GLOBAL QUALITY
AND STANDARDS PROGRAMME

Instituto Nacional
de Metrología
de Colombia

Validación de Métodos en Análisis Químico Cuantitativo

Instituto Nacional de Metrología
Subdirección de Metrología Química y Biología

Instituto Nacional de Metrología de Colombia

María del Rosario González Márquez
Directora General

José Álvaro Bermúdez Aguilar
Secretario General

Laura Lorena Rivera Roa
Jefa Oficina Asesora de Planeación

Edna Julieth Villarraga Farfán
Subdirectora de Metrología Química y Biología

Claudia Angélica Guillén
Subdirectora de Servicios Metrológicos y Relación con el Ciudadano

Jairo Gustavo Ayala Forero
Subdirector de Metrología Física

Rodolfo Manuel Gómez Rodríguez
Jefe Oficina de Informática y Desarrollo Tecnológico

GQSP Colombia - Programa de Calidad para la Cadena Química

Juan Pablo Díaz-Castillo
Gerente de Proyecto y Oficial de Desarrollo Industrial de la ONUDI

Helen Jhoana Mier Giraldo
Asesora Técnica Regional

Javier Francisco Fernández Rodríguez
Coordinador Técnico Nacional

Autores

Diego A. Ahumada, Cristhian Paredes, Johanna Abella, Ivonne González

Revisión

Mesa de Trabajo Técnico-Científica de la Subdirección de Metrología Química y Biología del INM

Diagramación

Cristhian Paredes, Diego A. Ahumada

ISBN: 978-958-53805-9-2
Primera edición, Abril de 2023

Para mayor información, contacte a:

Instituto Nacional de Metrología - INM, Colombia
Av. Cra. 50 No. 26 – 55 Int. 2 CAN, Bogotá / Tel: +57 (601) 254 22 22 / www.inm.gov.co / contacto@inm.gov.co

Organización de las Naciones Unidas para el Desarrollo Industrial - ONUDI, Colombia
Calle 115 No. 5-50, Bogotá / Tel: +57 (601) 477 98 88 / www.gqspcolombia.org / info@gqspcolombia.org

Portada: Antorcha de plasma acoplado inductivamente, por Rebecca Dzombak, con la unidad **mol** del Sistema Internacional de Unidades.

RESUMEN

Los resultados de las mediciones se utilizan para tomar decisiones sobre asuntos relacionados con el comercio nacional e internacional de productos y servicios, la salud de los consumidores, el establecimiento de políticas o regulaciones, entre otros. La información que generan los métodos químicos de medición puede, por ejemplo, ayudar a establecer si la concentración de un contaminante en el agua potable excede el nivel al partir del cual su presencia se considera peligrosa, soportar procesos de exportación de alimentos frescos o procesados, determinar si un lote de producción de un cosmético no presenta un riesgo para el consumidor, o si una muestra biológica de un deportista contiene marcadores que evidencian el consumo de sustancias prohibidas. El impacto que pueden tener este tipo de decisiones obliga a que los métodos analíticos generen resultados de la calidad adecuada, lo cual se puede garantizar a través de la validación del método.

Esta guía tiene como objetivo describir algunas de las metodologías más empleadas en la validación de métodos químicos de tipo cuantitativo y brindar recomendaciones de tipo práctico que permitan realizar una evaluación adecuada y objetiva de los métodos de medición. En la actualidad existen distintos enfoques desde los cuales puede abordarse una validación, pero el propósito común siempre es definir requisitos analíticos y documentar pruebas objetivas que demuestren el alcance de esos requisitos, evaluando los factores que influyen sobre el proceso de medición. Dependiendo de la aplicación, para cada caso el proceso de validación tendrá características diferentes, por lo que en esta guía se pretende mostrar algunas de las estrategias más comunes empleadas en la validación de métodos.

El presente documento es una guía completa para la validación de métodos analíticos que puede ser aplicada en diversos campos de la ciencia y la tecnología. El documento se encuentra dividido en tres secciones que cubren los aspectos fundamentales de la validación de métodos. En la primera sección se presenta una visión holística del proceso de validación de métodos, explicando su importancia, los niveles de validación y algunos de los conceptos básicos que se necesitan para llevar a cabo este proceso de manera efectiva. En la segunda sección se presentan conceptos básicos de estadística y las pruebas estadísticas más comunes que se utilizan en la validación de métodos. Finalmente, la tercera sección cubre en detalle todos los parámetros de validación, incluyendo la precisión, la exactitud, la linealidad, la selectividad y la robustez, entre otros. Cada parámetro se explica en detalle y se ofrecen ejemplos y casos de estudio para ayudar al lector a comprender mejor la aplicación práctica de estos conceptos.

AGRADECIMIENTOS

Esta publicación se desarrolla en el marco del Programa de Calidad para la Cadena de Químicos (GQSP Colombia) en conjunto con el Instituto Nacional de Metrología de Colombia (INM). El GQSP Colombia es ejecutado por la Organización de las Naciones Unidas para el Desarrollo Industrial (ONUDI), y es financiado por la Secretaría de Estado para Asuntos Económicos de la Confederación Suiza (SECO) y el Ministerio de Comercio, Industria y Turismo (MinComercio), a través de Colombia Productiva.

Queremos agradecer a las personas que han participado en los cursos, talleres y diplomados ofertados por el INM desde su creación, pues su participación es las diferentes actividades de formación y su pasión por la metrología química inspiraron este documento. Agradecemos también a las personas que nos compartieron sus observaciones durante la consulta pública del documento, en particular a Patricia Heredia, Carlos García, María Gómez, Julián Velasco, Edgar Avendaño, Rocío Bojacá, Katherin Holguin y al profesor Jesús Ágreda de la Universidad Nacional de Colombia, entre otros.

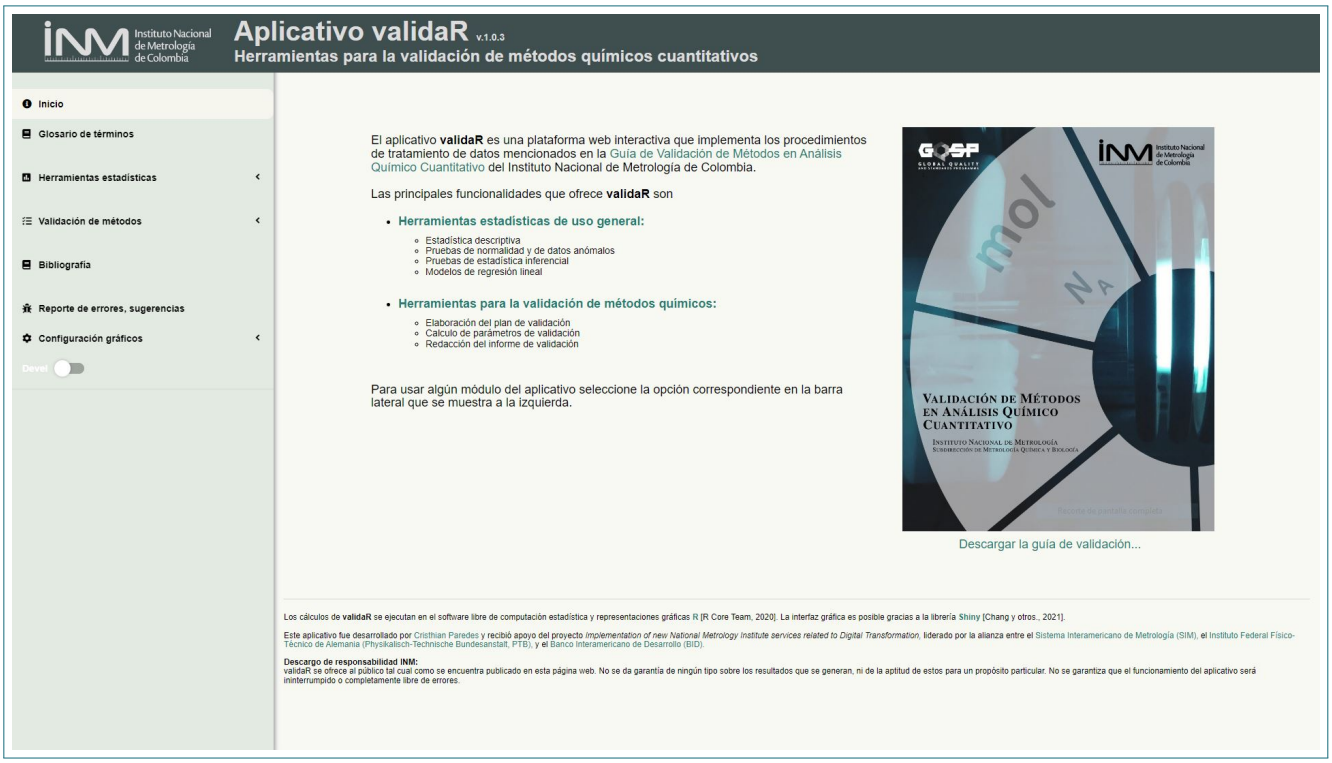
Finalmente, agradecemos de manera especial al Instituto Nacional de Metrología de Colombia por habernos brindado la oportunidad de impartir los cursos que han sido la principal inspiración para la creación de este libro. Esperamos que este libro sea de utilidad para todos aquellos que buscan mejorar sus conocimientos en validación de métodos de análisis químico cuantitativo.

APLICACIÓN WEB VALIDAR

Los procedimientos estadísticos mencionados en esta guía se implementaron en el aplicativo web validaR, disponible en la dirección `https://validar.inm.gov.co/validaR/`.

El propósito de validaR es facilitar el tratamiento estadístico de datos de laboratorio. Este aplicativo contiene herramientas generales de estadística descriptiva, representación gráfica de datos, estadística inferencial (pruebas de comparación), análisis de regresión lineal, interpolación de valores y cálculo de parámetros de validación, entre otras funciones. Se espera que el aplicativo sea de utilidad para las personas que validan métodos analíticos cuantitativos.

La aplicación funciona con el software libre R para computación estadística y representaciones gráficas (R Core Team, 2019). La implementación hace uso de la librería `Shiny`, que sirve para crear aplicaciones web interactivas utilizando R (Chang y col., 2020). La siguiente imagen muestra la página de inicio del aplicativo web.



En el desarrollo de validaR se recibió apoyo por parte del proyecto *Implementation of new National Metrology Institute services related to Digital Transformation*, liderado por la alianza entre el Sistema Interamericano de Metrología (SIM), el Instituto Federal Físico-Técnico de Alemania (*Physikalisch-Technische Bundesanstalt*, PTB) y el Banco Interamericano de Desarrollo (BID).

Agradecemos a la Oficina de Informática y Desarrollo Tecnológico del INM, por su colaboración en la publicación del aplicativo a través de la dirección web `https://validar.inm.gov.co/validaR/`.

ABREVIATURAS

Organizaciones

AOAC	Association of Official Analytical Chemists *(Organización Oficial de Químicos Analíticos)*
BIPM	Bureau International des Poids et Mesures *(Oficina Internacional de Pesas y Medidas)*
INM	Instituto Nacional de Metrología de Colombia
IUPAC	International Union of Pure and Applied Chemistry *(Unión Internacional de Química Pura y Aplicada)*
FDA	U.S. Food and Drug Administration *(Administración de medicamentos y alimentos de los Estados Unidos)*
JCGM	Joint Committee for Guides in Metrology *(Comité Conjunto para las Guías de Metrología)*

Términos de metrología, química analítica y estadística

MR	Material de referencia
MRC	Material de referencia certificado
VIM	Vocabulario Internacional de Metrología
LC	Límite de cuantificación
LD	Límite de detección
LMR	Límite máximo de residuos
RSD	Desviación estándar relativa
CV	Coeficiente de variación
% CV	Coeficiente de variación expresado en porcentaje
OLS	Mínimos cuadrados ordinarios
WLS	Mínimos cuadrados ponderados
ODR	Mínimos cuadrados de distancia ortogonal
GLS	Mínimos cuadrados generalizados

Técnicas analíticas

ICP-MS	Espectrometría de masas con plasma acoplado inductivamente
ICP-OES	Espectrometría de emisión atómica con plasma acoplado inductivamente
FAAS	Espectrometría de absorción atómica con llama
GC-μECD	cromatografía de gases con detector de micro captura de electrones
GFAA	Espectrometría de absorción atómica con horno de grafito
HPLC	Cromatografía líquida de alto desempeño
LC-MS/MS	Cromatografía líquida con detector támdem de espectrometría de masas

Compuestos químicos

MMA	Monometil arsénico

Índice general

1. Generalidades en la Validación de Métodos

La validación de metodologías analíticas tiene como objetivo garantizar que un método de análisis permite generar resultados confiables bajo las condiciones de trabajo disponibles en el laboratorio. Existen múltiples definiciones para la validación de métodos dentro de las que se destacan la del Vocabulario Internacional de Metrología (VIM) y la de la Unión Internacional de Química Pura y Aplicada (IUPAC, de sus siglas en ingles).

El VIM define la validación como una *"verificación, donde los requisitos especificados son adecuados para un uso previsto"*, entendiendo verificación como el proceso de aportar evidencia objetiva de que se satisfacen unos requisitos especificados (JCGM, 2012). Por otra parte, la IUPAC ha definido validación como *"confirmar la aptitud de un método de análisis para su uso y demostrar que un protocolo definido es adecuado para un propósito analítico particular, cuyo procedimiento es aplicable a un tipo específico de material y a una relación de concentración del analito definida"* (Thompson, Ellison y Wood, 2002). En la definición anterior se incluyen tres componentes importantes en las mediciones químicas: un protocolo definido, un tipo específico de material y una relación de concentraciones. Es decir que una validación se realiza para el conjunto de estos tres componentes, que según la IUPAC hacen parte de la definición de un sistema analítico (protocolo + matriz + analito/matriz).

Una tercera definición, acorde con la tendencia actual a trabajar bajo sistemas de gestión de calidad y que facilita los procesos de acreditación de los laboratorios, la confiere la administración de drogas y alimentos de Estados Unidos (FDA, de sus siglas en inglés), quienes definen la validación como el *"establecimiento de evidencia documental, de que un proceso específico proporciona un alto grado de seguridad de obtener un producto/servicio que cumpla con sus especificaciones y atributos de calidad"*. En esta definición es de resaltar que todo el proceso de validación debe venir acompañado de una serie de documentos que evidencien los procesos de planeación, obtención de los resultados, análisis estadístico e interpretación de los resultados obtenidos en la validación de los sistemas analíticos; todo esto con el propósito de brindar la evidencia a la que se refiere el Vocabulario Internacional de Metrología en su definición, y la confirmación del alcance del método al que se refiere la IUPAC.

El proceso de validación genera información que permite establecer las limitaciones de un método (por ejemplo, para conocer a que matrices no es aplicable o a partir de que niveles de concentración ya no produce resultados confiables) (Magnusson y Örnemark, 2014). El concepto de validación puede aplicarse considerando variables propias del método de medición y también otras variables, como el personal de laboratorio, el tipo de instrumentación, el tipo de reactivos y las condiciones ambientales, entre otros.

En las últimas décadas se han publicado un gran número de trabajos acerca de la validación de metodologías analíticas. La mayoría de estas describen diferentes definiciones y estrategias para realizar dicho procedimiento, pero en general se considera que los parámetros de desempeño mínimos de validación pueden ser: límite de detección (*LD*), límite de cuantificación (*LC*), intervalo de trabajo, linealidad, precisión, veracidad, selectividad y

robustez. Por otro lado, el creciente número de investigaciones relacionadas a la validación de sistemas analíticos ha hecho que existan diferentes aproximaciones para la evaluación de cada uno de estos parámetros. En el presente documento se explican diferentes alternativas que pueden ser empleadas en métodos de análisis químico tradicional y de análisis químico instrumental.

El primer capítulo de este documento brinda las generalidades y conceptos más relevantes relacionados con la validación de métodos, el segundo capítulo expone los procedimientos estadísticos más empleados en el cálculo de los parámetros de validación, y el tercer capítulo compila y propone diferentes estrategias para la evaluación de cada uno de los parámetros mencionados en el párrafo anterior, dando las principales recomendaciones en el uso y la interpretación de las diferentes aproximaciones existentes.

1.1 La importancia de validar un método analítico

Un método de medición debe ser apropiado para cumplir el objetivo que se le plantea. Los métodos químicos pueden tener distintos tipos de objetivos, pero en la mayoría de los casos el propósito es estimar la concentración de una o varias especies en una muestra analítica. Las implicaciones que giran en torno a la determinación cuantitativa de numerosas sustancias en distintas matrices requieren que los métodos analíticos empleados sean los adecuados, que la confiabilidad del resultado que genera sea conocida, y que la evidencia de todo esto se encuentre documentada correctamente. La confiabilidad de los resultados que genera un método puede evaluarse o entenderse a través de sus parámetros de desempeño, como su sesgo, precisión y selectividad.

Validar un método es verificar que sus parámetros de desempeño son aptos para el propósito u objetivo planteado. La verificación se hace por medio del aporte de evidencia objetiva que respalde las afirmaciones que se hagan sobre la aptitud del método (JCGM, 2012). En una validación se pretende evaluar si un método en particular genera resultados confiables bajo las condiciones de trabajo disponibles en el laboratorio. La evidencia objetiva por lo general comprende el análisis y la documentación de los resultados experimentales para cada uno de los parámetros de desempeño del método.

Lo parámetros de desempeño se seleccionan de acuerdo con los requerimientos que deben estar especificados en función del alcance del método, es decir, su aplicación final (Barwick y Prichard, 2011). Por lo anterior, es importante conocer las decisiones que se tomarán con la información que genere el método de medición. Adicionalmente, la información que se genera durante una validación sirve para (i) realizar una adecuada estimación de la incertidumbre, (ii) establecer herramientas de control dentro del laboratorio, (iii) asegurar el correcto funcionamiento del método en el tiempo, y (iv) mejorar continuamente el desempeño del mismo (Thompson, Ellison y Wood, 2002).

La importancia de evidenciar la aptitud de los métodos recae sobre el hecho de que los resultados producidos en un laboratorio con frecuencia se usan en la toma de decisiones importantes que pueden afectar la integridad de personas y ecosistemas, o que pueden tener serias connotaciones económicas. Considere, por ejemplo, las siguientes preguntas cuya respuesta está relacionada con algunos métodos químicos analíticos:

- ¿Tiene un producto agropecuario una concentración de residuos de plaguicidas por encima de los máximos permitidos para la ingesta humana? (European Union, 2020).
- ¿Es segura para la fauna acuática una determinada vertiente de aguas residuales? (Zhang y col., 2016).
- ¿Los plastificantes presentes en determinados juguetes para niños los hacen inadmisibles para su importación? (Ashworth y col., 2018)

Producir resultados cuya confiabilidad sea conocida disminuye el riesgo de que se tomen decisiones incorrectas con base en información errónea. Entender el alcance de un resultado dado permite evaluar si la información que contiene es lo suficientemente buena como para justificar la toma de una decisión. Por otro lado, usar métodos validados es un requisito para la acreditar la competencia de los laboratorios de ensayo (Norma ISO/IEC 17025, 2017).

1.2 Tipos de métodos: Validación y verificación

Los métodos analíticos se pueden clasificar en cuatro grupos:

i. Métodos nuevos o desarrollados por el laboratorio que provienen del desarrollo interno de una metodología que satisface una necesidad particular.
ii. Métodos normalizados que ya atravesaron un proceso de validación interno y externo.
iii. Métodos desarrollados y publicados por terceros, por ejemplo, en revistas especializadas.
iv. Métodos normalizados que se modifican o que se aplican fuera de su alcance, por ejemplo, cuando se aplica a nuevos tipos de muestra.

En el caso de los métodos normalizados, el proceso de normalización asegura que se conozcan las características de desempeño que se esperan para dichos métodos, debido a que ya se ha evaluado si dichas características son lo suficientemente buenas para el propósito del método. Esto quiere decir que los métodos normalizados ya se encuentran validados. En este caso el laboratorio solo debe verificar que el método funciona bien cuando lo adapta a sus condiciones particulares de trabajo. Este proceso suele ser más sencillo que validar un método analítico de otro tipo. Es común que los métodos normalizados incluyan información respecto a los parámetros de desempeño del método que deben evaluarse y sus criterios de aceptación respectivos. Esto simplifica enormemente el proceso de validación porque el laboratorio no debe establecer el requisito analítico y usualmente no se deben verificar todos los parámetros que aplican al método, como si es normal durante una validación completa.

Por otro lado, considerando que dentro de la validación de métodos es indispensable el establecimiento de un requisito que permita asegurar que la medición es apta para su uso, algunas áreas o sectores cuentan con regulaciones o normas que indican criterios de aceptación típicos. En muchos otros casos esto no es una tarea sencilla y se debe acudir a valores aceptados por consenso, como es el caso de la relación de Horwitz para predecir la reproducibilidad máxima típica de un método analítico (ver Sección 3.5.1.1.2).

Durante el desarrollo de métodos nuevos de medición algunos parámetros de desempeño se evalúan cada cierto tiempo, hasta que el método satisface los requerimientos de su fin propuesto. Sin embargo, la determinación de los parámetros de desempeño en la etapa de desarrollo del método no suelen ser válidos para soportar su validación. Las condiciones en las que se desarrolla y optimiza el método suelen diferir a las condiciones de rutina, por ejemplo, respecto al personal o el nivel de control de las variables de influencia. Esto hace necesario que una vez se haya optimizado el protocolo de medición se deba determinar los parámetros de desempeño del método bajo las condiciones de un análisis de rutina, y se establezca hasta qué punto los parámetros obtenidos son adecuados para el fin propuesto.

Los métodos modificados (en el grupo *iv*) se encuentran entre los métodos normalizados y los métodos desarrollados por el laboratorio. Un método puede modificarse para ser usado en un intervalo de concentraciones diferente al establecido, para incluir nuevos analitos en la misma matriz, o para considerar nuevas matrices. En este caso el método debe validarse, pero se tiene la ventaja de que en la mayoría de las veces el alcance, la regulación o las decisiones del método normalizado son similares a las del modificado, y por lo tanto se pueden emplear los criterios de aceptación del método normalizado, es decir nuevamente se simplifica el proceso.

Finalmente, los métodos que ya se encuentran validados deben someterse a una verificación cada vez que se produzca un cambio brusco en sus condiciones, para asegurar que sus parámetros de desempeño se mantienen aceptables (por ejemplo, cuando se ha adquirido un nuevo instrumento para realizar el análisis). La Organización Oficial de Químicos Analíticos (AOAC, por sus siglas en inglés) resalta que la estabilidad de la validación (es decir, la permanencia de los parámetros de desempeño del método) debe verificarse periódicamente utilizando idealmente un material de referencia de estabilidad conocida (AOAC, 2012). Esto también aplica para monitorear el desempeño de los métodos normalizados cuando estos son utilizados en el laboratorio (Norma ISO/IEC 17025, 2017).

1.3 Niveles de validación

Un método analítico puede validarse bajo distintos criterios según el nivel de validación que se escoja. Los niveles de validación varían de acuerdo con las necesidades que busca cubrir el método. Un nivel de validación alto involucra condiciones y procedimientos más elaborados y costosos que un nivel de validación bajo; pero de la misma manera, conduce a evidencia más completa y objetiva sobre el desempeño del método y permite que este pueda ser empleado en distintos contextos. En estos términos, se debe escoger un nivel de validación que permita optimizar la cantidad de tiempo y recursos que se dedican a este proceso, asegurándose de que se pueda concluir de manera objetiva sobre si el método en cuestión es o no adecuado para el propósito de la medición.

La FDA define cuatro niveles jerárquicos para el nivel de validación que se ilustran en la Figura 1.1 (FDA, 2019). Los niveles de emergencia e intralaboratorio se desarrollan por un único laboratorio mientras los niveles multilaboratorio (o interlaboratorio) y de estudio colaborativo involucran un conjunto de laboratorios capaces de aplicar el método. Para lograr la validación completa de un método analítico es necesario que este proceso se realice en el marco de un estudio colaborativo que conforma el nivel más alto de la validación.

Figura 1.1: Niveles jerárquicos de validación para un método de medición analítico (FDA, 2019). Un método se considera completamente validado cuando ha alcanzado el tope de la pirámide.

El nivel uno (nivel de emergencia) se utiliza cuando el método analítico debe quedar disponible de manera urgente por alguna situación especial. En este nivel se contempla el mínimo número de parámetros de validación necesarios que permitan evaluar si los resultados del método son apenas confiables para el alcance establecido del método. Este nivel es común cuando se hacen extensiones del alcance de métodos normalizados para considerar matrices adicionales.

En el segundo nivel (nivel intralaboratorio) se evalúan todos los parámetros de validación recomendados para el tipo de método que se está validando. Este proceso involucra únicamente un laboratorio. La validación de un método analítico al nivel uno o dos solo considera las condiciones particulares del laboratorio que realiza la validación. Parámetros como la reproducibilidad del método pueden llegar a ser estimados de manera un poco optimista, porque no es posible estimar como podría ser el desempeño del método cuando sea aplicado por otro laboratorio. Este es el nivel más común que se escoge para la validación de métodos desarrollados por el laboratorio.

El nivel tres de validación (interlaboratorio o multilaboratorio) requiere la participación de al menos un laboratorio adicional al que ya ha realizado la validación del método a nivel intralaboratorio. Idealmente debe contarse con un método de referencia para hacer una comparación de los resultados que produce el método que se está validando. Este nivel proporciona una visión más amplia de la variabilidad del método al introducir un conjunto completo de nuevas condiciones que están disponibles en los laboratorios adicionales. A partir de este nivel puede considerarse que el estudio tiene un enfoque dirigido hacia el desempeño del método como tal, en lugar del desempeño de los laboratorios que lo utilizan.

El nivel cuatro de validación consta de un estudio colaborativo en el que participan al menos ocho laboratorios que previamente ya han realizado la validación del método hasta el nivel dos o tres. Esta estrategia es la necesaria para lograr la validación total de un método analítico. Según las conclusiones que se generen durante el proceso

de validación, en el marco de un estudio colaborativo, se genera suficiente información para obtener un método normalizado que puede distribuirse entre los laboratorios que lo necesiten. Así, los laboratorios que adaptan un método normalizado solo deberán verificar que los parámetros de desempeño del método no se deterioren al adaptar el método a sus condiciones de trabajo.

Adicionalmente, la IUPAC resalta que si un método va a ser utilizado por un gran número de laboratorios resulta más conveniente realizar su validación en un estudio colaborativo, en lugar de que los laboratorios realicen la validación a nivel intralaboratorio de manera independiente (Thompson, Ellison y Wood, 2002). Adicionalmente, en algunos contextos es necesaria la validación de un método hasta el cuarto nivel para que dicho método pueda ser utilizado en concordancia con la regulación vigente.

1.4 Plan e informe de validación

Antes de iniciar la validación de un método se deben definir los requerimientos necesarios para su desarrollo, los parámetros a evaluar, los experimentos que se deben realizar, los tratamientos estadísticos y los criterios de aceptación de los parámetros que permitan evidenciar si el método es apropiado para el fin previsto. Esta información se documenta en el plan o protocolo de validación, el cual contiene todos los detalles relacionados con el proceso de validación, incluyendo las metodologías de tratamiento de datos para obtener las conclusiones.

Una vez se desarrollen todos los experimentos definidos en el plan de validación y se cuente con los resultados obtenidos, es necesario documentar en un informe o reporte de validación el tratamiento estadístico y cálculos realizados, los cuales permitirán determinar el cumplimiento frente a los criterios de aceptación definidos para cada uno de los parámetros evaluados.

En el Anexo A se describen algunos lineamientos generales respecto al contenido del plan e informe de validación.

2. Herramientas Estadísticas para la Validación

Los resultados que se obtienen en las mediciones repetidas de un analito en una alícuota de muestra presentan variación a pesar de que las mediciones se realicen bajo condiciones experimentales aparentemente controladas y constantes.[1] Muchas variables desconocidas influyen en el valor de un resultado de medición y son la causa del **error aleatorio**: una variación intrínseca del sistema analítico que compone la parte impredecible del error de medida (JCGM, 2012).

El uso de herramientas estadísticas es necesario durante la validación de métodos químicos porque la estadística permite recolectar, organizar, relacionar, representar y analizar datos con error aleatorio. En otros términos, la evidencia objetiva que se requiere para demostrar si un método de medición es adecuado para un propósito establecido se obtiene por medio de tratamientos estadísticos que se aplican a los datos de validación.

En la Sección 1.4 se menciona que cuando se elabora un plan de validación, se deben definir los protocolos para el tratamiento estadístico de los datos de la validación y los criterios de aceptación de los valores que resulten de estos tratamientos. Este capítulo describe algunos procedimientos estadísticos útiles para el tratamiento de datos de mediciones y que conforman la base de los cálculos de los parámetros de validación que se detallan en el siguiente capítulo.

2.1 Variables estadísticas y descriptores estadísticos

El valor real de una propiedad en un material puede ser único y exacto, pero dicho valor es imposible de conocer cuando se realiza una medición porque todas las mediciones son afectadas por error aleatorio. Debido a esto conviene ver y tratar a todos los mensurandos como **variables estadísticas aleatorias continuas**: atributos cuyos valores están afectados por fenómenos aleatorios y que pueden tomar cualquier valor numérico dentro de un intervalo.

Un valor que toma una variable estadística aleatoria representa solo un **individuo** que hace parte de un conjunto de elementos más grande, denominado **población estadística** y que comprende todos los posibles valores que puede tomar la variable. Un subconjunto conformado por algunos individuos de una misma población estadística se denomina **muestra estadística**. Las muestras estadísticas heredan características de las poblaciones de las que provienen, solo que con pequeñas discrepancias que surgen como consecuencia del muestreo aleatorio.

En las mediciones químicas los conceptos de población estadística y muestra estadística pueden referirse a conjuntos tangibles (y finitos) o a conjuntos intangibles (e infinitos):

[1] Mediciones repetidas es diferente de réplicas de medición. En las repeticiones se suelen analizar alícuotas de un mismo extracto, pero las réplicas de medición implican aplicar todo el procedimiento de medición a una porción de muestra nueva de una misma muestra analítica. La dispersión en el primer caso es menor que en el segundo.

- Un ejemplo de conjunto tangible es un lote de producción de un medicamento que pasa por un determinado control de calidad. Analizar todas las unidades del lote suele ser muy inconveniente, en especial si el control implica un análisis químico destructivo o una gran cantidad de muestra. En la práctica se seleccionan aleatoriamente algunas unidades para el control de calidad: el conjunto de unidades recolectadas conforman una muestra estadística de las unidades de todo el lote. El lote de producción completo es la población estadística y tiene un número finito de elementos.
- Un ejemplo de conjunto intangible son los resultados de medición que se obtienen para una misma porción de muestra analítica en condiciones de repetibilidad. El conjunto finito de resultados de las mediciones repetidas conforman una muestra aleatoria de la población estadística que comprende todos los resultados de medición que se obtendrían si las mediciones se repitieran un número infinito de veces. En este caso la población estadística contiene un número de individuos infinito y el conjunto es intangible, imposible de alcanzar en su totalidad.

La distribución de las variables estadísticas aleatorias puede representarse gráficamente con un histograma como el que se muestra en la Figura 2.1. Este gráfico toma el **espacio muestral de la variable** (el intervalo de valores que puede tomar la variable), lo divide en secciones del mismo tamaño, y para cada sección levanta una barra vertical cuya altura es proporcional al número de datos que caen en cada intervalo (la frecuencia relativa). El histograma de la Figura 2.1 incluye una curva continua sombreada que representa la función de densidad de probabilidad de la variable. Esta curva describe la probabilidad de que la variable estadística presente un valor en un determinado intervalo de valores. El área total bajo la curva de densidad de probabilidad es igual a uno (o 100 %).

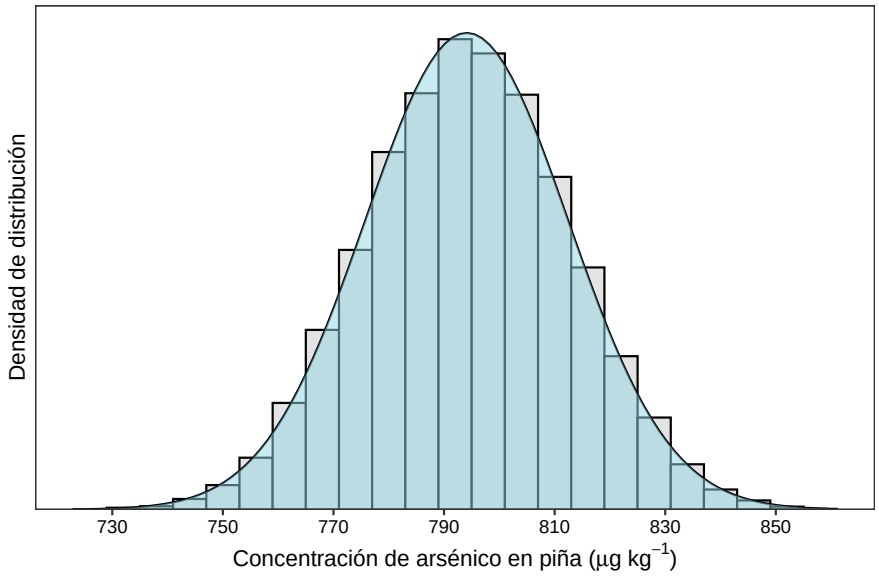
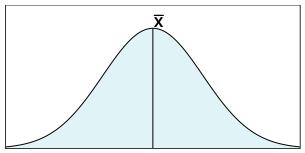
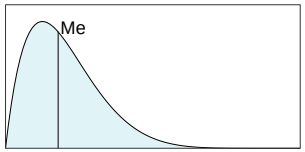
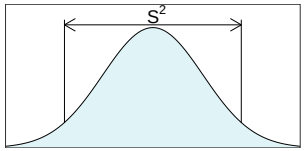
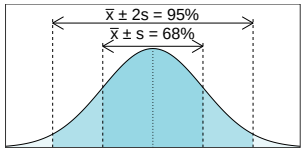
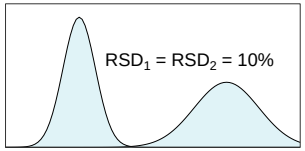
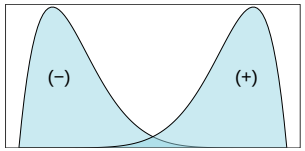
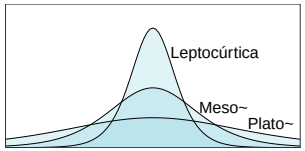


Figura 2.1: Población de resultados posibles de un mensurando en una muestra analítica. El diagrama de barras es el histograma y la curva sombreada es la curva de densidad de probabilidad de la distribución.

Las características más relevantes de una variable estadística se puede resumir numéricamente usando **descriptores estadísticos**. Los descriptores poblacionales resumen información de toda la población estadística mientras los descriptores muestrales se encargan solo de una muestra estadística. Los descriptores poblacionales se conocen como parámetros y se representan con letras del alfabeto griego (como μ y σ). Los descriptores muestrales se llaman estadísticos y se representan con letras del alfabeto latino (como $\bar{x}$ y s). Los descriptores muestrales se pueden calcular con resultados de una muestra estadística, mientras los descriptores poblacionales por lo general permanecen desconocidos.

Los descriptores estadísticos más comunes son **medidas de tendencia central**, que indican hacia que valor tienden a agruparse los individuos de un conjunto, y **medidas de dispersión**, que indican que tan agrupados se encuentran los datos entre sí. Otro tipo de descriptores menos comunes son las medidas de distribución, que dan información sobre la forma en la que se ordenan los datos en el espacio de la variable. La Tabla 2.1 resume los descriptores estadísticos más comunes.

Descriptor estadístico	Definición matemática	Descripción	Representación gráfica
Medidas de tendencia central			
Media aritmética o promedio	$\bar{x} = \frac{\sum x_i}{n}$	Representa el centro de gravedad de la distribución de la variable. Se ve muy afectado por valores extremos por lo que no es un estimador robusto frente a la presencia de datos anómalos	
Mediana (Me)	Es el valor de la mitad luego de organizar la serie de datos de manera ascendente.	Cuando el conjunto de datos es impar, la mediana corresponde al valor central. Cuando el número de datos es par, la mediana es el promedio de los dos valores centrales. La mediana se ve poco afectada por valores extremos o anómalos y se considera un estimador robusto. En una distribución normal la mediana coincide con la media aritmética.	
Medidas de dispersión			
Varianza	$s^2 = \frac{\sum (\bar{x} - x_i)^2}{n-1}$	Es una medida de las distancias de los valores del conjunto respecto a su promedio, elevadas al cuadrado. La varianza no tiene las mismas unidades que los datos del conjunto.	
Desviación estándar	$s = \sqrt{s^2}$	Es la raíz cuadrada de la varianza. La desviación estándar tiene las mismas unidades que los datos del conjunto y eso facilita su interpretación numérica. Una distribución normal presenta el 68.27 % de los datos en el intervalo $\bar{x} \pm s$, y aproximadamente 95.45 % de los datos en el intervalo $\bar{x} \pm 2s$.	
Desviación estándar relativa (RSD) (Coeficiente de variación, CV)	$CV\ (\%) = \frac{s}{\bar{x}}$	Es la desviación estándar del conjunto dividido entre su media aritmética. Es un valor proporcional adimensional que se suele representar en forma de porcentaje. El valor de CV de un conjunto no se suele afectar cuando se cambian las unidades en las que se expresa la variable estadística.	
Medidas de distribución			
Asimetría (Coeficiente de Fisher)	$CA_F = \frac{\sum (x_i - \bar{x})^3}{n \cdot s^3}$	Mide de que lado del promedio se encuentra el mayor número de datos. Valores positivos indican que la mayoría de los datos están por encima del promedio (asimetría a la derecha) mientras un valor negativo indica asimetría a la izquierda. Las distribuciones normales son simétricas ($CA_F = 0$).	
Curtosis	$K = \frac{\sum (x_i - \bar{x})^4}{n \cdot s^4} - 3$	Predice la distribución de los datos alejados de la media. Valores positivos implican una mayor acumulación de datos cerca del promedio (forma apuntada) y la muestra se denomina leptocúrtica. Valores negativos implican una distribución platicúrtica donde la forma es más aplanada. La distribución normal se dice mesocúrtica.	

En las ecuaciones, $\bar{x}$ es la media aritmética, x_i es el i-ésimo dato de la serie, n es el número total de datos, s^2 es la varianza, s es la desviación estándar, *RSD* es la desviación estándar relativa, *CV* es el coeficiente de variación, CA_F es el coeficiente de Fisher de asimetría y K representa la curtosis.

Tabla 2.1: Descriptores estadísticos de tendencia central, de dispersión, y de distribución.

Ejemplo 1: Cálculo de descriptores estadísticos.

El módulo **Estadística descriptiva** de la sección **Herramientas estadísticas** del aplicativo validaR calcula distintos descriptores estadísticos de una serie de datos que se haya ingresado en alguna de las columnas de la tabla del módulo **Inicio, ingreso de datos**, que se encuentra en la misma sección.

La siguiente tabla contiene datos de fracción másica de ion plomo (ν_{Pb}) en una disolución acuosa:

ν_{Pb} /[mg L^{-1}]	12.04	12.07	12.10	12.00	12.16	12.02	12.03	11.99

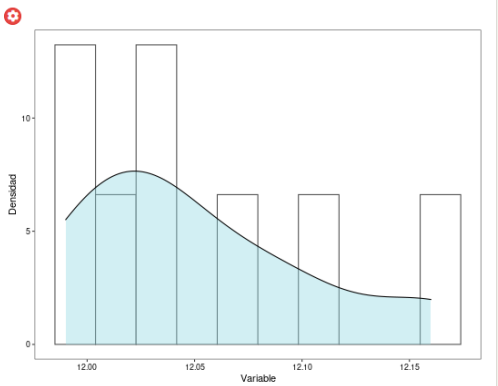
El siguiente recuadro muestra los resultados de validaR de los descriptores estadísticos mencionados en la Tabla 2.1, junto con un histograma que se construye con los datos ingresados. Tenga en cuenta que los datos de la tabla se deben ingresar verticalmente en la misma columna del aplicativo.

Resultados:

Descriptores estadísticos

Estadístico	Valor
Promedio	12.0512
Mediana	12.0350
Moda	12.0474
Desviación estándar	0.057
Desviación estándar relativa	0.5 %
Varianza	0.003
Número de datos	8
Rango	0.1700
Intervalo	
~ valor mínimo	11.99
~ valor máximo	12.16
Curtosis	2.65
Asimetría	0.841

Histograma



2.2 Estadística inferencial

En la Sección 2.1 se habló de población estadística y muestra estadística para referir al conjunto completo de individuos bajo estudio y a un subconjunto tomado aleatoriamente de dicho conjunto, respectivamente. La información de la muestra estadística por lo general es la única que se puede conocer experimentalmente mientras la información completa y detallada de la población estadística permanece inaccesible. La estadística inferencial es la rama de la estadística que utiliza la información de una muestra estadística para hacer predicciones sobre la población estadística de la que proviene la muestra, considerando el efecto indistinguible que tiene el error aleatorio sobre los datos.

Las conclusiones en estadística inferencial pueden generarse por medio de pruebas de hipótesis (o pruebas de comparación), en las que se contrastan dos afirmaciones o *hipótesis* que son contradictorias entre sí y que buscan anular o reforzar una pregunta de investigación dada. Las conclusiones sobre las hipótesis planteadas se caracterizan con un valor de probabilidad al que se le conoce como **nivel de confianza**. La pertinencia del uso de una determinada prueba de comparación suele depender de que se cumplan algunos supuestos que deberían comprobarse para el conjunto de datos que se analiza. Las pruebas de hipótesis más comunes se estudian en las Secciones 2.2.3 a la 2.2.7.1.

2.2.1 Distribución normal y probabilidad estadística

Variables estadísticas que presentan distribuciones de probabilidad similares pueden tratarse con herramientas de análisis de datos similares. La distribución de probabilidad más común entre las variables aleatorias continuas es similar a la que se mostró en el ejemplo de la Figura 2.1. Esta forma de agrupamiento de los datos se conoce como

distribución normal. La mayoría de las herramientas estadísticas que se describen en este capítulo parten del supuesto de que las muestras estadísticas provienen de poblaciones que se distribuyen de manera normal.

La distribución normal presenta la forma de una campana gaussiana en donde la mayoría de los datos se apilan cerca de la media aritmética. Por el centro pasa un plano vertical de simetría que divide la campana en dos partes iguales. La forma de la curva de densidad de probabilidad de la Figura 2.1 muestra que la probabilidad de encontrar un valor en un intervalo determinado disminuye considerablemente a medida se aleja de la media aritmética en cualquier dirección. La Figura 2.2 resume algunas características de este tipo de distribuciones: la media aritmética y la mediana quedan en el centro de la campana, cerca del 68.27 % de los valores caen en el intervalo de la media aritmética más o menos una desviación estándar ($\bar{x} \pm s$), el 95.45 % de los valores en el intervalo $\bar{x} \pm 2s$, y el 99.7 % de los valores en el intervalo $\bar{x} \pm 3s$.

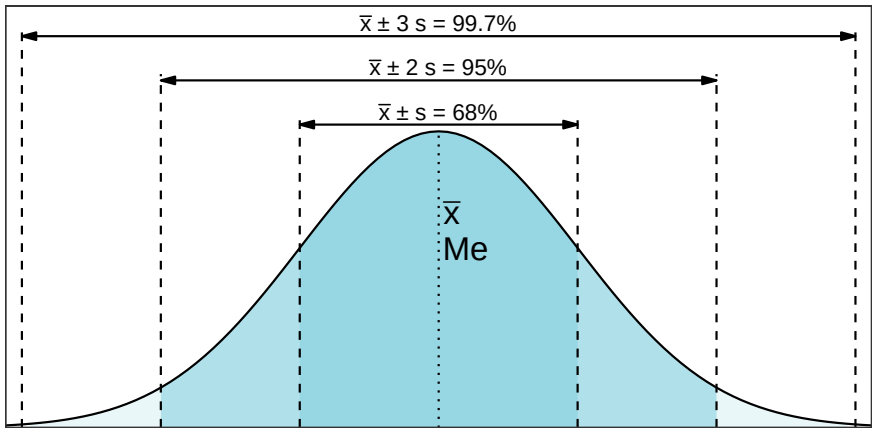


Figura 2.2: Agrupamiento de los datos en una distribución de probabilidad normal.

Las propiedades de la distribución normal sirven para ilustrar el concepto de **probabilidad estadística**: La probabilidad estadística es la frecuencia relativa con la que tiende a ocurrir un evento a largo plazo (Bulmer, 1979). El agrupamiento de los datos en la Figura 2.2 muestra que aproximadamente el 5 % de los datos de la población se encuentra por fuera del intervalo $\bar{x} \pm 2s$. En otras palabras, 5 de cada 100 datos que se toman aleatoriamente de una distribución normal caen por fuera del intervalo $\bar{x} \pm 2s$. La relación de 5/100 es la frecuencia relativa con la que ocurre el evento (obtener un dato en dicho intervalo), entonces la probabilidad de ese evento es de 0.05, o lo que es lo mismo, de 5 %.

2.2.2 Significancia estadística, nivel de confianza y valor P

Cuando la probabilidad de que ocurra un evento es demasiado pequeña, se hace razonable asumir que el evento no se presenta solo por azar. Esto quiere decir que los eventos con probabilidades muy pequeñas no se explican como consecuencia del error aleatorio. Esos casos se denominan estadísticamente significativos.

La **significancia estadística** (o nivel de significancia estadística) es un límite de probabilidad por debajo del cual la probabilidad de ocurrencia de un evento se considera demasiado pequeña. Este nivel se representa como α y siempre se debe definir al comienzo de un estudio, antes de iniciar a recopilar los datos. Los eventos estadísticamente significativos son los que tienen una probabilidad de ocurrir más pequeña que α. La elección de valor más común de α es 0.05 (5 %), pero en muchos casos puede resultar conveniente escoger niveles de significancia de 0.01 (1 %) o de 0.10 (10 %). La elección del nivel de significancia apropiado se describe en la Sección 2.2.4.1.

El término complementario a significancia estadística es el **nivel de confianza**. Si la región de lo estadísticamente significativo comprende los eventos que tienen una probabilidad muy pequeña de ocurrir, el resto del espacio muestral lo conforman los eventos que ocurren frecuentemente y que se consideran habituales dentro de lo que explica el error aleatorio. Esta región complementaria del espacio muestral tiene una probabilidad asociada de $1 - \alpha$ y se denomina nivel de confianza. El nivel de confianza más común es 95 %, pero como se describe en la Sección 2.2.4.1, en algunos casos puede ser mejor escoger valores de nivel de confianza de 90 % o 99 %.

El **valor P** es un indicador de la probabilidad de ocurrencia de un evento. El evento puede ser el obtener un conjunto de datos con las características que presenta la muestra estadística bajo estudio, considerando algunos supuestos. Los valores P son los resultados numéricos más importantes que se generan en las pruebas de estadística inferencial. Un valor P muy pequeño indica que la característica que se estudia de los datos tiene una probabilidad muy pequeña de ocurrir (menor al límite α) y por tanto no debería atribuirse al error aleatorio.

Suponga por ejemplo una hipótesis de que el valor 13.5 proviene de una población con distribución normal de media aritmética $\mu = 11.2$ y desviación estándar $\sigma = 0.75$. El valor 13.5 está alejado de la media aritmética por 2.3 unidades que es el equivalente a más de tres desviaciones estándar. La probabilidad de que un valor tan extremo haya surgido de una variable estadística con las características planteadas (Disttibución normal con $\mu = 11.2$ y $\sigma = 0.75$) es de 0.001 (0.1 %). Este es el valor P del evento estudiado. Si nuestro criterio de significancia estadística es de 5 %, podemos concluír a un nivel de confianza del 95 % que el valor 13.5 puede pertenecer a una población estadística diferente.

2.2.3 Pruebas de normalidad y pruebas de datos anómalos

En la Figura 2.2 se mostró el histograma y la curva de densidad de probabilidad de una población estadística de datos con distribución normal. De una población estadística se pueden tomar numerosas muestras estadísticas con características similares, pero que difieren entre sí y difieren a su vez de la población de la que provienen. La Figura 2.3 muestra tres histogramas de posibles muestras estadísticas tomadas de la población estadística representada en la Figura 2.2. Las curvas de densidad de probabilidad de las gráficas en la Figura 2.3 difieren un poco de la que presentan las distribuciones normales. La forma de campana gaussiana se hace más difícil de obtener a medida que el número de datos de la muestra estadística se hace más pequeña.

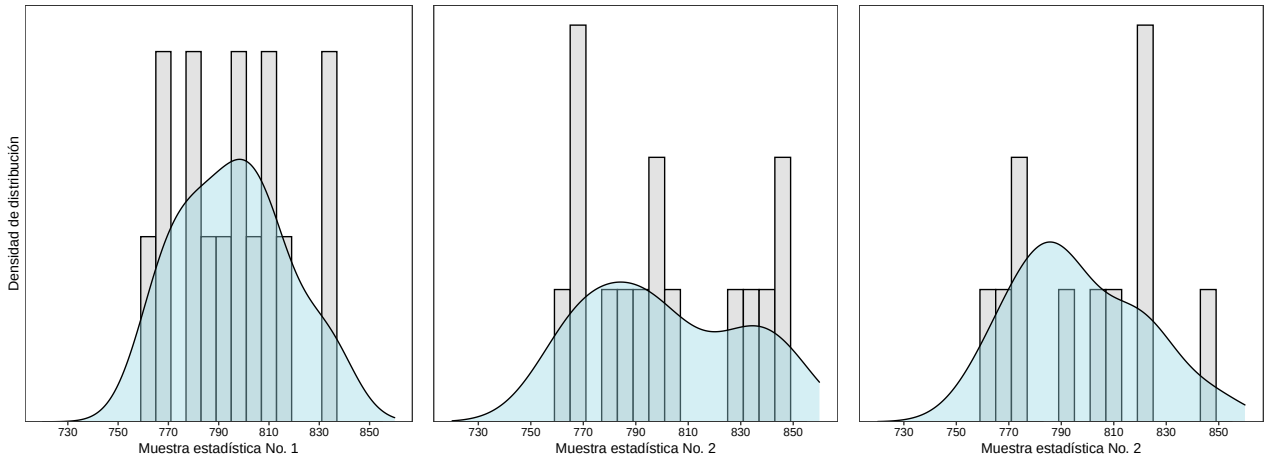


Figura 2.3: Histogramas y curvas de densidad de probabilidad para tres muestras estadísticas de 15 individuos tomados aleatoriamente de la población estadística representada en la Figura 2.2.

La inspección visual de un histograma no es una buena estrategia para evaluar la normalidad de la distribución de una muestra estadística. Los descriptores estadísticos de distribución que se incluyeron en la Tabla 2.1 (asimetría y curtosis) sirven en primera instancia para determinar si la forma de la distribución de los datos es simétrica y mesocúrtica, como cabría esperarse de una distribución normal, pero una alternativa más apropiada de evaluar la normalidad en muestras estadísticas de cualquier tamaño son las pruebas estadísticas de normalidad.

Las **pruebas de normalidad** permiten estimar la probabilidad de que una población estadística con distribución normal haya generado una muestra estadística con la distribución que tiene el conjunto de datos que se evalúa. Las pruebas de normalidad más comunes se listan a continuación (Lilliefors, 1967; Razali y Wah, 2011).

- **Prueba de Shapiro-Wilk**: Para muestras estadísticas de cualquier tamaño.
- **Prueba de Anderson-Darling**: Para muestras estadísticas de al menos ocho elementos.
- **Prueba de Kolmogorov-Smirnov**: Para muestras estadísticas de al menos 30 elementos.

En las pruebas de inferencia estadística se puede concluir comparando el valor P que se obtiene, contra el nivel de significancia estadística que se escoge para la prueba. Si el valor P es más pequeño que la significancia, la desviación del comportamiento normal se dice estadísticamente significativo y puede concluirse que la muestra estadística no tiene distribución normal. En algunos casos no se escoge directamente un nivel de significancia estadística sino que se define un nivel de confianza para la prueba. Ambas opciones son equivalentes si se considera que el nivel de confianza es igual a $1 - \alpha$.

Un enfoque similar puede usarse para evaluar la presencia de datos anómalos en una muestra estadística. Un **dato anómalo** es un valor extremo que se considera inusual en una muestra estadística, se alejan del resto de los valores en la muestra, y modifican su promedio muestral y su desviación estándar. Es importante identificar los datos anómalos porque muchas pruebas de comparación que se ven en las siguientes secciones requieren de que las series de datos bajo estudio no tengan datos anómalos significativos. Si se encuentra un dato anómalo se debería tratar de investigar la causa que lo pudo haber generado, en lugar de simplemente descartar el dato con base en únicamente la evidencia estadística.

Las pruebas de datos anómalos estiman la probabilidad del evento de que un dato tan extremo se haya originado por efecto del error aleatorio. Las pruebas más comunes de este tipo son la **prueba de Dixon** y las **pruebas de Grubbs**.

La prueba de Dixon evalúa si un dato en una muestra estadística (el más grande o el más pequeño) es anómalo. En esta prueba se usa la relación entre las distancias del dato más extremo a su valor más cercano, y su valor más lejano, para definir nuevos estadísticos (Q_{10} y Q_{01}) que tienen una función de densidad de probabilidad conocida (Q de Dixon). La función de densidad de probabilidad del estadístico permite obtener los valores P asociados a la prueba.

La prueba de Dixon es útil cuando la muestra presenta un único dato anómalo en solo uno de los dos extremos de la serie. Sin embargo, en muchos casos una muestra estadística puede presentar dos datos anómalos en un mismo extremo, o dos anómalos cada uno en cada extremo, entre otras posibilidades. Los casos posibles de hasta dos datos anómalos descritos se esquematizan en la Figura 2.4. Este tipo de casos de datos anómalos pueden evaluarse por medio de las pruebas de Grubbs. Estas pruebas se utilizan de manera similar a como se hace con la prueba de Dixon, en la que el valor P indicará la probabilidad de la una muestra estadística bajo estudio tenga valores tan anómalos como que se observan.

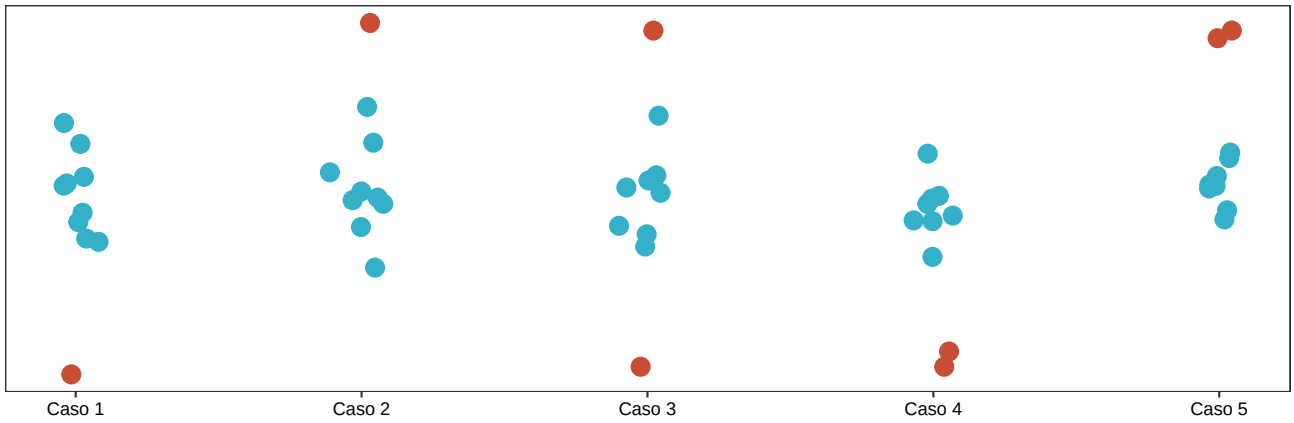
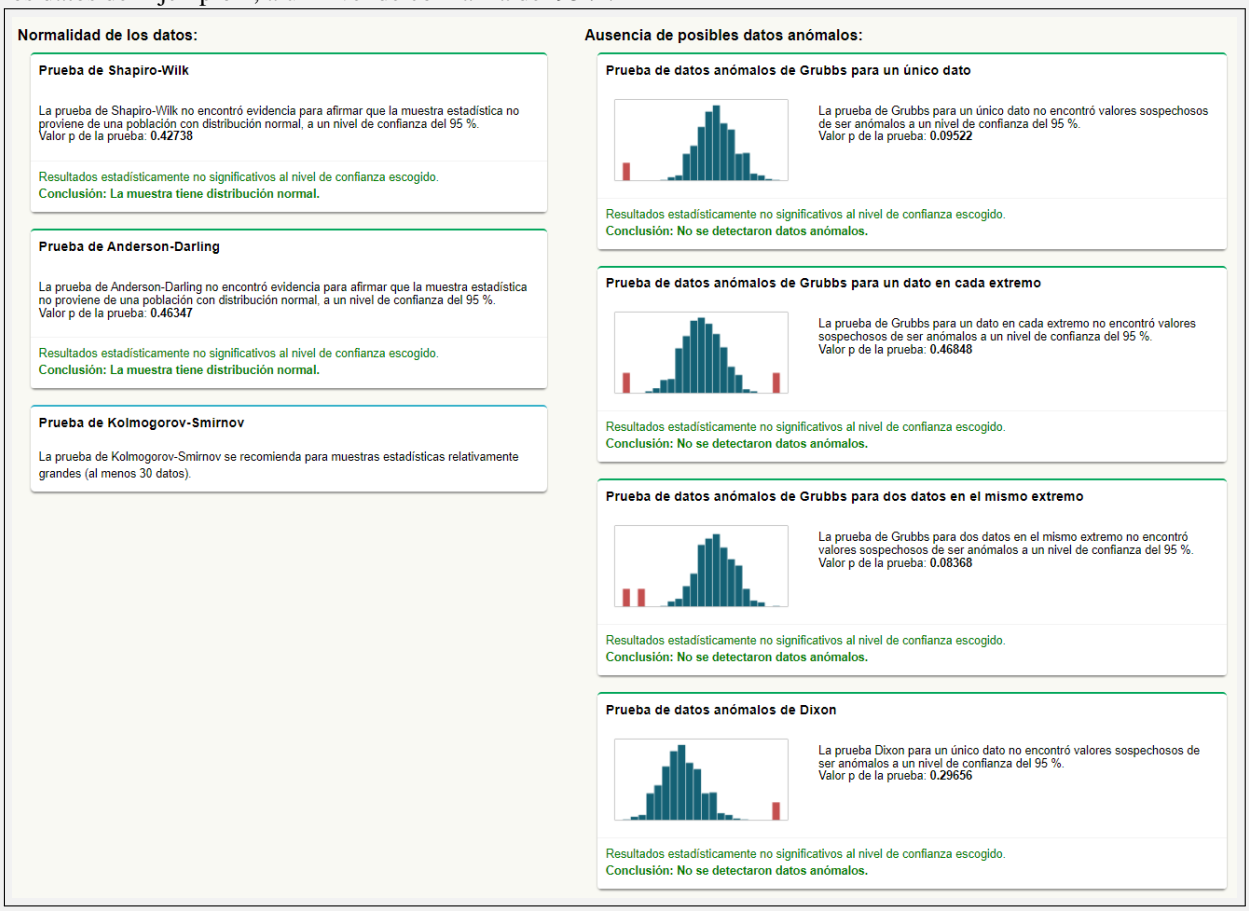


Figura 2.4: Casos típicos de datos anómalos (en rojo). Los casos 1 y 2 presentan un dato anómalo a cada extremo, el caso 3 tiene un dato anómalo en cada extremo, y los casos 3 y 4 muestran dos datos anómalos en el mismo extremo de la escala.

Ejemplo 2: Pruebas de normalidad y de datos anómalos.

El módulo **Normalidad y datos anómalos** de la sección **Herramientas estadísticas** del aplicativo validaR calcula las pruebas de normalidad y de datos anómalos descritas en esta sección.

El siguiente recuadro muestra los resultados de validaR al evaluar la normalidad y la ausencia de datos anómalos de los datos del Ejemplo 1, a un nivel de confianza del 95 %.



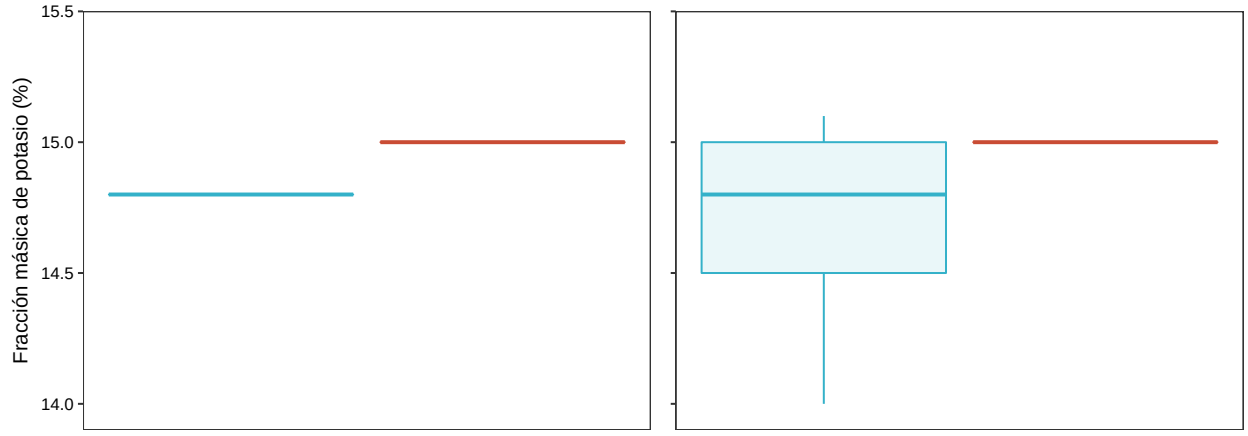
2.2.4 Generalidades de las pruebas de comparación: Inferencias sobre una media muestral

En esta sección se describen más a fondo las pruebas de hipótesis usando como ejemplo la prueba de comparación de una media muestral contra un valor de referencia: la prueba t de Student de una única muestra.

Suponga el escenario en el que se quiere verificar por medio de un control de calidad que la fracción másica de potasio en un lote de producción de un fertilizante corresponde con el indicado (15 %). El procedimiento habitual implicaría seleccionar aleatoriamente (muestrear) algún número de unidades (n) del lote y cuantificar potasio en ellas. Numéricamente se podría comparar de manera directa si el promedio de los n resultados ($\bar{x}$) es, o no es, igual al valor esperado para la propiedad. Sin embargo, $\bar{x}$ es un descriptor muestral afectado por error aleatorio, y su comparación puntual contra un valor de referencia no es adecuada. Para concluir sobre el control de calidad del fertilizante considerando la presencia del error aleatorio conviene hacer una prueba de comparación.

Una prueba de comparación estimaría la probabilidad de que una población estadística con media aritmética $\mu = 15\,\%$ haya generado una muestra estadística como la que se obtuvo durante el muestreo aleatorio. Si esta probabilidad es demasiado pequeña, razonablemente puede concluirse que las unidades de la muestra estadística provienen de una población estadística donde $\mu \neq 15\,\%$, y el lote no pasaría el control de calidad. Este tipo de comparación considera más información de la muestra estadística que la que consideraría una comparación puntual. Ambos esquemas de comparación se esbozan en la Figura 2.5.

Las pruebas de comparación contrastan dos afirmaciones contradictorias que buscan anular o reforzar una pregunta dada. Estas afirmaciones se denominan **hipótesis nula** e **hipótesis alternativa**, respectivamente. En el ejemplo que se estudia, la hipótesis nula ($\mathbf{H_0}$) parte del supuesto que la media aritmética de la población de la que proviene



(a) Comparación puntual (b) Comparación considerando el muestreo aleatorio

Figura 2.5: Esquemas de comparación de un promedio muestral (en azul) contra un valor de referencia (en rojo).

muestra de los datos (μ) es igual al valor de referencia (μ_0, en este caso de 15 %), y que la diferencia observada se debe únicamente al error aleatorio.

La hipótesis alternativa ($\mathbf{H_1}$) contradice a la hipótesis nula y suelen plantear en forma de desigualdad bidireccional o unidireccional (el tipo de desigualdad que conviene escoger se describe en la Sección 2.2.5.1):

$$\mathbf{H_0}: \quad \mu = \mu_0 \qquad\qquad \mathbf{H_1}: \quad \mu \neq \mu_0 \tag{2.1}$$

Al comienzo de una prueba de comparación se parte asumiendo la hipótesis nula y el proceso de inferencia estadística busca determinar si la evidencia experimental es suficiente para rechazarla, y aceptar en su lugar a la hipótesis alternativa. Esta evidencia para rechazar la hipótesis nula se considera suficiente cuando la probabilidad de ocurrencia de la muestra estadística bajo estudio es demasiado pequeña bajo el supuesto de que la hipótesis nula fuera cierta. En los casos en los que la evidencia no es suficiente para rechazarla, la hipótesis nula no se acepta sino que simplemente no se puede rechazar. En este caso se dice que los resultados no son estadísticamente significativos.

Para calcular la probabilidad de ocurrencia de una muestra estadística bajo el supuesto que propone una hipótesis nula, la información del conjunto de datos se suele transformar para dar lugar a nuevos valores denominados estadísticos. Los estadísticos de las pruebas de comparación están caracterizados y se les conoce su función de densidad de probabilidad, con lo que se pueden obtener los valores P asociados a cada caso.

Si una muestra estadística proviene de una población con distribución normal y no presenta datos anómalos significativos, la media muestral $\bar{x}$ puede compararse con el valor de referencia μ_0 por medio del estadístico t de Student, que se calcula como se muestra en la Ecuación 2.2. La función de densidad de probabilidad del estadístico t de Student se encuentra disponible en tablas y en programas estadísticos, de manera que se conoce el valor de probabilidad (P) asociado a cualquier valor de t, para un determinado número de grados de libertad.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{2.2}$$

El valor p asociado a un valor en la distribución t de Student depende del número de datos n que se utilizaron para calcular $\bar{x}$ y s. Esto se describe como que la distribución es sensible al número de **grados de libertad** (ν), que en este caso corresponde a $n-1$, donde n es el número de datos.

Si el valor P de la prueba t de Student es menor al valor de α, los resultados son estadísticamente significativos y se cuenta con evidencia de que la media aritmética de la población de la que proviene la muestra es diferente de μ_0. Se rechazaría la hipótesis nula. Por otro lado, el caso cuando el valor P es mayor a α no se debe malinterpretar.

No rechazar la hipótesis nula no implica que esta sea verdadera. La inferencia estadística busca evidencia para rechazar la hipótesis nula y el que los resultados no sean estadísticamente significativos solo implica que no hay evidencia que soporte el rechazo de la hipótesis nula. La hipótesis nula no se aceptaría sino simplemente no se podría rechazar.

En el ejemplo que se describió en esta sección, la prueba de inferencia estadística no permitiría establecer si la muestra estadística proviene de una población con media $\mu = 15\,\%$, sino que solo serviría para establecer si la muestra estadística proviene de una población estadística con media $\mu \neq 15\,\%$.

2.2.4.1 Selección del nivel de confianza: Error tipo I y error tipo II

En las pruebas de comparación se espera que la conclusión que se obtiene coincida con la realidad del fenómeno bajo estudio. Sin embargo, es probable que se rechace la hipótesis nula cuando esta es cierta (falso positivo), o que no se rechace la hipótesis nula cuando esta es falsa (falso negativo). Estos escenarios se denominan error tipo I y error tipo II, respectivamente.

El error tipo I suele coincidir con el nivel de significancia que se escoge para la prueba. Recuerde que el nivel de significancia comprende los escenarios que se consideran poco probables bajo el supuesto que plantea la hipótesis nula. Sin embargo, poco probable es diferente de imposible, por lo que todo rechazo conlleva un riesgo de que se asuma como falso un evento que es verdadero, solo porque su probabilidad de ocurrencia es baja. El error tipo I puede disminuirse escogiendo un nivel de confianza mayor. Pero aumentar el nivel de confianza implica considerar un rango más grande de no rechazo para la hipótesis nula, y esto aumenta linealmente la probabilidad de cometer un error tipo II. Es imposible disminuir simultáneamente la probabilidad de cometer ambos tipos de error. Al disminuir la probabilidad de cometer uno de los tipos de error, la probabilidad el otro tipo de error aumenta.

La probabilidad de cometer un error tipo II suele ser desconocida. Determinar el tipo de error que resulta menos perjudicial para el caso que se investiga es una decisión importante que debe tomarse en las primeras etapas del estudio.

Considere por ejemplo un análisis biológico en el que se pretende establecer si una persona sujeto de ensayo presenta una enfermedad infecto-contagiosa. Los resultados del análisis se pueden ver como una variable estadística aleatoria que se analizan como una inferencia estadística que puede presentar errores tipo I y tipo II. Supongamos que la hipótesis nula implica que el individuo está infectado, y el análisis busca aportar evidencia de que no lo está (hipótesis alternativa). En ese orden de ideas:

- El error tipo I sería rechazar erróneamente la hipótesis nula que dice que asume que el paciente está enfermo. Esto implica que se concluye que el individuo está sano cuando en realidad está infectado, y puede ocasionar que el paciente infecte a muchas personas en medio de la desinformación.
- El error tipo II sería no rechazar la hipótesis nula cuando esta es falsa, y que implicaría que el paciente entre a una cuarentena que realmente no necesita.

En la mayoría de los casos es más dañino causar la propagación de una enfermedad contagiosa (error tipo I) que someter a un paciente sano a una cuarentena (error tipo II). En este caso se debería minimizar la probabilidad de ocurrencia de un error tipo I por medio de la aplicación de una prueba de hipótesis *conservadora*, que se caracteriza por un valor de significancia pequeña (o lo que es equivalente, un nivel de confianza grande, como 99 %).

Las pruebas en las que el error tipo I es menos relevante que el error tipo II usan niveles de confianza más pequeños (como 90 %), y se les conoce como *liberales*. Una prueba liberal es conveniente, por ejemplo, cuando en la mejora de un proceso industrial se debe establecer si una variable afecta su rendimiento. En el cribado de variables relevantes, las hipótesis nulas suelen formular que la variable no influye y la hipótesis alternativa dice que si influye. En este caso es menos grave estudiar el efecto de variables que no son tan relevantes (rechazar erróneamente la hipótesis nula) en lugar de ignorar variables que potencialmente si lo son.

2.2.5 Pruebas de comparación de medias: prueba t de Student

La distribución t de Student se usa para modelar las medias aritméticas de muestras estadísticas pequeñas que provienen de poblaciones con distribución normal. Con esta prueba se pueden hacer comparaciones de la media de una muestra estadística contra un valor de referencia o contra la media aritmética de otra muestra estadística. Para comparar las medias aritméticas entre más de dos muestras estadísticas al tiempo se debe hacer un **Análisis de Varianza (ANOVA)**, que se trata en la Sección 2.2.7.

Las siguientes preguntas son ejemplos de problemas que se pueden abordar por medio de pruebas t de Student:

- ¿El nivel de concentración de un contaminante Se encuentra por debajo de un límite máximo permitido?
- ¿Los valores de concentración de un analito que se cuantifica en dos muestras analíticas independientes son diferentes entre sí?
- ¿Los resultados de medición de un analito cambian si se usan dos protocolos distintos de digestión de la muestra?

Cuando se comparan las medias de dos conjuntos de datos se puede tener el caso en el que los elementos del primer conjunto son independientes de los elementos del otro conjunto, o pueden guardar una estrecha relación con los elementos del oro conjunto. Las muestras estadísticas se denominan **emparejadas** cuando hay una relación directa entre los individuos que las conforman (por ejemplo cuando se trata de los mismos sujetos de prueba evaluados bajo dos condiciones diferentes). En el caso contrario se habla de **muestras estadísticas independientes**. El tipo de prueba de comparación de medias que se debe aplicar es diferente para cada caso.

2.2.5.1 Media muestral contra valor de referencia

Ese es el caso que se estudió en la Sección 2.2.4. La prueba t para comparar una media muestral ($\overline{x}$) con un valor de referencia (μ_0) asume inicialmente que μ y μ_0 son iguales (hipótesis nula). La hipótesis alternativa establece que los valores que se comparan son diferentes, pero las diferencias pueden ser de distintos tipos. El tipo de diferencia entre los valores puede ser de interés dependiendo de la pregunta que se quiere resolver, y da lugar a pruebas de comparación **bilaterales** y **unilaterales**.

En el ejemplo de la Sección 2.2.4 el valor de referencia era una especificación puntual que debía cumplir un producto. En este caso las diferencias por tanto por encima como por debajo representan un resultado desfavorable del control de calidad. En este caso se usa una comparación bilateral, en la que es de interés conocer si hay cualquier diferencia estadísticamente significativa respecto al valor de referencia. En este caso la hipótesis alternativa es una desigualdad en ambos sentidos ($H_1: \mu \neq \mu_0$).

En otros casos, el valor de referencia contra contra el que se hace la comparación puede ser un límite máximo permitido de un contaminante en una matriz, o puede ser una especificación de la concentración mínima garantizada de un nutriente en un suplemento alimenticio. Ambos escenarios requieren la aplicación de una prueba de comparación unilateral en la que la hipótesis de investigación se interesa en un tipo de diferencia particular (de tipo mayor que o menor que), entre la media muestral y el valor de referencia. Hay dos posibles hipótesis alternativas que se pueden plantear: (i) $H_1: \mu < \mu_0$ en el ejemplo en que se quiere comprobar que la matriz tiene una concentración del contaminante por debajo del límite máximo, y (ii) $H_1: \mu > \mu_0$ para el ejemplo en el que se quiere comprobar que el alimento tiene una concentración de nutriente por encima de la mínima garantizada en el producto.

La distribución t de Student define el valor P para los valores del estadístico t. Con el valor P se puede concluir si la evidencia es suficiente para afirmar que la media muestral difiere significativamente del valor de referencia, a un determinado nivel de confianza.

Para muestras estadísticas que no se distribuyen de manera normal se puede usar alternativamente la prueba de rangos de Wilcoxon que es robusta a la no normalidad de los datos (King y Eckersley, 2019). Los resultados de esta prueba se interpretan de la misma forma de la que se haría al ejecutar una prueba t de Student. Las pruebas de comparación que no asumen una distribución de probabilidad para el conjunto de datos se denominan **pruebas no paramétricas** (King y Eckersley, 2019).

Ejemplo 3. Comparación de una media muestral contra un valor de referencia.

El submódulo **Comparación de medias** del módulo **≥ Pruebas de comparación** de la sección **Herramientas estadísticas** del aplicativo validaR realiza la prueba de comparación de la media muestral de una muestra estadística contra un valor de referencia.

Suponga que los datos de la tabla del Ejemplo 1 corresponden a la medición de una muestra que debe contener una fracción másica de ion plomo menor a 12.2 mg L^{-1}.

Debe escogerse una hipótesis alternativa de manera tal que su posible aceptación otorgue evidencia de lo que se quiere comprobar: la media de los resultados debería ser significativamente menor al límite máximo permitido.

$$\mathbf{H_0}: \mu = 12.2\ mg\ L^{-1}, \qquad \mathbf{H_1}: \mu < 12.2\ mg\ L^{-1}$$

El recuadro a la derecha contiene los resultados de validaR al evaluar la prueba de comparación contra el valor de referencia con los datos de la tabla del Ejemplo 1, a un nivel de confianza del 95 %.

Resultados

Prueba t de Student para una media muestral

Resultados estadísticamente significativos:

La media muestral del conjunto es menor al valor de referencia.

Datos.prueba	Valor
Valor estadístico t	-7.423
Grados de libertad	7
Valor p	1e-04
Media muestral	12.05
Valor de referencia	12.2
Error estándar de la media	0.02004
Intervalo de confianza:	
~, límite inferior	-Inf
~, límite superior	12.09
Nivel de confianza (%)	95

La diferencia entre la media de la muestra estadística y el valor de referencia es estadísticamente significativa a un nivel de confianza del 95 %.

2.2.5.2 Dos medias muestrales independientes

La comparación de las medias de dos muestras estadísticas usando la distribución t de Student parte la hipótesis nula de que las muestras provienen de poblaciones con la misma media aritmética. Si dos grupos x y y vienen de poblaciones con medias aritméticas $\mu_x = \mu_0$ y $\mu_y = \mu_0$, la hipótesis nula puede escribirse de la siguiente forma:

$$\mathbf{H_0}: \mu_x = \mu_y \tag{2.3}$$

Si las muestras estadísticas provienen de la misma población estadística puede esperarse que su promedio sea similar pero también que su dispersión sea similar. En este caso el cálculo del estadístico t que se usa para la comparación es similar al que se mostró en la Sección 2.2.5.1, solo que en esta ocasión la diferencia entre las medias se normaliza considerando la desviación estándar combinada de ambos conjuntos de datos ($s_{x,y}$) y los tamaños de muestra de cada conjunto (n_x y n_y, respectivamente):

$$t = \frac{\overline{x}_x - \overline{x}_y}{s_{x,y}\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}}, \qquad \text{con} \quad s_{x,y} = \sqrt{\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}} \tag{2.4}$$

donde s_x y s_y son las desviaciones estándar de las muestras estadísticas x y y, respectivamente.

Una *dispersión similar* de ambos conjuntos de datos implica que la diferencia entre sus varianzas no es estadísticamente significativa (Johnson y Bhattacharyya, 2009). Esto se conoce como homogeneidad de varianzas (u **homocedasticidad**) y se trata en la Sección 2.2.6.2. En este caso la prueba t utiliza $(n_x + n_y - 2)$ grados de libertad.

Muestras estadísticas que no tienen homogeneidad de varianzas se denominan **heterocedásticas**. En este caso el estadístico t de Student y los respectivos grados de libertad se calculan con las siguientes ecuaciones:

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}}}, \qquad \text{con los grados de libertad} \quad \nu = \frac{\left(\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right)^2}{\frac{1}{n_x-1}\left(\frac{S_x^2}{n_x}\right)^2 + \frac{1}{n_y-1}\left(\frac{S_y^2}{n_y}\right)^2} \tag{2.5}$$

Las comparaciones entre dos medias muestrales también pueden ser unilaterales o bilaterales según sea de interés para la pregunta de investigación. Es importante considerar que casi nunca es necesario realizar los cálculos manualmente porque estos ya están implementados en diversos programas estadísticos. La parte importante del análisis consiste en plantear adecuadamente la hipótesis alternativa seleccionar y el nivel de confianza que más convenga. El aplicativo validaR establece automáticamente si se trata de muestras estadísticas homocedásticas o heterocedásticas para hacer los cálculos.

La versión no paramétrica de la prueba t de Student para dos medias muestrales se conoce como la prueba de Wilcoxon-Mann-Whitney (King y Eckersley, 2019).

Ejemplo 4. Comparación de dos medias muestrales independientes.

La misma sección del aplicativo validaR que se utilizó para el Ejemplo 3 sirve para realizar la comparación de medias muestrales de dos series de datos que se hayan ingresado en columnas independientes del aplicativo.

En este ejemplo se comparan los resultados que obtienen dos analistas para la fracción másica de nitrógeno amoniacal en la misma muestra analítica. Se quiere establecer si hay diferencias entre los resultados por cada analista. Los datos que obtuvieron los analistas (en porcentaje) se muestran en la tabla de abajo a la izquierda:

Analista 1	Analista 2
1.43	1.31
1.27	1.44
1.45	1.51
1.22	1.88
1.43	1.65
1.34	1.54
1.15	1.67

En este caso conviene escoger una hipótesis alternativa bilateral porque diferencias en cualquier dirección pueden implicar un problema: que los resultados de los analistas no coinciden.

$$\mathbf{H_0}: \mu_{Analista.1} = \mu_{Analista.2} \qquad \mathbf{H_1}: \mu_{Analista.1} \neq \mu_{Analista.2}$$

El recuadro a la derecha contiene los resultados de validaR al evaluar la prueba de comparación entre las dos medias muestrales.

2.2.5.3 Dos medias muestrales de conjuntos emparejados

Dos conjuntos se denominan emparejados si los individuos de una muestra estadística están relacionados con los individuos de otra muestra estadística. Este es el caso, por ejemplo, cuando se determina un mensurando en un conjunto de muestras analíticas utilizando dos procesos de medición diferentes. El conjunto de resultados de cada método de medición conforma una muestra estadística. La comparación de sus medias aritméticas directamente podría conducir a conclusiones erróneas, pues la variación de los valores de la propiedad dentro de los individuos de una misma muestra estadística puede opacar la variación entre las dos muestras estadísticas.

Cuando las muestras estadísticas se encuentran emparejadas toca analizar sus medias muestrales en términos de las diferencias que presenta cada individuo en ambos conjuntos de datos. Si no hay diferencias entre las muestras estadísticas, el promedio de las diferencias debería ser de cero. La distribución t puede emplearse para estudiar si la diferencia promedio entre los valores es significativamente diferente de cero. Para calcular el estadístico se utiliza la siguiente expresión:

$$t = \frac{\bar{x}_{dif}}{s_{dif}/\sqrt{n}} \tag{2.6}$$

donde $\bar{x}_{dif}$ y s_{dif} son la media aritmética y la desviación estándar de las diferencias en cada pareja ($dif_i = x_i - y_i$). La distribución del estadístico se evalúa considerando $n-1$ grados de libertad, donde n es el número de parejas.

La prueba t de Student es apropiada para los casos en los que las diferencias entre los valores de las parejas se distribuyen normalmente. Para casos en los que esto no se cumple se puede usar la prueba no paramétrica de Wilcoxon-Mann-Whitney para muestras emparejadas (King y Eckersley, 2019).

Ejemplo 5. Comparación de dos medias muestrales emparejadas.

La misma sección del aplicativo validaR que se utilizó para los Ejemplos 3 y 4 sirve para realizar la comparación de dos medias muestrales de conjuntos de datos emparejados que se hayan ingresado en columnas independientes del aplicativo. Para muestras estadísticas emparejadas el número de datos debe ser igual y deben estar en el mismo orden.

En este ejemplo se comparan los niveles de contaminación atmosférica entre la mañana y la tarde en distintos puntos de una ciudad. Se quiere comprobar si la contaminación aumenta durante el día. La concentración de partículas finas suspendidas medidas en ambos momentos del día se muestran en la tabla de abajo a la izquierda:

Punto	Mañana / $\mu g\ m^{-3}$	Tarde / $\mu g\ m^{-3}$
No. 1	12	12
No. 2	23	24
No. 3	8	15
No. 4	19	26
No. 5	33	26
No. 6	50	52
No. 7	26	28
No. 8	21	28

Como se busca evidencia que soporte que el nivel de contaminación en la tarde es mayor que en la mañana, las hipótesis de la prueba deben plantearse de la siguiente manera:

$$\mathbf{H_0}: \mu_{manana} = \mu_{tarde} \quad \mathbf{H_1}: \mu_{manana} < \mu_{tarde}$$

El recuadro a la derecha muestra los resultados de validaR al evaluar las hipótesis para las muestras emparejadas.

Resultados

Prueba t de Student para dos medias muestrales

Resultados estadísticamente no significativos.

La media muestral del primer conjunto no es menor de la del segundo grupo.

Datos.prueba	Valor
Valor estadístico t	-1.406
Grados de libertad	7
Valor p	0.1013
Media grupo 1	-2.375
Media grupo 2	NA
Diferencia de la medias	NA
Error estándar de la diferencia	1.69
Intervalo de confianza de la diferencia:	
~, límite inferior	-Inf
~, límite superior	0.8262
Nivel de confianza (%)	95

La diferencia entre las medias muestrales no es estadísticamente significativa a un nivel de confianza del 95 %

2.2.6 Pruebas de comparación de varianzas

El error aleatorio que afecta el promedio de una muestra estadística también afecta su varianza. Esta sección muestra las pruebas de hipótesis que sirven para comparar la dispersión de una muestra estadística contra un valor de referencia, contra la dispersión de otra muestra estadística, o en un grupo de varias muestras estadísticas.

2.2.6.1 Varianza muestral contra un valor de referencia: Prueba χ^2

En la Sección 2.2.5 se habla del uso del estadístico t de Student para modelar medias aritméticas de muestras estadísticas que provienen de poblaciones con distribución normal. Para modelar la varianza de muestras estadísticas que provienen de poblaciones con distribución normal se utiliza el estadístico χ^2 (*/ji cuadrado/*). El valor de referencia en este caso es una medida de dispersión de la población estadística de la que se presume que puede venir el conjunto de datos que se estudia. La prueba χ^2 sirve para estimar la probabilidad de que la muestra estadística que se obtuvo provenga de esa población. Esta prueba utiliza una relación entre la varianza muestral y la varianza de referencia para evaluar si el conjunto de datos viene de una población estadística con una varianza poblacional (σ^2) igual a una varianza de referencia (σ_0^2):

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{2.7}$$

A cualquier valor del estadístico χ^2 se le puede conocer el valor P asociado para un determinado número de grados de libertad ($\nu = n-1$). Nuevamente, este valor P se compara con el nivel de significancia que se escogió al comienzo del estudio para concluir sobre la hipótesis nula ($H_0: \sigma^2 = \sigma_0^2$).

La hipótesis alternativa depende del tipo de comparación que se requiera:

- $H_1: \sigma^2 \neq \sigma_0^2$, comparación bilateral para evaluar si la varianza muestral es diferente (en cualquier dirección) a la varianza de referencia.
- $H_1: \sigma^2 < \sigma_0^2$, comparación unilateral para evaluar si la dispersión de una muestra estadística es significativamente menor a un valor de referencia. Por ejemplo, cuando se quiere comprobar que un conjunto de réplicas instrumentales de medición tienen una desviación estándar relativa menor al 5 %.
- $H_1: \sigma^2 > \sigma_0^2$, comparación unilateral para evaluar si la dispersión de una muestra estadística es significativamente mayor a un valor de referencia. Por ejemplo, cuando se quiere evaluar si la repetibilidad de un proceso de medición empeora cuando se utilizan reactivos químicos de menor pureza.

Al comparar una varianza muestral contra una medida de dispersión de referencia es importante considerar que en algunos casos la referencia puede ser una desviación estándar o una desviación estándar relativa (coeficiente de variación). En estos casos puede ser necesario convertir el valor de referencia a unidades de varianza para calcular el estadístico χ^2.

Ejemplo 6. Comparación de varianza muestral contra valor de referencia.

En el Ejemplo 3 se comparó un conjunto de resultados de medición de ion plomo respecto a un valor máximo permitido. En este ejemplo se usan los mismos datos para probar si los resultados del método de medición tienen una desviación estándar relativa (RSD) menor a 5 %.

Se necesita una prueba unilateral en la que la hipótesis alternativa ayude a comprobar que la RSD de los resultados es menor a 5 %.

$$\mathbf{H_0}: RSD\ (\%) = 5\ \%, \qquad \mathbf{H_1}: RSD\ (\%) < 5\ \%$$

Si el cálculo del estadístico χ^2 se fuera a realizar manualmente sería necesario convertir el valor de 5 % de RSD a unidades de varianza para probar las siguientes hipótesis:

$$\mathbf{H_0}: \sigma^2 = \sigma_0^2, \qquad \mathbf{H_1}: \sigma^2 < \sigma_0^2$$

Si los cálculos se hacen en validaR, (submódulo **Comparación de varianzas** del módulo **Pruebas de comparación** de la sección **Herramientas estadísticas**) las conversión a unidades de varianza se hace automáticamente. El recuadro a la derecha muestra los resultados del aplicativo para la prueba de comparación.

Resultados

Prueba Chi cuadrado para una varianza muestral

Resultados estadísticamente significativos:

La varianza muestral del conjunto es menor al valor de referencia.

Datos.prueba	Valor
Estadístico \(\chi^2\)	0.06194
Grados de libertad	7
Valor p	< 1 e-04
Varianza muestral	0.003213
Intervalo de confianza de la varianza:	
~, límite inferior	0
~, límite superior	0.01038
Nivel de confianza (%)	95
Desviación estándar	0.05668
Desviación estándar relativa (%)	0.5

La diferencia entre la varianza de la muestra estadística y el valor de referencia es estadísticamente significativa a un nivel de confianza del 95 %.

2.2.6.2 Homogeneidad entre 2 varianzas muestrales: Prueba F de Fisher

Cuando se compara la dispersión de dos o más muestras estadísticas se habla de pruebas de homogeneidad de sus varianzas. Muestras estadísticas que presentan homogeneidad de varianzas se denominan **homocedásticas**, mientras las muestras estadísticas que no presentan homogeneidad de varianzas se denominan **heterocedásticas**.

La homogeneidad de varianzas entre dos muestras estadísticas se evalúa por medio de la prueba F de Fisher que usa la relación entre las varianzas de las muestras. Si las varianzas son similares, la relación entre estas debe aproximarse a uno. Este tipo de prueba se puede utilizar, por ejemplo, para probar si hay diferencias estadísticamente significativas en la dispersión de los resultados que producen dos analistas en condiciones de

trabajo normales. El estadístico F de Fisher se calcula con la siguiente ecuación:

$$F = \frac{s_1^2}{s_2^2} \tag{2.8}$$

donde s_1^2 y s_2^2 son las varianzas de las muestras estadísticas 1 y 2, respectivamente.

En la distribución F de Fisher los conjuntos de datos pueden tener diferente número de datos, lo que da lugar a dos valores de grados de libertad independientes: para el denominador y para el numerador. El valor P asociado a cada estadístico F de Fisher considera los dos valores de grados de libertad. Las pruebas F de Fisher también pueden ser bilaterales o unilaterales.

La prueba F de Fisher asume que las poblaciones de las que provienen las muestras tienen distribución normal. Para muestras que no cumplen este supuesto puede usarse la prueba no paramétrica de rangos al cuadrado (Conover, 1999).

Ejemplo 7. Homocedasticidad en resultados de medición obtenidos por dos analistas diferentes.

En el Ejemplo 4 se compararon las medias de los resultados obtenidos por dos analistas diferentes para un mismo mensurando en la misma muestra analítica y no se encontraron diferencias estadísticamente significativas entre los promedios de los resultados de cada grupo.

Ahora suponga que el Analista 2 está en entrenamiento y se quiere evaluar si sus resultados de medición presentan una dispersión mayor que los resultados que obtuvo el Analista 1 (un analista experimentado).

La pregunta planteada se resuelve por medio de una prueba de comparación de varianzas de tipo unilateral, que permita establecer si $\sigma^2_{analista.2} > \sigma^2_{analista.1}$:

$\mathbf{H_0}: \sigma^2_{analista.2} = \sigma^2_{analista.1}$, $\quad \mathbf{H_1}: \sigma^2_{analista.2} > \sigma^2_{analista.1}$

El recuadro a la derecha muestra los resultados del submódulo **Comparación de varianzas** del módulo **≥ Pruebas de comparación** de la sección **Herramientas estadísticas** del aplicativo validaR, para la prueba de comparación que se plantea.

Observe que en el aplicativo las hipótesis alternativas pueden estar en orden diferente al que se plantean. En este caso, recuerde que $\sigma^2_{analista.2} > \sigma^2_{analista.1}$ es lo mismo que decir $\sigma^2_{analista.1} < \sigma^2_{analista.2}$.

Resultados

Prueba F de Fisher para comparar dos varianzas muestrales

Resultados estadísticamente no significativos.

La varianza muestral del primer conjunto no es menor de la del segundo grupo.

Datos.prueba	Valor
Estadístico F	0.4097
Grados de libertad	
~, numerador	6
~, denominador	6
Valor p	0.1509
Relación de varianzas	0.4097
Intervalo de confianza:	
~, límite inferior	0
~, límite superior	1.755
Nivel de confianza (%)	95
Varianza grupo 1	0.01376
Varianza grupo 2	0.03358

La diferencia entre las varianzas muestrales no es estadísticamente significativa a un nivel de confianza del 95 %

2.2.6.3 Homocedasticidad entre más de dos muestras estadísticas y varianzas anómalas

Algunas pruebas estadísticas permiten evaluar si las varianzas de más de dos muestras estadísticas son iguales entre sí. En este tipo de pruebas estadísticas la hipótesis alternativa establece que al menos dos conjuntos de datos presentan una diferencia estadísticamente significativa en sus varianzas:

$$\mathbf{H_0}: \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \qquad \mathbf{H_1}: \sigma_i^2 \neq \sigma_j^2, \text{ para al menos una de las parejas } (i,\ j) \tag{2.9}$$

La Tabla 2.2 contiene las pruebas más comunes para determinar la homogeneidad de varianzas cuando se tienen más de dos muestras estadísticas. Es importante notar que en este tipo de conjuntos no conviene hacer comparaciones entre las parejas de muestras estadísticas porque esta práctica hace que aumente el error tipo I (ver Sección 2.2.4.1). El tipo de homocedasticidad que se trata en esta sección se denomina homocedasticidad entre grupos, y se aplica a grupos que están diferenciados por alguna variable categórica.

Calculo del estadístico	Características	Observaciones
Prueba de Barlett $T = \frac{(N-k)\ln S_p^2 - \sum_{i=1}^{k}((n_i-1)\ln S_i^2)}{1+\frac{1}{3k-3}((\sum_{i=1}^{k}\frac{1}{n_i-1})-\frac{1}{N-k})}$	El estadístico tiene una distribución χ^2 con $k-1$ grados de libertad.	Se ve muy afectado si las muestras estadísticas no tienen distribución normal.
Prueba de Hartley $F_{max} = \frac{S_{max}^2}{S_{min}^2}$	Es la prueba más sencilla de evaluar manualmente. El estadístico F_{max} tiene su propia distribución estadística.	Asume que las muestras estadísticas son del mismo tamaño.
Prueba de Cochran $C = \frac{S_{max}^2}{\sum_{i=1}^{k} S_i^2}$	Esta prueba evalúa si la varianza más grande es anómala.	Asume que las muestras estadísticas son del mismo tamaño.
Prueba de Levene $W = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k} n_i(\overline{Z}_i-\overline{\overline{Z}})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\overline{Z}_i-Z_{ij})^2}$	Es equivalente a realizar un ANOVA de las distancias entre los individuos de cada grupo respecto a una medida de tendencia central de su grupo (ver Sección 2.2.7).	Se puede hacer respecto a la media o a la mediana de los grupos. Es una prueba robusta frente desviaciones del comportamiento normal de las muestras cuando se utiliza la mediana.

En las ecuaciones, N es el número total de datos, k es el número de grupos, n_i y S_i^2 son el número de datos y la varianza del i-ésimo grupo, respectivamente. S_p^2 es la varianza combinada que se calcula como $(\sum_{i=1}^{k}(n_i-1)S_i^2)/(N-k)$.
S_{max}^2 y S_{min}^2 representan las varianzas más grande y más pequeña del conjunto de varianzas, respectivamente. Z_{ij} es la distancia (en valor absoluto) del j-ésimo dato en el i-ésimo grupo respecto a una medida de tendencia central (media, mediana o media recortada), $\overline{Z}_i$ es el promedio de las distancias en la i-ésima serie, y $\overline{\overline{Z}}$ es el promedio de los promedios de la media.

Tabla 2.2: Pruebas de homogeneidad de varianzas para varias muestras estadísticas.

Otra forma de evaluar la homocedasticidad de un conjunto de muestras estadísticas es por medio del **gráfico de homocedasticidad** y el **diagrama de cajas y bigotes**. El gráfico de homocedasticidad muestra las varianzas de cada grupo en función de sus respectivos promedios. Si las muestras estadísticas son homocedásticas el gráfico no debería mostrar ninguna tendencia muy marcada. Por otro lado, el diagrama de cajas y bigotes representa gráficamente las muestras estadísticas en términos de sus cuartiles. Si las muestras estadísticas son homocedásticas las distintas cajas deberán tener un tamaño similar. Ambos gráficos se incluyen en el ejemplo que se trata a continuación.

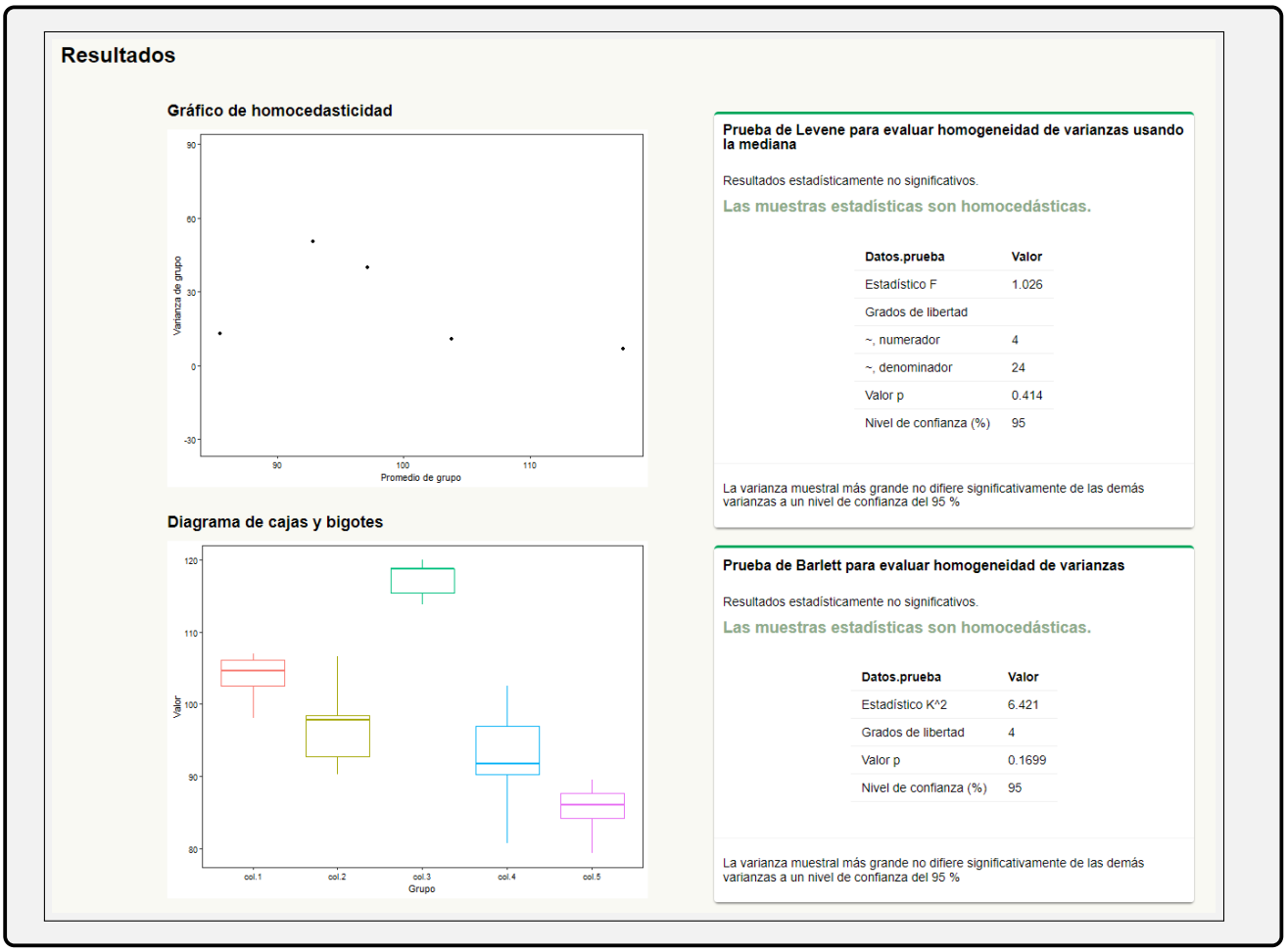
Ejemplo 8. Homocedasticidad de más de dos muestras estadísticas.

Las pruebas estadísticas y los gráficos mencionados en esta sección pueden realizarse en el submódulo **Comparación de varianzas** del módulo **≥ Pruebas de comparación** de la sección **Herramientas estadísticas** del aplicativo validaR.

La tabla a la derecha contiene los resultados de recuperación de distintos blancos de matrices de alimentos que se enriquecieron a un mismo nivel de concentración con un residuo de pesticida (en porcentaje). Se quiere conocer saber si los resultados para las distintas matrices son homocedásticos.

El recuadro a continuación muestra los resultados de validaR para el gráfico de homocedasticidad, el diagrama de cajas y bigotes, y dos de la pruebas de homocedasticidad mencionadas en la Tabla 2.2.

Matriz 1	Matriz 2	Matriz 3	Matriz 4	Matriz 5
105.09	92.75	118.78	89.32	79.40
104.08	97.82	113.81	94.63	83.67
98.14	90.24	119.94	102.48	88.13
102.00	98.46	118.71	99.38	86.56
106.47	106.61	115.45	91.83	85.64
106.95			80.70	89.49
			91.28	



2.2.7 ANOVA: Análisis de varianza para la comparación de varias medias

El Análisis de Varianza (ANOVA) se utiliza para comparar varias medias muestrales a la vez. La hipótesis nula del ANOVA indica que todas las muestras estadísticas provienen de la misma población, con lo que todas las medias poblacionales serían iguales entre sí. La hipótesis alternativa declara que hay diferencia estadísticamente significativa entre al menos dos de las medias de las poblaciones de las que provienen las muestras consideradas:

$$\mathbf{H_0}: \mu_1 = \mu_2 = ... = \mu_k \qquad \mathbf{H_1}: \mu_i \neq \mu_j, \text{ para al menos una de las parejas } (i,\ j) \tag{2.10}$$

El ANOVA es una prueba de comparación de medias, pero se llama análisis de varianza, porque la hipótesis de que todas las muestras estadísticas provienen de la misma población se evalúa en función de la dispersión de los datos: Si las muestras estadísticas provienen de la misma población, la dispersión de los datos entre grupos debería ser similar a la dispersión de los datos dentro de los grupos. En el ANOVA las medidas de dispersión utilizadas son las **sumas de cuadrados** y los **cuadrados medios**, que se calculan de una forma similar a como se calculan las varianzas (ver Tabla 2.1).

Una varianza puede verse como la media de las distancias al cuadrado de los valores frente al promedio del grupo. La media de las distancias al cuadrado se calcula como la sumatoria de los valores, dividida entre los grados de libertad de la sumatoria. La sumatoria de las distancias al cuadrado es lo que se conoce como suma de cuadrados, y el cuadrado medio es el resultado de dividir esa suma de cuadrados entre su número de grados de libertad.

En el ANOVA las medidas de tendencia central contra las que se calculan las sumas de cuadrados son los promedios de cada grupo ($\bar{x}_i$) o el promedio de todos los datos (promedio general, $\bar{\bar{x}}$). Un conjunto de n datos, agrupados en m muestras estadísticas de n_i elementos cada una, puede dar lugar a las siguientes sumas de cuadrados:

- Si se utilizan las distancias de todos los valores frente al promedio general se obtiene la suma de cuadrados totales (SS_{total}), que considera la dispersión total de todos los datos y tiene $n-1$ grados de libertad.

$$SS_{total} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2 \tag{2.11}$$

  donde x_{ij} es el j-ésimo elemento del i-ésimo grupo.

- Si se usan las distancias de los promedios grupales frente al promedio general la suma de cuadrados se denomina entre grupos (SS_{entre}), considera el componente de la dispersión de los datos como consecuencia de su agrupamiento y tiene $k-1$ grados de libertad.

$$SS_{entre} = \sum_{i=1}^{m}(\bar{x}_i - \bar{\bar{x}})^2 \tag{2.12}$$

- Si se usan las distancias de todos los valores frente a sus promedios grupales se obtiene una medida de dispersión dentro de los grupos que se conoce como suma de cuadrados residuales ($SS_{residuales}$) y es el componente de la dispersión de los datos que se atribuye al error aleatorio y tiene $n-k$grados de libertad.

$$SS_{residuales} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \tag{2.13}$$

  Esta suma de cuadrados residuales también se conoce como suma de cuadrados del error (SS_{error}) o suma de cuadrados dentro de los grupos (SS_{dentro}).

Las sumas de cuadrados y sus respectivos grados de libertad (DF) cumplen las siguientes condiciones:

$$SS_{total} = SS_{entre} + SS_{residuales}\ ; \qquad DF_{total} = DF_{entre} + DF_{residuales} \tag{2.14}$$

Las anteriores expresiones permiten obtener el cuadrado medio entre grupos (MS_{entre}) que resume el componente de varianza atribuible al agrupamiento de los datos, y el cuadrado medio de los residuales ($MS_{residuales}$) que resume el componente de varianza aleatorio que se observa dentro de los grupos.

$$MS_{entre} = \frac{SS_{entre}}{DF_{entre}}\ ; \qquad MS_{residuales} = \frac{SS_{residuales}}{DF_{residuales}} \tag{2.15}$$

La relación entre los componentes de varianza (los cuadrados medios) se estudia con una prueba F de Fisher (ver Sección 2.2.6.2).

$$F = \frac{MS_{entre}}{MS_{residuales}} \tag{2.16}$$

De esta manera se puede conocer el valor P asociado al ANOVA. Este valor P estima la probabilidad de que de una misma población estadística se hayan extraído aleatoriamiente las muestras estadísticas que están comparando. Los resultados del ANOVA se suelen presentar como la se muestra en la Tabla 2.3.

	Suma de cuadrados	**Grados de libertad**	**Cuadrado medio**	**Estadístico F**	**Valor p**
Entre grupos	SS_{entre}	DF_{entre}	MS_{entre}	$MS_{entre}/MS_{residuales}$	Valor P
Residuales	$SS_{residuales}$	$DF_{residuales}$	$MS_{residuales}$		

Tabla 2.3: Esquema de una tabla de ANOVA.

Al ejecutar una prueba de ANOVA se asume que las muestras estadísticas son **independientes**, **homocedásticas** y con **distribución normales**. La validez de las conclusiones que se obtengan del análisis depende de que se cumplan estos supuestos.

- La independencia de las muestras estadísticas se cumple si los datos corresponden a observaciones independientes obtenidas de manera aleatoria dentro de cada grupo.
- La homocedasticidad entre las series se puede analizar con alguna de las pruebas que se vieron en la Sección 2.2.6.3.
- La normalidad en la distribución de las muestras estadísticas puede evaluarse por medio de las pruebas que se describieron en la Sección 2.2.1.

Cuando el valor P que resulta del ANOVA es menor al valor de significancia estadística que se escogió para comparación se puede concluir que la diferencia entre al menos dos de las medias aritméticas de las muestras es estadísticamente significativa. Sin embargo, esta conclusión no permite discriminar cuales muestras estadísticas en particular son las que difieren entre sí. Para este fin se utilizan las pruebas post hoc que se presentan en la siguiente sección.

Ejemplo 9. ANOVA para comparar varias medias muestrales.

En el Ejemplo 8 se estudió la homocedasticidad de cinco conjuntos de datos de porcentajes de recuperación de un residuo de pesticida que se enriqueció en cinco distintos blancos de matriz al mismo nivel de concentración. En este ejemplo se utilizan los mismos datos para establecer si hay diferencias en las medias aritméticas de los grupos.

El ANOVA para identificar si al menos una pareja de grupos presenta diferencias estadísticamente significativas entre sí se puede realizar en el submódulo **ANOVA - Análisis de varianza**, del módulo **≥ Pruebas de comparación** de la sección **Herramientas estadísticas** del aplicativo validaR. Los resultados se muestran en el recuadro que aparece abajo.

Resultados

Resultados del Analisis de Varianza

Resultados estadísticamente significativos:

Las medias de al menos dos muestras estadisticas son diferentes.

Tabla del análisis de varianza:

Fuente.de.variación	G.L.	Suma.de.cuadrados	Cuadrado.medio	Estadístico.F	Valor.p
Entre grupos	4	3194.9	798.73	31.621	< 1 e-04
Dentro de los grupos	24	606.23	25.26		

Se recomienda hacer pruebas de comparaciones múltiples

La diferencia entre al menos dos de las medias muestrales es estadísticamente significativa a un nivel de confianza del 95 %.

Supuestos del análisis

Independencia

- Los datos deben corresponder a observaciones independientes obtenidas de manera aleatoria dentro de cada grupo.

Normalidad de los residuales

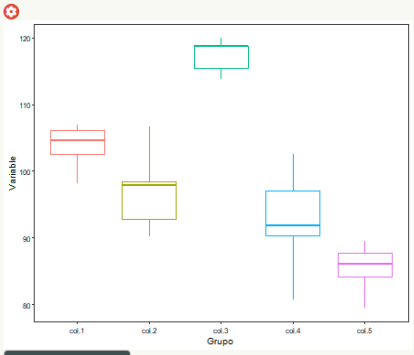
- Los residuales del ANOVA parecen seguir una distribución normal
  Valor p de la prueba Shapiro-Wilk de los residuales: 0.4988
  Revise los gráficos de residuales por grupo.

Homocedasticidad entre series

- No se encontró evidencia en contra del supuesto de homogeneidad de varianzas entre los grupos.

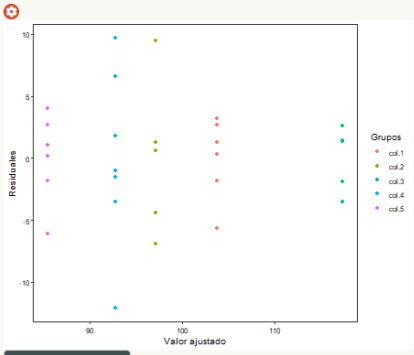
Gráficos del ANOVA

Cajas y bigotes



Descargar gráfico

Residuales por grupo



Descargar gráfico

2.2.7.1 Pruebas post hoc

Cuando los resultados de un ANOVA son estadísticamente significativos se puede concluir que al menos dos medias muestrales presentan entre sí una diferencia estadísticamente significativa. En estos casos se entiende que la probabilidad de que alguna de las diferencias observadas entre las medias muestrales se haya presentado por mero error aleatorio es muy baja. Cuando esto ocurre procede establecer cuales parejas de conjuntos de datos presentan estas diferencias estadísticamente significativas entre sí. Este tipo de análisis se denomina *post hoc* y deben ejecutarse luego de que los resultados del ANOVA son estadísticamente significativos.

Las pruebas de comparación múltiple inicialmente definen un valor límite para la diferencia entre medias muestrales a partir del cual la diferencia se considera como estadísticamente significativa. En un segundo paso se comparan las diferencias entre las posibles combinaciones de medias muestrales contra tal límite. Los límites que establecen las pruebas de comparación múltiple pueden ser fijos para evaluar todas las diferencias, o pueden ser variables, en cuyo caso sus valores dependen de la pareja de muestras que se están comparando.

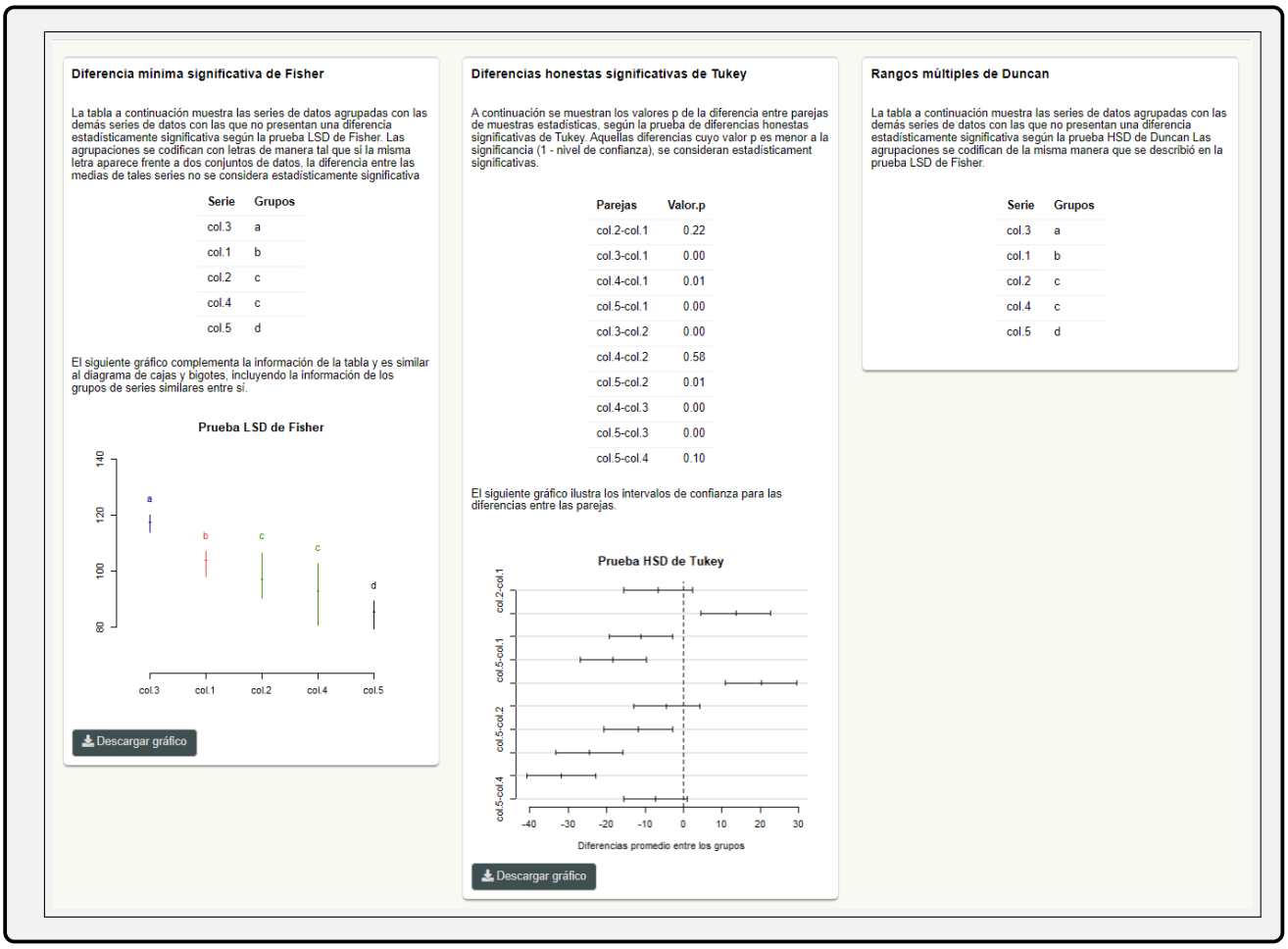
Las pruebas de comparación múltiple más comunes son la **prueba de diferencia mínima significativa de Fisher** (LSD, por sus siglas en inglés), la **prueba de diferencia mínima honesta de Tukey** (HSD, por sus siglas en inglés) y la **prueba de rangos múltiples de Duncan**. Estas pruebas similares a la prueba *t* de Student de comparación de una media muestral contra otra media muestral (Sección 2.2.5.2), solo que hacen uso de estadísticos diferentes.

Los resultados de las pruebas mencionadas se muestran en el ejemplo que se trata en esta sección. La prueba LSD de Fisher y la prueba de rángos múltiples de Duncan agrupan las muestras estadísticas que no muestran diferencias significativas en sus medias aritméticas. Los resultados de estas pruebas suelen coincidir entre sí. Una misma muestra estadística puede encontrarse en más de un grupo que comparte con diferentes muestras estadísticas. Las pruebas dan como resultado una tabla en la que relacionan a que grupos se pueden asociar las muestras estadísticas y un gráfico donde las medias de las muestras estadísticas se ordenan de mayor a menor y se colorean de acuerdo a los grupos al los que pertenecen.

La prueba HSD de Tukey estudia las diferencias entre las muestras estadísticas contra un criterio variable que determina si las diferencias son estadísticamente significativas y que depende de que tan alejadas están las medias muestrales cuando se organizan de menor a mayor. De esta manera se puede obtener un valor P asociado a cada diferencia entre las parejas para establecer cuales son estadísticamente significativas. Los resultados de la prueba incluyen una tabla con los valores P para todas las parejas posibles de muestras estadísticas y un gráfico con los intervalos de confianza para dichas diferencias. Los intervalos de confianza que no incluyen el valor cero se consideran estadísticamente significativos.

Ejemplo 10. Pruebas post hoc de comparación múltiple

En el Ejemplo 9 se encontraron diferencias estadísticamente significativas entre algunos de los resultados de recuperación de un residuo de pesticida en cinco diferentes tipos de blancos de matriz que se enriquecieron al mismo nivel de concentración. En este ejemplo se buscan las parejas de conjuntos que presentan diferencias importantes entre sí por medio de comparaciones múltiples disponibles en el submódulo **Pruebas de contrastes post hoc**, del módulo ≥ **Pruebas de comparación** de la sección **Herramientas estadísticas** del aplicativo validaR. Los resultados se muestran en el recuadro que aparece abajo.



2.3 Modelos de regresión lineal: Relación entre variables cuantitativas

La calibración de un método analítico por lo general inicia con un conjunto de patrones de medición que se utiliza para conocer la relación entre la intensidad de una señal instrumental y la concentración de un analito. Esta información permite predecir los valores de concentración del analito que le correspondería a nuevas muestras a las que se les aplica el proceso de medición (JCGM, 2012). La relación entre las señales y los niveles de concentración correspondientes se pueden expresar en forma de una ecuación matemática donde la intensidad de la señal es función de la propiedad de interés ($y = F(x)$). En química analítica se suelen emplear modelos lineales para describir estas relaciones. La variable explicatoria es la concentración de una especie química en una porción de muestra analítica (o en una disolución que se prepara a partir de la muestra). La variable respuesta es la señal que se observa en un equipo de medición tras aplicar un protocolo de medición.

El uso de un modelo lineal en una calibración analítica considera que los valores de las señales instrumentales se pueden modelar como un valor constante más una cantidad proporcional al valor de la concentración de analito. Este tipo de relación se ilustra en la Figura 2.6 y su representación matemática se muestra en la Ecuación 2.17. El componente constante de la variable respuesta es el intercepto de la relación lineal con el eje y (β_0), y el factor de proporcionalidad frente a la magnitud de interés es la pendiente de la recta (β_1). Las desviaciones de la variable respuesta frente al modelo se asumen como consecuencia del error aleatorio (ε).

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon \tag{2.17}$$

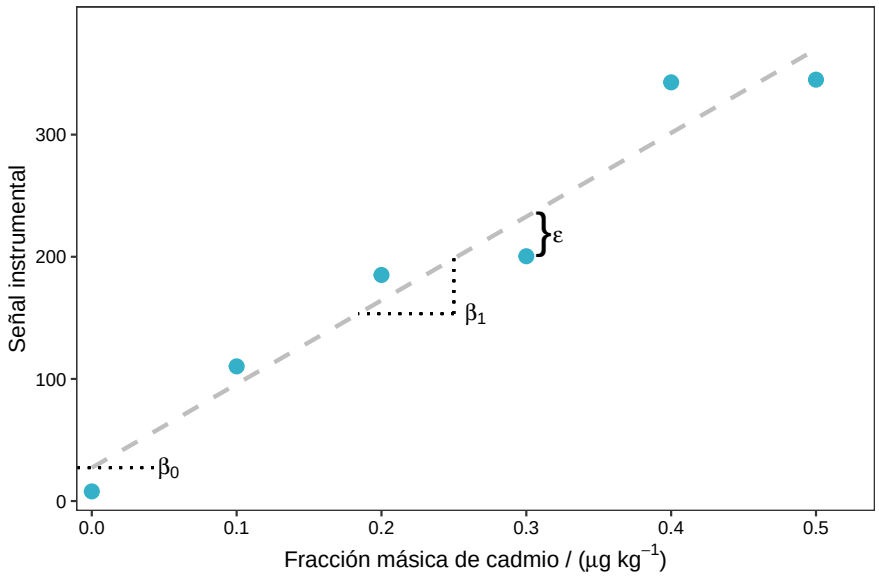


Figura 2.6: Gráfico de la señal instrumental de un método de medición en función de la fracción másica de ion cadmio en disolución. La línea descontinua gris corresponde a una recta de la forma $\beta_0 + \beta_1 \cdot x$.

Los parámetros reales de la regresión (β_0 y β_1) no se pueden conocer porque el error aleatorio (ε) es impredecible. Sin embargo, los estimadores del intercepto y la pendiente de regresión pueden obtenerse por medio de distintos procedimientos. Estos estimadores de intercepto y pendiente se suelen representar como a y b, respectivamente.[2] La Ecuación 2.18 representa la relación entre las variables de la Ecuación 2.17, pero reescrita en términos de los estimadores de los parámetros de regresión.

$$y = a + b \cdot x \tag{2.18}$$

Los métodos más comunes para estimar los parámetros de regresión hacen parte de una familia que se denomina algoritmos de optimización por mínimos cuadrados. Estos algoritmos buscan minimizar el error de predicción del modelo lineal, en términos de la suma de las diferencias entre los valores predichos por el modelo y los valores observados de las variables elevadas al cuadrado. La minimización de la suma de cuadrados puede hacerse de diversas formas, dependiendo de las características del conjunto de datos. Los algoritmos más representativos de esta familia se describen en las siguientes secciones (Ripley y Thompson, 1987; Therneau, 2018).

2.3.1 Mínimos cuadrados ordinarios (OLS)

Es el método más utilizado para regresión lineal. La estimación de los parámetros de regresión por mínimos cuadrados ordinarios (OLS, de *ordinary least squares*) describen la línea recta que minimiza la distancia promedio entre los valores observados de la variable respuesta (y_i) y los valores de variable respuesta que predice el modelo lineal (y_i^*), como se muestra en la Figura 2.7. El calculo de los parámetros de regresión por OLS se hace con las siguientes ecuaciones:

$$b = \frac{\sum_i^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_i^n (x_i - \bar{x})^2} \; ; \qquad a = \bar{y} - b \cdot \bar{x} \tag{2.19}$$

donde n es el número de datos, $\bar{x}$ es el promedio de los valores de la variable explicatoria, x_i es el i-ésimo valor de la variable explicatoria, $\bar{y}$ es el promedio de los valores de la variable respuesta, y_i es el valor de respuesta que corresponde al i-ésimo valor de la variable explicatoria.

[2] En algunos contextos se utiliza b_0 y b_1 para los estimadores del intercepto y la pendiente, respectivamente. En otros lugares se utiliza m para el intercepto y b para la pendiente.

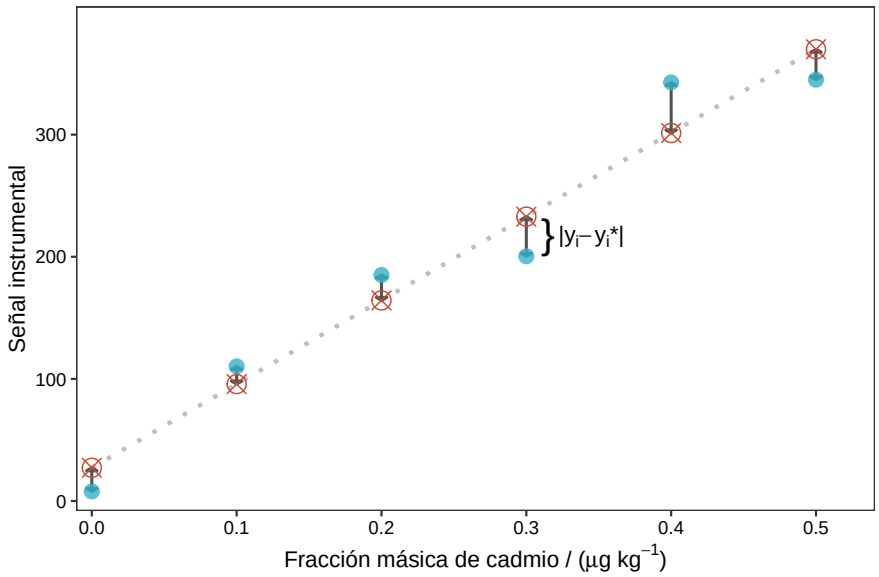


Figura 2.7: Regresión lineal por mínimos cuadrados ordinarios para los datos que se muestran en la Figura 2.6 (puntos azules). Los círculos rojos son los valores predichos por el modelo.

Este tipo de regresión es la apropiada para describir la relación entre variables que se correlacionan de manera lineal cuando se cumplen los siguientes supuestos:

1. Los residuales de regresión tienen distribución normal.
2. Los errores de las variables son independientes entre sí (los residuales no están autocorrelacionados).
3. El error experimental está principalmente en la variable respuesta (el error asociado a variable explicatoria es despreciable).
4. El error de la variable respuesta se puede considerar constante (homocedástico).

2.3.1.1 Supuestos del modelo de regresión por OLS

La mayoría de los supuestos del modelo de OLS se evalúan en función de los **residuales de regresión** (R_i), que corresponden a la diferencia entre los valores experimentales de la variable respuesta y los valores que predice el modelo lineal ($R_i = y_i - y_i^*$). Los residuales de regresión se pueden representar gráficamente en función de la variable explicatoria, como se muestra en los ejemplos de la Figura 2.8. La inspección visual del gráfico de residuales ayuda a identificar tendencias no aleatorias que pueden sugerir que el modelo lineal no describe adecuadamente la relación entre los datos. La Sección **3.6.3** describe estrategias para evaluar formalmente si un modelo lineal es adecuado para describir el ajuste de los datos de una curva de calibración.

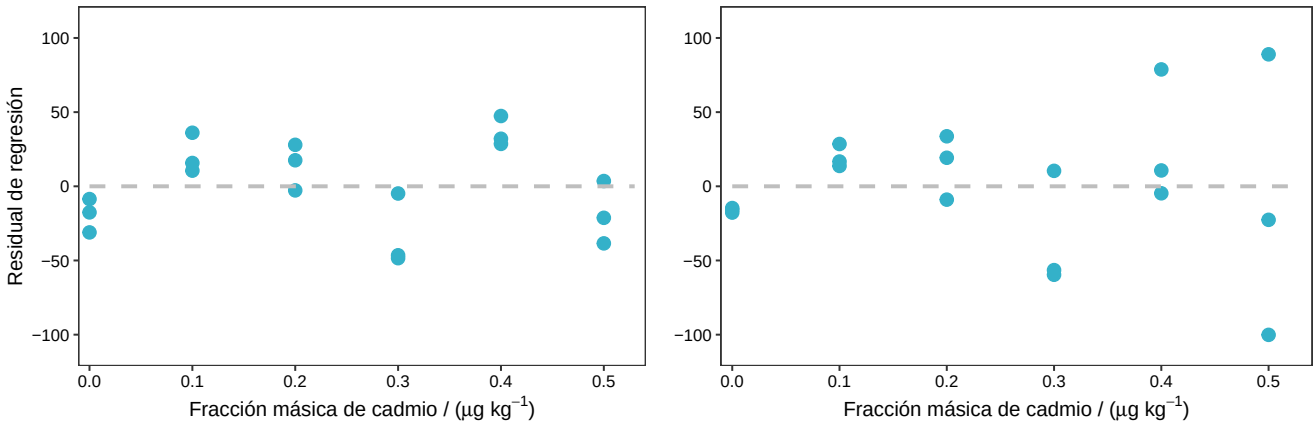
La **normalidad de los residuales de regresión** implica que los puntos del gráfico de residuales deberían organizarse aleatoriamente a ambos lados del eje x del plano cartesiano. Este supuesto se puede evaluar formalmente por medio de las pruebas de normalidad que se describieron en la Sección 2.2.3. Si los residuales de regresión no siguen una distribución normal puede ser útil aplicar un método de regresión no paramétrica como la regresión de Passing-Bablock que se expone en la Sección 2.3.5.

La **independencia de los residuales de regresión** se puede estudiar por medio de la prueba de autocorrelación de Durbin-Watson, en la que se estima la probabilidad de que los residuales de regresión experimentales se hayan generado de una población estadística de residuales que no están autocorrelacionados. Si los residuales de regresión muestran una marcada autocorrelación puede ser necesario estimar los parámetros de regresión por medio del algoritmo de mínimos cuadrados generalizados que se trata en la Sección 2.3.4

El **error significativo presente únicamente en la variable respuesta** depende de las características del conjunto de datos. Si este supuesto se cumple es razonable que las distancias que se minimizan entre los valores experimentales y los valores que predice el modelo sea únicamente en la dirección del eje y. Cuando el error en la variable explicatoria es comparable con el error en la variable respuesta la suposición no es correcta y la distancia que se

minimiza entre los valores experimentales y los valores del modelo es en ambas direcciones (eje x y eje y), dando lugar a la regresión por mínimos cuadrados de distancia ortogonal que se describe en la Sección 2.3.3.

El supuesto de **homocedasticidad de los residuales** se puede evaluar visualmente en el gráfico de residuales. Si los residuales son homocedásticos la dispersión de los residuales debería ser aproximadamente constante en todo el intervalo, como se muestra en la Figura 2.8(a). Por otro lado, la Figura 2.8(b) muestra un gráfico de residuales típico datos de regresión donde el error es heterocedástico. Alternativamente la homocedasticidad de los residuales de regresión se puede evaluar por medio de la prueba de Breush-Pagan que funciona de manera similar a las pruebas que se describieron en la Sección 2.2.6.3, pero que sirve para evaluar la homocedasticidad de un conjunto de datos en función de una variable cuantitativa. Si hay evidencia de que los residuales de regresión no son homocedásticos la estimación de los parámetros de regresión debe hacerse utilizando el algoritmo de mínimos cuadrados ponderados que se describe en la Sección 2.3.2



(a) Caso de datos homocedásticos.

(b) Caso de datos heterocedásticos.

Figura 2.8: Ejemplos de gráficos de residuales para mediciones por triplicado de los datos que se muestran en la Figura 2.7.

Cuando el análisis del gráfico de residuales sugiere que la relación entre las variables estudiadas no es lineal, se puede acotar el intervalo estudiado para considerar únicamente la región en la que la relación es lineal, se pueden utilizar modelos de regresión empíricos basados en polinomios de orden mayor (Raposo y Barceló, 2021), o se pueden utilizar modelos de regresión más apropiados que utilizan información concisa del fenómeno bajo estudio (Pagliano, Mester y Meija, 2015).

Ejemplo 11. Regresión lineal por OLS y supuestos del modelo de regresión.

El módulo **Regresión lineal** de la sección **Herramientas estadísticas** del aplicativo validaR incluye un submódulo para calcular modelos de regresión utilizando los algoritmos que se ven en esta sección.

La tabla a la derecha contiene datos de una curva de calibración de plomo que se midió por espectrometría de absorción atómica.

El recuadro que aparece a continuación muestra los resultados del aplicativo cuando calcula el modelo de regresión lineal con los datos de la tabla utilizando el algoritmo OLS.

Fracción másica / [mg/kg]	Señal de absorbancia / [UA]
0.0	0.009
2.0	0.158
4.0	0.301
6.0	0.472
8.0	0.577
10.0	0.739

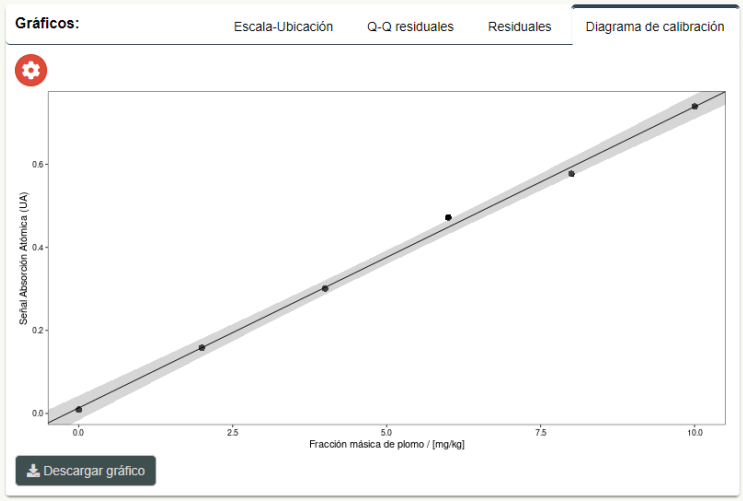
Resultados

Curva de regresión por mínimos cuadrados ordinarios:
Y = 0.0132857 + 0.0725429 * X

Error residual estándar de la regresión: 0.0145891

Detalle de los parámetros de regresión:

.	Intercepto	Pendiente
Parámetro	0.0132857	0.0725429
Error estándar	0.0105588	0.00174374
Intervalo de confianza		
~ límite inferior	-0.0160303	0.0677015
~ límite superior	0.0426018	0.0773842



El aplicativo hace una revisión de los supuestos de normalidad de los residuales (prueba de Shapiro-Wilk), ausencia de autocorrelación del error (prueba de Durbin-Watson) y de homocedasticidad de los residuales (prueba de Breush-Pagan). Los resultados de la revisión de supuestos se muestran en el siguiente recuadro:

Supuestos de los modelos:

Supuestos del análisis: Normalidad de los residuales

Los residuales parecen tener distribución normal. Valor p prueba de Shapiro-Wilk sobre los residuales de la regresión: 0.2328 Se recomienda verificar esta información con el gráfico Q-Q de los residuales.

Supuestos del análisis: Ausencia de autocorrelación de error

No hay una aparente autocorrelación del error:

.	Valor
Autocorrelación	-0.53
Estadístico Durbin-Watson	3.04
Valor p	0.29

Se recomienda verificar esta información con el gráfico de residuales.

Supuestos del análisis: Homocedasticidad del error

Hay homocedasticidad en los residuales de regresión. Valor p prueba de Breusch-Pagan: 0.4452 Se recomienda verificar esta información con el gráfico de residuales y el gráfico de ubicación de escala.

2.3.1.2 Interpolación en un modelo de regresión lineal

Una calibración termina estableciendo una relación que permite obtener nuevos valores del mensurando a partir de nuevos valores de señal (JCGM, 2012). Cuando se usa un modelo de regresión lineal este proceso se hace por interpolación. Para interpolar un nuevo valor de variable respuesta basta con despejar el factor x de la Ecuación 2.18, para lo cual es necesario conocer los estimados de los parámetros de regresión (a y b):

$$x_0 = \frac{y_0 - a}{b} \tag{2.20}$$

donde x_0 es el valor de variable explicatoria que le corresponde al valor de variable respuesta y_0.

El valor y_0 que se interpola en el modelo de regresión puede corresponder a un valor individual de variable respuesta de una muestra analítica que se midió solo una vez, o puede ser el promedio de varias réplicas de

medición que se tomaron de la misma muestra analítica. El valor interpolado carga una incertidumbre que proviene del error asociado al modelo de regresión ($u(x_i)$). Esta incertidumbre se calcula de acuerdo a la siguiente ecuación:

$$u(x_0) = \frac{s_{y/x}}{b_1}\sqrt{\frac{1}{p}+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{j=1}^{n}(x_j-\bar{x})^2}} \tag{2.21}$$

donde $s_{y/x}$ es el error residual estándar del modelo de regresión, p es el número de repeticiones de la señal que se promedian para obtener y_i, n es el número de estándares de calibración que se usaron para elaborar el modelo, $\bar{x}$ es la concentración promedio de los estándares de calibración y x_j es la concentración del j-ésimo estándar de calibración.

Ejemplo 12. Interpolación por OLS e incertidumbre del valor interpolado.

En el Ejemplo 11 se calculó el modelo de regresión lineal por OLS para una curva de calibración de plomo utilizando el módulo **Regresión lineal** del aplicativo validaR. En este ejemplo se usa la sección **Interpolar nuevos valores** de ese mismo módulo para predecir la fracción másica de plomo un una nueva disolución.

Luego de medir la curva de calibración del Ejemplo 11 se midió una muestra analítica por triplicado y los valores de absorbancia que se obtuvieron fueron 0.444, 0.448 y 0.447 UA.

Abajo a la izquierda se muestra el resultado del valor interpolado junto con la incertidumbre por interpolación cuando se considera que los tres valores de señal corresponden a réplicas de la misma muestra. El recuadro de abajo a la derecha muestra el resultado de interpolar los mismos valores pero asumiendo el escenario en el que corresponden a muestras independientes que se midieron una sola vez ($p = 1$).

Observe que la incertidumbre del valor interpolado es menor cuando se utilizan varias repeticiones de medición de la misma alícuota. Adicionalmente, si se tienen réplicas de la curva de calibración, el número de estándares de calibración aumenta proporcionalmente y la incertidumbre del valor interpolado también disminuye.

Resultados:

Respuesta	Valor.interpolado	Incertidumbre.por.modelo
0.4463333	5.969542	0.1441

Resultados:

Respuesta	Valor.interpolado	Incertidumbre.por.modelo
0.444	5.937377	0.21839
0.448	5.992517	0.21853
0.447	5.978732	0.21849

2.3.2 Mínimos cuadrados ponderados (WLS)

La mayoría de calibraciones analíticas producen datos que son inherentemente heterocedásticos: las réplicas de las señales instrumentales tienden a presentar una dispersión más grande a medida que aumenta la concentración de los patrones de calibración (Ketkar y Bzik, 2000). Esto implica que el supuesto de homocedasticidad de los residuales visto en la sección anterior no se cumple en muchos casos. Ignorar esta característica de la mayoría de las calibraciones analíticas puede ocasionar un incremento en el sesgo de la cuantificación de analitos, particularmente en la región de bajas concentraciones (Funke, Sperling y Karst, 2021).

El algoritmo de mínimos cuadrados ponderados (WLS, de *weighted least squares*) considera la heterocedasticidad de los datos para que los valores con menor dispersión tengan un mayor efecto en el cálculo de los parámetros de regresión. El error asociado a cada punto de la calibración se utiliza para asignar factores de ponderación (w_i) que se usan para minimizar una suma de cuadrados ponderada. Los parámetros de regresión en WLS se estiman con las siguientes ecuaciones:

$$b = \frac{\sum_i^n (w_i \cdot x_i \cdot y_i) - n \cdot \bar{x}_w \cdot \bar{y}_w}{\sum_i^n (w_i \cdot x_i^2) - n \cdot \bar{x}_w^2}\ ; \qquad a = \bar{y}_w - \beta_1 \cdot \bar{x}_w \tag{2.22}$$

donde n es el número de datos, $\bar{x}_w$ es el promedio ponderado de los valores de la variable explicatoria, x_i es el i-ésimo valor de la variable explicatoria, $\bar{y}_w$ es el promedio ponderado de los valores de la variable respuesta, y_i es

el valor de respuesta que corresponde al i-ésimo valor de la variable explicatoria y w_i es el factor de ponderación:

$$w_i = \frac{1}{s_i^2}\ ; \qquad \bar{x}_w = \frac{\sum_i^n w_i \cdot x_i}{\sum_i^n w_i}\ ; \qquad \bar{y}_w = \frac{\sum_i^n w_i \cdot y_i}{\sum_i^n w_i} \tag{2.23}$$

donde s_i es la desviación estándar del i-ésimo punto.

La Figura 2.9 presenta un ejemplo típico de resultados de una regresión lineal por WLS. En los datos de este gráfico las señales instrumentales tienen un coeficiente de variación constante, de manera que la desviación estándar de los valores aumenta de manera lineal con la concentración de analito. La recta que predice el modelo tiende a pasar más cerca de los puntos experimentales que tienen un error estándar más pequeño, en este caso, ubicados en la región de baja concentración de la curva de calibración.

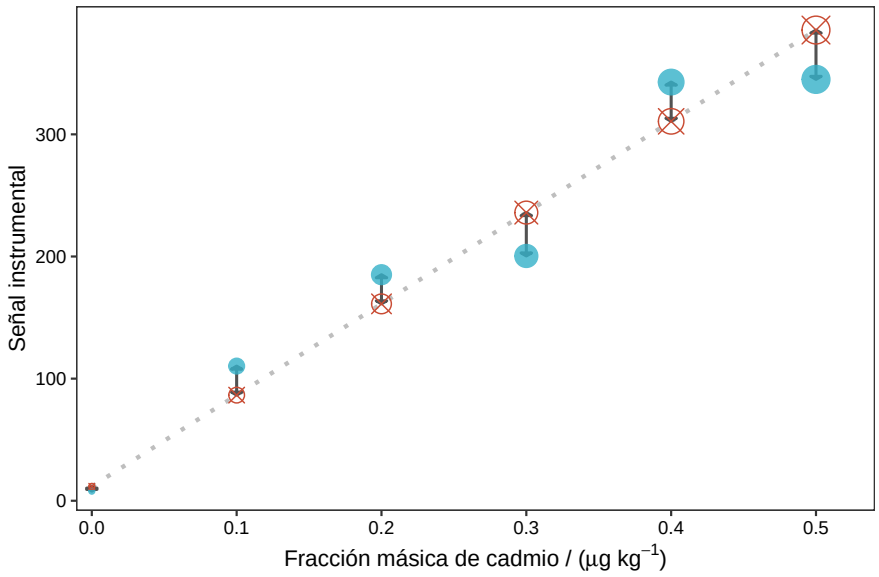


Figura 2.9: Regresión lineal por mínimos cuadrados ponderados para los datos que se muestran en la Figura 2.6 (puntos azules). El tamaño de los puntos es proporcional a su error experimental. Los círculos rojos son los valores predichos por el modelo.

El aplicativo validaR puede calcular modelos de regresión por WLS si se ingresan los datos de desviación estándar correspondiente a cada punto.

2.3.3 Mínimos cuadrados de distancia ortogonal (ODR)

En algunos casos el error en el eje x es comparable en magnitud al error en el eje y. Esto implica que no se cumple el supuesto de OLS de que el error significativo solo está presente en la variable respuesta. En esta situación la recta de OLS que minimiza las distancias verticales entre los puntos experimentales y los puntos que predice el modelo no es adecuada (Haeckel, Wosniok y Klauke, 2013). Un ejemplo común de cuando hay error en los datos de ambos ejes es cuando se usa un análisis de regresión para comparar los resultados que producen dos métodos de medición independientes, para un conjunto de muestras analíticas en un intervalo de concentraciones de analito. La solución para trabajar con este tipo de datos consiste en minimizar la distancia ortogonal entre los puntos experimentales y la recta de la regresión. Este tipo de regresión se conoce como mínimos cuadrados de distancia ortogonal (ODR, de *ortogonal distance regression*) y se ilustra en la Figura 2.10.

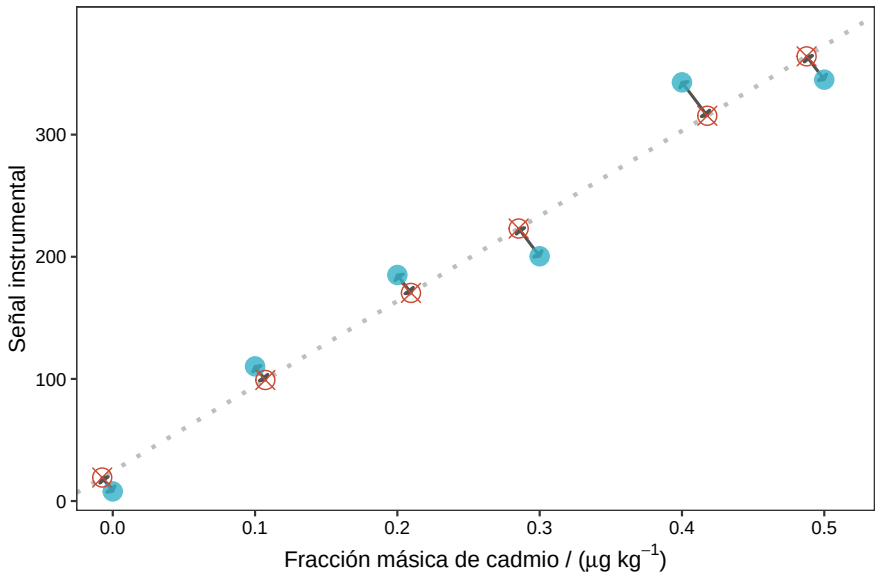


Figura 2.10: Regresión lineal por mínimos cuadrados de distancia ortogonal para los datos que se muestran en la Figura 2.6 (puntos azules). Los círculos rojos son los valores de la proyección ortogonal de los puntos en la recta del modelo.

En la estimación de parámetros de regresión por ODR se asume que el error de los datos experimentales sigue una distribución normal, es similar para cada punto en ambas variables y es constante en el intervalo de valores considerado. Si el error de cada punto experimental es igual para ambas variables pero no es constante entre todos los puntos puede utilizarse una versión ponderada del algoritmo que incorpora la característica que se estudió en la sección anterior.

2.3.4 Mínimos cuadrados generalizados (GLS)

La regresión lineal por mínimos cuadrados generalizados (GLS, de *generalized least squares*) es de utilidad cuando se conoce el error asociado a cada punto para ambas variables que se relacionan (Haeckel, Wosniok y Klauke, 2013). Con esta información ya no es necesario asumir que los errores en el eje x y en el eje y son iguales para todos los puntos y constantes en el intervalo de valores que se analiza. Este tipo de regresión se ejemplifica en la Figura 2.11.

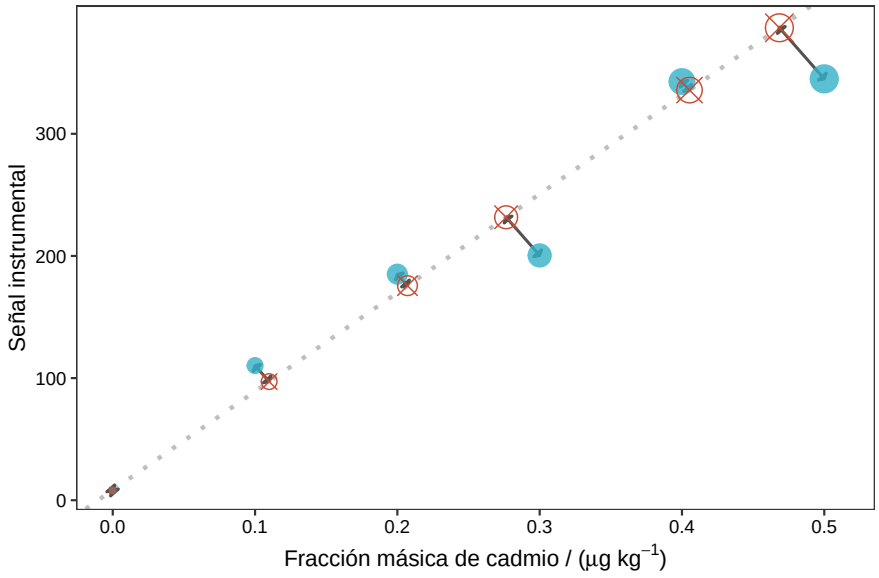


Figura 2.11: Regresión lineal por mínimos cuadrados generalizados para los datos que se muestran en la Figura 2.6 (puntos azules). El tamaño de los puntos es proporcional a su error experimental en ambas variables. Los círculos rojos son los valores de la proyección de los puntos en la recta del modelo GLS.

El único supuesto de la regresión lineal por GLS es que los errores de las variables presentan una distribución normal. Si este no es el caso, lo más conveniente es utilizar algún método de regresión lineal no paramétrico.

2.3.5 Regresión no paramétrica: Método de Passing-Bablok

Los métodos de regresión no paramétricos son útiles en la estimación de parámetros cuando no puede asumirse que los errores en las variables tienen una distribución normal. Este método es muy robusto a la presencia de datos anómalos dado que los valores extremos casi no afectan los resultados de los parámetros de regresión. Un método no paramétrico de regresión lineal muy utilizado en la comparación de resultados de métodos analíticos es la regresión de Passing-Bablok.

En la regresión de Passing-Bablok se calculan las pendientes de las rectas que conectan todas las parejas de puntos que pueden tomarse del conjunto de datos, y la mediana de las pendientes se toma como la pendiente de la curva de regresión lineal (b). El intercepto de la curva se calcula de la misma manera a como se hace en la regresión por OLS, como se mostró en la Ecuación 2.19. La Figura 2.12 ilustra el proceso de regresión no paramétrica usando el método de Passing-Bablok.

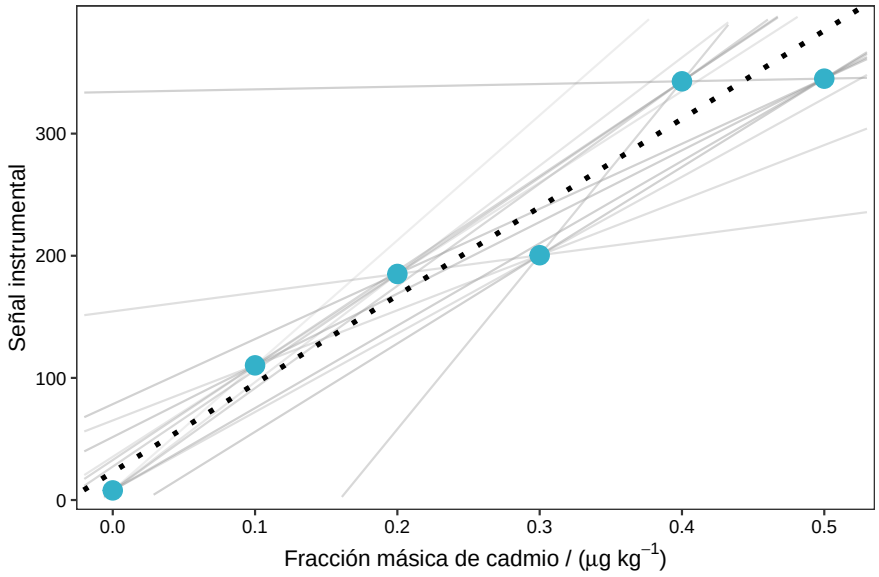


Figura 2.12: Regresión no paramétrica de Passing-Bablok para los datos que se muestran en la Figura 2.6 (puntos azules). Las líneas grises son las posibles rectas entre parejas de datos que pueden formarse y la línea punteada negra tiene como pendiente la mediana de las pendientes de esas curvas.

3. Parámetros de validación de métodos analíticos

3.1 Generalidades

La validación de los métodos analíticos se ha constituido en la base de los sistemas de aseguramiento y control de calidad; por ejemplo, una vez finalizada la validación los métodos es posible establecer estrategias para controlar la variabilidad de los resultados, evaluar el personal del laboratorio, establecer criterios para aceptar o rechazar repeticiones de medición de muestras, aceptación de porcentajes de recuperación de muestras fortificadas, gráficos de control, estimar la incertidumbre de los resultados, entre otros.

Sin embargo, previo al establecimiento de estos planes, programas o actividades fundamentados en los resultados de validación, e inclusive previo a la ejecución de los diseños experimentales que conducen a estos resultados, es indispensable que se realice un proceso de planeación en el que se incluyan las regulaciones, normativas, guías, entre otros documentos emitidos por diferentes entidades, comités o asociaciones, con el propósito de establecer los criterios de aceptación y los parámetros que son necesarios validar.

La validación de métodos analíticos es un proceso que se ha realizado por décadas, sin embargo, el sector farmacéutico fue el primero en desarrollar guías y normatividad, razón por la cual en la literatura es relativamente fácil de encontrar los parámetros que se deben validar en este sector, los criterios de aceptación, el número de réplicas, entre otros; lo anterior facilita enormemente realizar el proceso de planeación. Sin embargo, para muchos otros sectores, los métodos analíticos no se cuentan con este tipo de normatividad, regulaciones o guías que permitan planear la validación y establecer los criterios de aceptación para determinar si un parámetro de la validación cumple o no. Por lo anterior, se han creado ecuaciones empíricas o guías generales que permiten extrapolar estos criterios y requisitos a muchos métodos analíticos sin importar la aplicación.

Antes de seleccionar los parámetros de acuerdo a las recomendaciones de este numeral, se debe revisar que dentro de regulaciones aplicables al alcance del método no se establezcan los parámetros a validar o requisitos de aceptación. Posteriormente, se debe revisar si no existen normas o reglamentos técnicos o de producto que puedan aplicar.

En el caso en que no existan ni regulaciones o guías aplicables, la selección de los parámetros se debe realizar de la siguiente manera:

- Clasificar el método de medición.
- Seleccionar los parámetros de validación.

3.2 Clasificación de métodos: establecimiento de parámetros de desempeño

Los métodos de medición se pueden clasificar en 4 grandes grupos, de acuerdo con el tipo de prueba a realizar:

- Pruebas específicas: son pruebas destinadas a determinar propiedades fisicoquímicas de las muestras; por ejemplo, humedad, densidad, pH, entre otros.
- Pruebas cualitativas: son pruebas destinadas a determinar la identidad de un analito o grupo de analitos, por ejemplo: infrarrojo, cromatografía (tiempo de retención), espectrometría de masas.
- Pruebas semi-cuantitativas: son pruebas destinadas a determinar la concentración aproximada de un analito o grupo de analitos en una muestra, por ejemplo: inmunoensayos.
- Pruebas cuantitativas: son pruebas destinadas a determinar la concentración de un analito en una muestra.

Una vez clasificado el método de medición, la selección de los parámetros de validación se puede realizar de acuerdo con la Figura 3.1.

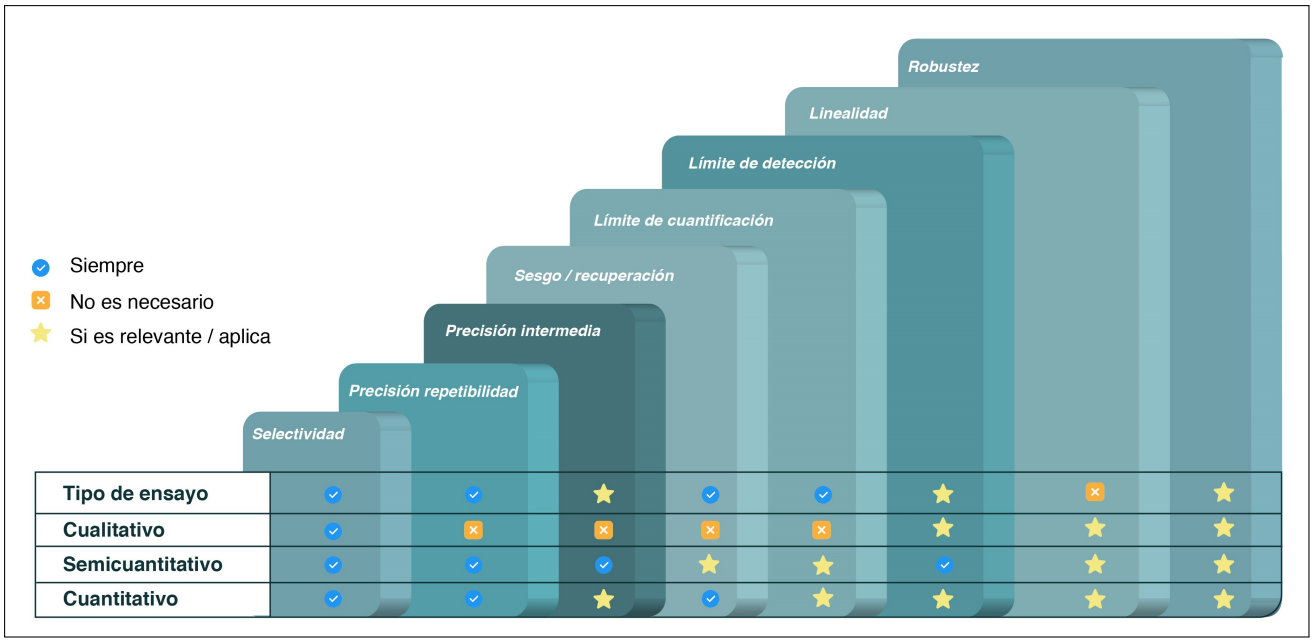


Figura 3.1: Selección de los parámetros de validación. Adaptado de van Zoonen y col., 1999.

Es importante considerar que la evaluación de la relevancia en la selección del parámetro depende de muchos factores como: la complejidad del método, el uso, la experticia y la disponibilidad del personal, los instrumentos de medición, las concentraciones que se desean cuantificar o identificar, y la disponibilidad de materiales de referencia en la matriz y concentración de interés. La Tabla 3.1 expone algunos ejemplos que pueden servir de guía para la selección de los parámetros *opcionales*.

Método	Parámetro	Situación	Decisión
Cualitativo	Límite de detección	Método de FRX para la identificación de Fe en aceros	No es relevante, debido a que el hierro se encuentra a concentraciones altas, lejos de la capacidad de detección del instrumento.
Cuantitativo	Robustez	Método para la medición directa de Pb por ICPMS	No es relevante, pues es un método que no tiende a sufrir pequeños cambios en su ejecución, a menos de que el analista establezca que pueden existir cambios que afecten el resultado asociado a componentes del equipo.
Cuantitativo	Precisión en condiciones intermedias	Un solo analista ejecuta el ensayo	No es relevante hacer interanalistas
Cuantitativo	Precisión en condiciones intermedias	El laboratorio debe hacer un monitoreo de un proceso y se realizarán varias mediciones en el tiempo	Es muy relevante realizar el estudio de precisión intermedia (entre días).

Tabla 3.1: Ejemplos en los que se puede omitir o no omitir la evaluación de un determinado parámetro de validación.

3.3 Selectividad

La selectividad es la propiedad de un sistema de medición para proporcionar valores medidos para uno o varios mensurandos que sean independientes entre sí o de otras magnitudes existentes en el fenómeno, cuerpo o sustancia en estudio (JCGM, 2012), empleando un procedimiento de medición especificado. La definición de la IUPAC habla de selectividad como el grado al cual el método puede ser usado para determinar un analito particular en mezclas o matrices, sin la interferencia de otros componentes que se comportan de manera similar (Vessman y col., 2001).

Los métodos analíticos por lo general gozan de selectividad como consecuencia de un proceso de separación (como en la cromatografía iónica con detector conductimétrico), gracias al uso de un detector selectivo (como en la espectrometría de fluorescencia de rayos X), o por medio de la combinación de ambos mecanismos (como es el caso de la cromatografía líquida acoplada a un espectrómetro de masas). La utilización simultánea de una técnica de separación diferencial con una técnica de detección selectiva, da lugar a las denominadas técnicas acopladas que, por lo general, presentan muy altos niveles de selectividad.

3.3.1 ¿Selectividad o especificidad?

Al consultar algunas de las normativas o guías existentes en lo que se refiere a validación de métodos se observa, en general, que el modo de referirse a cada uno de los parámetros de validación requeridos, es el mismo, por ejemplo, términos como repetibilidad, precisión intermedia, linealidad, entre otros, se denominan de igual manera. Sin embargo, el uso de los términos de selectividad y especificidad no es uniforme en todos los casos, como se puede observar en la Tabla 3.2, e inclusive algunas guías como la de EURACHEM, los tratan de manera equivalente.

Parámetro	Referencia
Especificidad	USP, EURACHEM, VAM, ICH
Selectividad	ISO 17025, EURACHEM, VAM, AOAC
Precisión	USP, ICH, EURACHEM, VAM, AOAC
Repetibilidad	ICH, ISO 17025, EURACHEM, AOAC
Precisión intermedia	ICH, EURACHEM, AOAC
Reproducibilidad	ICH, USP, ISO 17025, EURACHEM, AOAC
Veracidad	USP, ICH, ISO 17025, EURACHEM, VAM, AOAC
Linealidad	USP, ICH, ISO 17025, EURACHEM, VAM, AOAC
Rango	USP, ICH, VAM ,EURACHEM, AOAC
LD y LC	USP, ICH, VAM, ISO 17025, EURACHEM, AOAC

Tabla 3.2: Parámetros de validación que se piden en las diferentes normativas

Selectividad y especificidad son términos que se encuentran relacionados pero que no deben usarse intercambiablemente. La selectividad dictamina hasta qué grado un método analítico es capaz de determinar un mensurando en una muestra compleja a pesar de la presencia de otros componentes en la misma, mientras la especificidad es un término absoluto que puede verse como el nivel máximo de selectividad, es decir es un estado ideal y prácticamente inalcanzable (Vessman y col., 2001). Hablar de especificidad de un método analítico implica que sus resultados están influenciados única y exclusivamente por el valor del mensurando de interés, y aunque esto puede parecer muy atractivo, realmente no se puede tener la certeza de que se haya alcanzado.

A pesar de los avances modernos en química analítica instrumental, incluso las técnicas acopladas que proveen de resultados altamente selectivos son susceptibles de ser influenciadas por sustancias diferentes al analito de interés. No es posible asegurar que la respuesta que se obtiene tenga especificidad hasta que no se haya corroborado que el sistema es completamente selectivo frente a todos los potenciales interferentes existentes, lo cual es difícil de alcanzar debido al sinnúmero de sustancias que existen. Adicionalmente, hasta los epítomes de los métodos químicos instrumentales más selectivos incluyen la señal del ruido de fondo cuando se realiza un proceso de medición. Este defecto por lo general puede corregirse con facilidad, pero implica que el sistema responde a algo más aparte del analito de interés. En este orden de ideas, el término especificidad no es aplicable a los métodos analíticos por lo que su uso no se recomienda.

3.3.2 Interferentes químicos

Como se observó en las definiciones ya sea de especificidad o de selectividad, estos términos se refieren a la propiedad del método analítico de distinguir inequívocamente el analito de las demás sustancias. En este sentido, las "demás sustancias", dependiendo si afectan la respuesta o no del sistema de detección se pueden llamar interferentes. En general, estos interferentes ocasionan la presencia de errores sistemáticos, por lo cual se dice que la selectividad de un método es otro de los componentes de lo denominado como exactitud de un método (ver Sección 3.5).

Los interferentes en análisis químico se pueden clasificar en dos grupos. El primer grupo lo conforman los compuestos o elementos que ocasionan una respuesta similar a la del analito en el sistema de detección y, por lo tanto, se dice que estos siguen un mecanismo similar al del analito en el sistema de detección (interferentes tipo I, ver Figura 3.2). Por ejemplo, la coelución de un éster metílico de configuración *cis* en el tiempo de retención de su isómero *trans*, produce una respuesta idéntica en un detector de ionización de llama, por lo cual el isómero *cis* sería un interferente tipo I.

Por otro lado, los interferentes que no generan una respuesta en el sistema de detección, pero sí ocasionan que se aumente o disminuya la respuesta del analito de interés, se consideran interferentes tipo II. Por ejemplo, la presencia de fosfatos en la determinación de calcio por absorción atómica hace que la respuesta del calcio disminuya, pero el fosfato no genera ninguna señal en el detector.

Cuando un método analítico tiene presencia del primer tipo de interferentes se dice que el método tiene problemas de selectividad, mientras que, si el método tiene problemas de supresión o aumento de señal ocasionados por el segundo tipo de interferentes, se dice que el método tiene problemas de efecto matriz. La Figura 3.2 muestra un esquema que resume lo expuesto en esta sección.

Figura 3.2: Problemas ocasionados por los diferentes tipos de interferentes.

3.3.3 Evaluación de la selectividad

La evaluación de la selectividad de los métodos analíticos se realiza aplicando el procedimiento de medición repetidas veces a muestras que contengan los interferentes, pero que no contenga a los analitos (es decir, blancos de muestras[1]). Desde el punto de vista práctico, esto ocasiona un primer problema si se considera que para muchos casos no es posible contar con blancos de muestras. Por ejemplo, si se desea analizar sodio en suelos es prácticamente imposible conseguir un suelo que no contenga este elemento. En este sentido, como primera aproximación se emplean los blancos de reactivos, lo cual consiste en aplicar el procedimiento de medición sin la inclusión de la etapa de pesado o la toma de la alícuota de la muestra, lo cual desde un punto de vista químico tiene muchos inconvenientes y es una aproximación un tanto irrazonable.

Para el caso de sistemas analíticos que permiten realizar la medición de manera directa; por ejemplo, en la medición de algunos elementos en agua a través de técnicas como ICP-OES, ICP-MS o GFAA, o en el análisis de contaminantes orgánicos en aguas a través de LC-MS/MS, el concepto de blanco de reactivos muchas veces no aplica; y es así que para mediciones en agua es necesario analizar un agua purificada tipo I, II o III, dependiendo del analito de interés a evaluar, o algún agua natural que, por un análisis previo, no presente el analito que se desea cuantificar. De hecho, una de las guías de validación de métodos analíticos de la Unión Europea para el análisis de compuestos orgánicos en aguas establece que es necesario realizar la validación con mínimo un agua potable y un agua natural que no tengan los analitos de interés del método analítico.

Por otro lado, cuando se conocen los posibles interferentes, pero no es posible contar con el blanco de muestra, una opción muy empleada es el simular las matrices de análisis, es decir crear un “blanco artificial”, lo cual dependiendo de la rigurosidad del análisis y los posibles interferentes que pueden existir puede ser un proceso extremadamente largo y costoso, como es el caso de los métodos empleados para determinar algunos analitos en plasma sanguíneo.

Otra alternativa, en el caso de conocer los interferentes de los métodos y no contar con el blanco de muestra, es realizar la evaluación mediante la adición de estos interferentes a muestras (previamente cuantificadas) o MR de concentración conocida del analito, para posteriormente evaluar el sesgo (ver Sección 3.5.2) que los interferentes producen sobre el análisis de estos analitos (depende del mecanismo del interferente).

Por el contrario, cuando no se conocen los interferentes y no se cuenta con un blanco de muestra, la opción es realizar el análisis de muestras de concentración conocida mediante dos métodos, el método que se está validando y un método de referencia de exactitud conocida (ver Sección 3.5). Posterior a ello, se aplica una prueba de comparación como las presentadas en la Sección 2.2.5, y se concluye acerca de la selectividad del método.

[1] Blancos de muestras hacen referencia a una porción del material que contiene una composición química similar de una muestra real, pero el analito de interés no está presente.

3.3.4 Evaluación de la selectividad: aspectos prácticos

La selectividad de los métodos de medición se relaciona de manera directa con los interferentes de tipo químico, los cuales son elementos o compuestos que generan aumentos o disminuciones en las respuestas instrumentales y por lo tanto un error sistemático (interferentes tipo I y II). En el caso de conocer los interferentes más comunes al método se sugiere emplear el método de Danzer o evaluación del sesgo, y en el caso de desconocer dichos interferentes se recomienda el empleo de métodos de comparación. La Figura 3.3 muestra el esquema general para la evaluación de la selectividad.

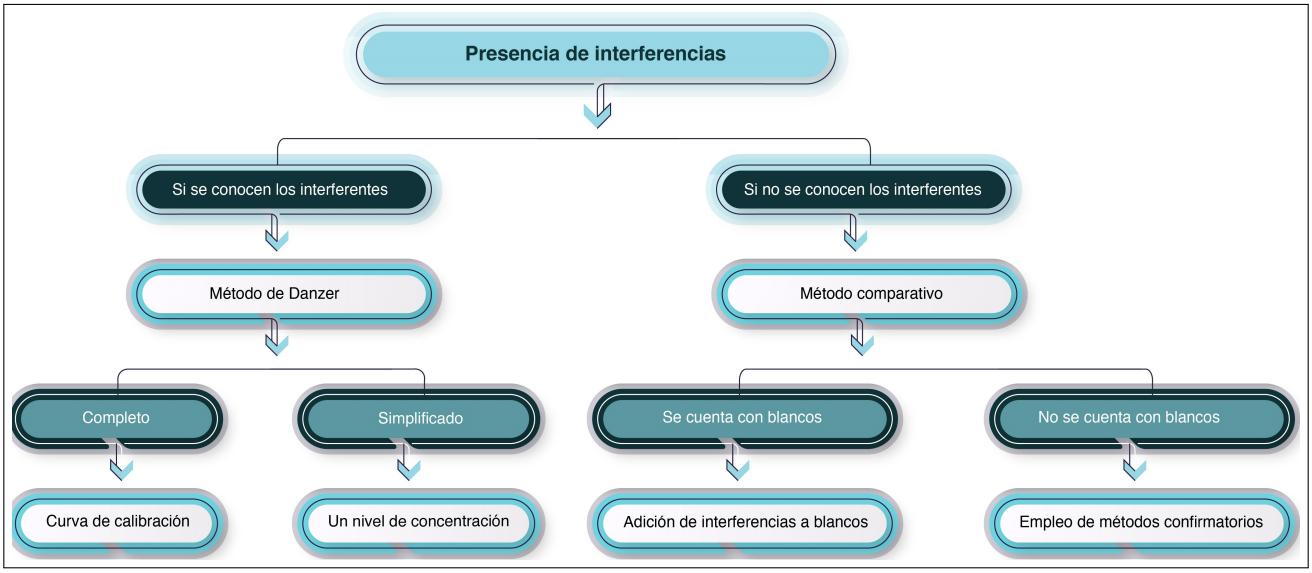


Figura 3.3: Esquema general para la evaluación de selectividad.

Para métodos de análisis químico multicomponente, la selectividad puede evaluarse utilizando una matriz de experimentos que considere cada componente con cada interferente. Ejemplos para evaluar la selectividad en este tipo de sistemas analíticos fue descrito por Otto y Wegscheider, 1986.

3.3.4.1 Método de Danzer

El método de Danzer se basa en la estimación del coeficiente de selectividad K_s, el cual viene dado por la siguiente ecuación:

$$K_s = \frac{K_I}{K_A} \tag{3.1}$$

donde K_I es el coeficiente de sensibilidad para el interferente (pendiente de la curva del interferente o respuesta promedio dividido sobre la concentración medida del interferente) y K_A es el coeficiente de sensibilidad del analito (pendiente de la curva del analito o respuesta promedio dividido sobre la concentración medida del analito). Se considera que un método es selectivo si $K_s \leq 0.3$.

La evaluación de la selectividad a través de este método se realiza de la siguiente manera:

a) Se prepara una curva de calibración del analito en solvente
b) Se prepara una curva de calibración del interferente en solvente.
c) Se miden las dos curvas de manera aleatoria, junto con un blanco de reactivos.
d) Se estima cada uno de los coeficientes de sensibilidad (pendiente), a través de un modelo de regresión lineal para cada una de las curvas.
e) Se estima el coeficiente Ks.
f) Compare contra el criterio establecido y concluya

Ejemplo 13: Selectividad: Método de Danzer completo.

Todos los ejemplos de este capítulo utilizan herramientas del módulo **Parámetros de validación** que se encuentra en la sección **Validación de métodos** del aplicativo validaR. Para los ejemplos de esta sección se utiliza el submódulo **Selectividad**.

Para la evaluación del parámetro de selectividad en la validación del método de medición de cadmio en extractos de cacao, se preparó una curva de calibración de 7 niveles en concentración entre 0 mg/L y 151 mg/L. Se sabe por regulación que esta validación tiene un requisito de sesgo del 2 %. Así mismo, este elemento presenta como interferente cobre, por lo cual se preparó una curva de calibración de 5 niveles entre 0 mg/L y 100 mg/L para este interferente. Se midió el blanco de reactivos el cual no presentó ninguna señal interferente. Las curvas se midieron de manera aleatoria en un equipo de absorción atómica de llama (FAAS). Los resultados se muestran en las siguientes tablas:

Curva de calibración del analito:

[Cd] / (mg/L)	Absorbancia
0	0.05
50	0.55
70	0.77
90	0.98
110	1.20
130	1.40
151	1.63

Curva de calibración del interferente:

[Cu] / (mg/L)	Absorbancia
0	0.06
10	0.04
20	0.07
50	0.15
100	0.31

Los resultados de la evaluación de la selectividad mediante el método de Danzer se muestran en el recuadro que aparece a continuación.

Resultados

Estimación del coeficiente de sensibilidad utilizando el enfoque de Danzer:

El coeficiente de selectividad es con el método de Danzer es 0.257

El método es selectivo.

Detalles de las curvas de calibración:

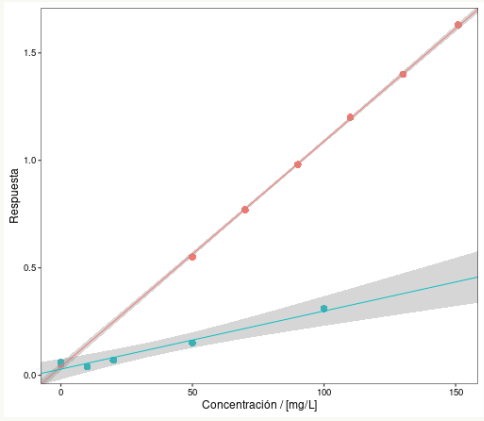
Analito:

```
                      Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)         0.03887731 7.250088e-03   5.362323 3.033593e-03
EstandarMatrix[, 1] 0.01049561 7.394036e-05 141.946913 3.291874e-10
```

Interferente:

```
                       Estimate   Std. Error  t value    Pr(>|t|)
(Intercept)         0.028711656 0.0150445452 1.908443 0.152359276
InterferMatrix[, 1] 0.002702454 0.0002950478 9.159377 0.002751338
```

Curvas de calibración del analito y el interferente



La línea roja corresponde a la curva de calibración del analito.
La línea azul corresponde a la curva de calibración del interferente.

Con los anteriores resultados se concluye que el método es selectivo para la medición de cadmio en presencia de cobre en la matriz de cacao por FAAS.

Por otro lado, se puede realizar una estimación simplificada de K_s, empleando únicamente un nivel de calibración, de la siguiente manera:

a) Se prepara el nivel más bajo de la curva de calibración del analito en solvente o en blancos de muestras.
b) Se prepara una solución del interferente a la concentración máxima en la cual este podría encontrarse en la muestra o en el extracto que se mide en el instrumento.

c) Se miden las dos soluciones por lo menos tres veces cada una (aleatorizando su medición) junto con un blanco de reactivos.
d) Se estima cada uno de los coeficientes (K_I y K_A) de sensibilidad dividiendo la respuesta promedio obtenida para cada solución por su respectiva concentración.
e) Se estima el coeficiente K_s.
f) Compare contra el criterio establecido y concluya.

Ejemplo 14: Selectividad: Método de Danzer simplificado.

En un laboratorio de análisis de aguas que utiliza la técnica ICP-OES se desea realizar la evaluación de la selectividad para níquel. Como punto de partida se conoce que el interferente para este elemento es cobalto, el cual suele encontrarse en muestras de aguas en una concentración de 100 μg/L y que el sesgo del método no debe ser mayor al 1 %.

Se preparó una disolución de níquel a 0.5 μg/L (el cual corresponde al nivel más bajo de su intervalo lineal) y una disolución de cobalto a 100 μg/L, ambas en presencia de germanio a una concentración de 1 μg/L. Posteriormente, se realizaron mediciones aleatorias de estas disoluciones por ICP-OES hasta obtener un total de 6 réplicas de medición para cada disolución. Se midió el blanco de reactivos el cual no presentó ninguna señal interferente. Los resultados se presentan en la tabla que aparece abajo a la izquierda. El recuadro que aparece abajo a la derecha contiene los resultados de validaR para la evaluación de selectividad utilizando los datos de la tabla.

Réplica	Señales relativas de emisión	
	Ni/Ge	Co/Ge
1	0.553	0.305
2	0.549	0.291
3	0.559	0.316
4	0.541	0.328
5	0.553	0.331
6	0.568	0.298

Resultados

Estimación del coeficiente de sensibilidad utilizando el enfoque de Danzer simplificado:
El coeficiente de selectividad es con el método de Danzer simplificado es 0.00281
El método es selectivo.

Con los anteriores resultados se concluye que el método es selectivo para la medición de cadmio en presencia de cobre en la matriz de cacao por FAAS.

3.3.4.2 Método comparación: cuando se cuenta con blancos de muestras

La evaluación de la selectividad a través de este método se basa en la comparación estadística de las respuestas obtenidas en diferentes escenarios. A continuación se explican las generalidades de este método.

a) Preparar el siguiente set de muestras (mínimo por triplicado):
   - Blanco de muestra: muestra que no tiene el analito.
   - Blanco de muestra fortificada: blanco de muestra al cual se le adicionan el (los) analito (s) de interés.

   Se sugiere que se realice la fortificación a la concentración más baja del intervalo del método (ver Sección 3.4).
b) Medir las muestras preparadas a través del método en evaluación.
c) Estimar los valores de concentración.
d) Determinar el porcentaje de error o alternativamente puede realizar una prueba estadística:

$$Error\ (\%) = \frac{|\ C_{bl.fortificado} - C_{bl.muestra}\ |}{C^{*}_{bl.fortificado}} \tag{3.2}$$

   donde $C^{*}_{bl.fortificado}$ es la concentración fortificada en el blanco de muestra fortificado, $C_{bl.fortificado}$ es la concentración de analito determinada en el blanco de muestra fortificado y $C_{bl.muestra}$ es la concentración de analito determinada en el blanco de muestra.
e) Comparar el porcentaje de error con el criterio establecido de selectividad.
   - Nota 1: debido a que la selectividad se relaciona con el error sistemático se puede emplear el criterio establecido para sesgo o porcentaje de recuperación (ver Sección 3.5.2).

- Nota 2: para métodos multianalito, específicamente los cromatográficos sin espectrometría de masas, se debe tener una resolución mayor a 1.5 entre la señal del analito de interés y las señales de otros componentes presentes en la muestra.

Ejemplo 15: Selectividad: Método comparativo cuando se cuenta con blancos.

En un laboratorio de análisis de contaminantes en sedimentos que utiliza ICP-MS, se debe realizar la evaluación de la selectividad para arsénico. Para ello el responsable de la validación cuenta con una muestra de sedimento proveniente de una mina de oro en la cual se cuantificó el arsénico y se encontró que su concentración era menor al límite de detección de la técnica ICP-MS (≤ 0.5 μg/kg). Esta muestra se dividió en dos porciones, para tomar una de ellas como blanco de muestra y la otra como blanco fortificado, para lo cual se humedeció con agua tipo I, luego se fortificó gravimétricamente con As en 0.1 mg/kg y finalmente, se secó a 70^{o}C durante dos horas.

Una vez preparadas las muestras, se realizó la extracción mediante la adición de 10 mL de HNO_3 concentrado sobre 0.5 g de muestra y se dejó el sistema en agitación durante 45 minutos. Posteriormente, los extractos filtrados se llevaron a 50 g con agua tipo I y se adicionó Rodio en 1 μg/kg como estándar interno. La medición de las muestras se realizó por quintuplicado. Los resultados se muestran en la tabla que aparece abajo a la izquierda. El recuadro que aparece abajo a la derecha contiene los resultados del aplicativo validaR para la evaluación de la selectividad mediante el método comparativo cuando se cuenta con blancos de muestra.

Réplica	Señal relativa de As/Rh		
	Bl. reactivos	Bl. muestra	Bl. fortificado
1	0.014	0.135	2.618
2	0.013	0.126	2.606
3	0.011	0.115	2.738
4	0.011	0.129	2.635
5	0.014	0.146	2.610

Resultados

Evaluación de selectividad utilizando la comparación de blancos:

La diferencia entre la media de las respuestas del blanco fortificado y del blanco de muestra es estadísticamente significativa.
Valor p prueba t de Student: < 1 e-04
El método es selectivo.

La diferencia entre la media de las respuestas del blanco de muestra y del blanco de reactivos es estadísticamente significativa.
Valor p prueba t de Student: < 1 e-04
Se recomienda utilizar blancos de muestra.

Se concluye que el método es selectivo para la medición de As en sedimentos en concentraciones mayores a 0.1 mg/kg. Con base en los resultados se recomienda usar el blanco de muestra para realizar las respectivas correcciones en las respuestas.

3.3.4.3 Método comparación: cuando no se cuenta con blancos de muestras

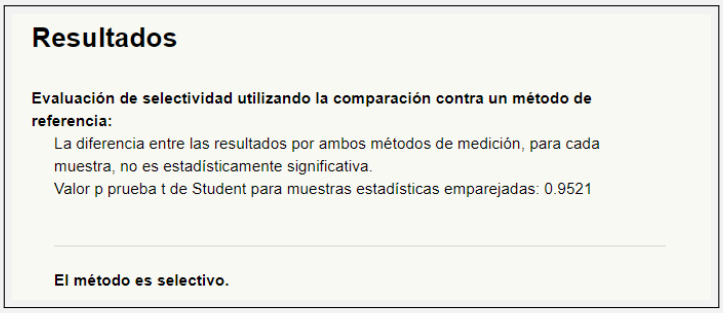
La evaluación de la selectividad a través de este método se basa en la comparación estadística de las respuestas con un método confirmatorio. A continuación se explican las generalidades de este método.

a) Preparar por lo menos tres muestras de diferentes características (si aplica), las cuales deben estar a una baja concentración (idealmente en el punto inferior del intervalo de trabajo del método).
b) Realizar la medición de las muestras mediante los dos métodos (el evaluado y el confirmatorio o de referencia). Por lo menos debe medir por triplicado las muestras.
c) Realizar un análisis estadístico en el que se comparen los resultados obtenidos por cada uno de los métodos. Se sugiere emplear una prueba t de comparación de medias de muestras emparejadas a un 95 % de nivel de confianza.
d) Se considera que el método a validar es selectivo, si no se encuentran diferencias significativas entre los resultados obtenidos por cada uno de los métodos en los diferentes tipos de muestras.

Ejemplo 16: Selectividad: Método comparativo cuando no se cuenta con blancos.

Un laboratorio de análisis de aguas presta el servicio de medición para hierro basado en la técnica FAAS, acorde con la norma SM 3111. Como parte de sus procesos de mejora en la calidad de sus servicios el laboratorio adquirió un equipo de ICP-OES. Con este nuevo equipo, el laboratorio ha desarrollado un método para la medición de hierro y otros elementos de forma simultánea. En el marco de esta nueva validación, el personal designado debe evaluar la selectividad, para lo cual realiza la preparación de muestras de agua de río, un material de referencia certificado (MRC) comercial de agua y los correspondientes blancos de reactivos. Las disoluciones obtenidas fueron medidas por triplicado en cada uno de los métodos que el laboratorio posee. Los resultados obtenidos se presentan en la tabla abajo a la izquierda.

Resultados ICP-OES	Resultados AAS
Blanco de reactivos	
0.02	0.11
0.08	0.04
0.06	0.14
Agua de río	
9.74	9.05
9.36	9.99
9.25	9.63
MRC de agua comercial	
6.62	6.85
6.89	6.28
6.71	6.56



En caso de que las diferencias de los valores medidos por ambos métodos no se distribuyan normalmente, estas se pueden analizar por medio de la prueba no paramétrica de Wilcoxon-Mann-Whitney para muestras emparejadas (King y Eckersley, 2019).

3.4 Intervalo de trabajo

El intervalo de trabajo corresponde al conjunto de concentraciones del analito en la muestra dentro de las cuales el método permite obtener valores de medición confiables o adecuados para el propósito; al tratarse de un intervalo, se encuentra compuesto por un límite inferior o límite de cuantificación (LC) y un límite superior o máxima concentración que puede ser cuantificada. La definición del intervalo de trabajo del método debe realizarse previo a planear los experimentos de precisión y sesgo; pues este intervalo define el alcance del método y a través de la evaluación de estos dos últimos parámetros se puede demostrar que el método produce resultados adecuados para el uso.

De manera general, el intervalo de trabajo del método debe cubrir:

- Los límites de regulación que apliquen.
- Las normas técnicas que apliquen.
- Las necesidades del cliente.
- Las concentraciones esperadas o la experiencia del laboratorio.

La Figura 3.4 presenta algunas recomendaciones para el establecimiento de estos límites. Es importante señalar que en el caso en que exista una regulación se debe emplear el criterio que indique esta (cuando aplique) o establecer los límites sugeridos de la Figura 3.4.

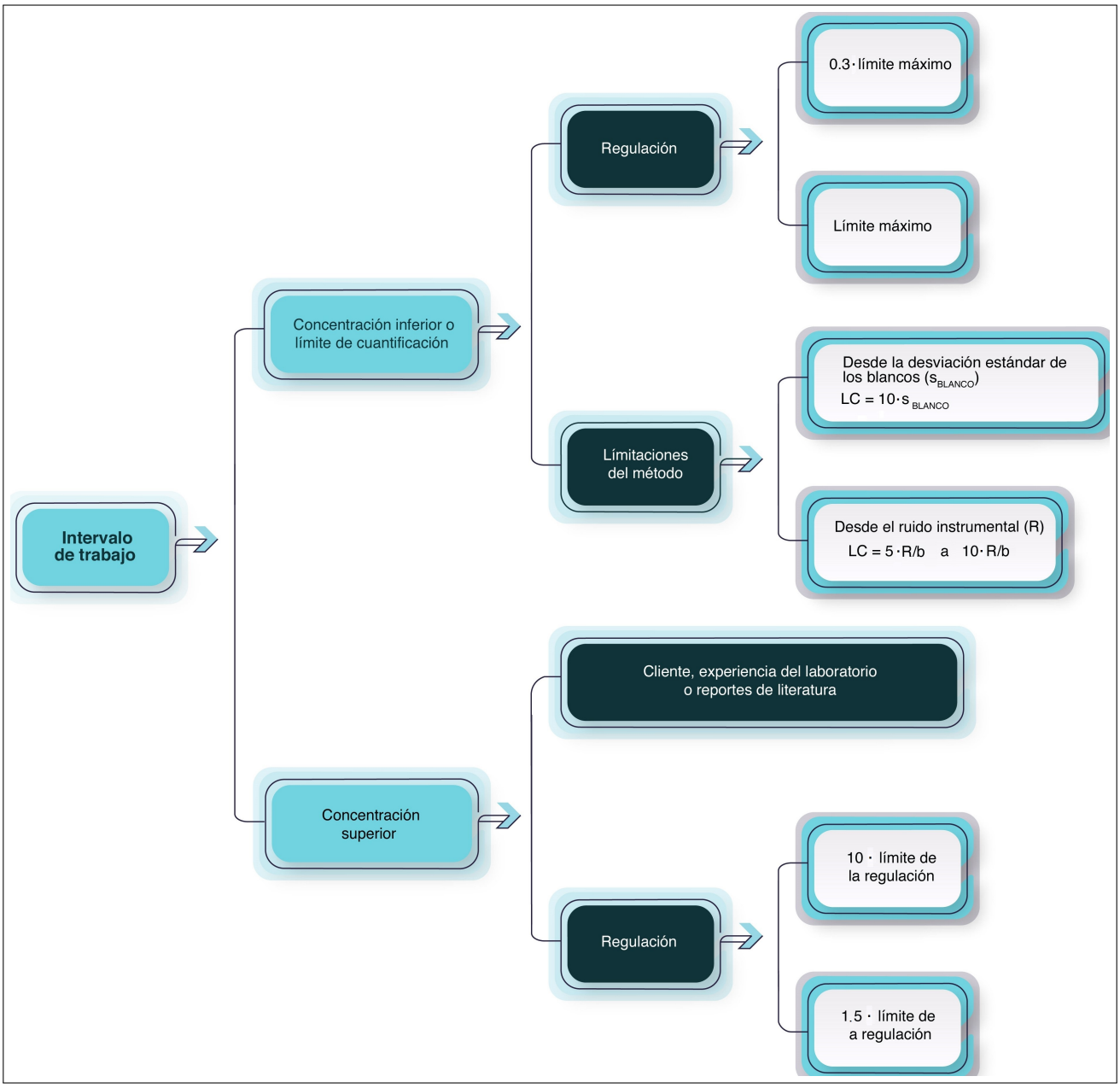


Figura 3.4: Esquema general para el establecimiento del intervalo de trabajo.

A continuación se dan algunas recomendaciones para la selección del límite inferior o límite de cuantificación:

- La estimación del límite de cuantificación basada en la estimación de la desviación estándar del blanco (s_{blanco}), se puede realizar a través de cualquiera de los métodos empleados para el límite de detección, pero en lugar de multiplicar por 3 se debe multiplicar por un factor entre 5 y 10, lo cual depende de la relación señal/ruido.
- El límite de cuantificación se debe reportar en unidades de la muestra, por ejemplo mg/kg de maíz, μg/kg de agua, ng/L de agua, entre otros.
- Se puede emplear la concentración equivalente al nivel más bajo de la curva de calibración, siempre y cuando el sesgo no sea significativo.
  - Nota 1: la concentración equivalente implica que se deben considerar los factores de dilución que apliquen.
  - Nota 2: en el caso en que se tengan porcentajes de recuperación inferiores al 100 % y se debe considerar este valor en el cálculo de la concentración equivalente. Por ejemplo, si el nivel más bajo de concentración de la curva de calibración es 10 μg/mL, el factor de dilución del método es 7 y el porcentaje de recuperación es

90 %, el límite de cuantificación (estimado) correspondería a:

$$LC = \frac{(10\ [\mu\text{g/mL}] \cdot 7\ [\text{mL/mL}])}{90\,\%} \cdot 100\,\% = 77.78\ [\mu\text{g/mL}] \tag{3.3}$$

Una vez establecidas las concentraciones mínima y máxima en unidades de la muestra, se procede a demostrar que el método funciona adecuadamente a través de los estudios de precisión y veracidad (ver Sección 3.5). En el caso en que se cumpla con los criterios establecidos para precisión y veracidad se puede concluir que el intervalo del método es el seleccionado; en el caso en que no se cumpla para alguno de los criterios se deberá redefinir el intervalo y repetir los experimentos de veracidad y precisión.

Ejemplo 17: Selección del intervalo de trabajo.

Un laboratorio de referencia experto en la medición de residuos de plaguicidas desea implementar una metodología por cromatografía de gases acoplada a espectrometría de masas (GC-MS) para la determinación de endosulfán en granos de café. Para establecer el intervalo de trabajo en su validación, se hace uso de la regulación, en la cual se tiene que el límite máximo de residuos (LMR) para esa molécula es de 0.2 mg/kg. De acuerdo con los lineamientos de la Figura 3.7, el límite inferior del intervalo de trabajo se tomará de la siguiente manera:

$$\text{Límite inferior} = 0.3 \cdot LMR \tag{3.4}$$

Para este caso se aplica esta ecuación dado que:

- Se conoce que la recuperación en la extracción y determinación de este plaguicida en el café es del 90 %
- Es necesario que el método de medición garantice la cuantificación del plaguicida por debajo del LMR, debido al potencial impacto que tienen los residuos de esta molécula y al riesgo en la población; por lo cual se recomienda validar desde un nivel de concentración menor al LMR. Así, el límite inferior de la validación corresponde a 0.06 mg/kg de endosulfán en granos de café.

De otra parte, el límite superior se establecerá con base en la siguiente ecuación:

$$\text{Límite superior} = 1.5 \cdot LMR \tag{3.5}$$

Al aplicar esta ecuación se tiene que el límite superior del intervalo de trabajo corresponde a 0.3 mg/kg de endosulfán en granos de café.
Respecto al límite superior, se considera suficiente este criterio dado que:

- Un lote de granos de café que contengan endosulfán a partir de 0.2 mg/kg dejará de tener valor comercial en el mercado internacional al representar un riesgo para los usuarios.
- Concentraciones mayores implica más costos asociados a tiempo de personal, uso de materiales de referencia, tiempos de medición, entre otros, que no representan valor agregado para el laboratorio.

Conclusión: el intervalo de trabajo del método para la validación de endosulfán en granos de café corresponde al intervalo comprendido entre 0.06 mg/kg y 0.3 mg/kg.

3.5 Exactitud: precisión y veracidad

La exactitud se define según el Vocabulario Internacional de Metrología como la proximidad entre un valor medido y un valor verdadero de un mensurando (JCGM, 2012); sin embargo, el valor verdadero del mensurando no se puede conocer, por lo cual la exactitud se convierte en un estado ideal que en la práctica no es posible evidenciar, por lo menos numéricamente. Por otro lado, la ISO 3534 define la exactitud como la proximidad de concordancia entre el resultado de una medición y el valor de referencia aceptado, y cuando este término es aplicado a una serie de resultados de prueba, involucra una combinación de componentes aleatorios y una componente de error sistemático o sesgo (Norma ISO 3534-1, 2006).

En este contexto, la exactitud se evalúa por medio de la estimación de los efectos sistemáticos y aleatorios sobre los resultados obtenidos de una serie de mediciones, los cuales se representan por los parámetros de desempeño conocidos como veracidad y precisión, respectivamente. La Figura 3.5 presenta un esquema que muestra la relación entre la exactitud, la incertidumbre, los tipos de errores y los parámetros del método.

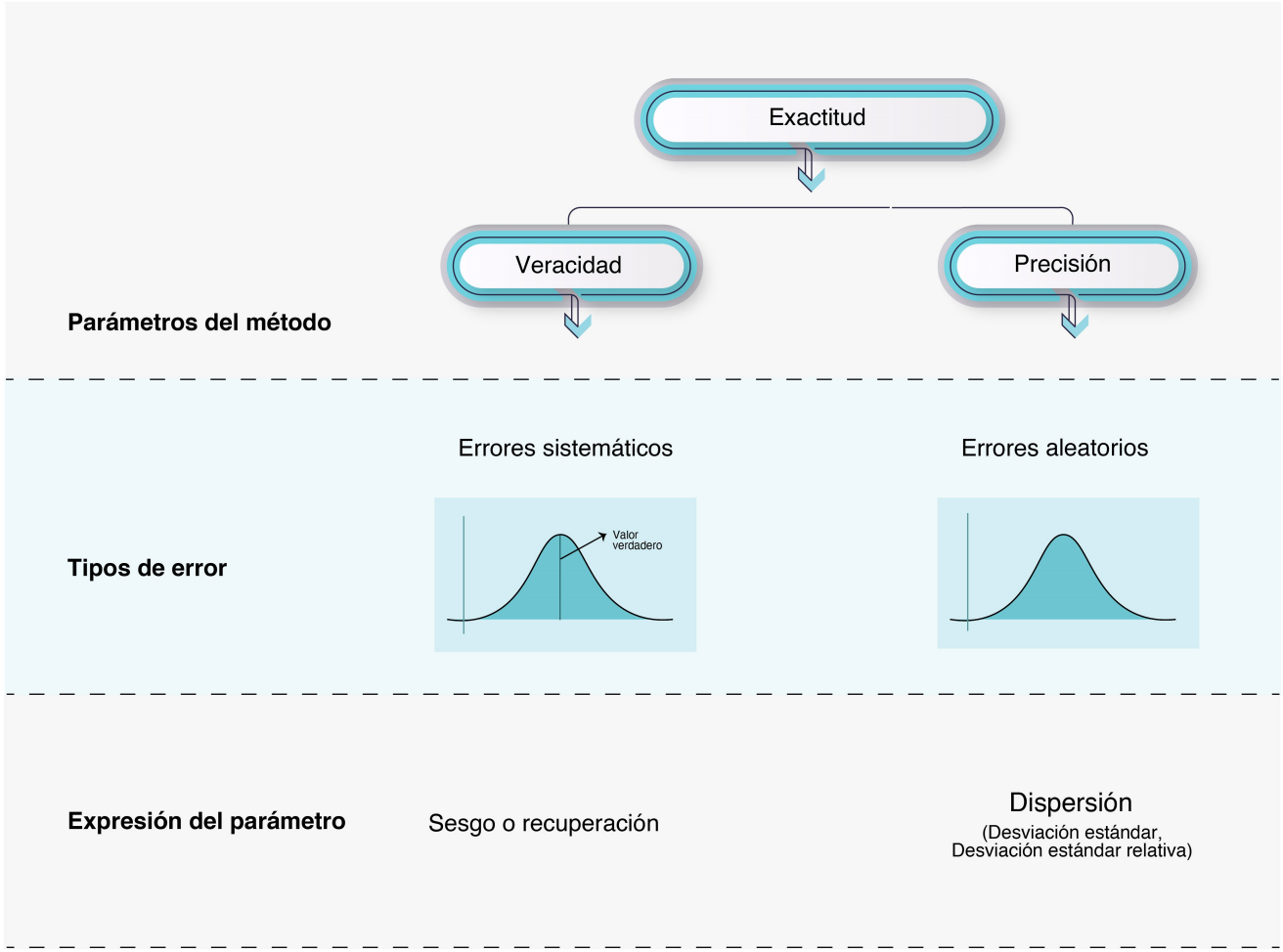


Figura 3.5: Relación de la exactitud y el error de medición.

La evaluación de la exactitud debe cubrir todo el intervalo de trabajo especificado, es decir durante su evaluación debe considerar por lo menos dos niveles de concentración, los cuales deben concordar con los límites superior e inferior del intervalo definido. Las siguientes secciones, presentarán los conceptos y aspectos prácticos más relevantes para la evaluación de cada uno de los parámetros relacionados con la exactitud.

3.5.1 Precisión

La precisión es definida por el Vocabulario Internacional de Metrología como la proximidad entre las indicaciones o los valores medidos obtenidos en mediciones repetidas de un mismo objeto, o de objetos similares, bajo condiciones específicas (JCGM, 2012). Por su parte, la norma ISO 5725-1 y la guía EURACHEM definen la precisión como el grado de concordancia entre los resultados de pruebas, independientes obtenidas al aplicar un procedimiento experimental bajo unas condiciones establecidas (Magnusson y Örnemark, 2014; Norma ISO 5725-1, 1994).

De acuerdo con estas definiciones es importante resaltar lo siguiente:

- La precisión depende sólo de la distribución de los errores aleatorios y no es indispensable conocer el valor verdadero o contar con un valor de referencia de la muestra o del ítem bajo prueba. Lo anterior es considerando que los factores que aportan a la variación de los resultados de las mediciones son: el analista, las condiciones ambientales, el instrumento de medición, el tiempo transcurrido entre mediciones, la concentración, entre otras.
- Al tratarse de una evaluación del error aleatorio, el número de repeticiones debe ser tal que garantice que la muestra represente a la población; Magnusson y Örnemark, 2014 recomienda un número de 10 repeticiones como mínimo para cada concentración. Sin embargo, en otros casos se recomienda que el estudio de exactitud

tenga en cuenta cual es la máxima variación que se puede aceptar en el método que se está validando y es posible que se necesite un mayor número de repeticiones.

- La precisión se expresa como una medida de dispersión, usualmente como desviación estándar o desviación estándar relativa.
- La precisión depende de la concentración del analito y por lo tanto se debe evaluar en todo el intervalo de trabajo especificado. Es decir, se debe demostrar que el método funciona de manera adecuada, como mínimo dos concentraciones, las cuales corresponderán a los extremos del intervalo.
- La precisión se evalúa bajo condiciones específicas, principalmente, repetibilidad y reproducibilidad. Por lo anterior, se puede señalar que existen diferentes tipos de precisión, las cuales se detallan a continuación.
  - La precisión en condiciones de repetibilidad hace referencia a la variación del método en condiciones donde los resultados de medición son obtenidos sobre muestras de ensayo idénticas (alícuotas de una muestra suficientemente homogénea), en el mismo laboratorio, por el mismo operador, utilizando el mismo equipo dentro de intervalos de tiempo cortos (Norma ISO 3534-1, 2006).
  - La precisión como reproducibilidad corresponde a la máxima variación del método en condiciones en las que los resultados de medición se obtienen sobre objetos de prueba idénticos, en diferentes laboratorios, por diferentes operadores, usando diferentes equipos, durante un período en el que se hubieran dado diferencias entre los materiales y equipos, lo que implica la máxima dispersión de un mismo proceso analítico (Norma ISO 5725-1, 1994).
  - La precisión en condiciones de reproducibilidad indica el grado de dispersión entre los resultados de distintos laboratorios, por ejemplo, en el marco de un ensayo colaborativo. De esta manera se puede estimar la máxima variabilidad que presentaría el método de medición cuando cambian todas las condiciones experimentales durante su ejecución. Idealmente la validación del método analítico debería considerar la evaluación de la precisión como reproducibilidad, sin embargo, en muchos casos la evaluación de este parámetro en condiciones de reproducibilidad es difícil porque muchos de los laboratorios desarrollan sus propios métodos y coordinar un ensayo colaborativo o un estudio interlaboratorio implica una gran inversión de recursos y tiempo, por lo que obtener los resultados de precisión en condiciones de reproducibilidad no es una práctica común en la validación. En este contexto, es común no realizar la evaluación de la precisión en condiciones de reproducibilidad, sino que se realiza en condiciones de precisión intermedia.
  - La precisión en condiciones intermedias evalúa la dispersión de resultados entre ensayos mutuamente independientes utilizando el mismo método aplicado a la misma muestra, en el mismo laboratorio bajo diferentes condiciones: distintos operadores, diferente equipamiento o distintos días (Norma ISO 5725-1, 1994). Este parámetro hace referencia al grado de coincidencia o de variación de los resultados dentro del mismo laboratorio, pero en días distintos con analistas y equipos diferentes (según corresponda)(ICH-USP; ICH Q2A, CPMP/ICH/381/95).

La precisión se expresa en términos de dispersión, generalmente por desviación estándar, desviación estándar relativa y depende de las condiciones en las que se ejecute el estudio:

- Repetibilidad (s_r)
- Intermedia (s_i)
- Reproducibilidad (s_R)

La siguiente tabla presenta los casos en los que se evalúa la precisión en las diferentes condiciones:

Condiciones	Casos en los que aplica
Repetibilidad	• En todos los métodos de tipo cuantitativo.
Intermedias	• Seguimiento a materiales primas o producto final en periodos largos de tiempo (más de tres meses). • Alta variación instrumental o del método en el tiempo, identificada por experiencia previa. • Requisito de regulación/norma. • Cuando se espera que el método se emplee de manera frecuente y por largos periodos de tiempo. • Por acuerdo con el cliente. • Requisito de regulación/norma • Cuando se tiene un método de múltiples pasos y laborioso. • Cuando el método tiene etapas que dependen del analista. • Cuando hay alta rotación de personal. • Cuando hace parte de un ejercicio de evaluación del personal. • Por acuerdo con el cliente.

Tabla 3.3: Criterios para evaluar la precisión en las diferentes condiciones

3.5.1.1 Selección del criterio de aceptación

La máxima variación aceptable para el método, expresada como porcentaje de coeficiente de variación (% CV), puede ser establecida a través de diferentes estrategias:

- Empleo de la regulación o criterios aceptados por la comunidad científica.
- Empleo de la ecuación de Horwitz.
- Empleo de pruebas estadísticas.
- Experimentos o estudios previos a la validación.

3.5.1.1.1 Empleo de la regulación o guías:

En el caso en él que una determinada regulación o norma establezca una precisión esperada, siempre debe emplearse esta como criterio. Por otro lado, en ocasiones las regulaciones o normas establecen o citan algunos criterios de aceptación que son internacionalmente aceptados, por ejemplo: IUPAC, EURACHEM, AOAC, AEFI, entre otros. La Tabla 3.4 y la Tabla 3.5 presentan algunos de estos criterios de aceptación.

Nótese que para el caso de la AEFI (Tabla 3.4) se presenta un coeficiente de variación en función del intervalo de recuperación o de error sistemático, por lo cual es indispensable que se revise previamente si el porcentaje de recuperación obtenido se encuentra dentro del intervalo esperado (ver Sección 3.5.2). Una vez se encuentre que la veracidad o recuperación son adecuadas se puede proceder a seleccionar el % CV de aceptación que sugiere la AEFI.

Por otro lado, como se puede observar en la Tabla 3.5, la AOAC establece un criterio de aceptación de la precisión en función de la concentración en la cual se evalúo. Lo anterior implica que dependiendo de la amplitud del intervalo se pueden tener 2 o más criterios de aceptación; por ejemplo, si el método se evalúa desde 1 mg/kg hasta 100 mg/kg, los criterios de aceptación serían 11 % y 5.3 % respectivamente.

Intervalo de recuperación (%)	CV aceptable (%)			
	n = 2	n = 3	n = 4	n = 5
99.0 - 101.0	0.55	0.67	0.78	0.87
98.5 - 101.5	0.82	1.01	1.19	1.30
98.0 - 102.0	1.10	1.34	1.55	1.73
95.0 - 105.5	2.74	3.36	3.88	4.33
90.0 - 110.0	5.48	6.71	7.75	8.67
85.0 - 115.0	8.22	10.07	11.63	13.0

Tabla 3.4: Coeficientes de variación aceptables (porcentaje) en función de la recuperación del método y el número de determinaciones (AEFI, 2001).

Analito	Fracción másica	Unidad	CV aceptable (%)
100	1	100 %	1.3
10	10^{-1}	10 %	1.9
1	10^{-2}	1 %	2.7
0.1	10^{-3}	0.1 %	3.7
0.01	10^{-4}	100 ppm (mg/kg)	5.3
0.001	10^{-5}	10 ppm (mg/kg)	7.3
0.0001	10^{-6}	1 ppm (mg/kg)	11.0
0.000 01	10^{-7}	100 ppb (μg/kg)	15.0
0.000 001	10^{-8}	10 ppb (μg/kg)	21.0
0.000 000 1	10^{-9}	1 ppb (μg/kg)	30.0

Tabla 3.5: Coeficientes de variación aceptables (porcentaje) en función de la concentración del analito (AOAC, 2016).

3.5.1.1.2 Empleo de la ecuación de Horwitz

La ecuación o relación de Horwitz, es un índice que propuso el químico analítico William Horwitz, quién a partir del estudio del comportamiento de diferentes interlaboratorios propuso una ecuación que permite establecer el coeficiente de variación que se puede obtener en los experimentos de precisión en función de la fracción másica del analito (C). La ecuación presenta dos formas comúnmente conocidas:

$$\% \, CV_R = 2 \cdot C^{-0.15} \tag{3.6}$$

$$\% \, CV_R = 2^{1-0.5 \cdot \log C} \tag{3.7}$$

donde $\% \, CV_R$ corresponde al máximo coeficiente de variación que se puede obtener en condiciones de reproducibilidad y C corresponde a la concentración expresada en fracción másica.

Posteriormente, Thompson transformó la ecuación en algunas expresiones más compactas que permiten la estimación del máximo $\% \, CV$ en condiciones de repetibilidad ($\% \, CV_r$) o de precisión intermedia ($\% \, CV_i$):

$$\% \, CV_r = C^{-0.15} \tag{3.8}$$

$$\% \, CV_i = 1.5 \cdot C^{-0.15} \tag{3.9}$$

La ecuación de Horwitz ha sido validada y aceptada por diferentes campos como el farmacéutico, alimentos, ambiente, entre otros. Sin embargo, dicha ecuación tiene algunas limitantes que se listan a continuación:

- La concentración del analito se debe expresar en las mismas unidades, por ejemplo, mg/mg o kg/kg. Para el caso de matrices acuosas, usualmente se puede asumir que 1 mL tiene una masa aproximada de 1 g.
- Es necesario realizar el cálculo del $\% \, CV$ a través de las dos ecuaciones, con el propósito de validar que los cálculos se encuentran bien realizados.
- La ecuación no aplica para mensurandos operacionalmente definidos, tales como: humedad, fibra, grasa total, entre otros.
- La ecuación no aplica para analitos como enzimas, polímeros o biomoléculas.
- La ecuación no aplica para mensurandos de tipo físico como color, densidad, viscosidad, entre otros.
- La ecuación puede tener algunas desviaciones a concentraciones altas y bajas, por lo cual se debe emplear con precaución, de hecho, existe una versión contemporánea de la ecuación que establece diferentes alternativas dependiendo de la concentración. La Tabla 3.6 presenta dichas alternativas.

Fracción másica analito C	Expresión para desviación estándar aceptable s_R
$C < 1.2 \cdot 10^{-7}$	$s_R = 0.22 \cdot C$
$1.2 \cdot 10^{-7} < C < 0.138$	$s_R = 0.02 \cdot C^{-0.8495}$
$C < 1.2 \cdot 10^{-7}$	$s_R = 0.01 \cdot C^{0.5}$

Tabla 3.6: Alternativas contemporáneas a la ecuación de Horwitz.

Ejemplo 18: Uso de la ecuación de Horwitz.

En la planificación de la validación de un método para la determinación de cobre en agua potable por GF-AAS se debe establecer el criterio de máxima variación aceptable que se espera obtener en esta medición para evaluar la precisión bajo condiciones de repetibilidad ($\%\ CV_r$) y precisión intermedia ($\%\ CV_i$).

El intervalo de trabajo del método se encuentra entre 5 μg/L y 1500 μg/L, dentro del cual se seleccionaron cuatro niveles de concentración (8.00 μg/L, 85.0 μg/L, 500 μg/L y 1000 μg/L) para evaluar la precisión. De esta manera, la máxima variación aceptable obtenida usando la ecuación de Horwitz modificada por Thompson corresponde a:

Nivel de concentración /[μg/L]	Fracción másica equivalente /[g/g]	% CV_r	% CV_i
1000	1.0×10^{-6}	7.94	11.91
500	5.0×10^{-7}	8.81	13.22
85	8.5×10^{-8}	11.50	17.25
8	8.0×10^{-9}	16.39	24.58

3.5.1.1.3 Empleo de pruebas estadísticas

Algunas regulaciones o normas establecen el máximo coeficiente de variación que se puede obtener en un experimento de precisión, por lo cual es suficiente para la evaluación de la precisión del método que se demuestre que el coeficiente de variación experimental es inferior a este valor. Por otro lado, en el caso en que no se cuente con un coeficiente de variación máximo; por ejemplo, con la ecuación Horwitz, se debe proceder a aplicar una prueba estadística que permita evaluar si existen diferencias significativas con el valor esperado.

De manera general, la evaluación estadística de la precisión obtenida se realiza mediante una prueba chi-cuadrado (χ^2), según se explica en la Sección 2.2.6.1, donde la hipótesis nula establece que la varianza experimental (σ^2) es igual a la máxima precisión aceptable (σ_0^2).

Por otro lado, como se mencionó previamente la evaluación de la precisión se debe realizar en diferentes concentraciones, por lo menos en los dos valores extremos del intervalo de trabajo del método. Por lo anterior, es indispensable evaluar si la concentración de las muestras influye sobre la precisión del método. Para esto, se deben aplicar pruebas estadísticas que evalúen la heterocedasticidad de los resultados a las diferentes concentraciones. La Tabla 2.2 del Capítulo 2 sugiere algunas de estas pruebas.

Una vez realizada la prueba se tienen dos escenarios (i) la concentración no influye de manera significativa sobre la precisión del método o (ii) la concentración influye de manera significativa sobre la precisión del método. Para el primer caso se puede emplear cualquiera de las precisiones obtenidas en las diferentes concentraciones para estimar la incertidumbre de medición; y para el segundo caso se debe seleccionar la más alta dispersión para estimar la incertidumbre de medición.

Ejemplo 19: Precisión: Uso de la prueba estadística χ^2.

En un laboratorio de análisis de alimentos está realizando la validación de un método cromatográfico (HPLC-Fluorescencia) para la determinación de especies de arsénico en arroz.

En la evaluación del parámetro de precisión el personal realizó la extracción de la especie MMA (monometil arsénico) tomando siete porciones de 1 g de tres muestras de arroz que poseen MMA a diferentes concentraciones: 1.0 mg/kg, 5.0 mg/kg y 10 mg/kg. A las muestras se les adicionaron 10 mL de una mezcla de ácido nítrico (HNO_3) al 1 % y peróxido de hidrógeno (H_2O_2) al 2 %. Se dejaron en agitación durante 1 hora a 95 ^oC. Posteriormente, los extractos fueron centrifugados y las disoluciones sobrenadantes fueron inyectadas aleatoriamente en el sistema cromatográfico. En la cuantificación se obtuvieron los resultados que se muestran en la tabla a la derecha.

Réplica No.	**Concentración (mg/kg)**		
	Nivel bajo	**Nivel medio**	**Nivel alto**
1	1.34	4.94	11.55
2	1.51	5.11	9.51
3	1.30	5.22	10.41
4	1.69	4.37	9.64
5	1.42	4.89	8.03
6	1.41	4.95	9.39
7	1.43	4.87	11.26
CV experimental	8.87 %	5.47 %	12.13 %
CV Máximo (Horwitz)	8.0 %	6.28 %	5.66 %

Los resultados del aplicativo validaR para la evaluación de la precisión del método se muestran en el recuadro que aparece a continuación:

Resultados

Resultados de precisión en condiciones de Repetibilidad :
Los criterios de precisión se cumple para los niveles de concentración No. 1, No. 2
Las siguientes secciones detallan los resultados de las comparaciones para cada punto.

Comparación contra los criterios de referencia:

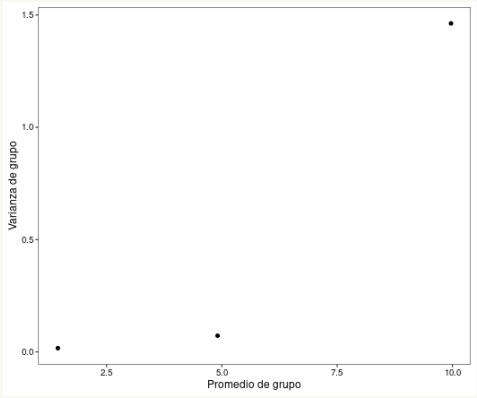
Nivel No. 1:
- Valor de concentración: 1 [mg/kg]
- RSD experimental: 8.87 %
- RSD máxima: 8 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 0.287 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados no es significativamente mayor al valor de RSD máxima. **Los resultados cumplen con el criterio de precisión.**

Nivel No. 2:
- Valor de concentración: 5 [mg/kg]
- RSD experimental: 5.47 %
- RSD máxima: 6.28 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 0.6024 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados no es significativamente mayor al valor de RSD máxima. **Los resultados cumplen con el criterio de precisión.**

Nivel No. 3:
- Valor de concentración: 10 [mg/kg]
- RSD experimental: 12.13 %
- RSD máxima: 5.66 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 1e-04 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados es significativamente mayor al valor de RSD máxima. **Los resultados no cumplen con el criterio de precisión.**

Prueba de homocedasticidad entre las series.
Se realizó la prueba de Levene para evaluar la homocedasticidad del método
La precisión del método NO es constante en el intervalo de concentraciones considerado.
Valor p de la prueba de Levene: 0.0011

Gráfico de homocedasticidad



El esquema general de la evaluación de la precisión se presenta a continuación:

a) Seleccionar las concentraciones en las cuales se realizará la evaluación.
   - Nota 1: la concentración máxima y mínima pueden ser sugeridas por la regulación o las recomendadas en la Sección 3.4 Intervalo de trabajo.
   - Nota 2: se sugiere seleccionar por lo menos tres concentraciones para la evaluación de este componente, las cuales pueden ser: límite inferior o de cuantificación, límite dado por la regulación o concentración esperada, y límite superior del intervalo de trabajo del método.

b) Seleccionar o preparar los materiales con los cuales se realizará la evaluación, estos pueden tratarse de materiales de referencia, muestras naturales (previamente medidas) o materiales fortificados.
   - Nota 1: para la evaluación de la precisión no es necesario conocer el valor exacto del analito, pero si se requiere realizar una cuantificación adecuada para demostrar que el método es lo suficientemente precisión en un intervalo de concentración.
   - Nota 2: el estudio de veracidad y precisión normalmente se realiza al mismo tiempo, es decir los datos que se compilen del estudio de veracidad normalmente sirven para demostrar la precisión.

c) Analizar el conjunto de muestras con el método que se está validando.
   - Nota 1: se sugiere que esto se realicé con un diseño completamente al azar y en el caso de ser posible se realicen por lo menos 7 réplicas independientes de medición.
   - Nota 2: si se trabaja con materiales de referencia se debe considerar todas las instrucciones que se dan en el certificado.

d) Estimar la concentración para cada una de las muestras o niveles de concentración.
   - Nota: en el caso en que conozca el valor esperado o de referencia, puede calcular el porcentaje de recuperación.

e) Estimar el % CV para cada concentración evaluada.

f) Comparar el % CV con el máximo coeficiente de variación esperado (ver Sección 3.5.1.1) o realizar el análisis estadístico de acuerdo con la Sección 3.5.1.1.3.

En el caso en que se requiera realizar la evaluación de la precisión en condiciones intermedias, la Tabla 3.7 presenta algunas recomendaciones en función de los resultados de homocedasticidad de los resultados obtenidos en el experimento de precisión bajo condiciones de repetibilidad.

Recomendación para ejecutar la prueba	**Recomendación para analizar la información**
Cuando se observó homocedasticidad entre las diferentes concentraciones	
Se puede realizar la prueba entre días o entre analistas a una sola concentración	Emplear la prueba de Levene para evaluar si hay influencia de los analistas o el tiempo. Si hay diferencias entre las varianzas, emplear un ANOVA para estimar la desviación estándar asociada a la precisión e incluir este componente en la incertidumbre de medición. En el caso en que no existan diferencias se recomienda emplear la máxima variación que se obtuvo de los analistas o del tiempo o emplear la Ecuación 2.4 de la Sección 2.2.5.2.
Cuando se observó heterocedasticidad entre las diferentes concentraciones	
Se deben emplear como mínimo dos concentraciones (baja y alta concentración) para los estudios de precisión intermedia (internalista o Inter días)	Emplear la prueba de Levene para evaluar si hay influencia de los analistas o el tiempo. Si hay diferencias entre las varianzas, emplear un ANOVA para estimar la desviación estándar asociada a la precisión e incluir este componente en la incertidumbre de medición En el caso en que no existan diferencias se recomienda emplear la máxima variación que se obtuvo a una de las concentraciones evaluadas o emplear la Ecuación 2.4 de la Sección 2.2.5.2.

Tabla 3.7: Recomendaciones del número de concentraciones y análisis estadístico a realizar en el estudio de precisión intermedia.

Ejemplo 20: Precisión: Evaluación de la precisión de un método analítico.

Un laboratorio de referencia nacional especializado en el análisis de alimentos está realizando la validación de un método por ICP-MS para la determinación de oligoelementos en harina de trigo. En la evaluación del parámetro de precisión en condiciones de repetibilidad se planteó aplicar un diseño completamente al azar.

El personal tomó siete porciones de 500 mg en tres muestras de harina de trigo que poseen hierro en tres concentraciones diferentes: 0.9 mg/kg, 5.0 mg/kg y 22 mg/kg. A las muestras se les adicionaron 4 mL de ácido nítrico concentrado, 2 mL de peróxido de hidrógeno al 30 % y 2 mL de una disolución de rhodio de 10 μg/kg. La mezcla se dejó en reposo durante 5 minutos y se llevó a un sistema de digestión asistida por microondas. Posteriormente, los extractos fueron llevados a 50 g con agua tipo I. Finalmente, las muestras fueron medidas aleatoriamente en el ICP-MS y la cuantificación fue realizada por el método de adición patrón. Los resultados se presentan en la tabla que se muestra a la derecha.

Réplica No.	**Fracción másica de hierro / (mg/kg)**		
	Harina A	**Harina B**	**Harina C**
1	0.825	5.23	21.60
2	0.912	4.93	20.84
3	0.950	5.21	21.35
4	0.816	4.51	20.79
5	0.832	4.84	21.57
6	0.904	5.52	19.22
7	0.980	5.33	20.53
CV experimental	7.33 %	6.72 %	3.96 %
CV Máximo (Horwitz)	8.1 %	6.3 %	5.0 %

Los resultados del aplicativo validaR para la evaluación de la precisión del método haciendo comparación directa contra el criterio de Horwitz se muestran en el siguiente recuadro. Adicionalmente se presenta el gráfico de homocedasticidad del método. Para esta serie de experimentos los resultados presentan heterocedasticidad:

Resultados

Resultados de precisión en condiciones de Repetibilidad :
Los criterios de precisión se cumple para los niveles de concentración No. 1, No. 2, No. 3
Las siguientes secciones detallan los resultados de las comparaciones para cada punto.

Comparación contra los criterios de referencia:

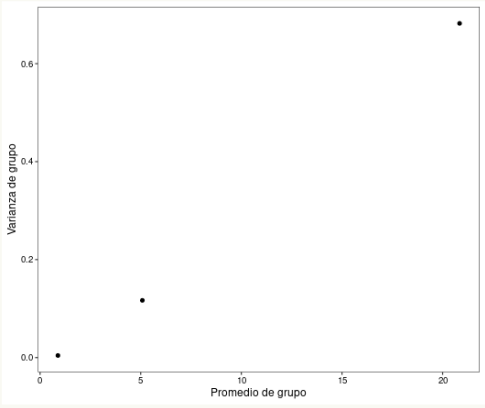
Nivel No. 1:
- Valor de concentración: 0.9 [mg/kg]
- RSD experimental: 7.33 %
- RSD máxima: 8.13 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 0.5597 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados no es significativamente mayor al valor de RSD máxima. **Los resultados cumplen con el criterio de precisión.**

Nivel No. 2:
- Valor de concentración: 5 [mg/kg]
- RSD experimental: 6.72 %
- RSD máxima: 6.28 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 0.332 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados no es significativamente mayor al valor de RSD máxima. **Los resultados cumplen con el criterio de precisión.**

Nivel No. 3:
- Valor de concentración: 22 [mg/kg]
- RSD experimental: 3.96 %
- RSD máxima: 5.02 % (criterio de Horwitz)
- Valor p prueba estadística Chi cuadrado: 0.7127 (con 6 grados de libertad)
- **Conclusión:** La dispersión de los resultados no es significativamente mayor al valor de RSD máxima. **Los resultados cumplen con el criterio de precisión.**

Prueba de homocedasticidad entre las series.
Se realizó la prueba de Levene para evaluar la homocedasticidad del método
La precisión del método NO es constante en el intervalo de concentraciones considerado.
Valor p de la prueba de Levene: 0.0321

Gráfico de homocedasticidad



Finalmente, se tiene que frente a la heterocedasticidad del método observada para las concentraciones evaluadas se recomienda tomar en la estimación de incertidumbre el aporte de mayor variación, como fuente asociada a la precisión del método (repetibilidad).

3.5.2 Veracidad

La veracidad se define como la proximidad entre la media de un número infinito de valores medidos repetidos y un valor de referencia (JCGM, 2012). Esta definición implica que en la práctica la estimación de este parámetro es imposible, pues no se puede realizar un número infinito de mediciones. La veracidad se expresa en términos de sesgo (Norma ISO 5725-1, 1994) y recuperación.

La evaluación del componente de veracidad mediante las diferentes alternativas presenta una jerarquía en lo que se refiere a confianza, comparabilidad e incertidumbre. La Figura 3.6 presenta un diagrama de dicha jerarquía.

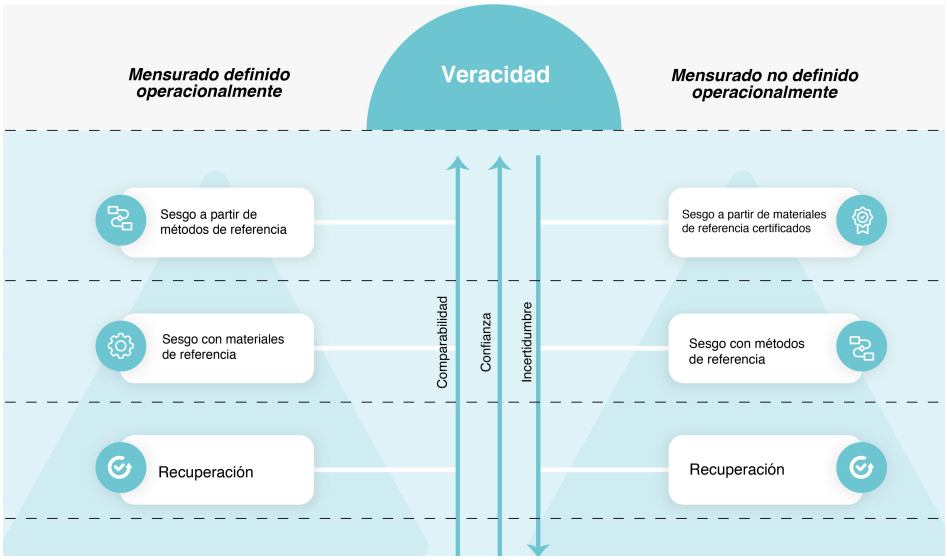


Figura 3.6: Jerarquía de las alternativas en la evaluación de la veracidad.

En términos prácticos, los experimentos de recuperación en muchos casos resultan ser más económicos y rápidos de realizar que los experimentos de sesgo; sin embargo, desde una perspectiva metrológica la evaluación de la veracidad a través de esta alternativa tiene varios inconvenientes como lo son:

- No se cuenta con trazabilidad metrológica, por lo menos a la magnitud de cantidad de sustancia.
- No se puede asegurar la estabilidad de las muestras y en muchos casos la homogeneidad.
- No se asegura la conmutabilidad de los materiales fortificados.
- No siempre es posible adicionar el analito en la forma en la que se encuentra en la matriz; por ejemplo, la determinación de ácidos grasos totales.

Las siguientes secciones presentan algunos aspectos generales de las diferentes alternativas para la evaluación de la veracidad del método.

3.5.2.1 Sesgo

El sesgo se define como la diferencia entre el valor esperado de los resultados de prueba y un valor de referencia aceptado (Norma ISO 3534-1, 2006):

$$Sesgo = \bar{x}_i - x_{referencia} \tag{3.10}$$

donde $\bar{x}_i$ corresponde al promedio de las n mediciones realizadas durante la evaluación del sesgo y $x_{referencia}$ corresponde a un valor de referencia aceptado. Cabe resaltar que la anterior ecuación es aplicable cuando se cuenta con un material caracterizado, suficientemente estable y homogéneo (por ejemplo, un MRC), pues a través de este se puede obtener un valor de referencia aceptado.

Si por el contrario se cuenta con un método de referencia la ecuación en este caso correspondería a la diferencia dada entre los valores promedio de las mediciones realizadas con dos métodos:

$$Sesgo = \bar{x}_i - \bar{x}_r \tag{3.11}$$

donde $\bar{x}_r$ corresponde al promedio del resultados del método de referencia.

Como se puede observar en cualquiera de las expresiones anteriores, el sesgo puede ser positivo o negativo, por lo cual estas diferencias se pueden atribuir a algunas fuentes error como son: la presencia de interferencias analíticas, pérdidas de analito por extracción incompleta, eficiencias de reacción en procesos de derivatización, efecto matriz, volatilización o adsorción, contaminación de la muestra por los reactivos o por el material empleado, entre otras (Hernández Revilla, 2013).

3.5.2.2 Recuperación

Los experimentos de recuperación se realizan comúnmente cuando (i) no se cuenta con materiales de referencia, (ii) el número de analitos es alto, por ejemplo, los métodos multiresiduo, (iii) el método tiene un amplio alcance, por ejemplo, diferentes tipos de alimentos, (iv) se realizan experimentos para la ampliación del alcance del método, por ejemplo, la inclusión de nuevos analitos.

La recuperación hace referencia a la relación entre la cantidad medida de analito y la cantidad conocida o añadida a la muestra, expresada como porcentaje; es decir, la recuperación hace referencia a la estimación del error sistemático y, al igual que para el caso del sesgo, la recuperación puede tener valores superiores al 100 % (errores positivos) y valores inferiores al 100 % (errores negativos).

En los experimentos de recuperación se adiciona una cantidad de analito conocida sobre una muestra de la que se conoce previamente que no hay presencia del analito (blanco de muestra), y posteriormente esta muestra fortificada se analiza utilizando el procedimiento de medición en evaluación. El porcentaje de recuperación se estima mediante la siguiente ecuación:

$$\% R = \frac{Cantidad\ medida}{Cantidad\ adicionada} \cdot 100\ \% \tag{3.12}$$

Por otro lado, en el caso en que no fuera posible obtener un blanco de muestra, se puede realizar un experimento de recuperación mediante el uso de una muestra que idealmente contenga el analito en concentraciones inferiores al límite inferior del intervalo de trabajo seleccionado. En este caso, una porción de la muestra se mide previamente (natural), posteriormente, a otra porción se le adiciona una cantidad conocida del analito y se mide (fortificada). Finalmente, se estima el porcentaje de recuperación mediante la siguiente ecuación:

$$\% R = \frac{Cantidad\ medida - Cantidad\ natural}{Cantidad\ adicionada} \cdot 100\ \% \tag{3.13}$$

3.5.2.3 Selección del criterio de aceptación

De igual manera a lo presentado en la Sección 3.5.1.1 (precisión), en el caso en que la regulación correspondiente o norma que aplique establezca el sesgo máximo o recuperación aceptables, estas deben ser tomadas como el criterio de aceptación. Por otro lado, en el caso en que no existan criterios para estos parámetros se pueden emplear dos alternativas:

- Empleo de guías de organizaciones reconocidas.
- Empleo de pruebas estadísticas.

3.5.2.3.1 Empleo de la regulación o guías

Organizaciones como IUPAC, AOAC, EPA, ICH, CODEX, entre otras, establecen criterios de aceptación para el porcentaje de recuperación o máximo error permitido. La Tabla 3.8 presenta algunos ejemplos.

Concentración (%)	Porcentaje de recuperación esperado		
	CODEX	AOAC	AEFI
100 - 10	98 % - 102 %	98 % - 102 %	98 % - 102 %
< 10	97 % - 103 %		
< 1	97 % - 103 %	97 % - 103 %	97 % - 103 %
< 0.1		95 % - 105 %	95 % - 105 %
< 0.01	90 % - 107 %	90 % - 107 %	90 % - 107 %
< 0.001	80 % - 110 %	80 % - 110 %	80 % - 110 %

Tabla 3.8: Criterios de aceptación sugeridos en guías de validación de métodos.

Nótese que los criterios de recuperación que se presentan en las diferentes guías suelen venir dados en términos de intervalos, por lo cual en el caso de hacer uso de estas tablas no es necesario realizar pruebas estadísticas que permitan soportar que es el adecuado para el uso. Por otro lado, de manera general se encuentra que los intervalos de aceptación en función de la concentración son prácticamente los mismos, sin embargo, algunos documentos consideran la complejidad de los analitos y las matrices, lo que implica que se pueden tener intervalos de aceptación diferentes (SANTE 11312, 2021). Lo anterior indica que siempre que se seleccionen criterios de aceptación de guías se debe verificar que:

- Las guías seleccionadas son adecuadas para el método y sector correspondiente, por ejemplo, si el método es aplicado a muestras de alimentos el empleo de guías tipo AOAC son idóneas para esto. Sin embargo, pueden existir guías más especializadas, similar al caso de residuos de medicamentos o plaguicidas en alimentos (SANTE 11312, 2021), las cuales pueden ser más idóneas.
- Idealmente, la incertidumbre de la referencia (valor esperado) con la cual se evalúe la recuperación debe ser inferior a $1/4$ del intervalo de recuperación esperado. Lo anterior asegura que se realicé una evaluación adecuada del error sistemático, pero se vuelve un reto para los laboratorios, pues para ello se debe asegurar la homogeneidad y estabilidad del fortificado.
- La precisión obtenida en el experimento de recuperación debe cumplir con los criterios esperados (ver Sección 3.5.1.1).

3.5.2.3.2 Empleo de pruebas estadísticas

Por otro lado, en el caso en que (i) no aplique ninguna regulación, (ii) no se empleen materiales de referencia certificados, (iii) no se empleen métodos de referencia, (iv) no se cuente con ninguna guía con criterios de aceptación aplicables, es necesario demostrar si el sesgo o porcentaje de recuperación es el adecuado de acuerdo con el alcance de la validación, a través de pruebas estadísticas de significancia.

En la evaluación del sesgo por medio de MRC, es necesario involucrar dentro de la evaluación del sesgo la incertidumbre asociada a dicho material, puesto que el valor de referencia se encuentra dentro del intervalo de valores asociado a la incertidumbre. Por esta razón como criterio de aceptación del sesgo se emplea la siguiente ecuación donde se establece que el sesgo $(\bar{x}_i - \mu)$, se debe encontrar entre los límites definidos por la incertidumbre estándar del MRC $(u(MRC))$ y la desviación estándar (s) del experimento de evaluación del sesgo:

$$-2 \cdot \sqrt{u(MRC)^2 + s^2} \leq \bar{x}_i - \mu \leq 2 \cdot \sqrt{u(MRC)^2 + s^2} \tag{3.14}$$

Por su parte, la guía ISO 33 propone una evaluación del sesgo que considera la incertidumbre $(u(\bar{x}_i))$ en lugar de la precisión (s):

$$| \bar{x}_i - \mu | \leq 2 \cdot \sqrt{u(MRC)^2 + u(\bar{x}_i)^2} \tag{3.15}$$

Donde $u(\bar{x}_i)$ corresponde a la incertidumbre de medición obtenida durante la evaluación del sesgo, es decir la incertidumbre del promedio de mediciones. Por otro lado, en el caso en que el componente de incertidumbre

que más aporte sobre la incertidumbre de medición ($u(\bar{x}_i)$) corresponda al proveniente de la precisión de las mediciones, la ecuación se transforma a:

$$| \bar{x}_i - \mu | \leq 2 \cdot \sqrt{u(MRC)^2 + s^2} \tag{3.16}$$

Por otro lado, la prueba t de comparación de medias es quizás la prueba estadística que más frecuentemente se sugiere en los textos de validación, dicha prueba se presenta en la Sección 2.2.5. La siguiente tabla presenta algunas sugerencias para su uso durante la evaluación del sesgo o recuperación:

Caso que aplica	**Ecuación**	**Información adicional**
Evaluación de sesgo contra un MRC	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$	Sección 2.2.5.1
Comparación de un porcentaje de recuperación contra 100 %	$t = \frac{\%\, R - 100\,\%}{s_{\%\, R}/\sqrt{n}}$	Sección 2.2.5.1
Evaluación de sesgo o recuperación contra un método de referencia	$t = \frac{\bar{x}_x - \bar{x}_y}{s_{x,y}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$	Sección 2.2.5.2
Evaluación de sesgo o recuperación contra un método de referencia en caso de heterocedasticidad	$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$	Sección 2.2.5.2

Tabla 3.9: Pruebas sugeridas para la evaluación del sesgo o la recuperación.

3.5.2.4 Aspectos prácticos en la evaluación de la veracidad

El esquema general de la evaluación de la veracidad se presenta a continuación:

a) Seleccionar las concentraciones en las cuales se realizará la evaluación.
   - Nota 1: la concentración máxima y mínima puede ser sugerida por la regulación o las recomendadas en la Sección 3.4 - intervalo de trabajo.
   - Nota 2: se sugiere seleccionar por lo menos tres concentraciones para la evaluación de este componente, las cuales pueden ser el límite de cuantificación, el límite fijado por la regulación y el límite superior del intervalo de trabajo del método.

b) Seleccionar o preparar los materiales con los cuales se realizará la evaluación, puede tratarse de materiales de referencia certificados o materiales fortificados.
   - Nota: en el caso de requerir evaluar el sesgo a través de un método de referencia o de más alta jerarquía (comparación de métodos) se debe asegurar que este se encuentra bien implementado y su alcance es el adecuado. Para este caso en particular, no se requiere conocer el valor de referencia, pues este se estima a partir del método de mayor jerarquía.

c) Analizar el conjunto de muestras con el método que se está validando.
   - Nota 1: se sugiere que esto se realicé con un diseño completamente al azar y en el caso de ser posible se realicen por lo menos entre 5 y 10 réplicas independientes de medición.
   - Nota 2: si se trabaja con materiales de referencia se debe considerar todas las instrucciones que se dan en el certificado.

d) Para el caso de materiales de referencia certificados, evaluar si el sesgo es significativo en concordancia con las recomendaciones del numeral 3.5.2.3.2.

e) Para el caso de muestras fortificadas, se debe estimar el porcentaje de recuperación (% R), mediante las ecuaciones 3.6 o 3.7, según aplique.

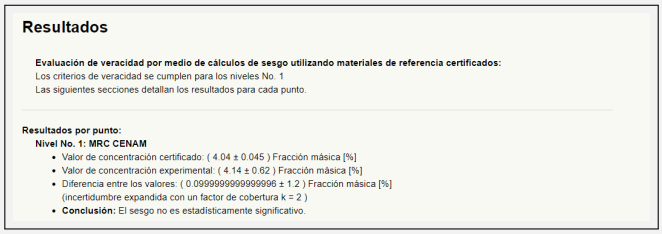
f) Posteriormente, se debe comparar el % R obtenido con el criterio de aceptación seleccionado de acuerdo con el numeral 3.5.2.3.

Ejemplo 21: Veracidad: Evaluación de la veracidad con información de sesgo.

Un laboratorio de análisis de alimentos desea prestar el servicio de medición del contenido de Nitrógeno total en leche en polvo empleando el método Kjeldahl. Durante la planeación de la validación el personal asignado para la evaluación del sesgo usará un MRC de leche cuyo valor de nitrógeno total en base seca es de 4.04 % ± 0.09 % (k = 2, con un 95 % de nivel de confianza). De esta manera, se planteó un experimento con un diseño completamente al azar para el cual se tomaron 7 porciones que se trataron acorde con el procedimiento de Kjeldahl. La tabla que aparece abajo a la izquierda presenta los porcentajes de nitrógeno obtenidos para cada una de las submuestras estudiadas.

La significancia del sesgo para la determinación del contenido de Nitrógeno total en base seca se puede realizar con ayuda del aplicativo validaR. Los resultados se muestran en el recuadro que aparece abajo a la derecha.

Submuestra No.	**Nitrógeno total (% base seca)**
1	4.14
2	4.12
3	4.15
4	4.14
5	4.13
6	4.15
7	4.16
Promedio	**4.14**
$u(\bar{x}_i)$	**0.62**


Resultados

Evaluación de veracidad por medio de cálculos de sesgo utilizando materiales de referencia certificados:
Los criterios de veracidad se cumplen para los niveles No. 1
Las siguientes secciones detallan los resultados para cada punto.

Resultados por punto:
Nivel No. 1: MRC CENAM
- Valor de concentración certificado: (4.04 ± 0.045) Fracción másica [%]
- Valor de concentración experimental: (4.14 ± 0.62) Fracción másica [%]
- Diferencia entre los valores: (0.0999999999999996 ± 1.2) Fracción másica [%]
(incertidumbre expandida con un factor de cobertura k = 2)
- **Conclusión:** El sesgo no es estadísticamente significativo.

El sesgo para esta validación no se considera estadísticamente significativo por lo que no necesita ser corregido. Por otro lado, es necesario que el método se verifique periódicamente para asegurar que se cumple la vigencia de este resultado.

3.6 Intervalos instrumentales y linealidad

En métodos analíticos el intervalo hace referencia a las concentraciones de analito comprendidas entre un nivel inferior y uno superior. En los métodos analíticos es posible caracterizar diferentes intervalos, dentro de los que se encuentran el intervalo de trabajo del método, el intervalo dinámico del instrumento de medición y el intervalo lineal del instrumento.

El intervalo de trabajo del método (Sección 3.4) hace referencia al intervalo de concentraciones en el cual el método brinda resultados confiables. Este intervalo se expresa en términos de la matriz o muestra de trabajo, por ejemplo 1 a 100 mg de Na/kg de suelo.

Por su parte, el intervalo dinámico es la zona donde la respuesta analítica del instrumento[2] varía con la concentración del analito. Los extremos de dicho intervalo están dados normalmente por el límite de detección como límite inferior y un límite superior el cual depende de la respuesta máxima que el instrumento puede obtener. Como se puede detallar, este intervalo dinámico depende completamente del fabricante del instrumento y por consiguiente no se suele evaluar. Por último, el intervalo lineal corresponde a una zona del intervalo dinámico en la que se encuentra una relación lineal entre la concentración del analito y la respuesta del instrumento, en otras palabras, la sensibilidad es constante. La Figura 3.7 presenta una representación de estos intervalos.

[2]Nótese que se habla del instrumento. En este punto, es importante aclarar que la evaluación de la linealidad del sistema aplica sólo para los métodos de medición que hacen uso de la calibración analítica para poder cuantificar las muestras, por ejemplo, curvas de calibración, *bracketing*, adición patrón, entre otros, que hacen uso de modelos de regresión o asumen linealidad en el sistema de medición en un determinado intervalo.

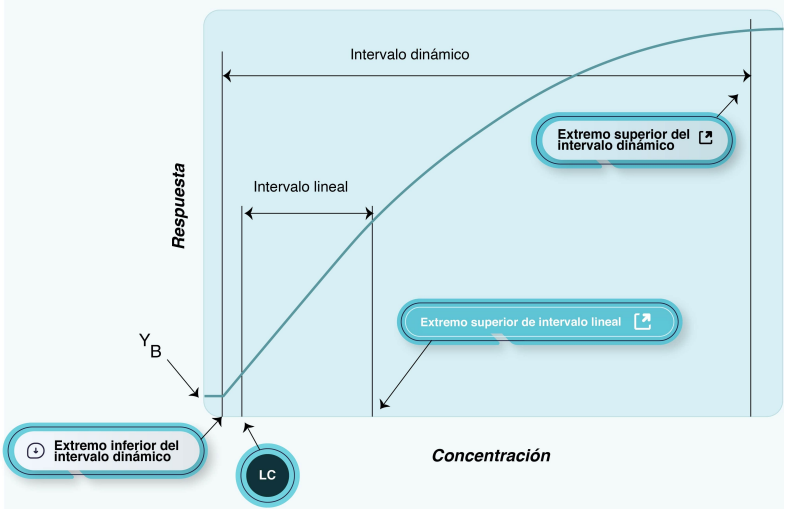


Figura 3.7: Representación del intervalo lineal y el intervalo dinámico.

3.6.1 Intervalo lineal

El intervalo lineal es establecido por el personal del laboratorio durante la etapa de diseño del método y usualmente el límite inferior puede estar dado por una concentración equivalente al límite de cuantificación del método, aunque idealmente este límite inferior debería encontrarse por debajo de la concentración equivalente al límite de cuantificación, en especial considerando que algunos analitos tienen pérdidas durante las diferentes etapas del proceso de medición, por ejemplo, extracciones, digestiones y derivatizaciones. Por su parte, el límite superior corresponderá a la máxima concentración en la cual se tenga un comportamiento lineal con la respuesta analítica y en el caso ideal esta concentración debe ser igual o mayor a la concentración equivalente que se puede determinar cuando el método analítico es exacto.

Ejemplo 22: Intervalo lineal: Definición de los límites.

En la validación de un método de medición para la determinación de residuos de plaguicidas en aguacate por cromatografía líquida acoplada a espectrometría de masas se debe validar el intervalo lineal. Para ello, es necesario plantear de manera preliminar el intervalo lineal a evaluar de cada plaguicida. En el análisis de los plaguicidas en esta matriz, se tiene que el límite de cuantificación del método es de 8.0 μg/kg de plaguicida en aguacate. Por su parte, el método tiene un factor de conversión a extracto de 10. Si se tiene en cuenta que el primer nivel de la curva de calibración se debe preparar entre un 10 % y 30 % por encima de la concentración equivalente al intervalo de trabajo, es decir, entre 0.8 μg/kg y 3 μg/kg (conversión de concentración en aguacate a concentración de extracto); el punto más bajo de la curva debe encontrarse entre 0.88 μg/L y 1.04 μg/L. Por otro lado, si el límite superior del intervalo de trabajo del método es 30 μg/kg de plaguicida en aguacate, el límite superior de la curva debería encontrarse entre 3.3 μg/L y 3.9 μg/L.

3.6.2 Selección del modelo de regresión

En química analítica el objetivo es establecer una ecuación que permita obtener una adecuada predicción de una variable (normalmente concentración) para un valor determinado de la otra (normalmente respuesta instrumental). Dicha ecuación típicamente es de la forma:

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon \tag{3.17}$$

donde β_0 y β_1 son los parámetros de la regresión lineal, conocidos como intercepto y pendiente, respectivamente, y ε es la variable que representa el error aleatorio del resultado de la medida. El establecimiento de estos parámetros se puede lograr a través del empleo de diferentes modelos de regresión, los cuales tienen diferentes supuestos (ver Sección 2.3). En este sentido, es indispensable que se revisen dichos supuestos con el propósito de seleccionar el mejor modelo de regresión. La Figura 3.8 presenta un esquema que sugiere un mecanismo para la selección del modelo a emplear.

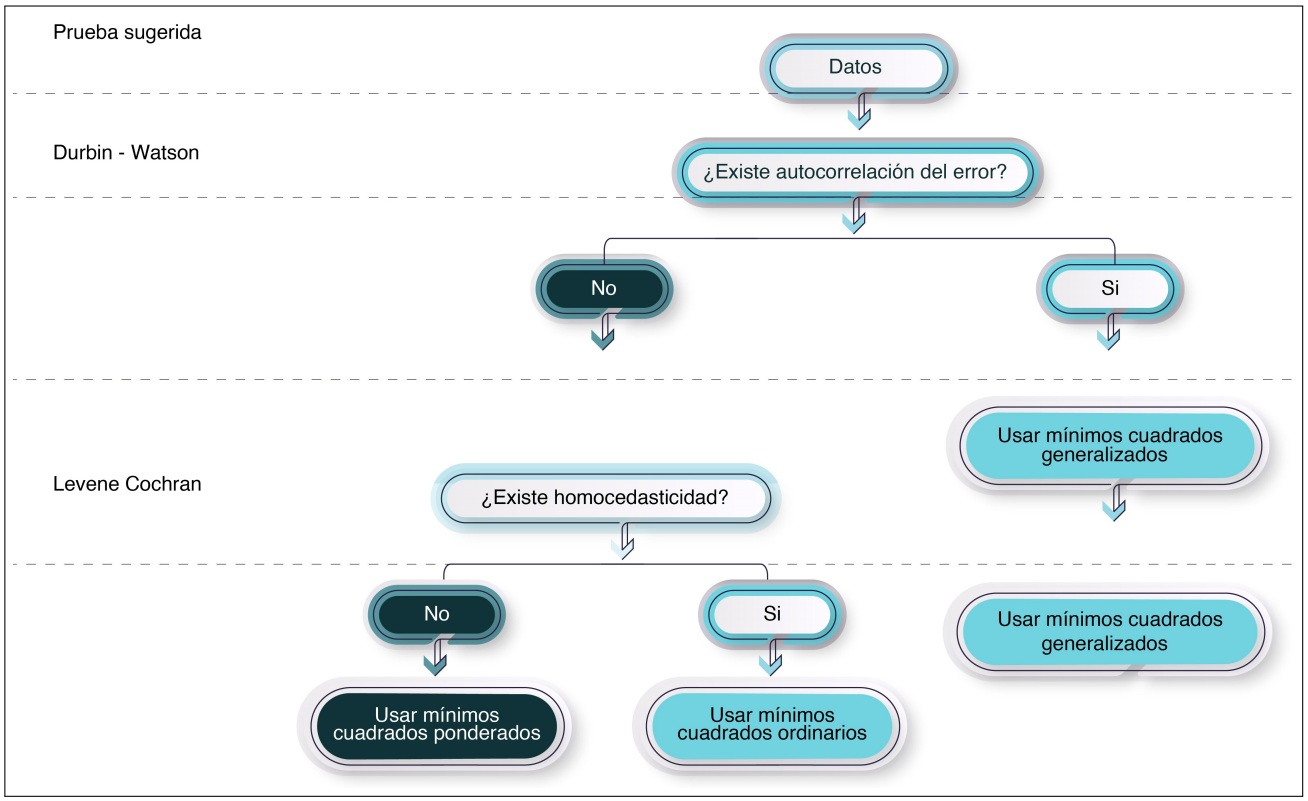


Figura 3.8: Selección del modelo de regresión de acuerdo con el comportamiento de los datos.

Ejemplo 23: Selección del modelo de regresión.

En un laboratorio de análisis de aguas que utiliza la técnica ICP-OES se desea obtener el intervalo lineal para plomo como parte de una validación. Para ello, la persona responsable preparó una disolución madre de 500 μg/L en HNO_3 al 0.5 %, a partir de una ampolleta del SRM 3128 de NIST. Con esta disolución obtuvo diez niveles de calibración entre 0.26 μg/L y 31.21 μg/kg; todas las disoluciones contienen germanio de 10 μg/L como estándar interno. Posteriormente se realizaron aleatoriamente 10 repeticiones de medición para cada nivel. La respuesta relativa promedio (respecto a germanio) y su desviación estándar se muestran en la siguiente tabla.

Concentración Pb /[μg/kg]	**Respuesta promedio** (n = 10)	**Desv. estándar**
0.25	0.29	0.0070
0.50	0.49	0.0136
1.00	0.76	0.0081
5.04	4.14	0.0617
9.93	8.10	0.1523
14.99	11.62	0.2874
20.01	15.47	0.3164
26.00	20.44	0.4149
31.21	24.18	0.5527
35.12	27.85	0.5291

El siguiente recuadro muestra los resultados de regresión por OLS (ver Sección 2.3.1) utilizando el módulo **Regresión lineal** de la sección **Herramientas estadísticas** del aplicativo validaR.

Resultados

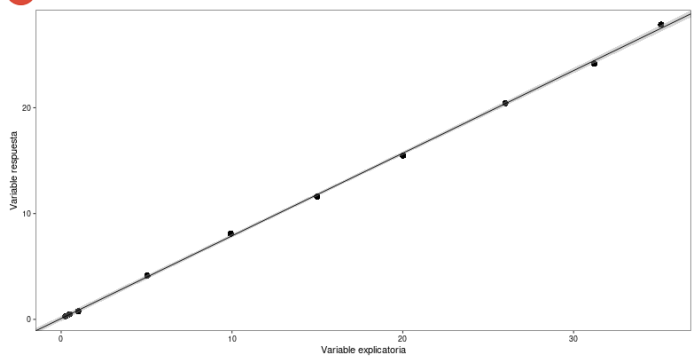
Curva de regresión por mínimos cuadrados ordinarios:
Y = 0.0820868 + 0.781112 * X

Error residual estándar de la regresión: 0.216661

Detalle de los parámetros de regresión:

.	Intercepto	Pendiente
Parámetro	0.0820868	0.781112
Error estándar	0.104608	0.00548753
Intervalo de confianza		
~ límite inferior	-0.159139	0.768457
~ límite superior	0.323312	0.793766

Gráficos: Escala-Ubicación | Q-Q residuales | Residuales | Diagrama de calibración



Descargar gráfico

Supuestos de los modelos:

Supuestos del análisis: Normalidad de los residuales

Los residuales parecen tener distribución normal. Valor p prueba de Shapiro-Wilk sobre los residuales de la regresión: 0.7877 Se recomienda verificar esta información con el gráfico Q-Q de los residuales.

Supuestos del análisis: Ausencia de autocorrelación de error

No hay una aparente autocorrelación del error:

.	Valor
Autocorrelación	-0.28
Estadístico Durbin-Watson	2.26
Valor p	0.99

Se recomienda verificar esta información con el gráfico de residuales.

Supuestos del análisis: Homocedasticidad del error

Hay homocedasticidad en los residuales de regresión. Valor p prueba de Breusch-Pagan: 0.1016 Se recomienda verificar esta información con el gráfico de residuales y el gráfico de ubicación de escala.

Los supuestos del modelo de OLS parecen estarse cumpliendo, pero la tabla muestra que la dispersión de las señales aumenta para valores altos de concentración de plomo. Debido a esto es más adecuada una regresión lineal por WLS:

Resultados

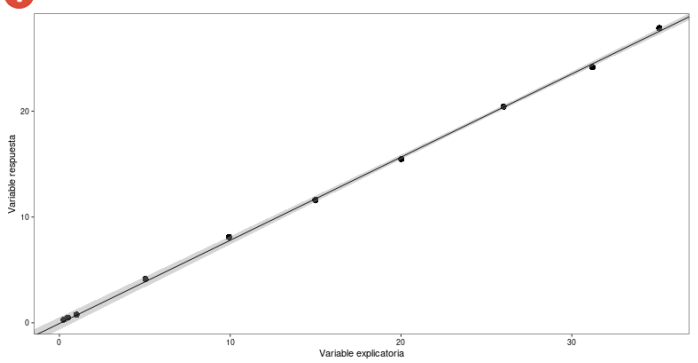
Curva de regresión por mínimos cuadrados ponderados:
Y = 0.0637673 + 0.764536 * X

Error residual estándar de la regresión: 0.00634393

Detalle de los parámetros de regresión:

.	Intercepto	Pendiente
Parámetro	0.0637673	0.764536
Error estándar	0.0246102	0.0230883
Intervalo de confianza		
~ límite inferior	0.00701614	0.711294
~ límite superior	0.120518	0.817777

Gráficos: Escala-Ubicación | Q-Q residuales | Residuales | Diagrama de calibración



Descargar gráfico

3.6.3 Evaluación de la linealidad

En la evaluación de la linealidad se busca demostrar que el intervalo seleccionado es el adecuado, así como demostrar que el modelo de regresión seleccionado se ajusta de manera adecuada a los datos, permitiendo una adecuada predicción del comportamiento de los resultados experimentales. De acuerdo con la guía EURACHEM, la linealidad es una propiedad importante de los métodos utilizados para efectuar mediciones en un intervalo de concentraciones. Sin embargo, esta linealidad no es cuantificada, por lo cual es comprobada mediante inspección y utilizando pruebas de significancia (Magnusson y Örnemark, 2014). Lo anterior indica que este proceso de evaluación de la linealidad puede ser abordado a través de las siguientes estrategias:

- Análisis Exploratorio: basado en un análisis de correlación.
- Pruebas de significancia: basado en la evaluación del modelo y sus parámetros.

Es importante resaltar que a pesar de las recomendaciones de algunos textos la evaluación de la linealidad se debe realizar mediante el empleo de las dos estrategias, de lo contrario se puede llegar a conclusiones erróneas. A continuación, se presenta una breve descripción de cada una de las estrategias.

3.6.3.1 Análisis exploratorio

En esta primera fase, se evalúa el grado o intensidad de la asociación o relación entre variables, pues de manera general, se tiene que la confiabilidad en la predicción es alta cuando la correlación entre las variables es mayor; por lo anterior diferentes textos de validación sugieren la evaluación del coeficiente de correlación o el coeficiente de determinación.

Previo a esta evaluación se recomienda elaborar una gráfica de dispersión de los valores de concentración contra las respuestas de cada una de las repeticiones para cada concentración. Una vez obtenida la gráfica se procede a observar el comportamiento de la curva y determinar de manera cualitativa si la tendencia de la relación entre variables es lineal o no.

Adicionalmente, esta gráfica permite visualizar si se cuenta con posibles puntos anómalos de la regresión, con lo cual se puede decidir si es necesario repetir los experimentos, si se aplican pruebas de descarte de datos de regresión lineal, o si se selecciona un modelo de regresión robusto como la regresión no paramétrica (disponible en el aplicativo validaR). En caso de que la gráfica cuente con una linealidad aparentemente buena, se procede a estimar el valor del coeficiente de correlación (R).

El coeficiente de correlación representa el grado de asociación de las dos variables, lo que se traduce en que representa de una manera indirecta el grado de ajuste del modelo de regresión a los datos. En el caso en que se tenga un modelo de regresión perfecto este coeficiente tomará un valor de 1. Por lo anterior, una vez se tenga este coeficiente, algunos autores sugieren el empleo de una prueba t de Student con el propósito de evaluar si la relación entre las variables es significativa o no.

$$t_{cal} = \frac{|R|\sqrt{m-2}}{\sqrt{1-R^2}} \tag{3.18}$$

donde m corresponde al número de niveles de la curva y R al coeficiente de correlación.

Ejemplo 24: Selección del modelo de regresión.

En un laboratorio de ensayo se está montando un método para la determinación de hierro en agua potable por FAAS. Para ello, la persona responsable preparó una disolución madre de 5 mg/L en HNO_3 al 1 %, a partir de una disolución calibrante comercial de hierro de 1000 mg/L. Con esta disolución se obtuvieron diez niveles de calibración entre 100 y 550 μg/L que fueron leídos aleatoriamente en el sistema de medición. La absorbancia promedio (cinco réplicas instrumentales) se muestra en la tabla de abajo a la izquierda. Los resultados del análisis exploratorio se muestran en la tabla de abajo a la derecha:

Concentración Fe /[μg/L]	Absorbancia (n = 5) /[UA]
100	0.1558
150	0.1619
200	0.1809
250	0.2105
300	0.2143
350	0.2998
400	0.2652
450	0.2746
500	0.3183
550	0.3294

R	0.9565
R^2	0.9149
t_{cal}	9.2762
$t(0.05, 8)$	2.306
$\mathbf{H_0}$	El coeficiente de correlación obtenido procede de una población cuya correlación es cero.
Resultado	$t_{cal} > t(0.05, 8)$ por lo tanto se rechaza $\mathbf{H_0}$.

Se concluye que la correlación obtenida no procede de una población con una correlación igual a cero. Por tanto las variables están relacionadas.

Como se puede apreciar en el ejemplo anterior, a pesar de tener un coeficiente de correlación relativamente bajo, la prueba t de Student indica que el coeficiente es estadísticamente igual a 1, por lo cual en primera instancia se podría encontrar que el sistema es lineal. Sin embargo, y aunque es una práctica común, se debe evitar llegar a esta conclusión debido a que es una prueba que no evalúa el ajuste del modelo, sino sólo el grado de asociación de las variables.

La Figura 3.9 muestra la simulación de la Ecuación 3.18 para una curva de 10 niveles de concentración; en esta figura se puede observar que siempre que se tengan coeficientes de correlación superiores a 0.65, se concluirá que la regresión es significativa (a un nivel de confianza del 95 %). En este sentido, se puede observar que esta resulta ser una prueba poco idónea para evaluar la linealidad del sistema, por lo cual se debe recurrir a una evaluación más integral de la linealidad.

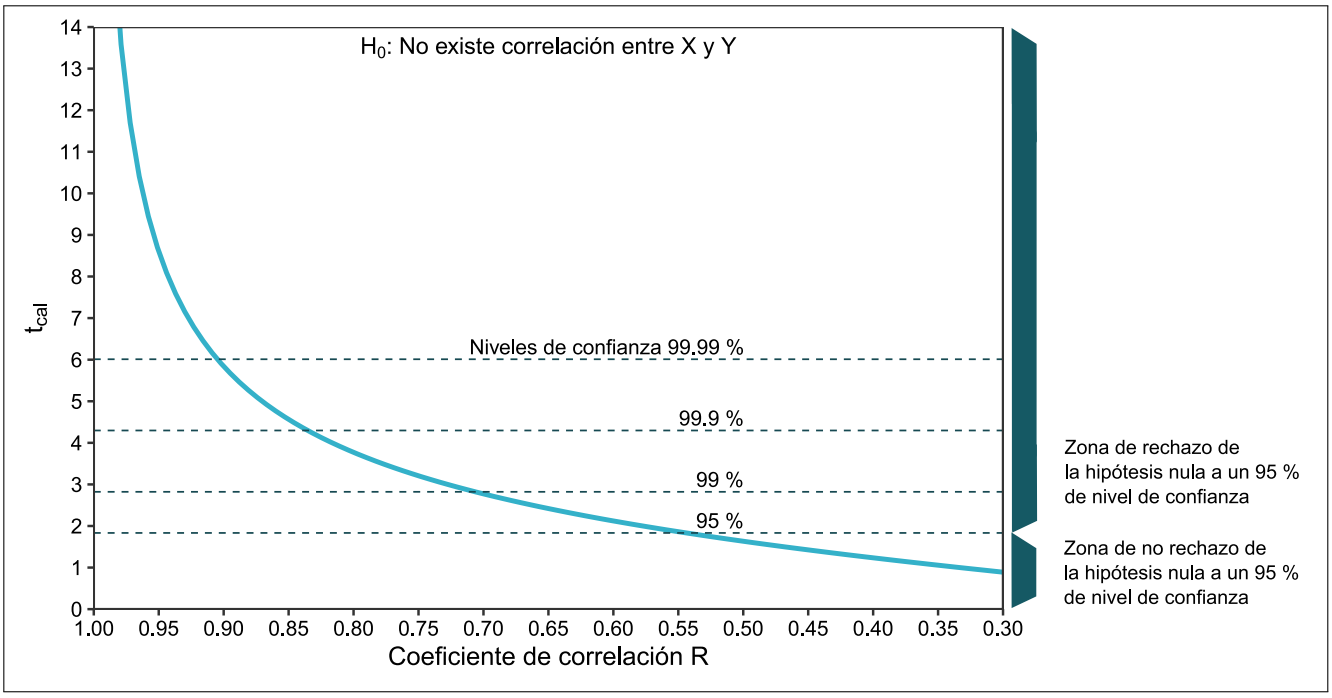


Figura 3.9: Variación del valor de t_{cal} (Ecuación 3.18) en función del coeficiente de correlación. Las líneas horizontales representan el valor de t_{crtico} de las tablas, a diferentes niveles de confianza.

La Figura 3.9 presenta valores de t_{tabla} a diferentes niveles de confianza (95 %, 99 %, 99.9 %, 99.99 %). Esta figura busca ilustrar como a medida que aumenta el nivel de confianza, el valor de t aumenta y por consiguiente la prueba

se vuelve más exigente. Lo anterior implica que, aunque de manera general se emplea un nivel de confianza del 95 %, la selección de este nivel depende de diferentes variables como el tamaño de la muestra y el contexto de la prueba, por lo que se puede recurrir a cambiar el nivel de confianza con el propósito de obtener pruebas más acordes con la necesidad o que permita asegurar con mayor confianza que, para este caso, los resultados presentan una relación entre sí (Anita Nanda and Bibhuti Bhusan and Abikesh Kumar and Abiresh Kumar and Abinash Kumar, 2021).

Finalmente, es de resaltar que la linealidad no es una propiedad cuantitativa, por lo cual un mayor coeficiente de correlación no implica que se tenga un mejor modelo de regresión, por lo anterior contrario a algunos documentos o guías, en el presente documento se recomienda evitar el uso de este coeficiente como un criterio de calidad de la curva de calibración y mucho menos como evidencia objetiva de la linealidad del sistema.

3.6.3.2 Pruebas de significancia

Una vez se ha realizado el análisis exploratorio y confirmado que la curva de calibración es en apariencia lineal, se debe realizar un conjunto de pruebas estadísticas que permiten soportar dicha linealidad de una manera objetiva. La Figura 3.10 presenta el esquema propuesto para llevar a cabo la evaluación de la linealidad.

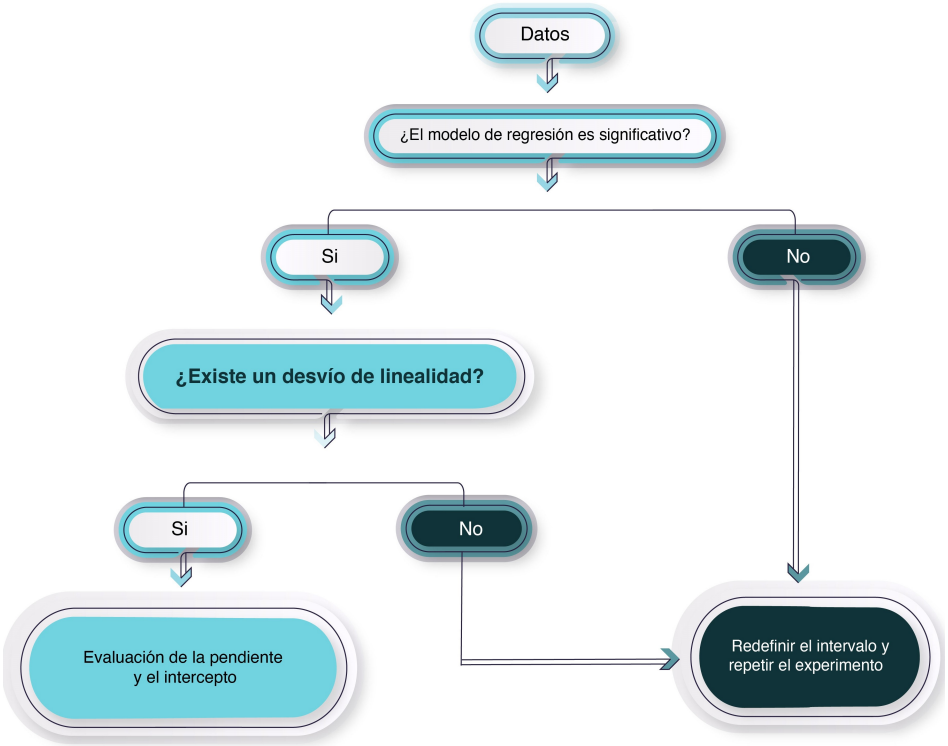


Figura 3.10: Esquema propuesto para la evaluación de la linealidad del sistema.

El análisis de varianza permite verificar que la varianza de los datos es explicada por el modelo de regresión (ANOVA de regresión), y evaluar la bondad del ajuste (ANOVA de falta de ajuste) (Norma ISO 11095, 1996). Es importante mencionar que para realizar el ANOVA de falta de ajuste se debe contar con mediciones repetidas de cada uno de los niveles que componen la curva de calibración, por lo menos triplicados.

El ANOVA de regresión permite demostrar que el modelo planteado (Ecuación 3.17) se ajusta de manera significativa a los datos. Por su parte, con el ANOVA de desvío de linealidad se demuestra que el modelo no tiene ningún tipo de curvatura o desvío en los extremos de la curva de calibración, por lo cual se pueden hacer estimaciones con el modelo en todo el intervalo y no se presentarán errores significativos a bajas o altas concentraciones.

Ejemplo 25: Linealidad: Desvío del comportamiento lineal.

En un laboratorio de ensayo se está implementando un método para la determinación de potasio en agua potable por emisión atómica (FAES). Como parte de esta validación se evalúa el parámetro de linealidad, para lo cual se preparan 7 niveles de calibración en un intervalo entre 5.0 mg/L y 20 mg/L a partir de una disolución calibrante de potasio de concentración 500 mg/kg. De cada nivel de calibración se realizaron 7 medidas repetidas. La medición de estos niveles se realizó de manera aleatoria. A continuación se presentan los resultados de las mediciones:

	Niveles de concentración /[mg/kg]						
	4.95	**7.41**	**9.88**	**12.35**	**14.82**	**17.28**	**19.78**
Repetición 1	4219	6117	7890	9568	11113	12626	13916
Repetición 2	4230	6144	7925	9586	11110	12651	13998
Repetición 3	4220	6055	7946	9585	11121	12721	14114
Repetición 4	4236	6066	7937	9560	11225	12665	14041
Repetición 5	4203	6119	7978	9533	11166	12622	13937
Repetición 6	4191	6126	7914	9619	11222	12619	14149
Repetición 7	4221	6152	7857	9549	11138	12646	14116

En la evaluación de pruebas de significancia se obtienen los siguientes resultados:

Resultados

Análisis exploratorio de la regresión

- Coeficiente de correlación R: 0.99841
- Estadístico t asociado: 121.45 , con 47 grados de libertad
- Valor p: < 1 e-05
- El coeficiente de correlación es estadísticamente significativo

- Pendiente de regresión: 661.883
- Error estándar: 5.45
- Estadístico t asociado: 121
- Valor p: < 1 e-04
- La pendiente de regresión es estadísticamente significativo

- Intercepto de regresión: 1204.05
- Error estándar: 72.51
- Estadístico t asociado: 16.6
- Valor p: < 1 e-04
- El intercepto de regresión es estadísticamente significativo

Prueba de ajuste lineal

El modelo lineal describe adecuadamente el comportamiento de los datos de la curva de calibración.
Valor P de la regresión: < 1 e-04

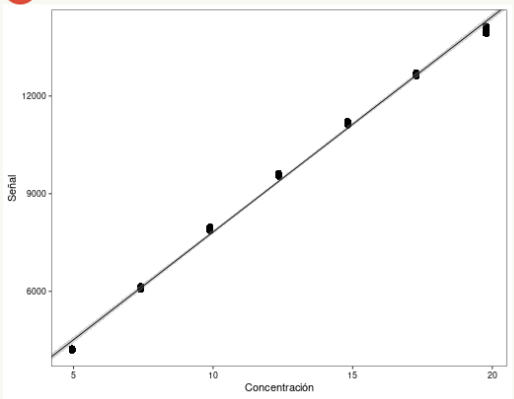
(Un valor P pequeño indica que la pendiente de la regresión es estadísticamente significativa)

Prueba de falta de ajuste

El modelo lineal u falta de ajuste.
Valor P del error puro: < 1 e-04

(Un valor P muy pequeño indica que el modelo lineal presenta falta de ajuste)

Diagrama de calibración

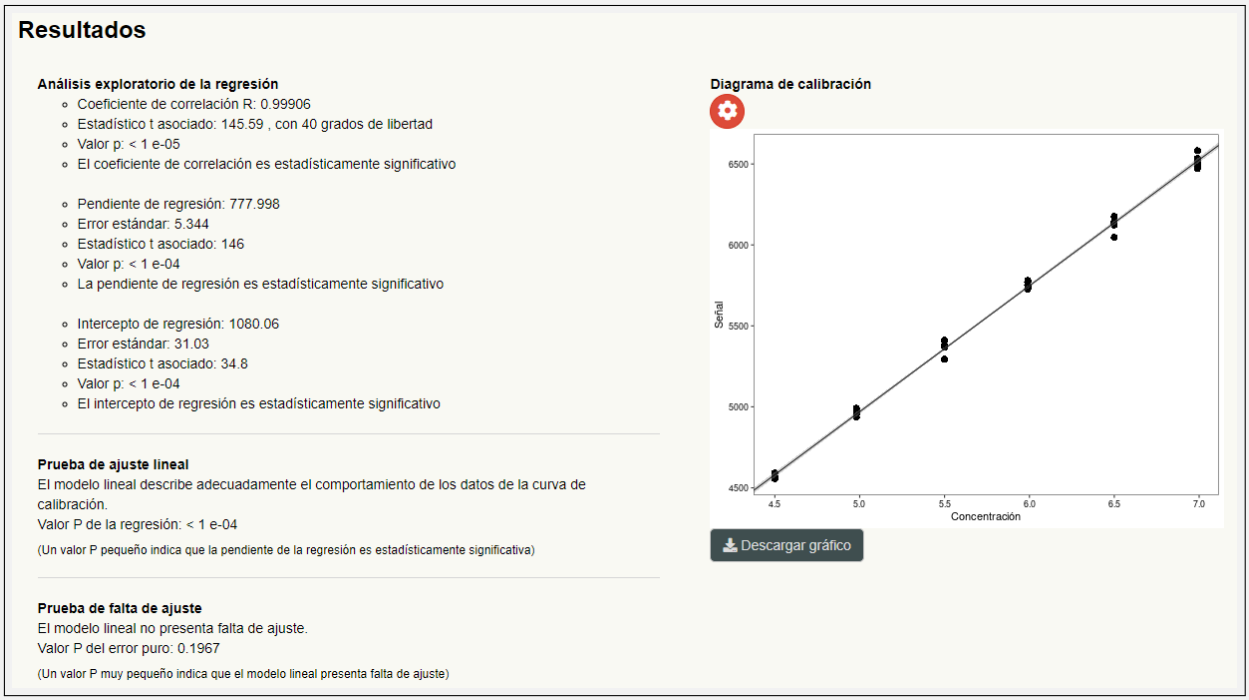


Descargar gráfico

Conclusión: Es necesario redefinir el intervalo evaluado y repetir el experimento

Con base en los resultados obtenidos se evalúa nuevamente el intervalo lineal comprendido entre 4.5 mg/L y 7.0 mg/kg, para lo cual se preparan 6 niveles de los cuales se obtienen 7 repeticiones de medición para cada nivel. Los resultados se presentan a continuación.

	Niveles de concentración /[mg/kg]					
	4.50	**4.98**	**5.50**	**5.99**	**6.50**	**6.99**
Repetición 1	4558	4982	5294	5727	6047	6472
Repetición 2	4592	4991	5411	5781	6176	6534
Repetición 3	4562	4960	5380	5770	6130	6513
Repetición 4	4556	4969	5378	5741	6145	6582
Repetición 5	4564	4971	5373	5735	6123	6489
Repetición 6	4556	4955	5376	5745	6131	6504
Repetición 7	4557	4937	5369	5752	6139	6504



Conclusión: El intervalo lineal del método se encuentra entre 4.5 mg/L y 7.0 mg/kg. Dentro de este intervalo es adecuado realizar la cuantificación de muestras.

Por otro lado, en concordancia con la Figura 3.10, posterior a demostrar que el modelo seleccionado es significativo no se tienen desvíos de la linealidad, se procede a evaluar la significancia de la pendiente y el intercepto. Para tal fin, se usa una prueba t, el cual se halla dividiendo el valor absoluto de la pendiente o el intercepto sobre su respectiva desviación estándar:

$$t_{\beta_0} = \frac{|\beta_0|}{s_{\beta_0}} \tag{3.19}$$

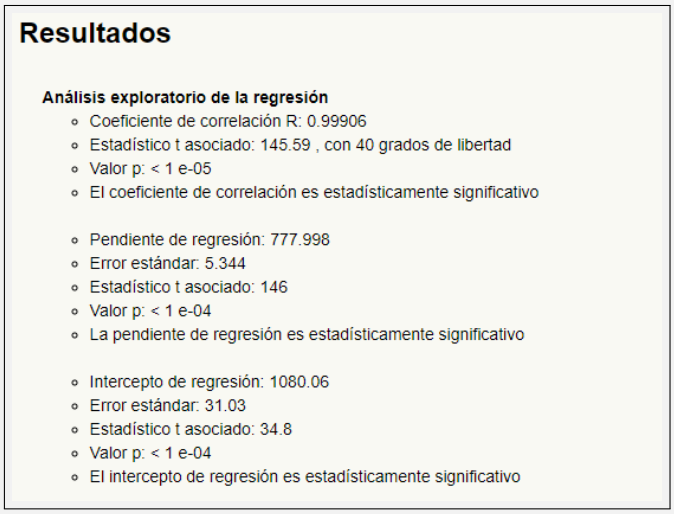
$$t_{\beta_1} = \frac{|\beta_1|}{s_{\beta_1}} \tag{3.20}$$

Para comprobar que la pendiente es significativamente diferente de cero se busca rechazar la hipótesis nula (la pendiente no es significativamente diferente de cero), siendo el valor t_{β_1} mayor al valor $t_{tabulado}$.

Para el caso del intercepto se espera aceptar la hipótesis nula, la cual establece que el valor del intercepto sea cercano a cero, siendo el valor t_{β_0} menor al valor $t_{tabulado}$.

Ejemplo 26: Linealidad: Pruebas t de la pendiente y el intercepto.

Se continua con los datos utilizados en el Ejemplo 3.6.3.2. Luego de acotar el intervalo de concentraciones se observó que los datos se ajustan a una regresión lineal y no presentan falta de ajuste al modelo. El recuadro que aparece abajo a la izquierda muestra los resultados de las pruebas t de la pendiente y el intercepto con el aplicativo validaR.



Conclusión: En el intervalo evaluado, el sistema de detección identifica los cambios cuando hay variaciones en las concentraciones. El intercepto de la curva debe ser incluido en la ecuación del modelo en caso de realizar cuantificación de muestras por interpolación en la curva de calibración.

3.6.4 Aspectos prácticos en la evaluación de la linealidad

La linealidad es una propiedad de los instrumentos de medición, que puede ser afectada por algunos aspectos propios del método de medición, tal es el caso del efecto matriz. A continuación se dan algunas recomendaciones de tipo práctico para la evaluación de este parámetro.

- La linealidad se evalúa en condiciones de repetibilidad.
- El modelo de regresión más empleado (mínimos cuadrados ordinarios) asume que los errores del eje X (concentración) son despreciables, por lo cual es indispensable que cuente con un excelente sistema de control de calidad que asegure que se minimicen los errores de la preparación de los diferentes niveles o estándares de calibración.
- No prepare los niveles de calibración inferiores a partir de la dilución de los niveles superiores de la curva.
- Trate de diseñar curvas con niveles equidistantes. Por ejemplo, si el intervalo a evaluar es de 10 ng/mL a 110 ng/mL y se desean preparar 6 niveles, la curva estaría compuesta por los siguientes puntos: 10 ng/mL, 30 ng/mL, 50 ng/mL, 70 ng/mL, 90 ng/mL y 110 ng/mL.
- Evite que la relación de la concentración entre niveles consecutivos sea superior a 4. Por ejemplo, si el nivel más bajo corresponde a 10 ng/mL, el siguiente nivel no debe exceder los 40 ng/mL.
- El análisis estadístico se realiza para las mediciones de una sola curva, es decir no se tienen que preparar varios niveles de la misma concentración, sino que en la práctica se deben medir varias veces cada uno de los niveles preparados. Sin embargo, si desea confirmar la linealidad del sistema, puede preparar una nueva curva, realizar las respectivas mediciones y volver a realizar el análisis estadístico.
- El experimento de linealidad no se limita por el número de niveles de calibración que en la rutina se van a emplear. Es decir, puede que en la rutina se desee trabajar con 5 puntos de calibración, sin embargo, el experimento de linealidad se puede realizar con diez puntos de calibración. De hecho, el establecimiento de los 5 puntos que se van a trabajar rutinariamente se debe realizar posterior a la evaluación de la linealidad.

A continuación se presentan el esquema general para ejecutar el experimento de evaluación de linealidad del sistema:

a) Establecer el intervalo lineal del método. Teniendo en cuenta las consideraciones de la Sección 3.6.1
   - Nota: una vez establezca el intervalo verificar que cumple con lo requerido por las etapas de preparación del método, la regulación y la aplicabilidad del método.

b) Establecer el número de niveles de la curva de calibración y la concentración de cada uno de estos niveles, de acuerdo con las recomendaciones del numeral 3.6.4.

c) Prepare cada uno de los niveles, tratando de evitar al máximo los errores y minimizando la incertidumbre por la preparación de estos niveles.

- Nota: considere las recomendaciones que se brindan en el numeral 3.6.4.

d) Mida cada uno de los niveles como mínimo por triplicado.
   - Nota: la medición de todos los niveles se debe realizar de manera aleatoria.
e) Evalúe la autocorrelación del error, la homocedasticidad del sistema y seleccione el modelo de regresión más adecuado de acuerdo con el numeral 3.6.2.
f) Realice el análisis exploratorio de la curva de acuerdo con las recomendaciones del numeral 3.6.3.1.
g) Realice el análisis estadístico de la curva de acuerdo con las recomendaciones del numeral 3.6.3.2.

3.7 Límite de detección

El Vocabulario Internacional de Metrología define el límite de detección como "el valor medido, obtenido mediante un procedimiento de medición dado, con una probabilidad β de declarar erróneamente la ausencia de un constituyente en un material, dada una probabilidad α de declarar erróneamente su presencia" (JCGM, 2012). Lo primero que es de resaltar de esta definición es que se trata de un valor medido, de esta manera es importante tener en cuenta que para evaluar los límites de detección no basta con realizar su estimación, sino que es indispensable realizar su confirmación mediante mediciones.

Por otro lado, en la definición se hace referencia a que es un valor medido que tiene asociadas unas probabilidades (α y β), lo cual implica que el establecimiento de este límite debe estar basado en métodos estadísticos que permitan estimar dichas probabilidades de cometer errores tipo I o tipo II (ver Sección 2.2.4.1).

De acuerdo con la IUPAC y la ISO, los límites de detección se basan en la teoría de la prueba de hipótesis y las probabilidades de falsos positivos α y falsos negativos β. En este contexto, para explicar la teoría del límite de detección, supongamos que tenemos un método analítico con precisión conocida, invariante a lo del intervalo de trabajo, y que sus resultados siguen una distribución normal. Por lo tanto, si realizamos la medición de varios blancos, se puede obtener una distribución como la presentada en la Figura 3.11.

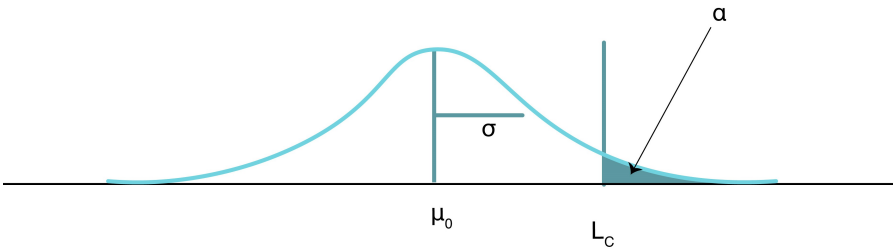


Figura 3.11: Distribución de valores alrededor de cero obtenidos para un método con precisión invariante.

Donde μ_0 corresponde al valor del blanco, σ_0 representa la desviación estándar del blanco, L_c corresponde al límite crítico y α corresponde a la probabilidad de error tipo I. La Figura 3.11 indica que si se mide un blanco, se puede obtener un resultado diferente de cero debido a los errores experimentales del proceso de medición.

Por lo tanto, es necesario establecer un límite crítico (L_c), el cual corresponde a un valor que permite diferenciar las medidas del blanco con las medidas que no son él. Este L_c permite determinar si una señal corresponde a un blanco o si la señal es de un analito, es decir, tomar una decisión posterior. Sin embargo, como se puede observar en la Figura 3.11, existe una probabilidad (sombra naranja) de que un blanco pueda dar una señal por encima de la L_c; por lo tanto, podemos concluir erróneamente que hay analito, cuando no lo hay, es decir confundir una señal del blanco con una señal del analito. La teoría de la prueba de hipótesis utiliza la siguiente definición para L_c:

$$Pr(L > L_c \mid L = 0) \leq \alpha \tag{3.21}$$

donde L representa la cantidad de interés (respuesta del analito o concentración).

Usualmente este valor de L es reemplazada por S cuando se trata de la señal instrumental del analito, y X, cuando se trata la concentración o cantidad del analito. Por otro lado, matemáticamente L_c se representa por:

$$L_c \leq K_{\alpha-1}\sigma_0 \tag{3.22}$$

donde $K_{\alpha-1}$ y α dependen de la distribución (a una cola) del blanco correspondiente a los niveles de probabilidad, $1-\alpha$. Es importante resaltar que este L_c corresponde a un valor en el cual se tiene una reducción de los errores tipo I o α, pero no los errores tipo β. Por ejemplo, si se estableciera este valor como el valor del límite de detección, la probabilidad de error tipo β es aproximadamente el 50 %.

Por lo anterior es indispensable establecer un límite de detección (LD), especificando un nivel aceptable β y de α. Por lo tanto, dicho límite debe ser mayor al límite crítico, así por la teoría de la prueba de hipótesis, se puede plantear:

$$Pr(L < L_c \mid L = LD) = \beta \tag{3.23}$$

por lo cual, el límite de detección (LD) vendría dado por la ecuación:

$$LD = K_{\alpha-1} \cdot \sigma_0 + K_{\beta-1} \cdot \sigma_D \tag{3.24}$$

donde $K_{\beta-1}$ y β, dependen de la distribución (a una cola) del límite de detección a un nivel de probabilidad, $1-\beta$. En este punto es importante resaltar que la ecuación tiene dos componentes principales, el primero es el asociado a la medición del blanco ($K_{\alpha-1} \cdot \sigma_0$) y el segundo es el componente asociado a la medición del límite ($K_{\beta-1} \cdot \sigma_D$), lo cual en la práctica representa la medición de un blanco de muestra y la muestra de rutina. Es decir, el límite de detección se debe abordar considerando si el laboratorio realiza corrección por blanco o no.

Por otro lado, asumiendo que: (i) las señales analíticas (o concentraciones) siguen una distribución normal, (ii) se comportan de manera homocedástica y (ii) los valores por defecto para β y α son 0.05, la Ecuación 3.24 se transforma en:

$$LD = 2 \cdot z_{\alpha-1} \cdot \sigma_0 \qquad \text{ó} \qquad LD = 2 \cdot z_{\beta-1} \cdot \sigma_D \tag{3.25}$$

donde z corresponde a 1.65 para una distribución normal, lo cual lleva a la conocida ecuación:

$$LD = 3.3 \cdot \sigma_0 \tag{3.26}$$

Finalmente, como puede observar las expresiones para la estimación del límite crítico (Ecuación 3.22) o límite de detección (Ecuación 3.26), son relativamente sencillas. Sin embargo, la estimación de los valores de α son el tema de mayor discusión y por el cual existen diferentes alternativas para la estimación de los límites. La Figura 3.12 muestra un diagrama en el que se muestran algunos de los métodos más empleados para la estimación de este valor, así como los casos en los que es más idóneo realizar esto.

Las siguientes secciones exponen los métodos presentados en la Figura 3.12 para la estimación de los límites de detección.

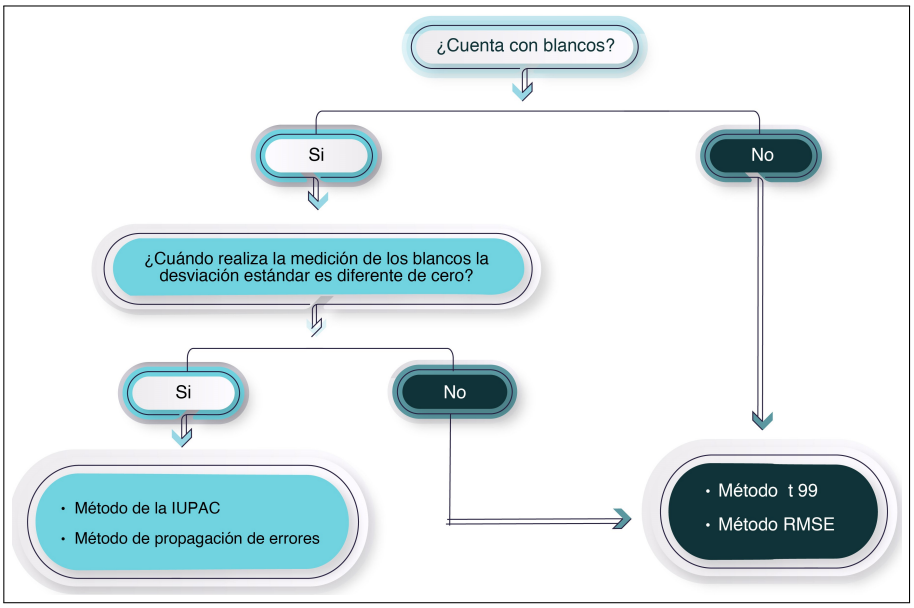


Figura 3.12: Recomendaciones para la selección del método para la estimación del límite de detección.

3.7.1 Generalidades en la estimación del LD

La estimación del límite de detección depende de diferentes factores como el instrumento de medición, las matrices o interferentes que estas puedan contener, el deterioro del sistema de medición entre otros factores. En este sentido, a continuación se presentan algunas recomendaciones que buscan asegurar la correcta estimación de estos límites:

- Diseñe un sistema de control de calidad para asegurar la correcta estimación del límite de detección. Por ejemplo: control de ruido, "*system suitability*", características de las señales, entre otros.
- Emplee herramientas que permitan evidenciar cambios en la línea base del instrumento o posibles contaminaciones, por ejemplo, blancos de reactivos, de muestras, entre otros.
- Para el caso de los métodos que requieren una curva de calibración en la cercanía del límite de detección, establezca criterios de calidad para dichas curvas, los cuales no implican necesariamente evidenciar una linealidad adecuada.
- Siempre realice los experimentos de manera aleatoria.
- En el caso en que el laboratorio realice la medición en diferentes matrices, se sugiere que se realice una agrupación de estas matrices y realice la estimación de los límites de detección por cada grupo. La guía ONAC-INM de Ensayos de Aptitud, presenta algunas propuestas de agrupación de matrices.
- Una vez estime los límites de detección verifíquelos periódicamente, para lo cual se pueden incluir en las herramientas de control de calidad del laboratorio tales como: cartas de control, evaluación de personal, verificación de los instrumentos.
- La estimación de los límites de detección se realiza usualmente en condiciones de repetibilidad, sin embargo, si el método es de uso extendido y el sistema de medición se ve afectado de manera significativa durante el tiempo se sugiere que esta estimación se realice en condiciones de precisión intermedia.
- Una vez estime los límites de detección, se sugiere que realice la corrección del sesgo o la recuperación. Similar a lo mostrado en la Ecuación 3.28.
- En la medida de lo posible, evalúe el supuesto de homocedasticidad que se planteó a lo largo de la Sección 2.2.6.3. En caso de que tenga un sistema heterocedástico ($\sigma_0 \neq \sigma_D$), emplee la Ecuación 3.24 para la estimación de los límites, la cual para una distribución normal se convierte en:

$$LD = z_{\alpha-1} \cdot \sigma_0 + z_{\beta-1} \cdot \sigma_D \tag{3.27}$$

3.7.2 Método de la IUPAC

La estimación del límite de detección, a través de este método, es quizás la más empleada por la comunidad científica y es puede decirse que es la que considera en mejor medida el procedimiento de medición y los interferentes que puedan tener las matrices. A continuación se presenta el esquema general para aplicar esta aproximación:

a) Seleccionar 10 blancos de muestra.
b) Aplicar el procedimiento de preparación de muestra a cada uno de los blancos.
c) Junto con los blancos de muestra preparar una curva de calibración de concentraciones bajas.
d) Determinar las respuestas de las soluciones de la curva de calibración en las condiciones definidas por el método.
e) Determinar la respuesta de cada uno de los blancos de muestra en las condiciones definidas por el método.
f) Calcular el promedio ($\bar{y}_{bl}$) y la desviación estándar (s_{bl}), de la respuesta de los blancos.
g) Calcular la pendiente (m) de la curva de calibración.
h) Calcular el LD así:
   - Para métodos en los que no se realiza la corrección por blanco:

$$LD = \frac{\bar{y}_{bl} + 3 \cdot s_{bl}}{m} \tag{3.28}$$

   - Para métodos en los cuales se realiza la corrección por blanco:

$$LD = \frac{3.3 \cdot s_{bl}}{m} \tag{3.29}$$

- Nota: en el caso en que el laboratorio realice la medición de las muestras y de los blancos repetidas veces de manera rutinaria y dichas determinaciones sean independientes, la estimación de la s_{bl} puede ser mejorada por esta práctica. Para más información se recomienda revisar la guía EURACHEM.

i) Fortifique unas muestras al límite de detección estimado.
j) Mida las muestras fortificadas a través del método.
k) Estime la relación señal ruido y verifique que sea como mínimo 2.5.

Ejemplo 27: Límite de detección: Enfoque de la IUPAC sin corrección de blanco.

Para los ejemplos de esta sección se utiliza el submódulo **Límite de detección** del aplicativo validaR.

Un laboratorio de ensayo de análisis de residuos de plaguicidas se encuentra realizando la validación del herbicida glifosato en gulupa por cromatografía líquida con detección de espectrometría de masas. Como parte de las actividades establecidas en el plan de validación, el personal clave debe evaluar el límite de detección del método, para lo cual se tienen 10 muestras blanco (libre de plaguicidas) y una curva de calibración en el intervalo de concentración entre 30 μg/L y 70 μg/L. A continuación se presentan los resultados de la medición para curva de calibración y los 10 blancos de muestra.

Curva de calibración

Conc. glifosato /[μg/L]	**Respuesta** /[cuentas*s]
30	4188447
40	8545791
50	15897083
60	21350029
70	26136852
80	31076354

Respuestas de los blancos

Muestra	**Respuesta** /[cuentas*s]
Blanco 1	1109
Blanco 2	1145
Blanco 3	1193
Blanco 4	1177
Blanco 5	1064

Muestra	**Respuesta** /[cuentas*s]
Blanco 6	1137
Blanco 7	1164
Blanco 8	1108
Blanco 9	1066
Blanco 10	1118

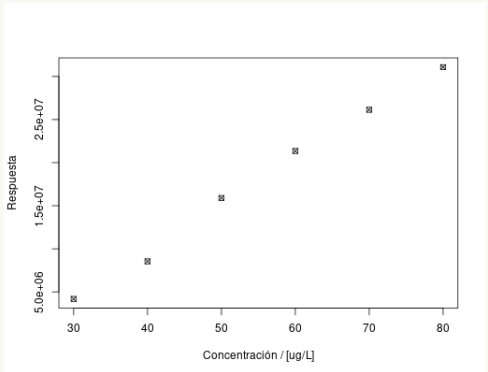
El recuadro que se muestra a continuación contiene los resultados de validaR para la estimación del *LD* usando el enfoque de la IUPAC.

Resultados

Estimación del límite de detección utilizando el enfoque de la IUPAC basado en blancos de muestra:

El límite de detección es **0.0023106 [ug/L]**
La pendiente de la curva de calibración fue 550473 / [ug/L]

Representación gráfica del límite de detección



Conclusión: El límite de detección para la determinación de glifosato en gulupa por LC-MS es de 0.002 μg/L. Teniendo en cuenta el procedimiento de extracción y cuantificación es necesario aplicar los factores necesarios para expresar el límite de detección en μg de glifosato/kg de gulupa. Aplicando esto, el *LD* para la determinación de glifosato en gulupa por LC-MS es de 0.02 μg/kg.

Recomendación: Una vez se confirme el valor del *LD*, acorde con los lineamientos de los ítems g a h de este numeral, tener en cuenta que el *LD* estimado por el método IUPAC se encuentra muy por debajo del nivel 1 de la curva de calibración empleada para el cálculo; por lo que se recomienda evaluar y de ser necesario iniciar la curva de calibración en un nivel que este por debajo del planteado en este ejemplo.

Ejemplo 28: Límite de detección: Enfoque de la IUPAC con corrección de blanco.

Un laboratorio de análisis de contaminantes en cosméticos se encuentra validando la metodología para la determinación de cadmio en labial por GF-AAS. Para esto se realizó la preparación y medición de una curva de calibración en el intervalo de concentración comprendido entre 10 μg/L a 35 μg/L. Adicionalmente se realizó la medición de 12 muestras definidas como blanco. En esta medición se realiza corrección de la respuesta de la muestra con la respuesta del blanco. A continuación se presentan los resultados de la medición de la curva de calibración y la medición de blancos de muestra.

Curva de calibración

Conc. cadmio /[μg/L]	**Respuesta** /[UA*s]
10.0	0.121
15.0	0.195
21.0	0.262
25.0	0.348
30.0	0.429
35.0	0.503

Respuestas de los blancos

Muestra	**Respuesta** /[UA*s]
Blanco 1	0.0897
Blanco 2	0.0345
Blanco 3	-0.0445
Blanco 4	-0.0299
Blanco 5	0.0747
Blanco 6	-0.0573

Muestra	**Respuesta** /[UA*s]
Blanco 7	0.0263
Blanco 8	-0.0441
Blanco 9	-0.0391
Blanco 10	0.0392
Blanco 11	0.0724
Blanco 12	0.0285

El siguiente recuadro muestra los resultados de validaR para la estimación del *LD* usando el enfoque de la IUPAC.

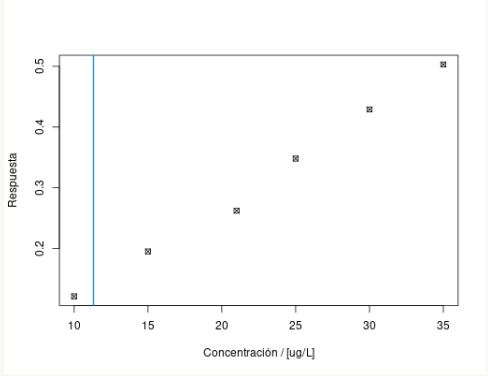
Resultados

Estimación del límite de detección utilizando el enfoque de la IUPAC basado en blancos de muestra:

El límite de detección es **11.307 [ug/L]**

La pendiente de la curva de calibración fue 0.0154554 / [ug/L]

Representación gráfica del límite de detección



La línea vertical azul indica el límite de detección estimado.

Conclusión: El límite de detección para la determinación de cadmio en labial por GF-AAS es de 11.31 μg/L. Si se tiene en cuenta los factores relacionados con la digestión del material, al aplicar la conversión (factor de 20) se tiene que el límite de detección para la determinación de cadmio en labial es de 226.2 μg de Cd /kg de labial.

Recomendación: Una vez se confirme el valor del *LD*, acorde con los lineamientos de los incisos g a h de este numeral, tener en cuenta que el *LD* estimado por el método IUPAC se encuentra por encima del nivel 1 de la curva de calibración empleada para el cálculo; por lo que se recomienda que la curva de calibración inicie arriba del valor de *LD*.

3.7.3 RMSE: desviación estándar del intercepto

El intercepto de una curva de calibración representa la respuesta que se obtendría cuando el valor de la variable independiente es cero, es decir cuando la concentración es igual a cero. El intercepto representa el valor del ruido o del blanco (μ_0 de la Figura 3.11) y la desviación estándar de este representaría σ_0 (ver ecuación 3.23). En este

contexto, la EPA propuso el método conocido como RMSE (de sus siglas en inglés *Root Mean Square Error*), el cual consiste en estimar σ_0 a través una curva de calibración.

A continuación se presenta el esquema general para aplicar esta aproximación:

a) Preparar por lo menos cuatro curvas de calibración de mínimo cuatro niveles en un intervalo bajo de concentración (cercano al límite de detección).
b) Medir cada una de las curvas.
c) Estimar los parámetros del modelo de regresión (pendiente e intercepto).
d) Con los valores de concentración de cada uno de los niveles y el modelo de regresión, calcule el valor de respuesta estimado ($y_{estimados}$) para cada punto de la curva.
e) Estime cada uno de los residuales (E_j).
f) Estime el RMSE, mediante la siguiente ecuación:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} E_j^2}{n-2}} \tag{3.30}$$

g) Estime el límite de detección (LD) mediante la siguiente ecuación:

$$LD = \frac{3.3 \cdot RMSE}{m} \tag{3.31}$$

- Nota: para métodos en los que el intercepto (i) es estadísticamente diferente de cero (evaluación de linealidad, Sección 3.6.1) se debe considerar el intercepto de la siguiente manera:

$$LD = \frac{i + 3.3 \cdot RMSE}{m} \tag{3.32}$$

h) Fortifique unas muestras al límite de detección estimado.
i) Mida las muestras fortificadas a través del método.
j) Estime la relación señal ruido y verifique que sea como mínimo 2.5.

Ejemplo 29: Límite de detección: Enfoque RMSE, desviación estándar del intercepto.

Un laboratorio nacional de referencia experto en el análisis de contaminantes en aguas se encuentra validando la metodología para la determinación de cadmio por ICP-MS. En la planeación de la validación se plantea la evaluación del LD mediante el método RMSE. Para ello, se prepararon 4 curvas de calibración, cada una con 5 niveles de concentraciones entre 0.25 μg/L y 1.25 μg/L. A continuación se presentan los resultados de medición para cada una de las curvas preparadas.

Curva 1:

Cd / (μg/kg)	**Señal** (I_{Cd}/I_{Rh})
0.255	0.1829
0.500	0.3290
0.756	0.4896
1.000	0.6565
1.250	0.8017

Curva 2:

Cd / (μg/kg)	**Señal** (I_{Cd}/I_{Rh})
0.248	0.1896
0.506	0.3259
0.748	0.5035
0.994	0.6692
1.260	0.8346

Curva 3:

Cd / (μg/kg)	**Señal** (I_{Cd}/I_{Rh})
0.250	0.1848
0.482	0.3230
0.751	0.4988
1.020	0.6582
1.250	0.8408

Curva 4:

Cd / (μg/kg)	**Señal** (I_{Cd}/I_{Rh})
0.253	0.1855
0.490	0.3326
0.743	0.4867
1.070	0.6546
1.270	0.8284

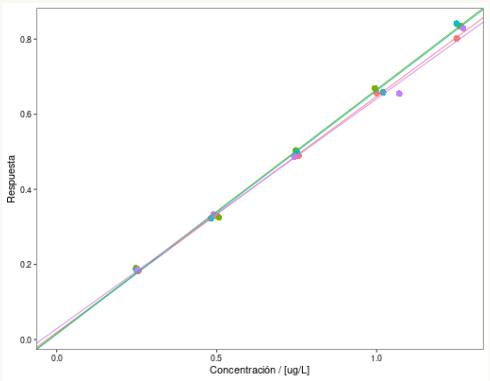
Los resultados del aplicativo validaR se muestran en el recuadro que aparece a continuación.

Resultados

Estimación del límite de detección utilizando el enfoque del RMSE:

- El límite de detección es **0.06403 [ug/L]**
- Valor RMSE: 0.01232 (mismas unidades de las señales de las curvas de calibración)
- Los interceptos de las curvas de calibración no son estadísticamente significativos

Representación gráfica de las curvas de calibración



Conclusión: El límite de detección para la determinación de cadmio en agua potable por ICP-MS es de 0.06 μg/L.

Recomendación: Una vez se confirme el valor del LD, acorde con los lineamientos de los ítems g a h de este numeral, tener en cuenta que el LD estimado por el método RMSE se encuentra muy por debajo del nivel 1 de las curvas de calibración empleadas para el cálculo; por lo que se recomienda evaluar y de ser necesario iniciar la curva de calibración en un nivel que esté por debajo del primer nivel planteado en este ejemplo.

3.7.4 Método de la propagación de errores

El método de propagación de errores propone estimar el σ_D de la Ecuación 3.24 mediante la estimación de la desviación estándar del intercepto (s_i), similar a la propuesta del método RMSE. Sin embargo, en esta aproximación no sólo se hace uso de la información obtenida desde el modelo de regresión, sino que se hace uso del valor de la desviación estándar del blanco s_{bl} (método de la IUPAC). El esquema general para este método corresponde a:

a) Seleccionar 10 blancos de muestra.
b) Aplicar el procedimiento de medición a cada uno de los blancos mediante el método.
c) Junto con los blancos, preparar cuatro curvas de calibración de concentraciones bajas.
d) Determinar las respuestas de las soluciones de la curva de calibración en las condiciones definidas por el método.
e) Estimar los parámetros del modelo de regresión (pendiente e intercepto).
f) Estimar la desviación estándar de las mediciones de los diferentes blancos (s_{bl}).
g) Estimar el promedio de las pendientes ($\overline{b_1}$).
h) Estimar la desviación estándar de los interceptos (s_i).
i) Estimar el límite de detección mediante la siguiente ecuación:

$$LD = \frac{3.3 \cdot \sqrt{s_{bl} + s_i}}{\overline{b_1}} \tag{3.33}$$

- Nota: Este método hace uso de los dos componentes de la ecuación 3.18, es decir contiene el componente del blanco y de la desviación estándar del límite. Por lo anterior, es una modificación del método de la IUPAC, que en el caso en que la desviación del intercepto sea mucho menor a la desviación estándar de los blancos la anterior ecuación se simplifica al método de la IUPAC.

j) Fortifique unas muestras al límite de detección estimado.
k) Mida las muestras fortificadas a través del método.
l) Estime la relación señal ruido y verifique que sea como mínimo 2.5.

Ejemplo 30: Límite de detección: Método de la propagación de errores.

Un laboratorio de análisis de contaminantes en alimentos se encuentra validando la metodología para la determinación de arsénico en material foliar de cannabis sativa por la técnica ICP-MS. En la planeación de la validación se plantea la evaluación del LD mediante el método de propagación de errores, para lo cual ha realizado la digestión de 10 muestras blanco de muestra (material foliar cultivado bajo condiciones controladas); y se han preparado cuatro curvas de calibración en concentraciones de 0.2 μg/kg a 1.0 μg/kg. A continuación, se presentan los resultados obtenidos en la medición de las curvas de calibración y los blancos de muestra.

Curva 1:

As / (μg/kg)	Señal (I_{As}/I_{Rh})
0.198	0.5420
0.399	0.8944
0.601	1.3832
0.795	1.9136
1.100	2.7028

Curva 2:

As / (μg/kg)	Señal (I_{As}/I_{Rh})
0.200	0.5506
0.402	0.8966
0.597	1.3824
0.799	1.9849
1.090	2.6824

Curva 3:

As / (μg/kg)	Señal (I_{As}/I_{Rh})
0.197	0.5313
0.402	0.8848
0.604	1.3289
0.805	1.9958
1.110	2.7857

Curva 4:

As / (μg/kg)	Señal (I_{As}/I_{Rh})
0.199	0.5601
0.400	0.9132
0.601	1.3965
0.800	1.9721
1.090	2.7292

Respuestas de los blancos:

Muestra	Señal (I_{As}/I_{Rh})
Blanco de muestra 1	0.1729
Blanco de muestra 2	0.1771
Blanco de muestra 3	0.1786
Blanco de muestra 4	0.1833
Blanco de muestra 5	0.1826

Muestra	Señal (I_{As}/I_{Rh})
Blanco de muestra 6	0.1780
Blanco de muestra 7	0.1831
Blanco de muestra 8	0.1729
Blanco de muestra 9	0.1708
Blanco de muestra 10	0.1763

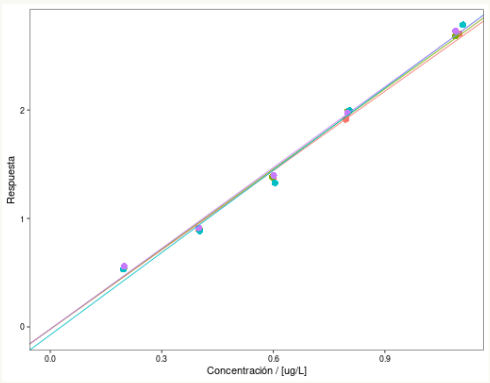
Los resultados del aplicativo validaR se muestran en el recuadro que aparece a continuación.

Resultados

Estimación del límite de detección utilizando el enfoque de la propagación de errores:

- El límite de detección es **0.06634 [ug/L]**
- Desviación estándar de los interceptos de regresión: 0.02646 (mismas unidades de las señales de las curvas de calibración)
- Desviación estándar de las señales de los blancos: 0.004488 (mismas unidades de las señales de las curvas de calibración)
- Promedio de las pendientes de regresión: 2.478 / [ug/L]

Representación gráfica de las curvas de calibración



Conclusión: El límite de detección para la determinación de As en material foliar de cannabis sativa es de 0.066 μg/L; pero realizando la conversión a unidades de μg de As por kg de material foliar se tiene un límite de detección de 1.32 μg/kg.
NOTA: Se debe realizar la conversión con base en el procedimiento (diluciones, digestiones, extracciones) para expresar el LD en las unidades del intervalo de trabajo.

Recomendación: Una vez se confirme el valor del LD, acorde con los lineamientos de los incisos g a h de este numeral, tener en cuenta que el LD estimado por el método de propagación del error se encuentra muy por debajo del nivel 1 de las curvas de calibración empleadas para el cálculo; por lo que se recomienda evaluar y de ser necesario iniciar la curva de calibración en un nivel que esté por debajo del primer nivel planteado en este ejemplo.

3.7.5 Método t_{99}

En el caso en que no sea posible contar con blancos para la respectiva desviación estándar de estos o en el caso en que la estimación de dicha desviación estándar corresponda a cero, se recomienda hacer uso del método conocido como EPA 40 CFR PART 136 o t99 (EPA) En este método, al igual que en el método de la IUPAC, se asume que los resultados del instrumento o las concentraciones son homocedásticos ($\sigma_0 = \sigma_D$), por lo cual la ecuación 3.19 se convierte en la siguiente ecuación:

$$LD = 3.3 \cdot \sigma_{LD} \tag{3.34}$$

donde la σ_{LD} corresponde a la desviación estándar del límite de detección, en lugar de la desviación estándar del blanco (σ_0). En este sentido, el método **t_{99}** se basa en la estimación del límite a partir de muestras que idealmente se encuentren en el límite de detección o cercanas a este.

a) Estime la concentración un nivel de concentración cercano al límite de detección (N_f).
   - Nota 1: puede realizar una estimación a partir de la relación señal ruido que se obtiene en una curva de calibración previa (esperando una relación ruido > 3).
   - Nota 2: puede corresponder el límite inferior del intervalo de trabajo que usted definió (límite de cuantificación).

b) Seleccione una muestra que se encuentre cercana a este límite (N_f).
   - Nota: puede realizar fortificación de una muestra y luego corregir por la concentración de la muestra de acuerdo con lo ilustrado en la ecuación 3.7.

c) Mida por lo menos 7 veces la muestra.

d) Cuantifique cada muestra.

e) Estime la desviación estándar (sLD).

f) Ubique el t99 a n-1 grados de libertad y a una cola, donde n corresponde al número de muestras que midió.

g) Estime el límite de detección de acuerdo con la siguiente ecuación:

$$LD = t_{0.99,\ n-1} \cdot s_{LD} \tag{3.35}$$

h) Puede corregir el límite de detección por el porcentaje de recuperación ($\%\ R$), mediante la siguiente ecuación:

$$LD = \frac{t_{0.99,\ n-1} \cdot s_{LD}}{\%\ R} \cdot 100\,\% \tag{3.36}$$

i) Fortifique unas muestras al límite de detección estimado.

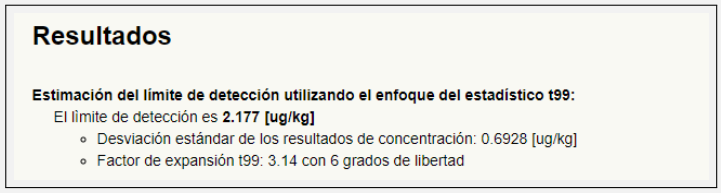
j) Mida las muestras fortificadas a través del método.

k) Estime la relación señal ruido y verifique que sea como mínimo 2.5.

Ejemplo 31: Límite de detección: Método de la EPA (t_{99}).

Un laboratorio de ensayo de análisis de residuos de plaguicidas se encuentra realizando la validación de clorotalonil en miel de abejas por cromatografía de gases con detector de micro captura de electrones (GC-μECD). Como parte de las actividades establecidas en el plan de validación, el personal clave debe evaluar el límite de detección del método mediante la aplicación del estadístico t_{99}. De esta manera, realizaron la medición de 7 muestras fortificadas en una concentración de 8.0 μg/kg. Se definió esta concentración dado que permite una relación señal/ruido (S/N) de 3.2 en el sistema de medición; además, se conoce del método que tiene un porcentaje de recuperación del 85 %. Las concentraciones obtenidas por la cuantificación con un nivel de calibración se muestran en la tabla de abajo a la derecha. Los resultados de *LD* se muestran en el recuadro que aparece abajo a la izquierda.

Muestra	[Clorotalonil] /(μg/kg)
Muestra 1	8.5
Muestra 2	9.4
Muestra 3	8.3
Muestra 4	8.0
Muestra 5	9.2
Muestra 6	7.9
Muestra 7	9.6



El límite de detección de este método es de 2.177 μg/kg. La corrección de este valor por el porcentaje de recuperación de 85 % usando la Ecuación 3.36 es:

$$LD = \frac{2.717\ \mu\text{g/kg}}{85\,\%} \cdot 100\,\% = 2.561\ \mu\text{g/kg}$$

Conclusión: El límite de detección para la determinación de clorotalonil en miel de abejas es de 2.561 μg/kg
Recomendación: Confirmar el valor del LD, acorde con los lineamientos de los ítems g a h de este numeral.

3.8 Robustez

Las pruebas de robustez son bien conocidas actualmente en el sector farmacéutico, posiblemente debido a la amplia experiencia de este sector en la validación de métodos y la necesidad de transferencia de métodos entre los diferentes laboratorios. Por lo cual, la mayoría de las definiciones y metodologías existentes, por ejemplo, las del ICH, se pueden encontrar en ese campo. Sin embargo, esto no debería tener implicaciones para la aplicación de la prueba de robustez en otros campos.

La robustez de un método analítico es la capacidad de poder reproducir dicho método en diferentes circunstancias, sin que se presenten diferencias inesperadas o significativas en los resultados obtenidos. La evaluación de la robustez del método indica el efecto de las variaciones sobre la precisión y veracidad del método en función de variables propias del método; por lo cual este parámetro al igual que el parámetro de selectividad se encuentran íntimamente relacionados con la exactitud del método.

El resultado de la robustez, por considerarse dentro de la precisión del método analítico, indica que tanto deben ser controlados los parámetros del método. Detectar los factores que influyen considerablemente pueden conllevar a un método más reproducible, una mejor estimación de incertidumbre, un mejor control de las variables de influencia, entre otras ventajas.

3.8.1 Definiciones de robustez

El término de robustez fue introducido en química analítica por Youden y Steiner (Youden y Steiner, 1975). Estos autores encontraron en los resultados de los estudios interlaboratorio que siempre se tenían variaciones asociadas a la implementación dentro de cada laboratorio, aunque los laboratorios emplearán el mismo método. Por esto propusieron realizar un experimento en donde se pudiera evaluar la influencia de algunos factores del método sobre los resultados de medición. Posteriormente, se comenzaron a popularizar los experimentos de robustez, los cuales tradicionalmente se realizaron a través de diseños experimentales, especialmente factoriales completos y fraccionados.

Hoy en día la robustez se conoce como la "*medida de la capacidad de un método analítico de no ser afectado considerablemente por variaciones pequeñas pero deliberadas de los principales parámetros del método analítico*" (Thompson, Ellison y Wood, 2002; USP 36, 2009). Por otro lado, el término *solidez* (en inglés *ruggedness*) se

utiliza con frecuencia como sinónimo, sin embargo, este término se encuentra relacionado en la actualidad con el concepto de reproducibilidad que se encuentra en el VIM (Mulholland, 1988; van Leeuwen y col., 1991); desafortunadamente hoy en día, aún persisten confusiones en este parámetro, por lo cual, en ocasiones, se asocian estudios colaborativos a ensayos de robustez.

Por lo anterior, es importante señalar que este parámetro de validación hace parte del conjunto de pruebas que usualmente se realizan al interior de cada laboratorio, es decir bajo la denominada validación "*in house*". Por lo anterior, organizaciones como la AOAC han definido la prueba de robustez como un "*Estudio dentro del laboratorio para evaluar el comportamiento de un proceso analítico cuando se efectúan pequeños cambios en las condiciones ambientales o de operación, semejantes a aquéllos que pudieran surgir en los diferentes ambientes de la prueba*"

En la actualidad no existe un consenso acerca de un procedimiento específico para evaluar la robustez o para examinar los parámetros del método analítico. Sin embargo, se han propuesto unas etapas generales que son presentadas a continuación.

3.8.2 Esquema general para la evaluación de robustez de un método analítico

Considerando que la prueba de robustez buscar evaluar los diferentes factores que pueden influir sobre los resultados de medición, usualmente en el estudio de robustez, se seleccionan varios factores del procedimiento de medición, los cuales se evalúan en un intervalo que se pueden esperar ocurra en las condiciones normales de operación del método de manera rutinaria. Posteriormente, se analiza la información obtenida y se procede a determinar los factores que podrían perjudicar el rendimiento del método. Lo anterior permite:

- Mejorar un mejor sistema de control de variables.
- Mejorar la exactitud del método.
- Identificar las variables de influencia del método y considerarlas en el presupuesto de incertidumbre.

En este contexto, la evaluación de la robustez requiere siempre los siguientes pasos (Vander Heyden y Massart, 1996):

- Selección e identificación de los factores operacionales y/o condiciones ambientales para ser evaluados.
- Seleccionar los niveles de los factores. En robustez generalmente se evalúan 2 ó 3 niveles para cada factor.
- Seleccionar el diseño de experimento adecuado, acorde con el número de variables y el tipo de variable.
- Ejecución del diseño de experimentos. Esto se denomina evaluación de la robustez.
- Cálculo de los efectos de los factores sobre la(s) respuesta(s) del método, para determinar cuáles factores tienen efectos experimentalmente relevantes.
- Análisis de los resultados. En esta parte se deben identificar los efectos estadísticamente significativos.
- Establecer las conclusiones relevantes.

Algunos de estos pasos se explican con mayor detalle a continuación.

3.8.2.1 Selección de los factores para la evaluación de la robustez

Los factores en el ensayo de robustez se deben seleccionar a partir de la descripción del método analítico o a partir de las condiciones ambientales que no están especificadas explícitamente dentro del método analítico, sino que pueden ocurrir en un momento dado (de acuerdo con la experiencia del laboratorio). Los factores que se seleccionen pueden ser cuantitativos (continuos), cualitativos (discretos) o mixtos, por ejemplo, en un método analítico de determinación de mercurio por absorción atómica algunos factores pueden corresponder a la concentración de borohidruro de sodio, la marca de la lámpara a emplear, el flujo de la muestra, entre otros.

Los factores que se deben evaluar son aquellos que representan cambios potencialmente significativos en la respuesta del método analítico, siempre deben seleccionarse reflexionando que durante la vigencia del método en el laboratorio el factor puede cambiar o ha cambiado (experiencia previa). Se debe evitar la evaluación de factores que en la práctica es poco probable que sucedan, por ejemplo, el cambio de la temperatura en un detector de cromatografía de gases, el cambio de un tipo de inyector, entre otros.

3.8.2.2 Selección de los niveles de los factores para la evaluación de la robustez

Una vez se seleccionan los factores, para factores de tipo cuantitativo, se recomienda seleccionar tres niveles (los extremos y el nominal) si no se puede excluir un comportamiento no lineal de la respuesta en función del cambio de los factores. Sin embargo, usualmente se asume que el factor tiene la misma influencia si este aumenta o si este disminuye, respecto al valor nominal del método.[3] Por lo cual, durante la fase de experimentación se suelen escoger sólo dos niveles.

Ejemplo 32: Robustez: Selección de niveles para la evaluación de la robustez.

En un procedimiento de medición por HPLC, se identifica que el valor de pH de una fase móvil es uno de los factores de influencia. El valor nominal de este corresponde a 4.6, por lo cual en un experimento ideal se deberían evaluar por lo menos tres diferentes valores de pH: el nominal, uno superior y uno inferior; lo anterior suponiendo que en la práctica la fase móvil puede quedar preparada en un valor mayor o menor de pH. Sin embargo, de acuerdo con la experiencia (registro de control de pH que se tiene durante la preparación de la fase móvil), se encuentra que esta fase móvil siempre se prepara en un valor igual o mayor a 4.6 y que el valor de pH más alto que se ha preparado corresponde a 4.9. Por lo anterior, se decide seleccionar sólo dos niveles del factor (pH) con valores de 4.6 y 4.9.

Como se puede observar en el ejemplo anterior, siempre se debe buscar que los niveles de los factores se deben seleccionar que representen la máxima diferencia entre los valores de los factores, sin embargo, siempre debe primar la lógica o la experiencia, es decir siempre se deben evaluar niveles que en la rutina del laboratorio puedan ocurrir. No es aconsejable seleccionar niveles muy alejados uno del otro, porque si no se conoce el efecto del factor con anterioridad se puede introducir al diseño experimental la posibilidad de encontrar efectos significativos que no son relevantes para la evaluación de la robustez.

Por otro lado, cuando se identifican factores de tipo cualitativo se debe revisar con precaución si es posible realizar los experimentos, pues en muchas ocasiones se tiene que estos factores no son controlables por el laboratorio sino por un tercero, por ejemplo: lote de un reactivo, columna, entre otros. De igual manera, como se podrá observar en la siguiente sección, es posible simplificar enormemente la experimentación a realizar a través de la selección adecuada de un diseño de experimentos, sin embargo, la ejecución de estos diseños se imposibilita en el momento en que se incluyen algunos factores de tipo cualitativo.

Finalmente, una vez se seleccionen los factores y los niveles, se tiene que revisar que en la práctica se pueden controlar dichos factores en sus correspondientes niveles, antes de proceder a seleccionar el respectivo diseño de experimentos. Por ejemplo, experimentos que impliquen temperaturas superiores a las condiciones ambientales, típicamente son más fáciles de controlar que aquellos que se encuentran algunos grados Celsius por debajo o por encima de la temperatura ambiental.

3.8.2.3 Selección del diseño de experimentos para la evaluación de la robustez

Desde una perspectiva minimalista y asumiendo que todos los factores son variables independientes, la fase experimental más elemental consideraría la variación a cada uno de los factores independientes, lo cual permitirá establecer el grado de influencia sobre los resultados de medición para cada factor de una manera directa. Sin embargo, esta práctica resulta laboriosa, costosa y omite la sinergia que puede ocurrir entre dos o más factores. Por lo anterior, se recomienda que siempre se empleen diseños de experimentos que consideren estos factores, o por lo menos desde la planificación del experimento se consideren los aspectos a despreciar.

La tendencia en estudios de robustez es el empleo de los denominados diseños de "*screening*", los cuales permiten evaluar el efecto de varios factores en muy pocos experimentos (NIST/SEMATECH, 2022). De manera general, los diseños más empleados corresponden a los factoriales fraccionados y a los diseños de Plackett-Burman.

Los diseños de Plackett-Burman son más fáciles de construir que los diseños factoriales y existe una gran cantidad de literatura al respecto, así como diversidad de ejemplos. Sin embargo, los diseños factoriales fraccionados

[3] El valor nominal es aquel que está descrito en el procedimiento o aquel que es más probable que ocurra en el caso que no esté especificado en el procedimiento analítico.

son una gran alternativa y en muchas ocasiones resultan más adecuados para la evaluación de diversos factores. Por lo anterior, dentro del aplicativo validaR se ha incluido una herramienta que permite construir los diseños experimentales más frecuentemente empleados en los estudios de robustez.

Los diseños factoriales son diseños en los que se eligen adecuadamente una parte o fracción de los tratamientos de un factorial completa (2^n), con la intención de estudiar el efecto de los factores utilizando menos corridas experimentales. Por ejemplo, si se trata de un diseño factorial completo en el cual se desean evaluar 5 factores (2^5) se deben desarrollar 32 puntos experimentales,[4] lo cual, en la mayoría de los casos, lo convierte en un diseño inviable o muy costoso. Por lo anterior en los estudios de robustez, se prefiere trabajar con diseños fraccionados; los cuales para el mismo ejemplo lograr reducir los puntos experimentales considerablemente, por ejemplo, a 16 y 8.

Los dos tipos de diseños tiene la opción de trabajar con dos versiones, la primera corresponde al diseño minimalista que permite identificar los efectos de los diferentes factores, la segunda opción requiere un mayor número de experimentos,[5] sin embargo, permite obtener mejores conclusiones acerca de los efectos de los factores evaluados, es decir evalúan en mejor medida la influencia de los factores sobre los resultados de medición.

La Tabla 3.10 presenta una relación del número de factores a evaluar y los diseños experimentales sugeridos.

Número de factores	Diseños recomendados	Número de puntos experimentales
1	Completamente al azar	2*
2	Factorial completo 2^2	4*
3	Factorial completo 2^3	8
	Factorial fraccionado 2^{3-1}	4*
4	Factorial completo 2^4	16
	Factorial fraccionado 2^{4-1}	8
5	Factorial completo 2^5	32
	Factorial fraccionado 2^{5-1}	16
	Plackett-Burman	12
	Factorial fraccionado 2^{5-2}	8
6	Plackett-Burman	12
	Factorial fraccionado 2^{6-3}	8
	Plackett-Burman	12
7	Factorial fraccionado 2^{7-3}	16
	Plackett-Burman	12
	Factorial fraccionado 2^{7-4} (Youden-Steiner)	8*

*En algunos diseños es necesario hacer réplicas de los puntos experimentales para evaluar la significancia estadística de las variables.

Tabla 3.10: Diseños de experimentos sugeridos para el ensayo de robustez.

3.8.2.4 Ejecución del diseño de experimentos

Previo a la ejecución del experimento se deben realizar ensayos o verificar que cada una de las condiciones establecidas en la fase de planeación del diseño experimental se pueden realizar en la práctica, en especial debido a que se desean evaluar condiciones que pueden cambiar en la rutina. Es decir, suponga que en la rutina un medio isotérmico que se emplea para realizar una derivatización opera a $(40 \pm 1)\ ^oC$ (Condición base), y la temperatura

[4] Los puntos experimentales son el conjunto particular de condiciones experimentales que deben configurarse a una unidad experimental dentro de los confines del diseño seleccionado.

[5] Las réplicas en diseño de experimentos es correr nuevamente todo el diseño. Es decir, si se tenían 8 puntos experimentales y se desea realizar una réplica se debe realizar un total de 16 experimentos.

que se desea evaluar corresponde a 37 ^{o}C, con el mismo baño, por lo cual debe verificar que puede controlar el baño a por lo menos (37 $\pm$ 0.5) ^{o}C y (40 $\pm$ 0.5) ^{o}C. En caso de que no pueda realizar esto en la práctica, es recomendable que cambien la temperatura, por ejemplo, a 35 ^{o}C, de lo contrario es posible que tenga una conclusión errónea.

Por otro lado, durante la ejecución de los experimentos es importante que se controlen muy bien los factores que se incluyeron en el diseño factorial y en general todos los factores del método. Por ejemplo, de acuerdo con la experiencia del laboratorio se conoce que la humedad del laboratorio puede estar entre el 10 % y el 90 % y los resultados de medición no se ven afectados; por lo cual el laboratorio decide no incluir dicho factor en el ensayo de robustez. Sin embargo, que no se incluya el factor no quiere decir que el factor no genere ningún ruido en el experimento de robustez, por lo cual el laboratorio debe asegurar que dicha variable no tenga una gran variación, por ejemplo, ejecutar el experimento en horas de la tarde, donde de acuerdo con los registros del laboratorio, la humedad siempre se encuentra en el intervalo de 10 % a 90 % y tiene la menor variación del día, por ejemplo (10 %).

De otra parte, es importante que previo a la ejecución del experimento establezca la respuesta con la cual va a evaluar los diferentes factores, es decir la variable con la cual va a realizar el análisis estadístico. La respuesta, debe presentar en lo posible una relación directa con el resultado analítico y en lo posible debe ser de fácil interpretación. Asimismo, esta respuesta no debe ser afectada por otras variables que no se encuentren en el diseño. Por ejemplo, si un laboratorio selecciona la respuesta instrumental como la variable respuesta, se debe asegurar que esta respuesta no importe otros efectos, como la deriva del instrumento. En su lugar, si el laboratorio es consciente de que tiene un efecto de deriva, puede emplear la concentración en lugar de la respuesta instrumental e implementar un esquema de calibración analítica que le permita reducir dicho efecto. Finalmente, es importante realizar una aleatorización de todos los experimentos de tal manera que se asegure que otros tipos de error, en especial los sistemáticos, no afecten la evaluación de los datos y la posterior conclusión.

Ejemplo 33: Robustez: Creación de la matriz de diseño.

Suponga que en un proceso de digestión de muestras en el laboratorio desea evaluar dos factores: temperatura de digestión y concentración del ácido. De acuerdo con la experiencia del laboratorio, la temperatura del sistema de digestión usualmente puede ser un poco más alta de la indicada en el método (80 ^{o}C), lo cual depende del tipo de muestra y la cantidad, por lo cual para este factor se selecciona una temperatura superior (85 ^{o}C) a la del método. Por su parte, el método indica que el ácido debe estar en una concentración del 63 %, sin embargo, el laboratorio al subdestilar sus propios ácidos, encuentra que el ácido obtenido se suele encontrar entre estos ácidos se suele encontrar entre el 60 % y 65 %, por lo cual se decide evaluar este factor al 60 %.

En este contexto, para realizar el ensayo de robustez se decide seleccionar un diseño factorial completo con dos réplicas. El diseño se puede crear en el módulo **Robustez** del aplicativo validaR. El siguiente recuadro muestra la información que se debe ingresar al aplicativo.

1. Escoja una acción:

(●) Diseñar una matriz de experimentos nueva para evaluar robustez del método
( ) Cargar una matriz de experimentos creada anteriormente en el aplicativo

2. Indique cuantas variables considerará para el estudio de robustez (hasta siete variables).

Número de variables:* 2

3. (Opcional) llene la siguiente tabla con los nombres de las variables y sus unidades de medición respectivas.

	Variable.1	Variable.2
Nombre:	Temperatura	Concentración
Unidades:	ºC	%

4. (Opcional) llene la siguiente tabla con los valores alto/bajo que tomará cada variable.

	Variable.1	Variable.2
Nivel.Bajo	80.00	60.00
Nivel.Alto	85.00	65.00

- Para variables cualitativas se recomienda utilizar -1 y 1, respectivamente.
- Para variables cuantitativas se recomienda utilizar niveles que se distancien simétricamente del valor nominal declarado en el procedimiento de medición.

5. Escoja una opción de diseño experimental que le parezca conveniente:

(●) 4 puntos experimentales: Diseño factorial completo

6. Indique las siguientes opciones de la matriz de experimentos:

Número de réplicas por cada punto experimental:* 2

Numero de semilla para generar el orden aleatorio: 958

Aleatorizar orden de los experimentos [on]

7. Presione el siguiente botón para generar una matriz de diseño experimental:

Proponer diseño experimental para evaluar la robustez

El recuadro que aparece a la izquierda muestra la matriz de experimentos que propone el aplicativo validaR. Note que se seleccionaron los niveles de los factores de acuerdo con la experiencia del laboratorio, por lo cual en un caso se seleccionó un nivel superior y en otro un nivel inferior, respecto a la temperatura indicada en el método. En la matriz de experimentos se puede observar que se aleatorizó el orden de los 8 experimentos que se deben realizar.

	Punto.exp	Replica	Temperatura	Concentración	Resultado
1	No.2	1	85	60	
2	No.3	1	80	65	
3	No.4	1	85	65	
4	No.1	1	80	60	
5	No.3	2	80	65	
6	No.4	2	85	65	
7	No.1	2	80	60	
8	No.2	2	85	60	

Descargar información del diseño experimental

3.8.2.5 Análisis y evaluación de los efectos de las variables

Existen diferentes herramientas que permiten realizar un análisis visual de los resultados de los diseños experimentales. Se destacan el gráfico de efectos principales, el diagrama de Pareto, el gráfico de probabilidad normal de los efectos estimados y los gráficos de interacción. Estas herramientas permiten visualizar los posibles efectos de los factores sobre las variables, así como su relación (inversa o directa). Sin embargo, es recomendable complementar los análisis gráficos con un análisis de varianza (ANOVA, ver Sección 2.2.7), el cual permite estimar la significancia estadística de los efectos de cada una de los factores sobre la variable respuesta seleccionada.

Dependiendo de la complejidad del diseño experimental seleccionado, la evaluación de los resultados obtenidos de los diferentes diseños experimentales se puede realizar a través de diferentes programas especializados y aunque es posible realizar hojas de cálculo que determinen los efectos no se recomienda esta práctica, debido a la complejidad de dichos cálculos y su posterior validación.

Por otro lado, en el caso en que se empleen diseños sin réplicas no es posible realizar un análisis de la significancia de las variables, por lo cual en estos casos se sugiere que se compare con la precisión del método, para demostrar si las variables afectan. La comparación de la magnitud de los efectos con la precisión del método, es quizás el criterio más recomendada en muchas guías, sin embargo, dicha práctica no es recomendada, por diversas razones dentro de las que se encuentran: (i) no considera que el método se puede encontrar cercano a los límites del error sistemático, (ii) la precisión del método en condiciones de repetibilidad puede tener diversos valores (heterocedasticidad), (iii) la precisión del método es baja (altos %CV) lo cual no permite evaluar el efecto de manera adecuada con certeza y (iii) no permite evaluar si los factores son significativos a nivel de confianza dado.

Por lo anterior, dentro de validaR se ha diseñado un módulo que permite evaluar los diferentes diseños experimentales de una manera rápida y segura, para diseños con réplicas y sin réplicas. Dentro de este módulo se encuentra la opción de incluir criterios de guías para los diseños sin réplicas.

Ejemplo 34: Robustez: Análisis de resultados de la evaluación de robustez en un diseño sin réplicas y con réplicas.

El método QuEChERS, es quizás el método multiresiduo más empleado para el análisis de residuos de contaminantes de tipo orgánico en alimentos, a continuación, se presentan los factores que se seleccionaron para un estudio de robustez de este método (Ahumada y Zamudio, 2011). En un laboratorio se seleccionaron los siguientes factores:

Variable	Nivel bajo	Nivel alto	Unidades	Nombre corto
Masa de sulfato de magnesio	4.5	5.0	g	MSO4
Masa de amina primaria/secundaria	20	25	mg	MPSA
Flujo del gas de secado (cambio de disolvente)	10	15	L/min	FGS
Temperatura de secado	30	35	^{o}C	TS
Temperatura del bloque de calentamiento	200	220	^{o}C	TBC
Temperatura de la línea de desolvatación	220	250	^{o}C	TLD
Voltaje del capilar	4.0	4.5	kV	VC

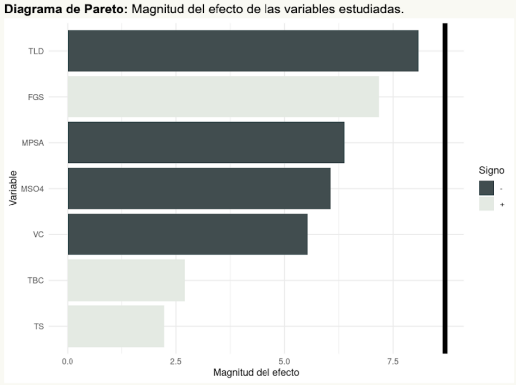
Para este estudio de robustez, se selecciona el diseño de factorial fraccionado (sin réplicas) y como variable respuesta se emplea el porcentaje de recuperación del método. La Figura 3.15, muestra los resultados obtenidos a través de validaR.

Resultados:

El diseño experimental está saturado y no es posible evaluar la significancia estadística de las variables.

Concluya a partir de lo que se muestra en el diagrama de Pareto considerando el valor de precisión típica del método.

Diagrama de Pareto: Magnitud del efecto de las variables estudiadas.



El valor de precisión típica del método ingresado se representa como una línea vertical negra en el gráfico.
Se considera que el método no es robusto frente a las variables cuyos efectos sobrepasan este límite.

Los resultados de la figura anterior, específicamente el diagrama de Pareto, indican que 4 de los factores (TLD, MPSA, MSO4 y VC) tienen una relación directa positiva con el porcentaje de recuperación, por el contrario, los 3 factores restantes producen un descenso en los porcentajes de recuperación. Por otro lado, al comparar los resultados con el criterio establecido (línea negra en el diagrama de Pareto), se encuentra que debido a que la magnitud del efecto de ninguno de los factores es superior al límite, se puede concluir que el método es robusto a los factores y los niveles evaluados.

De igual manera, como se puede apreciar en la Figura 3.16, no es posible realizar un análisis de la significancia de los efectos, por lo cual se procede a realizar el mismo diseño de experimentos, pero con una réplica. A continuación, se presentan los resultados obtenidos en la evaluación de la robustez del mismo método, pero con el empleo de réplicas de medición.

Resultados:

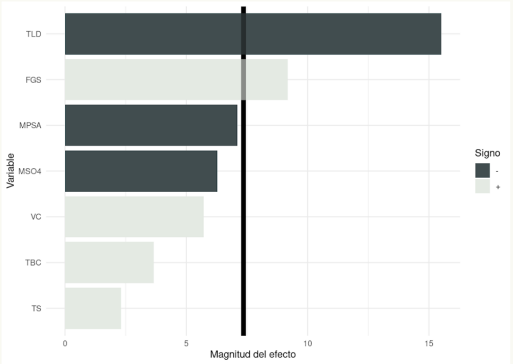
El método de medición es robusto solo frente a cambios en las siguientes variables:

- **MSO4** entre los niveles [4.5 y 5]
  Valor P del efecto: 0.3091
- **MPSA** entre los niveles [20 y 25]
  Valor P del efecto: 0.2534
- **FGS** entre los niveles [10 y 15]
  Valor P del efecto: 0.1505
- **TS** entre los niveles [30 y 35]
  Valor P del efecto: 0.699
- **TBC** entre los niveles [200 y 220]
  Valor P del efecto: 0.5445
- **VC** entre los niveles [4 y 4.5]
  Valor P del efecto: 0.3516

Importante: El método de medición no es robusto frente a cambios en las siguientes variables que presentan un efecto estadísticamente significativo a un 95 % de confianza:

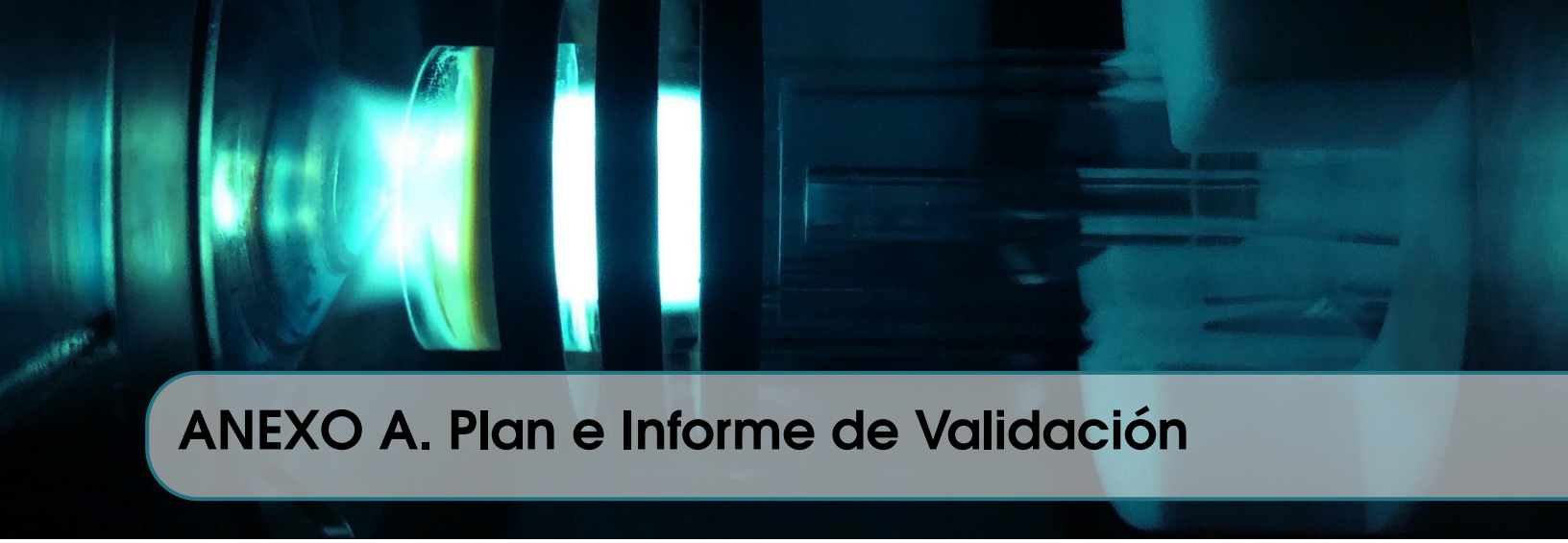
- **TLD** entre los niveles [220 y 250]
  Valor P del efecto: 0.0278

Diagrama de Pareto: Magnitud del efecto de las variables estudiadas.



El valor de precisión típica del método ingresado se representa como una línea vertical negra en el gráfico.
Se considera que el método no es robusto frente a las variables cuyos efectos sobrepasan este límite.

Como puede observarse, por medio de la realización de una réplica del diseño, se observan grandes diferencias en los resultados. La primera y más relevante es que el análisis estadístico permite encontrar que la variable TLD afecta al método de una manera significativa, lo cual indica que se debe controlar en mejor medida la variable o esta se debe incluir en el presupuesto de incertidumbre del método. Por otro lado, otra diferencia encontrada es que mientras el diseño de experimentos sin réplicas indicó que VC tenía una relación inversa con el porcentaje de recuperación, el diseño con réplicas permitió identificar que su relación era directamente positiva. Por último, pero de resaltar, es que al emplear el criterio de la precisión se encontraría que la FGS también resultaría ser una variable que afecta el método de medición, lo cual indicaría que es un criterio más exigente, sin embargo, esto puede ser engañoso pues depende del estudio de precisión.



ANEXO A. Plan e Informe de Validación

El documento **plan de validación** hace referencia a un protocolo, previo a la ejecución de la validación, donde se definen los parámetros de validación a evaluar, el alcance de la validación, y la manera en la que se llevará a cabo experimentalmente la evaluación de cada uno de ellos.

Por su parte, el **informe de validación** se elabora una vez se cuenten con todos los resultados obtenidos de la experimentación, y los análisis estadísticos que permitan verificar el cumplimiento de los criterios de aceptación o requisitos previamente definidos en el plan de validación.

Método de medición	Desarrollo y documentación del método de medición
Planificación de la validación	Elaboración del plan de validación, previo a la ejecución
Ejecución de la validación	Desarrollo experimental y análisis de resultados
Informe de la validación	Elaboración del documento con el análisis de los resultados y conclusiones

El plan e informe de la validación deberán ser elaborados por el responsable de todo el proceso de validación, quien a su vez estará encargado de coordinar el desarrollo y realizar el análisis de resultados.

Antes de elaborar el plan de validación y de su ejecución, es necesario:

- Tener el método de medición documentado.
- Contar con personal capacitado y competente en el método de medición a evaluar
- Asegurar la disponibilidad de los reactivos, consumibles y materiales necesarios para la ejecución de los experimentos
- Asegurar el adecuado funcionamiento de los equipos e instrumentos, y contar con las instalaciones y condiciones ambientales adecuadas.

Plan de validación

En este plan de validación se deberá establecer:

- Los parámetros de validación.
- Los experimentos que se llevarán a cabo para la evaluación de cada uno de los parámetros seleccionados.
- Los tratamientos estadísticos y los criterios de aceptación de acuerdo con los requerimientos del método a validar.

Este documento puede incluir, como mínimo:

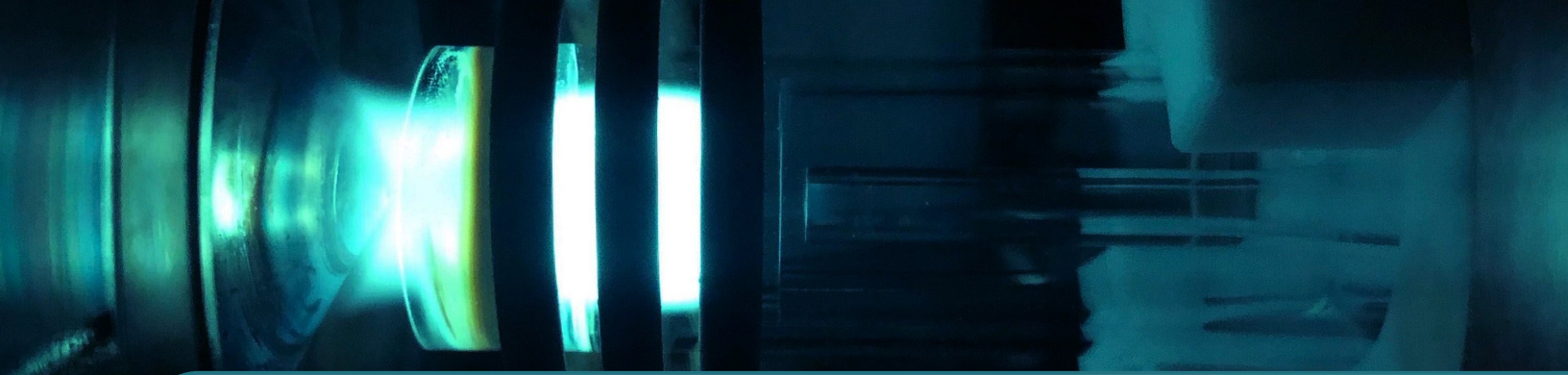
- Objetivo y alcance de la validación: dentro del objetivo y el alcance se deberá mencionar el método de medición que se va a validar junto con sus criterios de aceptación, la descripción del mensurando a ser analizado con sus niveles de concentración.
- Método de medición: hacer una descripción del procedimiento de medición con sus limitaciones y precauciones, el cual deberá estar documentado. Cabe aclarar que este procedimiento no debe ser modificado durante el desarrollo de la validación.
- Lista de materiales y equipos: identificar los equipos y materiales necesarios, y verificar que se encuentren calificados, calibrados o verificados, según sea el caso.
- Metodología experimental: indicar lo más detallada posible los experimentos a realizar, donde se mencione el tipo de muestra o muestras a ser analizadas (blanco de muestra, MRC, muestras fortificadas, entre otros), los materiales, insumos y equipos necesarios para el desarrollo de la validación, los parámetros de validación, el diseño experimental y el orden en el cual serán evaluados.

Informe de validación

El informe de validación debe contener como mínimo:

- Objetivo y alcance de la validación: dentro del objetivo y el alcance se deberá mencionar el método de medición que se validó y la descripción del mensurando con sus niveles de concentración.
- Método de medición: descripción del procedimiento de medición que fue validado.
- Metodología experimental: resumen de los experimentos realizados.
- Parámetros de validación evaluados, los requisitos o criterios de aceptación definidos, el resultado obtenido y el concepto donde se indique el cumplimiento de los requisitos o criterios de aceptación.
- Conclusión y declaración sobre si el método de medición es adecuado o apto para su uso previsto.
- Nombre del responsable y fecha de la elaboración

Si los criterios de aceptación o requisitos definidos en el plan de validación no se cumplen, se podría concluir que el método de medición no satisface el fin propuesto, y por lo tanto sería necesario regresar a la etapa de desarrollo y optimización del método para su mejora, lo cual implicaría realizar de nuevo el proceso de validación.



Bibliografía

Ahumada, Diego A. y Adriana M. Zamudio (2011). «Análisis de residuos de plaguicidas en tomate mediante el uso de QuEChERS y cromatografía líquida ultrarápida acomplada a espectrometría de masas». En: *Revista Colombiana de Química* 40.2, páginas 227-246. URL: https://revistas.unal.edu.co/index.php/rcolquim/article/view/26091 (véase página 86).

Anita Nanda and Bibhuti Bhusan and Abikesh Kumar and Abiresh Kumar and Abinash Kumar (2021). «Multiple comparison test by Tukey's honestly significant difference (HSD): Do the confident level control type I error.» En: *]International Journal of Statistics and Applied Mathematics* 6.1. DOI: 10.22271/maths.2021.v6.i1a.636 (véase página 66).

AOAC (2012). *Guidelines for Single Laboratory Validation of Chemical Methods for Dietary Supplements and Botanicals*. AOAC. Rockville, United States: AOAC International (véase página 3).

Ashworth, M. J. y col. (ene. de 2018). «Analysis and Assessment of Exposure to Selected Phthalates Found in Children's Toys in Christchurch, New Zealand». En: *Int J Environ Res Public Health* 15.2. DOI: 10.3390/ijerph15020200 (véase página 2).

Barwick, Vicki y Elizabeth Prichard, editores (2011). *Eurachem: Terminology in Analytical Measurement*. European Union. ISBN: 978-0-948926-29-7. URL: http://www.eurachem.org (véase página 2).

Bulmer, M.G. (1979). *Principles of Statistics*. Dover Books on Mathematics Series. Dover Publications. ISBN: 9780486637600 (véase página 10).

Chang, Winston y col. (2020). *shiny: Web Application Framework for R*. R package version 1.4.0.2. URL: https://CRAN.R-project.org/package=shiny (véase página IV).

Conover, W. J. (1999). *Practical Nonparametric Statistics*. 9.ª edición. Wiley. ISBN: 978-0-471-16068-7 (véase página 21).

European Union (2020). *Analytical quality control and method validation procedures for pesticide residues analysis in food and feed*. European Union (véase página 2).

FDA, Food and Drug Administration (2019). *Guidelines for the Validation of Chemical Methods in Food, Feed, Cosmetics, and Veterinary Products*. FDA Guide. Maryland, United States: U.S. Food y Drug Administration (véase página 4).

Funke, Sabrina K. I., Michael Sperling y Uwe Karst (2021). «Weighted Linear Regression Improves Accuracy of Quantitative Elemental Bioimaging by Means of LA-ICP-MS». En: *Analytical Chemistry* 93.47. PMID: 34784194, páginas 15720-15727. DOI: 10.1021/acs.analchem.1c03630. URL: https://doi.org/10.1021/acs.analchem.1c03630 (véase página 32).

Haeckel, Rainer, Werner Wosniok y Rainer Klauke (2013). «Comparison of ordinary linear regression, orthogonal regression, standardized principal component analysis, Deming and Passing-Bablok approach for method validation in laboratory medicine». En: *Laboratoriumsmedizin* 37.3, páginas 147-163. DOI: doi:10.1515/labmed-2013-0003. URL: https://doi.org/10.1515/labmed-2013-0003 (véanse páginas 33, 34).

Hernández Revilla, Marta (2013). «Validación de métodos de ensayo y estimación de la incertidumbre de medida conforme a la norma ISO/IEC 17025. Aplicación al análisis de aguas residuales». Tesis doctoral. URL: http://uvadoc.uva.es/handle/10324/4284 (véase página 57).

JCGM, Joint Committee for Guides in Metrology (2012). *International Vocabulary of Metrology - Basic and General Concepts and Associated Terms*. Volumen 3. Bureau International des Poids et Measures (véanse páginas 1, 2, 6, 27, 31, 38, 47, 48, 56, 70).

Johnson, Richard A. y Gouri K. Bhattacharyya (2009). *Statistics: Principles and Methods*. 6.ª edición. Wiley. ISBN: 9780470409275 (véase página 17).

Ketkar, S. N. y T. J. Bzik (2000). «Calibration of Analytical Instruments. Impact of Nonconstant Variance in Calibration Data». En: *Analytical Chemistry* 72.19, páginas 4762-4765. ISSN: 0003-2700. DOI: 10.1021/ac000018s. URL: https://doi.org/10.1021/ac000018s (véase página 32).

King, Andrew P. y Robert J. Eckersley (2019). «Chapter 6 - Inferential Statistics III: Nonparametric Hypothesis Testing». En: *Statistics for Biomedical Engineers and Scientists*. Editado por Andrew P. King y Robert J. Eckersley. Academic Press, páginas 119-145. ISBN: 978-0-08-102939-8. DOI: https://doi.org/10.1016/B978-0-08-102939-8.00015-3 (véanse páginas 16, 18, 19, 45).

Lilliefors, Hubert W. (1967). «On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown». En: *Journal of the American Statistical Association* 62.318, páginas 399-402. DOI: 10.1080/01621459.1967.10482916 (véase página 11).

Magnusson, B. y U. Örnemark, editores (2014). *Eurachem Guide: The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics*. 2.ª edición. European Union. ISBN: 978-91-87461-59-0. URL: http://www.eurachem.org (véanse páginas 1, 48, 64).

Mulholland, M. (1988). «Ruggedness testing in analytical chemistry». En: *TrAC Trends in Analytical Chemistry* 7.10, páginas 383-389. ISSN: 0165-9936. DOI: https://doi.org/10.1016/0165-9936(88)85089-1 (véase página 80).

NIST/SEMATECH (2022). *NIST/SEMATECH e-Handbook of Statistical Methods*. URL: https://doi.org/10.18434/M32189 (visitado 24-05-2022) (véase página 81).

Norma ISO 11095 (1996). *Linear calibration using reference materials*. ISO. Geneva, Switzerland: International Organization for Standardization (véase página 66).

Norma ISO 3534-1 (2006). *Statistics - Vocabulary And Symbols - Part 1: General Statistical Terms And Terms Used In Probability*. ISO. Geneva, Switzerland: International Organization for Standardization (véanse páginas 47, 49, 56).

Norma ISO 5725-1 (1994). *Accuracy (Trueness and Precision) of Measurement Methods and Results - Part 1: General Principles and Definitions*. ISO. Geneva, Switzerland: International Organization for Standardization (véanse páginas 48, 49, 56).

Norma ISO/IEC 17025 (2017). *General requirements for the competence of testing and calibration laboratories*. ISO/IEC. Geneva, Switzerland: International Organization for Standardization e International Electrotechnical Comission (véanse páginas 2, 3).

Otto, Matthias y Wolfhard Wegscheider (1986). «Selectivity in multicomponent analysis». En: *Analytica Chimica Acta* 180, páginas 445-456. ISSN: 0003-2670. DOI: https://doi.org/10.1016/0003-2670(86)80024-1 (véase página 41).

Pagliano, Enea, Zoltán Mester y Juris Meija (2015). «Calibration graphs in isotope dilution mass spectrometry». En: *Analytica Chimica Acta* 896, páginas 63-67. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2015.09.020. URL: https://www.sciencedirect.com/science/article/pii/S0003267015011599 (véase página 30).

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/ (véase página IV).

Raposo, Francisco y Damià Barceló (2021). «Assessment of goodness-of-fit for the main analytical calibration models: Guidelines and case studies». En: *TrAC Trends in Analytical Chemistry* 143, página 116373. ISSN: 0165-9936. DOI: https://doi.org/10.1016/j.trac.2021.116373. URL: https://www.sciencedirect.com/science/article/pii/S0165993621001965 (véase página 30).

Razali, N. M. e Y. B. Wah (2011). «Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests». En: *Journal of Statistical Modeling and Analitics* 2.1, páginas 21-33 (véase página 11).

Ripley, Brian D. y Michael Thompson (1987). «Regression techniques for the detection of analytical bias». En: *Analyst* 112 (4), páginas 377-383. DOI: 10.1039/AN9871200377. URL: http://dx.doi.org/10.1039/AN9871200377 (véase página 28).

SANTE 11312 (2021). *Guidance Document SANTE 11312/2021 – Analytical quality control and method validation procedures for pesticide residues analysis in food and feed*. Informe técnico. European Commission (véase página 58).

Therneau, Terry (2018). *deming: Deming, Theil-Sen, Passing-Bablock and Total Least Squares Regression*. R package version 1.4. URL: https://CRAN.R-project.org/package=deming (véase página 28).

Thompson, Michael, Stephen L. R. Ellison y Roger Wood (2002). «Harmonized guidelines for single-laboratory validation of methods of analysis (IUPAC Technical Report)». En: *Pure and Applied Chemistry* 74.5, páginas 835-855. DOI: 10.1351/pac200274050835 (véanse páginas 1, 2, 5, 79).

USP 36 (2009). *General Chapter 1225, Validation of Compendial Procedures*. USP. Rockville, United States: United States Pharmacopoeia (véase página 79).

van Leeuwen, J.A. y col. (1991). «RES, an expert system for the set-up and interpretation of a ruggedness test in HPLC method validation: Part 1: The ruggedness test in HPLC method validation». En: *Chemometrics and Intelligent Laboratory Systems* 10.3, páginas 337-347. ISSN: 0169-7439. DOI: https://doi.org/10.1016/0169-7439(91)80098-B (véase página 80).

van Zoonen, Piet y col. (1999). «Some practical examples of method validation in the analytical laboratory». En: *TrAC Trends in Analytical Chemistry* 18.9, páginas 584-593. ISSN: 0165-9936. DOI: https://doi.org/10.1016/S0165-9936(99)00159-4. URL: https://www.sciencedirect.com/science/article/pii/S0165993699001594 (véase página 37).

Vander Heyden, Y. y D.L. Massart (1996). «Chapter 3 Review of the use of robustness and ruggedness in analytical chemistry». En: *Robustness of analytical chemical methods and pharmaceutical technological products*. Editado por Margriet M.W.B. Hendriks, Jan H. de Boer y Age K. Smilde. Volumen 19. Data Handling in Science and Technology. Elsevier, páginas 79-147. DOI: https://doi.org/10.1016/S0922-3487(96)80016-5 (véase página 80).

Vessman, Jörgen y col. (2001). «Selectivity in analytical chemistry (IUPAC Recommendations 2001)». En: *Pure and Applied Chemistry* 73.8, páginas 1381-1386. DOI: 10.1351/pac200173081381 (véanse páginas 38, 39).

Youden, W.J. y E.H. Steiner (1975). *Statistical Manual of the Association of Official Analytical Chemists*. The Association of Official Analytical Chemists. URL: https://books.google.com.co/books?id=IfdBAAAAIAAJ (véase página 79).

Zhang, Mengtao y col. (2016). «Validation and application of an analytical method for the determination of selected acidic pharmaceuticals and estrogenic hormones in wastewater and sludge». En: *Journal of Environmental Science and Health, Part A* 51.11, páginas 914-920. DOI: 10.1080/10934529.2016.1191304 (véase página 2).

GLOBAL QUALITY AND STANDARDS PROGRAMME

PROGRAMA DE CALIDAD PARA LA CADENA DE QUÍMICOS

Instituto Nacional de Metrología de Colombia

Un programa de:

ONUDI ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA EL DESARROLLO INDUSTRIAL

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Confederación Suiza

Departamento Federal de Economía,
Formación e Investigación DEFI
Secretaría de Estado para Asuntos Económicos SECO

MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO

Colombia Productiva
PRODUCTIVIDAD • CALIDAD • VALOR AGREGADO