 es seguido, dando lugar al resumen estadístico listado. La aptitud para el fin propuesto σ_{pt} para el As en el chocolate es del 15%, por lo que la estimación de la variabilidad de la muestra se comprueba contra 0,3 veces el σ_{pt} .

- Tabla E.2 Datos de homogeneidad para los artículos de la prueba de aptitud de arsénico en el chocolate

Botella ID	Réplica 1	Réplica 2
3	0,185	0,194
111	0,187	0,189
201	0,182	0,186
330	0,188	0,196
405	0,191	0,181
481	0,188	0,180
599	0,187	0,196
704	0,177	0,186
766	0,179	0,187
858	0,188	0,196

Promedio general: 0,18715
SD de promedios: 0,00398
 $s_w:$ 0,00556
 $s_s:$ 0,00060
 $\sigma_{pt}:$ $= 0,18715 \times 0,15 = 0,02807$
Comprobar valor: $0,3\sigma_{pt} = 0,00842$
 $s_s:$ es menor que el valor de comprobación, por lo que la homogeneidad es suficiente.

Prueba de Estabilidad: 2 artículos del ensayo de aptitud son seleccionados al azar y se almacenan a una temperatura elevada (60 °C) durante la ejecución de la ronda del programa de ensayos de aptitud (6 semanas). Los artículos de la prueba de aptitud fueron ensayados por duplicado (Tabla E.3), y los cuatro resultados se comparan con los valores de homogeneidad.

Tabla E.3 - Datos de estabilidad para los artículos de la prueba de aptitud para el arsénico en el chocolate

Estabilidad de la muestra	Réplica 1	Réplica 2
164	0,191	0,198
732	0,190	0,196

Promedio general: = 0,19375
Diferencia de Homogeneidad significa: $0,19375 - 0,18715 = 0,00660$

Valor de Comprobación: $0,3\sigma_{pt} = 0,00842$

La diferencia es menor que el valor de comprobación, por lo que la estabilidad es suficiente.

E.3 Ejemplo ampliado de atrazina en agua potable

Un esquema de ensayos de aptitud para un herbicida (atrazina) en el agua potable cuenta con 34 participantes. Estos datos en bruto tal como fue presentado en la Tabla E.4, ordenan por valores para mayor claridad. La tabla muestra los valores calculados para la media y la desviación típica robusta siguiente algoritmo A, después de 6 iteraciones hasta la media robusta y la desviación típica no cambian en sus terceros cifras significativas. Los datos se muestran como gráfico de datos clasificado en la figura E.1 y en histograma y la densidad del núcleo parcela en las figuras E.2 y E.3 figura, respectivamente correspondiente.

La Tabla E.5 muestra las estimaciones de la ubicación (promedio) y la desviación típica utilizando diversas técnicas clásicas y robustas. También se muestra la incertidumbre de la estimación de ubicación. Las estadísticas para el método de remuestreo (bootstrap) se derivan de los procedimientos en las referencias [17,18] y el paquete de software R [ver R3.1.1 abajo]. Figura E.4 muestra las diferentes estimaciones de la ubicación y la estimación de la incertidumbre expandida ($2u(XPT)$) como la barra de error.

Tabla E.4 - Cálculo de la media robusta y desviación típica de atrazina en el agua potable

x_i	1 st iteración	2 nd iteración	3 rd iteración	4 th iteración	5 th iteración	6 th iteración
$x^* - \delta$	0,204163	0,199732	0,198466	0,198037	0,197865	0,197790
$x^* + \delta$	0,319837	0,315969	0,315871	0,316065	0,316185	0,316243
1	0,0400	0,2042	0,1997	0,1985	0,1980	0,1979
2	0,0550	0,2042	0,1997	0,1985	0,1980	0,1979
3	0,1780	0,2042	0,1997	0,1985	0,1980	0,1979
4	0,2020	0,2042	0,2020	0,2020	0,2020	0,2020
5	0,2060	0,2060	0,2060	0,2060	0,2060	0,2060
6	0,2270	0,2270	0,2270	0,2270	0,2270	0,2270
7	0,2280	0,2280	0,2280	0,2280	0,2280	0,2280
8	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300
9	0,2300	0,2300	0,2300	0,2300	0,2300	0,2300
10	0,2350	0,2350	0,2350	0,2350	0,2350	0,2350

11	0,2360	0,2360	0,2360	0,2360	0,2360	0,2360	0,2360
12	0,2370	0,2370	0,2370	0,2370	0,2370	0,2370	0,2370
13	0,2430	0,2430	0,2430	0,2430	0,2430	0,2430	0,2430
14	0,2440	0,2440	0,2440	0,2440	0,2440	0,2440	0,2440
15	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450	0,2450
16	0,2555	0,2555	0,2555	0,2555	0,2555	0,2555	0,2555
17	0,2600	0,2600	0,2600	0,2600	0,2600	0,2600	0,2600
18	0,2640	0,2640	0,2640	0,2640	0,2640	0,2640	0,2640
19	0,2670	0,2670	0,2670	0,2670	0,2670	0,2670	0,2670
20	0,2700	0,2700	0,2700	0,2700	0,2700	0,2700	0,2700
21	0,2730	0,2730	0,2730	0,2730	0,2730	0,2730	0,2730
22	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740
23	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740	0,2740
24	0,2780	0,2780	0,2780	0,2780	0,2780	0,2780	0,2780
25	0,2811	0,2811	0,2811	0,2811	0,2811	0,2811	0,2811
26	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870
27	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870	0,2870
28	0,2880	0,2880	0,2880	0,2880	0,2880	0,2880	0,2880
29	0,2890	0,2890	0,2890	0,2890	0,2890	0,2890	0,2890
30	0,2950	0,2950	0,2950	0,2950	0,2950	0,2950	0,2950
31	0,2960	0,2960	0,2960	0,2960	0,2960	0,2960	0,2960
32	0,3110	0,3110	0,3110	0,3110	0,3110	0,3110	0,3110
33	0,3310	0,3198	0,3160	0,3159	0,3161	0,3162	0,3162
34	0,4246	0,3198	0,3160	0,3159	0,3161	0,3162	0,3162
Promedio	0,2512	0,2579	0,2572	0,2571	0,2570	0,2570	0,2570
SD	0,0672	0,0342	0,0345	0,0347	0,0348	0,0348	0,0348
δ		0,0578	0,0581	0,0587	0,0590	0,0592	0,0592
Nuevo x^*	0,2620	0,2579	0,2572	0,2571	0,2570	0,2570	0,2570
Nuevo s^*	0,0386	0,0387	0,0391	0,0393	0,0394	0,0395	0,0395

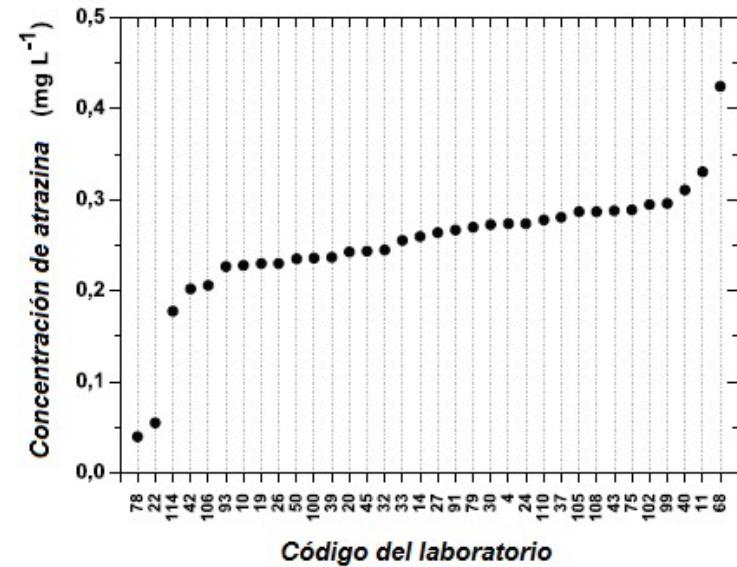


Figura E.1 Clasificación de los resultados de participantes para la atrazina (datos de la Tabla E.4)

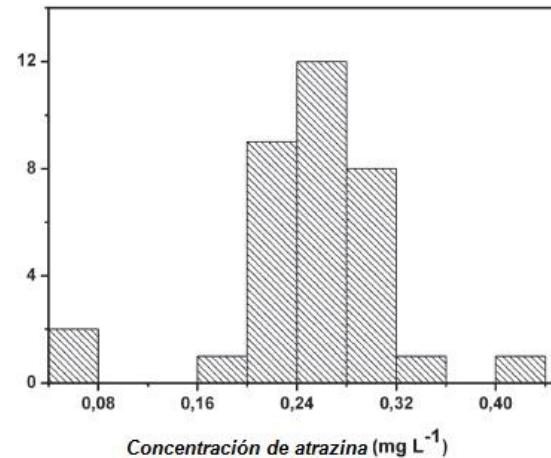


Figura E.2 - Histograma de resultados de participantes

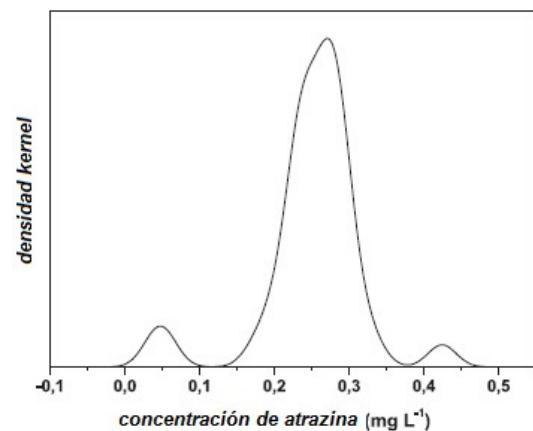


Figura E.3 – Diagrama de densidad Kernel de resultados de participantes

Tabla E.5 – Ejemplo resumen de Estadísticos para atrazina

Procedimiento	Ubicación (Promedio)	Desviación típica	$u(x_{pi})$
robusto: Mediana, nIQR (MADE)	0,2620	0,0402 (0,0386)	0,0086
Robusto: Algoritmo A (x^*, s^*)	0,2570	0,0395	0,0085
Robusto: Q/Hampel	0,2600	0,0426	0,0091
Manos a la Obra (por media)	0,2503	0,0667	0,0113
Aritmética, valores atípicos removidos	0,2588	0,0337	0,0061
Aritmética, valores atípicos incluidos	0,2512	0,0672	0,0115

NOTA diferentes paquetes de software comerciales tienen diferentes procedimientos para el cálculo de cuartiles, que pueden causar diferencias NOTables en nIQR. Las pequeñas diferencias de las cifras anteriores podrían ser causadas por esas diferencias, o por diferentes procedimientos de redondeo

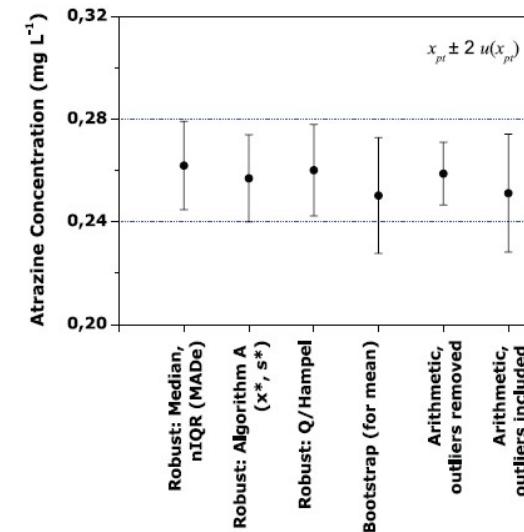


Figura E.4 - Resumen de las estadísticas robustas de la Tabla E.5

E.4 Ejemplo ampliado sobre mercurio en alimentos para animales.

En una ronda de ensayo de aptitud se instruye a los participantes que informen sus resultados como ellos lo hacen cotidianamente y que informen su incertidumbre expandida (U_{lab}) y el factor de cobertura (k). La incertidumbre típica (u_{lab}) se calcula entonces por el proveedor del ensayo de aptitud como U_{lab}/k . Se asignan indicadores a las incertidumbres informadas, siguiendo los criterios discutidos en la sección 9.8. La información mostrada en las Tablas E.6 y E.7 se refiere a mercurio total en alimentos. En la Tabla E.6 la incertidumbre típica u_{lab} se calcula a partir de la incertidumbre expandida de los participantes U_{lab} , dividiendo por el factor de cobertura informado k , y son mostrados aquí con valores redondeados. Para el cálculo de la estadística del desempeño en la Tabla E.7, se utilizan los valores no redondeados. El factor de cobertura, del participante con código L23, no fue informado, y se utilizó 1,732 (la raíz cuadrada redondeada de 3).

Los valores del desempeño fueron calculados utilizando las técnicas descritas en la sección 9. Para todos los cálculos se utilizó un valor de referencia como x_{pt} y σ_{pt} fue el valor de la aptitud-para-el fin-propuesto basado en experiencias anteriores. La incertidumbre del valor asignado es la incertidumbre típica combinada del valor de referencia más la incertidumbre debido a la homogeneidad (diferencias frasco a frasco).

$$x_{pt} = 0,044 \text{ mg/kg}; U(x_{pt}) = 0,0082 \text{ mg/kg}; \sigma_{pt} = 0,0066 \text{ mg/kg} (=15\%);$$

La densidad Kernel ploteada en la Figura E.6 muestra una muy clara distribución bimodal, debido a las diferencias de métodos. Esto no tuvo impacto en la evaluación del desempeño, porque se utilizó un valor de referencia como x_{pt} y un valor de la aptitud-para-el fin-propuesto de σ_{pt} . Para este análisis fueron retirados los resultados con valores "menores que un valor dado" (<).

Tabla E.6 — Resultados del Ensayo de Aptitud con 24 participantes en el estudio IMEP 111

Código Lab	Valor	U_{lab}	k	u_{lab}	Indicador	Método
L04	0,013	0,003	2	0,002	b	AMA
L05	0,013	0,007	2	0,004	a	AMA
L23	0,0135	0,00108	1,732	0,00062	b	AMA
L02	0,014	0,004	2	0,002	b	AMA
L15	0,014	0,0005	2	0,0003	b	AMA
L17	<0,015					CV-ICP-AES
L06	0,016	0,003	2	0,002	b	AMA
L09	0,017	0,008	2	0,004	a	AMA
L26	0,019	0,003	2	0,002	b	AAS
L12	0,0239	0,0036	2	0,0018	b	AMA
L13	<0,034					TDA-AAS
L03	0,037	0,013	2	0,007	a	CV-AAS
L29	0,039	0,007	2	0,004	a	CV-AAS
L07	0,04	0,008	2	0,004	a	ICP-MS
L21	0,04	0,03	2	0,02	c	HG-AAS
L25	0,040	0,010	2	0,005	a	CV-AAS
L16	0,0424	0,008	2	0,004	a	CV-AAS
L08	0,044	0,007	2	0,004	a	CV-AAS
L10	0,045	0,007	2	0,004	a	ICP-MS
L24	0,045	0,005	2	0,003	a	HG-AAS
L18	0,046	0,007	2	0,004	a	CV-AAS
L28	0,049	0,0072	2	0,0036	a	CV-AAS
L01	0,053	0,007	2	0,004	a	CV-AAS
L14	<0,1					ICP-MS

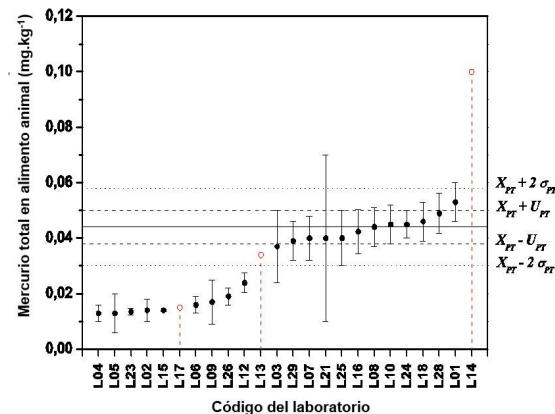


Figura E.5 — Resultados por participante e incertidumbres por resultado en IMEP 111 (información de la Tabla E.6)

Las líneas de guiones representan a $\pm U_{(xp)}$ y las líneas de puntos a $\pm 2\sigma_{pt}$
Los círculos abiertos y las líneas verticales de guiones corresponden a resultados expresados como "menor que un valor"

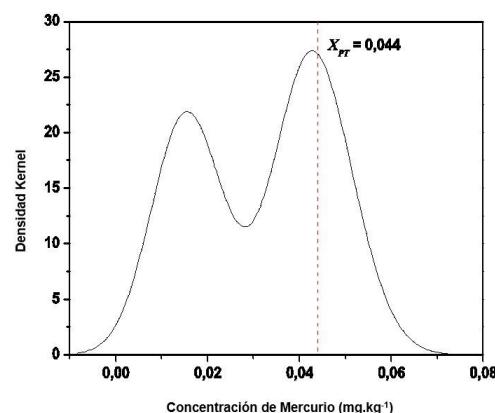


Figura E.6 — Ploteo de la densidad Kernel por resultado de los participantes

Tabla E.7 — Estadística del desempeño por varios métodos

Código Lab	D %	P _A	z	z'	ζ	E _n
L04	-70,5%	-156,6%	-4,70	-3,99	-7,10	-3,55
L05	-70,5%	-156,6%	-4,70	-3,99	-5,75	-2,88
L23	-69,3%	-154,0%	-4,62	-3,93	-7,35	-3,69
L02	-68,2%	-151,5%	-4,55	-3,86	-6,58	-3,29
L15	-68,2%	-151,5%	-4,55	-3,86	-7,30	-3,65
L17						
L06	-63,6%	-141,4%	-4,24	-3,60	-6,41	-3,21
L09	-61,4%	-136,4%	-4,09	-3,47	-4,71	-2,36
L26	-56,8%	-126,3%	-3,79	-3,22	-5,73	-2,86
L12	-45,7%	-101,5%	-3,05	-2,59	-4,49	-2,24
L13						
L03	-15,9%	-35,4%	-1,06	-0,90	-0,91	-0,46
L29	-11,4%	-25,3%	-0,76	-0,64	-0,93	-0,46
L07	-9,1%	-20,2%	-0,61	-0,51	-0,70	-0,35
L21	-9,1%	-20,2%	-0,61	-0,51	-0,26	-0,13
L25	-9,1%	-20,2%	-0,61	-0,51	-0,62	-0,31
L16	-3,6%	-8,1%	-0,24	-0,21	-0,28	-0,14
L08	0,0%	0,0%	0,00	0,00	0,00	0,00
L10	2,3%	5,1%	0,15	0,13	0,19	0,09
L24	2,3%	5,1%	0,15	0,13	0,21	0,10
L18	4,5%	10,1%	0,30	0,26	0,37	0,19
L28	11,4%	25,3%	0,76	0,64	0,92	0,46
L01	20,5%	45,5%	1,36	1,16	1,67	0,83
L14						

*Este ejemplo es cortesía de European Commission Joint Research Centre, Institute for Reference Materials and Measurements, and International Measurement Evaluation Program (IMEP®), estudio 111.

E.5 Valor de referencia de un solo laboratorio: Valor de agregado de Los Ángeles (sección 7.5)

La tabla E.8 (en la siguiente página) muestra un ejemplo de datos que podrían obtenerse en una serie de ensayos de un ensayo de aptitud y un material de referencia certificado muy similar (MRC) con un valor de propiedad certificado de 21.62 unidades LA y la incertidumbre asociada 0.26 unidades LA. Este ejemplo muestra cómo se obtiene un valor de referencia y de incertidumbre mediante un ensayo de aptitud. Observe que la incertidumbre del valor certificado para el MRC incluye la incertidumbre debido a la no homogeneidad, transporte, y estabilidad a largo plazo.

$$x_{pt} = 21,62 + 1,73 = 23,35 \text{ unidades LA}$$

Y,

$$u(x_{pt}) = \sqrt{0,26^2 + 0,24^2} = 0,35 \text{ unidades LA}$$

donde 0.26 es la incertidumbre típica del valor certificado del MRC, y 0.24 es la incertidumbre típica de \bar{d} .

Tabla E.8 — Cálculo de la diferencia media entre un MRC y un ensayo de aptitud y de la incertidumbre típica de esta diferencia

Muestra	Ensayo de Aptitud		MRC		Diferencia en los valores medios EA – MRC unidades LA
	Ensayo 1 Unidades LA	Ensayo 2 Unidades LA	Ensayo 1 Unidades LA	Ensayo 2 Unidades LA	
1	20,5	20,5	19,0	18,0	2,00
2	21,1	20,7	19,8	19,9	1,05
3	21,5	21,5	21,0	21,0	0,50
4	22,3	21,7	21,0	20,8	1,10
5	22,7	22,3	20,5	21,0	1,75
6	23,6	22,4	20,3	20,3	2,70
7	20,9	21,2	21,5	21,8	-0,60
8	21,4	21,5	21,9	21,7	-0,35
9	23,5	23,5	21,0	21,0	2,50
10	22,3	22,9	22,0	21,3	0,95
11	23,5	24,1	20,8	20,6	3,10
12	22,5	23,5	21,0	22,0	1,50
13	22,5	23,5	21,0	21,0	2,00
14	23,4	22,7	22,0	22,0	1,05
15	24,0	24,2	22,1	21,5	2,30
16	24,5	24,4	22,3	22,5	2,05
17	24,8	24,7	22,0	21,9	2,80
18	24,7	25,1	21,9	21,9	3,00
19	24,9	24,4	22,4	22,6	2,15
20	27,2	27,0	24,5	23,7	3,00
Diferencia media, \bar{d}					1,73
Desviación típica					1,07
Incertidumbre típica de \bar{d} (desviación típica / $\sqrt{20}$)					0,24
NOTA: Los datos son mediciones de la Fortaleza mecánica del agregado, obtenidas del ensayo Los Ángeles (LA).					

E.6 Ejemplo de la técnica de secuencia para Coliformes en una muestra de alimento (sección 7.7.2)

Un esquema de ensayo de aptitud para Coliformes en una muestra de alimento (leche) contó con 35 participantes que realizaron cinco mediciones como réplicas independientes. La media de los datos del "log UFC" de cada participante fue utilizada para estimar el valor asignado y su incertidumbre. Un valor de la aptitud-para-el fin-propuesto igual a "0.25 log UFC/ml" fue tomado como σ_{pt} , mientras que la desviación típica de la función Kernel fue de 0.75 σ_{pt} (cf. "bw" en el código R). El ploteo de la densidad Kernel (Figura E.7) presenta una distribución asimétrica. El método de secuencia (1000 réplicas) fue aplicado para estimar el modo y el error típico correspondiente de la función de densidad Kernel de la distribución de los datos, fijados como x_{pt} y $u(x_{pt})$, respectivamente. Los valores siguientes fueron derivados:

$$x_{pt} = 3.79 \text{ y } u(x_{pt}) = 0.0922 \text{ en log UFC/ml}$$

NOTA: Ya que $u(x_{pt}) > 0.3 \sigma_{pt}$, el desempeño de los laboratorios fue evaluado utilizando los valores-z'.

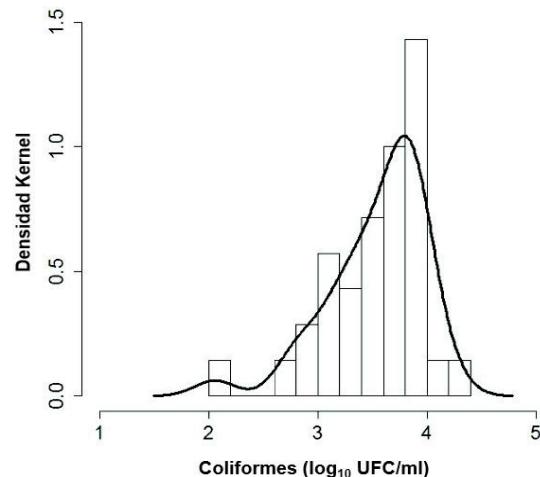


Figura E.7 — Ploteo de la densidad Kernel por resultado de los participantes

R.3.1.1 Código

```
#####
#PROGRAMA PARA DESCARGAR Y UTILIZAR
Se ha mantenido el idioma original del programa.
#####

library(boot)      #for bootstrap estimates
library(pastecs)  #for descriptive statistics

#DATA
#DATA
colif<-c(3.80, 3.90, 3.07, 3.64, 4.06, 3.40, 3.59, 3.39, 3.47, 3.47, 3.77, 3.53, 2.83,
2.75, 2.06, 3.75, 3.73, 3.82, 3.86, 3.88, 3.97, 3.96, 3.80, 3.88, 3.25, 3.45, 3.64, 2.86,
3.17, 3.19, 3.17, 4.22, 3.82, 3.82, 3.95)

#DESCRIPTIVE STATISTICS
options(digits = 3) #number of decimal
stat.desc(colif)

#CONDITIONS
sigmat<-0.25          #standard deviation "fitness for purpose"
bw=0.75*sigmat         #standard deviation of kernel density

#HISTOGRAM AND KERNEL DENSITY GRAPH
hist(colif, freq=F,main="", cex.axis= 1.5,cex.lab=1.5, xlim=c(1,5) , ylim=c(0,1.5),
xlab="Coliforms (log10CFU/ml)",ylab="Kernel density", breaks=10)
lines(density(colif, kernel="gaussian", bw), col="black", lwd=3)

#FUNCTION TO DEFINE THE STATISTICS
theta<- function(y,i)
{
  dens<-density(y[i], kernel="gaussian", bw=bw)
  mode<-dens$x[which.max(dens$y)]
}

#BOOTSTRAP MODE CALCULATION AND ITS UNCERTAINTY
set.seed(220)           #START POINT OF BOOTSTRAP
boot.statistics<- boot(colif,theta,R=1000)
boot.statistics          #MODE AND STANDARD ERROR
```

Cortesía del Instituto Zooprofiláctico Experimental de Venecia – EA Microbiología de Alimentos "AQUA"

E.7 Comparación del valor de referencia y la media por consenso (sección 7.8)

Considere el ejemplo E.4 y los datos de la Tabla E.6 como una demostración del procedimiento en la sección 7.8 para comparar el valor de referencia con la media robusta de los resultados de los participantes.

En esta ronda de un esquema de ensayo de aptitud la media robusta x^* es 0,03161 y la desviación típica robusta s^* es 0,0164, calculadas con el Algoritmo A, después de eliminar 3 resultados que

reportaron valores "menor que" (n=24). Por consiguiente la incertidumbre de la media robusta es calculada como:

$$u(x^*) = 1,25(s^* / \sqrt{n})$$

$$u(x^*) = 1,25(0,0164 / \sqrt{24}) = 0,0042$$

De la sección 7.8, ecuación 8, la incertidumbre de la diferencia entre el x_{ref} y x^* es como sigue:

$$u_{diff} = \sqrt{u^2(x_{ref}) + u^2(x^*)} = \sqrt{0,0041^2 + 0,0042^2} = 0,0059$$

$$U_{diff} = 2(0,0059) = 0,012$$

$x_{diff} = x_{ref} - x^* = 0,044 - 0,032 = 0,012$ así que la diferencia es dos veces la incertidumbre de la diferencia.

No se recomienda ninguna acción, ya que se comprende el sesgo en algunos métodos.

E.8 Determinación de los criterios de evaluación por experiencia de rondas anteriores: Toxafeno en agua potable (sección 8.3)

Hay dos proveedores de ensayos de aptitud organizando esquemas de ensayos de aptitud para el pesticida Toxafeno (un pesticida) en agua potable. Por un período de más de 5 años se han desarrollado 20 rondas de ensayos de aptitud donde había 20 o más participantes, cubriendo los niveles regulados de Toxafeno de 3 a 20 µg/L. La Tabla E.9 muestra a los resultados de las 20 rondas de ensayos de aptitud, ordenadas de abajo hacia arriba según los valores asignados. Las figuras E.8 y E.9 muestran la dispersión al plotear la desviación típica robusta relativa (RSD%) y la desviación típica robusta (SD) para cada ronda del esquema de ensayo de aptitud, comparado con el valor asignado (de la formulación). En cada figura se muestra la fórmula para la regresión lineal por el método de los mínimos cuadrados simples. Las líneas de regresión por mínimos cuadrados pueden determinarse con el software, generalmente disponible, en una hoja de cálculo. (NOTA: un modelo polinómico de segundo orden también fue verificado para la relación entre la desviación típica y el valor asignado, pero el término cuadrático no resultó significativo, indicando que no hay ninguna curva significativa en la línea; por tanto se considera apropiado el modelo lineal simple).

Está claro que la RSD es bastante constante en aproximadamente un 19% para todos los niveles, y que la línea de la regresión para la desviación normal es bastante fiable ($R^2 = 0,82$). Un cuerpo regulador puede escoger requerir que la desviación normal para la evaluación del desempeño sea 19% del valor asignado (o quizás 20%), o ellos pueden requerir el cálculo de la desviación típica esperada, utilizando la ecuación de la regresión para la desviación típica.

Tabla E.9 — Rondas de ensayos de aptitud para Toxafeno en agua potable; $p \geq 20$ resultados

Código del proveedor de EA	Valor Asignado	Media Robusta	Desviación Típica	Recobrado Medio	RSD (% del VA)	p
P004	3,96	3,98	0,639	100,5 %	16,1 %	25
P001	4,56	5,18	0,638	113,6 %	14,0 %	23
P001	5,99	5,98	0,995	99,8 %	16,6 %	22
P004	6,08	5,80	1,48	95,4 %	24,3 %	20
P001	6,20	6,66	0,97	107,4 %	15,7 %	23
P001	6,72	7,13	1,43	106,1 %	21,3 %	22
P004	8,10	7,09	2,23	87,5 %	27,5 %	21
P001	8,73	8,15	1,80	93,4 %	20,6 %	22
P001	9,57	8,60	1,45	89,9 %	15,2 %	23
P001	12,1	12,4	1,44	102,5 %	11,9 %	23
P001	12,5	13,8	2,25	110,4 %	18,0 %	24
P004	13,1	12,0	2,41	91,6 %	18,4 %	20
P004	15,6	13,3	3,57	85,3 %	22,9 %	27
P004	15,9	13,6	2,44	85,5 %	15,3 %	28
P004	16,3	13,5	3,60	82,8 %	22,1 %	31
P004	16,3	14,2	3,09	87,1 %	19,0 %	40
P004	17,0	15,6	2,63	91,8 %	15,5 %	24
P004	17,4	16,0	2,85	92,0 %	16,4 %	23
P004	17,4	16,0	3,36	92,0 %	19,3 %	23
P004	19,0	16,4	3,20	86,3 %	16,8 %	27

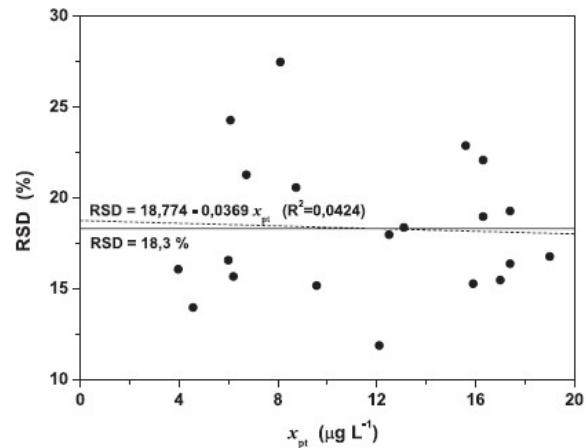


Figura E.8 — Desviación típica relativa de los resultados de los participantes (%) vs valor asignado ($\mu\text{g/L}$)

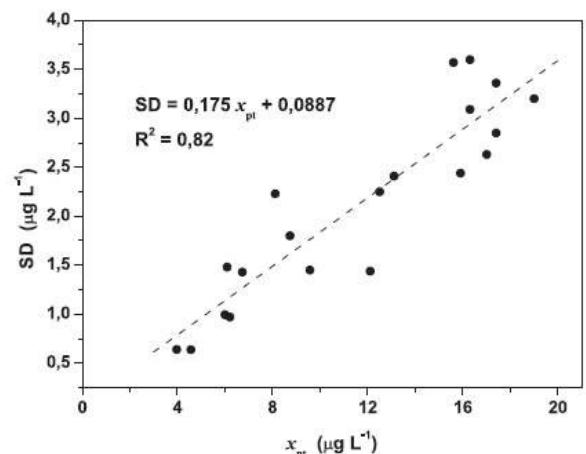


Figura E.9 — Desviación típica de los participantes ($\mu\text{g/L}$) vs valor asignado ($\mu\text{g/L}$)

E.9 A partir de un modelo general: la ecuación de Horwitz (sección 8.4)

Horwitz^[22,31] describió un modelo general común para las aplicaciones químicas. Este enfoque da a un modelo general para la desviación típica de la reproducibilidad de los métodos analíticos, que pueden utilizarse en la derivación de la expresión siguiente para la desviación típica de la reproducibilidad:

$$\sigma_R = 0,02' c^{0,8495}$$

donde c es la concentración de la especie química a determinar como fracción mísica.

Por ejemplo, un esquema de ensayo de aptitud para la melamina en leche en polvo utiliza dos artículos en ensayos de aptitud con los niveles de referencia $A = 1,195 \text{ mg/kg}$ y $B = 2,565 \text{ mg/kg}$ ($0,000\ 001\ 195$ y $0,000\ 002\ 565$). Esto produce la siguiente desviación típica de la reproducibilidad esperada:

Ensayo de aptitud A a $1,195 \text{ mg/kg}$: $\sigma_R = 0,186 \text{ mg/kg}$ o relativa $\sigma_R = 15,6 \%$

Ensayo de aptitud B a $2,565 \text{ mg/kg}$: $\sigma_R = 0,356 \text{ mg/kg}$ o relativa $\sigma_R = 13,9 \%$

E.10 Cómo determinar el desempeño a partir de un experimento de precisión: Determinación del contenido de cemento en un hormigón endurecido (sección 8.5)

El volumen de cemento en el hormigón es normalmente moderado en lo que se refiere a la masa en kilogramos de cemento por metro cúbico de hormigón (es decir kg/m^3). En la práctica, se produce el hormigón en grados de calidad que tienen volúmenes diferenciados de cemento de 25 kg/m^3 entre ellos, y se desea que los participantes puedan identificar la calidad correctamente. Por esta razón, es deseable que el valor escogido de σ_{pt} deba ser no más de la mitad de 25 kg/m^3 ($\sigma_{pt} < 12,5 \text{ kg/m}^3$).

Un experimento de precisión produjo los resultados siguientes, para un hormigón con un contenido de cemento medio de 260 kg/m^3 : $\sigma_R = 23,2 \text{ kg/m}^3$ y $\sigma_r = 14,3 \text{ kg/m}^3$. Asuma que se van a hacer $m=2$ réplicas de las mediciones.

Así, la siguiente ecuación (9):

$$\sigma_{pt} = \sqrt{23,2^2 - 14,3^2 (1 - 1/2)} \text{ kg/m}^3 = 20,9 \text{ kg/m}^3$$

Así, el objetivo de lograr $\sigma_{pt} < 25/2 \text{ kg/m}^3 = 12,5 \text{ kg/m}^3$ puede no ser práctico.

NOTA: En ISO 5725-2, $\sigma_R = \sqrt{\sigma_L^2 + \sigma_r^2}$ donde σ_L es el componente de la varianza debida a las diferencias entre laboratorios.

En este ejemplo σ_L podría ser calculada como $\sigma_L = \sqrt{\sigma_R^2 - \sigma_r^2} = \sqrt{(23,2^2 - 14,3^2)} = 18,3 \text{ kg/m}^3$.

E.11 Ploteo-de-barras de sesgos normalizados: Concentraciones de anticuerpo (sección 10.4)

En la Figura E.10 se muestran los valores-z de una ronda de un ensayo de aptitud con tres mesurandos relacionados (anticuerpos), ploteados en un gráfico de barras. En la Tabla E.10 se muestran los datos para dos de los tres alérgenos. De este gráfico, se puede ver que los laboratorios B y Z (por ejemplo) deben buscar la causa del sesgo que afecta a los tres niveles aproximadamente en la misma cantidad, mientras que según se aprecia en el propio gráfico, en el caso de los laboratorios K y P (por ejemplo), la señal del valor-z depende del tipo de anticuerpo.

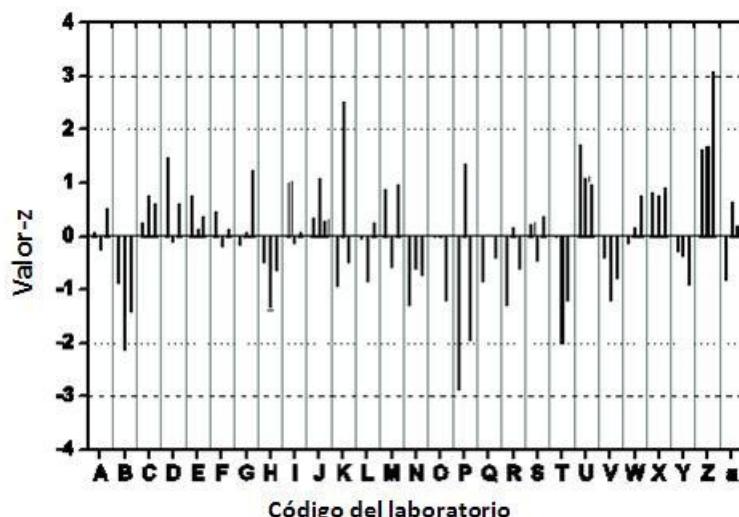


Figura E.10 — Gráfico de barras de valores-z (4,0 a -4,0) para una ronda de un esquema de ensayo de aptitud en el cual los participantes determinaron las concentraciones de tres alérgenos específicos anticuerpos IgE

E.12 Ploteo Youden – concentraciones de anticuerpos (sección 10.5)

La Tabla E.10 muestra los datos obtenidos al ensayar dos artículos similares de ensayos de aptitud para concentraciones de anticuerpo. En la Figura E.11 se muestra el desempeño de los valores (z) que está basado en la media robusta y la desviación típica, utilizando el Algoritmo A.

La inspección de la Figura E.11 revela dos participantes (con números 5 y 23) en la parte superior del cuadrante derecho, que por consiguiente podrían tener un consistente sesgo positivo. El laboratorio 26 tiene un valor-z elevado en el artículo B del ensayo de aptitud y un valor-z negativo de -0,055 en el artículo A del ensayo de aptitud, de manera que podría tener una pobre repetibilidad.

Los resultados de los participantes 5, 23 y 26 deben ser interpretados como generadores de "señales de advertencia", y se debe verificar donde fallan sus resultados en la próxima ronda del esquema de EA. La revisión visual y el coeficiente de correlación indican una tendencia para los valores-z consistentes (positivo o negativo), de manera que podría haber una oportunidad de mejora del método de medición con instrucciones más detalladas.

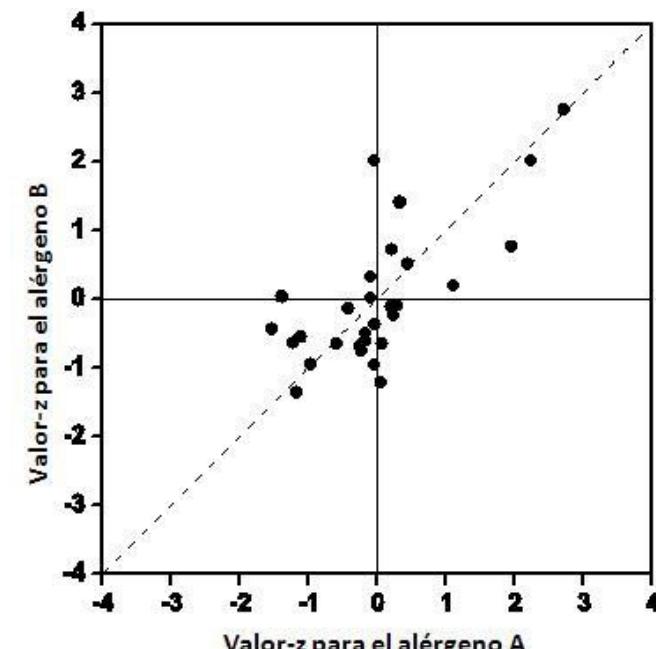


Figura E.11 — Ploteo Youden de valores-z a partir de la Tabla E.10

Tabla E.10 — Datos y cálculos de la concentración de anticuerpos para dos alérgenos similares

Laboratorio	Datos		Valor-z	
<i>i</i>	Alérgeno A: $x_{A,i}$	Alérgeno B: $x_{B,i}$	Alérgeno A: $z_{A,i}$	Alérgeno B: $z_{B,i}$
1	12,95	9,15	0,427	0,515
2	6,47	6,42	-1,540	-0,428
3	11,40	6,60	-0,043	-0,366
4	8,32	4,93	-0,978	-0,942
5	18,88	13,52	2,228	2,023
6	15,14	8,22	1,092	0,194
7	10,12	7,26	-0,432	-0,138
8	17,94	9,89	1,942	0,770
9	11,68	4,17	0,042	-1,204
10	12,44	7,39	0,272	-0,093
11	6,93	7,78	-1,400	0,042
12	9,57	5,80	-0,599	-0,642
13	11,73	5,77	0,057	-0,652
14	12,29	6,97	0,227	-0,238
15	10,95	6,23	-0,180	-0,493
16	10,95	5,90	-0,180	-0,607
17	11,17	7,74	-0,113	0,028
18	11,20	8,63	-0,104	0,335
19	7,64	3,74	-1,185	-1,353
20	12,17	7,33	0,190	-0,114
21	10,71	5,70	-0,253	-0,676
22	7,84	6,07	-1,124	-0,549
23	20,47	15,66	2,710	2,762
24	12,60	11,76	0,321	1,415
25	11,37	4,91	-0,052	-0,949
26	11,36	13,51	-0,055	2,019
27	10,75	5,48	-0,241	-0,752
28	12,21	9,77	0,203	0,729
29	7,49	5,82	-1,230	-0,635
Promedio	11,54	7,66	0,00	0,00
Desv. típica	3,29	2,90	1,00	1,00
Coef. Correlación	0,706		0,706	

NOTA 1: Los datos son números de unidades (U) en miles (k) por litro (L) de muestra donde una unidad se define por la concentración de un material de referencia internacional.

NOTA 2: Los valores-z en esta tabla han sido calculados utilizando valores no redondeados de los promedios robustos y desviaciones típicas, no utilizando los valores redondeados mostrados en el trasfondo de la tabla.

E.13 Ploteo de las desviaciones típicas de la repetibilidad: Concentraciones de anticuerpo (sección 10.6)

La tabla E.11 muestra los resultados de concentraciones determinantes de cierto anticuerpo en muestras de suero en ensayos de aptitud. Cada participante hizo cuatro réplicas bajo condiciones de repetibilidad. Se utiliza la fórmula que aparece arriba para obtener el ploteo mostrado en la Figura E.12. El ploteo muestra que algunos laboratorios reciben acciones o señales de advertencia.

Tabla E.11 — Concentraciones de ciertos anticuerpos en suero en ensayos de aptitud (4 réplicas por determinación en una muestra de ensayo de aptitud por cada participante)

Laboratorio	Promedio kU/L	Desviación típica kU/L
1	2,15	0,13
2	1,85	0,21
3	1,80	0,08
4	1,80	0,24
5	1,90	0,36
6	1,90	0,32
7	1,90	0,14
8	2,05	0,26
9	2,35	0,39
10	2,03	0,53
11	2,08	0,25
12	1,25	0,24
13	1,13	0,72
14	1,00	0,26
15	1,08	0,17
16	1,20	0,32
17	1,35	0,4
18	1,23	0,36
19	1,23	0,33
20	0,90	0,43
21	1,48	0,40
22	1,20	0,55
23	1,73	0,39
24	1,43	0,30
25	1,28	0,22
Media robusta	1,57	
Desv. típica robusta		0,34

NOTA: Los datos son números de unidades (U) en miles (k) por litro (L) de muestra, donde una unidad se define por la concentración de un material de referencia internacional.

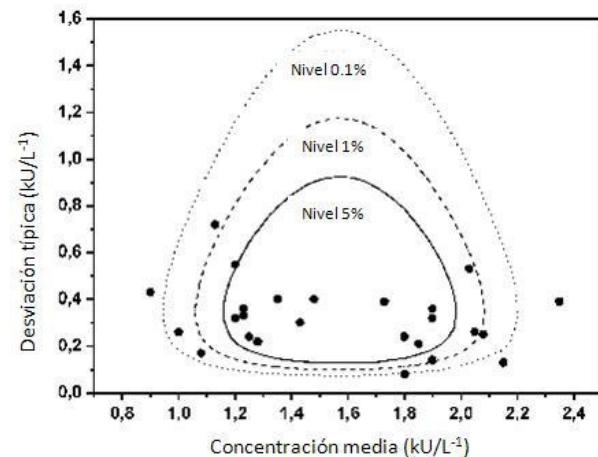


Figura E.12 — Ploteo de las desviaciones típicas contra los promedios de 25 participantes (datos de la Tabla E.10)

E.14 Métodos gráficos para rastrear el desempeño en el tiempo (sección 10.8)

Puede ser útil para un participante rastrear su propio desempeño en el tiempo, o tener esto preparado por el proveedor del ensayo de aptitud. Una herramienta simple y convencional es una carta de control de la calidad, o ploteo de Shewhart. Esto requiere tener un valor del desempeño normalizado, como el valor-z o el valor- P_A así como haber participado en varias rondas. Este ejemplo corresponde a un esquema de ensayo de aptitud médica, para potasio en suero.

Este proveedor de ensayo de aptitud utiliza un intervalo fijo de aceptación del 5%, sin embargo redondeando al próximo valor reportable (0,1mmol/L), y no menor que $\pm 0,2$ mmol/L. El proveedor de ensayo de aptitud utiliza los valores- P_A en lugar de los valores-z.

Tabla E.12 — Valores- P_A para 5 rondas de un esquema de ensayo de aptitud, cada uno con 3 objetos de ensayo (muestras) de Potasio en Suero.

Código de la Ronda	Objeto de ensayo (muestra)	Resultado	Valor asignado	Valor- P_A	Media de P_A
101	A	6,4	6,2	75	42
101	B	4,2	4,1	50	
101	C	4,1	4,1	0	
102	A	6,0	5,9	25	8
102	B	4,3	4,4	-33	
102	C	5,5	5,4	33	
103	A	4,1	4,2	-33	-28
103	B	3,6	3,7	-50	
103	C	4,2	4,2	0	
104	A	5,7	5,8	-25	11
104	B	3,9	4,0	-50	
104	C	6,3	5,9	110	
105	A	3,6	3,7	-50	-19
105	B	4,5	4,6	-33	
105	C	5,3	5,2	25	

Los resultados pueden plotearse fácilmente para la revisión. Se recomiendan 2 tipos de ploteos:

- Carta de control de la calidad del valor de desempeño normalizado para cada ronda, mostrando múltiples objetos de muestra en la misma ronda del ensayo de aptitud. Esto resaltaría el desempeño en el tiempo, incluyendo cualquier tendencia; mostrado en la Figura E.13.
- Ploteo de dispersión de valores del desempeño normalizados contra valores asignados, para ver si el desempeño se relaciona con el nivel de concentración, con el objetivo de mostrar cualquier tendencia relacionada con el nivel del mesurando; mostrado en la Figura E.14.

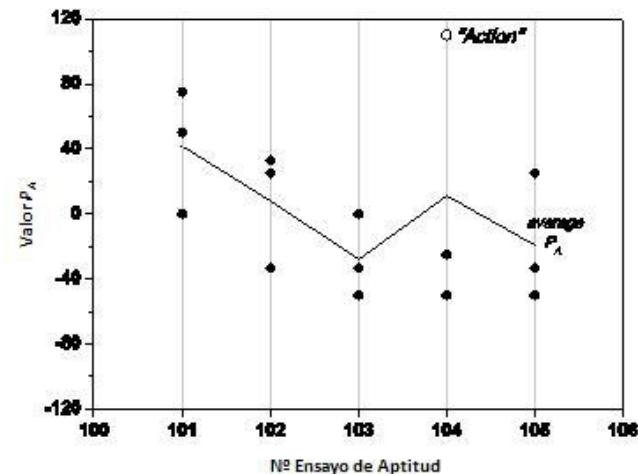


Figura E.13 — Valores del desempeño para cada ronda (datos de la Tabla E.12)

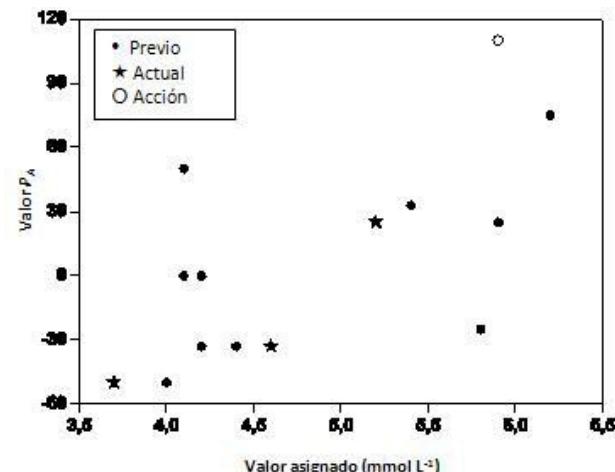


Figura E.14 — Valores del desempeño para diferentes niveles del mesurando

E.15 Datos de Análisis Cualitativo; ejemplo de una cantidad ordinal: reacción de la piel a un cosmético (sección 11)

Un esquema de ensayo de aptitud involucra el análisis de la reacción a un producto para el cuidado de la piel, cuando se aplica a un sujeto animal patrón. Cualquier reacción inflamatoria se evalúa según la escala siguiente:

1. no reacción
2. enrojecimiento moderado
3. irritación significativa o hinchazón
4. reacción severa, incluyendo supuración o sangramiento

Se distribuyeron dos objetos de ensayo que consisten en dos productos diferentes, etiquetados como productos A y B, y hay 50 participantes para cada producto. Los resultados de los participantes se relacionan en la Tabla E.13 y se muestran gráficamente en la Figura E.15. La moda y la mediana se listan por resultado de los participantes para cada objeto de ensayo del ensayo de aptitud.

Tabla E.13 — Resultados para el desempeño de dos objetos de ensayo, irritación de la piel

Reacción	Producto A	Producto B
1	20 (40 %) [#]	8 (16 %)
2	18 (36 %) [@]	12 (24 %)
3	10 (20 %)	20 (40 %) ^{#@}
4	2 (4 %)	10 (20 %)
#moda		
@mediana		

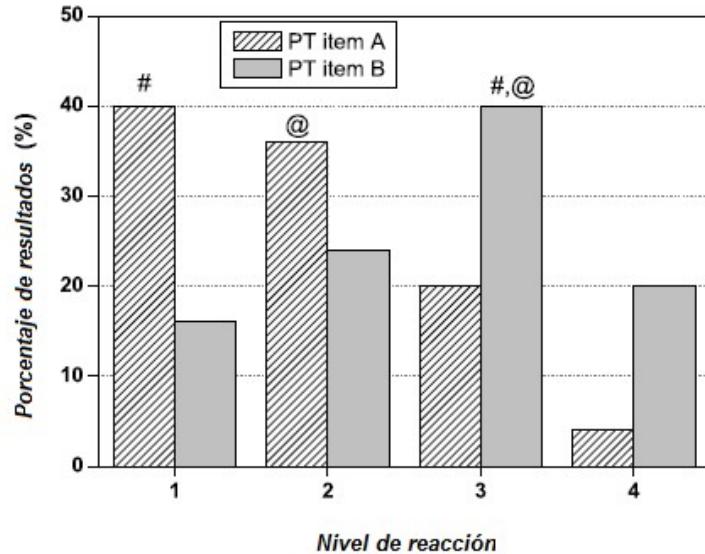


Figura E.15 — Carta de barras del por ciento de respuestas del desempeño en dos objetos de ensayo para evaluar la irritación de la piel — # moda, @ mediana

Observe que la mediana o moda pueden usarse como estadística resumen para éstos objetos de ensayo del ensayo de aptitud, y ellos sugieren que el nivel de reacción al producto B fue más severo que la reacción al producto A. El proveedor del ensayo de aptitud puede determinar qué "señales de acción" ocurrirían para cualquier resultado que esté más allá de una unidad ordinal de la mediana, en cuyo caso para el producto A, las señales de acción ocurren para los 2 resultados (4%) de "4" y para el producto B, las señales de acción ocurren para los 8 resultados (16%) de "1."

Bibliografía

- [1] ISO 5725-2, *Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method*.
- [2] ISO 5725-3, *Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method*.
- [3] ISO 5725-4, *Accuracy (trueness and precision) of measurement methods and results — Part 4: Basic methods for the determination of the trueness of a standard measurement method*.
- [4] ISO 5725-5, *Accuracy (trueness and precision) of measurement methods and results — Part 5: Alternative methods for the determination of the precision of a standard measurement method*.
- [5] ISO 5725-6, *Accuracy (trueness and precision) of measurement methods and results — Part 6: Use in practice of accuracy values*.
- [6] ISO 7870-2, (2013), *Control charts — Part 2: Shewhart control charts*.
- [7] ISO 11352, *Water quality — Estimation of measurement uncertainty based on validation and quality control data*.
- [8] ISO 11843-1, *Capability of detection — Part 1: Terms and definitions*.
- [9] ISO 11843-2, *Capability of detection — Part 2: Methodology in the linear calibration case*.
- [10] ISO 16269-4, *Statistical interpretation of data — Part 4: Detection and treatment of outliers*.
- [11] ISO/IEC 17011, *Conformity assessment — General requirements for accreditation bodies accrediting conformity assessment bodies*.
- [12] ISO/IEC 17025, *General requirements for the competence of testing and calibration laboratories*.
- [13] ISO Guide 35, *Reference materials — General and statistical principles for certification*.
- [14] ISO/IEC Guide 98-3, *Uncertainty of measurement — Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)*
- [15] ANALYTICAL METHOD COMMITTEE. Royal Society of Chemistry Accred Qual Assur. 2010, **15** pp. 73–79
- [16] CCQM Guidance note: Estimation of a consensus KCRV and associated Degrees of Equivalence. Version 10. Bureau International des Poids et Mesures, Paris (2013)
- [17] DAVISON, A. C., HINKLEY, D. V. *Bootstrap Methods and Their Application*. Cambridge University Press, 1997
- [18] EFRON, B., TIBSHIRANI, R. *An Introduction to the Bootstrap*. Chapman & Hall, 1993
- [19] FRES, J. Anal Chem 360_359-361
- [20] GOWER, J. C. A general coefficient of similarity and some of its properties. *Biometrics*. 1971, **27** (4) pp. 857–871
- [21] HELSEL, D. R. *Non-detects and data analysis: statistics for censored environmental data*. Wiley Interscience, 2005
- [22] HORWITZ, W. Evaluation of analytical methods used for regulations of food and drugs. *Anal. Chem.* 1982, **54** pp. 67A–76A
- [23] JACKSON, J. E. Quality control methods for two related variables. *Industrial Quality Control*. 1956, **7** pp. 2–6

- [24] KUSELMAN, I., FAJGELJ, A. IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for a limited number of participants — chemical analytical laboratories (IUPAC Technical Report). *Pure Appl. Chem.* 2010, **82** (5) pp. 1099–1135
- [25] MARONNA, R. A., MARTIN, R. D., YOHAI, V. J. *Robust Statistics: Theory and methods*. John Wiley & Sons Ltd, Chichester, England, 2006
- [26] MÜLLER, C. H., UHLIG, S. Estimation of variance components with high breakdown point and high efficiency; *Biometrika*; **88**: Vol. 2, pp. 353-366, 2001.
- [27] ROUSSEEUW, P. J., VERBOVEN, S. *Comput. Stat. Data Anal.* 2002, **40** pp. 741–758
- [28] SCOTT, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992
- [29] SHEATHER, S. J., JONES, M. C. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Stat. Soc., B*. 1991, **53** pp. 683–690
- [30] SILVERMAN B. W. *Density Estimation*. Chapman and Hall, London, 1986
- [31] THOMPSON, M. *Analyst (Lond.)*. 2000, **125** pp. 385–386
- [32] THOMPSON, M., ELLISON, S. L. R., WOOD, R. "The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories" (IUPAC Technical Report). *Pure Appl. Chem.* 2006, **78** (1) pp. 145–196
- [33] THOMPSON, M., WILLETTS, P., ANDERSON, S., BRERETON, P., WOOD, R. Collaborative trials of the sampling of two foodstuffs, wheat and green coffee. *Analyst (Lond.)*. 2002, **127** pp. 689–691
- [34] UHLIG, S. *Robust estimation of variance components with high breakdown point in the 1-way random effect model*. In: Kitsos, C. P. and Edler, L.; Industrial Statistics; Physica, S. 65-73, 1997.
- [35] UHLIG, S. Robust estimation of between and within laboratory standard deviation measurement results below the detection limit, *Journal of Consumer Protection and Food Safety*, 2015
- [36] VAN NULAND, Y. ISO 9002 and the circle technique. *Qual. Eng.* 1992, **5** pp. 269–291
- [37] <http://quodata.de/en/web-services/QHampel.html>