NORMA TÉCNICA COLOMBIANA

NTC-ISO 13528

2017-12-06

MÉTODOS ESTADÍSTICOS PARA USO EN ENSAYOS DE APTITUD POR COMPARACIÓN INTERLABORATORIO

E: STATISTICAL METHODS FOR USE IN PROFICIENCY TESTING BY INTERLABORATORY COMPARISON

CORRESPONDENCIA:	esta norma es adopción Idéntica por traducción (IDT) a la ISO 13528: 2015
DESCRIPTORES:	método estadístico; ensayos de aptitud; comparación interlaboratorio.

I.C.S.: 03.120.30

Editada por el Instituto Colombiano de Normas Técnicas y Certificación (ICONTEC)
Apartado 14237 Bogotá, D.C. - Tel. (571) 6078888 - Fax (571) 2221435

Editada 2017-12-13

PRÓLOGO

El Instituto Colombiano de Normas Técnicas y Certificación, **ICONTEC**, es el organismo nacional de normalización, según el Decreto 1595 de 2015.

ICONTEC es una entidad de carácter privado, sin ánimo de lucro, cuya Misión es fundamental para brindar soporte y desarrollo al productor y protección al consumidor. Colabora con el sector gubernamental y apoya al sector privado del país, para lograr ventajas competitivas en los mercados interno y externo.

La representación de todos los sectores involucrados en el proceso de Normalización Técnica está garantizada por los Comités Técnicos y el período de Consulta Pública, este último caracterizado por la participación del público en general.

La norma NTC-ISO 13528 fue ratificada por el Consejo Directivo de 2017-12-06.

Esta norma está sujeta a ser actualizada permanentemente con el objeto de que responda en todo momento a las necesidades y exigencias actuales.

A continuación se relacionan las empresas que colaboraron en el estudio de esta norma a través de su participación en el Comité Técnico 04 Aplicación de métodos estadísticos.

2C DESIGN
CALIDAD Y ESTADÍSTICA
CCOLSMI LTDA.
CORPOICA
HIDROMÉTRICA
INSTITUTO NACIONAL DE METROLOGÍA
INSTITUTO NACIONAL DE SALUD
METROLOGÍA Y SERVICIOS DE ENSAYOS DE APTITUD (MYSA S.C.)
MINISTERIO DE COMERCIO
ORGANISMO NACIONAL DE ACREDITACIÓN (ONAC)
PROASECAL SAS
PROASEM
SUELOS Y PAVIMENTOS GREGORIO ROJAS

Además de las anteriores, en Consulta Pública el Proyecto se puso a consideración de las siguientes empresas:

ACERÍAS PAZ DEL RÍO S.A.
ACUAVIVA S.A. ESP
AGUAS DE BUGA S.A. ESP
ALIMENTOS CÁRNICOS SAS
ALPINA PRODUCTOS ALIMENTICIOS S.A.
ATLANTIC MINERALS AND PRODUCTS CORPORATION
BAVARIA S.A.
C.I INDUSTRIAS HUMCAR SAS
C.I. CONFECCIONES BALALAIKA S.A.
CÁLCULO Y CONSTRUCCIONES S.A.
CARULLA VIVERO S.A.
CENTRO COMERCIAL CHIPICHAPE
CHALLENGER S.A.
CLÍNICA DE OCCIDENTE S.A.
COLOMBIANA DE EXTRUSIÓN S.A.
COMPAÑÍA COLOMBIANA DE CERÁMICA S.A.
CORPORACIÓN DE CIENCIA Y TECNOLOGÍA PARA EL DESARROLLO DE LA INDUSTRIA NAVAL, MARÍTIMA Y FLUVIAL
DEPARTAMENTO DE ANTIOQUIA
ECOPETROL S.A.
EMPRESA COLOMBIANA DE SOPLADO E INYECCIÓN S.A.
ESCOBAR & MARTÍNEZ S.A.
ETERNA S.A.

FERTIABONOS S.A.
FRIGORÍFICO GUADALUPE S.A.
FUNDACIÓN SALAMANDRA
GASEOSAS POSADA TOBÓN S.A.
GLOBAL PLASTIK S.A.
HOSPITAL SAN VICENTE ESE DE MONTENEGRO
INDUSTRIA DE ALIMENTOS ZENÚ SAS
INDUSTRIAL DE TINTAS LTDA. (INDUSTINTAS)
INGENIERÍA DESARROLLO Y TECNOLOGÍA (IDT) LTDA.
INMOBILIARIA LLERAS E.U.
INTRAMAR SHIPPING S.A.
LEONISA S.A.
MATPEL DE COLOMBIA S.A.
MATRICES TROQUELES Y MOLDES CÍA. LTDA.
OFFIMÓNACO SAS
ORGANISMO NACIONAL DE ACREDITACIÓN DE COLOMBIA
PROFESIONALES CONTABLES EN ASESORÍA EMPRESARIAL Y DE INGENIERÍA LTDA. - (PROASEM) S.A.
PROFESIONALES DE LA SALUD S.A.
RENTASISTEMAS S.A.
SERVICIO OCCIDENTAL DE SALUD S.A.
SERVICIOS INDUSTRIALES TÉCNICOS S.A.
SERVIREDES LTDA.
SHELL COLOMBIA S.A.
SIDERURGICA DE CALDAS SAS.
SIEMENS S.A.
SIKA COLOMBIA S.A.
STARTECH LTDA.
SYNTOFARMA S.A.
TECNOLOGÍA EMPRESARIAL DE ALIMENTOS S.A.
UNIVERSIDAD DE LOS ANDES
UNIVERSIDAD NACIONAL DE COLOMBIA

ICONTEC cuenta con un Centro de Información que pone a disposición de los interesados normas internacionales, regionales y nacionales y otros documentos relacionados.

DIRECCIÓN DE NORMALIZACIÓN

CONTENIDO

Página

Página

Página

Página

0. INTRODUCCIÓN

0.1 LOS PROPÓSITOS DE LOS ENSAYOS DE APTITUD

El ensayo de aptitud incluye el uso de comparaciones interlaboratorio para determinar el desempeño de los participantes (que pueden ser laboratorios, organismos de inspección o individuos) para ensayos o mediciones específicas y para monitorear su desempeño continuo. Existen varios propósitos habituales para el ensayo de aptitud, según se describe en la Introducción a la NTC-ISO/IEC 17043. Entre estos se encuentran la evaluación del desempeño del laboratorio, la identificación de problemas en los laboratorios, el establecimiento de la eficacia y comparabilidad de métodos de ensayo o medición, el brindar confianza adicional a los clientes del laboratorio, la validación de quejas relacionadas con incertidumbre y la educación de los laboratorios participantes. El diseño estadístico y las técnicas analíticas aplicadas deben ser apropiadas para los propósitos establecidos.

0.2 FUNDAMENTOS PARA LA CALIFICACIÓN EN PROGRAMAS DE ENSAYO DE APTITUD

Se encuentra disponible y en uso una variedad de estrategias de calificación para los ensayos de aptitud. Aunque los cálculos detallados difieren, la mayoría de programas de ensayos de aptitud comparan la desviación de los participantes de un valor asignado con un criterio numérico que se emplea para decidir si la desviación es causa de preocupación o no. Por consiguiente, las estrategias empleadas para asignación de valores y selección de un criterio para valoración de las desviaciones del participante son muy importantes. En especial, resulta importante considerar si el valor y criterio asignados para evaluar las desviaciones deberían ser independientes de los resultados de los participantes o deberían derivarse de los resultados presentados. En la presente norma, se presentan ambas estrategias. No obstante, se llama la atención hacia el análisis que aparece en los numerales 7 y 8 sobre las ventajas y desventajas de seleccionar valores o criterios asignados para evaluar desviaciones que no se derivan de los resultados de los participantes. En general, se apreciará que la selección de valores asignados y criterios de evaluación independientemente de los resultados de los participantes ofrece ventajas. En especial, éste es el caso del criterio empleado para evaluar desviaciones del valor asignado (tal como la desviación estándar para la evaluación de la aptitud o una previsión para error de medición) para lo cual resulta muy útil una opción coherente con base en un uso final particular de los resultados de la medición.

0.3 LAS NORMAS NTC-ISO 13528 Y NTC-ISO/IEC 17043

La NTC-ISO 13528 apoya la implementación de la NTC-ISO/IEC 17043 en especial, en relación con los requisitos del diseño estadístico, la validación de ítems de ensayo de aptitud, la revisión de los resultados y la presentación de informes con resumen de estadísticas. El Anexo B de la NTC-ISO/IEC 17043 describe brevemente los métodos estadísticos generalmente empleados en programas de ensayos de aptitud. Esta norma tiene como intención ser complementaria de la NTC-ISO/IEC 17043 y ofrece orientación detallada que hace falta en dicho documento sobre métodos estadísticos particulares para ensayos de aptitud.

La definición de ensayo de aptitud de la NTC-ISO/IEC 17043 se repite en la NTC-ISO 13528, con las Notas que describen diferentes tipos de ensayos de aptitud y la gama de diseños que se pueden emplear. Esta norma no puede cubrir de manera específica todos los propósitos, diseños, matrices y mensurandos. Se pretende que las técnicas presentadas en la NTC-ISO 13528 tengan aplicación generalizada, en especial para programas de ensayos de aptitud recién establecidos. Se espera que las técnicas estadísticas empleadas para un programa particular de ensayo de aptitud evolucionen a medida que el esquema madure, y que los puntajes, criterios de evaluación y técnicas gráficas se refinen para que sirvan de mejor manera a las necesidades específicas de un grupo objeto de participantes, organismos de acreditación y autoridades reglamentarias.

La NTC-ISO 13528 incorpora publicaciones de orientaciones para ensayos de aptitud de laboratorios analíticos químicos [32], pero adicionalmente incluye un mayor espectro de procedimientos que permiten el uso de métodos de medición válidos e identificaciones cualitativas. Esta actualización de la ISO 13528:2005 contiene la mayoría de métodos estadísticos y orientación de la primera edición, ampliados como sea necesario de acuerdo con los documentos previamente referenciados y el alcance ampliado de la NTC-ISO/IEC 17043. La NTC-ISO/IEC 17043 incluye ensayos de aptitud para individuos y organismos de inspección, y el Anexo B, incluye consideraciones para resultados cualitativos.

La presente norma incluye técnicas estadísticas que son coherentes con otras normas internacionales, en especial aquellas del Comité Técnico TC69 SC6, correspondientes a la serie de Normas ISO 5725 (NTC 3529) sobre *Exactitud: veracidad y precisión*. Las técnicas también tienen como propósito reflejar otras normas internacionales, cuando sea apropiado, y se busca que sean coherentes con la Guía 98-3 ISO/IEC (GUM) y la GTC ISO/IEC 99 (VIM).

0.4 EXPERIENCIA ESTADÍSTICA

La NTC ISO/IEC 17043 exige que, a fin de ser competente, un proveedor de ensayos de aptitud debe tener acceso a experiencia estadística y autorizar a personal específico para que realice análisis estadístico. Ni la NTC-ISO/IEC 17043 ni la presente norma pueden dar más especificaciones sobre cuál es la experiencia necesaria. Para algunas aplicaciones, resulta útil un grado avanzado en experiencia estadística; pero, por lo general, individuos con experiencia técnica en otras áreas, que estén familiarizados con conceptos estadísticos básicos y que tengan experiencia o formación en las técnicas comunes aplicables al análisis de datos de programas de ensayo de aptitud, pueden satisfacer las necesidades de experiencia. Si un individuo está a cargo del diseño y/o análisis estadístico, es muy importante que esta persona tenga experiencia en comparaciones interlaboratorio, incluso si dicha persona tiene un grado avanzado en estadística. Con frecuencia, la formación estadística convencional no incluye ejercicios con comparaciones interlaboratorio y las únicas causas del error en la medición que ocurre en los ensayos de aptitud pueden parecer muy confusas. La orientación de esta norma no puede proveer toda la experiencia necesaria para considerar todas las aplicaciones y no puede remplazar la experiencia obtenida al trabajar con comparaciones interlaboratorio.

0.5 SOFTWARE DE ORDENADOR

Los programas informáticos que se requieren para el análisis estadístico de los datos de ensayos de aptitud, pueden variar en gran medida, ya que van desde la aritmética simple de una hoja de cálculo para los pequeños programas de ensayos de aptitud usando valores de referencia conocidos, hasta software estadístico sofisticado, empleado para los métodos estadísticos que dependen de cálculos iterativos u otros métodos numéricos avanzados. La mayoría de las técnicas de la presente norma pueden llevarse a cabo mediante aplicaciones de hojas de cálculo convencionales, tal vez con rutinas ajustadas para un esquema o análisis

particular. Algunas técnicas requerirán aplicaciones de ordenador que se encuentran disponibles de forma gratuita (en el momento de publicación de esta norma). En todos los casos, los usuarios deberían comprobar la exactitud de sus cálculos, en especial cuando el usuario ha ingresado rutinas especiales. Sin embargo, incluso cuando las técnicas de la presente norma resulten apropiadas y se implementen de forma correcta mediante aplicaciones de ordenador adecuadas, no se pueden aplicar sin la atención de un individuo con experiencia técnica y estadística que sea suficiente para identificar e investigar anomalías que puedan ocurrir en cualquier ronda de ensayos de aptitud.

MÉTODOS ESTADÍSTICOS PARA USO EN ENSAYOS DE APTITUD POR COMPARACIÓN INTERLABORATORIO

1. OBJETO Y CAMPO DE APLICACIÓN

La presente norma ofrece descripciones detalladas de métodos estadísticos para que proveedores de ensayos de aptitud los empleen, para diseñar programas de ensayos de aptitud y analizar los datos obtenidos de dichos programas. Esta norma brinda recomendaciones sobre la interpretación de datos de ensayos de aptitud que hacen los participantes de tales esquemas y los organismos de acreditación.

Los procedimientos de esta norma pueden aplicarse para demostrar que los resultados de la medición obtenidos por laboratorios, organismos de inspección e individuos cumplen criterios especificados para desempeño aceptable.

Esta norma es aplicable para los ensayos de aptitud donde los resultados reportados son tanto para mediciones cuantitativas, u observaciones cualitativas sobre ítems de ensayo.

NOTA Los procedimientos de esta norma también se pueden aplicar en la evaluación de la opinión de expertos, donde las opiniones o juicios se reportan de forma que se pueden comparar objetivamente con un valor de referencia independiente o una estadística de consenso. Por ejemplo, cuando se clasifican los ítems de ensayo de aptitud en categorías conocidas por inspección -o cuando se determina por inspección si los ítems de ensayo de aptitud provienen, o no, de la misma fuente original - y se comparan objetivamente los resultados de la clasificación, se pueden aplicar las disposiciones de esta norma relacionadas con las propiedades nominales (cualitativas).

2. REFERENCIAS NORMATIVAS

Los siguientes documentos normativos referenciados, en su totalidad o en una parte, son indispensables para la aplicación de este documento normativo. Para referencias fechadas, se aplica únicamente la edición citada. Para referencias no fechadas, se aplica la última edición del documento normativo referenciado (incluida cualquier corrección).

NTC 2062-1, Estadística. Vocabulario y símbolos. Parte 1. Términos estadísticos generales y términos utilizados en el cálculo de probabilidades. (ISO 3534-1).

NTC 2062-2, Estadística. Vocabulario y símbolos. Parte 2. Estadística aplicada (ISO 3534-2).

NTC 3529-1, Exactitud (veracidad y precisión) de los métodos de medición y de los resultados. Parte 1: principios generales y definiciones. (ISO 5725-1).

NTC-ISO/IEC 17043, Evaluación de la conformidad. Requisitos generales para los ensayos de aptitud. (ISO/IEC 17043).

GTC-ISO 30, Términos y definiciones usados en relación con los materiales de referencia (ISO GUIDE 30).

GTC-ISO/IEC 99, Vocabulario internacional de metrología (ISO/IEC Guide 99),

3. TÉRMINOS Y DEFINICIONES

Para los propósitos de este documento, se aplican los términos y definiciones presentados en las normas NTC 2062-1 (ISO 3534-1), NTC 2062-2 (ISO 3534-2), NTC 3529-1 (ISO 5725-1), NTC-ISO/IEC 17043, GTC ISO/IEC 99, GTC ISO 30 y los siguientes. En caso de que haya diferencias entre estas referencias sobre el uso de los términos, se aplican las definiciones de la NTC 2062, Partes 1-2. En el Anexo A hay una lista de símbolos matemáticos.

3.1 comparación interlaboratorio (*interlaboratory comparison*). Organización, realización y evaluación de mediciones o ensayos sobre el mismo ítem o ítems similares por dos ó más laboratorios, de acuerdo con condiciones predeterminadas.

3.2 ensayo de aptitud (*proficiency testing*). Evaluación del desempeño de los participantes con respecto a criterios previamente establecidos por medio de comparaciones interlaboratorio.

NOTA 1 Para los propósitos de esta norma el término "ensayo de aptitud" se interpreta en su sentido más amplio e incluye:

- programa cuantitativo - cuando el objetivo es cuantificar uno ó más mensurandos por cada ítem de ensayo de aptitud;
- programa cualitativo - cuando el objetivo es identificar o describir una ó más características cualitativas del ítem de ensayo de aptitud;
- programa secuencial - donde uno ó más ítems de ensayo de aptitud se distribuyen de forma secuencial para ensayo o medición y regresan al proveedor de ensayos de aptitud a intervalos;
- programa simultáneo - donde los ítems de ensayo de aptitud se distribuyen de forma simultánea para ensayo o medición dentro de un período de tiempo definido;
- ejercicio de única ocasión - cuando los ítems de ensayo de aptitud se proveen en una única ocasión;
- programa continuo - donde los ítems del ensayo de aptitud se presentan en intervalos regulares;
- muestreo - donde se toman muestras para análisis posterior y el propósito del programa de ensayos de aptitud incluye la evaluación de la ejecución del muestreo, e
- interpretación de datos - donde se suministran conjuntos de datos u otra información y se procesa la información para proporcionar una interpretación (u otro resultado)

3.3 valor asignado (*assigned value).* Valor que se atribuye a una propiedad particular de un ítem de ensayo de aptitud.

3.4 desviación estándar para evaluación de la aptitud (*standard deviation for proficiency assessment*). Medida de la dispersión utilizada en la evaluación de resultados de ensayos de aptitud.

NOTA 1 Ésta puede interpretarse como la desviación estándar de la población de resultados provenientes de una población hipotética de laboratorios que se desempeñan exactamente de conformidad con los requisitos.

NOTA 2 La desviación estándar para evaluación de aptitud se aplica sólo a resultados en escala de intervalo y relación.

NOTA 3 No todos los programas de ensayos de aptitud evalúan el desempeño con base en la dispersión de los resultados.

[FUENTE: NTC-ISO/IEC 17043, *modificada* - En la definición se ha eliminado ", con base en la información disponible". La Nota 1 ha agregada, y las Notas 2 y 3 editadas levemente].

3.5 error de medición (*measurement error*). Valor medido de la magnitud menos un valor de de referencia.

[FUENTE: GTC-ISO/IEC 99, *modificada* - Se han eliminado las notas.]

3.6 error máximo permitido (*maximum permissible error*). Valor extremo del error de medición, con respecto a un valor cuantitativo de referencia conocido, permitido por especificaciones o regulaciones para una medición determinada, un instrumento de medición o sistema de medición.

[FUENTE: GTC-ISO/IEC 99, *modificada* - Se han eliminado las notas.]

3.7 puntaje *z* (*z score*). Medida normalizada de desempeño, calculada mediante el uso del resultado del participante, el valor asignado y la desviación estándar para evaluación de la aptitud.

NOTA Una variación común en el puntaje z, algunas veces denotado como z' (que comúnmente se pronuncia como z-prima) se determina combinando la incertidumbre del valor asignado con la desviación estándar para la evaluación de aptitud antes de calcular el puntaje z.

3.8 puntaje zeta (*zeta score*). Medida normalizada de desempeño, calculada mediante el uso del resultado del participante, el valor asignado y las incertidumbres estándar combinadas para el resultado y el valor asignado.

3.9 proporción de puntaje límite permitido (*proportion of allowed limit score*). Medida normalizada de desempeño, calculada empleando el resultado del participante, el valor asignado y el criterio para el error de la medición en un ensayo de aptitud.

NOTA Para resultados únicos, se puede expresar el desempeño como la desviación del valor asignado (*D* ó *D %*).

3.10 señal de acción (*action signal*). Indicación de una necesidad de acción que surge de un resultado de ensayo de aptitud.

EJEMPLO Un puntaje z por encima de 2 se toma convencionalmente como una indicación de la necesidad de investigar las causas posibles; Un puntaje z por encima de 3, es convencionalmente tomado como señal de acción que indica la necesidad de acción correctiva.

3.11 valor de consenso (*consensus value*). Valor derivado de un conjunto de resultados en una comparación interlaboratorio.

NOTA La frase "valor de consenso" se emplea típicamente para describir estimados de ubicación y dispersión derivados de los resultados de los participantes en una ronda de ensayos de aptitud; pero también se puede utilizar para referirse a los valores derivados de los resultados de un subconjunto especificado de tales resultados o, por ejemplo, de una cantidad de laboratorios expertos.

3.12 valor atípico (*outlier*). Miembro de un conjunto de valores que es inconsistetente con otros miembros de ese grupo.

NOTA 1 Un valor atípico puede surgir al azar de la población esperada, originarse de una población diferente o ser el resultado de un registro incorrecto u otra equivocación.

NOTA 2 Muchos esquemas emplean el término valor atípico para designar un resultado que genera una señal de acción. Pero éste no es el uso previsto del término. Si bien, por lo general, los valores atípicos generan señales de acción, es posible tener señales de acción de resultados que no son valores atípicos.

[FUENTE: ISO 5725-1, 1994, modificada - Se han eliminado las notas a la entrada.]

3.13 participante (*participant*). Laboratorio, organización o individuo que recibe ítems de ensayo de aptitud y entrega resultados para revisión por el proveedor de ensayo de aptitud.

3.14 ítem de ensayo de aptitud (*proficiency test ítem*). Muestra, producto, artefacto, material de referencia, pieza de equipo, patrón de medición, conjunto de datos u otra información empleada para evaluar el desempeño del participante en el ensayo de aptitud.

NOTA 1 En la mayoría de casos, los ítems de ensayo de aptitud cumplen la definición de “material de referencia” de la Guía 30 de ISO (3.17).

3.15 proveedor del ensayo de aptitud (*proficiency testing provider*). Organización que es responsable de todas las tareas relacionadas con el desarrollo y operación de un programa de ensayos de aptitud.

3.16 programa de ensayos de aptitud (*proficiency testing scheme*). Ensayo de aptitud diseñado y operado en una o más rondas para un área especificada de ensayo, medición, calibración o inspección.

NOTA 1 Un programa de ensayo de aptitud podría cubrir un tipo particular de ensayo, calibración, inspección o una cantidad de ensayos, calibraciones o inspecciones en ítems de ensayo de aptitud.

3.17 material de referencia MR (*reference material RM*). Material, suficientemente homogéneo y estable con respecto a una o más propiedades especificadas, sobre el cual se ha establecido que se ajusta a su uso previsto en un proceso de medición.

NOTA 1 MR es un término genérico.

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

NOTA 3 Entre sus usos puede estar la calibración de un sistema de medición, evaluación de un procedimiento de medición, asignación de valores a otros materiales y control de calidad.

[FUENTE: Guía ISO 30:2015, modificada - Se ha eliminado la Nota 4]

3.18 material de referencia certificado, MRC (*certified reference material, CRM*). Material de referencia (MR) caracterizado por un procedimiento metrológicamente válido para una o más propiedades especificadas, acompañado de un certificado de MR que presenta el valor de la propiedad especificada, su incertidumbre asociada y una declaración de trazabilidad metrológica.

NOTA 1 El concepto de valor incluye una propiedad nominal o un atributo cualitativo tal como identidad o secuencia. Las incertidumbres de tales atributos pueden expresarse como probabilidades o niveles de confianza.

[FUENTE: Guía ISO 30:2015, modificada - Se han eliminado las Notas 2,3 y 4]

4. PRINCIPIOS GENERALES

4.1 REQUISITOS GENERALES PARA MÉTODOS ESTADÍSTICOS

4.1.1 Los métodos estadísticos empleados deben ajustarse al propósito y ser estadísticamente válidos. Cualquier suposición estadística en la que se basen los métodos o el diseño debe establecerse en el diseño o en una descripción escrita del programa de ensayos de aptitud; y se debe demostrar que estas suposiciones son razonables.

NOTA Un método estadísticamente válido cuenta con una base teórica confiable, tiene desempeño conocido bajo las condiciones de uso esperadas y se basa en suposiciones o condiciones que se pueden demostrar para aplicar los datos de modo suficientemente bien para el propósito disponible.

4.1.2 El diseño estadístico y las técnicas de análisis de datos deben ser coherentes con los objetivos establecidos para el programa de ensayos de aptitud.

4.1.3 El proveedor del ensayo de aptitud debe brindar a los participantes una descripción de los métodos de cálculo empleados, una explicación de la interpretación general de los resultados, y una declaración de cualquier limitación relacionada con la interpretación. Esto debe aparecer ya sea en cada informe de cada ronda del programa de ensayos de aptitud o en un resumen por separado de los procedimientos disponibles para los participantes.

4.1.4 El proveedor del ensayo de aptitud debe garantizar que todo el software esté validado de forma adecuada.

4.2 MODELO BÁSICO

4.2.1 Para resultados cuantitativos en programas de ensayos de aptitud donde se informa un resultado único para un ítem de ensayo de aptitud determinado, el modelo básico se presenta en la ecuación (1).

$$x_i = \text{«} + £_i \qquad (1)$$

en donde

x_i = resultado del ensayo de aptitud del participante *i*

μ = valor verdadero para el mensurando

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

NOTA 1 Los modelos comunes para ε incluyen: la distribución normal $\varepsilon_i \sim N(0, \sigma^2)$ con media 0 y varianza ya sea constante o diferente para cada laboratorio; o, de forma más común, una distribución “normal contaminada por valor atípico" que consta de 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 puntajes *z* y σ_{pt} es que en una población “idealizada” de laboratorios competentes, la desviación estándar interlaboratorio sería σ_{pt} ó menor.

NOTA 3 Este modelo difiere del modelo básico de la NTC 3529 (ISO 5725) en que no incluye el término B_i correspondiente al sesgo del laboratorio. Esto se debe a que los términos sesgo de laboratorio y error residual no pueden diferenciarse cuando sólo se reporta una observación. Cuando se consideran los resultados de un participante en varias rondas o ítems de ensayo, no obstante, puede ser útil incluir un término separado para sesgo de laboratorio.

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

4.3 ENFOQUES GENERALES PARA LA EVALUACIÓN DEL DESEMPEÑO

4.3.1 Existen tres enfoques generales para evaluar el desempeño en un programa de ensayos de aptitud. Estos enfoques se emplean para cumplir diferentes propósitos para el programa de ensayos de aptitud. A continuación se enuncian los enfoques:

a) desempeño evaluado por comparación con criterios derivados externamente;

b) desempeño evaluado por comparación con otros participantes;

c) desempeño evaluado por comparación con incertidumbre de la medición declarada;

4.3.2 Los enfoques generales se pueden aplicar de diferente manera para determinar el valor asignado y para determinar los criterios para la evaluación del desempeño. Por ejemplo, cuando el valor asignado es la media robusta de los resultados de los participantes y la evaluación del desempeño se deriva de σ_{pt} ó $¢_E$, donde $¢_E$ es una previsión predefinida para error de medición y $\sigma_{pt} = ¢_E /3$. De forma similar, en algunas situaciones, el valor asignado puede ser un valor de referencia, pero σ_{pt} puede ser una desviación estándar robusta de los resultados de los participantes. En el enfoque c) empleando la incertidumbre de la medición, el valor asignado es típicamente un valor de referencia apropiado.

5. DIRECTRICES PARA EL DISEÑO ESTADÍSTICO DE PROGRAMAS DE ENSAYOS DE APTITUD

5.1 INTRODUCCIÓN AL DISEÑO ESTADÍSTICO DE PROGRAMAS DE ENSAYOS DE APTITUD

El ensayo de aptitud tiene que ver con la evaluación del desempeño del participante y como tal no trata específicamente el sesgo o la precisión (aunque estos pueden evaluarse con diseños específicos). El desempeño de los participantes se evalúa por medio de la evaluación estadística de sus resultados, con base en las mediciones o interpretaciones que éstos hacen sobre los ítems de ensayo de aptitud. Con frecuencia, el desempeño se expresa en forma de puntajes que permiten la interpretación coherente a través de un rango de mensurandos y puede permitir que se comparen los resultados para diferentes mensurandos sobre una misma base. Por lo general, los indicadores de desempeño se derivan comparando la diferencia entre el resultado reportado de un participante y un valor asignado con una desviación permisible o con un estimado de la incertidumbre de medición de la diferencia. El examen de los indicadores de desempeño en múltiples rondas de un programa de ensayo de aptitud puede proporcionar información sobre si laboratorios individuales muestran evidencia de efectos sistemáticos constantes ("sesgo") o precisión deficiente a largo plazo.

Los numerales de las secciones 5 a 10 ofrecen orientación sobre el diseño de programas cuantitativos de ensayos de aptitud y sobre el tratamiento estadístico de los resultados, incluido el cálculo e interpretación de varios indicadores de desempeño. En el numeral 11 se presentan consideraciones sobre programas cualitativos de ensayos de aptitud (incluidos programas ordinales).

5.2 BASES DE UN DISEÑO ESTADÍSTICO

5.2.1 De acuerdo con la NTC-ISO/IEC 17043, numeral 4.4.4.1, el diseño estadístico debe desarrollarse para cumplir los objetivos del programa de ensayo de aptitud, con base en la naturaleza de los datos (cuantitativos o cualitativos incluidos ordinales y categóricos), las hipótesis estadísticas, la naturaleza de los errores y la cantidad esperada de resultados. Por

consiguiente, los programas de ensayo de aptitud con diferentes objetivos y con diferentes fuentes de error podrían tener diseños diferentes.

A continuación se enuncian consideraciones de diseño para objetivos comunes. Son posibles otros objetivos.

EJEMPLO 1 Para que un programa de ensayos de aptitud compare el resultado de un participante contra un valor de referencia predeterminado y dentro de los límites que se especifican antes de iniciar la ronda, el diseño requiere un método para la obtención de un valor de referencia definido externamente, un método para fijar límites y un método de puntaje;

EJEMPLO 2 Para que un programa de ensayos de aptitud compare el resultado de un participante con resultados combinados de un grupo en la misma ronda y límites que se especifican antes de iniciar la ronda, el diseño necesitará tener en cuenta cómo se determinará el valor asignado a partir de los resultados combinados al igual que los métodos para fijar límites y puntaje;

EJEMPLO 3 Para que un programa de ensayos de aptitud compare el resultado de un participante con resultados combinados de un grupo en la misma ronda y límites determinados por la variabilidad de los resultados de los participantes, el diseño necesitará tener en cuenta el cálculo de un valor asignado y una medición apropiada de la dispersión al igual que el método para el puntaje;

EJEMPLO 4 Para que un programa de ensayos de aptitud compare el resultado de un participante con el valor asignado, empleando la incertidumbre de la medición propia del participante, el diseño necesitará tener en cuenta cómo se obtendrán el valor asignado y su incertidumbre y cómo se emplearán las incertidumbres de la medición de los participantes en el puntaje.

EJEMPLO 5 Para que un programa de ensayos de aptitud con un objetivo de comparar el desempeño de diferentes métodos de medición, el diseño necesitará tener en cuenta las estadísticas de resumen pertinentes y los procedimientos para calcularlas.

5.2.2 Existen varios tipos de datos empleados en el ensayo de aptitud, incluidos los cuantitativos, nominales (categóricos) y ordinales. Entre las variables cuantitativas, algunos resultados podrían estar sobre una escala de intervalo; o una escala relativa, o una escala de relación. Para algunas mediciones sobre una escala cuantitativa, sólo se puede realizar sobre un conjunto discreto y discontinuo de valores (por ejemplo, diluciones secuenciales). No obstante, en muchos casos estos resultados pueden tratarse por técnicas que son aplicables a variables cuantitativas continuas.

NOTA 1 Para valores cuantitativos, una escala de intervalo es una escala en la que los intervalos (diferencias) son significativas pero las relaciones no, tal como la escala de temperatura Celsius. Una escala de relación es una escala en la que tanto los intervalos como las relaciones son significativos, tal como la escala de temperatura Kelvin o las unidades más comunes para longitud.

NOTA 2 Para valores cualitativos, una escala categórica tiene distintos valores para los cuales el ordenamiento no es significativo, tales como los nombres de especies bacterianas. Los valores sobre una escala ordinal tienen un ordenamiento significativo, pero las diferencias no son significativas; por ejemplo, una escala tal como "grande, mediano, pequeño" se puede ordenar pero las diferencias entre los valores no se definen sino en términos del número de los valores que intervienen.

5.2.3 Los programas de ensayos de aptitud se pueden emplear para otros propósitos además de los anteriores, como se analiza en el numeral 0.1 y en la NTC-ISO/IEC 17043. El diseño debe ser apropiado para todos los propósitos establecidos para el programa de ensayos de aptitud particular.

5.3 CONSIDERACIONES PARA LA DISTRIBUCIÓN ESTADÍSTICA DE LOS RESULTADOS

5.3.1 La NTC-ISO/IEC 17043, 4.4.4.2, exige que las técnicas de análisis estadístico sean coherentes con las hipótesis estadísticas para los datos. Las técnicas de análisis más comunes para ensayos de aptitud asumen que un conjunto de resultados de participantes competentes estará distribuido aproximadamente de forma normal o al menos será unimodal y

razonablemente simétrico (después de transformación, si es necesario). Una suposición adicional común es que la distribución de los resultados de mediciones determinadas de forma competente es mezclada (o es "contaminada") con resultados de una población de valores erróneos que pueden generar valores anómalos. Por lo general, la interpretación del puntaje se basa en la suposición de normalidad, pero sólo para el supuesto de que la distribución corresponde a participantes competentes.

5.3.1.1 Por lo general, no es necesario verificar que los resultados están distribuidos de forma normal, pero es importante comprobar la simetría aproximada, por lo menos de forma visual. Si no se puede comprobar la simetría, entonces el proveedor del ensayo de aptitud debería emplear técnicas que sean robustas para la asimetría (véase el Anexo C).

5.3.1.2 Cuando la distribución esperada para el programa de ensayos de aptitud no es suficientemente simétrica (permitiendo la contaminación con valores atípicos), el proveedor del ensayos de aptitud debería seleccionar métodos de análisis de datos que tengan la debida consideración de la asimetría esperada y que sean resistentes a valores atípicos y métodos para determinar el puntaje que también tengan la debida consideración de la distribución esperada para los resultados de participantes competentes. Lo cual podría incluir:

- transformación para brindar simetría aproximada;

- métodos de estimación que sean resistentes a la asimetría;

- métodos de estimación que incorporen las suposiciones de distribución apropiadas (por ejemplo, ajuste de máxima verosimilitud con suposiciones de distribución adecuadas y, de ser necesario, rechazo de valores atípicos)

EJEMPLO 1 Los resultados que se basan en la dilución, tales como recuentos microbiológicos cuantitativos o para técnicas de inmunoensayo, con frecuencia se distribuyen de acuerdo con la distribución logarítmica normal, y por lo tanto, una transformación logarítmica puede ser apropiada como el primer paso del análisis.

EJEMPLO 2 Los recuentos de pequeñas cantidades de partículas pueden distribuirse de acuerdo con una distribución Poisson y, por consiguiente, los criterios para la evaluación del desempeño pueden determinarse empleando una tabla de probabilidades de Poisson, con base en el recuento promedio para el grupo de participantes.

5.3.1.3 En algunas áreas de calibración, los resultados de los participantes pueden seguir distribuciones estadísticas que se describen en el procedimiento de medición (por ejemplo, exponenciales o en forma de onda). Distribuciones definidas se deberían considerar en cualquier protocolo de evaluación.

5.3.2 De acuerdo con la NTC-ISO/IEC 17043, numeral 4.4.4.2, el proveedor de ensayos de aptitud debe establecer la base para cualquier suposición (hipótesis) estadística y demostrar que las suposiciones son razonables. Esta demostración puede basarse en, por ejemplo, los datos observados, los resultados de rondas previas del programa de ensayo de aptitud, o la literatura técnica.

NOTA La demostración de lo razonable de una suposición de distribución es menos rigurosa que la demostración de la validez de dicha suposición.

5.4 CONSIDERACIONES PARA CANTIDADES PEQUEÑAS DE PARTICIPANTES

5.4.1 El diseño estadístico de un programa de ensayos de aptitud debe tener en cuenta el número mínimo de participantes que se requieren para cumplir los objetivos del diseño y establecer enfoques alternativos que se emplearán si no se logra la cantidad mínima (NTC-ISO/IEC 17043, numeral 4.4.4.3 b)). Los métodos estadísticos que son apropiados para

grandes cantidades de participantes puede que no sean apropiados con cantidades limitadas de participantes. Causa preocupación el hecho de que los estadísticos determinados a partir de resultados de pequeñas cantidades de participantes no sean suficientemente confiables por lo que se podría evaluar un participante contra un grupo de comparación inapropiado.

NOTA El informe técnico IUPAC/CITAC: *Selección y uso de programas de ensayo de aptitud para una cantidad limitada de participantes* [24] ofrece orientación útil para programas de ensayos de aptitud donde hay pocos participantes. En resumen, el informe IUPAC/CITAC recomienda que el valor asignado debería basarse en mediciones independientes confiables; por ejemplo, usando un material de referencia certificado, de asignación independiente por un instituto de calibración o metrología nacional o mediante preparación gravimétrica. Además, el informe establece que es posible que la desviación estándar para la evaluación del desempeño no se base en la dispersión observada entre los resultados de los participantes para una ronda única de un esquema de ensayo de desempeño.

5.4.2 La cantidad mínima de participantes necesarios para los diversos métodos estadísticos dependerá de una variedad de situaciones:

- los métodos estadísticos empleados; por ejemplo, la robustez del método particular o la estrategia de remoción de valores atípicos seleccionados;

- la experiencia de los participantes con el programa de ensayos de aptitud particular;

- la experiencia del proveedor de ensayos de aptitud con la matriz, mensurando, métodos y grupo de participantes;

- si la intención es determinar el valor asignado o la desviación estándar (o ambos).

En el Anexo D.1 se presenta orientación adicional sobre técnicas para el manejo de una cantidad pequeña de participantes.

5.5 DIRECTRICES PARA SELECCIONAR EL FORMATO DEL INFORME

5.5.1 Un requisito de la NTC-ISO/IEC 17043, numeral 4.6.1.2 es que los proveedores de ensayo de aptitud instruyan a los participantes para realizar mediciones y reportar resultados sobre ítems de ensayo de aptitud de la misma forma que para la mayoría de mediciones que se realizan de forma rutinaria, excepto en circunstancias especiales.

Este requisito puede, en algunas situaciones, dificultar el obtener una evaluación exacta de la precisión y veracidad de los participantes o la competencia con un procedimiento de medición. El proveedor de ensayos de aptitud debería adoptar un formato de presentación de informes coherente para el programa de ensayo de aptitud pero debería, en la medida de lo posible, emplear unidades conocidas por la mayoría de participantes y seleccionar un formato de informe que minimice los errores de transcripción y otros. Esto puede incluir la advertencia automatizada de unidades inapropiadas cuando se sabe que los participantes realizan informes de forma rutinaria en unidades diferentes a las requeridas por el programa.

NOTA 1 Para algunos programas de ensayo de aptitud, un objetivo es evaluar la capacidad de un participante para seguir un método normalizado, que podría incluir el uso de una unidad de medición particular o una cantidad de dígitos significativos.

NOTA 2 Los errores de transcripción en la recopilación de resultados por el proveedor de ensayos de aptitud pueden reducirse o eliminarse de forma sustancial mediante el uso de sistemas de presentación de informes electrónicos que permitan a los participantes ingresar sus propios datos directamente.

5.5.2 Si un programa de ensayo de aptitud requiere mediciones replicadas en ítems de ensayo de aptitud, se debería pedir al participante reportar todos los valores replicados. Esto puede ocurrir, por ejemplo, si un objetivo es evaluar la precisión de un participante en ítems de

ensayo de aptitud replicados conocidos o cuando un procedimiento de medición exige la presentación independiente de informes de observaciones múltiples. En estas situaciones, el proveedor de ensayos de aptitud puede también solicitar al participante el valor promedio (u otro estimado de ubicación) y la incertidumbre como ayuda para el análisis de datos del programa de ensayos de aptitud

5.5.3 Cuando la práctica convencional en la redacción de informes es reportar resultados como "menor que" o "mayor que" un límite (tal como un nivel de calibración o un límite de cuantificación) y donde se requieren resultados numéricos para asignación de puntajes, el proveedor de ensayos de aptitud debe determinar cómo se procesarán los resultados.

5.5.3.1 El proveedor de ensayos de aptitud debería adoptar un tratamiento de datos validados y procedimientos de asignación de puntajes que se acomoden a datos censurados (véase el Anexo E.1) o exigir a los participantes reportar el valor medido del resultado ya sea en lugar del, o adicionalmente al, valor convencional.

NOTA 1 Una opción de procedimiento de asignación de puntaje podría ser no dar indicador (puntaje) a dichos datos.

NOTA 2 El pedirle a los participantes que reporten valores numéricos por fuera del rango reportado normalmente (por ejemplo, por debajo del límite de cuantificación del participante) permitirá el uso de métodos estadísticos que exijan valores numéricos pero puede originar puntajes que no reflejen el servicio rutinario de los participantes a los clientes.

5.5.3.2 Cuando se emplean estadísticas de consenso, de pronto no sea posible evaluar el desempeño si la cantidad de valores censurados es lo suficientemente grande que la robustez del método se vea afectado por la censura. En circunstancias donde la cantidad de resultados censurados es suficiente para afectar la robustez del método, entonces se deberían evaluar los resultados empleando métodos estadísticos que permitan el cálculo no sesgado en la presencia de datos censurados [21], o no se deberían evaluar los resultados. Cuando exista duda sobre el efecto del procedimiento seleccionado, el proveedor de ensayos de aptitud debería calcular las estadísticas de resumen 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.

5.5.3.3 Cuando se esperan o se han observado resultados censurados tales como declaraciones de "menor que", el diseño del programa de ensayo de aptitud debería incluir disposiciones para la asignación de puntaje y/u otra acción sobre valores censurados reportados por participantes y éstos deberían ser notificados de tales disposiciones.

NOTA En el Anexo E.1 hay un ejemplo de algunos enfoques de análisis para datos censurados. Este ejemplo muestra estadísticas robustas de consenso con tres enfoques diferentes: con los valores censurados retirados, con los valores retenidos pero el signo ‘<’ retirado y con los resultados reemplazados por la mitad del valor límite.

5.5.4 Por lo general, el diseño del programa de ensayos de aptitud determinará la cantidad de dígitos significativos que se van a reportar.

5.5.4.1 Cuando se especifican los números de dígitos significativos a ser reportados, el error de redondeo debería ser despreciable en comparación con la variación esperada entre los participantes.

NOTA En algunas situaciones, la presentación correcta de informes hace parte de la determinación de la competencia del participante y el número de dígitos significativos y lugares decimales puede variar.

5.5.4.2 Cuando el número de dígitos reportados bajo condiciones de medición rutinarias tiene un efecto adverso notorio en el tratamiento de los datos por parte del proveedor de ensayos de aptitud (por ejemplo, cuando los procedimientos de medición exigen la presentación de

informes con una pequeña cantidad de dígitos significativos), el proveedor de ensayos de aptitud puede especificar la cantidad de dígitos que se deben reportar.

EJEMPLO Un procedimiento de medición podría especificar la presentación de informes a 0,1 g, llevando con esto a una gran proporción (> 50 %) de resultados idénticos y a su vez comprometiendo el cálculo de medias robustas y desviaciones estándar robustas. El proveedor de ensayos de aptitud podría entonces exigir a los participantes reportar con dos o tres cifras decimales a fin de obtener estimados suficientemente confiables de ubicación y variación.

5.5.4.3 Si se permite que diferentes participantes reporten resultados empleando diferentes números de dígitos significativos, el proveedor de ensayos de aptitud debería tener esto en consideración al generar cualquier estadístico de consenso (tal como el valor asignado y la desviación estándar para la evaluación de la aptitud).

6. DIRECTRICES PARA LA REVISIÓN INICIAL DE LOS ÍTEMS DE ENSAYO DE APTITUD Y RESULTADOS

6.1 HOMOGENEIDAD Y ESTABILIDAD DE ÍTEMS DE ENSAYO DE APTITUD

6.1.1 El proveedor de ensayos de aptitud debe garantizar que los lotes de ítems de ensayo de aptitud sean lo suficientemente homogéneos y estables para los propósitos del programa de ensayos de aptitud. El proveedor debe evaluar la homogeneidad y la estabilidad empleando criterios que garanticen que la falta de homogeneidad y la inestabilidad de los ítems de ensayo de aptitud no afecten de forma adversa la evaluación del desempeño. En la evaluación de la homogeneidad y la estabilidad se debería emplear uno de los siguientes enfoques:

a) estudios experimentales como se describe en el Anexo B o métodos experimentales alternativos que ofrezcan un aseguramiento equivalente o superior de homogeneidad y estabilidad;

b) experiencia con el comportamiento de ítems de ensayo de aptitud muy similares a los empleados en rondas previas del programa de ensayo de aptitud, comprobada según sea necesario para la ronda actual;

c) evaluación de datos de participantes en la ronda actual del programa de ensayos de aptitud para tener evidencia de coherencia con rondas previas, y evidenciar el cambio a través del tiempo de los informes o el orden de producción o cualquier dispersión no esperada atribuible a la falta de homogeneidad o inestabilidad.

NOTA 1 Se pueden adoptar estos enfoques sobre una base de caso por caso, empleando técnicas estadísticas apropiadas y justificación técnica. Con frecuencia, el enfoque cambiará durante el tiempo de vida de un programa de ensayos de aptitud; por ejemplo, la experiencia acumulada reduce el requisito inicial de estudio experimental.

NOTA 2 Confiar en la experiencia (como se menciona en el literal b anterior) sólo es razonable en la medida que:

1) El proceso para producir lotes de los ítems de aptitud no cambien de ninguna manera que pudiera causar impacto a la homogeneidad;

2) Los materiales empleados en la producción de los ítems de aptitud no cambien de ninguna manera que pudiera causar impacto a la homogeneidad;

3) No exista "falla" en la homogeneidad identificada ya sea mediante ensayo de homogeneidad o respuestas de los participantes y

4) Se revisen los requisitos de homogeneidad para el material de forma regular, teniendo en cuenta el uso previsto del material en el momento de la revisión, a fin de garantizar que la homogeneidad lograda por el proceso de producción sigue siendo apta para el propósito.

EJEMPLO Si en rondas previas de un programa de ensayos de aptitud se emplearon ítems de aptitud que fueron ensayados y se demostró que eran suficientemente homogéneos y estables y con los mismos participantes de rondas anteriores, entonces si una desviación estándar interlaboratorio no es mayor que la desviación estándar de rondas previas, existe evidencia de homogeneidad y estabilidad suficientes en la ronda actual.

6.1.2 Para programas de ensayo de aptitud en calibración, donde se emplea el mismo artefacto por múltiples participantes, el proveedor del ensayo de aptitud debe asegurar la estabilidad durante la ronda o contar con procedimientos para identificar e considerar- la inestabilidad durante el avance de una ronda del programa de ensayos de aptitud. Esto debería incluir la consideración de las tendencias para ítems y mensurandos de ensayo de aptitud particulares, tales como la deriva. Cuando resulte apropiado, el aseguramiento de la estabilidad debería tener en cuenta los efectos de múltiples envíos del mismo artefacto.

6.1.3 Por lo general, debería comprobarse la homogeneidad y estabilidad de todos los mensurandos (o propiedades). No obstante, cuando se puede demostrar que el comportamiento de un subconjunto de propiedades proporciona una buena indicación de estabilidad y/o homogeneidad para todas las propiedades reportadas en una ronda, se puede limitar la evaluación descrita en el numeral 6.1.1 a dicho subconjunto de propiedades. Los mensurandos que se comprueban deberían ser sensibles a fuentes de falta de homogeneidad o inestabilidad en el procesamiento del ítem de ensayo de aptitud. Algunos casos importantes son:

a) cuando la medición es una proporción, una característica que sea una proporción pequeña, puede ser más difícil de homogeneizar y por lo tanto, ser más sensible a una comprobación de homogeneidad.

b) si un ítem de ensayo de aptitud se calienta durante el procesamiento, entonces se elige un mensurando que sea sensible a cambios en el calentamiento;

c) si una propiedad medida se puede ver afectada por asentamiento, precipitación u otros efectos dependientes del tiempo durante la preparación del ítem de ensayos de aptitud, entonces se debería comprobar esta propiedad durante el llenado.

EJEMPLO En un programa de ensayos de aptitud para el contenido de metales tóxicos al suelo, el contenido metálico medido se ve principalmente afectado por el contenido de humedad. Entonces una comprobación del contenido de humedad constante se puede considerar suficiente para asegurar la adecuada estabilidad de metales tóxicos.

NOTA En el Anexo E.2 se presenta un ejemplo de homogeneidad y comprobaciones de estabilidad, empleando 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, normalmente el valor asignado debería ser el mismo para todos los participantes. No obstante, cuando se permite a los participantes seleccionar su propio método de medición, es posible que un valor asignado único por cada analito o propiedad no sea apropiado para todos los participantes. Esto puede ocurrir, por ejemplo, cuando diferentes métodos de medición dan resultados que no son comparables. En este caso, el proveedor de ensayos de aptitud puede emplear un valor asignado diferente para cada método de medición.

EJEMPLOS

a) los ensayos médicos, donde se sabe que diferentes métodos de medición aprobados responden de forma diferente al mismo material de ensayo y emplean diferentes intervalos de referencia para el diagnóstico;

b) mensurandos definidos de forma operacional, tales como metales tóxicos lixiviables en suelos, para los cuales se encuentran disponibles diferentes métodos estándar y no se espera que se comparen de forma

directa, pero donde el programa de ensayo de aptitud especifica el mensurando sin referencia a un método de ensayo específico.

6.2.2 En el diseño del programa de ensayos de aptitud, se debería considerar la necesidad de valores asignados diferentes para subconjuntos de participantes (por ejemplo, tener disposiciones para la generación de informes de métodos específicos) y también debería considerarse cuando se revisen datos de cada ronda.

6.3 REMOCIÓN DE DATOS EQUIVOCADOS

6.3.1 La NTC-ISO/IEC 17043, literal B.2.5 y el Protocolo Armonizado IUPAC recomiendan eliminar las equivocaciones obvias de un conjunto de datos en una etapa temprana de un análisis, antes de emplear cualquier procedimiento robusto o cualquier prueba a fin de identificar valores atípicos estadísticos. Por lo general, estos resultados deberían ser tratados por separado (por ejemplo, contactando al participante). Es posible corregir algunas equivocaciones, pero esto sólo debería hacerse de acuerdo con una política y un procedimiento aprobado.

NOTA Las equivocaciones obvias, tales como el reporte de resultados en unidades incorrectas o el cambio de resultados entre diferentes ítems de aptitud, ocurren en la mayoría de rondas de ensayo de aptitud y estos resultados sólo perjudican el desempeño posterior de los métodos estadísticos.

6.3.2 Si existe duda sobre si un resultado es una equivocación, se debería retener en el conjunto de datos y someterse a tratamiento posterior, como se describe en los numerales 6.4 a 6.6.

6.4 REVISIÓN VISUAL DE DATOS

6.4.1 Como primer paso de cualquier análisis de datos, el proveedor debería organizar una revisión visual de los datos, realizada por una persona que tenga la experiencia técnica y estadística adecuada. Con esta comprobación se busca confirmar la distribución de resultados esperada e identificar anomalías o fuentes de variabilidad inesperadas. Por ejemplo, una distribución bimodal podría ser evidencia de una población mixta de resultados causados por métodos diferentes, muestras contaminadas o instrucciones con redacción deficiente. En esta situación, se debería resolver el problema antes de proceder con el análisis o la evaluación.

NOTA 1 Un histograma es un procedimiento de revisión útil y ampliamente disponible para buscar una distribución que sea unimodal y simétrica y para identificar valores atípicos no habituales (véase el numeral 10.2). No obstante, los intervalos empleados para combinar los resultados en un histograma son sensibles a los números de resultados y puntos de corte y por tanto pueden ser difíciles de crear. Con frecuencia resulta más útil una gráfica de densidad de kernel para identificar posibles bimodalidades o falta de simetría (véase el numeral 10.3).

NOTA 2 Pueden ser útiles otras técnicas de revisión, tales como la gráfica de distribución acumulada o el diagrama de tallos y hojas. En los Anexo E.3 y E.4 se ilustran algunos métodos gráficos para la revisión de datos.

6.4.2 Cuando no es factible llevar a cabo una revisión visual de todos los conjuntos de datos de interés, debe haber un procedimiento para advertir la variabilidad inesperada en un conjunto de datos. Por ejemplo, mediante la revisión de la incertidumbre del valor asignado en comparación con las rondas previas del programa de ensayo de aptitud.

6.5 MÉTODOS ESTADÍSTICOS ROBUSTOS

6.5.1 Se pueden emplear métodos estadísticos robustos para describir la tendencia central de un conjunto de resultados distribuidos de forma normal, pero sin exigir la identificación de valores específicos como valores atípicos y su exclusión de análisis posteriores. Muchas de las técnicas robustas que se emplean se basan (en el primer paso) en la mediana y el rango central del 50 % de los resultados. Estas son medidas de centramiento y dispersión de los

datos, similares a la media y la desviación estándar. En general, se deberían emplear preferencialmente métodos robustos con respecto a los métodos que eliminan resultados etiquetados como valores atípicos.

NOTA Por lo general, las estrategias que aplican estadísticos clásicos tales como la desviación estándar después de eliminar los valores atípicos, conducen a subestimados de dispersión para datos casi normales. Las estadísticas robustas, usualmente son utilizadas para arrojar estimados no sesgados de dispersión.

6.5.2 La mediana, la desviación absoluta de mediana ajustada (*MADe*) y el IQR normalizado (*nIQR*) son estimadores permitidos. El algoritmo A transforma los datos originales mediante un proceso denominado transformación (*winsorisation*) para ofrecer estimadores alternativos de media y desviación estándar para datos cercanos a una distribución normal y es más útil cuando la proporción esperada de valores atípicos está por debajo de 20 %. Los métodos Q_n y Q (descritos en el Anexo C) para calcular la desviación estándar son especialmente útiles para situaciones donde una gran proporción (> 20 %) de resultados pueden ser discrepantes o cuando los datos no pueden ser confiables al ser revisados por expertos. Otros métodos descritos en el Anexo C también ofrecen buen desempeño cuando la proporción esperada de valores extremos está por encima del 20 % (véase el Anexo D).

NOTA La mediana, el rango intercuartilico y la desviación estándar de la mediana ajustada tienen mayor varianza que la media y la desviación estándar cuando se aplican a datos distribuidos de forma aproximadamente normal. Los estimadores robustos más sofisticados brindan mejor desempeño para datos distribuidos de forma aproximadamente normal, al tiempo que son más resistentes para retener los resultados anómalos comparados con la mediana y el rango intercuartílico

6.5.3 La selección de métodos estadísticos es responsabilidad del proveedor de ensayos de aptitud. La media y la desviación estándar robustas pueden emplearse para varios propósitos de los cuales la evaluación del desempeño es sólo una. También se pueden usar como estadísticas de resumen para diferentes grupos de participantes o para métodos específicos.

NOTA En el Anexo C se presentan detalles de procedimientos robustos. En los Anexos E.3 y E.4 se incluyen ejemplos globales 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 se pueden emplear para apoyar la revisión visual de anomalías o, junto con el rechazo de valores atípicos, con el fin de proporcionar un grado de resistencia a valores extremos cuando se calculan estadísticas de resumen. Cuando se emplean técnicas de detección de valores atípicos, se deberían demostrar las suposiciones correspondientes a la prueba estadística de tal forma que se apliquen, de forma suficiente, para el propósito del programa de ensayo de aptitud; en especial muchas pruebas de valores atípicos suponen normalidad.

NOTA Las normas ISO 16269-4 [10] e ISO 5725-2 (NTC 3529-2) [1] brindan varios procedimientos de identificación de valores atípicos que son aplicables en datos interlaboratorio.

6.6.2 Las estrategias de rechazo de valores atípicos, que se basan en el rechazo de valores atípicos detectados por una prueba con un nivel de confianza alto, seguido de la aplicación de estadísticas simples, tales como la media y la desviación estándar, son permitidas cuando los métodos robustos no son aplicables (véase el numeral 6.5.1). Cuando se emplean estrategias de rechazo de valores atípicos, el proveedor de ensayo de aptitud debe:

a) documentar las pruebas y el nivel de confianza requerido para el rechazo;

b) fijar límites para la proporción de datos rechazados por pruebas de valores atípicos sucesivos, si se emplean;

c) demostrar que los estimados resultantes de ubicación y (si es apropiado) escala tienen un desempeño suficiente (incluida eficiencia y sesgo) para los propósitos del programa de ensayos de aptitud.

NOTA En la NTC 3529-2 se presentan recomendaciones para el nivel de confianza apropiado para rechazo de valores atípicos en estudios interlaboratorio para la determinación de la precisión de métodos de ensayo. En especial, la NTC 3529-2 recomienda rechazo sólo en el nivel del 99 % a menos que exista otra razón poderosa para rechazar un resultado particular.

6.6.3 Cuando el rechazo de valores atípicos hace parte de un procedimiento de manejo de datos y se elimina un resultado como valor atípico, el desempeño del participante aún debería ser evaluado de acuerdo con los criterios empleados para todos los participantes en el programa de ensayo de aptitud.

NOTA 1 Con frecuencia, los valores atípicos que se encuentran entre los valores reportados se identifican empleando la prueba de Grubbs para valores atípicos, como se presenta en la NTC 3529-2. La evaluación en este procedimiento se aplica empleando la desviación estándar de todos los participantes, incluyendo los potenciales valores atípicos. Por consiguiente, debería aplicarse este procedimiento cuando el desempeño de los participantes es coherente con las expectativas de rondas previas y existe una pequeña cantidad de valores atípicos (uno o dos valores atípicos en cada lado de la media). Las tablas convencionales para la prueba de Grubbs asumen una aplicación única para un valor atípico posible (ó 2) en una ubicación definida, no una aplicación secuencial ilimitada. Si se aplican las tablas de Grubbs de forma secuencial, es posible que no se apliquen las probabilidades de error Tipo I para las pruebas.

NOTA 2 Cuando se retornan resultados replicados o se incluyen ítems idénticos de ensayos de aptitud en una ronda de programa de ensayos de aptitud, es común que se emplee la prueba de Cochran para valores atípicos de repetibilidad, también descrito en la NTC 3529-2.

NOTA 3 Los valores atípicos también pueden identificarse mediante técnicas robustas o no paramétricas; por ejemplo, si se calculan una media y desviación estándar robustas, los valores que se desvían de la media robusta en más de 3 veces la desviación estándar robusta podrían identificarse como valores atípicos.

7. DETERMINACIÓN DEL VALOR ASIGNADO Y SU INCERTIDUMBRE ESTÁNDAR

7.1 SELECCIÓN DE MÉTODO PARA DETERMINAR EL VALOR ASIGNADO

7.1.1 En los numerales 7.3 a 7.7 se describen cinco formas de determinar el valor asignado x_{pt}. La selección de estos métodos es responsabilidad del proveedor de ensayos de aptitud.

NOTA Los numerales 7.3 - 7.6 son muy similares a los enfoques empleados para determinar los valores de propiedad de materiales de referencia certificados descritos en la Guía 35 de la ISO [13].

7.1.2 Se pueden utilizar métodos alternos para determinar el valor asignado y su incertidumbre siempre que tengan una base estadística sólida y que el método utilizado esté descrito en el plan documentado para el programa de ensayos de aptitud y se describa por completo a los participantes. Sin importar el método empleado para determinar el valor asignado, es apropiado siempre comprobar la validez del valor asignado para una ronda de un programa de ensayos de aptitud. En el numeral 7.8 se analiza este tema.

7.1.3 En el numeral 11.3 se analizan los enfoques para determinar valores asignados cualitativos.

7.1.4 El método de determinación del valor asignado y su incertidumbre asociada en cada informe para los participantes o se debe describir de forma clara en un protocolo del programa 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 de Medida (Guía ISO/IEC 98-3[14]) brinda orientación sobre la evaluación de incertidumbres en la medición. La Guía 35 de la ISO ofrece orientación sobre la incertidumbre del valor asignado para valores certificados del atributo o propiedad, que se pueden aplicar para muchos diseños de programas de ensayos de aptitud.

7.2.2 En las ecuaciones (2) y (3) se describe un modelo general para el valor asignado y su incertidumbre: El modelo para el valor asignado puede expresarse de la siguiente manera:

$$x_{pt} = x_{char} + ¢_{hom} + ¢_{trans} + ¢_{stab} \qquad (2)$$

en donde

x_{pt} denota el valor asignado;

x_{char} denota el valor de la propiedad obtenido de la caracterización (determinación de valor asignado);

$¢_{hom}$ denota un término de error debido a la diferencia entre ítems de ensayo de aptitud;

$¢_{trans}$ denota un término de error debido a la inestabilidad bajo condiciones de transporte;

$¢_{stab}$ denota un término de error debido a la inestabilidad durante el período de ensayos de aptitud.

El modelo asociado para la incertidumbre del valor asignado puede expresarse de la siguiente manera:

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

en donde

$u(x_{pt})$ denota la incertidumbre estándar del valor asignado;

u_{char} denota la incertidumbre estándar debida a la caracterización;

u_{hom} denota la incertidumbre estándar debida a la diferencia entre ítems de ensayo de aptitud;

u_{trans} denota la incertidumbre estándar debida a la inestabilidad causada por transporte de ítems de ensayo de aptitud;

u_{stab} denota la incertidumbre estándar debida a la inestabilidad durante el período de ensayos de aptitud.

NOTA 1 La covarianza entre fuentes de incertidumbre, o fuentes despreciables, puede conducir a un modelo diferente para aplicaciones específicas. En algunas situaciones, cualquiera de los componentes de la incertidumbre puede ser cero o despreciable.

NOTA 2 Cuando se calcula σ_{pt} como la desviación estándar de los resultados de participantes, los componentes de la incertidumbre debidos a la homogeneidad, transporte e inestabilidad se reflejan en gran medida en la variabilidad de los resultados de los participantes. En este caso, es suficiente la incertidumbre de la caracterización, como se describe en los numerales 7.3 a 7.7.

NOTA 3 Por lo general, se espera que el proveedor de ensayos de aptitud se asegure que los cambios relacionados con la inestabilidad o incurridos en el transporte sean despreciables en comparación con la desviación estándar para evaluación de la aptitud; es decir, que asegure que δ_{trans} y δ_{stab} son despreciables. Cuando se cumple este requisito, u_{stab} y u_{trans} pueden ajustarse a cero.

7.2.3 Puede existir sesgo en el valor asignado el cual no está considerado en la expresión anterior. Este debe ser considerado, siempre que sea posible, en el diseño del programa de ensayos de aptitud. Si existe un ajuste por sesgo en el valor asignado, se debe incluir la incertidumbre de este ajuste en la evaluación de la incertidumbre del valor asignado.

7.3 FORMULACIÓN

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

7.3.1.1 El valor asignado x_{pt} se deriva mediante cálculo a partir de las masas de propiedades utilizadas. Este enfoque es especialmente valioso cuando los ítems individuales del ensayo de aptitud se preparan de este modo y es la proporción de las propiedades lo que se debe determinar.

7.3.1.2 Se debería tener cuidado razonable para asegurar que:

a) el material de base esté efectivamente libre del componente adicionado, o que se conozca con exactitud la proporción de dicho componente en el material de base.

b) los componentes se mezclan de manera homogénea (cuando se requiera);

c) todas las fuentes significativas de error son identificadas (por ejemplo, no siempre se tiene en cuenta que el vidrio absorbe compuestos de mercurio, de modo que la concentración de una solución acuosa de un compuesto de mercurio puede verse alterada por su envase de vidrio);

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

e) el comportamiento de los ítems de ensayo de aptitud que contienen material adicionado, sea similar a las muestras del cliente que se ensayan de forma rutinaria. Por ejemplo, los materiales puros que se agregan a una matriz natural, con frecuencia, se extraen con mayor facilidad que la misma sustancia que está presente de forma natural en el material. Si existe alguna preocupación sobre este evento, el proveedor del ensayo de aptitud debería asegurar la idoneidad de los ítems de ensayo de aptitud para los métodos que se emplearán.

7.3.1.3 Cuando la formulación genera ítems de ensayo de aptitud en la cual la adición presenta enlaces más débiles que en las muestras ensayadas de forma rutinaria, o de forma diferente, puede ser preferible emplear otro enfoque para preparar los ítems de ensayo de aptitud.

7.3.1.4 La determinación del valor asignado mediante formulación es un caso de un enfoque general para la caracterización de materiales de referencia certificados descritos por la Guía ISO 35 donde un laboratorio único determina un valor asignado empleando un método de medición primario. Entre otros usos, un método primario realizado por un único laboratorio puede ser usado para determinar el valor asignado para ensayos de aptitud (véase el numeral 7.5).

7.3.2 Cuando se calcula el valor asignado a partir de la formulación del ítem de ensayos de aptitud, la incertidumbre estándar para la caracterización (u_{char}) se estima mediante la combinación de incertidumbres empleando un modelo apropiado. Por ejemplo, en ensayos de aptitud para mediciones químicas, por lo general las incertidumbres serán aquellas asociadas con mediciones gravimétricas y volumétricas y la pureza de cualquier material empleado en la formulación. La incertidumbre estándar del valor asignado ($u(x_{pt})$) se calcula entonces de acuerdo con la ecuación (3).

7.4 MATERIAL DE REFERENCIA CERTIFICADO

7.4.1 Cuando un ítem de ensayos de aptitud es un material de referencia certificado (MRC), su valor de propiedad certificada x_{MRC} se emplea como el valor asignado x_{pt}.

Las siguientes son las limitaciones de este enfoque:

- puede ser costoso proporcionar a cada participante una unidad de material de referencia certificado;

- con frecuencia, los MRC tienen un procesamiento muy complejo a fin de asegurar la estabilidad a largo plazo, que puede comprometer la conmutabilidad de los ítems de ensayo de aptitud.

- es posible que los participantes conozcan un MRC, haciendo que sea importante ocultar la identidad del ítem de ensayo de aptitud.

7.4.2 Cuando se emplea un material de referencia certificado como el ítem de ensayo de aptitud, la incertidumbre estándar del valor asignado se deriva de la información sobre la incertidumbre del valor de la propiedad declarado en el certificado. La información del certificado debería incluir los componentes de la ecuación (3) y tener un uso previsto apropiado para el fin del programa de ensayos de aptitud.

7.5 RESULTADOS DE UN LABORATORIO

7.5.1 Un laboratorio único puede determinar un valor asignado empleando un método de referencia, tal como, por ejemplo, un método primario. El método de referencia empleado debería describirse y entenderse por completo e ir acompañado de una declaración de incertidumbre completa y trazabilidad metrológica documentada que sea apropiada para el programa de ensayos de aptitud. El método de referencia debería ser conmutable para todos los métodos de medición empleados por los participantes.

7.5.1.1 El valor asignado debería ser el promedio de un estudio diseñado empleando más de un ítem de ensayo de aptitud o condiciones de medición y una cantidad suficiente de replicas en la mediciones.

7.5.1.2 La incertidumbre de la caracterización es el estimado apropiado de la incertidumbre para el método de referencia y las condiciones del estudio diseñado.

7.5.2 El valor asignado x_{pt} del ítem de ensayo de aptitud puede derivarse por un laboratorio único empleando un método de medición adecuado, a partir de una calibración contra los valores de referencia de un material de referencia certificado que tenga una correspondencia cercana. Este enfoque da por sentado que el MRC es conmutable para todos los métodos de medición empleados por los participantes.

7.5.2.1 Esta determinación requiere realizar una serie de ensayos, en un laboratorio, de los ítems de ensayo de aptitud y del MRC, empleando el mismo método de medición y bajo condiciones de repetibilidad.

Cuando:

x_{MRC} es el valor asignado para el MRC

x_{pt} es el valor asignado para el ítem de ensayos de aptitud

d_i es la diferencia entre los resultados promedio para el ítem de ensayos de aptitud y el MRC en las i ésimas muestras

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

entonces,

$$x_{pt} = x_{MRC} + \bar{d} \qquad (4)$$

NOTA x_{MRC} y $\bar{d}$ son independientes excepto en la rara situación que el laboratorio experto también produjera el MRC.

7.5.2.2 La incertidumbre estándar de la caracterización se deriva de la incertidumbre de la medición empleada para la asignación del valor. Este enfoque permite establecer el valor asignado de manera que sea metrológicamente trazable al valor certificado del MRC, con una incertidumbre estándar que se puede calcular a partir de la ecuación (5).

$$U_{char} = \sqrt{u_{MRC}^2 + u_{\bar{d}}^2} \qquad (5)$$

El ejemplo del Anexo E.5 ilustra cómo se puede calcular la incertidumbre requerida en el caso simple cuando se establece el valor asignado de un ítem de ensayo de aptitud por comparación directa con un MRC único.

7.5.3 Cuando se asigna un valor de referencia antes de iniciar una ronda de un programa de ensayos de aptitud secuencial, y se comprueba el valor de referencia posteriormente empleando el mismo sistema de medición, la diferencia entre los valores debe ser inferior a dos veces la incertidumbre de dicha diferencia (es decir, los resultados deben ser metrológicamente compatibles). En dichos casos, el proveedor de ensayos de aptitud puede optar por emplear un promedio de las mediciones como el valor asignado con la incertidumbre apropiada. Si los resultados no son compatibles metrológicamente, el proveedor de ensayos de aptitud debería investigar la razón de la diferencia y tomar medidas apropiadas, incluido 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 LABORATORIOS EXPERTOS

7.6.1 Se pueden determinar valores asignados empleando un estudio de comparación interlaboratorio con laboratorios expertos, como se describe en la Guía ISO 35 para uso de comparaciones interlaboratorio a fin de caracterizar el MRC. En primer lugar se preparan los ítems de ensayo de aptitud y se alistan para distribución a los participantes. Luego seleccionen algunos de estos ítems de ensayo de aptitud de forma aleatoria y un grupo de expertos los analizan empleando un protocolo que especifica las cantidades de ítems de ensayo de aptitud, el número de replicas y cualquier otra condición pertinente. Se requiere que cada laboratorio experto provea una incertidumbre estándar con sus resultados.

7.6.2 Cuando los laboratorios expertos reportan un resultado único y el protocolo de medición no les exige proveer información suficiente de la incertidumbre con los resultados o cuando la evidencia de los resultados reportados u otros sugiere que las incertidumbres reportadas no son suficientemente confiables, normalmente debería obtenerse el valor de consenso con los métodos del numeral 7.7, aplicado al conjunto de resultados de laboratorios expertos. Cuando cada uno de los laboratorios expertos reportan más de un resultado (por ejemplo, incluyendo replicas), el proveedor del programa de ensayos de aptitud debe establecer un método alternativo para determinar el valor asignado y la incertidumbre asociada que es estadísticamente valida (véase el numeral 4.1.1) y permite la posibilidad de valores atípicos u otras desviaciones de la distribución esperada de resultados.

7.6.3 Cuando los laboratorios expertos reportan incertidumbres con los resultados, el estimado de un valor por consenso de resultados constituye un problema complejo y se ha sugerido una amplia variedad de enfoques, incluidos, por ejemplo, promedios ponderados, promedios no ponderados, procedimientos que permiten mayor dispersión y procedimientos que permiten posibles resultados anómalos o erróneos y estimados de incertidumbre [16]. Por consiguiente, el proveedor de ensayos de aptitud debe establecer un procedimiento para el cálculo que:

a) debería incluir comprobaciones de validez de estimados de incertidumbre reportados, por ejemplo, verifican si las incertidumbres reportadas tienen en cuenta completamente la dispersión observada de los resultados;

b) debería emplear un procedimiento de ponderación apropiado para la escala y confiabilidad de las incertidumbres reportadas, el cual puede incluir iguales ponderaciones si las incertidumbres reportadas son similares o de confiabilidad deficiente o desconocida (véase el numeral 7.6.2);

c) debería permitir la posibilidad de que las incertidumbres reportadas no tengan en cuenta plenamente la dispersión observada ("mayor dispersión"), por ejemplo, incluyendo un término adicional para permitir mayor dispersión;

d) debería permitir la posibilidad de valores anómalos para el resultado reportado de la incertidumbre;

e) debería tener una base teórica sólida;

f) debe tener desempeño demostrado (por ejemplo sobre datos de ensayo o en simulaciones) suficiente para los propósitos del programa de ensayos de aptitud.

7.7 VALOR DE CONSENSO DE RESULTADOS DE PARTICIPANTES

7.7.1 Con este enfoque, el valor asignado x_{pt} para el ítem de ensayo de aptitud empleado en una ronda de un programa de ensayos de aptitud es el estimado de localización (por ejemplo, media robusta, mediana o media aritmética) determinado a partir de los resultados reportados por los participantes en la ronda, calculados empleando un procedimiento apropiado de acuerdo con el diseño, como se describe en el Anexo C. Las técnicas descritas en los numerales 6.2 al 6.6 se deberían emplear para confirmar que exista suficiente acuerdo, antes de combinar los resultados.

7.7.1.1 En algunas situaciones, es posible que el proveedor de ensayos de aptitud quiera emplear un subconjunto de participantes determinado como confiable, de acuerdo con algunos criterios predefinidos tales como el estado de acreditación o sobre la base de desempeño anterior. Las técnicas de este el numeral se aplican a aquellas situaciones que incluyen consideraciones del tamaño del grupo.

7.7.1.2 Se pueden emplear otros métodos de cálculo en lugar de los indicados en el Anexo C, siempre y cuando tengan una base estadística sólida y el informe establezca el método que se emplea.

7.7.1.3 Las siguientes son las ventajas de este enfoque:

a) no se requieren mediciones adicionales para obtener el valor asignado;

b) el enfoque puede ser especialmente útil con un mensurando estandarizado, definido de forma operacional, puesto que con frecuencia no existe un método más confiable para obtener resultados equivalentes.

7.7.1.4 Las siguientes son las limitaciones de este enfoque:

a) puede no existir acuerdo suficiente entre los participantes;

b) el valor de consenso puede incluir sesgo desconocido debido al uso general de metodología errónea y este sesgo no se verá reflejado en la incertidumbre estándar del valor asignado;

c) el valor de consenso podría sesgarse debido al efecto del sesgo en los métodos que se emplean para determinar el valor asignado.

d) Puede ser difícil determinar la trazabilidad metrológica del valor de consenso. Si bien el resultado siempre es trazable a los resultados individuales de los laboratorios, sólo se puede hacer una declaración de trazabilidad más extensa cuando el proveedor de ensayos de aptitud cuenta con información completa sobre los patrones de calibración empleados y control de otras condiciones relevantes de los métodos para todos los participantes que contribuyen al valor de consenso

7.7.2 La incertidumbre estándar del valor asignado dependerá del procedimiento empleado. Si se requiere un enfoque general completo, el proveedor del ensayo de aptitud debería considerar el uso de técnicas de re muestreo (*bootstrapping*) para estimar un error estándar del valor asignado. En las Referencias [17,18] se presentan detalles de técnicas de re muestreo.

NOTA En el Anexo E.6 se presenta un ejemplo en el que se emplea una técnica de re muestreo

7.7.3 Cuando el valor asignado se deriva como un promedio robusto calculado empleando procedimientos del Anexo C.2, C.3 ó C.5, la incertidumbre estándar del valor asignado x_{pt} se puede calcular de la siguiente manera:

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

donde s^* es la desviación estándar robusta de los resultados. (Aquí un “resultado” para un participante es el promedio de todas sus mediciones en el ítem de ensayo de aptitud.)

NOTA 1 En este modelo, cuando el valor asignado y la desviación estándar robusta se determinan a partir de resultados de participantes, se puede suponer que la incertidumbre del valor asignado incluye los efectos de la incertidumbre debida a homogeneidad, transporte e inestabilidad.

NOTA 2 El factor de 1,25 se basa en la desviación estándar de la mediana o la eficiencia de la mediana como un estimado de la media, en un conjunto grande de resultados extraídos de una distribución normal. Se aprecia que la eficiencia de los métodos robustos más sofisticados puede ser superior que la de la mediana, justificando un factor de corrección menor de 1,25. No obstante, se ha recomendado este factor debido a que, por lo general, los resultados de los ensayos de aptitud no se distribuyen estrictamente de forma normal y contienen proporciones desconocidas de resultados de distribuciones diferentes ("resultados contaminados"). El factor de 1,25 se considera un estimado conservador (alto), para tener en cuenta la posible contaminación. Los proveedores de ensayos de aptitud también pueden justificar el empleo de un factor menor, o de una ecuación diferente, dependiendo de la experiencia y del procedimiento robusto empleado.

NOTA 3 En el Anexo E.3 se presenta un ejemplo del uso de un valor asignado a partir de los resultados de los participantes.

7.8 COMPARACIÓN DEL VALOR ASIGNADO CON UN VALOR DE REFERENCIA INDEPENDIENTE

7.8.1 Cuando se emplean los métodos descritos en el numeral 7.7 para establecer el valor asignado (x_{pt}), y se cuenta con un estimado independiente y confiable (denotado como x_{ref}), por ejemplo, a partir del conocimiento de la preparación o a partir de un valor de referencia, se debería comparar el valor de consenso x_{pt} con x_{ref}.

Cuando se emplean los métodos descritos en los numerales 7.3 a 7.6 para establecer el valor asignado, el promedio robusto x^* derivado de los resultados de la ronda se debería comparar con el valor asignado, después de cada ronda del programa de ensayos de aptitud.

Calcule la diferencia como $x_{diff} = (x_{ref} - x_{pt})$ (ó $(x^* - x_{pt})$) y la incertidumbre estándar de la diferencia, como:

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

en donde

$u(x_{ref})$ es la incertidumbre del valor de referencia para la comparación, y

$u(x_{pt})$ es la incertidumbre del valor asignado.

NOTA En el Anexo E.7 se incluye un ejemplo de la comparación de un valor de referencia con un valor de consenso.

7.8.2 Si la diferencia es mayor a dos veces su incertidumbre estándar, se debería investigar la razón. A continuación se presentan las posibles razones:

- un sesgo en el método de medición de la referencia;
- un sesgo común en los resultados de los participantes;
- una falla en la determinación de las limitaciones del método de medición cuando se emplea el método de formulación descrito en el numeral 7.3;
- un sesgo en los resultados de los “expertos” cuando se emplean los enfoques de las secciones 7.5 ó 7.6, y
- el valor de la comparación y el valor asignado no son trazables a la misma referencia metrológica.

7.8.3 Dependiendo de la razón de la diferencia, el proveedor del ensayo de aptitud debería decidir si evalua o no los resultados y (para programas de ensayos de aptitud continuos), si corrige el diseño para posteriores programas de ensayos de aptitud. Cuando la diferencia es suficientemente grande para afectar la evaluación del desempeño o sugerir un sesgo importante en los métodos de medición empleados por los participantes, se debería registrar la diferencia en el informe de la ronda. En tales casos, la diferencia se debería considerar en el diseño de programas de ensayos de aptitud futuros.

8. DETERMINACIÓN DE CRITERIOS PARA LA EVALUACIÓN DEL DESEMPEÑO

8.1 ENFOQUES PARA DETERMINAR LOS CRITERIOS DE EVALUACIÓN

8.1.1 El enfoque básico para todos los propósitos, consiste en comparar el resultado de un ítem de ensayo de aptitud (x_i) con un valor asignado (x_{pt}). Para fines de evaluación, se compara la diferencia con un error de medición permitido. Esta comparación se realiza comúnmente por medio de una estadística de desempeño estandarizada (por ej., z, z', ζ, En), como se analiza en las numerales 9.4 a9.7. También puede hacerse mediante comparación de la diferencia con un criterio definido (D ó D% comparada con ¢$_E$) como se analiza en el numeral 9.3. Una alternativa apropiada para la evaluación, consiste en comparar la diferencia con la incertidumbre declarada de los resultados de los participantes combinada con la incertidumbre del valor asignado (E_n y ζ).

8.1.2 Si un requisito reglamentario o una meta de idoneidad para el fin previsto se presenta como una desviación estándar, se puede usar directamente como σ_{pt}. Si el requisito o meta es para un error de medición máximo permitido, se puede dividir el criterio por el límite de acción para obtener σ_{pt}. Un error máximo admisible preestablecido, se puede emplear directamente como ¢$_E$ para utilizarlo con D ó D%. Las ventajas de este enfoque para los esquemas continuos son:

a) los indicadores de desempeño tienen una interpretación coherente en términos de idoneidad para el fin previsto entre una ronda y otra;

b) los indicadores de desempeño no están sujetos a la variación esperada cuando se calcula la dispersión a partir de los resultados reportados.

EJEMPLO Si se especifica un criterio reglamentario como un error máximo permitido y 3,0 es un límite de acción para la evaluación con un puntaje z, entonces, el criterio especificado se divide por 3,0 para determinar σ_{pt}.

8.1.3 Cuando el criterio para la evaluación de desempeño se basa en estadísticas de consenso a partir de la ronda actual o rondas anteriores del programa de ensayos de aptitud, entonces la estadística preferida es un estimado robusto de la desviación estándar de los resultados de los participantes. Por lo general, cuando se emplea este enfoque, resulta más conveniente utilizar un puntaje de desempeño tal como el puntaje z y fijar la desviación estándar para la evaluación de desempeño (σ_{pt}) como el estimado calculado de la desviación estándar.

8.2 POR PERCEPCIÓN DE EXPERTOS

8.2.1 El error máximo permitido o la desviación estándar para la evaluación de aptitud puede fijarse en un valor que corresponda con el nivel de desempeño que una autoridad reglamentaria, un organismo de acreditación o los expertos técnicos del proveedor de ensayo de aptitud crean que es razonable para los participantes.

8.2.2 Un error máximo permitido se puede transformar en una desviación estándar para evaluación de aptitud dividiendo el límite por el número de múltiplos del σ_{pt} que se empleen para definir una señal de acción (o un resultado inaceptable). De manera similar, un σ_{pt} especificado se puede transformar en ¢$_E$.

8.3 POR EXPERIENCIA DE RONDAS PREVIAS DE UN PROGRAMA DE ENSAYO DE APTITUD

8.3.1 La desviación estándar para la evaluación del ensayo de aptitud (σ_{pt}) y el error máximo permitido (δ_E) se pueden determinar por experiencia con otras rondas previas de ensayo de aptitud para el mismo mensurando con valores de propiedad comparables y donde los participantes emplean procedimientos de medición comparables. Este es un enfoque útil cuando no existe acuerdo entre expertos sobre la aptitud para el fin previsto. Las siguientes son las ventajas de este enfoque:

- las evaluaciones estarán basadas en expectativas de desempeño razonables;

- los criterios de evaluación del programa de ensayos de aptitud no variarán de una ronda a otra debido a variación aleatoria o cambios en la población de participantes;

- los criterios de evaluación no variarán entre diferentes proveedores de ensayo de aptitud, cuando hayan dos o más proveedores de ensayos de aptitud aprobados para un área de ensayo o calibración.

8.3.2 La revisión de rondas previas de un programa de ensayos de aptitud debería incluir la consideración del desempeño que pueden alcanzar participantes competentes y que no se vean afectado por nuevos participantes o por la variación aleatoria debida a, por ejemplo, grupos de menor tamaño u otros factores únicos de una ronda particular. Se pueden realizar determinaciones subjetivas mediante el examen de la consistencia de rondas previas u objetivas con promedios o con un modelo de regresión que se ajuste para el valor del mensurando. La ecuación de regresión podría ser una línea recta o podría ser una línea curva [31]. Se deberían considerar las desviaciones estándar y las desviaciones estándar relativas, con una selección basada en cuál es más constante a lo largo del rango apropiado de niveles del mensurando. También se puede obtener el error máximo permitido de esta manera.

8.3.3 Cuando el criterio para la evaluación de desempeño se basa en las estadísticas de consenso de rondas previas de un programa de ensayos de aptitud, se deberían emplear estimados robustos de la desviación estándar.

NOTA 1 El algoritmo S (véase el Anexo C.4) determina una desviación estándar robusta que es aplicable cuando todas las rondas previas a un programa de ensayos de aptitud en consideración tienen la misma desviación estándar esperada o (si se emplean desviaciones relativas para la evaluación) tienen la misma desviación estándar relativa.

NOTA 2 En el Anexo E.8 se presenta un ejemplo de un valor derivado de la experiencia de rondas previas de un programa de ensayo de aptitud.

8.4 MEDIANTE EL USO DE UN MODELO GENERAL

8.4.1 El valor de la desviación estándar para evaluación de aptitud puede derivarse de un modelo general para la reproducibilidad del método de medición. Este método tiene la ventaja de la objetividad y consistencia en los mensurandos y de tener bases empíricas. Dependiendo del modelo usado, este enfoque podría considerarse como un caso especial de un criterio de idoneidad para el fin previsto.

8.4.2 Cualquier desviación estándar esperada, seleccionada mediante un modelo general, debe ser razonable. Si se asignan señales de acción o de advertencia a proporciones muy grandes o muy pequeñas de participantes, el proveedor de ensayo de aptitud debería garantizar que ésta es coherente con el propósito del programa de ensayos de aptitud.

8.4.3 Por lo general, se prefiere una estimación específica que tenga en cuenta las especificidades del problema de medición, a un enfoque genérico. Consecuentemente, antes de emplear un modelo general, se debería explorar la posibilidad de emplear los enfoques descritos en los numerales 8.2, 8.3 y 8.5.

EJEMPLO Curva Horwitz.

Un modelo general común para aplicaciones químicas fue descrito por Horwitz [22] y modificado por Thompson [31]. Este enfoque ofrece un modelo general para la reproducibilidad de métodos analíticos que se pueden emplear para derivar la siguiente expresión para la desviación estándar de la reproducibilidad:

$$\pm_R = \begin{cases} 0{,}22c & donde \quad c < 1{,}2 \times 10^{-7} \\ 0{,}02c^{0{,}8495} & donde \quad 1{,}2 \times 10^{-7} \le c \le 0{,}138 \\ 0{,}01c^{0{,}5} & donde \quad c > 0{,}138 \end{cases} \qquad (8)$$

en donde c es la fracción másica de los elementos o compuestos químicos que se van a determinar, donde $0 \le c \le 1$.

NOTA 1 El modelo Horwitz es empírico, y está basado en observaciones de ensayos de cooperación para muchos parámetros en un largo período de tiempo. Los valores σ_R son los límites superiores esperados de la variabilidad de un interlaboratorio cuando los ensayos de cooperación no presentaron problemas significativos. Por consiguiente, los valores σ_R podrían no ser criterios apropiados para determinar la competencia en un programa de ensayos de aptitud.

NOTA 2 En el Anexo E.9 se presenta un ejemplo en el que se deriva un valor del modelo Horwitz modificado.

8.5 POR EL USO DE LAS DESVIACIONES ESTÁNDAR DE REPETIBILIDAD Y REPRODUCIBILIDAD A PARTIR DE ESTUDIOS DE COOPERACIÓN PREVIOS DE LA PRECISIÓN DE UN MÉTODO DE MEDICIÓN

8.5.1 Cuando el método de medición que se va a emplear en el programa de ensayos de aptitud es normalizado y se cuenta con información disponible sobre la repetibilidad (σ_r) y la reproducibilidad (σ_R) del método, se puede calcular la desviación estándar para evaluación de aptitud (σ_{pt}) empleando esta información, de la siguiente manera:

$$\pm_{pt} = \sqrt{\pm_R^2 - \pm_r^2 (1 - 1/m)} \qquad (9)$$

donde m es el número de mediciones replicadas que cada participante va a realizar en una ronda del programa de ensayos de aptitud.

NOTA Esta ecuación se deriva de un modelo básico de efectos aleatorios de la NTC 3529-2.

8.5.2 Cuando las desviaciones estándar de repetibilidad y reproducibilidad son dependientes del valor promedio de los resultados de ensayo, deberían derivarse relaciones funcionales de los métodos descritos en la NTC 3529-2. Estas relaciones deberían emplearse para calcular los valores de las desviaciones estándar de repetibilidad y reproducibilidad apropiados para el valor asignado que se va a emplear en el programa de ensayos de aptitud.

8.5.3 Para que las anteriores técnicas sean validas, es necesario que el estudio de cooperación se haya llevado a cabo de acuerdo con los requisitos de la NTC 3529-2 ó con un procedimiento equivalente.

NOTA En el Anexo E.10 se presenta un ejemplo.

8.6 A PARTIR DE DATOS OBTENIDOS EN LA MISMA RONDA DE UN PROGRAMA DE ENSAYO DE APTITUD

8.6.1 Con este enfoque, se calcula la desviación estándar para la evaluación de aptitud, σ_{pt}, a partir de los resultados de los participantes en la misma ronda del programa de ensayos de aptitud. Por lo general, cuando se emplea este enfoque resulta más conveniente utilizar un indicador de desempeño tal como el puntaje z. Normalmente, para calcular σ_{pt} se debería emplear un estimado robusto de la desviación estándar de los resultados reportados por todos los participantes, calculado mediante el uso de una de las técnicas enunciadas en el Anexo C. En general, la evaluación con *D* ó *D%* y empleando ¢$_E$ no son apropiados en estas situaciones. No obstante, aún puede emplearse *PA* como puntaje estandarizado para la comparación entre los mensurandos (véase el numeral 9.3.6).

8.6.2 El uso de los resultados de los participantes puede conducir a criterios para la evaluación del desempeño que no son apropiados. El proveedor del ensayo de aptitud debería asegurar que el σ_{pt} empleado para las evaluaciones de desempeño sea idóneo para el fin previsto.

8.6.2.1 El proveedor de ensayos de aptitud debería poner un límite en el valor inferior de σ_{pt} que se empleará, cuando la desviación estándar robusta es muy pequeña. Este límite debería ser seleccionado de modo que cuando el error de medición es adecuado para el uso previsto más exigente, el indicador de desempeño será $z < 3{,}0$.

EJEMPLO En un programa de ensayos de aptitud para tela, un mensurando es la cantidad de hilos por centímetro. La desviación estándar 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 estándar robusta se emplea como σ_{pt}, a menos que sea inferior a 1,3 hilos/cm, en cuyo caso se emplea σ_{pt} =1,3.

8.6.2.2 El proveedor de ensayos de aptitud debería poner un límite para el valor máximo de σ_{pt} que se empleará, o en los resultados de medición que pueden evaluarse como "aceptables" (sin señal), para el caso en que la desviación estándar robusta sea muy grande. Este límite debería seleccionarse de tal modo que los resultados que no son aptos para el fin previsto reciban una señal de acción.

8.6.2.3 En algunos casos, el proveedor de ensayos de aptitud puede fijar límites superiores o inferiores en el intervalo de resultados que se pueden evaluar como "aceptables" (sin advertencia ni señal de acción), cuando los intervalos simétricos incluyen resultados que no serían aptos para el fin previsto.

EJEMPLO Para un programa de ensayos de aptitud reglamentario para agua no potable, las regulaciones especifican que los resultados deben estar dentro de $3\sigma_{pt}$ de la media robusta de los resultados de los participantes. No obstante, puesto que en algunos casos el rango de resultados aceptables pudiera incluir 0 µg/L, cualquier resultado inferior a 10 % de un valor formulado generará una señal de acción (o "inaceptable"). El ítem de un ensayo de aptitud es formulado con 4,0 µg/L de una sustancia regulada. La media robusta de los participantes es 3,2 µg/L y σ_{pt} es 1,1 µg/L. Por consiguiente, es posible que un participante presente un resultado de 0,0 µg/L y esté dentro de $3\sigma_{pt}$, pero cualquier resultado inferior a 0,4 µg/L será evaluado como “inaceptable”.

8.6.3 Las ventajas principales de este enfoque son la simplicidad y la aceptación convencional debido al uso exitoso en muchas situaciones. Este puede ser el único enfoque factible.

8.6.4 Existen varias desventajas con este enfoque:

a) El valor de σ_{pt} puede variar sustancialmente de una ronda a otra de un programa de ensayos de aptitud, dificultando que un participante emplee los valores del puntaje z para buscar tendencias que persistan en varias rondas.

b) Las desviaciones estándar pueden ser no confiables cuando la cantidad de participantes en el programa de ensayos de aptitud es pequeña o cuando se combinan resultados de diferentes métodos. Por ejemplo, si p = 20, la desviación estándar para datos distribuidos de forma normal puede variar aproximadamente ± 30 % de su valor verdadero de un ronda del programa de ensayos de aptitud con respecto a la siguiente.

c) El empleo de medidas de dispersión derivadas de los datos puede conducir a una proporción aproximadamente constante de puntajes aparentemente aceptables. Por lo general, el desempeño deficiente no será detectado mediante inspección de los puntajes y el buen desempeño conllevará a que los participantes buenos reciban calificaciones deficientes.

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

NOTA En el Anexo E.3 se presentan ejemplos completos del uso de los datos de los participantes

8.7 POR SEGUIMIENTO DE ACUERDOS INTERLABORATORIO

8.7.1 Como comprobación del desempeño de los participantes y para evaluar el beneficio del programa de ensayos de aptitud para los participantes, el proveedor del ensayos de aptitud debería aplicar un procedimiento para monitorear el acuerdo interlaboratorio, para hacer un seguimiento a los cambios en el desempeño y asegurar la validez de los procedimientos estadísticos.

8.7.2 Los resultados obtenidos en cada ronda de un programa de ensayo de aptitud deberían emplearse para calcular estimados de las desviaciones estándar de reproducibilidad del método de medición (y de repetibilidad, si están disponibles), empleando los métodos robustos descritos en el Anexo C. Estos estimados deberían representarse en gráficas de forma secuencial o como una serie en el tiempo, junto con valores de desviaciones de repetibilidad y reproducibilidad obtenidos en experimentos de precisión de la NTC 3529-2 (si los hay), y/ó σ_{pt}, si se emplean las técnicas de los numerales 8.2 al 8.4.

8.7.3 Después el proveedor de ensayo de aptitud debería examinar estas gráficas. Si las gráficas muestran que los valores de precisión obtenidos en una ronda específica de ensayo de aptitud son mayores en un factor de dos ó más por encima de los valores esperados de datos o experiencia anteriores, entonces el proveedor de ensayo de aptitud debería investigar por qué el acuerdo en esta ronda fue peor que antes. De forma similar, una tendencia hacia valores de precisión mejores o peores debería activar una investigación para la mayoría de causas probables.

9. CÁLCULO DE ESTADÍSTICAS DE DESEMPEÑO

9.1 CONSIDERACIONES GENERALES PARA DETERMINAR EL DESEMPEÑO

9.1.1 Las estadísticas empleadas para determinar el desempeño deben ser coherentes con el/los objetivo(s) para el programa de ensayos de aptitud.

NOTA Las estadísticas de desempeño son más útiles si los participantes y otras partes interesadas entienden las estadísticas y de donde se derivan.

9.1.2 Los indicadores de desempeño deberían ser fácilmente revisados a través de los niveles del mensurando y de las diferentes rondas de un programa de ensayos de aptitud.

9.1.3 Los resultados de los participantes deberían ser revisados y determinar su coherencia con las suposiciones empleadas en el diseño de dicho programa de ensayos de aptitud para permitir estadísticas de desempeño significativas. Por ejemplo, que no haya evidencia de deterioro del ítem de ensayo de aptitud ni una mezcla de poblaciones de participantes ni violaciones graves de ninguna suposición estadística sobre la naturaleza de los datos.

9.1.4 En general, no resulta apropiado emplear métodos de evaluación que clasifiquen de forma intencional una proporción fija de resultados como generadores de una señal de acción.

9.2 LIMITACIÓN DE LA INCERTIDUMBRE DEL VALOR ASIGNADO

9.2.1 Si la incertidumbre estándar $u(x_{pt})$ del valor asignado es grande en comparación con el criterio de evaluación del desempeño, entonces existe un riesgo de que algunos participantes reciban señales de acción y advertencia debido a la inexactitud en la determinación del valor asignado, y no debido a cualquier causa del participante. Por esta razón, la incertidumbre estándar del valor asignado debe ser determinada e informada a los participantes (véase la NTC-ISO/IEC 17043, numerales 4.4.5 y 4.8.2).

Si se cumple el siguiente criterio, entonces la incertidumbre del valor asignado puede considerarse despreciable y no se debe incluir en la interpretación de los resultados de la ronda del ensayo de aptitud.

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

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

9.2.2 Si no se cumple este criterio, entonces el proveedor de ensayo de aptitud debería considerar lo siguiente, garantizando que cualquier acción que se emprenda sea coherente con la política de evaluación de desempeño acordada para el programa de ensayo de aptitud en cuestión.

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

b) Emplear la incertidumbre del valor asignado para interpretar los resultados del programa de ensayos de aptitud (véanse las numerales 9.5 sobre el puntaje z', ó la 9.6 sobre puntajes ζ ó la 9.7 sobre puntajes E_n).

c) Si el valor asignado se deriva de los resultados de los participantes y surge gran incertidumbre a partir de las diferencias entre sub poblaciones identificables de participantes, se reportan valores por separado y las incertidumbres para cada sub población (por ejemplo, los participantes que emplean diferentes métodos de medición).

   NOTA El Protocolo armonizado IUPAC [32] describe un procedimiento específico para detectar bimodalidad, con base en una inspección de una gráfica de densidad kernel con un ancho de banda especificado.

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

Si ninguno de los literales a) - d) aplica, entonces se deberá informar a los participantes que no se puede determinar un valor asignado confiable y que no se pueden proveer indicadores de desempeño.

NOTA En los Anexos E.3 y E.4 se demuestran las técnicas presentadas en este numeral.

9.3 ESTIMADOS DE LA DESVIACIÓN (ERROR DE MEDICIÓN)

9.3.1 Si x_i representa el resultado (o el promedio de las repeticiones) reportados por un participante *i* para la medición de una propiedad del ítem de ensayo de aptitud en una ronda de un programa de ensayo de aptitud. Entonces, una medida simple de desempeño del participante puede calcularse como la diferencia entre el resultado x_i y el valor asignado x_{pt}:

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

D_i puede interpretarse como el error de medición para dicho resultado, en la medida en que el valor asignado pueda considerarse como un valor convencional o de referencia.

La diferencia D_i puede expresarse en las mismas unidades que el valor asignado o como una diferencia de porcentaje, que se calcula así:

$$D_i\% = 100\left(x_i - x_{pt}\right)/\, x_{pt}\% \qquad (12)$$

9.3.2 Por lo general, la diferencia *D* ó *D%* se compara con un criterio $¢_E$ con base en la idoneidad para el fin previsto o con experiencia de rondas anteriores de un programa de ensayos de aptitud; el criterio se anota aquí como $¢_E$, una tolerancia para el error de medición. Sí $-¢_E < D < ¢_E$ entonces se considera que el desempeño es 'aceptable' (o 'sin señal'). (El mismo criterio se aplica para *D%,* dependiendo de la expresión de $¢_E$.).

9.3.3 $¢_E$ está estrechamente relacionado con σ_{pt} como empleado para puntajes z (véase el numeral 9.4), cuando se determina σ_{pt} por idoneidad para el fin previsto o expectativas de rondas anteriores. La relación está determinada por los criterios de evaluación para puntajes z. Por ejemplo, si $z \geq 3$ crea una señal de acción $¢_E = 3\,\sigma_{pt}$, ó equivalente a $\sigma_{pt} = ¢_E /3$ equivalente. Varias expresiones de $¢_E$ son convencionales en el ensayo de aptitud para aplicaciones médicas y en especificaciones de desempeño para métodos de medición y productos.

9.3.4 La ventaja de *D* como estadística de desempeño y $¢_E$ como criterio de desempeño es que los participantes tienen una comprensión intuitiva de tales estadísticas puesto que se vinculan directamente al error de medición y son criterios comunes para determinar la idoneidad para el fin previsto. La ventaja de *D%* es que la comprensión es intuitiva, es normalizada para el nivel de mensurando y se relaciona con causas comunes de error (por ejemplo, calibración incorrecta o sesgo en la dilución).

9.3.5 Las desventajas son, que no es convencional para ensayos de aptitud en muchos países o campos de medición; y que *D* no está estandarizado para permitir exploración simple de informes en busca de señales de acción en programas de ensayos de aptitud con múltiples analitos o cuando los criterios de idoneidad para el fin previsto pueden variar por nivel del mensurando.

NOTA Por lo general, el uso de *D* y *D%* supone simetría de la distribución de los resultados de los participantes en el sentido en que el rango aceptable es $-¢_E < D < ¢_E$.

9.3.6 Para propósitos de comparaciones a través de niveles del mensurando, donde los criterios de idoneidad para el fin previsto pueden variar; o para combinación a través de rondas o a través de mensurandos, *D* y *D%* pueden ser transformados en un indicador de desempeño normalizado que muestre las diferencias relacionadas con los criterios de desempeño de los mensurandos. Para hacerlo, calcule el "Porcentaje de desviación permitida" (P_A) para cada resultado, de la siguiente manera:

$$P_{Ai} = \left(D_i / ¢_E\right) \text{ x } 100\,\% \qquad (13)$$

Por consiguiente, $P_A \geq 100$ % ó $P_A \leq -100$ % indica una señal de acción (o "desempeño inaceptable").

NOTA 1 Los indicadores P_A se pueden comparar a través de los niveles y rondas diferentes de un programa de ensayos de aptitud, o rastrearse en tablas. Estos indicadores de desempeño tienen un uso e interpretación similares a los puntajes z que tienen un criterio de evaluación común tal como $z \leq -3$ ó $z \geq 3$ para señales de acción.

NOTE 2 Las variaciones de esta estadística son de uso común, en especial en aplicaciones médicas, donde por lo general existe una frecuencia mayor de ensayos de aptitud y una gran cantidad de analitos.

NOTA 3 Puede ser apropiado emplear el valor absoluto de P_A para reflejar resultados coherentemente aceptables (o inaceptables) en relación con el valor asignado.

9.4 PUNTAJES z

9.4.1 El puntaje z para un resultado de ensayos de aptitud x_i *se* calcula así:

$$z_i = \frac{\left(x_i - x_{pt}\right)}{\pm_{pt}} \qquad (14)$$

en donde

x_{pt} es el valor asignado, y

σ_{pt} es la desviación estándar para evaluación de la aptitud

9.4.2 La interpretación convencional de los puntajes z es la siguiente (véase laNTC-ISO/IEC 17043, literal B.4.1.1):

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

Se debería aconsejar a los participantes que comprueben sus procedimientos de medición siguiendo las señales de advertencia en el caso en que los puntajes indiquen un problema emergente o recurrente.

NOTA 1 En algunas aplicaciones, los proveedores de ensayos de aptitud emplean 2,0 como una señal de acción para puntajes z.

NOTA 2 La elección del criterio σ_{pt} debería hacerse normalmente para permitir la interpretación anterior, que se emplea ampliamente para evaluación de aptitud y además es muy similar a los límites de gráficos de control conocidos.

NOTA 3 La justificación para el uso de los límites de 2,0 y 3,0 para puntajes z es la siguiente: Se supone que las mediciones que se realizan de forma correcta generan resultados que pueden describirse (después de transformación si es necesaria) mediante una distribución normal con una media x_{pt} y desviación estándar σ_{pt}. Entonces. Los puntajes z se distribuirán de forma normal con una media de cero y una desviación estándar de 1,0. Bajo estas circunstancias se esperaría que sólo el 0,3 % de los indicadores quedara por fuera del rango $-3{,}0 \leq z \leq 3{,}0$ y sólo cerca del 5 % quedaría por fuera del rango $-2{,}0 \leq z \leq 2{,}0$. Puesto que la probabilidad de que z quede por fuera de ± 3,0 es muy baja, es poco probable que ocurran señales de acción aleatorias cuando no exista un problema real, de modo que es probable que haya una causa identificable de una anomalía cuando se presenta una señal de acción.

NOTA 4 La suposición sobre la cual se basa esta interpretación se aplica sólo a una distribución hipotética de laboratorios competentes y no una suposición sobre la distribución de los resultados observados. No es necesario hacer suposiciones sobre los resultados observados.

NOTA 5 Si la variabilidad interlaboratorio verdadera es menor que σ_{pt} entonces se reducen las probabilidades de clasificación errónea.

NOTA 6 Cuando se fija la desviación estándar para evaluación de aptitud por cualquiera de los métodos descritos en el numeral 8.2 u 8.4, ésta puede diferir sustancialmente de la desviación estándar (robusta) de resultados, y las proporciones de resultados que quedan por fuera de ± 2,0 y ± 3,0 pueden diferir considerablemente en 5 % y 0,3 % respectivamente.

9.4.3 El proveedor de ensayo de aptitud debe determinar el redondeo apropiado para los puntajes z reportados, con base en la cantidad de dígitos significativos para el resultado y para el valor asignado y la desviación estándar para el ensayo de aptitud. Se deben incluir las reglas para redondeo en la información disponible para los participantes.

NOTA Rara vez es útil tener más de dos dígitos después del decimal para puntajes z.

9.4.4 Cuando la desviación estándar de los resultados de los participantes se emplea como σ_{pt} y los programas de ensayo de aptitud incluyen cantidades muy grandes de participantes, es posible que el proveedor de ensayo de aptitud desee comprobar la normalidad de la distribución, empleando los resultados reales o los puntajes z. En el otro extremo, cuando sólo existe una pequeña cantidad de participantes, es posible que no se presente señal de acción. En este caso, los métodos gráficos que combinan indicadores de desempeño de varias rondas pueden proporcionar indicaciones más útiles del desempeño de los participantes que los resultados de rondas individuales.

9.5 PUNTAJES z´

9.5.1 Cuando existe preocupación sobre la incertidumbre de un valor asignado $u(x_{pt})$, por ejemplo cuando $u(x_{pt})$ > 0,3 σ_{pt}, entonces se puede tener en cuenta la incertidumbre expandiendo el denominador del indicador de desempeño. Esta estadística se denomina puntaje z' y se calcula de la siguiente manera (con la misma notación como se muestra en el numeral 9.4):

$$z_i' = \frac{x_i - x_{pt}}{\sqrt{\pm_{pt}^2 + u^2(x_{pt})}} \qquad (15)$$

NOTA Cuando se calculan x_{pt} y/ó σ_{pt} de los resultados de los participantes, el indicador de desempeño se correlaciona con resultados de participantes individuales, porque éstos tienen un impacto tanto en una media robusta como en la desviación estándar. La correlación para un participante individual depende de la ponderación dada a dicho participante en la estadística combinada. Por esta razón, los indicadores de desempeño que incluyen la incertidumbre del valor asignado sin tener una previsión para correlación representan subestimados de los puntajes que resultarían si se incluyera la covarianza. Por ejemplo, cuando $u(x_{pt})$ = 0,3 σ_{pt} entonces existe un subestimado de aproximadamente 10 % del puntaje z'. Por consiguiente, se puede emplear la ecuación (15) cuando se determinan x_{pt} y/ó σ_{pt} a partir de los resultados de los participantes.

9.5.2 Los puntajes *D* y *D%* también pueden modificarse para tener en cuenta la incertidumbre del valor asignado con la siguiente fórmula a fin de expandir $¢_E$ a $¢_E$'

$$¢_E' = \sqrt{¢_E^2 + U^2(x_{pt})} \qquad (16)$$

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

9.5.3 Los puntajes z' pueden interpretarse de la misma forma que los puntajes z (véase el numeral 9.4) y empleando los mismos valores críticos de 2,0 y 3,0, dependiendo del diseño del programa de ensayos de aptitud. De manera similar, los puntajes *D* y *D%* se compararían luego con $¢_E'$ (véase el numeral 9.3).

9.5.4 La comparación de las fórmulas para el puntaje z y el puntaje z' en el numeral 9.4 y 9.5 demuestra que los puntajes z' para una ronda de un programa de ensayos de aptitud siempre serán menores que los puntajes z correspondientes por un factor constante de:

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

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

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

Así, en este caso, los puntajes z' serán casi idénticos a los puntajes z y se puede concluir que la incertidumbre del valor asignado es despreciable para la evaluación de desempeño.

Cuando no se cumple la directriz del numeral 9.2.1 para la incertidumbre del valor asignado, la diferencia en la magnitud de los puntajes z' y z puede ser tal que algunos puntajes z superen los valores críticos de 2,0 ó 3,0 y que presenten "señales de advertencia" ó "señales de acción", mientras que los puntajes z' correspondientes no superen estos valores críticos y no arrojen señales.

En general, para situaciones en las que el valor asignado y/o σ_{pt} no se determinen a partir de los resultados de los participantes, se puede preferir z' puesto que cuando el criterio del numeral 9.2.1 se cumple, la diferencia entre z y z' será despreciable.

9.6 PUNTAJES ZETA (ζ)

9.6.1 Los puntajes zeta pueden ser útiles cuando un objetivo para el programa de ensayos de aptitud es evaluar la capacidad de un participante para obtener resultados cercanos al valor asignado dentro de su incertidumbre declarada.

De acuerdo con las notaciones del numeral 9.4, los puntajes ζ se calculan como:

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

en donde

$u(x_i)$ es el estimado propio del participante de la incertidumbre estándar de su resultado x_i, y

$u(x_{pt})$ es la incertidumbre estándar del valor asignado x_{pt}.

NOTA 1 Cuando el valor asignado x_{pt} se calcula como el valor de consenso a partir de los resultados de los participantes, entonces x_{pt} se correlaciona con los resultados de participantes individuales. La correlación para un participante individual depende de la ponderación dada a dicho participante en el valor asignado y, en menor grado, en la incertidumbre del valor asignado. Por esta razón, los indicadores de desempeño que incluyen la incertidumbre del valor asignado sin tener una tolerancia para correlación representan subestimados de los puntajes que resultarían si se incluyera la covarianza. La subestimación no es grave si la incertidumbre del valor asignado es pequeña; cuando se emplean métodos robustos es menos grave para los participantes más alejados con mayor probabilidad de recibir indicadores de desempeño adversos. Por consiguiente, la ecuación (17) puede emplearse con estadísticas de consenso sin ajuste para la correlación.

NOTA 2 Los puntajes ζ difieren de puntajes E_n (numeral 9.7) mediante el uso de incertidumbres estándar $u(x_i)$ y $u(x_{pt})$, en vez de incertidumbres expandidas $U(x_i)$ y $U(x_{pt})$. Los resultados de puntajes ζ por encima de 2 ó por debajo de -2 pueden ser causados por métodos sistemáticamente sesgados o por un cálculo deficiente de la incertidumbre de la medición por el participante. Por consiguiente, los puntajes ζ ofrecen una evaluación rigurosa del resultado completo presentado por el participante.

9.6.2 El uso de puntajes ζ permite la evaluación directa de si los laboratorios pueden entregar resultados correctos, es decir, resultados que concuerden con x_{pt} dentro de sus incertidumbres de medición. Los puntajes ζ pueden interpretarse empleando los mismos valores críticos de 2,0 y 3,0 al igual que para puntajes z, o con múltiplos del factor de cobertura del participante, empleado cuando se calcula la incertidumbre expandida. No obstante, un puntaje adverso del puntaje ζ puede indicar o una gran desviación de x_i con respecto a x_{pt}, o un subestimado de incertidumbre de parte del participante, o una combinación de ambos.

NOTA Puede ser útil para el proveedor de ensayos de aptitud brindar información adicional sobre la validez de las incertidumbres reportadas. En el numeral 9.8 se sugieren directrices útiles para dicha evaluación.

9.6.3 Se pueden usar los puntajes ζ en conjunto con puntajes z como una ayuda para mejorar el desempeño de los participantes, de la siguiente manera. Si un participante obtiene resultados del puntaje z que superen repetidamente el valor crítico de 3,0, puede resultarle útil examinar su procedimiento de ensayo paso a paso y derivar una evaluación de la incertidumbre para dicho procedimiento. En la evaluación de la incertidumbre se identificarán los pasos del procedimiento donde surgen las incertidumbres más grandes, de modo que el participante pueda ver donde aumentar sus esfuerzos para lograr mejora. Si los resultados del puntaje ζ del participante superan repetidamente el valor crítico de 3,0, esto implica que la evaluación de la incertidumbre del participante no incluye todas las fuentes significativas de incertidumbre (es decir, falta algo importante). Por el contrario, si un participante repetidamente obtiene resultados del puntaje $z \geq 3$ pero los resultados del puntajes $\zeta < 2$, esto demuestra que el participante pudo haber evaluado la incertidumbre de sus resultados con precisión pero que sus resultados no cumplen el desempeño esperado para el programa de ensayos de aptitud. Este puede ser el caso, por ejemplo, de un participante que emplea un método de tamizaje en los procedimientos de medición donde los otros participantes aplican métodos cuantitativos. No se requiere acción si el participante considera que la incertidumbre de sus resultados es suficiente.

NOTA Cuando se emplea un solo puntaje ζ, éste se puede interpretar sólo como un ensayo de si la incertidumbre del participante es coherente con la desviación particular observada y no se puede interpretar como una indicación de la idoneidad para el fin previsto de los resultados de un participante particular. La determinación de la idoneidad para el fin previsto podría realizarse por separado (por ejemplo, por el participante o por un organismo acreditador) examinando la desviación (x-x_{pt}) o las incertidumbres estándar combinadas en comparación con una incertidumbre objetivo.

9.7 PUNTAJES E_n

9.7.1 Los puntajes E_n pueden ser útiles cuando un objetivo para el programa de ensayos de aptitud es evaluar la capacidad de un participante para obtener resultados cercanos al valor asignado dentro de su incertidumbre expandida declarada. Esta estadística es convencional para los ensayos de aptitud en calibración; pero puede emplearse para otros tipos de ensayos de aptitud.

Esta estadística de desempeño se calcula como:

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

en donde

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

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

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

NOTA La combinación directa de incertidumbres expandidas no es coherente con el requisito de la Guía ISO/IEC 98-3 y no es equivalente al cálculo de una incertidumbre expandida combinada a menos que ambos factores de cubrimiento y los grados efectivos de libertad sean idénticos para $U(x_i)$ y $U(x_{pt})$.

9.7.2 Los puntajes E_n deberían interpretarse con precaución, puesto que son relaciones de dos medidas de desempeño separadas (pero relacionadas). El numerador es la desviación del resultado del valor asignado y tiene una interpretación analizada en el numeral 9.3. El denominador es una incertidumbre expandida combinada que no debería ser superior a la desviación en el numerador, si el participante ha determinado $U(x_i)$ de forma correcta y si el proveedor de ensayo de aptitud ha determinado $U(x_{pt})$ de forma correcta. Por lo tanto, los puntajes de $\boldsymbol{E_n} \geq 1{,}0$ ó $\boldsymbol{E_n} \leq -1{,}0$ podrían indicar una necesidad de revisar los estimados de incertidumbre o corregir un problema de medición; de manera similar $-1{,}0 < \boldsymbol{E_n} < 1{,}0$ debería tomarse como un indicador de desempeño exitoso sólo si las incertidumbres son validas y la desviación $(x_i - x_{pt})$ es más pequeña que la requerida por los clientes de los participantes.

NOTA Si bien la interpretación de los puntajes $\boldsymbol{E_n}$, puede ser difícil, esto no impide su uso. La incorporación de la información sobre la incertidumbre en la interpretación de resultados de ensayos de aptitud puede cumplir una función importante en la mejora de la comprensión de los participantes sobre la incertidumbre de la medición y su evaluación.

9.8 EVALUACIÓN DE INCERTIDUMBRES DE PARTICIPANTES EN EL ENSAYO

9.8.1 Con la creciente aplicación de la norma NTC-ISO/IEC 17025 existe mejor comprensión de la incertidumbre de la medición. El uso de evaluaciones de laboratorio sobre incertidumbre en la evaluación de desempeño ha sido común en programas de ensayos de aptitud en diferentes áreas de calibración, tales como con los puntajes $\boldsymbol{E_n}$, pero no ha sido común en ensayos de aptitud para laboratorios de ensayo. Los puntajes ζ descritos en el numeral 9.6, y los puntajes $\boldsymbol{E_n}$ del numeral 9.7, son opciones para la evaluación de los resultados contra la incertidumbre declarada.

9.8.2 Algunos proveedores de ensayos de aptitud han reconocido la utilidad de pedir a los laboratorios que reporten la incertidumbre de los resultados en los ensayos de aptitud. Esto puede ser útil cuando las incertidumbres no se emplean en la determinación del puntaje i. Existen varios propósitos para reunir dicha información:

a) los organismos de acreditación pueden asegurar que los participantes estén reportando incertidumbres que sean coherentes con su alcance de acreditación;

b) los participantes pueden revisar su incertidumbre reportada junto con la de otros participantes, para evaluar la coherencia (o no) y por ende tener la oportunidad de identificar si su evaluación de la incertidumbre no tiene en cuenta todos los componentes pertinentes o está sobre estimando algunos componentes;

c) los ensayos de aptitud pueden ser empleados para confirmar declaraciones de incertidumbre y esto resulta más fácil cuando se reporta la incertidumbre con el resultado.

NOTA En el Anexo E.3 se presenta un ejemplo del análisis de datos cuando se reportan incertidumbres.

9.8.3 Cuando se determina x_{pt} empleando los procedimientos de los numerales 7.3 - 7.6 y $u(x_{pt})$ cumple el criterio del numeral 9.2.1 entonces es improbable que un resultado de participante tenga incertidumbre estándar menor que ésta, de modo que se podría emplear $u(x_{pt})$ como un límite inferior para tamizaje, denominado u_{min}. Si se determina el valor asignado a partir de los resultados de participante (véase el numeral 7.7), entonces el proveedor de ensayo de aptitud debería determinar límites de tamizaje prácticos para u_{min}.

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

9.8.4 También es improbable que la incertidumbre estándar reportada de cualquier participante sea mayor de 1,5 veces la desviación estándar robusta de los participantes (1,5s*), de modo que se podría usar esta como límite superior práctico para filtrar incertidumbres reportadas, denominadas u_{max}.

NOTA El factor 1,5 es el límite superior de la variabilidad en las desviaciones estándar que pueden ser esperadas para una desviación estándar de consenso con 10 ó más resultados, con base en la raíz cuadrada de percentiles de la distribución F. Cualquier proveedor de ensayos de aptitud que adopte este procedimiento tal vez considere emplear un multiplicador diferente.

9.8.5 Si se emplean u_{min} ó u_{max}, u otros criterios para identificar incertidumbres aberrantes, el proveedor del ensayo de aptitud debería explicar esto a los participantes y aclarar que una incertidumbre reportada, $u(x_i)$, puede ser válida incluso si es inferior que u_{min} o superior a u_{max}; y cuando esto ocurra los participantes y cualquier parte interesada debería comprobar el resultado o el estimado de incertidumbre. De manera similar, una incertidumbre reportada puede ser mayor que u_{min} y menor que u_{max} y aún no ser válida. Estos son solamente puntajes informativos.

9.8.6 Los proveedores de ensayos de aptitud también pueden llamar la atención con respecto a incertidumbres inusualmente altas o bajas con base en, por ejemplo:

- cuartiles especificados para las incertidumbres reportadas (por ejemplo, por debajo del quinto percentil y por encima del 95^{avo} percentil de incertidumbres estándar o expandidas reportadas);

- límites basados en una distribución supuesta con base en la dispersión de las incertidumbres reportadas;

- una incertidumbre de medición requerida.

NOTA Puesto que es improbable que las incertidumbres se distribuyan de forma normal, probablemente será necesaria una transformación cuando se empleen límites que dependan de una normalidad aproximada o subyacente; por ejemplo, los límites de bigote del diagrama de caja con base en el rango intercuartil tienen una interpretación probabilística sólo cuando la distribución es aproximadamente normal.

9.9 INDICADORES DE DESEMPEÑO COMBINADOS

9.9.1 Dentro de una ronda única de un programa de ensayos de aptitud, es común que se obtengan resultados para más de un ítem de ensayo de aptitud o para más de un mensurando. En esta situación, los resultados para cada ítem del ensayo de aptitud y para cada mensurado deberían interpretarse como se describe en los numerales 9.3 a 9.7; es decir, los resultados de cada ítem de ensayo de aptitud y de cada mensurando se deberían evaluar por separado.

9.9.2 Existen aplicaciones en las que se incluyen dos o más ítems de ensayo de aptitud con niveles diseñados de forma especial en un programa de ensayos de aptitud para medir otros aspectos de desempeño, tales como investigar la repetibilidad, el error sistemático o la

linealidad. Por ejemplo, se pueden emplear dos ítems de ensayo de aptitud similares en un programa de ensayos de aptitud con la intención de tratarlos con una gráfica de Youden, como se describe en el numeral 10.5. En tales casos, el proveedor de ensayos de aptitud debería proporcionar a los participantes descripciones completas del diseño estadístico y de los procedimientos que se emplean.

9.9.3 Se deberían emplear los métodos gráficos del numeral 10 cuando se obtienen resultados para más de un ítem de ensayo de aptitud o para varios mensurandos, siempre que estén estrechamente relacionados y/o se obtengan mediante el mismo método. Estos procedimientos combinan resultados de indicadores de desempeño de tal forma que no ocultan valores elevados de resultados de indicadores individuales y pueden revelar información adicional sobre el desempeño de los participantes - tales como, la correlación entre los resultados para mensurandos diferentes- que no es evidente en tablas de indicadores individuales.

9.9.4 En programas de ensayos de aptitud que incluyen una gran cantidad de mensurandos, se puede emplear un recuento o proporción de las cantidades de señales de acción y advertencia para evaluar el desempeño.

9.9.5 Los indicadores de desempeño combinado o de recompensa o penalización se deberían emplear sólo de manera cuidadosa, ya que puede ser difícil describir las suposiciones estadísticas subyacentes a los indicadores. Mientras que los indicadores de desempeño combinados para resultados sobre ítems de ensayo de aptitud diferentes sobre el mismo mensurando pueden tener distribuciones esperadas y ser útiles para detectar sesgo persistente, los indicadores promediados o resumidos a través de los diferentes mensurandos para los mismos o diferentes ítems de ensayo de aptitud pueden ocultar el sesgo en resultados para mensurandos únicos. El método de cálculo, la interpretación y las limitaciones de cualquier puntaje combinado o de penalización empleado, debe ser claro para los participantes.

10. MÉTODOS GRÁFICOS PARA ILUSTRAR INDICADORES DE DESEMPEÑO

10.1 APLICACIÓN DE MÉTODOS GRÁFICOS

Normalmente, el proveedor de ensayos de aptitud debería emplear los indicadores de desempeño obtenidos en cada ronda de un programa de ensayos de aptitud para preparar gráficas tales como las descritas en los numerales 10.2 y 10.3. El uso de indicadores de desempeño, tales como P_A, z, z', ζ, ó el indicador de desempeño E_n en estas gráficas tiene la ventaja que se pueden diagramar empleando ejes normalizados, por ende simplificando su presentación e interpretación. Las gráficas deberían ponerse a disposición de los participantes, permitiéndole a cada uno observar donde encajan sus propios resultados en relación con aquellos obtenidos por otros participantes. Se pueden emplear códigos alfabéticos o numéricos para representar a los participantes, de modo que cada uno pueda identificar sus propios resultados pero no pueda determinar cuál participante obtuvo cualquier otro resultado. Las gráficas también pueden ser usadas por el proveedor de ensayos de aptitud y cualquier organismo de acreditación, para poder juzgar la eficacia general del programa de ensayos de aptitud y ver si hay necesidad de revisar los criterios empleados para evaluar el desempeño.

10.2 HISTOGRAMAS DE RESULTADOS O DE INDICADORES DE DESEMPEÑO

10.2.1 El histograma es una herramienta estadística común que es útil en dos puntos diferentes del análisis de resultados de un ensayo de aptitud. La gráfica es útil en la etapa de análisis preliminar, para comprobar si las suposiciones estadísticas son razonables o si existe una anomalía - tal como una distribución bimodal, una gran proporción de valores atípicos, o asimetría inusual que no se había previsto.

Los histogramas también pueden ser útiles en informes del programa de ensayos de aptitud para describir los indicadores de desempeño o para comparar los resultados en, por ejemplo, diferentes métodos o diferentes ítems del ensayo de aptitud. Son particularmente útiles en informes individuales para programas de ensayos de aptitud de tamaño pequeño o moderado (con menos de 100 participantes) para permitir a los participantes evaluar cómo se compara su desempeño con el de otros participantes, por ejemplo, cuando se resalta un bloque dentro de una barra vertical para representar el resultado de un participante o, en programas de ensayo de aptitud pequeños (con menos de 50 participantes), empleando una gráfica con caracteres individualizados para cada participante.

10.2.2 Los histogramas se pueden preparar empleando los resultados reales de los participantes o indicadores de desempeño. Los resultados de los participantes, tienen la ventaja de estar directamente relacionados con los datos presentados y se pueden evaluar sin cálculos o transformaciones adicionales desde el indicador de desempeño hasta el error de medición. Los histogramas basados en indicadores de desempeño tienen la ventaja de relacionarse de forma directa con evaluaciones de desempeño y se pueden comparar fácilmente a través de los mensurandos y las rondas de un programa de ensayos de aptitud.

El rango y el tamaño de caja (columna, cuadro) empleado para un histograma se deben determinar para cada conjunto de datos, con base en la variabilidad y el número de resultados. A menudo, es posible realizar esto con base en la experiencia en ensayos de aptitud, pero en la mayoría de situaciones los intervalos de los grupos necesitarán ser ajustados después de la primera revisión. Si se emplean indicadores de desempeño en el histograma, resulta útil tener una escala con base en la desviación estándar para la evaluación de aptitud y puntos de corte para señales de advertencia y acción.

10.2.3 La escala y los intervalos de la gráfica se deberían seleccionar de modo que se pueda detectar dicha bimodalidad (si la hay), sin crear advertencias falsas debido a la resolución de los resultados de medición o a cantidades pequeñas de resultados.

NOTA 1 La apariencia de los histogramas es sensible al ancho de caja seleccionado y a la ubicación de los límites de la caja (para ancho de caja constante que depende en gran medida del punto de partida). Si el ancho de caja es demasiado pequeño, la gráfica mostrará muchas modas pequeñas; si es demasiado grande, es posible que las modas apreciables cerca del cuerpo principal no se distingan suficientemente. La apariencia de modas angostas y las alturas relativas de barras adyacentes pueden cambiar de forma apreciable al cambiar la posición de partida o el ancho de caja, en especial donde el conjunto de datos es pequeño y/o muestra algún agrupamiento.

NOTA 2 En el Anexo E.3 se presenta un ejemplo de representación gráfica con histograma.

10.3 GRÁFICAS DE DENSIDAD KERNEL

10.3.1 Una gráfica de densidad Kernel, a menudo denominada en forma abreviada como "gráfica de densidad" presenta una curva uniforme que describe la forma general de la distribución del conjunto de datos. La idea que subyace al estimador kernel es que cada punto de datos es remplazado por una distribución especificada (por lo general normal), centrada en el punto y con una desviación estándar σ_k; por lo general, σ_k se denomina "ancho de banda". Estas distribuciones se suman y la distribución resultante, ajustada para tener un área unitaria, proporcionan un "estimador de densidad" que se puede representar como una curva uniforme.

10.3.2 Se pueden seguir los siguientes pasos para preparar una gráfica de densidad kernel. Se supone que se va a incluir un conjunto de datos X que consta de p valores x_1, x_2, ..., x_p en la gráfica. Por lo general, estos son resultados de los participantes pero pueden ser indicadores de desempeño derivados de los resultados.

i) Seleccione un ancho de banda σ_k apropiado. Dos opciones resultan particularmente útiles:

a) Para inspección general, se fija σ_k = 0,9 $s^*/p^{0,2}$ donde s^* es una desviación estándar robusta de los valores x_1, ..., x_p calculada empleando los procedimiento del Anexo C.2 ó C.3.

b) A fin de examinar los conjuntos de datos de modas groseras que son importantes comparados con el criterio para evaluación del desempeño, se establece σ_k = 0,75 σ_{pt} si se emplean puntajes z ó ζ, ó σ_k = 0,25 ¢$_E$ si se emplea D ó $D\%$.

NOTA 1 La opción a) anterior concuerda con Silverman[30] que recomienda s^* con base en el rango intercuartilico normalizado ($nIQR$). Otras reglas de selección de ancho de banda que proporcionan resultados similares incluyen la de Scott[29], que remplaza el multiplicador 0,9 con 1,06. La referencia [29] describe un método de selección de ancho de banda casi óptimo, pero mucho más complejo. En la práctica, las diferencias por inspección visual son leves y la selección depende de la disponibilidad de software.

NOTA 2 La opción b) anterior concuerda con la orientación IUPAC [32].

ii) Se fija un rango de diagramación q_{min} a q_{max} de modo que $q_{min} \leq \min(x_1, \ldots x_p) - 3\,\sigma_k$ y $q_{max} \geq \max(x_1, \ldots x_p) + 3\,\sigma_k$.

iii) Seleccione la cantidad de puntos n_k para diagramar o graficar la curva. n_k = 200 por lo general es suficiente a menos que existan valores atípicos extremos dentro del rango del diagrama.

iv) Calcule las ubicaciones de diagramación q_1 a q_{nk} a partir de:

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

v) Calcule las n_k densidades h_1 a $h_{n\mathrm{k}}$ a partir de

$$h_i = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{x_j - q_i}{\pm_k}\right) para\ i=1\ a\ i=n_k \qquad (20)$$

en donde $\phi(.)$ denota la densidad normal estándar.

vi) Graficar h_i contra q_i .

NOTA 1 Puede ser útil agregar las ubicaciones de los puntos de datos individuales a la gráfica. La forma más común de hacerlo es representando las ubicaciones por debajo de la curva de densidad trazada como marcadores verticales cortos, pero también puede ser apropiado diagramar los puntos de datos en los puntos apropiados a lo largo de la curva de densidad calculada.

NOTA 2 Las gráficas de densidad son mejor elaboradas con software. El cálculo por pasos descrito anteriormente puede realizarse en una hoja de cálculo para conjuntos de datos con tamaños modestos. Con frecuencia, el software patentado y disponible de forma gratuita incluye gráficas de densidad con base en opciones de ancho de banda similares predeterminados. Las implementaciones de software avanzado de gráficas de densidad pueden emplear este algoritmo o cálculos más rápidos con base en métodos para calcular convolución.

NOTA 3 En los Anexos E.3, E.4 y E.6 se presentan ejemplos de gráficas de densidad kernel.

10.3.3 La forma de la curva se toma como una indicación de la distribución a partir de la cual se diagramaron los datos. Las modas distintas aparecen como picos separados. Los valores anómalos aparecen como picos muy separados del cuerpo principal de los datos.

NOTA 1 Las gráficas de densidad son sensibles al ancho de banda σ_k seleccionado. Si el ancho de banda es demasiado pequeño, la gráfica mostrará muchas modas pequeñas; si es demasiado grande, es posible que las modas apreciables cerca del cuerpo principal no se distingan suficientemente.

NOTA 2 Al igual que los histogramas, las gráficas de densidad se emplean mejor en conjuntos de datos moderados a grandes, ya que, conjuntos de datos pequeños (diez o menos) de repente pueden incluir valores atípicos moderados o modas aparentes, en especial cuando se emplea una desviación estándar robusta como la base para el ancho de banda.

10.4 GRÁFICAS DE BARRAS DE INDICADORES DE DESEMPEÑO NORMALIZADOS

10.4.1 Las gráficas de barras son un método adecuado para presentar los indicadores de desempeño para un número de características similares en un gráfico. Ellas podrán revelar si existe una característica común en los puntajes de un participante; por ejemplo, si un participante obtiene puntajes z elevados, por lo general indica un desempeño deficiente del participante, con un posible sesgo positivo.

10.4.2 Para preparar un gráfico de barras, se recolectan los indicadores de desempeño normalizados en un gráfico de barras como el que se ilustra en la Figura E.10, en el cual se agrupan los resultados de los puntajes para cada participante. Otros indicadores de desempeño normalizados, tales como *D%* ó P_A pueden graficarse para el mismo propósito.

10.4.3 Cuando se realizan determinaciones replicadas en una ronda de un programa de ensayos de aptitud, se pueden emplear los resultados para calcular una gráfica de medidas de precisión; por ejemplo el estadístico k como se describe en la NTC 3529-2, o una medida ajustada contra la desviación estándar del promedio robusto tal como la que se define en el Algoritmo S (Anexo C.4).

NOTA En el Anexo E.11 se presenta un ejemplo de gráfica de barras con puntaje *z*.

10.5 GRÁFICA DE YOUDEN

10.5.1 Cuando se han ensayado dos ítems similares en una ronda de un programa de ensayos de aptitud, la Gráfica de Youden brinda un método gráfico muy informativo para estudiar los resultados. Puede ser útil para demostrar correlación (o independencia) de resultados en diferentes ítems del ensayo de aptitud y para guiar las investigaciones sobre razones para señales de acción.

10.5.2 La gráfica es construida al trazar los resultados de los participantes, o los resultados de los puntajes z, obtenidos para uno de los ítems del ensayo de aptitud frente a los resultados obtenidos para el otro ítem. Una línea horizontal y una vertical son típicamente graficadas para crear cuatro cuadrantes de valores, que ayudan a la interpretación. Se trazan las líneas en los valores asignados o en las medianas para las dos distribuciones de resultados, o en 0 si se diagraman los resultados de puntajes z.

NOTA Para una apropiada interpretación de las gráficas de Youden, es importante que los dos ítems del ensayo de aptitud tengan niveles similares (o idénticos) del mensurando; es tan así, que la naturaleza de un error de medición sistemático es la misma que la del área del intervalo de medición. Las gráficas de Youden pueden ser útiles para niveles muy diferentes de un mensurando en presencia de error sistemático constante, pero pueden ser engañosas si un error de calibración no es constantemente positivo o negativo a lo largo del rango de niveles del mensurando.

10.5.3 Cuando se elabora una gráfica de Youden se interpreta así:

a) Se inspecciona en la gráfica los puntos que están muy separados del resto de los datos. Si un participante no sigue el método de ensayo correctamente, de modo que sus resultados estén sujetos al error sistemático, se encontrará un punto muy alejado en los cuadrantes inferior izquierdo o superior derecho. Tales puntos alejados de los demás,

en los cuadrantes superior izquierdo o inferior derecho, representan a los participantes cuya repetibilidad es mayor que la mayoría de los demás participantes, cuyos métodos de medición muestran sensibilidad diferente a la composición del ítem del ensayo de aptitud o, en algunas veces, participantes que han intercambiado accidentalmente los ítems del ensayo de aptitud.

b) Se inspecciona la gráfica para ver si existe evidencia de una relación general entre los resultados para los dos ítems de ensayo (por ejemplo, si la línea se ajusta aproximadamente a lo largo de la línea inclinada). Si hay evidencia de una relación, entonces muestra que hay evidencia de sesgo del participante que afecta a diferentes ítems de ensayo de aptitud de una manera similar. Si no hay una relación visual aparente entre los resultados (por ejemplo, los puntos se distribuyen aproximadamente en forma uniforme en una región circular, usualmente con mayor densidad hacia el centro), que los errores de medición para los dos ítems de ensayo de aptitud son en gran medida independientes. Esto se puede verificar con una estadística de correlación de rango, si el examen visual no es concluyente.

c) Se inspecciona la gráfica para buscar grupos cercanos de participantes, ya sea a lo largo de las diagonales o en otra parte. Es probable que grupos dispersos indiquen diferencias entre métodos.

NOTA 1 En los estudios donde todos los participantes emplean el mismo método de medición o las gráficas de resultados provienen de un método de medición único, si los resultados se encuentran a lo largo de una línea, esto puede ser evidencia de que no se ha especificado adecuadamente el método de medición. La investigación del método de ensayo podría permitir la mejora general de la reproducibilidad del método.

NOTA 2 En el Anexo E.12 se presenta un ejemplo de la gráfica de Youden.

10.6 GRÁFICAS DE DESVIACIONES ESTÁNDAR DE REPETIBILIDAD

10.6.1 Cuando los participantes realizan mediciones replicadas en una ronda de un programa de ensayos de aptitud, los resultados pueden ser utilizados para producir una gráfica que identifique los participantes cuyo promedio y desviaciones estándar son inusuales.

10.6.2 El gráfico se elabora trazando la desviación estándar intra-participante s_i para cada participante contra el promedio correspondiente xi para el participante. De manera alternativa, se puede emplear el rango de resultados replicados en lugar de la desviación estándar. Sea:

x^* = promedio robusto de $x_1, x_2, ..., x_p$, se calcula mediante el Algoritmo A

w^* = promedio combinado robusto de $s_1, s_2, ..., s_p$, se calcula mediante el Algoritmo S

y se asume que los datos tienen distribución normal. Según la hipótesis nula de que no hay diferencia entre los participantes en los valores de la población bien sea de las medias de los participantes o las desviaciones estándar dentro de cada uno de ellos, el estadístico

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

tiene aproximadamente la distribución χ^2 con 2 grados de libertad. De este modo, una región crítica con un nivel de significancia de aproximadamente 1 % se puede trazar en el gráfico al representar:

$$s = w^*\,exp\left\{\pm\frac{1}{\sqrt{2(m-1)}}\sqrt{X^2_{2;0,99} - \left(\sqrt{m}\,\frac{x - x^*}{w^*}\right)^2}\right\} \qquad (22)$$

en el eje de desviación estándar contra x en el eje promedio para:

$$x = x^* - w^* \sqrt{\frac{X^2_{2;0,99}}{m}} \quad a \quad x^* + w^* \sqrt{\frac{X^2_{2;0,99}}{m}} \qquad (23)$$

NOTA Este procedimiento se basa en la Técnica del Círculo introducida por van Nuland [36]. El método descrito utilizó una aproximación normal simple para la distribución de la desviación estándar que podría suministrar una región crítica que contiene desviaciones estándar negativas. El método que se presenta aquí utiliza una aproximación para la distribución de la desviación estándar que evita este problema, pero la región crítica ya no es un círculo como en el original. Además, se utilizan valores robustos para el punto central en lugar de los promedios simples como en el método original.

10.6.3 La gráfica puede indicar participantes con sesgo que es inusualmente grande, dada su repetibilidad. Si existe una gran cantidad de replicas, esta técnica también puede identificar los participantes con repetibilidad excepcionalmente pequeña. No obstante, puesto que por lo general existe una pequeña cantidad de replicas, las interpretaciones son difíciles.

NOTA En el Anexo E.13 se presenta un ejemplo de una gráfica de desviaciones estándar de repetibilidad.

10.7 MUESTRAS DIVIDIDAS

10.7.1 Las muestras divididas se utilizan cuando es necesario realizar una comparación detallada de dos participantes o cuando no se encuentra disponible el ensayo de aptitud y se requiere alguna verificación externa. Se obtienen muestras de varios materiales que representan un intervalo amplio de la propiedad de interés, cada muestra se divide en dos partes y cada laboratorio obtiene una cantidad (por lo menos dos) de determinaciones replicadas en cada parte de la muestra.

En ocasiones, más de dos participantes pueden estar implicados, en cuyo caso uno se debería tratar uno como referencia y los otros se deberían comparar con éste utilizando las técnicas que aquí se describen.

NOTA 1 Este tipo de estudio es común aunque a menudo se denomina de manera diferente, tal como "muestra pareada" o "comparaciones bilaterales".

NOTA 2 No se debería confundir este diseño de muestra dividida con el diseño de "nivel dividido" empleado en la NTC 3529, que incluye dos ítems de ensayo con niveles levemente diferentes suministrados a todos los participantes.

10.7.2 Los datos de un diseño de muestras divididas pueden ser usados para producir gráficas que presenten la variación entre las mediciones replicadas para los dos participantes y las diferencias entre sus resultados promedio para cada ítem del ensayo de aptitud. Las representaciones bivariadas que emplean el rango completo de concentraciones, pueden tener una escala que dificulta identificar diferencias importantes entre los participantes, de modo que pueden ser más útiles las gráficas de las diferencias o porcentajes de diferencias entre los resultados de dos participantes. Un análisis posterior dependerá de las deducciones realizadas a partir de estas gráficas.

10.8 MÉTODOS GRÁFICOS PARA COMBINAR INDICADORES DE DESEMPEÑO EN VARIAS RONDAS DE UN PROGRAMA DE ENSAYOS DE APTITUD

10.8.1 Cuando se van a combinar indicadores de desempeño normalizados en varias rondas de un programa de ensayos de aptitud, el proveedor del ensayo de aptitud puede considerar la preparación de gráficas, como se describe en los numerales 10.8.2 ó 10.8.3. El uso de estas gráficas, en las cuales se combinan los indicadores de varias rondas de un programa de ensayos de aptitud, puede permitir la identificación de tendencias y otras características de los

resultados que no son evidentes cuando se examinan por separado los indicadores de desempeño para cada ronda.

NOTA Cuando se empleen "indicadores de ejecución" o "indicadores acumulativos" en los cuales los indicadores obtenidos por un participante se combinan en varias rondas de un programa de ensayos de aptitud, los indicadores de desempeño deberían mostrarse gráficamente. El participante puede tener una falla que se evidencia con el ítem de ensayo de aptitud utilizado en una ronda pero no en las otras. Un puntaje de ejecución puede esconder esta falla. No obstante, en algunas circunstancias (por ejemplo, en rondas frecuentes) el "aplanamiento" de indicadores anómalos (atípicos) puede ser útil para demostrar el desempeño subyacente con mayor claridad.

10.8.2 La gráfica de control Shewhart es un método eficaz para identificar los problemas que causan valores erráticos grandes de los puntajes z. Consulte en la Norma ISO 7870-2 [6] consejos para trazar gráficas de Shewhart y reglas para límites de acción.

10.8.2.1 Para preparar esta gráfica, los resultados de los puntajes estandarizados s, tales como los indicadores z ó P_A, para un participante se grafican como puntos individuales, con los límites de acción y advertencia establecidos de manera coherente en el diseño del programa de ensayos de aptitud. Cuando se miden varias características en cada ronda, se pueden representar los indicadores de desempeño para diferentes características en la misma gráfica, pero los puntos para las diversas características se deberían trazar utilizando diferentes símbolos y/o diferentes colores. Cuando se incluyen varios ítems de ensayo de aptitud en la misma ronda del programa de ensayos de aptitud, los indicadores de desempeño se pueden representar junto con puntos múltiples en cada período de tiempo. También se pueden agregar a la gráfica las líneas que unen los valores medios de los puntajes en cada punto de cada periodo de tiempo.

10.8.2.2 Las reglas convencionales para interpretar los gráficos de control Shewhart indican que se presenta una señal fuera de control cuando:

a) un solo punto está por fuera de los límites de acción (± 3,0 para puntajes z ó 100 % para P_A);

b) dos de tres puntos sucesivos están por fuera del límite de advertencia (± 2,0 para puntajes z ó 70 % para P_A);

c) se presentan seis resultados consecutivos ya sea positivos o negativos.

10.8.2.3 Cuando una gráfica de control de Shewhart presenta una señal fuera de control, el participante debería investigar las causas posibles.

NOTA La desviación estándar para la evaluación de aptitud σ_{pt} por lo general no es la desviación estándar de las diferencias, $(x_i - x_{pt})$, de modo que los niveles de probabilidad que normalmente se asocian con los límites de acción y advertencia en el gráfico de control Shewhart no se pueden aplicar.

10.8.3 Cuando el nivel de una propiedad varía de una ronda de un programa de ensayos de aptitud a otra, las gráficas de indicadores de desempeño normalizados, tales como z y P_A, contra el valor asignado mostrarán si el sesgo del participante cambia con el nivel. Cuando se incluye más de un ítem de ensayo de aptitud en la misma ronda, todos los indicadores de desempeño se pueden graficar de forma independiente.

NOTE 1 Puede ser útil tener un símbolo de representación diferente o color diferente para los resultados de la ronda actual de ensayos de aptitud, a fin de distinguir el punto o los puntos de rondas anteriores.

NOTA 2 En el Anexo E.14 se presenta un ejemplo de dicha gráfica, empleando puntajes P_A. Esta gráfica podría emplear fácilmente z, con sólo un cambio en la escala vertical.

11. DISEÑO Y ANÁLISIS DE PROGRAMAS DE ENSAYOS DE APTITUD CUALITATIVOS (INCLUYE PROPIEDADES NOMINALES Y ORDINALES)

11.1 TIPOS DE DATOS CUALITATIVOS

Una gran cantidad de ensayos de aptitud se dan para propiedades que se miden o identifican sobre escalas cualitativas. Lo cual incluye lo siguiente:

- Los programas de ensayos de aptitud que requieren informes sobre una escala categórica (algunas veces denominada "nominal"), donde el valor de la propiedad no tiene magnitud (tal como un tipo de sustancia u organismo);

- Los programas de ensayos de aptitud para presencia o ausencia de una propiedad, sea que se determine por criterios subjetivos o por la magnitud de una señal de un procedimiento de medición. Éste puede considerarse como un caso especial de una escala categórica u ordinal, con sólo dos valores (también denominados "dicotómicos" o binarios);

- Los programas de ensayos de aptitud que requieren informes de resultados sobre una escala ordinal, que se pueden ordenar de acuerdo con la magnitud, pero para la cual no existe relación aritmética entre resultados diferentes. Por ejemplo, "alto, medio y bajo" conforman una escala ordinal.

Dichos programas de ensayos de aptitud requieren consideración especial para las etapas de diseño, determinación del valor asignado y evaluación del desempeño (indicador) debido a que:

- con mucha frecuencia los valores asignados se basan en la opinión de expertos; y

- el tratamiento estadístico diseñado para datos de valores continuos y conteo de datos no es aplicable a datos cualitativos. Por ejemplo, no resulta significativo tomar medias y desviaciones estándar de resultados de escala ordinal incluso cuando se coloquen en un orden de clasificación.

De manera concordante, los siguientes párrafos ofrecen orientación sobre el diseño, la determinación del valor asignado y la evaluación del desempeño para programas de ensayos de aptitud cualitativos.

NOTA La orientación sobre datos ordinales no se aplica a resultados de la medición que se basan en una escala cuantitativa con indicaciones discontinuas (tales como diluciones o titulaciones); véase el numeral 5.2.2.

11.2 DISEÑO ESTADÍSTICO

11.2.1 Para programas de ensayos de aptitud en los cuales es esencial la opinión de un experto, ya sea para la determinación del valor asignado o para la evaluación de informes de los participantes, por lo general, será necesario conformar un panel de expertos calificados apropiadamente y disponer de tiempo para el debate a fin de lograr consenso sobre la asignación apropiada. Cuando haya necesidad de confiar en expertos individuales para la calificación o determinación del valor asignado, el proveedor de ensayos de aptitud debería adicionalmente, prever la evaluación y el control de la coherencia de la opinión entre expertos diferentes.

EJEMPLO En un programa de ensayos de aptitud clínico que depende del microscopio para el diagnóstico, se emplea la opinión de expertos para evaluar los portaobjetos del microscopio provistos a los participantes y entregar un diagnóstico apropiado para los ítems del ensayo de aptitud. El proveedor del ensayo de aptitud puede optar por hacer circular los ítems de ensayo "ciegos" para diferentes miembros del panel de expertos a fin de asegurar la consistencia del diagnóstico o realizar ejercicios periódicos para evaluar el acuerdo entre el panel.

11.2.2 Para programas de ensayos de aptitud que reportan resultados simples, valores únicos categóricos o resultados ordinales, el proveedor de ensayos de aptitud debería considerar:

- entregar dos ó más ítems de ensayos de aptitud por ronda: o

- solicitar los resultados de un número de observaciones replicadas para cada ítem de ensayo de aptitud con la cantidad de replicas especificadas por anticipado.

Cualquiera de estas estrategias permite contar los resultados por cada participante que se pueden emplear en la revisión de datos o en la asignación del puntaje. El suministro de dos ó más ítems de ensayo de aptitud puede ofrecer información adicional sobre la naturaleza de los errores y además permitir una puntuación más sofisticada del desempeño del ensayo de aptitud.

EJEMPLO 1 En un programa de ensayos de aptitud destinado a informar sobre la presencia o ausencia de un contaminante, la provisión de ítems de ensayo con un rango de niveles del contaminante permite que el proveedor del ensayo examine la cantidad de detecciones exitosas en cada nivel como una función del nivel de contaminante presente. Esto puede, por ejemplo, servir para proporcionar información a los participantes sobre la capacidad de detección de su método de ensayo seleccionado u obtener una probabilidad promedio de detección que puede, a su vez, permitir distribuir indicadores de desempeño a los participantes con base en las probabilidades estimadas de patrones de respuesta particulares.

EJEMPLO 2 El ensayo de aptitud en comparaciones forenses con frecuencia requiere verificar la concordancia de ítems de ensayo de aptitud en relación a si provienen de la misma o diferente fuente (por ejemplo, huellas dactilares, ADN, vainas de balas, huellas de los pies, etc.). En muchos casos, "indeterminado" es una respuesta permitida. Un programa de ensayos de aptitud podría incluir múltiples ítems de ensayo de aptitud de fuentes diferentes y se pide a los participantes establecer cuáles son de la "misma fuente", de "fuente diferente" o "indeterminada" por cada par. Esto permite puntajes objetivos de número (ó %) correcto ó incorrecto, o número (%) de correspondencias correctas o rechazos correctos. Entonces se pueden determinar criterios de desempeño sobre adecuación para el uso o sobre el grado de dificultad del reto.

11.2.3 Se debería demostrar la homogeneidad con la revisión de una muestra apropiada de ítems de ensayo de aptitud, todos los cuales deberían demostrar el valor de la propiedad esperado. Para algunas propiedades cualitativas, por ejemplo, presencia o ausencia, es posible verificar la homogeneidad con mediciones cuantitativas; por ejemplo, un recuento microbiológico o una absorbancia de espectro por encima de un umbral. En estas situaciones, puede ser apropiado un ensayo convencional de homogeneidad o una demostración de que todos los resultados están por encima o por debajo de un valor límite.

11.3 VALORES ASIGNADOS PARA PROGRAMAS DE ENSAYOS DE APTITUD CUALITATIVOS

11.3.1 Se pueden asignar valores a ítems de ensayos de aptitud:

a) por juicio de expertos;

b) mediante el uso de materiales de referencia como ítems de ensayo de aptitud;

c) a partir del conocimiento del origen o preparación del ítem (o ítems) de ensayo de aptitud;

d) empleando la moda o la mediana de los resultados de los participantes (la mediana es apropiada sólo para valores ordinales).

También se puede emplear cualquier otro método para la determinación del valor asignado que pueda demostrar que brinda resultados confiables. En los siguientes párrafos se considera cada una de las anteriores estrategias.

NOTA Por lo general, no es apropiado ofrecer información cuantitativa con respecto a la incertidumbre del valor asignado en programas de ensayos de aptitud cualitativos. No obstante, cada uno de los numerales 11.3.2 a 11.3.5 exige el suministro de información básica en relación con la confianza en el valor asignado de modo que los participantes puedan juzgar si un resultado deficiente pudiera ser atribuible de forma razonable a un error en el valor asignado.

11.3.2 Normalmente, los valores asignados por opinión de expertos deberían basarse en un consenso de un panel de expertos calificados adecuadamente. En el informe de la ronda se debería registrar cualquier desacuerdo significativo entre los miembros del panel. Si el panel no logra un consenso para un ítem de ensayo de aptitud particular, el proveedor de ensayos de aptitud puede considerar un método alternativo de determinación del valor asignado con respecto a los enunciados en el numeral 11.3.1. Si esto no resulta apropiado, el ítem del ensayo de aptitud no debería emplearse para la evaluación del desempeño de los participantes.

NOTA En algunos casos es posible que un único experto determine el valor asignado.

11.3.3 Cuando se proporciona un material de referencia a los participantes como ítem de ensayo de aptitud, normalmente se debería emplear el valor de referencia asociado o el valor certificado como el valor asignado para la ronda. Cualquier información de resumen suministrada con el material de referencia que se relacione con la confianza en el valor asignado debería estar disponible para los participantes después de la ronda.

NOTA En el numeral 7.4.1 se enuncian las limitaciones de este enfoque.

11.3.4 Cuando se preparan los ítems de ensayo de aptitud a partir de una fuente conocida, el valor asignado se puede determinar con base en el origen del material. El proveedor de ensayo de aptitud debería mantener registros del origen, transporte y manejo del material o materiales empleados. Se debe tener el debido cuidado para evitar la contaminación que podría causar resultados incorrectos de los participantes. La evidencia del origen y/o detalles de la preparación deberían estar disponibles para los participantes después de la ronda o a solicitud o como parte del informe de la ronda del ensayo de aptitud.

EJEMPLO Los ítems de ensayo de aptitud de vino, circulados para un programa de ensayos de aptitud para autenticidad, pueden obtenerse de forma directa de un productor adecuado en la región de origen designada, o por medio de un proveedor comercial capaz de suministrar garantía de autenticidad.

11.3.4.1 Se recomiendan ensayos o mediciones confirmatorias en la medida de lo posible, en especial cuando la contaminación puede comprometer el uso como ítem de ensayo de aptitud. Por ejemplo, un ítem de ensayo de aptitud identificado como un ejemplar de una especie única microbiana, vegetal o animal debería ensayarse normalmente, en relación con su respuesta a ensayos para otras especies pertinentes. Dichos ensayos deberían ser tan sensibles como sea posible para asegurar que la especie contaminante esté ausente o que el nivel de contaminación se cuantifique.

11.3.4.2 El proveedor de ensayos de aptitud debería proporcionar información sobre cualquier contaminación detectada o dudas sobre el origen que pudieran comprometer el uso del ítem de ensayo de aptitud.

NOTA Está por fuera del alcance de esta norma proveer detalles adicionales sobre la caracterización de dichos ítems de ensayos de aptitud.

11.3.5 La moda (que es la observación más común) se puede emplear como el valor asignado para resultados sobre una escala categórica u ordinal, mientras que la mediana puede usarse como el valor asignado para resultados en una escala ordinal. Cuando se emplean estas estadísticas, el informe de la ronda de ensayos de aptitud debería incluir una declaración de la proporción de los resultados empleados para la determinación del valor que encajan con el

valor asignado. Nunca es apropiado calcular medias o desviaciones estándar para los resultados de ensayos de aptitud de propiedades cualitativas, incluidos valores ordinales. Esto se debe a que no existe una relación aritmética entre valores diferentes en cada escala.

11.3.6 Cuando los valores asignados se basan en mediciones (por ejemplo, la presencia o ausencia), por lo general, el valor asignado puede estar determinado de forma definitiva, es decir, con baja incertidumbre. Los cálculos estadísticos de incertidumbre pueden resultar apropiados para niveles de mensurando en niveles "indeterminados" o "equívocos".

11.4 EVALUACIÓN DEL DESEMPEÑO Y ASIGNACIÓN DE PUNTAJES PARA PROGRAMAS DE ENSAYOS DE APTITUD CUALITATIVOS

11.4.1 La evaluación del desempeño de los participantes en un programa de ensayos de aptitud cualitativo 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 y las conclusiones exigen consideración y redacción cuidadosa, se pueden pasar los informes de los participantes a los expertos para su valoración y se les puede dar una nota general. En el otro extremo, se puede juzgar únicamente a los participantes con base en si su resultado coincide exactamente con el valor asignado para el ítem de ensayo de aptitud pertinente. De manera concordante, los siguientes parágrafos ofrecen orientación sobre evaluación de desempeño y asignación de puntajes para una serie de circunstancias.

11.4.2 La valoración de los informes de los participantes por parte de expertos exige que uno ó más individuos expertos revisen el informe de cada participante por cada ítem de ensayo de aptitud y asignen una nota o puntaje. En dicho programa de ensayos de aptitud, el proveedor debería garantizar que:

- el participante particular no se conozca con el experto. En especial, el informe remitido al experto o a los expertos no debería incluir ninguna información que pudiera permitir identificar al participante;

- la revisión, calificación y evaluación de desempeño sigan un conjunto de criterios acordados previamente que sean objetivos en la medida de lo posible;

- se cumplan las disposiciones del numeral 11.3.2 con respecto a la coherencia entre expertos;

- siempre que sea factible, se prevea la posibilidad para que los participantes apelen contra la opinión de un experto en particular y/o para revisión secundaria de opiniones cercanas a cualquier umbral de desempeño importante.

11.4.3 Se pueden emplear dos sistemas para calificar un único resultado cualitativo reportado con base en un valor asignado:

i) Cada resultado se califica como aceptable (o se le da un puntaje como exitoso) si corresponde exactamente con el valor asignado y en caso contrario, se califica como inaceptable o se le da un indicador de desempeño negativo.

EJEMPLO En un programa para determinar la presencia o ausencia de un contaminante, los resultados correctos se califican con 1 y los incorrectos con 0.

ii) Los resultados que corresponden exactamente con el valor asignado se califican como aceptables y se les da el puntaje correspondiente. Los resultados que no corresponden exactamente, se les da un puntaje que depende de la naturaleza de la falta de correspondencia. Con tales diseños de asignación de puntaje, se debería calificar con

puntajes inferiores al desempeño mejor, para ser coherentes con otros tipos de puntaje de desempeño (por ejemplo, puntaje z, puntaje P_A, ζ y E_n).

EJEMPLO 1 En un programa de ensayos de aptitud para patología clínica, un proveedor de ensayos de aptitud asigna un puntaje de '0' para una identificación exactamente correcta de una especie microbiológica, '1' punto para un resultado que sea incorrecto pero no alteraría el tratamiento clínico (por ejemplo, la identificación como especie microbiológica diferente pero relacionada que exige tratamiento similar) y 3 puntos para una identificación que es incorrecta y conduciría a un tratamiento incorrecto de un paciente. Por lo general, el esquema de asignación de puntaje exigirá el juicio de expertos sobre la naturaleza del error en correspondencia, tal vez obtenido antes de fijar el puntaje.

EJEMPLO 2 En un programa de ensayos de aptitud para el cual son posibles seis respuestas clasificadas en una escala ordinal, a un resultado que corresponda con el valor asignado se le da un puntaje de 0 y se incrementa en 2 unidades por cada diferencia en la calificación hasta alcanzar un máximo de 6 (de modo que un resultado adyacente al valor asignado implicaría un puntaje de 2).

Se debería brindar a los participantes indicadores de desempeño individuales por cada ítem de ensayo de aptitud. Cuando se realizan observaciones replicadas, se puede proporcionar un resumen de indicadores de desempeño por cada resultado.

11.4.4 Cuando se informan múltiples replicas por cada ítem de ensayo de aptitud o se proporcionan múltiples ítems de ensayo de aptitud a cada participante, el proveedor del ensayo aptitud puede calcular y emplear indicadores de desempeño combinados o resúmenes del puntaje en la evaluación de desempeño. Se pueden calcular los indicadores de desempeño combinados o los resúmenes, por ejemplo, de la siguiente manera:

- la simple suma de los resultados de los indicadores de desempeño a través de todos los ítems de ensayo de aptitud;
- el conteo de cada nivel de desempeño asignado;
- la proporción de resultados correctos;
- una métrica de distancia basada en las diferencias entre resultados y valores asignados.

EJEMPLO Una estadística métrica de distancia muy general que a veces se emplea es el coeficiente de Gower [20]. Este puede combinar variables cuantitativas y cualitativas con base en una combinación de puntajes para similitud. Para datos categóricos o binarios, el índice asigna un puntaje de 1 para categorías de correspondencia exacta y 0 para las demás; para escalas ordinales asigna un puntaje igual a 1 menos la diferencia en calificación dividido por el número de calificaciones disponibles, y para datos de intervalo o de escala de relación asigna un puntaje de 1 menos la diferencia absoluta dividida por el rango observado de todos los valores. Se suman estos puntajes, que necesariamente van todos de 0 a 1 y se divide la sumatoria por el número de variables empleadas. También se puede emplear una variante ponderada.

Los indicadores de desempeño combinados pueden asociarse con una evaluación de desempeño de resumen. Por ejemplo, una proporción particular (por lo general elevada) de puntajes de indicadores correctos puede considerarse un desempeño "aceptable" si es coherente con los objetivos del programa de ensayos de aptitud.

11.4.5 Se pueden emplear métodos gráficos para proporcionar información de desempeño a los participantes o brindar información de resumen en un informe de una ronda.

NOTA En el Anexo E.15 se presenta un ejemplo del análisis de datos ordinales.

ANEXO A
(Normativo)

SÍMBOLOS

d	Diferencia entre un valor de medición para un ítem de ensayo de aptitud y un valor asignado para un MRC
$\bar{d}$	Diferencia promedio entre valores de medición y el valor asignado para un MRC
D	Diferencia de los participantes a partir del valor asignado (x-x_{pt})
$D\%$	Diferencia de los participantes a partir del valor asignado expresado como un porcentaje de x_{pt}
$¢_E$	Criterio de error máximo permisible para las diferencias
$¢_{hom}$	Error debido a la diferencia entre ítems de ensayo de aptitud
$¢_{stab}$	Error debido a la inestabilidad durante el período del ensayo de aptitud
δ_{trans}	Error debido a la inestabilidad bajo condiciones de transporte
E_n	Indicador de “error, estandarizado” que incluye incertidumbres para el resultado del participante y el valor asignado
g	Cantidad de ítems de ensayo de aptitud ensayados en una comprobación de homogeneidad
m	Cantidad de mediciones repetidas efectuadas por ítem de ensayo de aptitud
p	Cantidad de participantes que participan de una ronda de un programa de ensayos de aptitud
P_A	Proporción de error permitido (D/δ_E), se puede expresar como un porcentaje
s_r	Estimado de la desviación estándar de repetibilidad
s_R	Estimado de la desviación estándar de reproducibilidad
s_s	Estimado de la desviación estándar entre muestras
s^*	Estimado robusto de la desviación estándar del participante
$s_{\bar{x}}$	Desviación estándar de promedios de muestra
s_w	Desviación estándar dentro de la muestra o dentro del laboratorio
σ_k	Desviación estándar de ancho de banda empleada para gráficas de densidad kernel
σ_L	Desviación estándar interlaboratorio (o entre participantes)
σ_{pt}	Desviación estándar para evaluación de la aptitud
σ_r	Desviación estándar de repetibilidad
σ_R	Desviación estándar de reproducibilidad
u_{hom}	Incertidumbre estándar debida a la diferencia entre ítems de ensayo de aptitud
u_{stab}	Incertidumbre estándar debida a la inestabilidad durante el período de ensayos de aptitud
u_{trans}	Incertidumbre estándar debida a la inestabilidad bajo condiciones de transporte
$u(x_i)$	Incertidumbre estándar de un resultado del participante i
$u(x_{pt})$	Incertidumbre estándar del valor asignado
$u(x_{ref})$	Incertidumbre estándar de un valor de referencia
$U(x_i)$	Incertidumbre expandida de un resultado reportado del participante i

$U(x_{pt})$	Incertidumbre expandida del valor asignado
$U(x_{ref})$	Incertidumbre expandida de un valor de referencia
w_t	Rango entre porción de ensayo
w^*	Estimado robusto para repetibilidad del participante
x	Resultado de medición (genérico)
x_{char}	Valor de la propiedad obtenido a partir de la determinación del valor asignado
x_{MRC}	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 establecido
x^*	Estimado robusto de la media del participante
$\overline{X}$	Promedio aritmético de un conjunto de resultados
z	Puntaje empleado para la evaluación de aptitud
z'	Puntaje z modificado que incluye la incertidumbre del valor asignado
ζ	Puntaje zeta -puntaje z modificado que incluye las incertidumbres para el resultado del participante y para el valor asignado

ANEXO B
(Normativo)

HOMOGENEIDAD Y ESTABILIDAD DE ÍTEMS DE ENSAYO DE APTITUD

B.1 PROCEDIMIENTO GENERAL PARA UNA COMPROBACIÓN DE HOMOGENEIDAD

B.1.1 Para realizar una evaluación de homogeneidad para una preparación a granel de ítems de ensayo de aptitud, se sigue el procedimiento que se indica a continuación:

Escoja una propiedad (o propiedades) o mensurando(s) para evaluarlos con la comprobación de homogeneidad.

Seleccione un laboratorio para realizar la comprobación de homogeneidad y el método de medición que se va a emplear. El método debería tener una desviación estándar de repetibilidad lo suficientemente pequeña (s_r) para que se pueda detectar cualquier falta de homogeneidad que pueda ser significativa. La relación de la desviación estándar de repetibilidad para el método con la desviación estándar para la evaluación de aptitud debería ser inferior a 0,5, como lo recomienda el Protocolo Armonizado IUPAC (ó 1/6 de $¢_E$). Se reconoce que esto no siempre es posible; por lo tanto, en dicho caso, el proveedor de ensayos de aptitud debería emplear más replicas.

Prepare y empaque los ítems del ensayo de aptitud para una ronda del programa de ensayos de aptitud, asegurándose que hay suficientes ítems para los participantes del programa y para la comprobación de homogeneidad.

Seleccione un número g de los ítems de ensayo de aptitud en sus empaques finales empleando un proceso de selección aleatoria adecuado, donde $g \geq 10$. El número de ítems de ensayo de aptitud que se incluyen en la comprobación de homogeneidad se puede reducir si se cuenta con datos adecuados de comprobaciones previas de la homogeneidad en ítems similares, preparados con los mismos procedimientos.

Prepare dos o más porciones de ensayo $m \geq 2$ de cada ítem de ensayo de aptitud, utilizando técnicas adecuadas para el ítem de ensayo con el fin de minimizar las diferencias entre las porciones de ensayo.

Tomando las porciones para ensayo $g \times m$ en orden aleatorio, obtenga un resultado de la medición para cada una, completando toda la serie de mediciones en condiciones de repetibilidad.

Calcule el promedio general $\overline{X}$, la desviación estándar intramuestras s_w y la desviación estándar entre las muestras s_s, como se indica en el literal B.3.

B.1.2 Cuando no es posible realizar mediciones replicadas, por ejemplo en ensayos destructivos, entonces se puede emplear la desviación estándar de los resultados como s_s. En esta situación es importante contar con un método con una desviación estándar de repetibilidad suficientemente baja s_r.

B.2 CRITERIOS GENERALES PARA UNA COMPROBACIÓN DE HOMOGENEIDAD

B.2.1 Se deberían emplear las siguientes tres comprobaciones para asegurarse de que los datos de ensayo de homogeneidad son validos para el análisis:

a) Examine los resultados por cada porción de ensayo en el orden de la medición para buscar una tendencia (o desviación) en el análisis; si existe una tendencia aparente, se emprenden las acciones correctivas apropiadas en relación con el método de medición, o se tiene cuidado en la interpretación de los resultados.

b) Examine los resultados promedio de los ítems de ensayo de aptitud por orden de producción; si existe una tendencia grave que hace que el ítem supere el criterio del literal B.2.2 o de otro modo prevenga el uso del ítem, entonces (i) se asignan valores individuales a cada ítem de ensayos de aptitud ó (ii) se descarta el subconjunto de ítems de ensayo de aptitud significativamente afectados y vuelva a ensayar los ítems restantes para lograr una homogeneidad suficiente; ó (iii) si la tendencia afecta todos los ítems de ensayo de aptitud, siga las disposiciones del literal B.2.4.

c) Compare a la diferencia entre replicas (o rango, si hay más de 2 replicas) y, de ser necesario, ensaye en búsqueda de una diferencia estadísticamente significativa, entre replicas, empleando la prueba de Cochran (NTC 3529-2). Si la diferencia entre replicas es grande para cualquier pareja, revise la explicación técnica para la diferencia y, si resulta apropiada, elimine el grupo de datos atípico del análisis o, si $m > 2$ y la alta varianza es causada por un único valor atípico, elimine el punto atípico (*outlier*).

NOTA Si $m > 2$ y se elimina una única observación, el cálculo posterior de s_w y s_s necesitará tener en cuenta el desequilibrio resultante.

B.2.2 Compare la desviación estándar s_s entre muestras con la desviación estándar para la evaluación de la aptitud σ_{pt}. Los ítems de ensayo de aptitud se pueden considerar adecuadamente homogéneos si:

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

NOTA 1 La justificación para el factor de 0,3 es que cuando se cumple este criterio, la desviación estándar entre muestras aporta menos del 10 % de la varianza para evaluación de desempeño, de modo que no es probable que la evaluación de desempeño se vea afectada.

NOTA 2 De manera equivalente, se puede comparar s_s con $¢_E$:

$$s_s \leq 0{,}1¢_E \qquad \text{(B.2)}$$

B.2.3 Puede resultar útil extender el criterio para permitir el error de muestreo real y la repetibilidad en la comprobación de homogeneidad. En estos casos, se siguen los siguientes pasos:

a) Calcule $\sigma^2_{permitido} = (0{,}3\ \sigma_{pt})^2$

b) Calcule $c = F_1\sigma^2_{permitido} + F_2 s^2_{\text{¬}\nu}$, donde

s_w es la desviación estándar intramuestra calculada según el literal B.3 y

F_1 y F_2 provienen de las tablas estadísticas, reproducidas en la Tabla B.1, para la cantidad de ítems de ensayo de aptitud seleccionados y con cada ítem ensayado por duplicado [33].

Tabla B.1 Factores *F1 y F2* para uso en ensayos 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

Donde $m > 2$, F_2 en el literal B.2.3 b) y la Tabla B.1 se debe remplazar con $F_{2m} = (F_{g\text{-}1,\ g(m\text{-}1),\ 0{,}95\text{-}1})/m$ donde $F_{g\text{-}1,\ g(m\text{-}1),\ 0{,}95\ -\ 1}$ es el valor crítico con probabilidad 0,05 para 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 las tablas estadísticas estándar, de la siguiente manera:

$F_1 = \chi^2_{0{,}95(g\ -\ 1)}$ donde $\chi^2_{0{,}95(g\ -\ 1)}$ es el valor crítico con probabilidad de 0,05 para una variable aleatoria de distribución de Chi cuadrado con *g* - 1 grados de libertad, y

$F_2 = (F_{0{,}95\ (g\text{-}\ 1;g)}\text{-}1)/2$ donde $F_{0{,}95(g\text{-}1;g)}$ es el valor crítico con probabilidad de 0,05 para una variable aleatoria con distribución *F* con *g* - 1 y *g* grados de libertad.

c) Si $s_s > \sqrt{c}$, entonces existe evidencia de que el lote de ítems de ensayos de aptitud no es suficientemente homogéneo.

B.2.4 Cuando no se conoce σ_{pt} por anticipado; por ejemplo, cuando σ_{pt} es la desviación estándar robusta de los resultados de los participantes, el proveedor del ensayo de aptitud debería seleccionar otros criterios para determinar la homogeneidad suficiente. Dichos procedimientos podrían incluir:

a) compruebe las diferencias estadísticamente significativas entre los ítems de ensayo de aptitud utilizando, por ejemplo, la prueba *F* del Análisis de Varianza con ~ = 0,05;

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

c) emplee los datos de un experimento de precisión (tales como una desviación estándar de reproducibilidad, descrita en la NTC 3529-2 (ISO 5725-2);

d) acepte el riesgo de que los ítems de ensayo de aptitud distribuidos no son suficientemente homogéneos, y compruebe el criterio después de haber calculado el σ_{pt} por consenso.

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

a) Incluir la desviación estándar entre muestras en la desviación estándar para la evaluación de aptitud, calculando σ'_{pt} como en la ecuación (B.3). Observe que esto necesita describirse por completo a los participantes.

$$\pm'_{pt} = \sqrt{\pm^2_{pt} + s_s^2} \qquad \text{(B.3)}$$

b) Incluir s_s en la incertidumbre del valor asignado y emplear z' ó $¢_E'$ para evaluar el desempeño (véase el numeral 9.5);

c) Cuando σ_{pt} es la desviación estándar robusta de los resultados de los participantes, entonces la falta de homogeneidad entre los ítems de ensayo de aptitud se incluye en σ_{pt} y entonces se puede flexibilizar, con precaución, el criterio para aceptabilidad de homogeneidad.

Si ninguno de los literales a) hasta c) aplican, se descarta el ítem de ensayo de aptitud y se repite la preparación después de corregir la causa de la falta de homogeneidad.

B.3 FÓRMULAS PARA COMPROBACIÓN DE HOMOGENEIDAD

Las estimaciones de las desviaciones estándar intramuestra s_w y entre muestras s_s se pueden calcular empleando el análisis de varianza como se muestra a continuación. El método mostrado es para un número seleccionado g de ítems de ensayo de aptitud, medidos en forma de replicas m veces.

Los datos de una comprobación de homogeneidad se representan para $x_{t,k}$

en donde,

t representa el ítem de ensayo de aptitud (t = 1,2……., g)

k representa la porción de ensayo (k = 1,2….., m)

Se define el promedio y la varianza del ítem de ensayo de la siguiente manera:

$$\overline{X}_t = \frac{1}{m}\sum_{k=1}^{m} X_k$$
$$s_t^2 = \frac{1}{m}\sum_{k=1}^{m}\left(X_k - \overline{X}_t\right)^2 \qquad \text{(B.4)}$$

y el estimado de varianza entre porciones de ensayo como:

$$w_t^2 = \frac{1}{(m-1)}\sum_{k=1}^{m}\left(X_k - \overline{X}_t\right)^2 \qquad \text{(B.5)}$$

Se calcula el promedio general:

$$\overline{\overline{X}} = \frac{1}{g}\sum_{t=1}^{g}\overline{X}_t \qquad \text{(B.6)}$$

el estimado de la varianza de promedios de muestra:

$$s_x^2 = \frac{1}{(g-1)}\sum_{t=1}^{g}\left(\overline{X}_t - \overline{\overline{X}}\right)^2 \qquad \text{(B.7)}$$

y la varianza intramuestral:

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

Estime la varianza combinada de *ss* y *sw*

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

Por último, estime la varianza entre muestras como:

$$s_s^2 = s_{s,w}^2 - s_w^2 = \frac{1}{(g-1)}\sum_{t=1}^{g}\left(\overline{X}_t - \overline{\overline{X}}\right)^2 - \frac{1}{m}s_w^2 \qquad \text{(B.10)}$$

NOTA En el caso que $s_s^2 < 0$, entonces resulta apropiado emplear $s_s = 0$.

Para un diseño común donde *m* es 2, se pueden emplear las siguientes fórmulas.

Se definen los promedios de la muestra como:

$$\overline{X}_t = \left(X_{t,1} + X_{t,2}\right)/2 \qquad \text{(B.11)}$$

– y los rangos entre porciones de ensayo como:

$$w_t = \left|X_{t,1} - X_{t,2}\right| \qquad \text{(B.12)}$$

Se calcula el promedio general:

$$\overline{\overline{X}} = \frac{1}{g}\sum_{t=1}^{g}\overline{X}_t \qquad \text{(B.13)}$$

Estime la desviación estándar de promedios de muestra

$$s_X = \sqrt{\sum_{t=1}^{g}\left(\overline{X}_t - \overline{\overline{X}}\right)^2/(g-1)} \qquad \text{(B.14)}$$

y la desviación estándar intramuestral:

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

en donde las sumatorias de las fórmulas B.13, B.14 y B.15 son sobre las muestras (t = 1, 2, ..., g).

Por último, estime la desviación estándar entre muestras como:

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

NOTA 1 El estimado de la varianza entre muestras s_s^2 con frecuencia se vuelve negativo cuando s_s es relativamente menor que s_w. Se puede esperar esto cuando los ítems de ensayo de aptitud son bastante homogéneos. En este caso $s_s = 0$.

NOTA 2 En lugar de emplear rangos, se podría emplear desviaciones estándar entre porciones ensayo tales como:

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

NOTA 3 En el Anexo E.2 se presenta un ejemplo.

B.4 PROCEDIMIENTOS PARA COMPROBAR LA ESTABILIDAD

B.4.1 CONSIDERACIONES GENERALES PARA COMPROBAR LA ESTABILIDAD

En los siguientes numerales se ofrece orientación para cumplir los requisitos de estabilidad incorporados en el numeral 6.1. Las disposiciones con respecto a las propiedades que se van a estudiar en cualquier comprobación experimental de estabilidad de acuerdo con la duración de la ronda de ensayos de aptitud y de estabilidad durante el transporte se presentan en el numeral 6.1.3.

B.4.1.1 Cuando existe una garantía razonable de estudios experimentales anteriores o conocimiento previo de que la inestabilidad es poco probable, las comprobaciones experimentales de estabilidad se pueden limitar a una verificación de cambio significativo en el curso de la ronda de ensayos de aptitud, realizada durante y después de la misma ronda. En otras circunstancias, los estudios de los efectos de transporte y estabilidad por la duración típica de una ronda de ensayos de aptitud pueden tomar la forma de estudios planeados antes de hacer circular los ítems de ensayo de aptitud, por cada ronda de ensayos de aptitud o durante la planeación inicial y los estudios de factibilidad para establecer condiciones consistentes de transporte y almacenamiento. Los proveedores de ensayos de aptitud también pueden comprobar la evidencia de inestabilidad revisando los resultados reportados para una tendencia con fecha de medición.

B.4.1.2 Las siguientes consideraciones se aplican a comprobaciones de estabilidad:

- Todas las propiedades que se emplean en el programa de ensayos de aptitud se deberían revisar o comprobar de otra manera su estabilidad. Esto se puede lograr con experiencia previa y justificación técnica con base en el conocimiento de la matriz (o artefacto) y el mensurando.

- Se deberían ensayar más de 2 ítems de ensayo de aptitud si la variabilidad entre los ítems es grande; y se deberían emplear más muestras o más replicas si se sospecha de la repetibilidad (por ejemplo, si s_w ó $s_r > 0{,}5\ \sigma_{pt}$).

NOTA En la Guía ISO 35 se presentan estrategias para minimizar el efecto en estudios de variación de estabilidad de largo plazo en el proceso de medición, tal como estudios isócronos 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 conveniente para ensayar la estabilidad en el ensayo de aptitud es tomar una pequeña muestra de ítems de ensayo de aptitud al finalizar una ronda y comparar estos ítems ensayados con los resultados obtenidos antes de la ronda, para asegurar que no se presentaron cambios durante el tiempo de la ronda. La verificación puede incluir una revisión de cualquier efecto o condición de transporte por exposición adicional de los ítems de ensayo de aptitud retenidos para la duración del estudio en condiciones que representen las condiciones de transporte. Para estudios exclusivamente destinados a comprobar los efectos

del transporte, la comparación se realiza entre ítems de ensayos de aptitud que son despachados con ítems de ensayo que se retienen bajo condiciones controladas.

NOTA 1 Los proveedores de ensayos de aptitud pueden emplear los resultados de los ensayos de homogeneidad antes de la ronda de ensayos de aptitud en lugar de seleccionar y medir un conjunto separado de ítems de ensayo de aptitud.

NOTA 2 Este modelo aplica igualmente a los programas de ensayos de aptitud, tanto en ensayos como en calibración.

B.4.2.2 Si un proveedor de ensayos de aptitud incluye los ítems de ensayo de aptitud enviados en la evaluación de estabilidad descrita en el literal B.4.2.1, entonces los efectos del transporte están incluidos en la evaluación de estabilidad. Si se verifican los efectos del transporte por separado, entonces se debería emplear el procedimiento descrito en el literal B.6.

B.4.2.3 El siguiente es un procedimiento para una comprobación de estabilidad básica empleando mediciones antes y después de una ronda de ensayos de aptitud:

a) Seleccione aleatoriamente un numero 2*g* de ítems de ensayo de aptitud, donde $g \geq 2$.

b) Seleccione un solo laboratorio empleando un método de medición único con buena precisión intermedia.

c) Mida los ítems de ensayo de aptitud *g* antes de la fecha planeada de distribución de los ítems a los participantes. Se deberían realizar mediciones replicadas en un orden completamente aleatorio.

d) Reserve los ítems de ensayo de aptitud restantes *g* bajo condiciones similares a las condiciones de almacenamiento esperadas en las instalaciones de los participantes.

e) Tan pronto como sea posible después de la fecha de cierre para entregar los resultados de los participantes, mida los ítems de ensayo de aptitud *g* restantes, empleando el mismo laboratorio, método de medición y cantidad de replicas como las del literal a) anterior, con todos los replicados en orden aleatorio.

f) Calcule 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 pueden emplear las siguientes variaciones al procedimiento del literal B.4.2.3:

a) Se puede omitir el primer grupo de ítems de ensayo de aptitud *g* si se cuenta con otras mediciones sobre el conjunto de ítems de ensayo de aptitud del mismo laboratorio y método de ensayo. Por ejemplo, se pueden emplear datos de una verificación de homogeneidad anterior.

b) Es posible utilizar condiciones que pueden acelerar el cambio para garantizar una mayor estabilidad.

c) Adicionalmente, se puede someter el segundo conjunto de ítems de ensayo de aptitud a condiciones esperadas en el despacho a fin de incluir una prueba sobre el efecto del envío.

d) Se pueden emplear otro diseño y condiciones que, junto con el criterio de comprobación de estabilidad seleccionado ofrecen igual o mayor seguridad en la estabilidad.

B.5 CRITERIO DE EVALUACIÓN PARA UNA COMPROBACIÓN DE ESTABILIDAD

B.5.1 Se compara el promedio general de las mediciones obtenidas en la comprobación antes de la distribución con el promedio general de los resultados obtenidos en la comprobación de estabilidad. Los ítems de ensayo de aptitud pueden considerarse adecuadamente estables si:

$$\left|\bar{y}_1 - \bar{y}_2\right| \leq 0{,}3\pm_{pt} \quad o \leq 0{,}1¢_E \qquad \text{(B.17)}$$

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

a) emplee un estudio de estabilidad isocrónica (véase la Guía 35 de la ISO);

b) incremente la incertidumbre del valor asignado para tener en cuenta la posible inestabilidad;

c) expanda el criterio de aceptación adicionando la incertidumbre de la diferencia para σ_{pt} con la siguiente fórmula:

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

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

B.5.3 Si no se cumple el criterio de las ecuaciones (B.17) ó (B.18) se deberían considerar las siguientes opciones:

- cuantifique el efecto de la inestabilidad y tengalo en cuenta en la evaluación (por ejemplo con puntajes z'); o

- examine la preparación del ítem de ensayo de aptitud y los procedimientos de almacenamiento para ver si hay mejoras posibles; o

- no evalúe el desempeño de los participantes.

B.5.4 El criterio de B.5.1 ó B.5.2 se puede remplazar por una prueba estadística apropiada para una diferencia entre dos conjuntos de datos siempre que la prueba tenga en cuenta la debida replicación y ofrezca la seguridad de identificar la estabilidad por lo menos igual a la proporcionada por la ecuación (B.18).

NOTA Por lo general, una prueba *t*-para diferencia significativa en el nivel de confianza del 95 %, empleando las medias de cada ítem de ensayo de aptitud, ofrecerá generalmente una garantía similar o mejor de detección de inestabilidad de acuerdo con la ecuación (B.18) siempre y cuando el número de unidades ensayadas sea mayor o igual que 3.

B.6 ESTABILIDAD EN CONDICIONES DE TRANSPORTE

B.6.1 El proveedor del ensayo de aptitud debería comprobar los efectos del transporte en los ítems de ensayo de aptitud al menos en las etapas iniciales del programa de ensayos de aptitud. Con esta verificación se deberían comparar los ítems de ensayo de aptitud retenidos en las instalaciones del proveedor de ensayos de aptitud con los ítems sometidos a envío y

retorno. También se pueden emplear, por ejemplo, los estudios que se basan en la exposición a condiciones previsibles de transporte.

B.6.2 Al evaluar el desempeño se deberían considerar todos los efectos conocidos del transporte. Cualquier incremento significativo en la incertidumbre debido al transporte se debería incluir en la incertidumbre del valor asignado.

B.6.3 Cuando la comprobación de estabilidad en el transporte incluye la comparación de resultados para dos grupos de ítems de ensayo de aptitud, un grupo expuesto a condiciones de transporte y el otro no, el criterio para estabilidad suficiente en el transporte es el mismo del literal B.5.1 ó B.5.2.

NOTA 1 Si el valor asignado y la desviación estándar para evaluación de aptitud se determinan a partir de los resultados de los participantes (por ej., por métodos robustos), entonces el promedio y la desviación estándar para evaluación de la aptitud reflejarán todo sesgo e incremento de variabilidad (respectivamente) ocasionado por las condiciones de transporte.

NOTA 2 En el Anexo E.2 se muestra un ejemplo de comprobación de estabilidad.

ANEXO C
(Normativo)

ANÁLISIS ROBUSTO

C.1 ANÁLISIS ROBUSTO: INTRODUCCIÓN

Las comparaciones interlaboratorio presentan desafíos únicos para el análisis de datos. Si bien la mayoría de comparaciones interlaboratorio ofrecen datos unimodales y aproximadamente simétricos, la mayoría de conjuntos de datos de ensayo de aptitud incluyen una proporción de resultados inesperadamente distantes de la mayoría. Estos pueden surgir por una variedad de razones; por ejemplo, por participantes menos experimentados, métodos de medición menos precisos o tal vez nuevos o por participantes que no entendieron las instrucciones o procesaron los ítems de ensayo de aptitud de forma incorrecta. Tales resultados atípicos pueden ser muy variables y hacer que las técnicas estadísticas convencionales, incluida la media y la desviación estándar, no sean confiables.

Se recomienda (véase el numeral 6.5.1) que los proveedores de ensayos de aptitud empleen técnicas estadísticas robustas frente a valores atípicos. Muchas de estas técnicas han sido propuestas en la literatura estadística y muchas de ellas han sido empleadas con éxito en ensayos de aptitud. Adicionalmente, las técnicas más robustas confieren resistencia a distribuciones asimétricas de valores atípicos.

Este anexo describe varias técnicas que se han aplicado en ensayos de aptitud y tienen diferentes capacidades en relación con la robustez en poblaciones contaminadas (por ejemplo, eficiencia y punto de ruptura) y diferente simplicidad de aplicación. Se presentan aquí en orden de simplicidad (la más simple primero y las más complejas de últimas), relacionadas aproximadamente de forma inversa con la eficiencia ya que existe la tendencia a desarrollar estimadores más complejos con el fin de mejorar la eficiencia.

NOTA 1 En el Anexo D se ofrece información adicional sobre eficiencia, punto de ruptura y sensibilidad a modas menores - tres puntajes importantes del desempeño de varios estimadores robustos.

NOTA 2 La robustez es una propiedad del algoritmo de cálculo, no de los estimados que produce; de modo que no es estrictamente correcto llamar “robusto” a los promedios y a las desviaciones estándar calculados por dicho algoritmo. No obstante, para evitar el uso de terminología excesivamente engorrosa, los términos "promedio robusto" y "desviación estándar robusta" deberían entenderse en esta norma con el significado de estimados de la media o de la desviación estándar de la población calculados empleando un algoritmo robusto.

C.2 ESTIMADORES SIMPLES RESISTENTES A VALORES ATÍPICOS PARA LA MEDIA Y DESVIACIÓN ESTÁNDAR DE LA POBLACIÓN

C.2.1 La mediana

La mediana es un estimador simple y altamente resistente a valores atípicos de la media de la población para distribuciones simétricas. Para determinar la mediana, denotada como $med(x)$:

i) Denote los ítems p de los datos, clasificados en orden ascendente, por:

$x\{1\}, x\{2\}, ..., x\{p\}$

ii) Calcule

$$med(x) = \begin{cases} X_{\{(p+1)/2\}} & p\ impar \\ \dfrac{\{X_{\{p/2\}} + X_{1+p/2}\}}{2} & p\ par \end{cases} \qquad \text{(C.1)}$$

C.2.2 Desviación absoluta de la mediana ajustada *MADe*

La desviación absoluta de la mediana ajustada $MADe(x)$ ofrece un estimado de la desviación estándar de la población para datos con distribución normal y es altamente resistente a valores atípicos. Para calcular $MADe(x)$:

i) Calcule las diferencias absolutas d_i (para i = 1 a p) de

$$d_i = |x_i - med(x)| \qquad \text{(C.2)}$$

ii) Calcule $MADe(x)$ de

$$MADe(x) = 1{,}483\ med(d) \qquad \text{(C.3)}$$

Si el 50 % ó más de los resultados de participantes son los mismos, entonces $MADe(x)$ será cero y puede ser necesario usar el *nIQR* en el literal C.2.3, una desviación estándar aritmética (después de la eliminación de valores atípicos), o el procedimiento descrito en el literal C.5.2.

C.2.3 Rango intercuartilíco normalizado *nIQR*

Un estimador robusto de la desviación estándar similar a $MADe(x)$ y levemente más simple de obtener ha probado ser útil en muchos programas de ensayos de aptitud y se puede obtener de la diferencia entre el percentil 75^{avo} (o 3er cuartil) y el 25^{avo} percentil (o 1er cuartil) de los resultados de los participantes. Este estadístico se denomina comúnmente "Rango Intercuartilíco normalizado" (ó *nIQR*), y se calcula como en la fórmula (C.4):

$$nIQR(x) = 0{,}7413\ (Q_3(x) - Q_1(x)) \qquad \text{(C.4)}$$

en donde

$Q_1(x)$ denota el 25^{avo} percentil de x_i (i =1,2,…,p)

$Q_3(x)$ denota el 75^{avo} percentil de x_i (i =1,2,…,p)

Si los percentiles 75^{avo} y 25^{avo} son los mismos, el *nIQR* será cero (como lo será $MADe(x)$) y se debería emplear un procedimiento alternativo tal como una desviación estándar aritmética (después de la remoción de valores atípicos) o el procedimiento del literal C.5.2 para calcular la desviación estándar robusta.

NOTA 1 El *nIQR* sólo requiere clasificación de datos una vez comparado con *MADe,* aunque tiene punto de ruptura de 25 % (véase el Anexo D), mientras que *MADe* tiene punto de ruptura de 50 %. Por consiguiente *MADe* puede tolerar una proporción apreciablemente superior de valores atípicos que *nIQR.*

NOTA 2 Tanto los estimadores *nIQR* como *MADe* muestran sesgo negativo apreciable en p < 30 que los puede afectar de forma adversa si se emplean estos estimados para asignar el puntaje a los resultados de los participantes.

NOTA 3 Diferentes paquetes estadísticos pueden emplear diferentes algoritmos para calcular cuartiles y por consiguiente pueden producir *nIQR* levemente diferentes.

NOTA 4 En el Anexo E.3 se incluye un ejemplo del uso de estimadores robustos simples.

C.3 ANÁLISIS ROBUSTO: Algoritmo A

C.3.1 Algoritmo A con escala iterada

Este algoritmo produce estimados robustos del promedio y de la desviación estándar de los datos a los cuales se aplica.

Denote los ítems p de los datos, clasificados en orden ascendente, por:

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

Denote el promedio robusto y la desviación estándar robusta de estos datos por x^* y s^*.

Calcule valores iniciales para x^* y s^* como:

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

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

NOTA 1 Los algoritmos A y S presentados en este anexo se reproducen de la NTC 3529-5, con una ligera adición al Algoritmo A para especificar un criterio de detención: sin cambio en las 3^{eras} cifras significativas de la media robusta y la desviación estándar robusta.

NOTA 2 En algunos casos, más de la mitad de los resultados x_i serán idénticos (por ejemplo, recuento de hilos en tela o los electrolitos en el suero). En estos casos, el valor inicial de s^* será cero y el procedimiento robusto no se desempeñará de forma correcta. En el caso de que s^* inicial sea igual a cero ($s^* = 0$), resulta aceptable reemplazar la desviación estándar de la muestra, después de comprobar cualquier valor atípico bruto que pudiera hacer que la desviación estándar de la muestra sea demasiado grande. Este remplazo se efectúa sólo para s^* inicial, y después de que el algoritmo iterativo puede proceder como se describe.

Actualice los valores x^* y s^* de la siguiente manera. Calcule:

$$¢ = 1{,}5s^* \tag{C.7}$$

Para cada x_i ($i = 1, 2, \ldots, p$), Calcule:

$$X_i^* \begin{cases} x^* - ¢ & cuando \quad x_i < x^* - ¢ \\ x^* + ¢ & cuando \quad x_i > x^* + ¢ \\ x_i & de\ otra\ forma \end{cases} \tag{C.8}$$

Calcule los nuevos valores de x^* y s^* a partir de:

$$x^* = \sum_{i=1}^{p} x_i^* / p \tag{C.9}$$

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

en donde la sumatoria es sobre i.

Los estimados robustos x^* y s^* se pueden derivar mediante un cálculo iterativo, es decir, mediante la actualización de los valores de x^* y s^* varias veces empleando los datos modificados de las ecuaciones C.7 a C.10, hasta que el proceso converja. Se puede asumir convergencia cuando no exista cambio entre dos iteraciones consecutivas en la tercera cifra significativa de la media robusta y de la desviación estándar robusta (x^* y s^*). Se pueden

determinar otros criterios de convergencia alternativos de acuerdo con el diseño y los requisitos de presentación del informe para los resultados del ensayo de aptitud.

NOTA En el Anexo E.3 y E.4 se presentan ejemplos de uso del Algoritmo A con escala iterada.

C.3.2 Variantes del Algoritmo A

El Algoritmo A con escala iterada en el literal C.3.1 presenta ruptura modesta (aproximadamente 25 % para conjuntos con gran cantidad de datos [25]) y el punto de partida s^* sugerido en C.3.1 para conjuntos de datos donde $MADe(x)$ es cero, puede degradar gravemente la resistencia del valor atípico cuando hay valores atípicos severos en el conjunto de datos. Las siguientes variaciones deberían tenerse en cuenta cuando se espera que la proporción de valores atípicos sea superior al 20 % en cualquier conjunto de datos o donde el valor inicial para s^* se vea afectado de forma adversa por valores atípicos extremos:

i) Reemplace *MADe* con *med* $(|x_i - \bar{x}|)$ cuando $MADe = 0$, o se emplea un estimador alternativo tal como el descrito en C.5.1 o la desviación estándar aritmética (después de la remoción de valores atípicos).

ii) Cuando no se emplea la desviación estándar robusta en la asignación de puntaje, se debe emplear $MADe$ (enmendado como en el literal i) anterior) y no se actualiza s^* durante la iteración. Cuando se emplea la desviación estándar robusta en asignación de puntaje, reemplace s^* con el estimador Q descrito en el literal C.5 y no actualice s^* durante la iteración.

NOTA La variante ii) mejora el punto de ruptura del Algoritmo A en un 50 % [25], permitiendo que el algoritmo enfrente una proporción mayor de valores atípicos.

C.4 ANÁLISIS ROBUSTO: Algoritmo S

Este algoritmo se aplica a las desviaciones estándar (o los rangos), las cuales se calculan cuando los participantes entregan m resultados replicados para un mensurando en un ítem de ensayo de aptitud o en un estudio con m ítems de ensayo de aptitud idénticos. Esto produce un valor robusto combinado de las desviaciones estándar o los rangos a los cuales se aplica.

Denote las p desviaciones estándar o los rangos, clasificados en orden ascendente, por:

$$w_{\{1\}}, w_{\{2\}}, ..., w_{\{p\}}$$

Denote el valor robusto combinado con 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 estándar de m resultados de ensayo, $v = m - 1$.). Obtenga los valores de ξ y η requeridos por el algoritmo a partir de la Tabla C.1.

Calcule un valor inicial para w^* como:

$$w^* = \text{mediana de } w_i \qquad (i = 1, 2, ..., p) \qquad \text{(C.11)}$$

NOTA Si más de la mitad de los w_i son cero, entonces el w^* inicial será cero y el procedimiento robusto no funcionará de forma correcta. Cuando el w^* inicial es cero, se remplaza la desviación estándar promedio combinada aritmética (o rango promedio) después de eliminar cualquier valor atípico extremo que pueda influir en el promedio. Este remplazo es sólo para el w^* inicial, después de lo cual el procedimiento debería continuar como se describe.

Actualice el valor w^* de la siguiente manera. Calcule:

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

Para cada w_i (i = 1, 2, ..., p), calcule:

$$w_i^* = \begin{cases} \psi & si \;\; w_i > \psi \\ w_i & de\ otra\ \ forma \end{cases} \qquad \text{(C.13)}$$

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

$$w^* = ¶\sqrt{\sum_{i=1}^{p} \left(w_i^*\right)^2 / p} \qquad \text{(C.14)}$$

El estimado robusto w^* se calcula mediante un cálculo iterativo actualizando el valor de w^* varias veces hasta que el proceso converja. Se puede asumir convergencia cuando no exista cambio entre iteraciones consecutivas en la tercera cifra significativa del estimado robusto.

NOTA El Algoritmo S ofrece un estimado de la desviación estándar de la población cuando se suministra con desviaciones estándar a partir de una distribución normal única (y por ende proporciona un estimado de la desviación estándar de repetibilidad cuando se aplican las suposiciones de la NTC 3529-2).

Tabla C.1 Factores requeridos para 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 del Anexo B de la ISO 5725-5: 1998.		

C.5 ESTIMADORES ROBUSTOS INTENSIVOS COMPUTACIONALMENTE: Método Q y estimador Hampel

C.5.1 Fundamentos para Estimadores de robustez intensivos computacionalmente

Los estimadores robustos de la media de la población y de la desviación estándar descrita en los literales C.2 y C.3 resultan útiles cuando los recursos computacionales son limitados o cuando es necesario proveer explicaciones concisas de los procedimientos estadísticos. Estos procedimientos han probado ser útiles en una amplia variedad de situaciones, incluidos los programas de ensayos de aptitud en áreas nuevas de ensayo o calibración y en economías donde no ha estado previamente disponibles los ensayos de aptitud. Sin embargo, estas técnicas pueden volverse poco confiables cuando más del 20 % de los resultados son valores atípicos o cuando existen distribuciones bimodales (o multimodales), y algunas pueden volverse inaceptablemente variables para cantidades más pequeñas de participantes. Además, ninguna puede manejar datos replicados de los participantes. La NTC-ISO/IEC 17043 exige

que estas situaciones sean previstas en el diseño o sean detectadass mediante una revisión competente antes de la evaluación del desempeño, pero existen ocasiones en que esto no es posible.

Además, algunas de las técnicas robustas descritas en los literales C.2 y C.3 fallan en términos de eficiencia estadística - si la cantidad de participantes es menor a 50, y se emplea la media y/o la desviación estándar robustas para asignar el puntaje, existe un riesgo considerable de clasificar erróneamente a los participantes debido al uso de métodos estadísticos ineficaces.

Las técnicas robustas que combinan eficiencia buena (es decir, variabilidad comparativamente baja) con tolerancia para una elevada proporción de valores atípicos tienden a ser más complejas y requieren más recursos computacionales, pero aparecen referenciadas en la literatura disponible y en normas internacionales. Algunas de estas adicionalmente proporcionan ganancias de desempeño útiles cuando la distribución subyacente de datos es sesgada o cuanto algunos resultados citados se encuentran por debajo de un límite de detección o reporte.

En los siguientes parágrafos se describen algunos métodos de eficiencia alta y ruptura elevada para la determinación de la desviación estándar y la ubicación (media) que resultan útiles para datos con mayores proporciones de valores atípicos y que muestran menor variabilidad que los estimadores más simples. Uno de los estimadores descritos también se puede emplear para calcular una desviación estándar de reproducibilidad cuando los participantes reporten múltiples observaciones.

C.5.2 Determinación de una desviación estándar robusta empleando métodos Q y Q_n

C.5.2.1 Q_n [34] es un estimador de punto de ruptura alto y eficiencia elevada de la desviación estándar de la población que no tiene sesgo para los datos con distribución normal (es decir, bajo la suposición de que no existan valores atípicos). Q_n emplea un único resultado reportado (incluida una media o mediana de las replicas) por cada participante. El cálculo se basa en el uso de diferencias por pares dentro del conjunto de datos y por consiguiente no depende de un estimado de la media o mediana de los datos. La implementación aquí descrita incluye correcciones para asegurar que el estimado no tenga sesgo para todos los tamaños posibles de conjuntos de datos.

Para calcular Q_n para un conjunto de datos (x_1, x_2, … x_p) con p resultados reportados:

i) Calcule las diferencias absolutas $p(p\text{-}1)/2$

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

ii) Denote las diferencias ordenadas d_{ij} por

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

iii) Calcule:

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

es decir, k es el número de distintas parejas seleccionadas de h objetos, donde:

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

iv) Calcule Q_n como

$$Q_n = 2{,}221\ 9 d_{\{k\}} b_p \qquad \text{(C.19)}$$

en donde se selecciona b_p de la Tabla C.2 para una cantidad particular p de puntos de datos o para $p > 12$, calcule a partir de:

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

en donde

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

NOTA 1 El factor de 2,2219 es un factor de corrección para dar un estimado sin sesgo de la desviación estándar para p grandes. Los factores de corrección b_p para p pequeños se encuentran en la Tabla C.2 y los cálculos para r_p para $p > 12$ son como los provistos en la referencia [34] de simulación extensiva y análisis de regresión posterior.

NOTA 2 El algoritmo simple descrito anteriormente requiere recursos de computación considerables para conjuntos de datos más grandes, por ejemplo $p > 1\ 000$. Se ha publicado una implementación rápida y de memoria eficiente capaz de manejar conjuntos de datos mucho más grandes, con código de computador completo [34]; en la referencia [34] se citó un desempeño aceptable para p sobre 8 000 en el momento de la 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 método Q produce un estimado de alta ruptura, alta eficiencia de la desviación estándar de los resultados de ensayos de aptitud reportados por diferentes laboratorios. El método Q no sólo es robusto contra resultados atípicos, sino además contra una situación donde muchos resultados de ensayo son iguales; por ejemplo, debido a datos cuantitativos en una escala discontinua o debido a distorsiones de redondeo. En dicha situación, otros métodos parecidos a Q pueden fallar porque muchas diferencias por pares son cero.

Se puede emplear el método Q para ensayos de aptitud tanto con resultados únicos por participante (incluida una media o mediana de replicas) y por replicas. El uso directo de replicas en el cálculo mejora la eficiencia del método.

El cálculo depende del uso de diferencias por pares dentro del conjunto de datos y por consiguiente no depende de un estimado de la media o la mediana de los datos. El método se conoce como Q/Hampel cuando se emplea junto con el algoritmo de paso finito para el estimador Hampel descrito en C.5.3.3.

Se denotan los resultados de medición reportados, agrupados por laboratorio, mediante:

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

Calcule la función de distribución acumulada de todas las diferencias interlaboratorio absolutas

$$H_1(x) = \frac{2}{p(p-1)} \sum_{1 \le i < j \le p} \frac{1}{n_i n_j} \sum_{k=1}^{n\,j} \sum_{m=1}^{n\,j} I\left\{ \left| y_{ik} - y_{jm} \right| \le x \right\} \qquad \text{(C.22)}$$

en donde $I\left\{ \left| y_{ik} + y_{jm} \right| \le x \right\} = \begin{Bmatrix} 1 & si & \left| y_{ik} - y_{jm} \right| \le x \\ 0 & de\ otra\ forma & \end{Bmatrix}$ denota la función del indicador.

Denote los puntos de discontinuidad de $H_1(x)$ mediante:

$x_1, \ldots, x_r$, donde $x_1 < x_2 < \ldots < x_r$.

Calcule para todos los puntos de discontinuidad positivos $x_1, \ldots, x_r$:

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

Y sea

$$G_1(0) = 0$$

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

Calcule la desviación estándar robusta s^* de los resultados de ensayo de los diferentes laboratorios

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

en donde $H_1(0)$ se calcula como en la ecuación (C.22) y es igual a cero a menos que hayan enlaces exactos en el conjunto de datos y donde $\Phi^{-1}(q)$ es el cuantil q^{avo} de la distribución normal estándar.

NOTA 1 Este algoritmo no depende de un valor promedio; se puede emplear junto con un valor de resultados combinados de los participantes o un valor de referencia especificado.

NOTA 2 Otras variantes del método *Q* ofrecen estimados robustos de desviación estándar tanto de repetibilidad como de reproducibilidad [25,34].

NOTA 3 En las referencias [26] y [34] se describe la base teórica para el método *Q*, incluido el desempeño asintótico y el punto de ruptura de una muestra finita. .

NOTA 4 Si los datos subyacentes de los participantes representan resultados de medición únicos obtenidos con un método de medición específico, la desviación estándar robusta es un estimado de la desviación estándar de la reproducibilidad como en la ecuación (C.21).

NOTA 5 La desviación estándar de reproducibilidad no necesariamente es la desviación estándar más apropiada para usar en ensayos de aptitud debido a que por lo general es un estimado de la dispersión de resultados únicos y no un estimado de la dispersión de medias o medianas de resultados replicados de cada participante. No obstante, la dispersión de las medias o medianas de resultados replicados se encuentra levemente por debajo de la dispersión de resultados únicos de laboratorios diferentes, si la relación de la desviación estándar de la reproducibilidad dividida por la desviación estándar de la repetibilidad es mayor que 2. Si esta relación es inferior a 2, para asignar el puntaje en el ensayo de aptitud puede considerarse reemplazar la desviación estándar de reproducibilidad s_R por el valor corregido $\sqrt{s_R^2 - \frac{m-1}{m} s_r^2}$, en donde m denota el número de replicas y s_r^2 la varianza de repetibilidad como se calcula en [35], o no usar las replicas sino la media de las replicas por participante para el método *Q*.

NOTA 6 La Nota 5 se aplica sólo si el puntaje se determina con base en las medias o medianas de los resultados replicados. Si las replicas son muestras ciegas de los ítems de ensayo de aptitud replicados , se deberían dar resultados de indicadores por cada ítem replicado. En este caso, la desviación estándar de reproducibilidad es la desviación estándar más apropiada.

NOTA 7 En el Anexo E.3 se muestra un ejemplo en el que se ha aplicado el método *Q*.

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

C.5.3.1 El estimado Hampel es un estimado altamente robusto y eficiente de la media general de resultados reportados por diferentes laboratorios. Puesto que no existe fórmula explícita para obtener el estimado Hampel, en este parágrafo se proveen dos algoritmos. El primero puede ser implementado con mayor facilidad pero puede conllevar a resultados desviados en diferentes implementaciones. El segundo ofrece resultados únicos dependiendo solamente de la desviación estándar subyacente.

C.5.3.2 El siguiente cálculo ofrece un esquema de repetición iterativa para obtener el estimado Hampel de ubicación.

i) Denote los datos como $x_1, x_2 \ldots x_p$

ii) Ajuste x^* a $med(x)$ (literal C.2.1)

iii) Ajuste s^* a un estimado robusto adecuado de desviación estándar, por ejemplo $MADe$, Q_n ó s^* del método *Q*.

iv) Por cada punto de datos x_i, calcule q_i a partir de

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

v) Calcule los pesos w_i de

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

vi) Recalcule x^* a partir de

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

vii) Repita los pasos iv) a vi) hasta que x^* converja. Se puede asumir la convergencia cuando el cambio en x^* de una iteración a la siguiente es menor de $0{,}01\, s^*/\sqrt{p}$, correspondiente a aproximadamente 1 % del error estándar en x*. Se pueden utilizar otros criterios de de convergencia más precisos..

No se garantiza que esta implementación del estimador Hampel tenga una solución única o resulte en la mejor solución puesto que una elección deficiente de ubicación inicial x^* y/o s^* puede excluir partes importantes del conjunto de datos. El proveedor de ensayos de aptitud

debería, por tanto, implementar medidas para comprobar la posibilidad de una solución deficiente o proporcionar reglas no ambiguas para la elección de la ubicación. La regla más común consiste en escoger la solución más cercana a la mediana. Para confirmar una solución viable, puede servir revisar los resultados a fin de garantizar que ninguna proporción grande del conjunto de datos esté por fuera del rango $|q|>4.5$.

NOTA 1 Esta implementación del estimador de Hampel tiene aproximadamente 96 % de eficiencia para datos con distribución normal.

NOTA 2 En el Anexo E.3 se incluye un ejemplo de uso de esta implementación.

NOTA 3 Se puede ajustar el estimador de Hampel para obtener mayor eficiencia o resistencia a valores atípicos cambiando la función de ponderación. La forma general de la función de ponderación es

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

en donde a, b y c son parámetros que se ajustan. Para la implementación aquí, a = 1,5, b = 3,0 y c = 4,5. Se obtiene mayor eficiencia aumentando el rango; y se mejora la resistencia a valores atípicos o modas menores, reduciendo el rango.

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

Calcule la media aritmética por cada laboratorio, ahora etiquetadas $y_1, y_2, \ldots y_p$.

Calcule la media robusta, x^*, resolviendo la ecuación

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

en donde

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

y s^* es la desviación estándar robusta de acuerdo con el método Q.

La solución exacta se puede obtener en una cantidad finita de pasos, lo que significa que no es de manera iterativa, empleando la propiedad de que ψ en el argumento de x^* es parcialmente lineal, teniendo en mente que los nodos de interpolación en el lado izquierdo de la ecuación (C.25) (interpretados aquí como una función de x^*) son como sigue:

Calcule 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 $y_3, \ldots, y_p$.

Clasifique estos datos $d_1, d_2, d_3, \ldots, d_{6 \cdot p}$ en orden ascendente. $d_{\{1\}}, d_{\{2\}}, d_{\{3\}}, \ldots, d_{\{6 \cdot p\}}$

Luego calcule para cada $m = 1, \ldots, (6 \cdot p - 1)$

$$p_m = \sum_{i=1}^{p} \mathrm{v}\left(\frac{y_i - d_{\{m\}}}{s^*}\right)$$ y compruebe 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) $p_m \cdot p_{m+1} < 0$. Si es así, $x_m = d_{\{m\}} - \dfrac{p_m}{\dfrac{p_{m+1} - p_m}{d_{(m+1)} - d_{(m)}}}$ es una solución de la ecuación (C.25).

Sea que S denote el conjunto de todas estas soluciones de la ecuación (C.25).

La solución $x^* \in S$ más cercana a la mediana se emplea como parámetro de ubicación x^*, por ejemplo,

$$\left|x^* - med\left(y_1, y_2, \ldots, y_p\right)\right| = m \quad en \left\{\left|x\right| - med\left(y_1, y_2, \ldots, y_p\right)\right|; x \in S\}$$

Pueden existir varias soluciones. Si existen dos soluciones más cercanas a la mediana o si no hay ninguna solución, se emplea la misma mediana como parámetro de ubicación x^*.

NOTA 1 Esta implementación del estimador de Hampel tiene aproximadamente 96 % de eficiencia para datos con distribución normal.

NOTA 2 Si se emplea este método de cálculo, los resultados de laboratorio que difieran de la media en más de 4,5 veces la desviación estándar de reproducibilidad dejarán de tener efecto en el resultado del cálculo; es decir, se tratan como valores atípicos.

C.5.4 El método *Q*/Hampel

El método conocido como *Q*/Hampel emplea el método *Q* descrito en C.5.3.2 para el cálculo de la desviación estándar robusta s^* junto con el algoritmo de pasos finitos para el estimador Hampel descrito en C.5.3.3 para el cálculo del parámetro de ubicación x^*.

Cuando los participantes reportan observaciones múltiples, se emplea el método *Q* descrito en C.5.3.2 para el cálculo de la desviación estándar de reproducibilidad robusta s_R. Para el cálculo de desviación estándar de repetibilidad robusta s_r, se aplica un segundo algoritmo empleando las diferencias por par dentro de los laboratorios.

NOTA Existe una aplicación web para el método *Q*/Hampel [37].

C.6 OTRAS TÉCNICAS ROBUSTAS

Los métodos descritos en este Anexo no constituyen una colección completa de enfoques válidos y no se garantiza que ninguno sea óptimo para todas las situaciones. Se pueden emplear otros estimadores robustos a discreción del proveedor de ensayos de aptitud, sujetos a demostración, por referencia a eficiencia conocida, punto de ruptura y otras propiedades apropiadas, que cumplan los requisitos particulares del programa de ensayos de aptitud.

ANEXO D
(Informativo)

ORIENTACIÓN ADICIONAL SOBRE PROCEDIMIENTOS ESTADÍSTICOS

D.1 PROCEDIMIENTOS PARA CANTIDADES PEQUEÑAS DE PARTICIPANTES

D.1.1 Consideraciones generales

Muchos programas de ensayos de aptitud tienen pocos participantes o tienen grupos de comparación con pequeñas cantidades de participantes, incluso si existe una gran cantidad de participantes en el programa. Esto puede ocurrir frecuentemente cuando los participantes se agrupan y califican por método, como se hace de forma común en ensayos de aptitud para laboratorios médicos, por ejemplo.

Cuando la cantidad de participantes es pequeña, el valor asignado debería estar determinado de forma ideal empleando un procedimiento metrológicamente válido, independiente de los participantes, tales como por formulación o de un laboratorio de referencia. Los criterios de evaluación de desempeño también deberían basarse en criterios externos, tales como el juicio de expertos o criterios basados en la idoneidad para el fin previsto. En estas situaciones ideales, se evalúa el desempeño empleando el valor asignado predeterminado y el criterio de desempeño; de modo que el ensayo de aptitud puede realizarse con sólo un participante. Este tipo de comparación interlaboratorio puede denominarse comparación bilateral o auditoría de medición y puede resultar muy útil en muchas situaciones; por ejemplo en calibración.

Cuando estas condiciones ideales no pueden cumplirse, es posible que se deba derivar el valor asignado o la dispersión, o ambos, de los resultados del participante. Si el número de participantes es demasiado pequeño para los procedimientos particulares empleados, la evaluación de desempeño puede volverse poco confiable. Por consiguiente, es importante considerar si se debería establecer una cantidad mínima de participantes para la evaluación de desempeño.

En los siguientes parágrafos se presenta orientación para situaciones de cantidades pequeñas, cuando los criterios de evaluación del desempeño se determinan empleando resultados de los participantes.

D.1.2 Procedimientos para identificar valores atípicos

Aunque los estadísticos robustos son muy recomendados para poblaciones contaminadas con valores atípicos, no se recomiendan a menudo, para conjuntos de datos muy pequeños (véanse las excepciones a continuación). No obstante, es posible realizar prueba de valores atípicos para conjuntos de datos muy pequeños. Por consiguiente, puede ser preferible el rechazo de valores atípicos seguido por el cálculo de la media o la desviación estándar, por ejemplo, en el caso de programas o grupos muy pequeños.

Diferentes pruebas de valores atípicos son aplicables a diferentes tamaños de conjuntos de datos. En la NTC 3529-2 (ISO 5725-2) se presentan tablas para la prueba de Grubbs para un único valor atípico y para dos valores atípicos simultáneos en la misma dirección. La prueba de Grubbs y otras pruebas exigen especificar la cantidad de valores atípicos posibles por anticipado y pueden fallar cuando existen múltiples valores atípicos, haciendo que sean más útiles para $p > 10$ (dependiendo de la proporción probable de valores atípicos).

NOTA 1 Se debería tener cuidado al calcular la dispersión después del rechazo de valores atípicos ya que los estimados de dispersión tendrán un sesgo bajo. Por lo general, el sesgo no es grave si el rechazo se efectúa sólo en el nivel de confianza del 99 % o superior.

NOTA 2 La mayoría de estimadores robustos univariados para ubicación y dispersión se desempeñan de forma aceptable para $p \geq 12$.

D.1.3 Procedimientos para estimados de ubicación

D.1.3.1 Los valores asignados derivados de pequeños conjuntos de datos de participantes deberían, en la medida de lo posible, cumplir con el criterio de incertidumbre del valor asignado presentado en el numeral 9.2.1. Para una situación en la que se emplee una media simple como el valor asignado y una desviación estándar de los resultados como la desviación estándar para evaluación de aptitud, este criterio no se puede cumplir para una distribución normal con $p \leq 12$, después de cualquier remoción de valores atípicos. Para usar la mediana como el valor asignado (tomando la eficiencia como 0,64), no se puede cumplir el criterio para $p \leq 18$.

Otros estimadores robustos, tales como el Algoritmo A (C.3), tienen eficiencia intermedia y pueden cumplir el criterio para $p > 12$ si se tienen en cuenta las disposiciones del numeral 7.7.3 Nota 2.

D.1.3.2 Existen limitaciones de tamaño del conjunto de datos en la aplicabilidad de algunos estimadores de ubicación. Pocos estimadores robustos computacionalmente intensivos se recomiendan para la media en conjuntos de datos pequeños. Un límite inferior típico es $p \geq 15$, aunque los proveedores pueden estar en la capacidad de demostrar desempeño aceptable para suposiciones específicas sobre conjuntos de datos más pequeños. La mediana es aplicable hasta $p = 2$ (cuando es igual a la media) pero para $3 \leq p \leq 5$ la mediana ofrece pocas ventajas sobre la media a menos que exista un riesgo inusualmente elevado de resultados deficientes.

D.1.4 Procedimientos para estimados de dispersión

D.1.4.1 No se recomienda usar criterios de desempeño con base en la dispersión de los resultados de los participantes para conjuntos pequeños de datos debido a la muy alta variabilidad de cualquier estimado de dispersión. Por ejemplo, para $p = 30$, se espera que los estimados de la desviación estándar para datos con distribución normal varíen en aproximadamente 25 % en cualquier lado de su valor verdadero (con base en un nivel de confianza del 95 %). Ningún otro estimador mejora en esto para los datos con distribución normal.

D.1.4.2 Cuando se requieren estimadores de dispersión para otros propósitos (por ejemplo como estadísticas de resumen o un estimado de dispersión para estimadores de ubicación robusta), o donde el programa de ensayos de aptitud puede tolerar la variabilidad elevada en estimadoress de dispersión, se deberían seleccionar estimadores de dispersión con la eficiencia más alta disponible cuando se manejen conjuntos de datos más pequeños.

NOTA 1 Con la frase "más alta disponible" se entiende que se tiene en cuenta la disponibilidad de software y la experiencia disponibles.

NOTA 2 El estimador Q_n de desviación estándar descrito en el literal C.5 es considerablemente más eficiente que el estimador $MADe$ ó $nIQR$ del Anexo C.1.

NOTA 3 Se han realizado recomendaciones específicas para estimadores robustos de dispersión en conjuntos de datos muy pequeños [24] como sigue:

- $p = 2$: emplee $|x_1 - x_2| / \sqrt{2}$;

- p = 3, ubicaciones y escala desconocidos: emplee $MADe$ para protección contra estimados excesivamente elevados de la desviación estándar o la desviación absoluta de la media para protección contra estimados excesivamente pequeños de la desviación estándar, por ejemplo cuando el redondeo puede dar dos valores idénticos;

- $p \geq 4$: En la referencia [27] se recomendó un estimado M específico de desviación estándar basado en una función de ponderación logarítmica; un equivalente cercano es el Algoritmo A sin iteración de ubicación, empleando la mediana como un estimador de ubicación.

NOTA 4 Para obtener un estimado de la desviación estándar a partir de la distancia absoluta hasta la mediana, se emplea

$$s^* = \frac{1}{0{,}798 \; x \; p} \sum_{i=1}^{p} \left| x_i - med\,(x) \right| \qquad \text{(D.1)}$$

D.2 EFICIENCIA Y PUNTOS DE RUPTURA PARA PROCEDIMIENTOS ROBUSTOS

D.2.1 Se pueden comparar diferentes estimadores estadísticos (por ejemplo, técnicas robustas) sobre tres características clave:

Punto de ruptura. La proporción de valores en el conjunto de datos que se puede remplazar por valores arbitrariamente grandes sin que el estimado también se vuelva arbitrariamente grande.

Eficiencia. La relación de la varianza del estimador dividida por la varianza de un estimador de varianza mínima para la distribución en cuestión.

Resistencia a modas menores**.** La capacidad de un estimador para resistir el sesgo causado por un grupo minoritario de resultados discrepantes (por lo general menor del 20 % del conjunto de datos).

Estas características dependen en gran medida de la distribución subyacente de los resultados para una población de participantes competentes y de la naturaleza de los resultados que provienen de participantes incompetentes (o de participantes que no siguieron las instrucciones o el método de medición). Los datos contaminantes pueden aparecer como valores atípicos, resultados con mayor varianza o resultados con una media diferente (por ejemplo, bimodales).

Los puntos de ruptura y las eficiencias para los diferentes estimadores serán diferentes para diferentes situaciones, y realizar una revisión exhaustiva va más allá del alcance de este documento. No obstante, se pueden realizar comparaciones simples bajo la suposición de una distribución normal para resultados de laboratorios competentes, con una media igual a x_{pt} y una desviación estándar igual a σ_{pt}.

D.2.2 Punto de ruptura

El punto de ruptura es la proporción de valores en el conjunto de datos que pueden ser valores atípicos sin que el estimado se vea afectado de forma adversa. El punto de ruptura es una medición de la resistencia a los valores atípicos; un punto de ruptura elevado está asociado con la resistencia a una proporción elevada de valores atípicos. Los puntos de ruptura y la resistencia a modas menores para los estimadores del Anexo C se presentan en la Tabla D.1. Se debería observar que los procedimientos exigidos en los numerales 6.3 y 6.4 deberían evitar el análisis de datos de conjuntos de datos con grandes proporciones de valores atípicos. No obstante, existen situaciones donde la revisión visual no es práctica.

Tabla D.1 Puntos de ruptura para estimados de la media y la desviación estándar (proporción de valores atípicos que pueden conllevar a falla del estimador)

Estimador estadístico	Parámetro de población estimado	Punto de ruptura	Resistencia a modas menores
Media de la muestra	Media	0 %	Deficiente
Desviación estándar de la muestra	Desviación estándar	0 %	Deficiente
Mediana de la muestra	Media	50 %	Buena
nIQR	Desviación estándar	25 %	Moderada
MADe	Desviación estándar	50 %	Moderada - Buena
Algoritmo A	Media y desviación estándar	25 %	Moderada
Q_n y *Q*/Hampel	Media y desviación estándar	50 %	Moderada (Muy buena para modas menores con distancia mayor que 6 *s**)

NOTA La definición de punto de ruptura empleada aquí es la proporción de un conjunto de datos grande con distribución normal que se puede mover a +infinito sin que el estimado también se mueva a infinito. Por ejemplo, si sólo por debajo del 50 % de un conjunto de datos se remplaza por +infinito, la mediana permanecerá dentro de los datos finitos restantes.

En resumen, la media y la desviación estándar de la muestra pueden presentar ruptura con sólo un valor atípico único. Los métodos robustos que emplean la mediana, *MADe*, y los métodos *Q*/Hampel pueden tolerar una muy grande proporción de valores atípicos. El Algoritmo A con desviación estándar iterada y *nIQR* tiene un punto de ruptura de 25 %. En cualquier situación con una gran proporción de valores atípicos (> 20 %), cualquier procedimiento convencional o robusto puede producir estimados no razonables de ubicación y de dispersión y se debería tener cuidado en la interpretación de tales valores.

D.2.3 Eficiencia relativa

Todos los estimados tienen varianza de muestreo; es decir, los estimados pueden variar entre rondas 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 varianzas. Los estimadores robustos modifican los resultados presentados que están excepcionalmente lejos de la mitad de la distribución, con base en suposiciones teóricas y por lo tanto estos estimadores tienen una mayor varianza que los estimadores de varianza mínima, en el caso en que el conjunto de datos tenga en efecto una distribución normal.

La media y la desviación estándar de la muestra son los estimadores de varianza mínima de la media y la desviación estándar de la población, y por ende tienen eficiencia del 100 %. Los estimadores con eficiencia inferior tienen mayor varianza; es decir, podrían variar más entre rondas, incluso si no hay valores atípicos o diferentes subgrupos de participantes. En la Tabla D.2 se presentan eficiencias relativas para los estimadores presentados en el Anexo C.

Tabla D.2 Eficiencia relativa de estimadores robustos para la media y la desviación estándar de la población para conjuntos de datos con distribución normal, con n = 50 ó 500 participantes:

Estimador estadístico	Media, n = 50	Media, n = 500	Desviación Estándar, n = 50	Desviación Estándar, n = 500
Media y desviación estándar de la muestra	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 sea perfecto para todas las situaciones. La media y la desviación estándar de la muestra son óptimas con una distribución normal, pero fracasa en el caso de valores atípicos. Los métodos robustos simples, tales como la mediana, $MADe$ ó $nIQR$ tienen un desempeño comparativamente pobre para los datos con distribución normal, pero pueden ser efectivos cuando hay presentes valores atípicos o el conjunto de datos es pequeño.

D.3 USO DE DATOS DE ENSAYOS DE APTITUD PARA EVALUAR LA REPRODUCIBILIDAD Y REPETIBILIDAD DE UN MÉTODO DE MEDICIÓN

D.3.1 En la Introducción de la NTC-ISO/IEC 17043 se establece que la evaluación de las características de desempeño de un método, por lo general, no es un propósito de los ensayos de aptitud. No obstante, es posible emplear los resultados de los programas de ensayos de aptitud para comprobar, y tal vez establecer la repetibilidad y reproducibilidad de un método de medición [15] cuando el programa de ensayos de aptitud cumpla las siguientes condiciones:

a) los ítems de ensayo de aptitud son suficientemente homogéneos y estables;

b) los participantes son capaces de lograr un desempeño satisfactorio coherente,

c) se ha demostrado la competencia de los participantes (o un subconjunto de participantes) antes de la ronda de ensayos de aptitud, y su competencia no se pone en duda por los resultados de la ronda.

D.3.2 Con el fin de proporcionar datos suficientes para la evaluación de repetibilidad y reproducibilidad de un método de ensayo a partir de un programa de ensayos de aptitud, se deben emplear las siguientes condiciones de diseño:

a) una cantidad suficiente de participantes para satisfacer un estudio de cooperación tiene competencia demostrada con un método de medición en rondas anteriores de un programa de ensayos de aptitud, y se ha comprometido para seguir el método de medición sin modificación;

b) cuando se va a evaluar la repetibilidad, cada ronda de ensayos de aptitud empleada para la evaluación de repetibilidad debería incluir por lo menos dos ítems de ensayo de aptitud o un requisito para observaciones replicadas;

c) siempre que sea posible, se debería proporcionar a los participantes replicas ciegas, en vez de solicitarles realizar mediciones replicadas sobre el mismo ítem de ensayo de aptitud;

d) los ítems de ensayos de aptitud empleados en una o varias rondas del programa de ensayos de aptitud cubren el rango de niveles y tipos de muestras de rutina para los cuales está destinado el método de medición;

e) los procedimientos de análisis de datos empleados para evaluar la repetibilidad y la reproducibilidad deberían ser coherentes con la NTC 3529 o con el protocolo del estudio de cooperación en uso.

ANEXO E
(Informativo)

EJEMPLOS ILUSTRATIVOS

Los siguientes ejemplos tienen como intención ilustrar los procedimientos especificados en la presente norma, a fin de que el lector pueda determinar que sus cálculos son correctos. Los ejemplos específicos no deberían considerarse como recomendaciones para uso en programas de ensayos de aptitud particulares.

E.1 EFECTO DE VALORES CENSURADOS (véase el numeral 5.5.3.3)

En la Tabla E.1 se muestran 23 resultados para una ronda de un programa de ensayos de aptitud, de los cuales 5 están indicados como "Menor que" alguna cantidad. La media robusta (x^*) y la desviación estándar (s^*) del Algoritmo A, se muestran para 3 cálculos diferentes, donde los signos '<' son descartados y los datos son analizados como datos cuantitativos; los resultados con valores '<' son ignorados; y donde 0,5 veces el resultado es insertado como un estimado del resultado cuantitativo. En cada escenario, los resultados que habrían estado por fuera del límite de aceptación se indican con '#'. Con esto se asume que la evaluación sería "inaceptable" (señal de acción) para cualquier resultado donde la parte cuantitativa esté por fuera de $x^* \pm 3s^*$. El proveedor del ensayo de aptitud podría tener reglas alternativas para evaluar los resultados con signos '<' or '>'.

Tabla E.1 Conjunto de datos con resultados truncados (<), y tres opciones para acomodar los resultados

Participante	Resultado	'<' ignorado	'< ' eliminado	0,5 x '<' 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

Continúa…

Tabla E.1 *(Final)*

Participante	Resultado	'<' ignorado	'< ' eliminado	0,5 x '<' valor
Y	45	45	45 #	45
Z	< 50	50 #	--	25
Resumen				
Número de resultados	23	23	18	23
*x**		26,01	26,81	23,95
*s**		7.23	5.29	8.60

La decisión de cómo manejar las muestras "menor que" tiene un efecto significativo en la media robusta y la desviación estándar robusta y en la evaluación de desempeño. Se espera que el proveedor de ensayos de aptitud determine un método apropiado.

E.2 ENSAYOS DE HOMOGENEIDAD Y ESTABILIDAD - Arsénico (As) en chocolate (véase el numeral 6.1)

Los ítems de ensayo se prepararon para emplearlos en un ensayo de aptitud internacional y para su posterior uso como materiales de referencia. Se fabricaron 1000 ampolletas.

Comprobación de homogeneidad: se seleccionaron 10 ítems de ensayo de aptitud empleando una selección aleatoria estratificada de ítems de ensayo de aptitud de diferentes porciones del proceso de manufactura. Dos de las porciones de ensayo fueron extraídas de cada botella y ensayadas en orden aleatorio, bajo condiciones de repetibilidad. Los datos obtenidos se presentan a continuación en la Tabla E.2. El procedimiento del Anexo B.3 se llevó a cabo, dando como resultado los estadísticos de resumen relacionadas. La adecuación para el propósito σ_{pt} para As en chocolate es de 15 %, de modo que el estimado de variabilidad de la muestra se comprueba contra 0,3 veces σ_{pt}.

Tabla E.2 Datos de homogeneidad para los ítems de ensayo de aptitud de arsénico en chocolate

Identificación de la botella	Replica 1	Replica 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

σ_{pt}: = 0,18715 x 0,15 = 0,02807

Valor de comprobación: $0,3\sigma_{pt}$ = 0,00842

s_s: es menor que el valor de comprobación, de modo que la homogeneidad es suficiente.

Comprobación de estabilidad: Se seleccionaron aleatoriamente 2 ítems de ensayos de aptitud y se almacenaron a una temperatura elevada (60°C) durante el tiempo de duración de la ronda del programa de ensayos de aptitud (6 semanas). Los ítems de ensayos de aptitud fueron ensayados por duplicado (Tabla E.3) y los cuatro resultados se verificaron contra los valores de homogeneidad.

Tabla E.3 Datos de estabilidad para ítems de ensayos de aptitud para arsénico en chocolate

Muestra de estabilidad	Replica 1	Replica 2
164	0,191	0,198
732	0,190	0,196

Promedio general: = 0,19375

Diferencia con respecto a la media de homogeneidad: 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 tanto, la estabilidad es suficiente.

E.3 EJEMPLO GLOBAL DE ATRAZINA EN AGUA POTABLE

Un programa de ensayos de aptitud para un herbicida (Atrazina) en agua potable tiene 34 participantes. Estos datos sin procesar, como se presentan en la Tabla E.4, se ordenaron por valor con fines de claridad. La Tabla muestra los valores calculados para la media y la desviación estándar robustas siguiendo el Algoritmo A, luego de 6 iteraciones hasta que la media y la desviación estándar robustas no cambien en sus terceras cifras significativas. Los datos se muestran como gráfico de datos clasificados en la Figura E.1 y en el histograma correspondiente y gráfico de densidad kernel en las Figuras E.2 y E.3, respectivamente.

En la Tabla E.5 se muestran los estimados de ubicación (promedio) y desviación estándar empleando varias técnicas clásicas y robustas. También se muestra la incertidumbre del estimado de ubicación. Los estadísticos para el método re muestreo (*bootstrap)* se derivaron de los procedimientos de las referencias [17,18] y del paquete de software R [véase R3.1.1 a continuación]. En la Figura E.4 se muestran los diferentes estimados de ubicación y el estimado de incertidumbre expandida ($2u(x_{pt})$) como la barra de error.

Tabla E.4 Cálculo del promedio robusto y la desviación estándar para Atrazina en agua potable

	x_i	1a iteración	2a iteración	3a iteración	4a iteración	5a iteración	6a iteración
x* - ¢		**0,204163**	**0,199732**	**0,198466**	**0,198037**	**0,197865**	**0,197790**
x* + ¢		**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	0,1978
2	0,0550	0,2042	0,1997	0,1985	0,1980	0,1979	0,1978
3	0,1780	0,2042	0,1997	0,1985	0,1980	0,1979	0,1978
4	0,2020	0,2042	0,2020	0,2020	0,2020	0,2020	0,2020
5	0,2060	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	0,2270
7	0,2280	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	0,2300
9	0,2300	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	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
¢		0,0578	0,0581	0,0587	0,0590	0,0592	0,0592
***x** nuevo**	0,2620	0,2579	0,2572	0,2571	0,2570	0,2570	**0,2570**
***s** nuevo**	0,0386	0,0387	0,0391	0,0393	0,0394	0,0395	**0,0395**

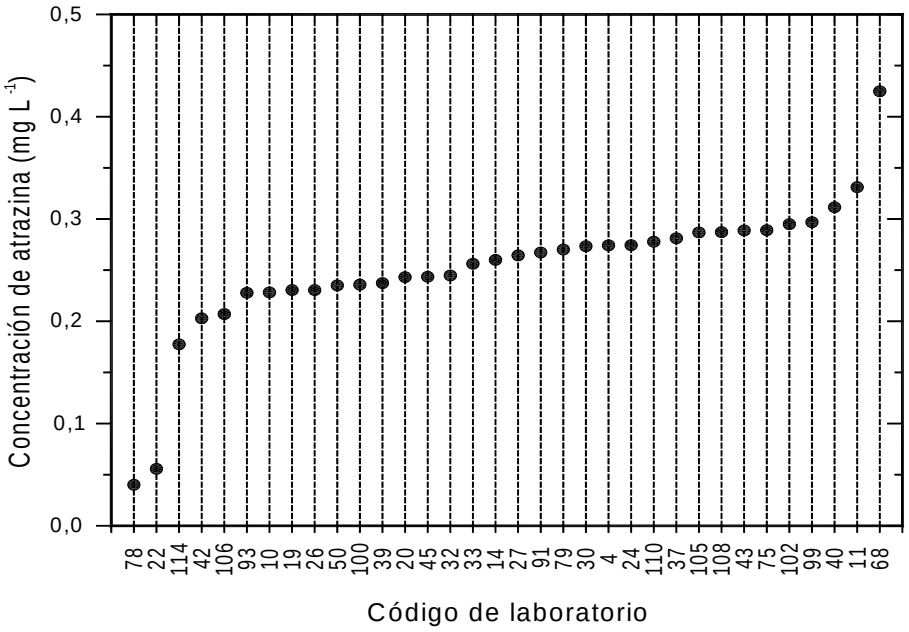

Figura E.1 Resultados de participantes clasificados para Atrazina (datos de la Tabla E.4)

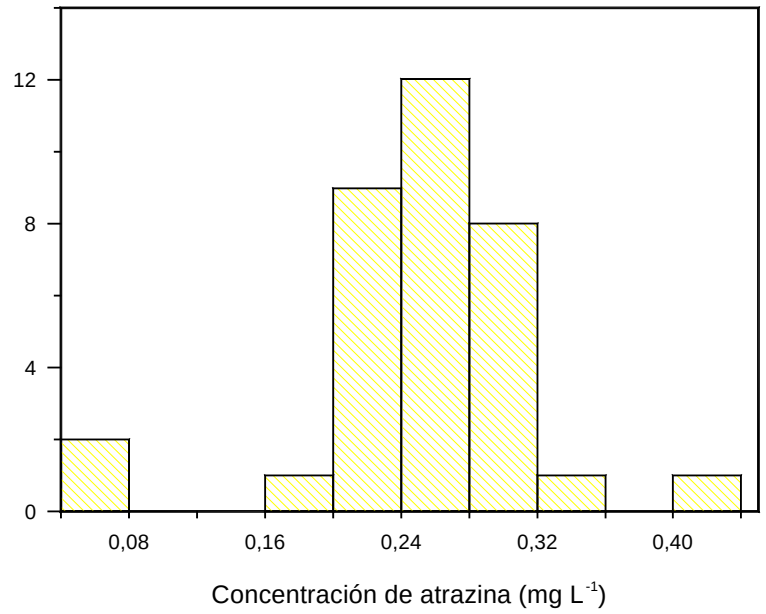

Figura E.2 Histograma de los resultados de los participantes

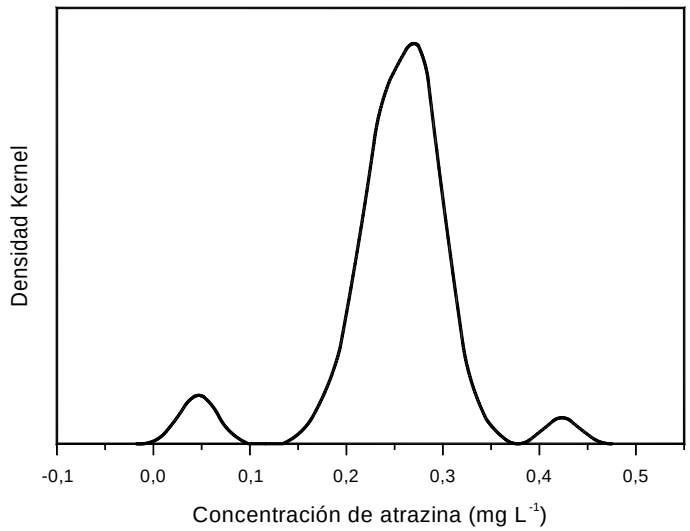

Figura E.3 Gráfico de densidad de Kernel para los resultados de los participantes

Tabla E.5 Estadísticas de resumen para el ejemplo de Atrazina

Procedimiento	Ubicación (promedio)	Desviación estándar	$u(x_{pt})$
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
Re muestreo (*Bootstrap*) (para media)	0,2503	0,0667	0,0113
Aritmético, Valores atípicos removidos	0,2588	0,0337	0,0061
Aritmético, Valores atípicos incluidos	0,2512	0,0672	0,0115

NOTA Diferentes paquetes de software comercial tienen diferentes procedimientos para calcular cuartiles, que pueden causar diferencias notables en *nIQR*. Las discrepancias menores de las cifras anteriores podrían ser causadas por tales diferencias o por diferentes procedimientos de redondeo.

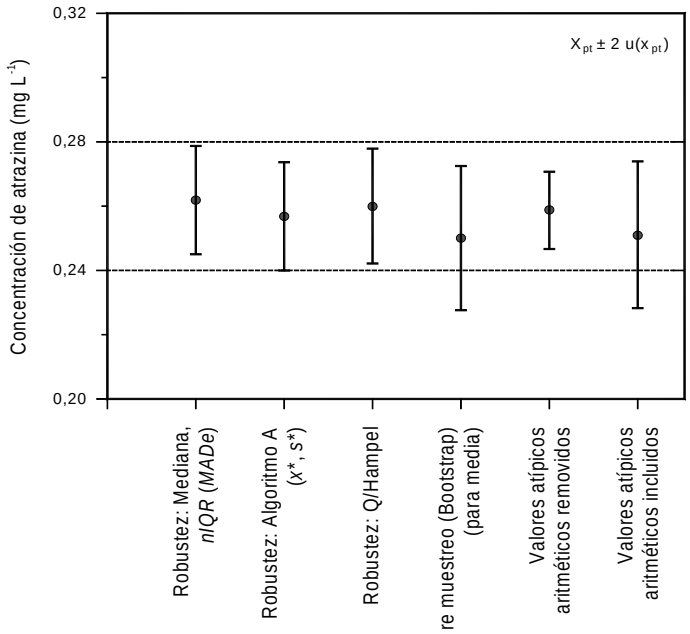

Figura E.4 Resumen de estadísticos robustos de la Tabla E.5

E.4 EJEMPLO GLOBAL PARA MERCURIO EN ALIMENTOS PARA ANIMALES

En una ronda de un programa de ensayos de aptitud, se dio instrucción a los participantes para que informaran sus resultados como lo harían de forma rutinaria e informaran su incertidumbre expandida (U_{lab}) y el factor de cubrimiento (k). Luego, el proveedor del ensayo de aptitud calculó la incertidumbre estándar (u_{lab}), como U_{lab}/k. Se asignaron los límites a las incertidumbres reportadas, siguiendo los criterios analizados en el numeral 9.8. En las Tablas E.6 y E.7 se presentan los datos correspondientes al total de mercurio en el alimento. En la Tabla E.6 se calculó la incertidumbre estándar u_{lab} a partir de la incertidumbre expandida del participante U_{lab}, dividiendo por el factor de cobertura reportado k; y se muestran como valores redondeados. Para el cálculo de las estadísticas de desempeño de la Tabla E.7, se emplearon valores sin redondear para u_{lab}. Para el participante con código L23 no se reportó el factor de cobertura y se empleó 1,732 (la raíz cuadrada de 3, redondeada).

Los indicadores de desempeño se calcularon empleando las técnicas descritas en la sección 9. Para todos los cálculos se empleo un valor de referencia como x_{pt} y σ_{pt} fue un valor de adecuación para el propósito con base en la experiencia previa. La incertidumbre del valor asignado fue la incertidumbre estándar combinada del valor de referencia más la incertidumbre debida a la homogeneidad (diferencias entre botellas).

x_{pt} = 0,044 mg/kg; $U(x_{pt})$ = 0,0082 mg/kg; σ_{pt} = 0,0066 mg/kg (=15 %);

En la gráfica de densidad de kernel Figura E.6, se muestra una distribución bimodal muy clara, debida a las diferencias de método. Ésta no causó impacto en la evaluación del desempeño, puesto que se empleó un valor de referencia como x_{pt} y se empleó un valor de adecuación para el propósito como σ_{pt}. Para este análisis, se eliminaron los resultados con un valor menor que (<).

Tabla E.6 Resultados del ensayo de aptitud de 24 participantes en el estudio IMEP 111

Código de lab.	Valor	U_{lab}	k	u_{lab}	Marca (Flag)	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

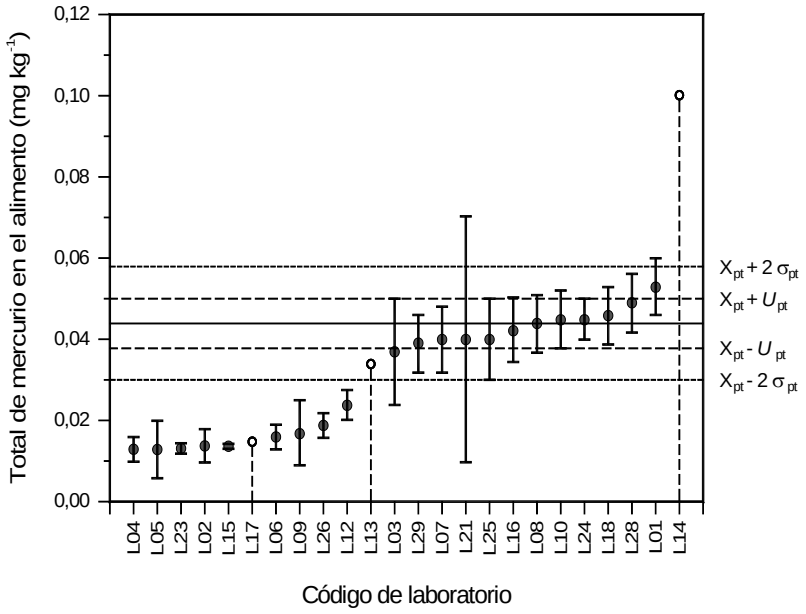

Figura E.5 Resultados de participantes e incertidumbres para resultados en IMEP 111 (datos de la Tabla E.6)

Las líneas discontinuas muestran $x_{pt} \pm U(xpt)$ y las líneas punteadas muestran $x_{pt} \pm 2\sigma_{pt.}$

Los círculos abiertos y las líneas verticales discontinuas muestran resultados ingresados como "menor que"

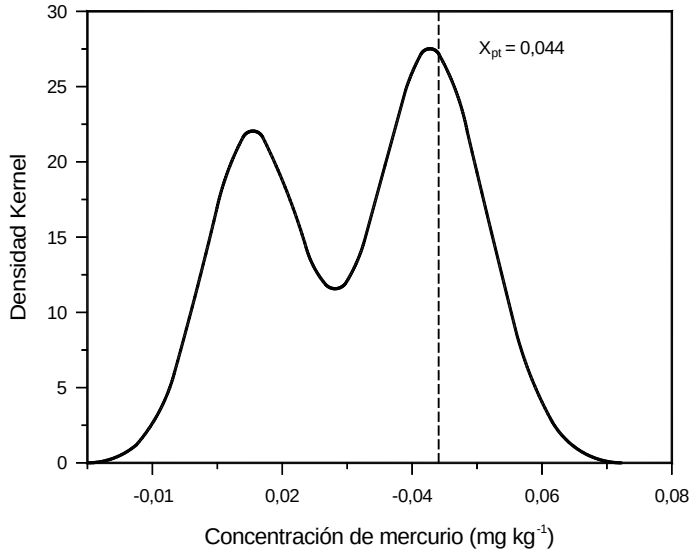

Figura E.6 Gráfica de densidad Kernel para resultados de participantes

Tabla E.7 Estadísticas de desempeño por varios métodos

Código de 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 del *European Commission Joint Research Centre, Institute for Reference Materials and Measurements, International Measurement Evaluation Program* (IMEP®), estudio 111.

E.5 Valor de referencia de un laboratorio único: Valor de agregados Los Ángeles (véase el numeral 7.5)

En la Tabla E.8 se presenta un ejemplo de datos que podrían obtenerse en una serie de ensayos sobre un ítem de ensayo de aptitud y un material de referencia certificado (MRC) muy similar, que tenga un valor de propiedad certificado de 21,62 unidades LA e incertidumbre asociada de 0,26 unidades LA. Este ejemplo muestra cómo se obtienen un valor de referencia y la incertidumbre para el ítem de ensayo de aptitud. Observe que la incertidumbre del valor certificado para el MRC incluye la incertidumbre debida a la falta de homogeneidad, transporte y estabilidad a largo plazo.

$$x_{pt} = 21{,}62 + 1{,}73 = 23{,}35 \quad \textit{unidades LA}$$

y,

$$u\left(x_{pt}\right) = \sqrt{0{,}26^2 + 0{,}24^2} = 0{,}35 \quad \textit{unidades LA}$$

donde 0,26 es la incertidumbre estándar del valor certificado del MRC y 0,24 es la incertidumbre estándar $\overline{d}$.

Tabla E.8 Cálculo de la diferencia promedio entre un MRC y un ítem de ensayo de aptitud y de la incertidumbre estándar de esta diferencia

Muestra	Ítem de ensayo de aptitud		MRC		Diferencia en valores promedio Ítem de ensayo de aptitud - MRC
	Ensayo 1 unidades *LA*	Ensayo 2 unidades *LA*	Ensayo 1 unidades *LA*	Ensayo 2 unidades *LA*	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 promedio, $\bar{d}$					1,73
Desviación estándar					1,07
Incertidumbre estándar $\bar{d}$ (desviación estándar / $\sqrt{20}$)					0,24

NOTA Los datos son mediciones de la resistencia mecánica del agregado, obtenidas del ensayo Los Ángeles (*LA*).

E.6 EJEMPLO DE TÉCNICA DE RE MUESTREO (*BOOTSTRAP*) PARA COLIFORMES EN UNA MUESTRA DE ALIMENTO (véase el numeral 7.7.2)

Un programa de ensayos de aptitud para coliformes en la muestra de alimento (leche) fue llevado a cabo con 35 participantes quienes realizaron cinco mediciones replicadas independientes. Se empleó la media de datos del registro CFU de cada participante para estimar el valor asignado y su incertidumbre. Se fijó un valor adecuado para el propósito igual a "0,25 log CFU/ml" como σ_{pt} mientras que la desviación estándar de la función kernel fue 0,75 σ_{pt} (cf. "bw" en el código R). La gráfica de densidad kernel (véase la Figura E.7) presenta una distribución asimétrica. Se aplicó el método de re muestreo (*Bootstrap)* (1 000 replicas) para calcular la moda y el error estándar correspondiente de la función de densidad kernel de la distribución de datos, establecidos como x_{pt} y $u(x_{pt})$, respectivamente. Se derivaron los siguientes valores:

$$x_{pt} = 3{,}79 \text{ y } u(x_{pt}) = 0{,}0922 \text{ en registro CFU/ml}$$

NOTA Puesto que $u(x_{pt}) > 0.3\ \sigma_{pt}$, se evaluaron los desempeños de los laboratorio empleando puntajes z'.

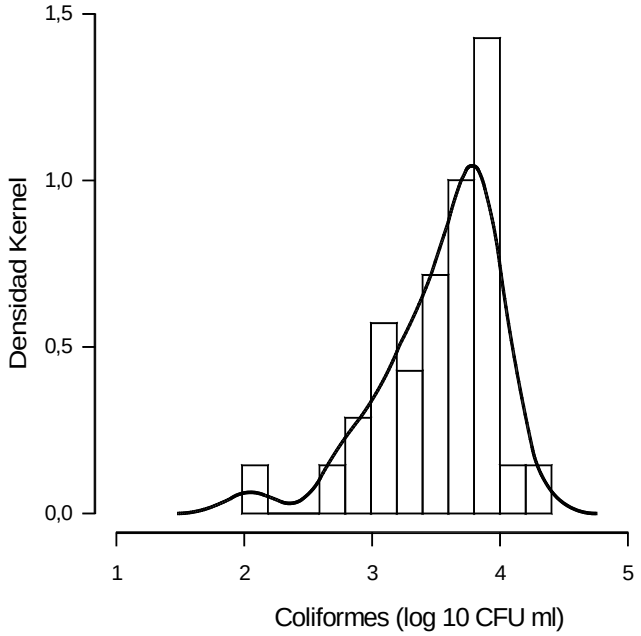

Figura E.7 Gráfica de densidad Kernel para los resultados de los participantes

Código R.3.1.1

```
###################################
(#DE BIBLIOTECA PARA DESCARGAR Y USAR)
###################################

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

#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)
<-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 Istituto Zooprofilattico Sperimentale delle Venezie - Food Microbiology PT “AQUA”

E.7 COMPARACIÓN DEL VALOR DE REFERENCIA Y LA MEDIA POR CONSENSO (Literal 7.8)

Como demostración del procedimiento en el numeral 7.8, para comparar un valor de referencia con la media robusta de los resultados de los participantes, considerar el ejemplo E.4 y los datos de la Tabla E.6.

En esta ronda de un programa de ensayos de aptitud, la media robusta x^* es 0,03161 y la desviación estándar robusta s^* es 0,0164, calculada con el Algoritmo A, después de la remoción de 3 resultados que tenían valores menor que (n = 24). Por lo tanto, se calcula la incertidumbre de la media robusta como:

$$u(x^*) = 1{,}25\left(s^* / \sqrt{n}\right)$$

$$u(x^*) = 1{,}25\left(0{,}0164 / \sqrt{24}\right) = 0{,}0042$$

A partir del numeral 7.8, ecuación 8, la incertidumbre de la diferencia entre x_{ref} y x^* es como sigue:

$$U_{diff} = \sqrt{u^2(x_{ref}) + u^2(x^*)} = \sqrt{0{,}0041^2 + 0{,}0042^2} = 0{,}0059$$

$$U_{diff} = 2(0{,}0059) = 0{,}012$$

$x_{diff} = x_{ref} - x^* = 0{,}044 - 0{,}032 = 0{,}012$ de modo que la diferencia es dos veces la incertidumbre de la diferencia.

No se recomienda ninguna acción, puesto que se entiende el sesgo en algunos métodos.

E.8 DETERMINACIÓN DE LOS CRITERIOS DE EVALUACIÓN POR EXPERIENCIA CON RONDAS ANTERIORES: TOXAFENO EN AGUA POTABLE (véase el numeral 8.3)

Existen dos proveedores de ensayos de aptitud que organizan programas de ensayos de aptitud para el pesticida Toxafeno (un pesticida) en agua potable. Durante un período de 5 años se han llevado a cabo 20 rondas de ensayos de aptitud donde hubo 20 ó más participantes, cubriendo los niveles regulados de toxafeno de 3 a 20 µg/L. En la Tabla E.9 se muestran los resultados de las 20 rondas de ensayos de aptitud, organizados de menor a mayor valor asignado. En las Figuras E.8 y E.9 se muestran los diagramas de dispersión para

la desviación estándar robusta relativa (DER o RSD por sus siglas en Inglés %) y la desviación estándar robusta (DE o SD por sus siglas en Inglés) para cada ronda de programas de ensayos de aptitud, comparadas con el valor asignado (a partir de la formulación). Las fórmulas para la línea de regresión lineal simple por mínimos cuadrados se muestran para cada figura. Las líneas de regresión por mínimos cuadrados se pueden determinar con software de hojas de cálculo disponible. (Es de anotar que también se verificó un modelo polimodal de segundo orden para la relación entre la desviación estándar y el valor asignado, pero el término cuadrático no fue significativo, indicando que la curva no es significativa en la línea; de modo que el modelo lineal simple resulta apropiado).

Es evidente que la DER (RSD) es muy constante en aproximadamente 19 % para todos los niveles, y que la línea de regresión para la desviación estándar es razonablemente confiable (R^2 = 0,82). Un organismo regulador puede optar por exigir que la desviación estándar para la evaluación del desempeño sea de 19 % del valor asignado (o quizás 20 %) o puede exigir el cálculo de la desviación estándar esperada, empleando la ecuación de regresión para la desviación estándar.

Tabla E.9 Rondas de ensayos de aptitud para Toxafeno en agua potable y con resultados *p* ≥ 20

Código del proveedor de ensayos de aptitud	Valor asignado	Media robusta	Desviación estándar	Recuperación de la media	DSR o RSD (% de AV)	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

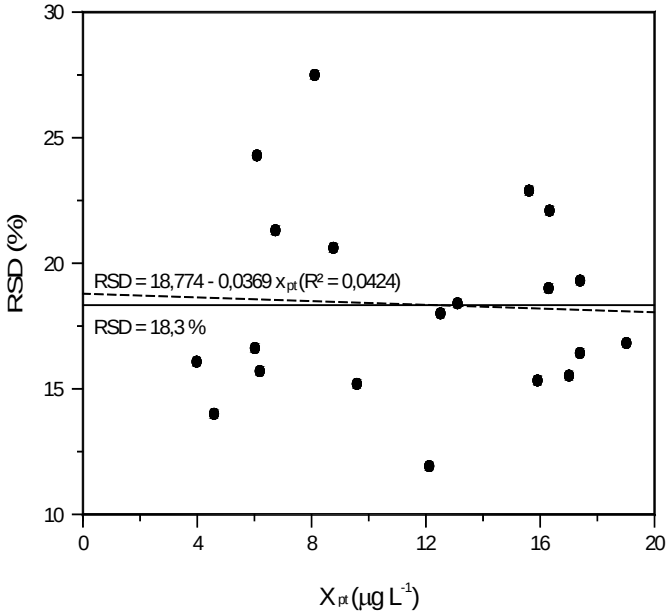

Figura E.8 Desviación estándar de los resultados de los participantes (%) vs valor asignado (µg/L)

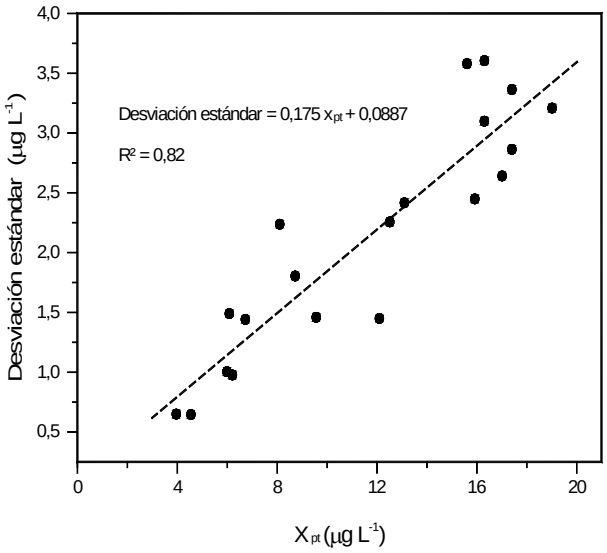

Figura E.9 Desviación estándar de participantes (µg/L) vs valor asignado (µg/L)

E.9 DE UN MODELO GENERAL: Ecuación Horwitz (véase el numeral 8.4)

Un modelo general común para aplicaciones químicas fue descrito por Horwitz[22,31]. Este enfoque ofrece un modelo general para la desviación estándar de reproducibilidad de métodos analíticos que se pueden emplear para derivar la siguiente expresión para la desviación estándar de la reproducibilidad:

$$\pm_R = 0{,}02 \times c^{0{,}8495}$$

en donde *c* es la concentración de los componentes químicos que se van a determinar en fracción másica.

Por ejemplo, para un programa de ensayos de aptitud para melanina en leche en polvo se emplean dos ítems de ensayo con niveles de referencia A = 1,195 mg/kg y B = 2,565 mg/kg (0,000 001 195 y 0,000 002 565). Esto produce las siguientes desviaciones estándar de reproducibilidad esperadas:

Ítem de ensayo de aptitud A en 1,195 mg/kg: σ_R = 0,186 mg/kg o σ_R relativo = 15,6 %

Ítem de ensayo de aptitud B en 2,565 mg/kg: σ_R = 0,356 mg/kg o σ_R relativo = 13,9 %

E.10 DETERMINACIÓN DEL DESEMPEÑO A PARTIR DE UN EXPERIMENTO DE PRECISIÓN: DETERMINACIÓN DEL CONTENIDO DE CEMENTO EN CONCRETO ENDURECIDO (véase el numeral 8.5)

Por lo general, el contenido de cemento en concreto se mide en términos de la masa en kilogramos de cemento por metro cúbico de concreto (es decir, en kg/m^3). En la práctica, el concreto se produce en grados de calidad que tienen contenido de cemento de 25 kg/m^3 por separado y es deseable que los participantes sean capaces de identificar el grado correctamente. Por esta razón, es deseable que el valor escogido σ_{pt} no sea mayor que la mitad de 25 kg/m^3 (σ_{pt} < 12,5 kg/m^3).

Un experimento de precisión produjo los siguientes resultados, para un concreto con un contenido de cemento promedio de 260 kg/m^3: σ_R = 23,2 kg/m^3 y σ_r = 14,3 kg/m^3. Se supone que se van a realizar m = 2 mediciones replicadas.

Por lo tanto, siguiendo la ecuación (9):

$$\pm_{pt} = \sqrt{23{,}2^2 - 14{,}3^2(1-1/2)}\ kg/m^3 = 20{,}9\ kg/m^3$$

Es posible que el objetivo de tener σ_{pt} < 25/2 kg/m^3 = 12,5 kg/m^3 no sea práctico.

NOTA En la NTC 3529-2 $\pm_R = \sqrt{\pm_L^2 + \pm_r^2}$ siendo σ_L el componente de la varianza debida a diferencias entre laboratorios.

En este ejemplo se podría calcular σ_L como:

$$\pm_L = \sqrt{\pm_R^2 - \pm_r^2} = \sqrt{\left(23{,}2^2 - 14{,}3^2\right)} = 18{,}3\ kg/m^3$$

E.11 GRÁFICAS DE BARRAS DE SESGOS ESTANDARIZADOS: CONCENTRACIONES DE ANTICUERPOS (véase el numeral 10.4)

En la Figura E.10 se muestran los puntajes z de una ronda de ensayos de aptitud con tres mensurandos relacionados (anticuerpos), representados como gráficas de barras. En la Tabla E.10 se muestran los datos para dos de los tres alérgenos. A partir de este gráfico, los laboratorios B y Z (por ejemplo) pueden ver que deberían buscar una causa del sesgo que afecta todos los tres niveles en aproximadamente la misma cantidad, mientras que los laboratorios K y P (por ejemplo) pueden ver en su caso que el signo del puntajes z depende del tipo de anticuerpo.

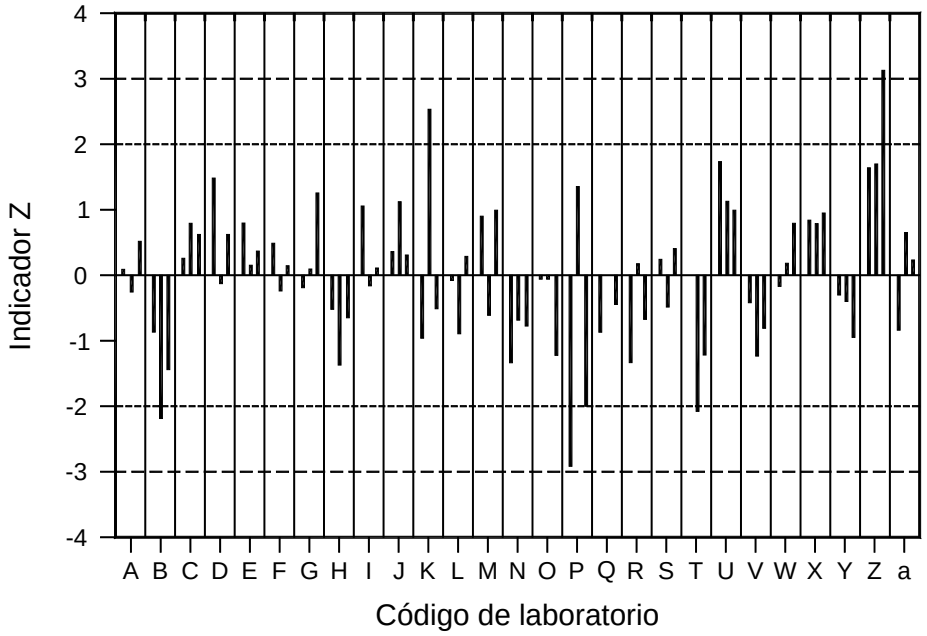

Figura E.10. Gráfica de barras para los puntajes z (4,0 a −4,0) para una ronda de programa de ensayos de aptitud en la que los participantes determinaron las concentraciones de tres anticuerpos IgE específicos para alérgenicos

E.12 GRÁFICA DE YOUDEN - CONCENTRACIONES DE ANTICUERPOS (véase el numeral 10.5)

En la Tabla E.10 se muestran los datos obtenidos al ensayar dos ítems de ensayo de aptitud similares para concentraciones de anticuerpos. En la Figura E.11 se muestran los resultados del indicador de desempeño (z) calculados con base en la media robusta y la desviación estándar robusta empleando el Algoritmo A.

La inspección de la Figura E.11 muestra a dos participantes (los números 5 y 23) en el cuadrante superior derecho y por consiguiente podrían tener sesgo positivo coherente. El laboratorio 26 tiene un resultado elevado del puntaje z en el ítem de ensayo de aptitud B y un resultado del puntaje z negativo de -0,055 en el ítem A y podría tener una repetibilidad deficiente.

Los participantes 5, 23 y 26 deberían tratar sus resultados como señales de “advertencia” y comprobar si sus resultados encajan en la siguiente ronda del programa. La revisión visual y el coeficiente de correlación indican una tendencia a coherencia de los resultados de los puntajes z (positivos o negativos), de modo que podría existir una oportunidad de mejorar en el método de medición con instrucciones más detalladas.

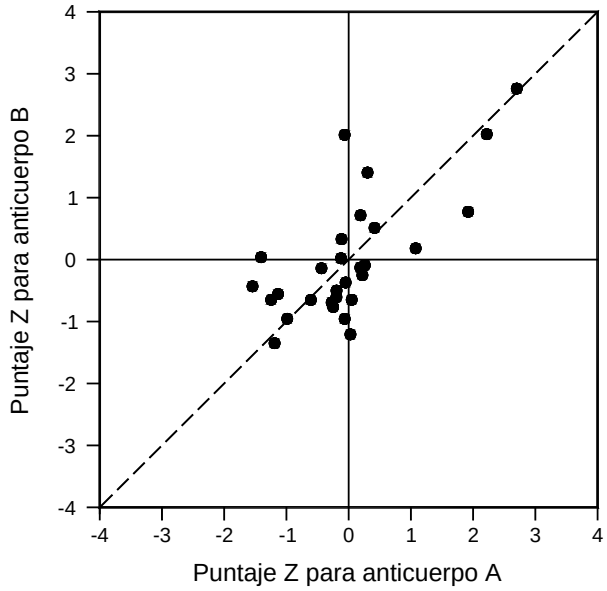

Figura E.11. Gráfica de Youden de los resultados del puntajes z (z-scores) a partir de la Tabla E.10

Tabla E.10 Datos y cálculos en concentraciones de anticuerpos para dos alérgenos similares

Laboratorio	Datos		puntajes z	
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

Continúa…

Tabla E.10 (Final)

Laboratorio	Datos		puntajes z	
i	Alérgeno A $x_{A,i}$	Alérgeno B $x_{B,i}$	Alérgeno A $z_{A,i}$	Alérgeno B $z_{B,i}$
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
Desviación estándar	3,29	2,90	1,00	1,00
Coeficiente de correlación	0,706		0,706	

NOTA 1 Los datos son cantidades de unidades (U) en miles (k) por litro (L) de muestra, donde se define una unidad por la concentración de un material de referencia internacional.

NOTA 2 Los resultados del puntaje z de esta tabla han sido calculados empleando valores no redondeados de los promedios y las desviaciones estándar robustas, sin usar los valores redondeados que se muestran al final de la tabla.

E.13 GRÁFICA DE LA DESVIACIÓN ESTÁNDAR DE REPETIBILIDAD: CONCENTRACIONES DE ANTICUERPOS (véase el numeral 10.6)

En la Tabla E.11 se muestran los resultados de la determinación de concentraciones de un cierto anticuerpo en ítems de ensayo de aptitud para suero. Cada participante realizó cuatro determinaciones replicadas, bajo condiciones de repetibilidad. Las fórmulas antes presentadas se emplean para obtener la representación gráfica de la Figura E.12. La gráfica muestra que varios de los laboratorios reciben señales de acción o advertencia.

Tabla E.11 Concentraciones de algunos anticuerpos en ítems de ensayo de aptitud para suero (cuatro determinaciones replicadas en un ítem de ensayo de aptitud por cada participante)

Laboratorio	Promedio kU/L	Desviación estándar 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

Continúa…

Tabla E.11 (Final)

Laboratorio	Promedio kU/L	Desviación estándar kU/L
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
Promedio robusto	1,57	
Desviación estándar robusta		0,34

NOTA Los datos son cantidades de unidades (U) en miles (k) por litro (L) de muestra, donde se define una unidad por la concentración de un material de referencia internacional.

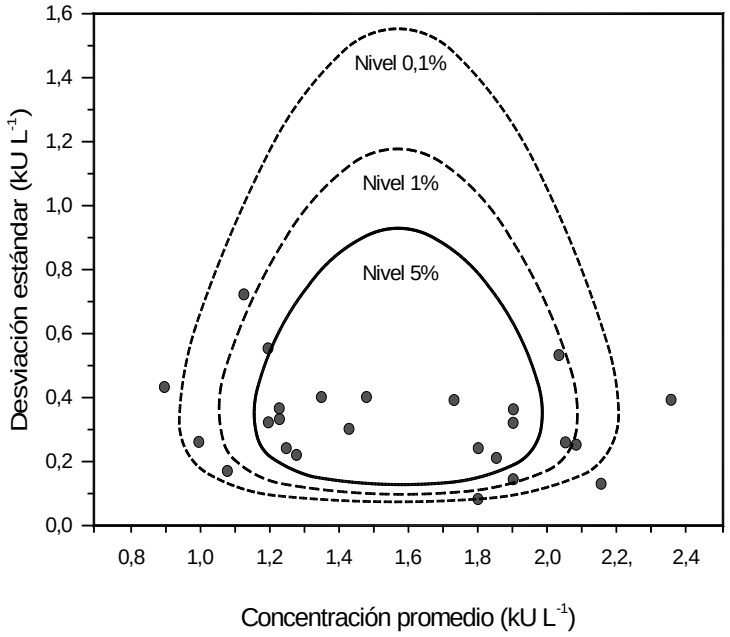

Figura E.12. Representación de las desviaciones estándar contra los promedios para 25 participantes (datos de la Tabla E.10)

E.14 MÉTODOS GRÁFICOS PARA HACER SEGUIMIENTO AL DESEMPEÑO EN EL TIEMPO (véase el numeral 10.8)

Para un participante puede ser útil rastrear su propio desempeño en el tiempo o que lo haga el proveedor del ensayo de aptitud. Una herramienta simple y convencional es una gráfica de control de calidad o diagrama Shewhart. Esta exige tener un indicador de desempeño normalizado, tal como el puntaje z ó el puntaje P_A y la participación en varias rondas. Este ejemplo es de un programa de ensayos de aptitud médica, para potasio sérico.

Este proveedor de ensayos de aptitud emplea un intervalo fijo para aceptación de 5 %, aunque con redondeo al próximo valor reportable (0,1 mmol/L), y no menor a ± 0,2 mmol/L. El proveedor de ensayos de aptitud emplea puntajes P_A en vez de puntajes z.

Tabla E.12. Puntajes P_A para 5 rondas de un programa de ensayos de aptitud, cada una con 3 ítems de ensayo de aptitud para potasio sérico

Código de la ronda	ítem del Ensayo de aptitud	Resultado	Valor asignado	Puntaje P_A	Promedio 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 se pueden diagramar con facilidad para revisión - se recomiendan dos tipos de gráficas:

- El diagrama de control de calidad del indicador de desempeño estandarizado para cada ronda, mostrando múltiples ítems de ensayo de aptitud en la misma ronda del ensayo de aptitud. Éste resaltará el desempeño en el tiempo, incluyendo algunas tendencias como se muestra en la Figura E.13.

- La gráfica de dispersión de indicadores de desempeño estandarizados contra valores asignados, para ver si el desempeño se relaciona con el nivel de concentración, con el fin de mostrar cualquier tendencia relacionada con el nivel del mensurando; como se muestra en la Figura E.14.

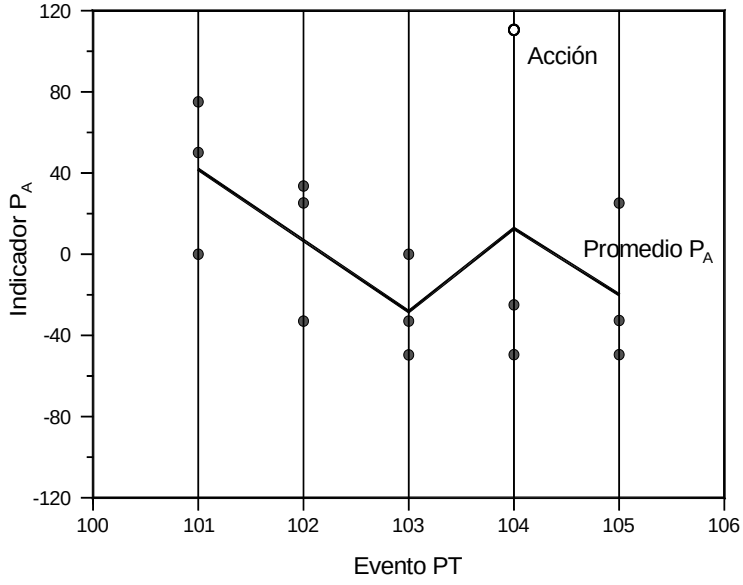

Figura E.13 Indicadores de desempeño para cada ronda (datos de la Tabla E.12)

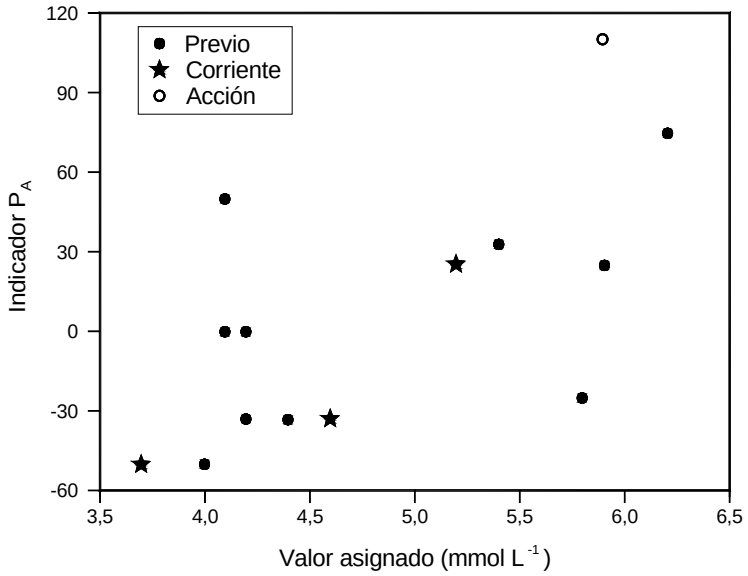

Figura E.14. Indicadores de desempeño para diferentes niveles del mensurando

E.15 ANÁLISIS DE DATOS CUALITATIVO; EJEMPLO DE UNA CANTIDAD ORDINAL: REACCIÓN DE LA PIEL A UN COSMÉTICO (véase el numeral 11)

Un programa de ensayos de aptitud incluye 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 clasifica de acuerdo con la siguiente escala:

1. sin reacción

2. enrojecimiento moderado

3. irritación o hinchazón significativa

4. reacción grave, incluida supuración o sangrado

Se distribuyeron dos ítems de ensayo de aptitud que constan de dos productos diferentes, etiquetados como producto A y producto B, y existen 50 participantes para cada producto. Los resultados de los participantes se presentan en la Tabla E.13 y se muestran gráficamente en la Figura E.15. La moda y la mediana se listan para los resultados de los participantes por cada ítem de ensayo de aptitud.

Tabla E.13. Resultados para dos ítems de ensayos de aptitud, irritación cutánea

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		

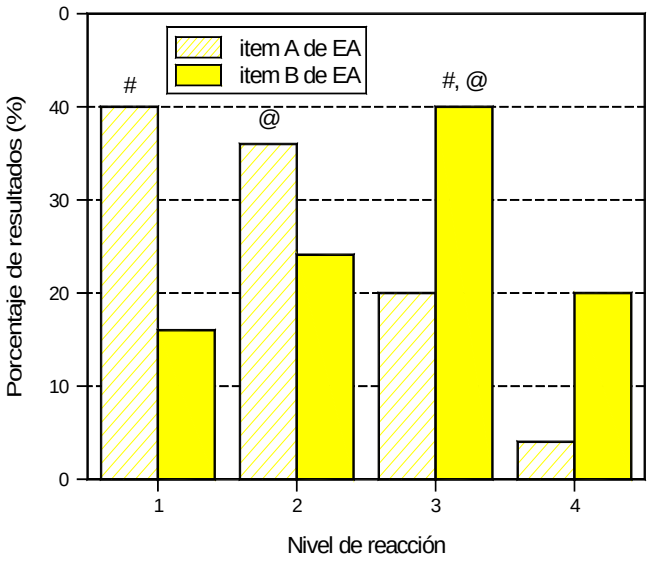

Figura E.15 Gráfica de barras de porcentaje de respuestas para dos ítems de ensayo de aptitud para irritación cutánea — # moda, @ mediana

Cabe anotar que se puede emplear la mediana o la moda como estadísticas de resumen para estos ítems de ensayo de aptitud y éstas sugieren que el nivel de reacción al producto B fue más grave que la reacción al producto A. El proveedor del ensayo de aptitud puede determinar que ocurrirían "señales de acción" para cualquier resultado que esté alejado más de una unidad ordinal con respecto a la mediana, en cuyo caso para el producto A, ocurren señales de acción para los 2 resultados (4 %) de “4” y para el producto B, ocurren señales de acción 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: Términos y Definiciones.*

[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. *Nondetects 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/e n/web-services/QHampel.html

DOCUMENTO DE REFERENCIA

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. *Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparison.* Geneve, Switzerland, ISO, 2015, 90p (ISO 13528:2015).