ESTADÍSTICA SIN MIEDO

Descifrando los números que explican el mundo

Primera Edición Digital

José Luis Villavicencio Guardia.
Yermmy Vasquez Salis.
Carlos Alberto Ramírez Chumbe.

Ho Nexus
EDITORIAL

ESTADÍSTICA SIN MIEDO

Descifrando los números que explican el mundo

Primera Edición Digital

José Luis Villavicencio Guardia.
Yermmy Vasquez Salis.
Carlos Alberto Ramírez Chumbe.

Ho Nexus
EDITORIAL

ESTADÍSTICA SIN MIEDO

Descifrando los números que explican el mundo

Editor de contenido:
Diseño de cubierta: Ho Nexus

1ª edición digital, marzo 2026

Editado por:

© HO NEXUS E.I.R.L.
Dirección legal: Urb. Paseo del Mar Mz L4, Lt 33
Nuevo Chimbote, Santa, Ancash - Perú
Correo electrónico; ed.honexus@gmail.com
teléfono: 978 653 152
https://books.honexus.org
DOI: https://doi.org//10.70504/978-612-99293-7-8

Depósito Legal: 2026-02886
ISBN: 978-612-99293-7-8

Revisión por pares:
Este libro (o monografía) fue sometido a evaluación de pares mediante el sistema de doble ciego (doubleblinded review), garantizando la calidad, pertinencia, ética y rigor académico de la obra, conforme a los estándares internacionales de revisión científica y las políticas editoriales de Ho Nexus.

ÍNDICE

PRÓLOGO

¿Alguna vez has escuchado una noticia que dice "4 de cada 5 dentistas recomiendan..." y te has preguntado de dónde sacan ese número? ¿O has visto una gráfica en las redes sociales que parece contar una historia, pero intuyes que quizás no es toda la verdad? Vivimos en un mundo saturado de datos. Desde las encuestas políticas hasta los estudios que nos dicen qué alimentos son buenos o malos para la salud, la estadística está en todas partes.

Pero, seamos honestos, para muchos de nosotros, la palabra "estadística" puede evocar recuerdos de fórmulas complicadas, clases aburridas y una pizarra llena de números que parecían hablar en otro idioma. Este libro nace con la misión de cambiar eso para siempre.

Para quienes han huido siempre de los números, "Estadística Sin Miedo" llega como ese amigo que se sienta a tu lado y te dice: "tranquilo, en realidad esto es más sencillo de lo que parece". Olvídate de esos manuales que parecen escritos en otro idioma, llenos de ecuaciones que abruman solo con verlas. Este libro propone algo distinto: un paseo curioso por el universo de los datos, donde las ideas fluyen con naturalidad, los ejemplos brotan de situaciones que vives a diario y, casi sin darte cuenta, terminas descubriendo que los números también pueden contar historias fascinantes.

Lo que sostienes entre las manos no ha surgido de la nada. Detrás de cada página hay un trabajo silencioso de destilación: tomar las enseñanzas de quienes verdaderamente dominan la materia sobre el arte de elegir una muestra que no engañe, de crear herramientas que midan lo que prometen, de interpretar resultados sin caer en trampas y devolverlas convertidas en algo cercano, útil y vivo. Hemos removido todo lo que sobraba, simplificado

lo complejo y construido una narrativa que, ojalá, te atrape tanto como a nosotros nos atrapó escribirla. Porque al final, de eso se trata: de que pierdas el miedo, te equivoques si hace falta, pero sobre todo, aprendas a sacar tus propias conclusiones con los pies bien puestos en la tierra.

Prepárate para descubrir que la estadística no es un monstruo de siete cabezas, sino una herramienta poderosa para darle sentido al caos. Bienvenido a un viaje donde los números cobran vida y te ayudan a ver el mundo con otros ojos. Empecemos.

RESUMEN

Vivimos rodeados de números. Cada vez que enciendes la tele, un político presume de una encuesta; cuando hojeas el periódico, un estudio revela qué alimentos alargan la vida; incluso en la conversación del café, alguien suelta un "el 70% de la gente piensa que...". Los datos están por todas partes, empujándonos a formar una opinión, a creer, a comprar o a dudar. En este escenario, saber moverse entre cifras ha dejado de ser una rareza académica para convertirse en algo tan básico como entender la letra pequeña de un contrato o saber por qué sube la luz.

El problema es que, para la mayoría, la estadística sigue sonando a chino. Nos la imaginamos como una habitación cerrada donde unos tipos con gafas y pizarras llenas de símbolos raros discuten sobre cosas que no tienen nada que ver con nuestro día a día. Este libro nace, precisamente, con la idea de tumbar esa puerta de una patada. Quiere demostrar que detrás de esos números que parecen fríos y lejanos, se esconden historias apasionantes, herramientas para entender mejor a nuestros vecinos, a nuestros clientes o incluso a nosotros mismos. No hace falta ser un genio para atravesar ese umbral; solo hace falta dejar el miedo en la puerta y tener curiosidad por lo que hay al otro lado.

Lo primero que descubrirás al asomarte a "Estadística Sin Miedo: Descifrando los Números que Explican el Mundo" es que los números, cuando se explican bien, dejan de ser esos bichos raros que dan ganas de espantar. El manuscrito funciona como una brújula para orientarse en ese laberinto de cifras en el que vivimos metidos sin pedirlo. Aquí no se trata de volverse experto en ecuaciones, sino de agarrar al vuelo ciertas ideas que, una vez que las entiendes, te cambian la mirada.

Por ejemplo, ¿cuántas veces has escuchado el resultado de una encuesta y te has preguntado si realmente representa lo que piensa la gente? Pues bien, entre sus páginas vas a encontrar las claves para diferenciar, sin necesidad de ser estadístico, cuándo una muestra es de fiar y cuándo te quieren colar gato por liebre. También te vas a llevar una sorpresa al descubrir que lo que verdaderamente importa no es tanto cuánta gente preguntes, sino a quién preguntas y cómo lo haces. Y ya puestos, te asomarás al trabajo que hay detrás de cualquier investigación seria: esa obsesión silenciosa de los expertos por asegurarse de que las preguntas que hacen realmente sirvan para medir lo que pretenden, y no cualquier otra cosa.

Todo esto explicado con los pies en la tierra, agarrado a situaciones que podrías vivir en cualquier esquina de Latinoamérica, donde las distancias, los sesgos y las realidades sociales tienen su propio color. Porque de eso se trata: de aprender a leer el mundo con otros ojos, sin miedo y con la tranquilidad de saber que, al final, los datos bien entendidos son una herramienta más para no dejarse engañar. Desde entender los márgenes de error en las encuestas electorales hasta evaluar la confiabilidad de un test psicológico o educativo, este libro proporciona las herramientas conceptuales necesarias para navegar con confianza en el océano de datos que nos rodea. Al finalizar la lectura, el lector no solo habrá adquirido un vocabulario técnico básico, sino que habrá desarrollado un pensamiento crítico que le permitirá cuestionar, interpretar y utilizar la información estadística con autonomía y criterio propio.

Palabras clave: Estadística aplicada, pensamiento crítico, muestreo probabilístico, validez de instrumentos, alfabetización cuantitativa.

INTRODUCCIÓN

¿Por qué este libro existe y por qué debería importarte?

Imaginemos por un momento la siguiente escena: es domingo por la noche y en la televisión anuncian los resultados de una encuesta electoral. La candidata del partido oficialista aparece con un 42% de intención de voto, mientras que su principal contendiente alcanza el 38%. Los conductores del programa celebran, analizan, especulan. Al día siguiente, todos comentan la noticia en el trabajo, en la universidad, en las redes sociales. Pero surge una pregunta incómoda: ¿a cuántas personas entrevistaron para llegar a esas conclusiones? ¿Dos mil, quinientas, quizás solo doscientas? ¿Y quiénes eran esas personas? ¿Cómo las eligieron? ¿Las preguntas estaban bien formuladas o inducían una respuesta determinada?

Estas interrogantes no son triviales. Detrás de cada número que leemos en las noticias, detrás de cada estudio que afirma que "el consumo de chocolate reduce el estrés" o que "los jóvenes peruanos prefieren las carreras técnicas", existe un complejo entramado de decisiones metodológicas que determinan si esa información es confiable o no. La estadística, lejos de ser un ejercicio abstracto y alejado de la realidad, es precisamente el conjunto de herramientas que nos permite responder a estas preguntas y, más importante aún, nos protege de ser engañados por quienes manipulan los datos para servir a sus propios intereses.

Este libro nace de la convicción de que la estadística debería ser tan accesible como la lectura o la escritura. No pretendemos formar matemáticos ni estadísticos profesionales; nuestro objetivo es mucho más ambicioso y, a la vez, más humilde: queremos ayudarte a desarrollar una mirada crítica y curiosa frente a los números que te rodean. Queremos que la próxima vez

que escuches una estadística en las noticias, puedas hacer las preguntas correctas: ¿de dónde vienen estos datos?, ¿cómo los obtuvieron?, ¿son confiables?, ¿qué me están queriendo decir realmente?

Para que todo esto no se quede en una simple declaración de intenciones, hemos armado el libro como quien construye una casa: primero los cimientos, luego las paredes, y así sucesivamente. Son cuatro capítulos que van llevando al lector de la mano, siguiendo más o menos los mismos pasos que daría cualquier investigador cuando se sienta frente a un problema de verdad, con los pies en la tierra y sin rodeos.

El primer capítulo, que lleva por título "La Base de Todo: Poblaciones, Muestras y Para Qué Sirven", es exactamente eso: el suelo que pisamos. Porque antes de meterse en honduras, hay que tener claro de qué hablamos cuando decimos "población" que no es solo la gente que vive en una ciudad, sino el grupo entero que nos interesa y "muestra", que es ese pedacito del pastel que realmente vamos a examinar con lupa. Aquí vamos a descubrir algo que parece de Perogrullo pero que mucha gente olvida: casi nunca hace falta preguntarle a todo el mundo para saber lo que el mundo piensa. El truco está en elegir bien a quién preguntas. También vamos a asomarnos a conceptos como el marco muestral que viene a ser la lista de la que sacamos a nuestra gente y, sobre todo, vamos a aprender a distinguir entre lo que queremos saber de verdad (los expertos le llaman "parámetro") y lo que finalmente calculamos con nuestros datos (eso es la "estadística"). Puede sonar a juego de palabras, pero no lo es: es la diferencia entre la realidad y lo que alcanzamos a ver de ella. Y sin tener clara esa frontera, mejor no seguir adelante.

Una vez que tenemos claro el terreno que pisamos, el Capítulo II, "El Arte de Elegir: Técnicas de Muestreo", nos mete de lleno en la faena. Porque

no todas las muestras valen lo mismo. Así de claro. Hay unas, las llamadas probabilísticas, que nos permiten sacar conclusiones que valen para el conjunto general; son las que usan los políticos cuando quieren vender que ganarán las elecciones. Y hay otras, las no probabilísticas, que son útiles para según qué cosas un estudio exploratorio, una prueba de producto pero con las que hay que andarse con cuidado porque lo que descubras solo vale para esa gente concreta que participó. En este capítulo también vamos a perderle el miedo a eso de "calcular el tamaño de la muestra", que siempre suena a fórmula mágica pero tiene su lógica. Y vamos a repasar las distintas maneras de echar a suertes o de seleccionar a los participantes: desde el clásico sorteo de la lotería hasta métodos más finos como el muestreo estratificado que es como asegurarte de que en tu muestra haya de todos los barrios o el muestreo por conglomerados, muy útil cuando andas justo de presupuesto o de tiempo. Cuando termines este capítulo, te vas a convertir en ese amigo incómodo que, cuando alguien suelte un "según una encuesta...", levantará la mano para preguntar: "un momento, ¿y esa muestra, cómo la eligieron?".

Pero ojo, que ni todo el monte es orégano ni tener una muestra de lujo te garantiza una buena investigación. Te puedes pasar meses seleccionando a los participantes perfectos, con cuotas y sorteo incluido, para luego tirarlo todo por la borda con un cuestionario mal planteado. Porque si las preguntas son un galimatías, llevan trampa o simplemente no apuntan a lo que te interesa, el dato, por bonito que sea, vale poco. De ese lío se ocupa el Capítulo III, que hemos titulado "El Desafío de Medir lo Invisible: Construyendo Instrumentos Confiables". Aquí el libro hace de puente: dejamos atrás el arte de elegir gente y nos adentramos en el arte de preguntar bien. Vamos a ver cómo los expertos hacen malabares para convertir ideas que no se pueden tocar la inteligencia, eso que llamamos "calidad de vida", el quemazón del trabajo en preguntas de esas que la gente entiende y puede

responder sin sentirse tonta. Hablaremos de escalas, de tipos de pregunta, del famoso "del 1 al 10"... y también de los errores de novato que hacen que una encuesta naufrague antes de empezar. Porque sí, preguntar parece fácil, pero preguntar bien es un oficio.

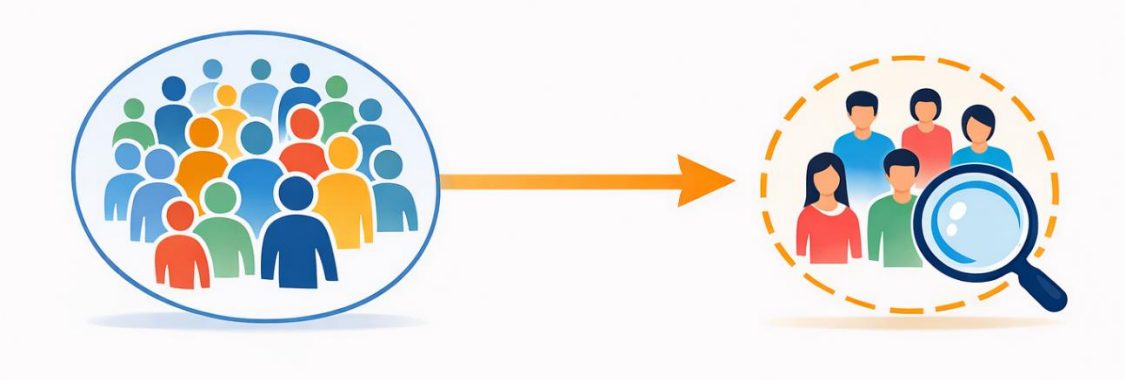
Y llegamos al final del trayecto, al Capítulo IV, que lleva por título "¿Es Confiable lo que Mide? Validez y Confiabilidad de un Instrumento". Aquí nos ponemos exquisitos. Porque ya tenemos nuestra muestra bien traída, nuestro cuestionario bien pulido... pero aún nos acechan dos preguntas incómodas, de esas que quitan el sueño a cualquier investigador que se precie. La primera: ¿estamos midiendo realmente lo que creemos medir, o nos estamos engañando con fuegos de artificio? Eso, en el argot, es la validez. Y la segunda: ¿nuestro invento mide siempre más o menos lo mismo, o es como esos termómetros de mercadillo que unos días te dan fiebre y otros hipotermia? Eso es la confiabilidad. En este capítulo vamos a desgranar los distintos tipos de validez de contenido, de constructo, de criterio... y nos asomaremos a las herramientas que usan los estadísticos para certificarlas, como la V de Aiken o ese tinglado un poco más complejo que llaman análisis factorial. Y en el mundo de la confiabilidad, nos haremos amigos de métodos como el test-retest (que es tan sencillo como volver a preguntar más adelante), el de mitades partidas y, cómo no, el famoso Alfa de Cronbach. Ese número que ves en todas las tesis y en todos los estudios, ese que a veces parece un fetiche, y que después de leer estas páginas podrás mirar a los ojos y decirle: "ya sé lo que escondes, colega".

Para cerrar la obra, ofrecemos una Conclusión que integra todos los conceptos aprendidos y reflexiona sobre la importancia de la estadística para la toma de decisiones informadas en la vida personal y profesional, así como un Apéndice con tablas y recursos prácticos que facilitarán la aplicación de lo aprendido.

Este texto está escrito para ti, estudiante universitario que te enfrentas por primera vez a un curso de metodología; para ti, profesional que necesitas interpretar estudios en tu campo laboral; para ti, ciudadano curioso que quieres entender mejor el mundo que te rodea. No necesitas ser un experto en matemáticas; solo necesitas curiosidad y ganas de descubrir que, detrás de cada número, hay una historia fascinante esperando ser contada.

Bienvenido a "Estadística Sin Miedo". Los números están a punto de cobrar vida.

CAPÍTULO I: LA BASE DE TODO: POBLACIONES, MUESTRAS Y PARA QUÉ SIRVEN



Pongamos un caso bien concreto, de esos que ayudan a entender por qué todo esto importa. Imagina que queremos saber qué piensan los peruanos de eso de la educación virtual, esa que llegó para quedarse después de la pandemia. ¿Te imaginas tener que salir a preguntarle a cada uno de los 33 millones de habitantes? Sería una locura. No solo por la plata que costaría, que sería un dineral, sino por el tiempo: cuando termináramos de entrevistar al último, el primero ya habría cambiado de opinión tres veces. Y total, para qué, si con una porción bien elegida podemos hacernos una idea más que decente.

Esa idea, tan sencilla como profunda, es la que está en el corazón de una de las herramientas más útiles que tienen los investigadores: el muestreo. Porque sí, a veces menos es más, pero siempre que ese "menos" esté bien escogido.

Ahora bien, antes de meternos en harina y empezar a ver técnicas para elegir muestras eso lo dejamos para el Capítulo II, con todo lujo de detalles y mucho antes de preocuparnos por si medimos bien o regular ese es el drama

del Capítulo IV, tenemos que ponernos de acuerdo en cómo llamamos a las cosas. Porque si cada cual entiende una palabra a su manera, esto se convierte en una torre de Babel y no avanzamos. Así que este primer capítulo va de eso: de agarrar bien los cimientos, de construir entre los dos un vocabulario común que nos sirva para andar sueltos por el resto del libro. Piensa que es como sentar las bases de una casa: si están firmes, lo que venga después se sostiene solo. Aquí definiremos qué es una población, qué es una muestra y, crucialmente, cuál es la diferencia entre lo que queremos saber (el parámetro) y lo que realmente podemos calcular (la estadística).

Como señalan (Hernández Sampieri & Mendoza Torres, 2018), la claridad en la definición de estos elementos desde el inicio de una investigación no es un mero formalismo académico, sino una condición indispensable para que los resultados tengan validez y puedan ser interpretados correctamente. Sin una base sólida, cualquier conclusión, por más atractiva que sea, se sostiene sobre arena movediza.

Acompáñenos en este recorrido para entender el ABC de la investigación cuantitativa. Al finalizar este capítulo, no solo dominará la terminología básica, sino que comprenderá por qué una muestra bien elegida puede decirnos mucho sobre un universo entero.

1.1 ¿De Quién Hablamos? Definiendo la Población

Cuando un investigador se propone realizar un estudio, lo primero que debe preguntarse es: ¿sobre quiénes o sobre qué quiero concluir algo? La respuesta a esta pregunta es lo que conocemos como población o universo.

En estadística, el concepto de población es más amplio y técnico que en el lenguaje cotidiano. No se refiere únicamente a un conjunto de personas que viven en un lugar geográfico. Una población es el conjunto total de individuos, objetos, eventos o medidas que poseen una o más características en común, son observables en un lugar y momento determinado, y sobre los cuales se desea inferir algo (Nencini, 2022).

Para que una población esté correctamente definida, debe ser delimitada con precisión mediante cuatro criterios fundamentales:

1) **Contenido**: ¿Qué características específicas debe tener el elemento? (ej., ser docente de educación primaria, tener diabetes tipo 2, ser una microempresa).
2) **Extensión**: ¿Dónde se ubican? (ej., en la ciudad de Arequipa, en los hospitales públicos de Lima Metropolitana, en el distrito de San Isidro).
3) **Tiempo**: ¿En qué período se les estudia? (ej., durante el año académico 2024, en el mes de marzo de 2025, en el primer semestre del año).

4) **Accesibilidad**: ¿Son todos accesibles o solo una parte? (ej., todos los estudiantes matriculados, todos los pacientes que acuden a consulta externa).

Ejemplo práctico:

Si un investigador desea estudiar la "satisfacción laboral de los enfermeros en hospitales públicos de Lima durante el año 2024", su población estará constituida por todos los enfermeros que laboraban en hospitales públicos de Lima Metropolitana a lo largo del año 2024. Cualquier enfermero que no trabaje en un hospital público, que trabaje en una clínica privada, o que haya laborado en el año 2023, quedaría fuera de esta población.

El tamaño de la población se denota comúnmente con la letra N. En nuestro ejemplo, si el Colegio de Enfermeros del Perú reportara que hay 8,500 enfermeros en esa condición, entonces N = 8,500.

Es importante destacar que una población puede ser finita (cuando conocemos su tamaño exacto, como el ejemplo anterior) o infinita (cuando el número de elementos es tan grande que no podemos contabilizarlos a todos, como el número de posibles gotas de lluvia en una tormenta). Sin embargo, en la práctica de las ciencias sociales y de la salud, trabajamos casi siempre con poblaciones finitas, aunque a veces sean muy grandes (Mode & García Garza, 2021).

1.2 La Unidad de Análisis: El Corazón de Nuestro Estudio

Si la población es el conjunto, la unidad de análisis (también llamada unidad estadística o elemento) es cada uno de los miembros individuales que conforman ese conjunto. Es la fuente primaria de la información, el "quién" o el "qué" específico que vamos a medir (Lohr, 2022).

Identificar correctamente la unidad de análisis es crucial porque de ella dependen el diseño del instrumento de medición (la encuesta, el test) y la forma en que recolectaremos los datos.

Ejemplos de unidades de análisis en diferentes contextos:

- En educación: Un estudiante, un docente, un aula de clases, una institución educativa.
- En salud: Un paciente, una historia clínica, un profesional de la salud, un establecimiento de salud.
- En marketing: Un consumidor, un hogar, un producto, una transacción de compra.

- En sociología: Una persona, una familia, una comunidad, una organización.

Sigamos con el ejemplo de los enfermeros, que es bien gráfico. En ese caso, la unidad de análisis es clara: cada enfermero o enfermera que echa horas en los hospitales públicos de Lima. Ellos son el centro de la diana, a quienes vamos a molestar con nuestra encuesta sobre satisfacción laboral. Cada cuestionario, cada respuesta, cada dato nuevo apunta a ellos.

Pero la cosa se puede complicar, y aquí viene un matiz importante que conviene tener claro desde ya. A veces, la persona que quieres estudiar no es la misma que seleccionas para llegar a ella. Suena a trabalenguas, pero es más sencillo de lo que parece. Los expertos llaman a esto la diferencia entre unidad de análisis y unidad de muestreo, y aunque lo dejaremos bien atado en el Capítulo II, vale la pena ir abriendo boca.

Pensemos en un estudio sobre niños menores de 5 años. Ellos son nuestra unidad de análisis, los protagonistas. Ahora bien, ¿cómo los encuentras? ¿Te plantas en la calle a ver si pasa uno? No tiene sentido. Lo que haces es ir por barrios, seleccionar hogares al azar eso sería la unidad de muestreo y, una vez que llamas a la puerta, preguntas si en esa casa vive algún pequeñín en edad de merecer. Si lo hay, lo entrevistas. Si no, pasas al siguiente hogar.

Parece un detalle menor, pero no lo es. En cuanto los diseños de muestreo se vuelven un poco más sofisticados y en investigación lo son a menudo, tener clara esta distinción te ahorra dolores de cabeza y, sobre todo, errores a la hora de interpretar los resultados. Porque no es lo mismo a quién quieres conocer que a través de quién llegas hasta él (Narayan & Sinha, 2023).

1.3 El Mapa del Tesoro: El Marco Muestral

Una vez que tenemos clara nuestra población (todos los enfermeros) y nuestra unidad de análisis (cada enfermero), necesitamos una herramienta que nos permita ubicarlos y seleccionarlos. Esta herramienta es el marco muestral (o marco de muestreo).

El marco muestral es una lista, un registro, un mapa o cualquier otro dispositivo que contenga e identifique a todas las unidades de muestreo de la población. Es, metafóricamente, el "mapa del tesoro" que nos guía para encontrar a los individuos que formarán parte de nuestro (Andrés Gutiérrez, 2016).

Características de un buen marco muestral:

- **Completo**: Debe incluir a todas las unidades de la población objetivo. Una omisión introduce un sesgo conocido como error de cobertura.
- **Preciso**: Los datos de identificación (nombres, direcciones, códigos) deben ser correctos y estar actualizados.

- **Excluyente**: Ninguna unidad de muestreo debe aparecer más de una vez (no debe haber duplicados).
- **Conveniente**: Debe estar organizado de manera lógica (numérica, alfabética, geográfica) para facilitar la selección.

Ejemplos comunes de marcos muestrales:

- El padrón electoral (para estudios con votantes).
- El directorio telefónico (cada vez menos usado).
- Los registros de pacientes de un hospital.
- La nómina de trabajadores de una empresa.
- El catastro de viviendas de una municipalidad.
- Una base de datos de clientes.

En nuestro ejemplo de los enfermeros, un marco muestral ideal sería el listado oficial de todos los enfermeros colegiados y habilitados que laboran en hospitales públicos de Lima, proporcionado por el Colegio de Enfermeros del Perú o por la planilla única del Ministerio de Salud.

La disponibilidad y calidad del marco muestral determinan, en gran medida, el tipo de muestreo que podremos utilizar. Si no disponemos de una lista completa, no podremos aplicar un muestreo aleatorio simple, pero quizás sí un muestreo por conglomerados (Arnab, 2017).

1.4 ¿Qué Queremos Saber? Parámetros vs. Estadísticas

Llegamos a uno de los conceptos más importantes y, a la vez, más confusos para quienes se inician en estadística: la diferencia entre un parámetro y una estadística. Entender esta distinción es la clave para comprender el propósito mismo de la inferencia estadística.

El Parámetro: La Verdad de la Población

Un parámetro es una medida descriptiva que se calcula utilizando los datos de todos los elementos de la población. Es un valor fijo, una constante (aunque casi siempre desconocida para nosotros) que describe alguna característica de la población. En notación estadística, los parámetros suelen representarse con letras griegas (Véliz Capuñay, 2011). Los parámetros más comunes son:

- **La media poblacional (μ - "mi"):** El promedio de una variable cuantitativa en toda la población. Por ejemplo, la edad promedio de todos los enfermeros de Lima.

- **La proporción poblacional (π - "pi"):** El porcentaje de elementos de la población que poseen una característica específica. Por ejemplo, la proporción de enfermeros que están satisfechos con su trabajo.
- **La desviación estándar poblacional (σ - "sigma"):** Una medida de cuánto varían los datos con respecto a la media poblacional.

La Estadística: Nuestra Mejor Estimación

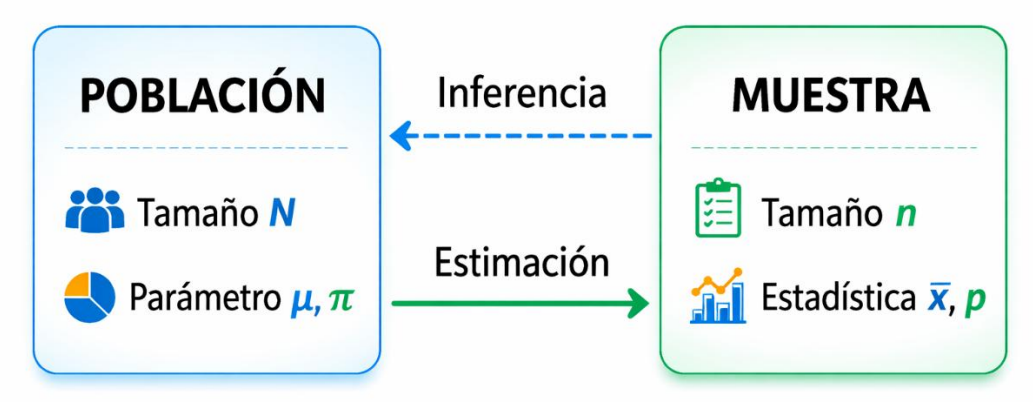
Una estadística (o estimador) es una medida descriptiva que se calcula utilizando únicamente los datos de una muestra. A diferencia del parámetro, la estadística es una variable; su valor cambia de una muestra a otra. Las estadísticas se utilizan precisamente para estimar los parámetros poblacionales desconocidos (Triola Mario F., 2018). Se representan, generalmente, con letras latinas.

Las estadísticas más comunes son:

- **La media muestral ($\bar{x}$ - "x barra"):** El promedio de la variable en la muestra. Se usa para estimar μ.
- **La proporción muestral (p):** El porcentaje de la muestra que tiene una característica. Se usa para estimar π.
- **La desviación estándar muestral (s):** La variabilidad de los datos en la muestra. Se usa para estimar σ.

La Relación Fundamental

La siguiente figura, adaptada de los conceptos de inferencia estadística de (Casella & Berger, 2024), resume esta relación:



- **Inferencia**: Es el proceso de sacar conclusiones sobre la población basándonos en la información de la muestra. Es el objetivo final.
- **Estimación**: Es el acto de usar el valor de la estadística (ej., $\bar{x} = 42$ años) para "adivinar" el valor del parámetro (ej., $\mu \approx 42$ años).

Volvamos a nuestro ejemplo:

Queremos saber cuál es la verdadera proporción de enfermeros satisfechos con su trabajo en todos los hospitales públicos de Lima (parámetro π). Como no podemos encuestar a los 8,500 (N), seleccionamos una muestra de 300 enfermeros (n). Al aplicar la encuesta, encontramos que 180 de ellos están satisfechos. Entonces, nuestra estadística (la proporción muestral, p) es 180/300 = 0.6 (60%).

Con base en esta evidencia, nuestro equipo de investigación inferirá que el verdadero parámetro poblacional (π) debe estar alrededor del 60%. Por supuesto, como veremos en el Capítulo II, esta estimación tiene un margen de error, y es precisamente la teoría del muestreo la que nos permite calcular ese margen y cuantificar nuestra confianza en él.

Esta distinción es la esencia de la estadística aplicada. Como bien señalan (Ciro Martínez, 2012), confundir un parámetro con una estadística

es uno de los errores más comunes y puede llevar a interpretaciones completamente equivocadas de los resultados de una investigación.

Para no perder el hilo: lo que llevamos hasta aquí

Vale, hagamos un alto en el camino y repasemos lo andado. Este primer capítulo ha sido como poner los cimientos de una casa: si están bien asentados, lo que construyamos después aguanta lo que le echen. Y como en toda obra que se precie, conviene tener claros los ladrillos con los que vamos a trabajar. Ahí van:

1) **Población** (que los expertos escriben con N mayúscula): No es la gente que vive en tu ciudad, ojo. En estadística, la población es el grupo entero que nos interesa estudiar, ya sean personas, empresas, votantes o mascotas. Lo importante es definirlo bien: ¿quiénes son, de dónde son, y de qué época hablamos? Si no acotas eso, todo lo demás cojea.
2) **Unidad de análisis**: Es cada pieza individual de ese puzle. Si tu población son los enfermeros de Lima, la unidad de análisis es Pepe, María, cada uno con su historia y su encuesta. El protagonista, vamos.
3) **Marco muestral**: Piénsalo como la guía telefónica de antes, pero puesta al día. Es la lista, el censo, el registro del que sacamos a nuestra gente. Si la lista está incompleta o desactualizada, podemos dar por bueno un estudio que no lo es. De la calidad del marco muestral depende medio chiste.
4) **Parámetro**: Este es el dato que nos quita el sueño, el que de verdad queremos conocer pero casi nunca podemos tocar. Es el valor real de la población, ese que los matemáticos visten con letras griegas como μ (la media) o π (la proporción). El tesoro escondido.

5) **Estadística**: Y esta es nuestra pala para encontrarlo. Como no podemos medir a toda la población, cogemos una muestra, calculamos con ella la media la llamamos $\bar{x}$, la proporción p y a partir de ahí hacemos nuestras apuestas. Es nuestra mejor aproximación al tesoro.

Con esto ya tenemos el equipo básico. Ahora sí, podemos dar el siguiente paso y meternos de lleno en el arte de elegir bien la muestra. Que de eso, de cómo hacer para que ese puñado de personas refleje fielmente a toda la población, va el Capítulo II. Y créeme, ahí empieza lo bueno.

CAPÍTULO II: EL ARTE DE ELEGIR: TÉCNICAS DE MUESTREO

En el capítulo anterior dejamos todo listo: ya sabemos qué es eso de la población, entendemos el papel de la muestra y por qué un puñado bien elegido de personas puede contarnos tanto sobre la multitud que no vimos. Pero seguro que hay una pregunta que te viene dando vueltas desde entonces, algo así como la mosca detrás de la oreja. Vale, muy bonita la teoría, pero ¿cómo se hace eso en la práctica? ¿Cómo se elige a esa gente que va a hablar por todos los demás? ¿Basta con salir a la calle, parar al primero que pasa y ya? ¿O hay un método, una especie de ciencia oculta, detrás de todo esto?

Pues mira, la respuesta tiene miga. Resulta que el muestreo es un poco de dos mundos: es ciencia y es arte al mismo tiempo. Es ciencia porque, no te voy a engañar, detrás hay matemáticas de las de verdad, probabilidades, fórmulas que permiten sentarse y decir: "con esta muestra, nuestro margen de error es de tantos puntos". Eso no es magia, es estadística. Pero también es arte, y de eso se habla menos, porque el investigador tiene que ir tomando decisiones sobre la marcha, adaptándose a lo que tiene, al dinero que le queda, al tiempo que le aprieta y a las vueltas que da la vida real, que nunca se parece a los manuales.

Como decía un sabio en esto del muestreo, el colombiano Andrés Gutiérrez, esto es casi una filosofía de vida: la idea de que no hace falta conocerlo todo para entenderlo todo. Suena a frase de coctel, pero encierra una verdad profunda. El truco no está en abarcarlo todo, que es imposible, sino en elegir bien esa porción de realidad para que, cuando la mires, lo que veas sea un reflejo fiel del conjunto.

Así que en este capítulo nos vamos a meter de lleno en el ajo. Vamos a ver las distintas maneras de pescar muestras, cada una con sus trucos, sus ventajas y sus trampas. Vamos a aprender a calcular cuánta gente necesitamos ese dolor de cabeza típico de quien empieza y, sobre todo, vamos a entrenar el ojo crítico. Porque cuando termines estas páginas, no volverás a leer una encuesta del periódico con los mismos ojos. Empezarás a fijarte en la letra pequeña, en cómo eligieron a la gente, en si la muestra es decente o te quieren colar un gol. Y créeme, una vez que aprendes a mirar así, ya no hay vuelta atrás.

2.1 Conceptos Básicos: La Parte por el Todo (Conceptos Básicos del Muestreo)

La experiencia con el muestreo es un hecho corriente que ocurre cotidianamente, aunque muchas veces no seamos conscientes de ello. Baste

observar cómo un docente puede verificar el conocimiento de sus alumnos sobre una materia determinada preguntando aleatoriamente a algunos de ellos, o cómo un médico detecta las condiciones de salud de un paciente a través de una serie de exámenes clínicos (análisis de sangre, orina, etc.) que no son más que muestras de un universo mucho más complejo. Se pueden mencionar otros ejemplos que usan procedimientos muestrales más sofisticados, pero todos tienen el mismo objetivo: obtener información sobre el todo basándonos en el conocimiento de una parte (Narayan & Sinha, 2023).

2.1.1 ¿Por qué no estudiar a todos? La magia de la muestra

La pregunta parece obvia, pero merece una reflexión profunda. Si queremos saber algo sobre una población, ¿por qué conformarnos con una parte? Las razones son múltiples y convincentes:

1) **Costo:** Estudiar a toda una población (realizar un censo) es extraordinariamente caro. El Instituto Nacional de Estadística e Informática (INEI) invierte miles de millones de soles en la realización de los censos nacionales, y eso ocurre solo cada varios años. Para un investigador individual o una pequeña empresa, un censo es sencillamente imposible.
2) **Tiempo:** Recolectar información de millones de personas lleva meses o incluso años. Para cuando se completara el censo, la información del principio ya estaría desactualizada. Las muestras, por el contrario, pueden proporcionar información oportuna y relevante en cuestión de días o semanas.
3) **Accesibilidad:** En muchos casos, es físicamente imposible acceder a todos los miembros de la población. Pensemos, por ejemplo, en un

estudio sobre la calidad del agua de todos los ríos de la Amazonía peruana. Simplemente no hay manera de llegar a cada recodo de cada río.

4) **Precisión:** Contraintuitivamente, estudiar una muestra bien seleccionada puede ser más preciso que estudiar a toda la población. Esto se debe a que, al trabajar con menos elementos, podemos dedicar más recursos y cuidado a cada medición, reduciendo los errores no muestrales (errores en la recolección, procesamiento o análisis de los datos).

2.1.2 El Riesgo de Equivocarse: El Error de Muestreo

Cuando tomamos una decisión basada en la información de una muestra, siempre existe el riesgo de cometer un error. Este riesgo se conoce como error de muestreo y es inherente a cualquier estudio que no examine a la población completa.

Para entenderlo mejor, consideremos el siguiente ejemplo propuesto por (Ciro Martínez, 2012): supongamos que se está experimentando con dos tipos de medicamentos para tratar la artritis reumatoide. Uno de ellos es el medicamento tradicional y el otro es un fármaco nuevo. Se selecciona una muestra de 40 pacientes con características similares y se asignan aleatoriamente 20 pacientes al tratamiento tradicional y los otros 20 al medicamento nuevo. Luego de un tiempo, los pacientes son evaluados y se observa que el 70% de los pacientes del grupo tradicional muestran mejoría, mientras que en el grupo del nuevo medicamento la mejoría alcanza el 74%.

Ante estos resultados, podríamos sentirnos tentados a concluir que el nuevo medicamento es mejor que el tradicional. Sin embargo, también podría ocurrir que ambos sean igualmente efectivos y que la diferencia

observada (apenas un 4%) sea debida simplemente a la casualidad o a las fluctuaciones propias del muestreo. Este riesgo de conclusiones erradas puede ser medido siempre que el muestreo sea probabilístico, es decir, siempre que hayamos utilizado métodos que nos permitan calcular la probabilidad de que nuestros resultados se deban al azar (Lohr, 2022).

La relación entre estos conceptos puede visualizarse de la siguiente manera:

El uso inadecuado de un procedimiento muestral puede llevar a un sesgo (error sistemático) en la interpretación de los resultados. Por ejemplo, si un docente solo pregunta a los estudiantes que se sientan en las primeras filas para verificar la comprensión de un tema, estará introduciendo un sesgo, pues probablemente esos estudiantes son los más atentos o participativos. Aun cuando el deseo del investigador es usar muestras que produzcan resultados confiables y libres de sesgos, en estudios sofisticados donde la información se obtiene a través de técnicas de muestreo, es común que el investigador quede tan entusiasmado por la prisa y la interpretación de los

datos que olvide verificar posibles sesgos originarios del diseño de la muestra (Arnab, 2017).

La teoría del muestreo describe las características, ventajas y desventajas de los diferentes diseños muestrales. Los conceptos no son triviales y es importante que sean establecidos correctamente para el uso científico de los procedimientos muestrales. Para ilustrar los conceptos que vamos a mencionar, utilizaremos un ejemplo a lo largo de todo el capítulo.

Ejemplo conductor:

Estamos interesados en averiguar cuál es la prevalencia de asma en niños de 5 a 15 años en el distrito de Chosica. Este ejemplo nos acompañará para entender cada uno de los conceptos.

Concepto	Definición	Aplicación al ejemplo
Unidad de análisis	Objeto o sujeto en el cual se realizan las mediciones	Cada niño entre 5 y 15 años que vive en el distrito de Chosica
Población	Conjunto total de individuos que poseen la característica de interés	Todos los niños entre 5 y 15 años que viven en el distrito de Chosica (N = ?)
Marco muestral	Lista de unidades de muestreo	Relación de hogares del distrito de Chosica con direcciones y nombre del jefe de hogar (obtenible del INEI)
Unidad de muestreo	Unidad seleccionada en el marco muestral	El hogar (luego, dentro del hogar, se entrevista a los niños que cumplan los criterios)
Muestra	Subconjunto representativo de la población	Conjunto de hogares seleccionados y, dentro de ellos, los niños que serán entrevistados (n = ?)
Parámetro	Medida descriptiva de la población	Prevalencia de asma en niños de 5 a 15 años en Chosica (π)
Estadística	Medida descriptiva de la muestra	Prevalencia de asma en la muestra de niños (p)

2.2 Tipos de Muestreo: Dos Caminos Posibles

La primera gran decisión que debe tomar un investigador es si utilizará un muestreo probabilístico o un muestreo no probabilístico. Esta determinación va a depender de los objetivos del estudio, la disponibilidad de recursos y, fundamentalmente, del alcance de las conclusiones que se deseen obtener.

La siguiente figura, adaptada de (Hernández Sampieri & Mendoza Torres, 2018), resume los tipos de muestreo que exploraremos:

2.2.1 Muestreo No Probabilístico: Cuando la Accesibilidad es Clave

El muestreo no probabilístico, también denominado muestreo intencional, muestreo dirigido o muestreo por juicio, se caracteriza porque el

procedimiento de selección de la muestra se realiza de manera informal y, en cierta medida, arbitraria. Esto conduce a que no todos los elementos de la población tengan una probabilidad conocida de ser seleccionados (Roldán & Oliva, 2015).

La principal ventaja del muestreo no probabilístico es su utilidad práctica en ciertos estudios, especialmente en aquellos de carácter exploratorio o cualitativo, donde no es indispensable que las muestras sean representativas de la población, sino que reúnan ciertas características previamente especificadas. Sin embargo, tiene una desventaja fundamental: no podemos evaluar el riesgo de decisiones y conclusiones erradas. Las inferencias realizadas con este tipo de muestras no tienen validez estadística para generalizar a toda la población, aunque pueden ser perfectamente válidas para los fines específicos del estudio (Klinger Angarita, 2024).

Los tipos más comunes de muestras no probabilísticas son:

a) Muestras de sujetos voluntarios

Son frecuentes en estudios donde los sujetos acceden voluntariamente a participar. Este tipo de muestreo es común en investigaciones médicas con nuevos tratamientos, donde los pacientes se ofrecen como voluntarios, o en estudios psicológicos que reclutan participantes mediante anuncios.

Ejemplo: Supongamos que un investigador afirma haber descubierto un medicamento que cura el SIDA. Un grupo de personas con la enfermedad se ofrece voluntariamente para recibir tratamiento con este nuevo fármaco. Los resultados, por muy positivos que sean, no podrán generalizarse a toda la población con SIDA, pues los voluntarios suelen tener características

diferentes (mayor motivación, mejor acceso a la información, etc.) que el promedio de los pacientes.

b) Muestra de expertos

Se utiliza cuando es necesaria la opinión de personas que son especialistas en un tema. Este tipo de muestreo es muy frecuente en estudios cualitativos, en investigaciones exploratorias y, como veremos en el Capítulo IV, en procesos de validación de instrumentos mediante juicio de expertos.

Ejemplo: Para validar un cuestionario sobre satisfacción laboral, recurrimos a un grupo de psicólogos organizacionales con amplia trayectoria para que evalúen si las preguntas son pertinentes y están bien formuladas.

c) Muestra por cuotas

Es una técnica muy utilizada en estudios de opinión y de mercados. Los encuestadores reciben instrucciones de entrevistar a un número determinado de personas (la cuota) que cumplan con ciertas características (por ejemplo, 50 mujeres de 30 a 40 años, 50 hombres de 30 a 40 años, etc.), pero la selección dentro de cada cuota queda a criterio del encuestador.

Ejemplo: Una empresa de cosméticos quiere conocer la opinión de las mujeres limeñas sobre una nueva crema facial. Establece cuotas por distritos: 100 encuestas en Miraflores, 100 en San Juan de Lurigancho, 100 en Comas, etc. Dentro de cada distrito, los encuestadores eligen a las mujeres que encuentran en la calle o en centros comerciales.

d) Estudio de casos

Este diseño muestral es muy utilizado en el área médica y en psicología clínica, donde el investigador, interesado en un tema particular, va registrando sistemáticamente los casos que van llegando a su consulta o servicio.

Ejemplo: Un neurólogo interesado en una enfermedad rara va documentando las características de todos los pacientes que llegan a su consulta con ese diagnóstico a lo largo de cinco años.

2.2.2 Muestreo Probabilístico: La Ciencia de la Representatividad

En el muestreo probabilístico, cada unidad de muestreo tiene una probabilidad conocida y mayor que cero de ser seleccionada para formar parte de la muestra. Dado que la probabilidad de seleccionar cada elemento es conocida, el investigador puede utilizar las diversas reglas y leyes de la probabilidad para evaluar la confiabilidad de las conclusiones que se obtengan a partir de las muestras (Andrés Gutiérrez, 2016).

En otras palabras, cuando una muestra es probabilística, el riesgo de decisiones y conclusiones incorrectas (el error de muestreo) puede ser cuantificado. Esta es la gran ventaja de este tipo de muestreo y la razón por la cual es el estándar de oro en la investigación científica.

El riesgo de decisiones y conclusiones incorrectas es inherente al muestreo, pero puede ser controlado si se planea cuidadosamente el diseño muestral y se respetan las reglas del muestreo probabilístico. En nuestro ejemplo sobre la prevalencia de asma en niños de Chosica, garantizamos la validez externa de nuestro estudio si obtenemos una muestra probabilística.

2.2.3 Características de un Marco Muestral

Antes de profundizar en los tipos de muestreo probabilístico, es necesario detenernos en un elemento crucial: el marco muestral. Como mencionamos en el Capítulo I, el marco muestral es la lista de todas las unidades de muestreo de las que se extraerá la muestra.

Las principales características de un buen marco muestral son:

1) **Completitud:** Toda unidad de muestreo de la población debe estar presente en el marco. Si faltan elementos, se produce un error de cobertura que sesga los resultados.
2) **Identificación clara:** Cada unidad debe tener un identificador lógico y único (un código, un número, un nombre completo y dirección).
3) **Organización lógica:** El marco debe estar organizado de manera sistemática (alfabéticamente, geográficamente, numéricamente) para facilitar la selección.
4) **Actualización:** Para que funcione correctamente, debe estar actualizado. Un marco muestral desactualizado (por ejemplo, un directorio telefónico de hace cinco años) introducirá sesgos importantes.
5) **Relevancia:** Debe contener toda la información relevante sobre las unidades para facilitar el diseño muestral (por ejemplo, si vamos a estratificar por distritos, el marco debe incluir esa información).

Cuando decidimos seleccionar una muestra aleatoria, surgen de inmediato dos preguntas fundamentales:

1) ¿Cuántos individuos son necesarios para que la muestra represente adecuadamente a la población? (el tamaño de la muestra)

2) ¿Cómo se deben seleccionar los individuos que conformarán la muestra? (el método de selección)

Las respuestas a ambas preguntas deben estar contenidas en el diseño muestral. A continuación, abordaremos la primera cuestión.

2.3 ¿De qué Tamaño la Hago? Calculando la Muestra Necesaria

La pregunta que surge de inmediato cuando se decide hacer una investigación por muestreo probabilístico es: ¿cuántos sujetos debo considerar en la muestra? O, dicho de otro modo, ¿cuál es el tamaño de muestra adecuado?

La respuesta, aunque parezca evasiva, es: ¡depende! Y depende de varios factores que podemos clasificar en prácticos y estadísticos.

2.3.1 Factores que Influyen en el Tamaño: Confianza, Error y Variabilidad

Factores prácticos:

- **Presupuesto disponible:** ¿con cuánto dinero contamos para realizar el estudio?
- **Tiempo disponible:** ¿de cuánto tiempo disponemos para recolectar los datos y presentar los resultados?
- **Tamaño de la población:** ¿qué tan grande es la población sobre la cual queremos hacer inferencias? (poblaciones más grandes requieren muestras más grandes, pero esta relación no es lineal).

Factores estadísticos:

- **Heterogeneidad de la población:** ¿qué tan variable es la característica que queremos medir? La varianza poblacional (σ^2) es un indicador clave. Si la población es homogénea (varianza pequeña), bastará una muestra pequeña para captar su variabilidad. Por el contrario, si la población es muy heterogénea (varianza grande), se requerirá una muestra grande para capturar toda su diversidad (Nencini, 2022).
- **Margen de error tolerado:** ¿qué tanta diferencia estamos dispuestos a aceptar entre el valor de la muestra y el verdadero valor poblacional? Este margen, denotado como ε (épsilon), es la máxima discrepancia que toleramos:

  Error = | Estadístico - Parámetro |

  Cuanto menor sea el error que toleramos, mayor deberá ser el tamaño de la muestra.

- **Nivel de confianza:** ¿qué tan seguros queremos estar de que nuestro margen de error se cumple? El nivel de confianza ($1-\alpha$) se expresa usualmente como 95% o 99%. Un nivel de confianza del 95% significa que si repitiéramos el estudio 100 veces, en 95 de ellas el intervalo de confianza contendría el verdadero valor poblacional. Cuanto mayor sea el nivel de confianza deseado, mayor deberá ser la muestra.

La relación entre estos elementos se expresa en el coeficiente de confianza Z_{α/2}, que es el valor en la distribución normal estándar que deja por debajo de sí una probabilidad de $1-\alpha/2$. Los valores más comunes son:

Nivel de Confianza	α (significación)	Z_{α/2}
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.58

Una vez respondidas estas cuestiones, podemos recurrir a la teoría del muestreo, que nos proporciona las fórmulas matemáticas para determinar el tamaño adecuado de muestra (Narayan & Sinha, 2023).

2.3.2 Tamaño de Muestra para Estimar una Media Poblacional (μ)

Cuando nuestro objetivo es estimar el valor promedio de alguna característica cuantitativa (como la talla, el peso, los ingresos económicos, etc.), utilizamos las siguientes fórmulas.

Para poblaciones finitas (cuando conocemos N):

$$n = (N * Z^2 * \sigma^2) / (\varepsilon^2 * (N - 1) + Z^2 * \sigma^2)$$

Donde:

N: Tamaño de la población

Z: Coeficiente de confianza (1.96 para 95% de confianza)

σ^2: Varianza poblacional (si no se conoce, se estima mediante una prueba piloto o estudios previos)

ε: Error máximo permisible (precisión deseada)

Para poblaciones infinitas (cuando N es muy grande o desconocido):

$$n = (Z^2 * \sigma^2) / \varepsilon^2$$

¿Cómo estimar σ^2 cuando no la conocemos?

En la práctica, rara vez conocemos la varianza poblacional. Podemos estimarla de tres maneras:

1) **Revisión bibliográfica:** Buscar estudios previos similares que reporten la desviación estándar.
2) **Estudio piloto:** Realizar una pequeña prueba con 30-50 sujetos para estimar la varianza.
3) **Rango aproximado:** Si la variable tiene distribución aproximadamente normal y conocemos sus valores mínimo y máximo, podemos estimar la desviación estándar como:

   $s \approx (\text{Valor máximo} - \text{Valor mínimo}) / 4$

   Este método se basa en que, en una distribución normal, el 95% de los datos se encuentran aproximadamente dentro de ±2 desviaciones estándar de la media.

Ejemplo práctico:

Queremos estimar el peso promedio de los niños de 5 a 15 años en Chosica, con un nivel de confianza del 95% y un margen de error de ±2 kg. No conocemos la varianza, pero un estudio piloto con 30 niños arrojó una desviación estándar de 5 kg.

Datos:

$Z = 1.96$ (para 95% de confianza)
$s = 5$ kg (estimación de σ)
$\varepsilon = 2$ kg

Como la población es grande (no conocemos N exacto, pero sabemos que es > 100,000), usamos la fórmula para poblaciones infinitas:

$n = (1.96^2 * 5^2) / 2^2 = (3.8416 * 25) / 4 = 96.04 / 4 = 24.01 \approx 25$ niños

Esto significa que necesitamos una muestra de al menos 25 niños para estimar el peso promedio con la precisión deseada.

2.3.3 Tamaño de Muestra para Estimar una Proporción Poblacional (π)

Cuando nuestro objetivo es estimar un porcentaje o proporción (como la prevalencia de asma en nuestro ejemplo), las fórmulas son similares, pero incorporan la variabilidad propia de las proporciones.

Para poblaciones finitas:

$n = (N * Z^2 * p * q) / (\varepsilon^2 * (N - 1) + Z^2 * p * q)$

Para poblaciones infinitas:

$n = (Z^2 * p * q) / \varepsilon^2$

Donde:

- p: Proporción estimada de elementos que poseen la característica de interés
- q: 1 - p (proporción que no posee la característica)

¿Cómo estimar p cuando no la conocemos?

Al igual que con la media, podemos estimar p de tres maneras:

1) **Revisión bibliográfica:** Buscar estudios previos que reporten proporciones similares.
2) **Estudio piloto:** Realizar una prueba pequeña para tener una estimación inicial.
3) **Asumir el caso más desfavorable:** Cuando no tenemos ninguna información, asumimos p = 0.5 y q = 0.5, pues este es el escenario que maximiza el producto p*q (0.5 * 0.5 = 0.25) y, por tanto, nos da el tamaño de muestra más grande y conservador.

Ejemplo práctico (nuestro caso):

Queremos estimar la prevalencia de asma en niños de 5 a 15 años en Chosica, con un nivel de confianza del 95% y un margen de error de ±3% (0.03). No tenemos estudios previos, así que asumimos el caso más desfavorable (p = 0.5). Según el INEI, la población de niños en ese rango de edad en Chosica es aproximadamente N = 15,000.

Datos:

- N = 15,000
- Z = 1.96
- p = 0.5; q = 0.5
- ε = 0.03

Primero, calculamos el numerador y denominador por separado:

Numerador = $N * Z^2 * p * q = 15{,}000 * (1.96)^2 * 0.5 * 0.5$

= 15,000 * 3.8416 * 0.25

= 15,000 * 0.9604

= 14,406

Denominador = $\varepsilon^2 * (N - 1) + Z^2 * p * q$

= $(0.03)^2 * 14,999 + 3.8416 * 0.25$

= 0.0009 * 14,999 + 0.9604

= 13.4991 + 0.9604

= 14.4595

Por lo tanto:

n = 14,406 / 14.4595 ≈ 996 niños

Necesitamos una muestra de aproximadamente 1,000 niños para estimar la prevalencia de asma con el nivel de precisión deseado.

Nota importante: Si no conociéramos N (población infinita), el cálculo sería:

n = $(1.96^2 * 0.5 * 0.5) / 0.03^2$ = (3.8416 * 0.25) / 0.0009 = 0.9604 / 0.0009 = 1,067 niños

Observemos que el tamaño es ligeramente mayor cuando no consideramos el factor de corrección por población finita.

2.4 Métodos de Selección de la Muestra: Poniendo en Práctica la Teoría

Una vez que hemos determinado cuántos individuos necesitamos, la siguiente pregunta es cómo seleccionarlos. La elección del método de selección depende de la forma como se encuentran organizadas geográfica o administrativamente las unidades de muestreo, así como de la disponibilidad del marco muestral.

2.4.1 Tipos de Muestreo Probabilístico

La teoría del muestreo nos proporciona varios diseños muestrales probabilísticos, cada uno con sus propias características, ventajas y limitaciones.

A) Muestreo Aleatorio Simple (MAS)

El muestreo aleatorio simple es el método más básico e importante para la selección de una muestra. Operacionalmente, podemos definirlo como: dada una lista de N unidades elementales (la población), se sortean

con igual probabilidad n unidades que conformarán la muestra (Narayan & Sinha, 2023).

Procedimiento:

1) Confeccionar una lista completa de todos los individuos de la población: el marco muestral.
2) Determinar el tamaño de la muestra n utilizando las fórmulas correspondientes.
3) Utilizando un mecanismo aleatorio (tabla de números aleatorios, programas informáticos, urnas), sortear con igual probabilidad las unidades que pasarán a formar parte de la muestra.
4) Repetir el proceso hasta completar las n unidades necesarias.

Formas de selección:

- **Con reemplazo:** Un elemento puede ser seleccionado más de una vez. Después de cada extracción, se devuelve el elemento a la población. Es teóricamente más simple pero poco usado en la práctica.
- **Sin reemplazo:** Los elementos seleccionados no se devuelven a la población. Es la forma más común en la investigación real.

Ventajas del MAS	Desventajas del MAS
Sencillo y de fácil comprensión	Requiere un listado completo de toda la población (marco muestral)
Cálculo rápido de medias y varianzas	Si la población está muy dispersa, puede resultar costoso
Existen software para analizar los datos	Con muestras pequeñas, puede no representar adecuadamente a subgrupos importantes
Base teórica sólida para la inferencia	Puede ser ineficiente si la población tiene estructura

La tabla de números aleatorios:

Independientemente del tipo de muestreo probabilístico, la tabla de números aleatorios nos permite seleccionar con igual probabilidad una muestra de cualquier tamaño. Para usarla correctamente:

1) Conocer el tamaño de la población (N) y determinar cuántos dígitos tiene (ej., si N = 3,000, necesitamos números de 4 dígitos).
2) Establecer un punto de inicio aleatorio en la tabla.
3) Definir una regla de seguimiento (vertical, horizontal, o combinada).
4) Seleccionar los números que estén dentro del rango 1 a N.
5) Identificar en el marco muestral los elementos correspondientes a los números seleccionados.

Ejemplo de tabla de números aleatorios:

10480	15011	01536	02011	81647	91646	69179	14194	62590	36207
22368	46573	25595	85393	30995	89198	27982	53402	93965	34095
24130	48360	22527	97265	76393	64809	15179	24830	49340	32081

B) Muestreo Sistemático

El muestreo sistemático es una alternativa al MAS que puede ser más práctica cuando el marco muestral está ordenado de alguna manera. El procedimiento exige numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios, solo se extrae un número de arranque y a partir de él se selecciona cada k-ésimo elemento (Andrés Gutiérrez, 2016).

Procedimiento:

1) Determinar el tamaño de la población (N) y el tamaño de la muestra (n).
2) Calcular el intervalo de muestreo:

   k = N / n (redondeado al entero más cercano)

3) Seleccionar aleatoriamente un número de arranque (i) entre 1 y k.
4) Los elementos que conforman la muestra son: i, i + k, i + 2k, i + 3k, ..., hasta completar n elementos.

Ejemplo práctico:

Supongamos que nuestra población está compuesta por 360 familias y hemos determinado que necesitamos una muestra de 30 familias.

1) Calculamos k = 360 / 30 = 12
2) Elegimos un número de arranque aleatorio entre 1 y 12. Usando la tabla de números aleatorios, obtenemos i = 10.
3) Las familias seleccionadas serán: 10, 22, 34, 46, 58, 70, 82, 94, 106, 118, 130, 142, 154, 166, 178, 190, 202, 214, 226, 238, 250, 262, 274, 286, 298, 310, 322, 334, 346, 358.

Ventaja principal: Es más fácil de implementar que el MAS cuando el marco muestral es una lista física. Precaución: Si la lista tiene algún patrón periódico que coincida con el intervalo k, se puede introducir un sesgo (por ejemplo, si la lista alterna sistemáticamente hombre-mujer y k es un número par).

C) Muestreo Estratificado

El muestreo estratificado consiste en dividir previamente la población en subgrupos o estratos que son internamente homogéneos respecto a la característica de interés, pero diferentes entre sí. Luego, se selecciona una muestra independiente de cada estrato (Arnab, 2017).

¿Por qué estratificar?

1) **Precisión**: Si las mediciones dentro de cada estrato son homogéneas, la estratificación produce estimaciones más precisas que el MAS.
2) **Costo**: Se puede reducir el costo por observación al estratificar la población en grupos convenientes.
3) **Información por subgrupos**: Permite obtener estimaciones separadas para cada estrato (por ejemplo, estimar la prevalencia de asma por separado para niños y niñas).

Procedimiento:

1) Identificar la variable de estratificación (sexo, edad, nivel socioeconómico, distrito, etc.).
2) Dividir la población en estratos mutuamente excluyentes.
3) Determinar el tamaño de muestra para cada estrato.
4) Seleccionar una muestra aleatoria (simple o sistemática) dentro de cada estrato.

Tipos de asignación:

- **Asignación proporcional:** El tamaño de la muestra en cada estrato es proporcional al tamaño del estrato en la población. Es la más común.
  n_h = (N_h / N) * n
- **Asignación óptima:** Se asigna más muestra a los estratos que tienen mayor variabilidad (para reducir el error total) o menor costo de medición.

Ejemplo práctico:

Para nuestro estudio de asma en Chosica, podríamos estratificar por sexo (niños y niñas). Si la población total es 15,000, con 7,500 niños y 7,500 niñas, y nuestra muestra total es 1,000, la asignación proporcional nos daría:

n_{niños} = (7,500 / 15,000) * 1,000 = 500 niños

n_{niñas} = (7,500 / 15,000) * 1,000 = 500 niñas

D) Muestreo por Conglomerados

El muestreo por conglomerados (o clústeres) es la cuarta técnica de muestreo probabilístico y, de las cuatro, es la que puede introducir mayor sesgo. Sin embargo, está considerada dentro del muestreo probabilístico y es extremadamente útil en ciertas circunstancias (Lohr, 2022).

Consiste en identificar grupos de unidades de estudio, llamados conglomerados o clústeres, donde cada grupo presenta toda la variabilidad que se observa en la población. Es decir, los conglomerados son como "mini-poblaciones" que reflejan la diversidad del conjunto total.

Diferencia clave con el muestreo estratificado:

- En el estratificado, los estratos son internamente homogéneos y diferentes entre sí, y se toma muestra de todos los estratos.

- En el de conglomerados, los conglomerados son internamente heterogéneos (reflejan la diversidad poblacional) y se selecciona solo algunos conglomerados, estudiando luego todos (o una muestra) de sus elementos.

Procedimiento (dos etapas):

1) Primera etapa: Seleccionar aleatoriamente una muestra de conglomerados (por ejemplo, centros de salud, escuelas, manzanas de viviendas).
2) Segunda etapa (opcional): Dentro de cada conglomerado seleccionado, se puede estudiar a todos los elementos o seleccionar una muestra aleatoria de ellos.

Ejemplo práctico:

Retomemos nuestro estudio sobre asma en niños de Chosica. Si quisiéramos utilizar muestreo por conglomerados:

1) Identificamos los conglomerados: podrían ser las instituciones educativas del distrito.
2) Seleccionamos aleatoriamente, por ejemplo, 10 de las 50 escuelas primarias de Chosica.
3) Dentro de cada escuela seleccionada, entrevistamos a todos los niños de 5 a 15 años (muestreo por conglomerados de una etapa) o seleccionamos aleatoriamente algunos niños de cada escuela (muestreo por conglomerados de dos etapas).

Ventajas:

- Muy útil cuando las unidades están geográficamente dispersas.
- Reduce significativamente los costos de traslado y logística.
- No requiere un marco muestral de todas las unidades individuales (solo de conglomerados).

Desventajas:

- Mayor error de muestreo que el MAS para el mismo tamaño de muestra.
- Los cálculos de precisión son más complejos.
- Si los conglomerados no son realmente heterogéneos, el sesgo puede ser considerable.

2.5 Consideraciones Éticas en el Muestreo

Antes de cerrar este capítulo, es importante reflexionar sobre las implicaciones éticas del muestreo. La forma en que seleccionamos nuestra muestra no es neutral; tiene consecuencias sobre quiénes son incluidos en los estudios y, por tanto, sobre qué conocimientos se generan y para quiénes (Israel, 2015).

Principios éticos fundamentales:

1) **Inclusión y representatividad:** Evitar sistemáticamente excluir a ciertos grupos (por razones de accesibilidad, idioma, etc.) puede perpetuar desigualdades en el conocimiento. Un estudio sobre salud que solo incluya a personas que viven cerca de hospitales urbanos no nos dirá nada sobre la salud en zonas rurales.

2) **Consentimiento informado**: Los participantes deben ser informados de que forman parte de una muestra y de los fines de la investigación. La aleatoriedad no exime del deber de obtener consentimiento.
3) **Confidencialidad:** La información obtenida de la muestra debe ser protegida. El hecho de que trabajemos con muestras no significa que los datos sean menos sensibles.
4) **Devolución de resultados:** Siempre que sea posible, los resultados del estudio deberían ser accesibles a las comunidades que participaron en la muestra.

Lo que aprendimos en este capítulo (y no deberías olvidar)

Hemos cubierto mucho terreno en estas páginas, así que vamos a parar un momento, tomar aire y repasar lo esencial. Porque una cosa es leerlo y otra muy distinta es quedarse con lo que de verdad importa. Ahí van las ideas clave:

1) Las muestras no probabilísticas: rápidas, útiles, pero con límites

Son esas que se eligen sin sorteos ni azar: el investigador va a lo que tiene a mano (conveniencia), busca cuotas de gente con cierto perfil, o recurre a expertos y voluntarios. Funcionan para tanteos, para estudios exploratorios, para entender cómo piensa un grupo concreto. Pero ojo, con ellas no puedes salir en la tele diciendo "el 60% de los peruanos piensa que...". No generalizan y no permiten calcular el margen de error. Punto.

2) Las muestras probabilísticas: el reino del azar bien entendido

Aquí cada persona tiene una probabilidad conocida de salir elegida, como en una tómbola pero con método. Esto es lo que permite hacer inferencias, es decir, hablar de la población con fundamento y, además, ponerle números al

error que podemos estar cometiendo. Si lo que buscas es representatividad de verdad, este es tu camino.

3) El tamaño sí importa (pero no es lo único que importa)

¿Cuánta gente necesito? La respuesta clásica: depende. Depende de la confianza que quieras tener (lo típico es 95%), del margen de error que estés dispuesto a tolerar (si es muy pequeño, necesitarás más gente) y de lo variada que sea la población. Además, no es lo mismo calcular el tamaño para sacar una media por ejemplo, el ingreso promedio que para estimar una proporción como el porcentaje que votaría por alguien. Las fórmulas son distintas, y en el capítulo las hemos visto con calma.

4) Cuatro formas de elegir bien, cada una con su aquel

- El **muestreo aleatorio simple** es el rey, el que todos tenemos en la cabeza: numeras a todos y haces un sorteo. El problema es que necesitas la lista completa, y eso no siempre se tiene.
- El **sistemático** es un primo práctico: ordenas a la gente y vas cogiendo uno de cada tantos. Más rápido, pero con algún que otro riesgo si la lista tiene patrones ocultos.
- El **estratificado** es para quienes quieren asegurarse de que ningún grupo se quede fuera. Divides la población por capas (por distritos, por edades...) y sorteas dentro de cada una. Ganas precisión y puedes comparar grupos.
- El **muestreo por conglomerados** es el comodín para cuando la gente está muy dispersa. En lugar de ir casa por casa, eliges manzanas enteras, colegios enteros, y dentro de ahí encuestas a todos o a unos pocos. Ahorras viajes, pero pierdes un poco de puntería.

5) Y nunca, nunca, olvidar la ética

Esto no es un detalle menor. Elegir la muestra también implica decidir a quién dejamos fuera y a quién incluimos. Hay que ser justos, respetar que la gente decida si participa o no, y cuidar sus datos como si fueran nuestros. Porque la estadística, cuando se hace bien, también es una forma de respeto.

Ahora que ya sabemos cómo elegir una muestra que sea un espejo fiel de la población, toca dar el siguiente paso. En el próximo capítulo vamos a meternos en otro jardín: una vez que tenemos a nuestra gente seleccionada, ¿cómo nos aseguramos de que las preguntas que les hacemos realmente sirvan para medir lo que queremos medir? Porque de nada sirve preguntarle a la persona indicada si la pregunta está mal hecha. Ahí nos vemos.

CAPÍTULO III: MIDIENDO LO INVISIBLE: INTRODUCCIÓN A LA MEDICIÓN EN CIENCIAS SOCIALES

Pongámonos en situación. Imagina que necesitas saber la estatura de un estudiante. Coges un metro, lo pones desde el suelo hasta el último pelo, y listo: 1,75. No hay discusión, no hay interpretación. La altura es de esas cosas que están ahí, se tocan, se ven, y todo el mundo acepta que el metro mide lo que tiene que medir.

Pero la vida rara vez es tan sencilla. Porque, ¿cómo le haces cuando lo que te interesa no se puede tocar? Ahí está el busilis de la cuestión. Piensa en cosas como la inteligencia. Nadie ha visto nunca una inteligencia paseándose por la calle, ¿verdad? Vemos a alguien resolver un problema complejo, oírle hablar con lucidez, y decimos: "este debe ser inteligente". Pero lo que vimos fueron manifestaciones, pistas, comportamientos. La inteligencia en sí misma es una especie de fantasma que asumimos que existe porque vemos sus huellas.

Lo mismo pasa con la satisfacción laboral de ese enfermero que trabaja turnos interminables. No podemos abrirle el pecho y ver un cartelito que ponga "satisfacción: 7 sobre 10". Lo que hacemos es preguntarle, observar su cara, notar si llega contento o arrastrando los pies. O con la actitud de un profesor hacia las nuevas tecnologías en el aula. O con la calidad de vida de alguien que vive con una enfermedad crónica. Todo eso son constructos, que es como los expertos llaman a estas realidades invisibles. También se dice variables latentes, porque están ahí, latiendo, pero escondidas.

Y ahí tenemos el dilema: ¿cómo mides lo que no se ve? ¿Cómo pones número a lo que no tiene cuerpo? Pues buscando sus señales, sus manifestaciones observables. Lo que la gente dice, lo que hace, cómo responde cuando le preguntas (Fitzner, 2007). De eso, justamente, va este capítulo.

Este capítulo viene a ser como ese puente que conecta dos orillas: del lado de allá quedó el muestreo, con todo lo que aprendimos sobre cómo elegir a la gente indicada. Del lado de acá nos espera lo que viene después, eso de la validez y la confiabilidad que exploraremos a fondo en el Capítulo IV. Pero para cruzar de una orilla a otra hace falta un tramo intermedio, y de eso vamos a hablar ahora. Porque de nada sirve tener una muestra de ensueño si luego no sabemos cómo preguntar.

El meollo del asunto es el siguiente: en ciencias sociales, en psicología, en educación, en salud, casi siempre queremos medir cosas que no se ven. La inteligencia, la motivación, el estrés, la calidad de vida, la vocación de servicio... todo eso son fantasmas, maravillosos pero fantasmas al fin. Y el desafío está en convertir esos espectros en algo tangible, en números que podamos analizar, en datos que nos permitan comparar, concluir y, con un poco de suerte, entender mejor el mundo.

Como apunta un experto en esto (Furr, 2022), medir en estos terrenos es básicamente ponerle números a la realidad, pero con reglas muy claras. Suena fácil, pero no lo es. Porque esas reglas tienen que ser tan sólidas y tan bien explicadas que cualquier otro investigador, en cualquier otro país y dentro de diez años, pueda aplicarlas y obtener resultados que tengan sentido comparar con los tuyos. Si cada cual mide a su manera, el caos está asegurado.

Y aquí viene una de esas diferencias que parecen pequeñas pero que lo cambian todo. No es lo mismo soltarle a alguien un "oye, ¿y tú estás satisfecho con tu trabajo?" y anotar lo que dice, que construir todo un tinglado de preguntas, una escala bien pensada, que explore distintos ángulos de esa satisfacción. Lo primero puede valer para una charla de café. Lo segundo es lo que permite hacer ciencia. Cuando termines este capítulo, vas a entender por qué esa diferencia no es un capricho de especialistas, sino el corazón mismo de cualquier investigación que se precie.

3.1 De Conceptos Abstractos a Variables Medibles: El Proceso de Operacionalización

El primer gran desafío de cualquier investigador es convertir una idea abstracta en algo que pueda ser observado y medido. Este proceso se conoce como operacionalización y es, posiblemente, el paso más creativo y, a la vez, más riguroso de toda investigación (Hernández Sampieri & Mendoza Torres, 2018).

3.1.1 Conceptos, Constructos y Variables

Para entender la operacionalización, necesitamos distinguir tres niveles de abstracción:

1. Concepto: Es una idea general, una abstracción que utilizamos en el lenguaje cotidiano. Por ejemplo: "inteligencia", "felicidad", "violencia", "calidad educativa". Los conceptos son útiles para la comunicación diaria, pero son demasiado vagos para la investigación científica.

2. Constructo: Es un concepto que ha sido deliberadamente inventado o adoptado para un propósito científico. Los constructos son conceptos pero con una definición más precisa, aunque siguen siendo abstractos. Por ejemplo, la psicología ha desarrollado el constructo de "inteligencia emocional" para referirse a un conjunto específico de habilidades relacionadas con el manejo de las emociones (Perloff, 1997).

3. Variable: Es un constructo que ha sido operacionalizado, es decir, que hemos definido cómo se va a medir. Una variable puede tomar diferentes valores. Por ejemplo, la "inteligencia emocional" puede ser medida a través de un test que arroje una puntuación numérica; esa puntuación es la variable.

Ejemplo del proceso de operacionalización:

Nivel	Ejemplo 1	Ejemplo 2	Ejemplo 3
Concepto cotidiano	Inteligencia	Estrés	Calidad de vida
Constructo científico	Inteligencia lógico-matemática (Gardner)	Estrés laboral (Karasek)	Calidad de vida relacionada con la salud (WHOQOL)
Variable (medición)	Puntuación en el test de Raven	Puntuación en la escala de demandas-control	Puntuación en el cuestionario WHOQOL-BREF

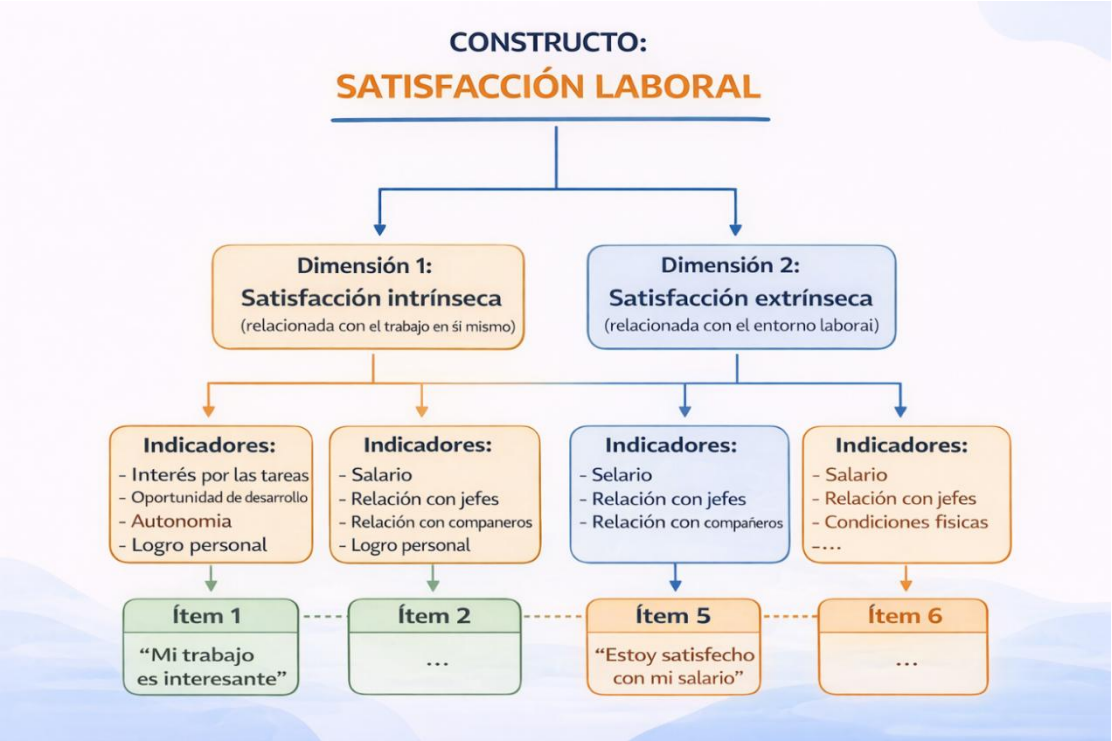
3.1.2 Dimensiones e Indicadores

La mayoría de los constructos en ciencias sociales son multidimensionales, es decir, están compuestos por varios aspectos o facetas. Por ejemplo, la "calidad de vida" no es una sola cosa; incluye dimensiones físicas, psicológicas, sociales y ambientales.

Para medir adecuadamente un constructo, necesitamos:

- **Dimensiones**: Los grandes componentes o facetas del constructo.
- **Indicadores**: Manifestaciones específicas y observables de cada dimensión. Los indicadores son la base para formular las preguntas o ítems del instrumento de medición.

Ejemplo de operacionalización de "Satisfacción Laboral"



Adaptado de (Warr et al., 1979).

3.1.3 Niveles de Medición: La Escalera de Stevens

Una vez que hemos definido nuestras variables, necesitamos entender qué tipo de información proporcionan. El psicólogo Stanley Smith Stevens (1946) propuso una clasificación de los niveles de medición que sigue siendo fundamental hasta hoy:

1. Nivel nominal: Las categorías son mutuamente excluyentes y no tienen un orden inherente. Solo podemos contar frecuencias.

Ejemplo: Sexo (hombre/mujer), estado civil (soltero/casado/divorciado/viudo), religión.

2. Nivel ordinal: Las categorías tienen un orden, pero no sabemos la distancia exacta entre ellas.

Ejemplo: Nivel socioeconómico (bajo/medio/alto), grado de acuerdo (muy en desacuerdo/en desacuerdo/indiferente/de acuerdo/muy de acuerdo).

3. Nivel de intervalo: Hay un orden y la distancia entre categorías es conocida y constante, pero el cero es arbitrario (no significa ausencia de la propiedad).

Ejemplo: Temperatura en grados Celsius (0°C no significa ausencia de temperatura), coeficiente intelectual.

4. Nivel de razón: Tiene todas las propiedades del intervalo y, además, el cero es absoluto (significa ausencia de la propiedad).

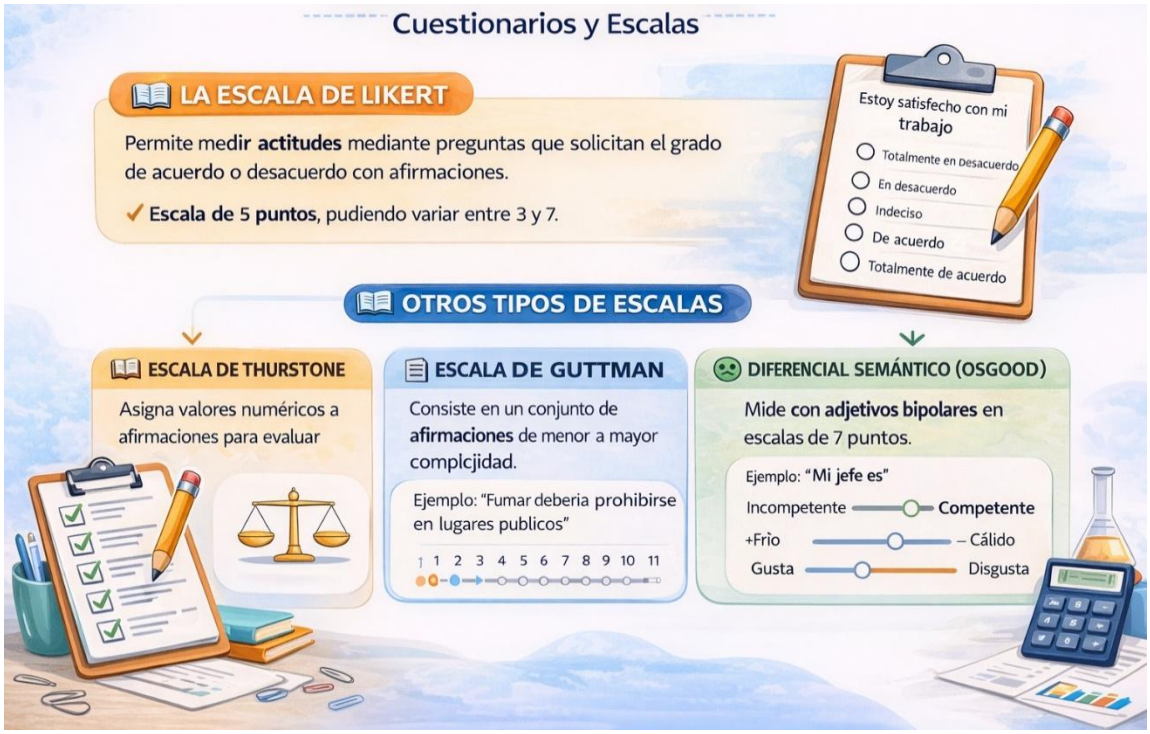
Ejemplo: Altura (0 cm significa ausencia de altura), peso, ingresos en soles, edad.

Importancia de esta clasificación: El nivel de medición determina qué análisis estadísticos podemos realizar. Con variables nominales podemos usar frecuencias y moda; con ordinales, medianas; con intervalo y razón, medias y desviaciones estándar (véase la tabla siguiente).

NOMINAL	ORDINAL	INTERVALO	RAZÓN
Sexo Estado Civil Religión	Nivel socioeconómico Grado de acuerdo Nivel educativo	Temperatura (°C) Coeficiente intelectual Fecha calendario	Edad Peso Ingresos
Etiqueta Sin orden	Etiqueta + Orden	Etiqueta + Orden + distancia constante	Etiqueta + orden + distancia + cero absoluto
Estadísticos: Frecuencias	Estadísticos: Mediana	Estadísticos: Media	Estadísticos:

Moda Chi-cuadrado	Percentiles Rangos	Desviación estándar Correlación de Pearson	Media Desviación estándar Coeficiente de variación

3.2 El Instrumento de Medición: Tipos y Características



Una vez que hemos operacionalizado nuestro constructo en dimensiones e indicadores, necesitamos un vehículo para registrar las observaciones: el instrumento de medición.

Un instrumento de medición es cualquier recurso que utiliza el investigador para registrar información sobre las variables que tiene en mente (Hernández Sampieri & Mendoza Torres, 2018). En ciencias sociales, los instrumentos más comunes son:

3.2.1 Cuestionarios y Escalas

El cuestionario: Es un conjunto de preguntas sobre una o más variables a medir. Puede incluir preguntas abiertas (el participante responde con sus propias palabras) o cerradas (selecciona entre opciones predefinidas).

Ejemplo de preguntas en un cuestionario:

- Pregunta abierta: "¿Qué opina sobre la calidad de la educación virtual en su universidad?"
- Pregunta cerrada: "¿Cómo califica la calidad de la educación virtual en su universidad?"

  a) Muy buena

  b) Buena

  c) Regular

  d) Mala

  e) Muy mala

Las escalas: Son un tipo especial de cuestionario diseñado para medir constructos psicológicos o sociales. La característica distintiva de una escala es que busca ubicar al individuo en un continuo, en una dimensión (o varias) previamente definida (DeVellis & Thorpe, 2022).

3.2.2 La Escala de Likert: La Reina de las Mediciones

La escala de Likert desarrollada por (Rensis Likert, 2017), es el formato de respuesta más utilizado en ciencias sociales. Consiste en

presentar una afirmación y pedir al participante que exprese su grado de acuerdo o desacuerdo en una escala que típicamente tiene 5 o 7 puntos.

Formato típico de escala Likert de 5 puntos:

Afirmación	**Muy en desacuerdo (1)**	**En desacuerdo (2)**	**Indiferente (3)**	**De acuerdo (4)**	**Muy de acuerdo (5)**
Me siento orgulloso de trabajar en esta institución					
Mi trabajo me permite desarrollar mis habilidades					
El ambiente laboral es agradable					

Ventajas de la escala Likert:

- Fácil de construir y administrar
- Familiar para la mayoría de los encuestados
- Permite capturar matices en las opiniones
- Produce datos que pueden ser tratados como intervalo (en la práctica)

Precaución: Aunque técnicamente los datos de Likert son ordinales, en la práctica investigadora se tratan frecuentemente como si fueran de intervalo, lo que permite calcular medias y desviaciones estándar. Esta práctica es aceptada siempre que se haga con cautela y se verifiquen ciertos supuestos (Norman, 2010).

3.2.3 Otros Tipos de Escalas

Escala de Thurstone: Busca crear una escala con intervalos aparentemente iguales. Un grupo de jueces evalúa la intensidad de cada afirmación, y solo se seleccionan aquellas en las que hay consenso sobre su posición en el continuo (Thurstone, 1928). Es más compleja de construir pero muy rigurosa.

Escala de Guttman (o escalograma): Asume que los ítems pueden ordenarse jerárquicamente. Si una persona está de acuerdo con un ítem, también estará de acuerdo con todos los ítems de menor intensidad. Por ejemplo, en una escala de participación política, estar de acuerdo con "participaría en una marcha" implica estar de acuerdo con "firmaría una petición" (Guttman, 1944).

Diferencial semántico (Osgood): Se presenta un concepto y se pide al participante que lo califique en una serie de escalas bipolares de 7 puntos (Osgood et al., 1978).

3.3 El Proceso de Construcción de un Instrumento

Construir un buen instrumento de medición no es algo que se improvise. Requiere un proceso sistemático que garantice que, al final, tendremos un instrumento que cumpla con los criterios de validez y confiabilidad que exploraremos en el Capítulo IV.

A continuación, presentamos las etapas del proceso de construcción, basadas en las recomendaciones de (DeVellis & Thorpe, 2022) y (Streiner et al., 2024):

Etapa 1: Definir claramente lo que se quiere medir

Antes de escribir la primera pregunta, debemos tener una definición clara y precisa del constructo. Esto implica:

- Revisar la literatura para entender cómo otros investigadores han definido el constructo.

- Delimitar qué aspectos del constructo incluiremos y cuáles excluiremos.
- Identificar las dimensiones del constructo (si es multidimensional).

Etapa 2: Generar un banco de ítems

Redactar entre 3 y 4 veces más ítems de los que se planea incluir en la versión final. Las reglas para una buena redacción de ítems incluyen:

Hacer	Evitar
Usar lenguaje simple y claro	Preguntas dobles ("¿Está satisfecho con su salario y con sus compañeros?")
Redactar frases cortas (menos de 20 palabras)	Términos ambiguos ("frecuentemente", "a menudo")
Incluir ítems positivos y negativos	Jerga técnica o modismos locales
Hay que asegurar que cada ítem exprese una sola idea	Afirmaciones con las que casi todos estarían de acuerdo
Hacer que los ítems sean relevantes para el constructo	Afirmaciones extremas que nadie respaldaría

Etapa 3: Determinar el formato de respuesta

Decidir si usaremos una escala Likert (5 o 7 puntos), una escala visual análoga, o algún otro formato. Para escalas Likert, es recomendable incluir una opción neutral (como "indiferente" o "ni de acuerdo ni en desacuerdo") para no forzar a los participantes a tomar una posición.

Etapa 4: Revisión por expertos (juicio de expertos)

Someter el instrumento preliminar a la evaluación de personas con experiencia en el tema (expertos en contenido) y en metodología (expertos en medición). Los expertos evaluarán:

- Claridad: ¿La redacción es comprensible?
- Coherencia: ¿El ítem es relevante para la dimensión que pretende medir?
- Suficiencia: ¿Los ítems cubren adecuadamente todas las dimensiones?
- Sesgo: ¿Algún ítem induce una respuesta particular?

Etapa 5: Prueba piloto

Aplicar el instrumento a una pequeña muestra (30-50 personas) con características similares a la población objetivo. El piloto permite:

- Identificar problemas de comprensión.
- Registrar el tiempo necesario para completar el instrumento.
- Obtener datos preliminares para evaluar la confiabilidad.

Etapa 6: Evaluación de propiedades psicométricas

Con los datos del piloto (y posteriormente con los datos definitivos), evaluamos:

- Validez: ¿El instrumento mide lo que pretende medir?
- Confiabilidad: ¿Las mediciones son consistentes?

Estos conceptos son tan importantes que les dedicaremos el próximo capítulo completo.

Etapa 7: Versión definitiva

Incorporar los ajustes derivados de las etapas anteriores y producir la versión final del instrumento, lista para ser aplicada a la muestra completa.

3.4 Errores Comunes en la Construcción de Instrumentos

A lo largo de nuestra experiencia docente e investigadora, hemos identificado una serie de errores frecuentes que cometen quienes se inician en la construcción de instrumentos. Conocerlos es el primer paso para evitarlos.

Error 1: No definir claramente el constructo

El error más común es comenzar a redactar preguntas sin tener una definición clara de qué se quiere medir. El resultado suele ser un instrumento que mezcla aspectos de diferentes constructos y termina midiendo "un poco de todo y nada en particular".

Error 2: Preguntas dobles (double-barreled questions)

Ocurre cuando una pregunta aborda dos temas diferentes, y el participante no sabe a cuál de ellos responder.

- Incorrecto: "¿Está satisfecho con el salario y las condiciones de trabajo?" (¿Qué pasa si está satisfecho con el salario pero no con las condiciones?)
- Correcto: Dividir en dos preguntas separadas.

Error 3: Preguntas tendenciosas (leading questions)

Son preguntas que, por su redacción, inducen una respuesta particular.

- Incorrecto: "¿No cree usted que la educación virtual es inferior a la presencial?"
- Correcto: "En su opinión, ¿cómo compararía la educación virtual con la presencial?"

Error 4: Uso de dobles negativos

Confunden al participante y dificultan la interpretación.

- Incorrecto: "No estoy en desacuerdo con que no se deba prohibir el consumo de alcohol" (¿?)
- Correcto: "Estoy de acuerdo con permitir el consumo de alcohol."

Error 5: Falta de equilibrio en ítems positivos y negativos

Si todos los ítems están redactados en la misma dirección (por ejemplo, todos positivos), algunos participantes pueden caer en el sesgo de "aquiescencia",

es decir, tender a estar de acuerdo con todo sin importar el contenido. Incluir ítems redactados en dirección contraria ayuda a detectar y controlar este sesgo.

Ejemplo de ítems equilibrados en una escala de autoestima:

- Positivo: "En general, me siento satisfecho conmigo mismo."
- Negativo: "A veces pienso que no soy bueno para nada." (invertir la puntuación)

Error 6: Longitud excesiva

Un instrumento demasiado largo puede causar fatiga en los participantes, lo que lleva a respuestas descuidadas o al abandono. Como regla general, un cuestionario no debería tomar más de 20-30 minutos en completarse.

Error 7: Formato confuso o poco claro

Instrucciones confusas, escalas mal diseñadas o saltos complejos entre preguntas pueden invalidar todo el instrumento, por buenas que sean las preguntas individuales.

3.5 La Importancia del Contexto Cultural en la Medición

Un aspecto que a menudo se descuida es la adecuación cultural de los instrumentos de medición. No podemos tomar un instrumento desarrollado

en Estados Unidos o Europa, traducirlo literalmente y aplicarlo en el Perú asumiendo que medirá lo mismo.

Consideraciones culturales:

1) **Equivalencia conceptual:** El constructo que queremos medir, ¿tiene el mismo significado en ambas culturas? Por ejemplo, el concepto de "familia" puede variar significativamente entre culturas.
2) **Equivalencia lingüística:** La traducción debe ser precisa y, además, comprensible. No basta con una traducción literal; debe ser una adaptación que preserve el significado original pero use expresiones naturales en el idioma de destino.
3) **Equivalencia métrica:** ¿La escala funciona de la misma manera en ambas culturas? Esto se evalúa mediante técnicas estadísticas como el análisis factorial confirmatorio multigrupo (Muñiz et al., 2013).

Proceso recomendado para la adaptación transcultural de instrumentos:

3.6 Instrumentos Existentes vs. Instrumentos Propios

Una decisión importante que enfrenta todo investigador es si construir su propio instrumento o utilizar uno ya existente. Ambas opciones tienen ventajas y desventajas:

Utilizar un instrumento existente

Ventajas:

- Ahorra tiempo y recursos.
- Ya cuenta con evidencia de validez y confiabilidad.
- Permite comparar resultados con otros estudios.
- Generalmente ha sido refinado a través de múltiples aplicaciones.

Desventajas:

- Puede no ajustarse perfectamente a los objetivos específicos de nuestra investigación.
- Puede estar desactualizado.
- Puede tener problemas de adaptación cultural.

Construir un instrumento propio

Ventajas:

- Se ajusta perfectamente a los objetivos de la investigación.
- Podemos incorporar aspectos específicos del contexto local.
- Generamos conocimiento nuevo sobre la medición del constructo.

Desventajas:

- Requiere mucho tiempo y recursos.
- Necesitamos demostrar su validez y confiabilidad (no es automático).
- Dificulta la comparación con otros estudios.

Recomendación práctica: Siempre que sea posible, utilice instrumentos existentes que cuenten con buenas propiedades psicométricas y que hayan sido adaptados a su contexto cultural. Reserve la construcción de nuevos instrumentos para cuando realmente no exista ninguno adecuado o cuando su objetivo de investigación sea precisamente desarrollar una nueva medida.

3.7 La Ética en la Medición

La medición en ciencias sociales implica una responsabilidad ética hacia los participantes. Algunos principios fundamentales son:

1) **Respeto por la dignidad humana:** Los participantes no deben ser reducidos a "números" o "datos". Son personas con derechos.
2) **Confidencialidad:** Las respuestas individuales no deben ser reveladas. Los datos deben ser reportados de manera agregada.
3) **Consentimiento informado:** Los participantes deben saber qué se va a medir, para qué fines y cómo se protegerá su información.
4) **Devolución de resultados:** Siempre que sea posible, los resultados deben ser compartidos con las comunidades que participaron.
5) **Minimización de daños:** Las preguntas no deben causar malestar innecesario. En temas sensibles (violencia, salud mental, etc.), se debe tener especial cuidado.

Lo que aprendimos en este capítulo (el arte de atrapar lo invisible)

Hemos llegado al final de este tramo del viaje, y la verdad es que hemos cubierto mucho terreno. Vale la pena hacer una pausa, echar la vista atrás y quedarnos con lo que de verdad importa. Porque convertir lo abstracto en medible no es un juego de niños, y todo lo que hemos visto aquí nos va a acompañar de ahora en adelante.

1) La operacionalización, o cómo darle cuerpo a un fantasma

Lo primero y más importante: ese proceso de bajar las ideas del cielo a la tierra tiene nombre. Se llama operacionalización, y consiste en coger un constructo la inteligencia, la satisfacción, lo que sea, partirlo en dimensiones más manejables y, de ahí, sacar indicadores concretos que podamos preguntar. Es como desmontar un reloj para entender cómo funciona.

2) Los niveles de medición no son un capricho

Lo de Stevens no es una ocurrencia de alguien con demasiado tiempo libre. Que los datos sean nominales, ordinales, de intervalo o de razón determina qué puedes hacer con ellos después. Mezclarlos sin criterio es como usar un destornillador para clavar un clavo: puedes intentarlo, pero el resultado será un desastre.

3) Las herramientas del oficio

En ciencias sociales, el martillo y el cincel son los cuestionarios y las escalas. Y entre todas ellas, la escala Likert es la reina: esas preguntas con opciones que van del "totalmente en desacuerdo" al "totalmente de acuerdo" están por todas partes, y con razón. Bien construidas, son una maravilla.

4) El camino para hacer bien las cosas (que tiene sus pasos)

No se llega a un buen instrumento por casualidad. Hay una ruta: primero defines bien el constructo, luego generas preguntas, eliges un formato, se lo enseñas a expertos para que lo critiquen, haces una prueba con gente de a pie, evalúas si funciona y, solo entonces, lanzas la versión definitiva. Saltarse pasos es pedir problemas.

5) Los errores que delatan al novato (y al que va con prisa)

Preguntar dos cosas a la vez ("¿está satisfecho con su sueldo y con su horario?"), llevar al encuestado hacia dónde queremos ("¿no cree usted que...?"), no dar opciones equilibradas, hacer el cuestionario eterno... Todo eso son pecados capitales que invalidan lo que tanto trabajo nos costó construir.

6) Lo que funciona en Madrid no tiene por qué funcionar en Lima

La cultura importa, y mucho. Usar un test hecho en otro país sin adaptarlo es como ponerse un traje ajeno: puede que te entre, pero seguro que no te queda bien. La adaptación transcultural no es un lujo, es una necesidad.

7) Y nunca, nunca, olvidar que detrás hay personas

Esto parece obvio, pero conviene repetirlo: cuando medimos, estamos pidiéndole a alguien que nos cuente algo de su vida. Eso merece respeto,

cuidado y confidencialidad. No es solo cuestión de técnica, es cuestión de humanidad.

Y ahora que ya sabemos cómo construir un instrumento que capture lo que queremos medir, llega el momento de ponernos exigentes. En el próximo capítulo vamos a someterlo a examen: ¿de verdad mide lo que dice medir? ¿Y lo hace siempre de la misma manera, o es como esos termómetros locos que unos días marcan una cosa y otros otra? Ahí nos espera la validez y la confiabilidad. Y créeme, merece la pena.

CAPÍTULO IV: ¿ES CONFIABLE LO QUE MIDE? VALIDEZ Y CONFIABILIDAD DE UN INSTRUMENTO

Hagamos un ejercicio de imaginación. Supón que ya pasaste por todo lo que hemos contado hasta ahora. Te tomaste el trabajo de definir bien a tu población, sin trampas ni atajos. Te sentaste a elegir la muestra con cuidado, aplicando esas técnicas de muestreo que parecían un trabalenguas pero que al final dominaste. Y luego, con la paciencia de un artesano, construiste tu instrumento, paso a paso, operacionalizando conceptos, afinando preguntas, probando y corrigiendo.

Llegó el gran día. Saliste al campo, aplicaste tu cuestionario, sudaste la gota gorda, y al final, ahí los tienes: tus datos. Una montaña de respuestas, números, opiniones, todo lo que habías estado buscando durante meses.

Y entonces, cuando los tienes delante, sentado frente a la pantalla o con los papeles desparramados sobre la mesa, llega el momento de la verdad. La pregunta que llevaba tiempo acechando en un rincón de tu cabeza se

planta frente a ti sin pedir permiso: ¿y ahora? ¿Todo esto sirve de algo? ¿Puedo fiarme de lo que tengo entre manos?

Porque una cosa es haber seguido los pasos al pie de la letra, y otra muy distinta es estar seguro de que lo que tienes refleja lo que querías capturar. ¿Estamos midiendo realmente lo que creemos medir, o nos estamos engañando con bonitas estadísticas que no significan nada? Y si mañana, pasado o dentro de un mes, volviéramos a aplicar el mismo cuestionario a la misma gente, ¿obtendríamos resultados parecidos o cada cosa cambiaría como el humor de un adolescente?

De eso, justamente, va este capítulo. De las preguntas incómodas, de las que nadie quiere hacerse cuando ya tiene los datos en la mano, pero que son las únicas que separan una investigación de verdad de un simple ejercicio de autoengaño.

Esa clase de preguntas, las que te desvelan por la noche, son las que te meten de lleno en el terreno de la psicometría. Una palabra que suena a chino, pero que en el fondo viene a decir algo bastante sencillo: ¿cómo sabemos si lo que medimos está bien medido? Y ahí aparecen dos viejos conocidos de cualquier investigador que se precie: la validez y la confiabilidad.

Ahora bien, ojo al dato, que conviene no simplificar demasiado. Porque como apuntan tipos con solera en esto, como (Mislevy, 2018) y (L. J. Cronbach, 2013), esto no va de tener un cartelito que ponga "válido" o "no válido" como quien pone etiquetas en un supermercado. No es blanco o negro. Es más bien una cuestión de grados, de matices. Un instrumento es válido en la medida en que la evidencia que vamos acumulando los datos, la teoría, los expertos, las pruebas respalda eso que queremos interpretar a partir de sus resultados. No es un sello que se pone de una vez y para siempre, sino una confianza que se construye paso a paso.

Así que en este capítulo nos vamos a ensuciar las manos con todo eso. Vamos a ver de verdad qué significan esos palabros, cómo se las arreglan los investigadores para evaluarlos y, sobre todo, cómo puedes tú, con tus datos, con tu estudio, aplicarlo sin morir en el intento. Vamos a hacer cuentas juntos, paso a paso, sin saltarnos ninguno. Y para que no se quede en teoría, nos vamos a asomar a cómo se hace todo esto en los programas que seguramente tengas a mano: el SPSS de toda la vida y el Excel, ese viejo amigo que a veces infravaloramos.

La idea es que cuando termines, no solo sepas qué son la validez y la confiabilidad, sino que seas capaz de sentarte delante de tus datos y decir: "esto vale, esto no vale tanto, y por esto lo sé". Y eso, en este oficio, es media vida.

4.1 Validez: La Pregunta Fundamental: ¿Estamos Midiendo lo que Creemos?

La validez está referida a la exactitud de la medición. Es decir, si el instrumento mide, de alguna manera demostrable, aquello que trata de medir, libre de distorsiones sistemáticas (Anastasi & Urbina, 1998). Al estimar la validez es necesario saber a ciencia cierta qué rasgo o característica se desea estudiar. La validez establece si el instrumento diseñado está midiendo realmente el atributo que dice medir.

Por ejemplo, un instrumento que pretende medir la inteligencia debe medir la inteligencia y no la memoria. Un termómetro debe medir temperatura, no presión atmosférica. Una escala de satisfacción laboral debe medir satisfacción con el trabajo, no satisfacción con la vida en general.

La validez de un instrumento, por lo general, no constituye un problema en el caso de la medida de objetos físicos, tales como la longitud, peso o capacidad. Por supuesto, la estatura de una persona se mide con una cinta métrica y la masa de un objeto con una balanza. Sin embargo, con los métodos usados para medir variables psicoeducativas, es necesario probar empíricamente que el instrumento es válido en todos los casos (Finch et al., 2022).

Cuando elaboramos una escala para medir la actitud de los docentes hacia la innovación educativa, debemos probar que los puntajes de la escala realmente distinguen entre aquellos docentes que tienen una actitud favorable hacia la innovación y aquellos cuya actitud es desfavorable.

La validez no es una propiedad del instrumento en sí mismo, sino de las interpretaciones que hacemos a partir de sus puntuaciones. Un mismo instrumento puede ser válido para un propósito y no para otro (American Educational Research Association [AERA], 2024), (American Psychological Association [APA], 2020) & (National Council on Measurement in Education, 2018).

Entre los diferentes tipos de validez se encuentran las siguientes: validez de contenido, validez de constructo y validez de criterio.

4.1.1 Validez de Contenido: ¿Cubrimos Todos los Aspectos Importantes?

La validez de contenido se refiere al grado en que un instrumento refleja un dominio específico del contenido de lo que se quiere medir (Haynes et al., 1995). Es decir, evalúa si los ítems o reactivos que componen

el instrumento son una muestra representativa y adecuada del universo de posibles contenidos relacionados con el constructo.

Al construir un test, elegimos determinados ítems de un conjunto de conductas que tiene un interés específico, por suponer que remiten al atributo a ser evaluado. En el instrumento no colocamos todas las conductas posibles, elegimos algunas de ellas; es decir, hacemos una muestra de conductas. Al validar la validez de contenido, lo que hacemos es evaluar si los ítems que hemos usado para construir el test son relevantes para el uso que se le va a dar, o sea, si todos los ítems están dentro del dominio de interés (Sireci, 1998).

La validez de contenido evalúa además la claridad, comprensión y congruencia de los reactivos o ítems que componen el instrumento.

El Juicio de Expertos: La Sabiduría Colectiva

Para hacer esta determinación se recurre a expertos o jueces. Un ejemplo ilustrativo: un cuestionario sobre la actitud de los alumnos ante la investigación no tendrá validez de contenido si explora la opinión de los alumnos sobre las características de los docentes dentro de la cátedra de estadística. Simplemente, no está cubriendo el dominio de interés.

Hay que considerar que la validez de contenido no puede expresarse cuantitativamente de manera directa; es más bien una cuestión de juicio, se estima de manera subjetiva o intersubjetiva empleando, usualmente, el denominado juicio de expertos (Escobar-Pérez & Cuervo-Martínez, 2008).

¿Qué es un experto?

El experto es una persona competente que ha sido invitada para dar solución a un problema que requiere de conocimientos especializados. El peritaje puede ser individual o colectivo; los expertos pueden expresar su opinión en forma oral o llenar un formato especial.

Selección de los expertos:

- Es una etapa muy importante de este peritaje.
- Al experto altamente calificado le es inherente la maestría o el doctorado en su especialidad, investigadores con trayectoria.
- El experto debe ser imparcial, con una gran amplitud de enfoques e independencia de juicios.

Procedimiento para realizar el juicio de expertos:

1) Seleccionar los expertos o jueces en la materia para evaluar los ítems del instrumento en términos de relevancia, congruencia, comprensión, adecuación y representatividad respecto al universo del contenido.
2) Cada experto recibe información escrita suficiente sobre el propósito de la investigación:
   - Propósito del estudio
   - Objetivo general
   - Objetivos específicos
   - Matriz de consistencia
   - Ficha técnica del instrumento
3) Cada juez recibe un formulario de validación para registrar su opinión con las observaciones del caso, si las hubiera.

4) Se obtiene una carta de validación del instrumento debidamente firmada por cada uno de los expertos.

Ejemplo de Ficha de Validación por Juicio de Expertos

VALIDACIÓN DE INSTRUMENTOS

FICHA DE EVALUACIÓN DEL INSTRUMENTO POR EXPERTOS

I. DATOS INFORMATIVOS

APELLIDOS Y NOMBRES DEL EXPERTO	CARGO O INSTITUCION DONDE LABORA EL EXPERTO	NOMBRE DEL INSTRUMENTO DE EVALUACION	AUTOR DEL INSTRUMENTO
		cuestionario	
TÍTULO:			

TÍTULO DEL ESTUDIO: Satisfacción laboral en enfermeros de hospitales públicos de Lima Metropolitana, 2024

II. ASPECTOS DE VALIDACIÓN (marque con una X)

INDICADORES	CRITERIOS	DEFICIENCIA	REGULAR	BUENA	MUY BUENA	EXCELENTE
CLARIDAD	Esta formulando con lenguaje apropiado					
OBJETIVIDAD	Esta expresado en conductas expresables					
ACTUALIDAD	Adecuado al avance de la ciencia					
ORGANIZACIÓN	Existe una organización lógica					
SUFICIENCIA	Comprende los aspectos en cantidad y calidad					
INTENCIONALIDAD	Adecuado para valorar aspectos de las estrategias					
CONSISTENCIA	Basado en aspectos teóricos científicos					
COHERENCIA	Entre los índices indicadores y las dimensiones					
METODOLOGÍA	La estrategia responde al propósito del diagnóstico					
OPORTUNIDAD	El instrumento a sido aplicado en el momento oportuno o más adecuado					

III. OPINIÓN DE APLICACIÓN: El instrumento es aplicable, aunque se sugiere revisar la redacción de los ítems 5 y 12 para mayor claridad.

IV. PROMEDIO DE VALORACIÓN: 42 puntos (sobre 50)

IV. PROMEDIO DE VALIACIÓN _____ PUNTOS			
LUGAR Y FECHA	DNI	SELLO Y FIRMA DE EXPERTO	TELEFONO FIJO, CELULAR
Lugar			

Cálculo de la Validez de Contenido: El Coeficiente V de Aiken

Una vez que los expertos han emitido sus juicios, necesitamos un método cuantitativo para resumir esa información y tomar decisiones sobre qué ítems conservar y cuáles eliminar o modificar. El coeficiente V de Aiken es una de

las herramientas más utilizadas para este propósito (Aiken, 1980), (Aiken, 1985).

La V de Aiken es un coeficiente que se computa como la razón de un dato obtenido sobre la suma máxima de la diferencia de los valores posibles. Puede ser calculado sobre las valoraciones de un conjunto de jueces con relación a un ítem, o como las valoraciones de un juez respecto a un grupo de ítems. Así mismo, las valoraciones asignadas pueden ser dicotómicas (valores de 0 ó 1) o politómicas (recibir valores de 0 a 5). Se recomienda contar con un mínimo de 3 y un máximo de 7 jueces (Merino Soto & Livia Segovia, 2009).

Fórmula de la V de Aiken:

$$V=\frac{S}{n(c-1)}$$

V: Coeficiente V de Aiken

S: Sumatorias de respuestas

n: número de jueces

c: número de valores en la escala de valoración

Interpretación: Este coeficiente puede obtener valores entre 0 y 1. A medida que se aproxima a 1, el ítem tiene una mayor validez de contenido. Generalmente, se considera que valores superiores a 0.70 indican una validez de contenido aceptable, aunque algunos autores sugieren umbrales más exigentes (0.80) para decisiones críticas (Vega et al., 2025).

Ejemplo Práctico de Cálculo de V de Aiken

Supongamos que tenemos una matriz que contiene información sobre la opinión de 7 expertos respecto a 7 ítems. La escala de valoración utilizada fue:

- A: Ítem aceptable = 2
- M: Ítem se modifica = 1
- R: Ítem que se rechaza = 0

Matriz de valoraciones originales:

ítems	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Juez 6	Juez 7
1	A	A	A	A	M	R	A
2	A	A	A	A	A	A	A
3	R	M	A	A	M	R	R
4	A	A	A	M	A	A	A
5	M	M	A	A	A	A	A
6	A	A	A	A	A	R	A
7	R	A	A	A	A	A	A

En la siguiente tabla se da las valoraciones a cada uno de los ítems, para calcular el del coeficiente V de Aiken. Conversión a valores numéricos (A=2, M=1, R=0):

ítems	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Juez 6	Juez 7	Suma (S)	V de Aiken
1	2	2	2	2	1	0	1	10	0,71
2	2	2	2	2	2	2	2	14	1
3	0	1	2	2	1	0	0	6	0,43
4	2	2	2	1	2	2	2	13	0,93
5	1	1	2	2	2	2	2	12	0,86
6	2	2	2	2	2	0	2	12	0,86
7	0	2	2	2	2	2	2	12	0,86

Cálculo detallado para el primer ítem:

S=10 n = 7 c = 3

$$V=\frac{S}{n(c-1)}=\frac{10}{7(3-1)}=\frac{10}{14}=0,71$$

Cálculo para el ítem 2:

$$V=\frac{S}{n(c-1)}=\frac{14}{7(3-1)}=\frac{14}{14}=1$$

Cálculo para el ítem 3:

$$V=\frac{S}{n(c-1)}=\frac{6}{7(3-1)}=\frac{6}{14}=0,43$$

Cálculo para el ítem 4:

$$V=\frac{S}{n(c-1)}=\frac{13}{7(3-1)}=\frac{13}{14}=0,93$$

Calculo para el ítem 5:

$$V=\frac{S}{n(c-1)}=\frac{12}{7(3-1)}=\frac{12}{14}=0,86$$

Cálculo para el ítem 6:

$$V=\frac{S}{n(c-1)}=\frac{12}{7(3-1)}=\frac{12}{14}=0,86$$

Cálculo para el ítem 7:

$$V=\frac{S}{n(c-1)}=\frac{12}{7(3-1)}=\frac{12}{14}=0,86$$

A medida que sea más elevado el valor computado, el ítem tendrá una mayor validez de contenido.

Interpretación de los resultados:

- El ítem 2 (V = 1.00) tiene una validez de contenido perfecta; todos los jueces coinciden en que es aceptable.
- Los ítems 4, 5, 6 y 7 (V > 0.85) tienen una validez de contenido muy alta.
- El ítem 1 (V = 0.71) tiene una validez aceptable, pero podría requerir alguna mejora.
- El ítem 3 (V = 0.43) tiene una validez deficiente; debería ser eliminado o completamente reformulado.

4.1.2 Validez de Constructo: ¿Respalda la Teoría Nuestros Resultados?

La validez de constructo es quizás la forma más importante y, a la vez, más compleja de validez. Define si una prueba o experimento está a la altura de sus pretensiones teóricas. Se refiere a si la definición operacional de una variable refleja realmente el significado teórico verdadero de un concepto (L. Cronbach & Meehl, 1998).

La forma más sencilla de pensar en ella es como una prueba de generalización, similar a la validez externa, pero evalúa si el experimento se ocupa de la variable que se está probando. La validez de constructo es un dispositivo utilizado casi exclusivamente en las ciencias sociales, la psicología y la educación, donde trabajamos con conceptos abstractos que no son directamente observables.

El Análisis Factorial: Buscando la Estructura Oculta de los Datos

El análisis factorial es la técnica estadística por excelencia para evaluar la validez de constructo. Permite ordenar los datos y facilitar la interpretación de las correlaciones entre los ítems. Se espera encontrar un factor explicativo del constructo con saturaciones altas de los ítems que miden aspectos parecidos, y con saturaciones bajas de aquellos que miden aspectos diferentes (Hair et al., 2019).

Con frecuencia se habla de la estructura factorial de un test como validez estructural o validez factorial. La idea fundamental es que si nuestro instrumento realmente mide el constructo teórico que hemos definido, entonces los ítems deberían agruparse de acuerdo con las dimensiones que postulamos en nuestra teoría.

Guía para realizar la validación de constructo mediante análisis factorial

Según (Valderrama, 2020), la validación de constructo se refiere al grado en que una medición se relaciona consistentemente con otras mediciones, de

acuerdo con las hipótesis derivadas teóricamente y que conciernen a los conceptos (constructos) que están siendo medidos.

Condiciones para poder aplicar el análisis factorial:

(Valderrey Sanz, 2010) señala que, para que exista una adecuación de los datos a un modelo de análisis factorial, la medida de KMO (Káiser-Meyer-Olkin) debe ser próxima a la unidad. Los valores de la prueba KMO por debajo de 0.5 no son aceptables (p-valor < 0.05), y se considera que los datos son inadecuados para realizar el análisis factorial. Si los valores son superiores a 0.5, se considera aceptable la adecuación de los datos a un modelo de análisis factorial.

Criterios de interpretación del KMO (Káiser, 1970):

VALOR DE KMO	INTERPRETACIÓN
KMO ≥ 0.90	Excelente
0.80 ≤ KMO < 0.90	Muy bueno
0.70 ≤ KMO < 0.80	Aceptable
0.60 ≤ KMO < 0.70	Mediocre
0.50 ≤ KMO < 0.60	Inaceptable
KMO < 0.50	Inadmisible

Según (Valderrama, 2020), se precisan que los valores más altos con signos positivos o negativos y, según Gorsuch (1983), los pesos factoriales de 0.35 a más son suficientes para asumir la relación entre las preguntas y el factor.

Ejemplo Práctico: Análisis Factorial con SPSS

Utilizaremos un ejemplo concreto para ilustrar el procedimiento. Se trata de un cuestionario dirigido a estudiantes sobre "Participación en la gestión curricular en proyectos formativos por competencias con enfoque socioformativo". El instrumento tiene 9 preguntas y fue aplicado a 79 estudiantes. La variable se mide en escala ordinal (1 = Insatisfactorio, 2 = Poco satisfactorio, 3 = Satisfactorio, 4 = Muy satisfactorio).

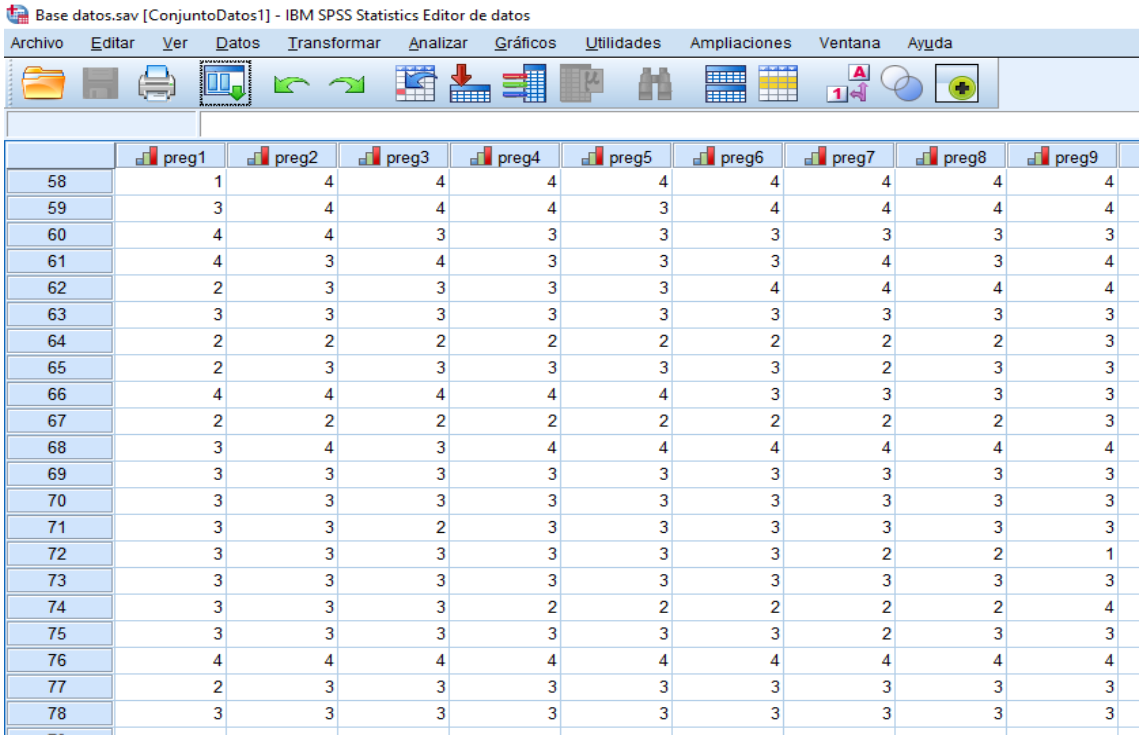
Cuestionario utilizado:

I. PARTICIPACIÓN EN LA PLANIFICACIÓN CURRICULAR

N°	Ítems	Valoración			
		1	2	3	4
1	Participa en la redacción de los objetivos de aprendizaje en la planificación del proyecto formativo.				
2	Participa en la descripción del proyecto formativo.				
3	Participa en la priorización de las competencias genéricas, específicas, de especialidad y desempeños esperados en los proyectos formativos.				
4	Participa en la descripción de los resultados de aprendizaje del proyecto formativo.				
5	Participa en la priorización del problema y el producto en el proyecto formativo.				
6	Participa en la redacción de la transversalidad en la planificación de los proyectos formativos.				
7	Contribuye en el planteamiento de las características de los estudiantes en la planificación de los proyectos formativos.				
8	Participa en la elaboración de la matriz de la planificación de la ejecución de los proyectos formativos				
9	Participa en la selección de las referencias bibliográficas según las normas Apa séptima edición.				

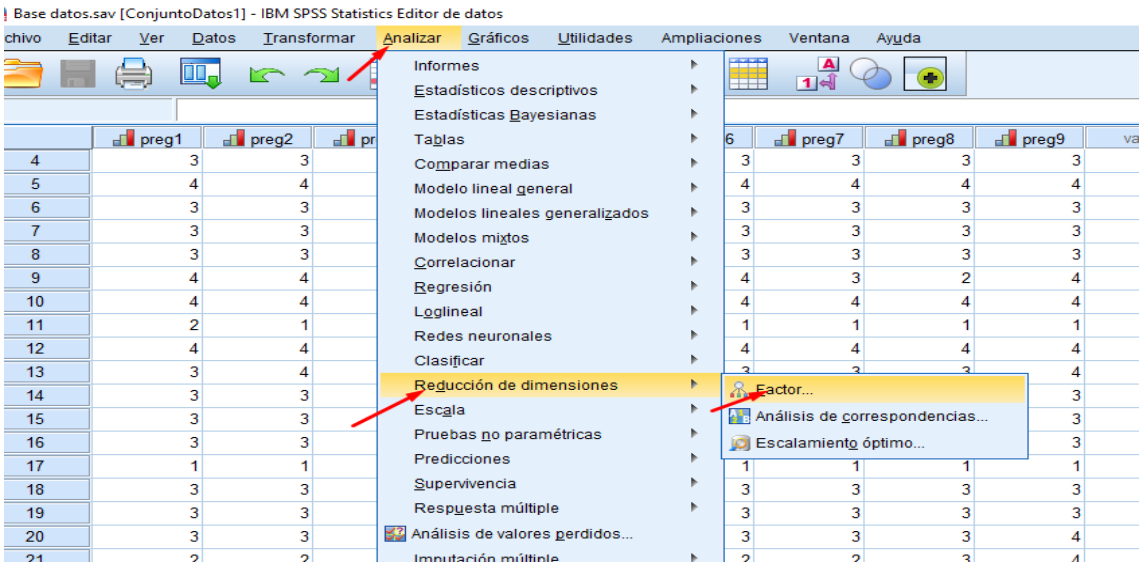
Procedimiento paso a paso en SPSS:

Paso 1: Cargar los datos en SPSS. La base de datos contiene 9 variables (preguntas) y 79 casos (encuestados).

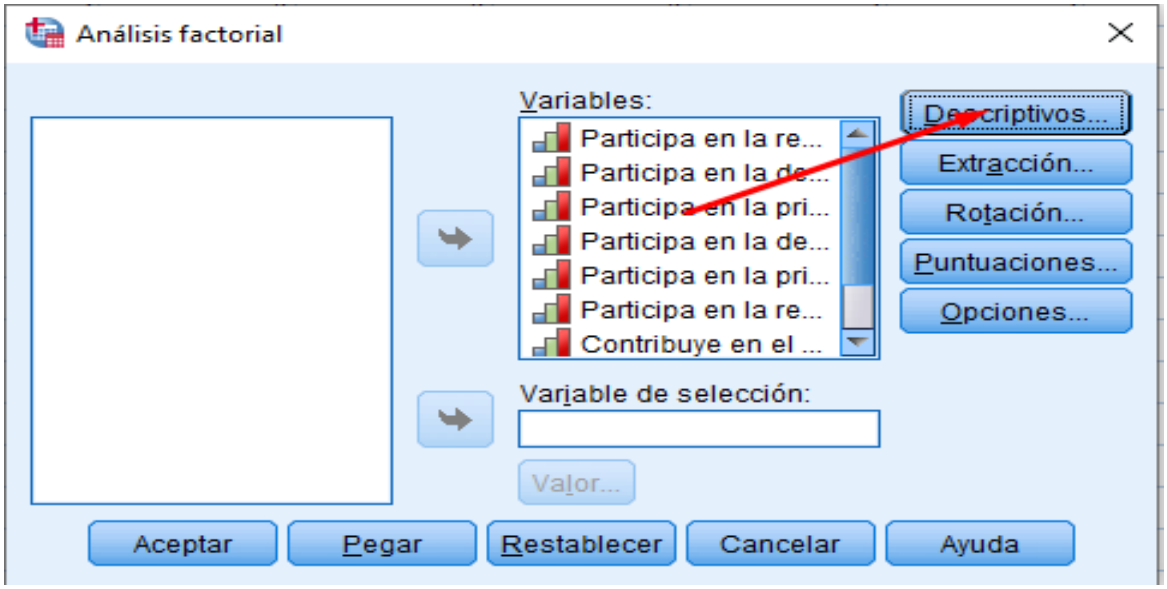


	preg1	preg2	preg3	preg4	preg5	preg6	preg7	preg8	preg9
58	1	4	4	4	4	4	4	4	4
59	3	4	4	4	3	4	4	4	4
60	4	4	3	3	3	3	3	3	3
61	4	3	4	3	3	3	4	3	4
62	2	3	3	3	3	4	4	4	4
63	3	3	3	3	3	3	3	3	3
64	2	2	2	2	2	2	2	2	3
65	2	3	3	3	3	3	2	3	3
66	4	4	4	4	4	3	3	3	3
67	2	2	2	2	2	2	2	2	3
68	3	4	3	4	4	4	4	4	4
69	3	3	3	3	3	3	3	3	3
70	3	3	3	3	3	3	3	3	3
71	3	3	2	3	3	3	3	3	3
72	3	3	3	3	3	3	2	2	1
73	3	3	3	3	3	3	3	3	3
74	3	3	3	2	2	2	2	2	4
75	3	3	3	3	3	3	2	3	3
76	4	4	4	4	4	4	4	4	4
77	2	3	3	3	3	3	3	3	3
78	3	3	3	3	3	3	3	3	3

Paso 2: Ir al menú Analizar → Reducción de dimensiones → Factor...

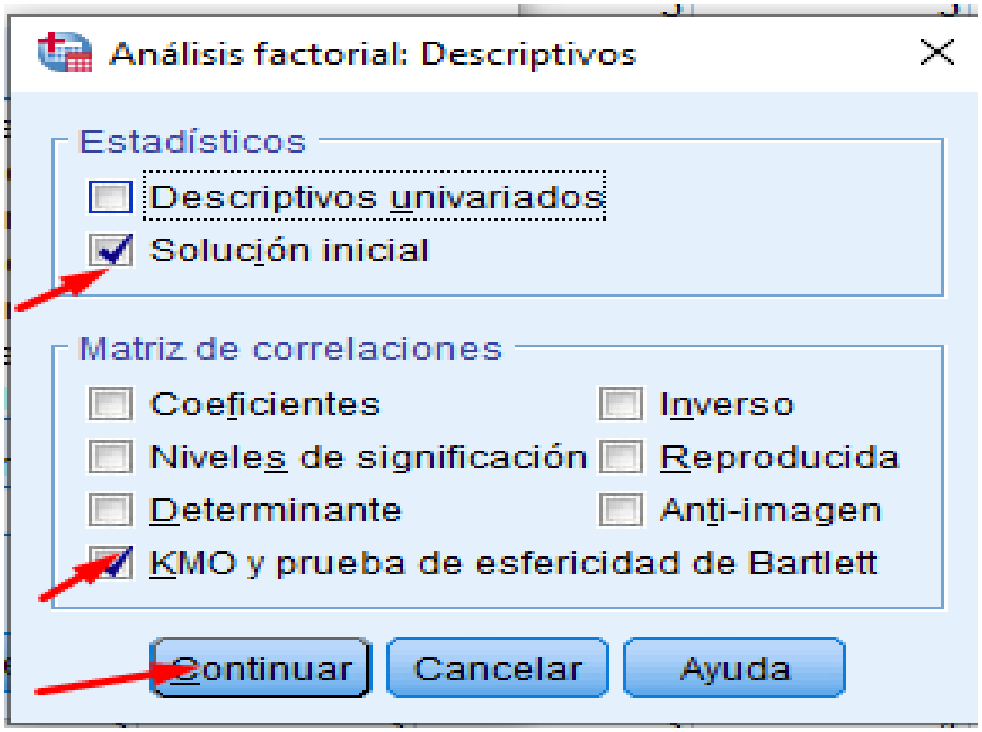


Paso 3: En el cuadro de diálogo, pasar las 9 variables a la lista de "Variables".

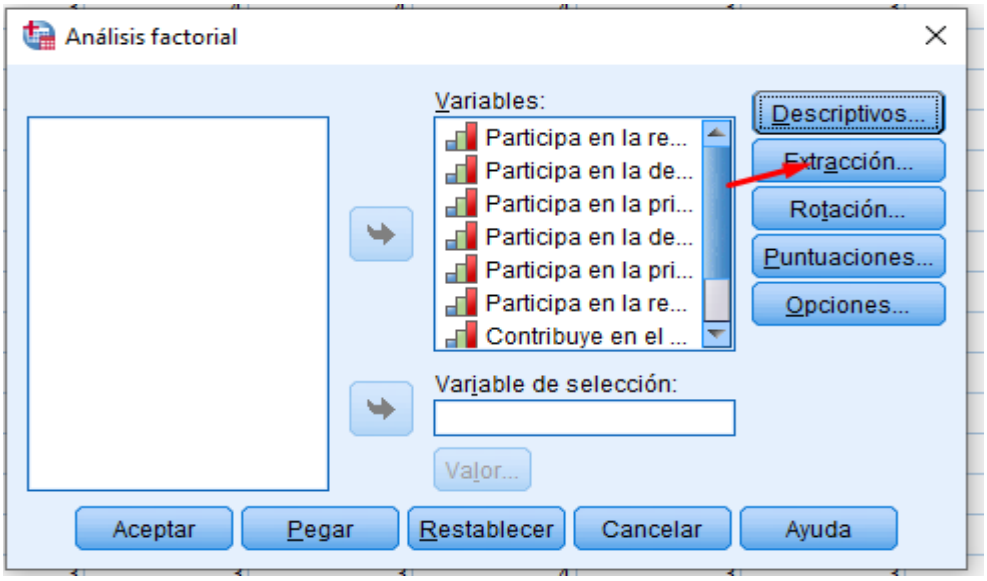


Paso 4: Hacer clic en el botón Descriptivos y marcar:

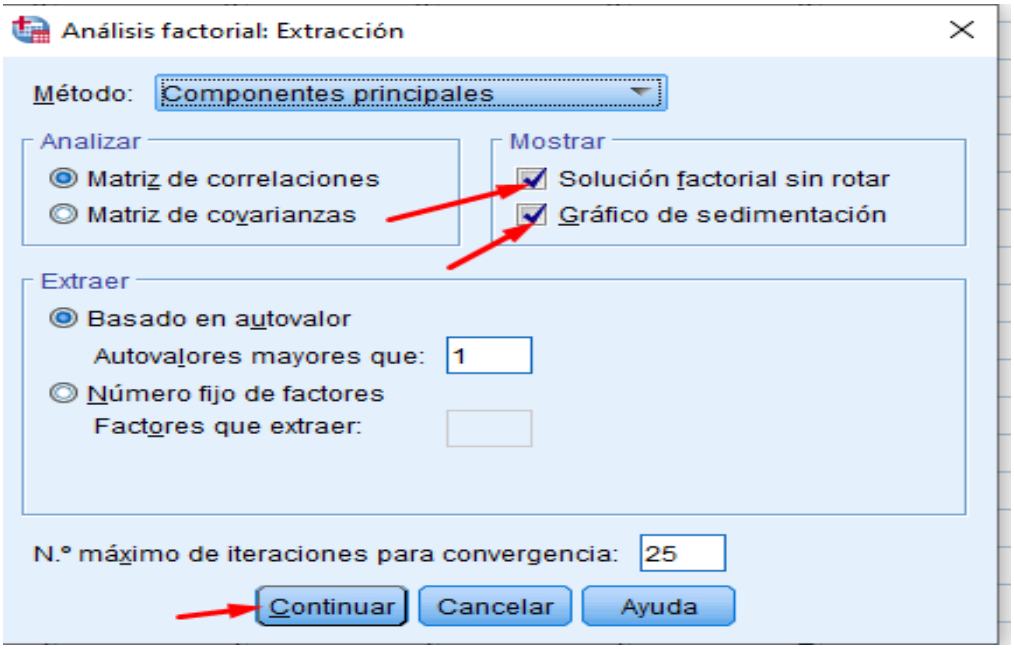
- Estadísticos univariados (opcional)
- Solución inicial
- KMO y prueba de esfericidad de Bartlett
- Hacer clic en Continuar



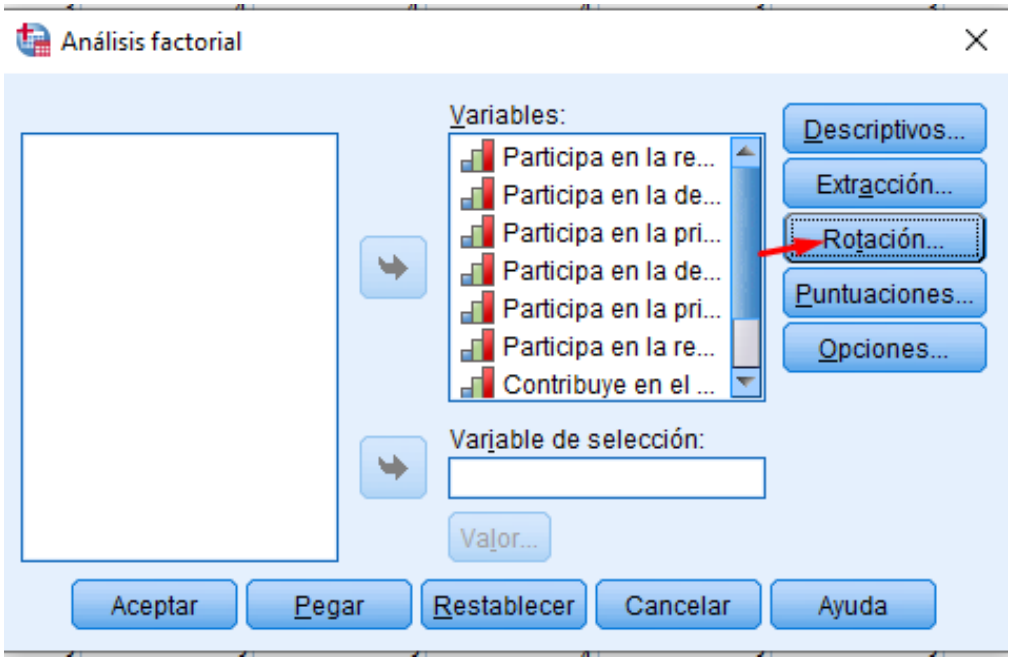
Paso 5: Hacer clic en el botón Extracción:



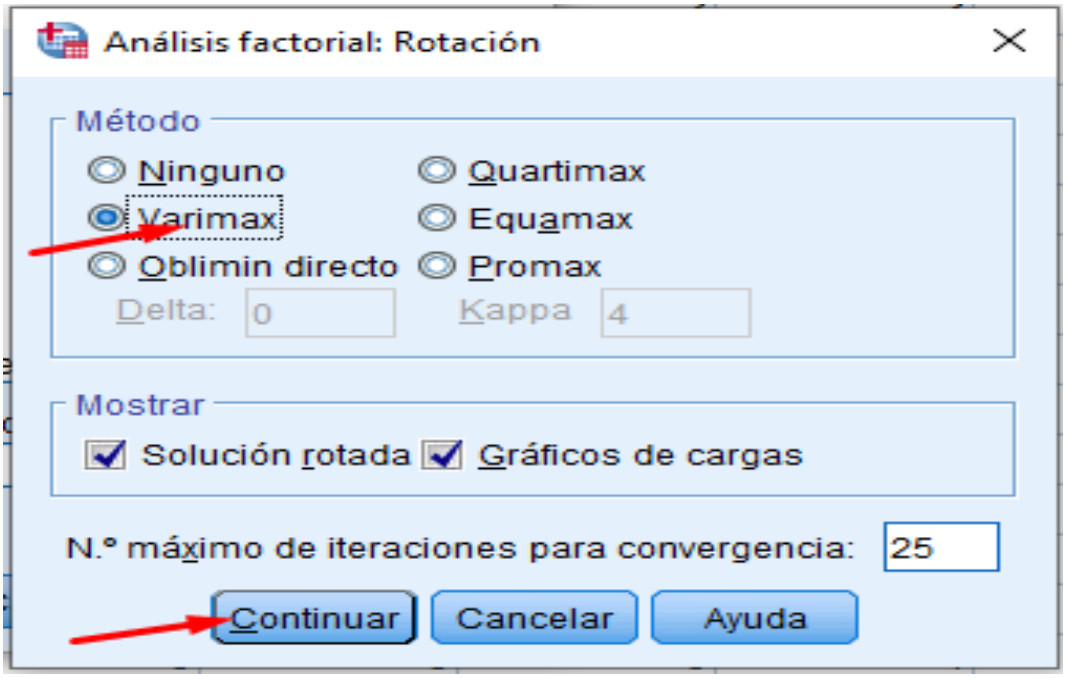
- Método: Componentes principales (por defecto)
- Analizar: Matriz de correlaciones
- Mostrar: Solución factorial sin rotar y Gráfico de sedimentación
- Hacer clic en Continuar



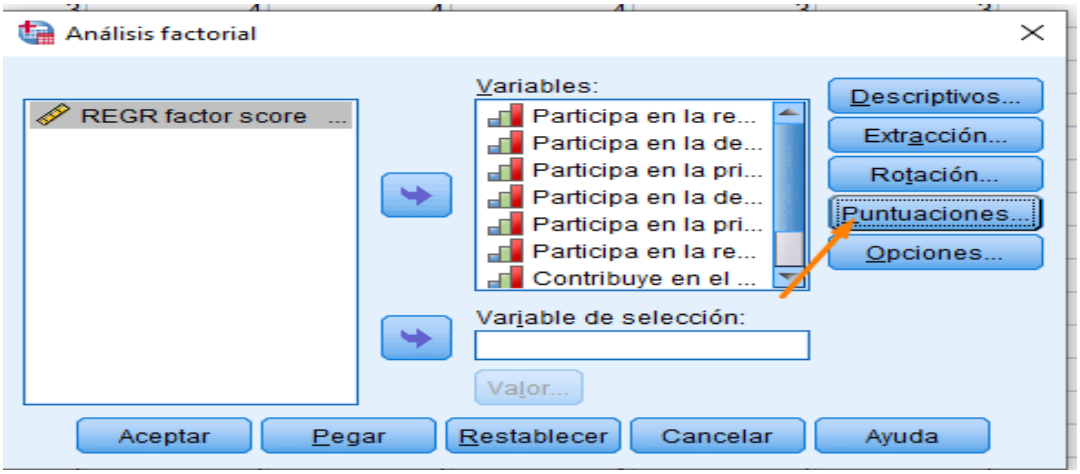
Paso 6: Hacer clic en el botón Rotación:



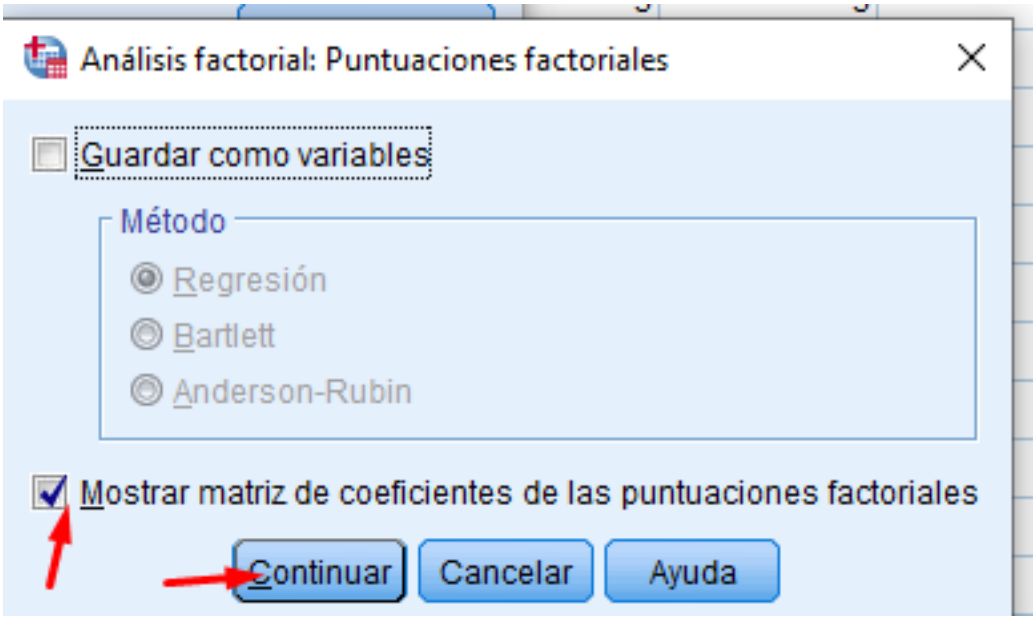
- Método: Varimax (rotación ortogonal que maximiza la varianza de las cargas factoriales)
- Mostrar: Solución rotada y Gráfico de cargas
- Hacer clic en Continuar



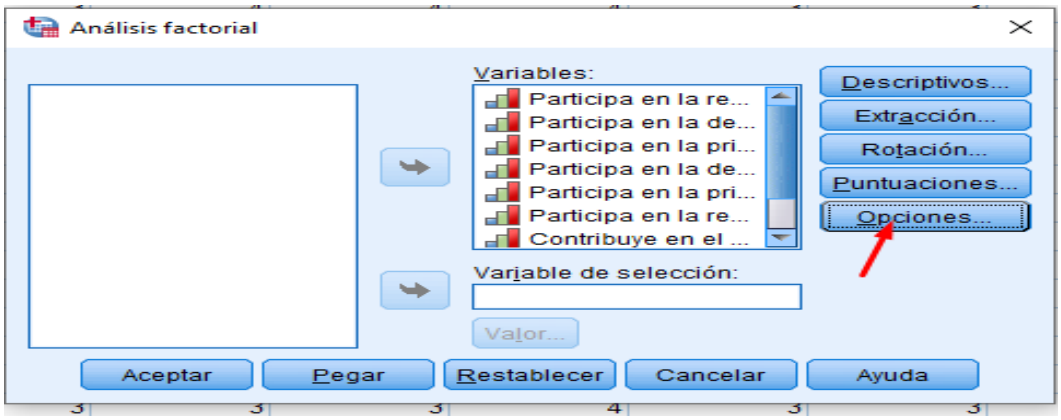
Paso 7: Hacer clic en el botón Puntuaciones:



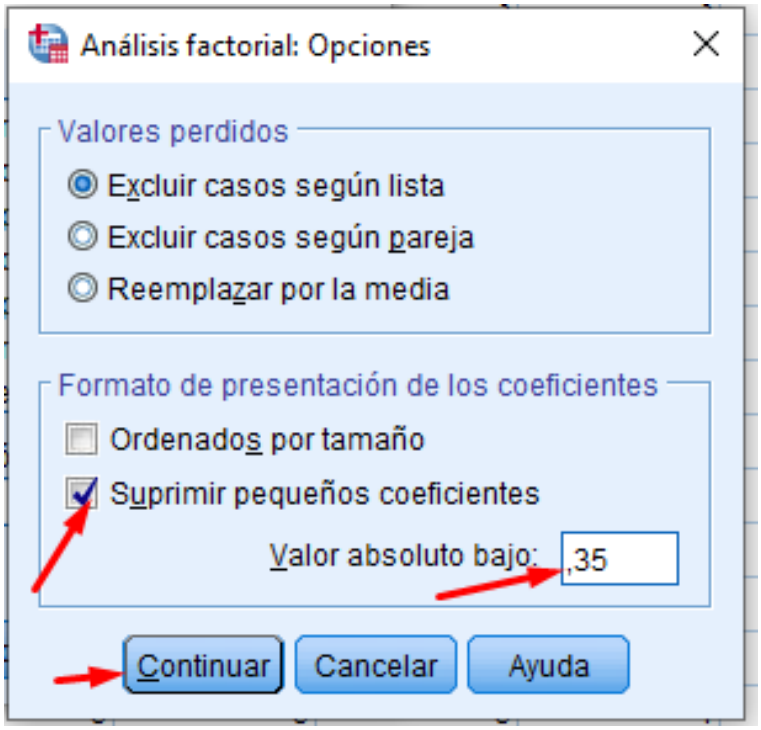
- Marcar Mostrar matriz de coeficientes de las puntuaciones factoriales
- Hacer clic en Continuar



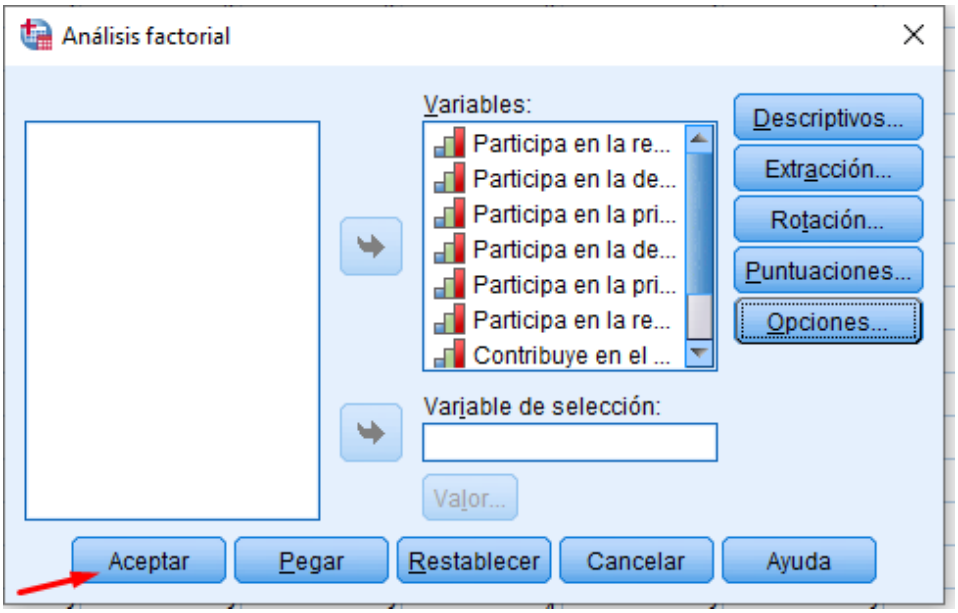
Paso 8: Hacer clic en el botón Opciones:



- Formato de presentación de coeficientes: Marcar Suprimir valores pequeños y establecer el valor absoluto por debajo de 0.35 (siguiendo la recomendación de Gorsuch)
- Hacer clic en Continuar



Paso 9: Hacer clic en Aceptar para ejecutar el análisis.



Interpretación de los Resultados del Análisis Factorial

1. Prueba KMO y esfericidad de Bartlett:

Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		,934
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	745,678
	gl	36
	Sig.	,000

Interpretación: El valor KMO = 0.856 es superior a 0.80, lo que indica que la correlación entre las variables es muy buena y los datos son adecuados para realizar el análisis factorial. La prueba de esfericidad de Bartlett es significativa ($p < 0.001$), lo que rechaza la hipótesis de que la matriz de correlaciones sea una matriz identidad (es decir, las variables están correlacionadas entre sí).

2. Comunalidades:

Las comunalidades indican la proporción de varianza de cada variable que es explicada por los factores retenidos.

	Inicial	Extracción
Participa en la redacción de los objetivos de aprendizaje	1,000	,473
Participa en la descripción del proyecto formativo	1,000	,771
Participa en la priorización de las competencias genéricas	1,000	,828
Participa en la descripción de los resultados	1,000	,746
Participa en la priorización del problema	1,000	,818
Participa en la redacción de la transversalidad	1,000	,880
Contribuye en el planteamiento de las características	1,000	,861
Participa en la elaboración de la matriz	1,000	,823
Participa en la selección de la referencia bibliográfica	1,000	,641

Interpretación: En la columna de extracción, todos los valores son mayores de 0.35, y la mayoría supera 0.70. Esto confirma que existe una buena correlación interna entre las variables y que los factores retenidos explican una proporción sustancial de la varianza de cada ítem. El ítem con comunalidad más baja es P1 (0.473), lo que sugiere que podría no estar tan alineado con el factor principal como los demás.

3. Varianza total explicada:

Componente	Autovalores Iniciales	Sumas de Extracción	Sumas de Rotación						
	Total	% varianza	% acumulado	Total	% varianza	% acumulado	Total	% varianza	% acumulado
1	5.24	58.22	58.22	5.24	58.22	58.22	4.12	45.78	45.78
2	1.38	15.33	73.55	1.38	15.33	73.55	2.62	29.11	74.89
3	0.89	9.89	83.44						

Interpretación: Se han extraído dos factores con autovalores (eigenvalues) mayores que 1 (criterio de Káiser), que explican conjuntamente el 73.55% de la varianza total. Esto es un resultado muy bueno, indicando que la estructura bidimensional captura la mayor parte de la información contenida en los 9 ítems.

4. Matriz de componentes rotados (cargas factoriales):

VARIABLE	COMPONENTE 1	COMPONENTE 2
P1: Participa en la redacción de objetivos	0.652	0.214
P2: Participa en la descripción del proyecto	0.834	0.278
P3: Participa en priorización de competencias	0.891	0.192
P4: Participa en descripción de resultados	0.798	0.325

VARIABLE	COMPONENTE 1	COMPONENTE 2
P5: Participa en priorización del problema	0.887	0.167
P6: Participa en redacción de transversalidad	0.912	0.211
P7: Contribuye en planteamiento de características	0.893	0.243
P8: Participa en elaboración de la matriz	0.856	0.298
P9: Participa en selección de referencias	0.231	0.769

Interpretación: Observamos que los ítems P1 a P8 tienen cargas factoriales altas (superiores a 0.65) en el Componente 1, mientras que el ítem P9 carga principalmente en el Componente 2. Esto sugiere que el constructo "Participación en la gestión curricular" tiene dos dimensiones:

- **Factor 1:** Participación en aspectos centrales de la planificación curricular (8 ítems).
- **Factor 2:** Participación en aspectos bibliográficos y formales (1 ítem, P9).

Esta estructura tiene sentido teórico: la selección de referencias bibliográficas según normas APA es un aspecto más técnico y específico que podría constituir una dimensión separada dentro de la participación curricular.

Conclusión del análisis factorial: Los resultados confirman la validez de constructo del instrumento. La estructura factorial encontrada es coherente

con la teoría, los ítems se agrupan de manera lógica y las cargas factoriales son suficientemente altas (≥ 0.35) para asumir la relación entre cada pregunta y su factor correspondiente.

4.1.3 Validez de Criterio: ¿Coincide con Otras Medidas?

La validez de criterio se estudia comparando los puntajes de un instrumento (variable independiente) con una o más variables externas (variable dependiente) denominadas variable criterio. Se asume que tales criterios, indicadores del desempeño futuro, están teórica y lógicamente relacionados con el rasgo representado en el instrumento bajo estudio (American Educational Research Association [AERA], 2024).

Esta comparación entre los puntajes de la variable en estudio y la variable criterio se expresa a través de un coeficiente de correlación, el cual se interpreta como un índice de validez. Entre más alta sea la correlación, mejor será la validez del instrumento.

Por ejemplo, si diseñamos una prueba de aptitud académica, esperaríamos que sus puntuaciones se correlacionen altamente con el promedio de notas de los estudiantes (tomado como variable criterio). Cuanto mayor sea esa correlación, mejor será la validez predictiva de la prueba de aptitud académica.

La obtención de criterios apropiados para validar una prueba a veces tiene sus dificultades. Por ejemplo, ¿cuál de las siguientes podría ser el mejor criterio para predecir el nivel de eficiencia de un docente?

- ¿El nivel de dominio que los alumnos tienen de los objetivos instruccionales?

- ¿El promedio de calificaciones de los alumnos?
- ¿La tasa de estudiantes aprobados o promovidos?
- ¿La claridad con la que el docente expone sus clases?
- ¿La responsabilidad del docente en el cumplimiento de las tareas inherentes a su cargo?

Todos los criterios mencionados pudieran estar, de alguna manera, relacionados con la eficiencia de un docente. El problema está en cómo determinar el más relevante, o los más relevantes, porque puede darse el caso de que el desempeño futuro que se aspira predecir se exprese no a través de un criterio sino de varios (Spoto et al., 2023).

La Validez de Criterio del Instrumento de Investigación

La validez de criterio puede ser de dos tipos:

a) Validez de criterio tipo concurrente: Hace relación al grado hasta el cual el instrumento mide lo mismo que miden otros instrumentos aceptados como válidos para interpretar el mismo asunto. Se aplica ambos instrumentos al mismo tiempo y se correlacionan sus puntuaciones.

Ejemplo: Si hemos desarrollado una nueva escala para medir depresión, la aplicamos junto con el Inventario de Depresión de Beck (un instrumento ya validado) a un grupo de pacientes. Si la correlación entre ambas medidas es alta, nuestra nueva escala tiene buena validez concurrente.

b) Validez de criterio tipo predictiva: Permite establecer el grado hasta el cual la aplicación del instrumento sirve para predecir eficientemente la

conducta futura de los sujetos que fueron evaluados. Hay un lapso de tiempo entre la medición del instrumento y la obtención del criterio.

Ejemplo: Aplicamos un test de aptitud académica a estudiantes que ingresan a la universidad. Un año después, correlacionamos esas puntuaciones con sus promedios de notas. Una correlación alta indica que el test tiene buena validez predictiva.

¿Cómo Evaluar el Criterio del Instrumento?

Cuando se tiene un instrumento para realizar el cotejo de criterio, se recurre a la correlación de resultados. Se pueden utilizar estadísticos como el coeficiente de contingencia, Spearman-Brown, Pearson, Alfa de Cronbach y la técnica de Aiken, dependiendo de los tipos de variables investigadas.

¿Cómo evaluar el criterio del instrumento mediante juicio de expertos?

Mediante el juicio de expertos, llamado también Kappa de Fleiss, podemos determinar la validez predictiva a través de la valoración de cada ítem del instrumento, en función de la pregunta: ¿Predicen las puntuaciones del instrumento una conducta futura del sujeto?

Escala de valoración (5 puntos):

1) En nada
2) Muy poco
3) De forma algo deficiente
4) Lo suficiente
5) Satisfactoriamente

Ejemplo Práctico: Validez de Criterio con Kappa de Fleiss

Contexto: Se desea validar un cuestionario sobre consumo de productos panificados integrales. El objetivo es: "Determinar en qué proporción las personas entre 20 y 40 años de un determinado estrato social aceptan consumir productos panificados de la línea integral, como parte de una alimentación saludable".

Cuestionario para validar:

ítem	Contenido y escala	1	2	3	4	5
Pregunta 1 (P1)	¿Con que frecuencia come usted productos del tipo “panes integrales”? 1) Nunca 2) Diariamente 3) Varias veces a la semana 4) Quincenalmente 5) Muy ocasionalmente					
Pregunta 2 (P2)	Al momento de comprar productos del tipo panes integrales ¿En qué rango de compra considera estar usted? 1) Compro un producto 2) Compro entre dos y cinco productos 3) Compro más de cinco unidades					
Pregunta 3 (P3)	Sus compras personales de productos del tipo pan integral son por valor de: 1) No consumo 2) Entre 100 y 500 soles por semana 3) Entre 500 y 1000 soles por semana 4) Más de 1000 soles semanales					
Pregunta 4 (P4)	Favor indicar que tan determinantes es para usted el precio del producto, al momento de comprar un producto que considere saludable. 1) No es determinante 2) Muy poco determinante 3) Medianamente determinante 4) Altamente determinante					

Pregunta 5 (P5)	¿Cuál de los siguientes elementos considera usted determinante para comprar un producto panificado? 1) La marca 2) El precio 3) Los ingredientes 4) El empaque					
Pregunta 6 (P6)	¿En qué rango de edades se encuentra usted actualmente? 1) Menos de 20 años 2) Entre 20 y 30 años 3) Entre 30 y 40 años 4) Mayores de 40 años					
Pregunta 7 (P7)	¿En qué estrato social vive actualmente? 1) Estrato 1 2) Estrato 2 3) Estrato 3 4) Estrato 4					

Procedimiento:

Cinco expertos valoran cada pregunta utilizando la escala de 1 a 5, respondiendo a la pregunta: ¿en qué medida este ítem contribuye a predecir la conducta futura de consumo de productos integrales?

Tabulación de las valoraciones de los expertos:

	ELEMENTO	NOMBRE	CALIFICACIÓN
Experto1	Pregunta 1	Prof. Aldo	4
	Pregunta 2	Prof. Aldo	5
	Pregunta 3	Prof. Aldo	4
	Pregunta 4	Prof. Aldo	4
	Pregunta 5	Prof. Aldo	3
	Pregunta 6	Prof. Aldo	4
	Pregunta 7	Prof. Aldo	5
Experto 2	Pregunta 1	Prof. Matias	4
	Pregunta 2	Prof. Matias	5
	Pregunta 3	Prof. Matias	4
	Pregunta 4	Prof. Matias	4
	Pregunta 5	Prof. Matias	4

	Pregunta 6	Prof. Matias	4
	Pregunta 7	Prof. Matias	5
	Pregunta 1	Prof. Hamilton	4
	Pregunta 2	Prof. Hamilton	5
	Pregunta 3	Prof. Hamilton	4
	Pregunta 4	Prof. Hamilton	4
	Pregunta 5	Prof. Hamilton	3
	Pregunta 6	Prof. Hamilton	4
Experto 3	Pregunta 7	Prof. Hamilton	4
	Pregunta 1	Dra. Ana	4
	Pregunta 2	Dra. Ana	5
	Pregunta 3	Dra. Ana	5
	Pregunta 4	Dra. Ana	4
	Pregunta 5	Dra. Ana	3
	Pregunta 6	Dra. Ana	4
Experto 4	Pregunta 7	Dra. Ana	5
	Pregunta 1	Mg. Arturo	4
	Pregunta 2	Mg. Arturo	5
	Pregunta 3	Mg. Arturo	5
	Pregunta 4	Mg. Arturo	4
	Pregunta 5	Mg. Arturo	3
	Pregunta 6	Mg. Arturo	4
Experto 5	Pregunta 7	Mg. Arturo	5

Tabla de frecuencias por ítem y categoría:

Ítem	Categoría 1	Categoría 2	Categoría 3	Categoría 4	Categoría 5	Total
P1	0	0	0	5	0	5
P2	0	0	0	0	5	5
P3	0	0	0	3	2	5
P4	0	0	0	5	0	5
P5	0	0	4	1	0	5
P6	0	0	0	5	0	5

Ítem	Categoría 1	Categoría 2	Categoría 3	Categoría 4	Categoría 5	Total
P7	0	0	0	1	4	5
Total	0	0	4	20	11	35

Interpretación de la tabla:

- La pregunta 1 recibió 5 valoraciones de 4 (lo suficiente).
- La pregunta 2 recibió 5 valoraciones de 5 (satisfactoriamente).
- La pregunta 3 recibió 3 valoraciones de 4 y 2 valoraciones de 5.
- La pregunta 5 recibió 4 valoraciones de 3 (de forma algo deficiente) y 1 de 4.
- En total, hubo 4 valoraciones de 3, 20 de 4 y 11 de 5, sobre un total de 35 valoraciones (7 ítems × 5 expertos).

Cálculo del Coeficiente Kappa de Fleiss

El coeficiente Kappa de Fleiss es una medida de concordancia entre múltiples evaluadores (más de 2) cuando las evaluaciones son en escala nominal u ordinal (Fleiss, 1971). A diferencia del Kappa de Cohen (que solo sirve para dos evaluadores), el Kappa de Fleiss permite evaluar el acuerdo entre varios jueces.

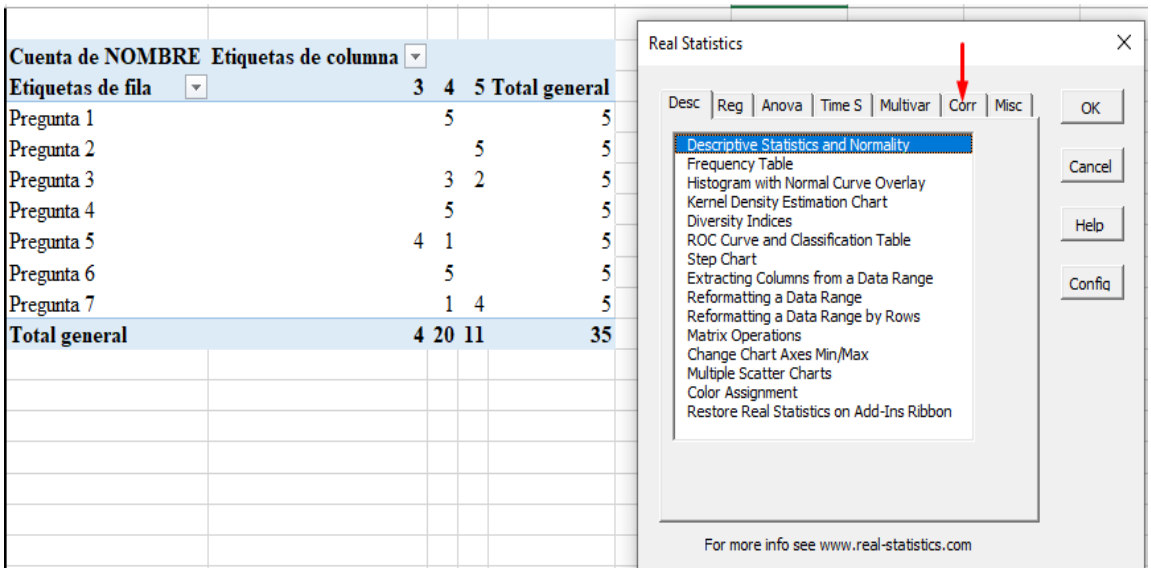
Interpretación del valor de Kappa (adaptado de Landis & Koch, 1977):

VALOR DE KAPPA	FUERZA DEL ACUERDO
< 0.00	Pobre

VALOR DE KAPPA	FUERZA DEL ACUERDO
0.00 - 0.20	Leve
0.21 - 0.40	Regular
0.41 - 0.60	Moderado
0.61 - 0.80	Sustancial
0.81 - 1.00	Casi perfecto

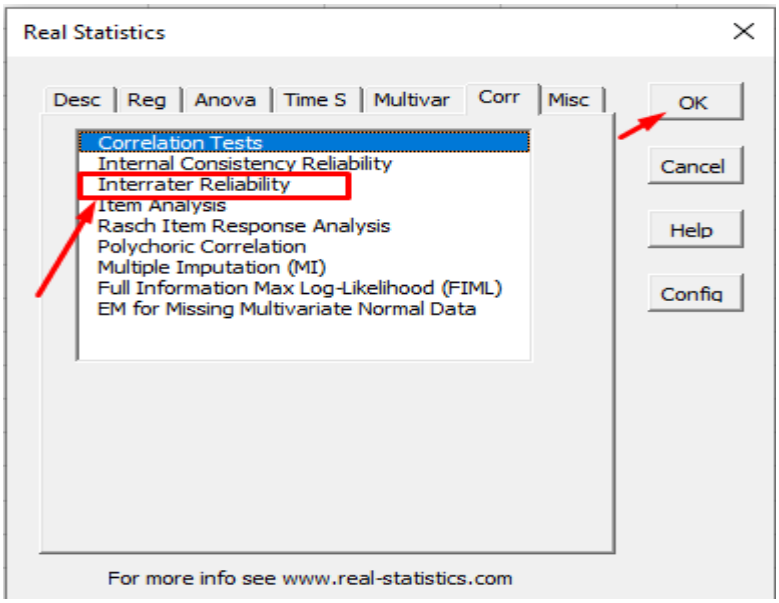
Procedimiento en Excel con complemento Real Statistics:

1) Organizar los datos en una tabla donde las filas sean los ítems y las columnas las categorías de respuesta (frecuencias por categoría).
2) Activar el complemento Real Statistics (Ctrl + M).
3) Seleccionar Corr en el menú.

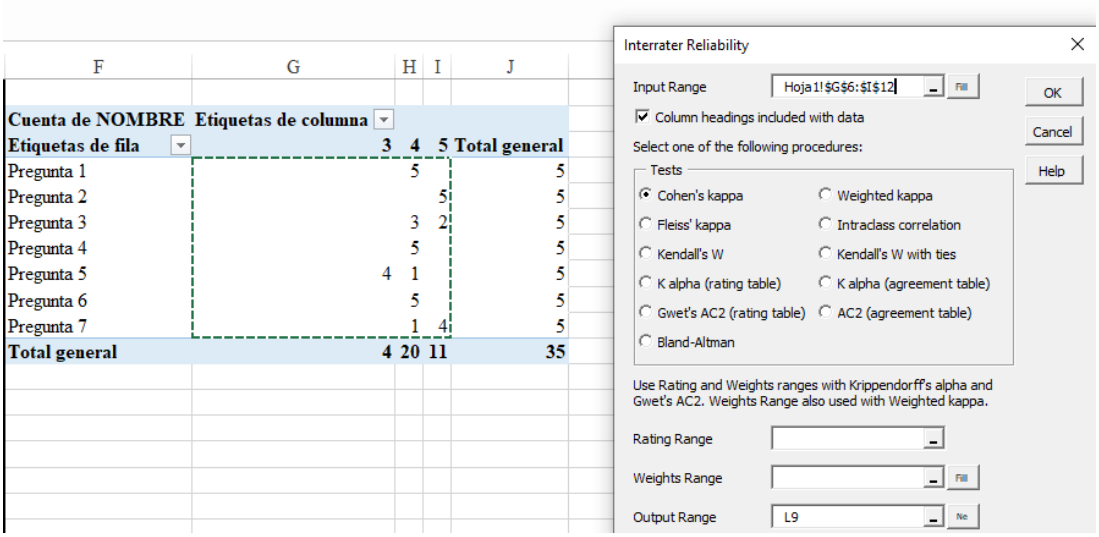


Cuenta de NOMBRE	Etiquetas de columna			
Etiquetas de fila	3	4	5	Total general
Pregunta 1		5		5
Pregunta 2			5	5
Pregunta 3		3	2	5
Pregunta 4		5		5
Pregunta 5	4	1		5
Pregunta 6		5		5
Pregunta 7		1	4	5
Total general	4	20	11	35

Ingresado a Corr. Nos muestra el siguiente cuadro

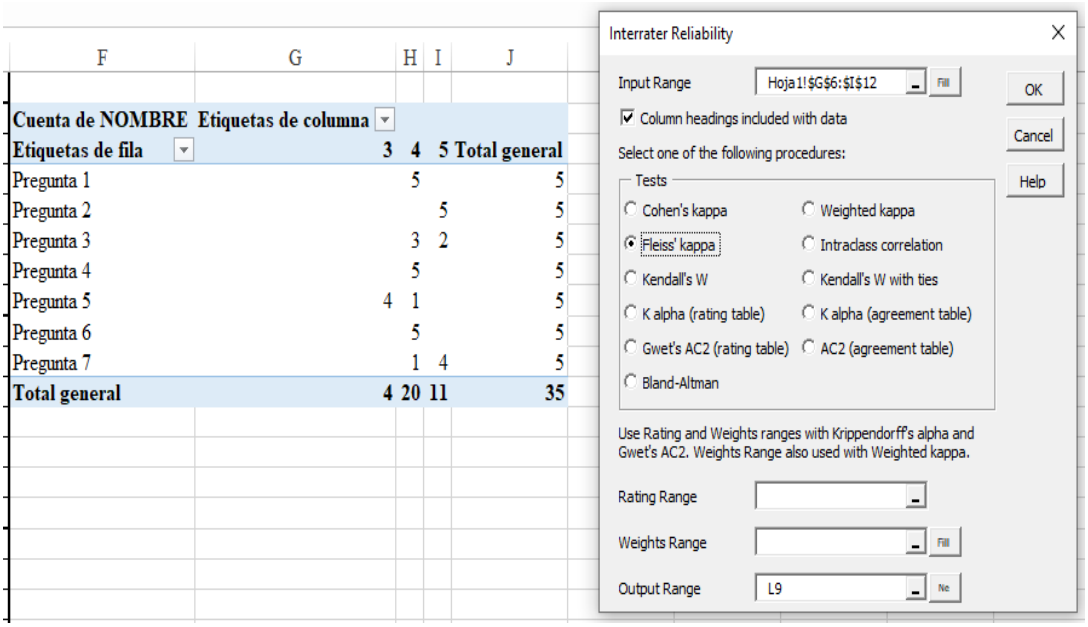


Luego marcar el rango



Cuenta de NOMBRE	Etiquetas de columna			
Etiquetas de fila	3	4	5	Total general
Pregunta 1		5		5
Pregunta 2			5	5
Pregunta 3		3	2	5
Pregunta 4		5		5
Pregunta 5	4	1		5
Pregunta 6		5		5
Pregunta 7		1	4	5
Total general	4	20	11	35

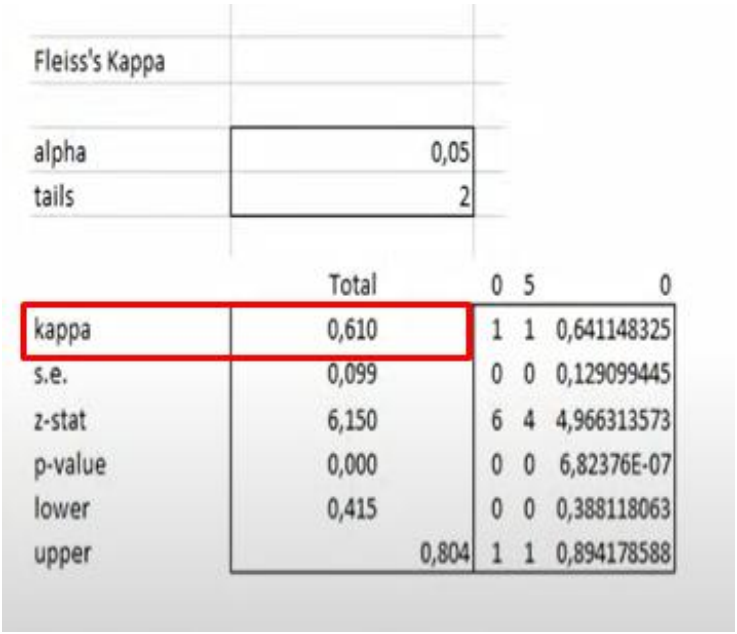
4) Una vez identificado el rango aparece por defecto Cohens Kappa se utiliza como es para dos juicios de expertos, vamos a utilizar Fleiss Kappa por tener más de 2 expertos
5) Elegir Fleiss' Kappa.
6) Seleccionar el rango de datos (incluyendo las frecuencias).
7) Hacer clic en Aceptar.



Cuenta de NOMBRE	Etiquetas de columna			
Etiquetas de fila	3	4	5	Total general
Pregunta 1		5		5
Pregunta 2			5	5
Pregunta 3		3	2	5
Pregunta 4		5		5
Pregunta 5	4	1		5
Pregunta 6		5		5
Pregunta 7		1	4	5
Total general	4	20	11	35

Resultado obtenido:

Kappa de Fleiss = 0.61

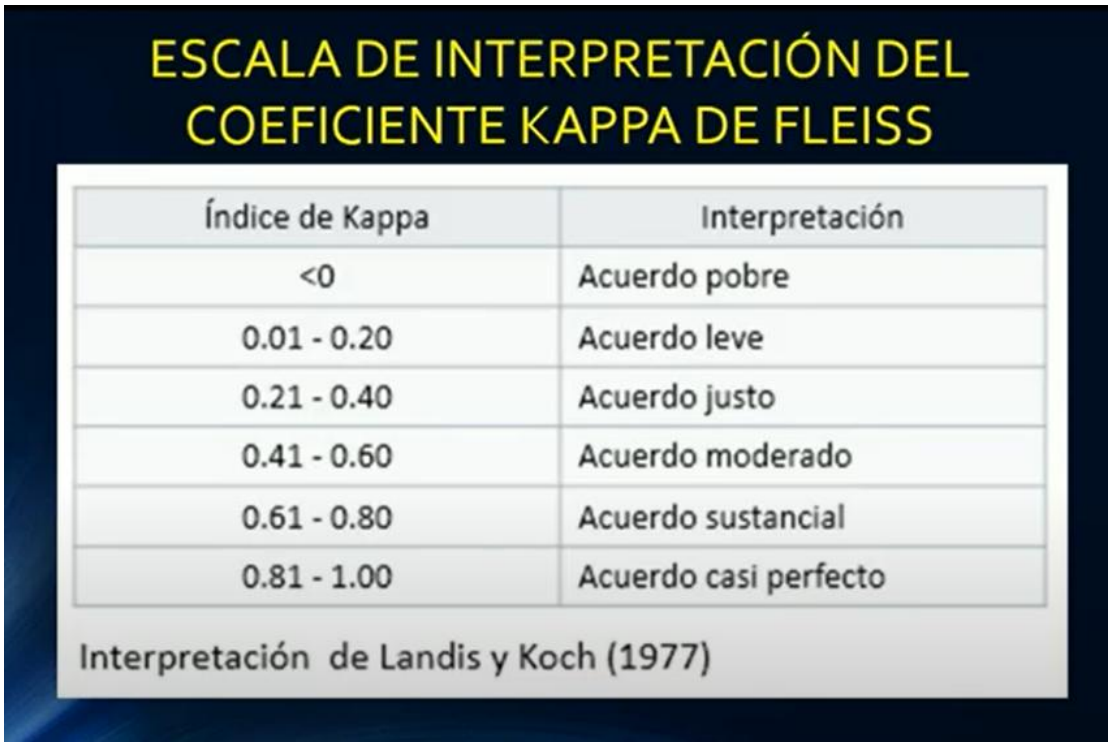


Fleiss's Kappa

alpha	0,05				
tails	2				
	Total	0	5		0
kappa	0,610	1	1		0,641148325
s.e.	0,099	0	0		0,129099445
z-stat	6,150	6	4		4,966313573
p-value	0,000	0	0		6,82376E-07
lower	0,415	0	0		0,388118063
upper	0,804	1	1		0,894178588

Interpretación: El valor de Kappa encontrado (0.61) equivale a un acuerdo sustancial de los expertos frente al criterio que presenta el instrumento. Esto significa que los cinco expertos coinciden, de manera significativa, en que

los ítems del cuestionario son predictivos de la conducta futura de consumo de productos integrales.



Conclusión: El instrumento presenta una validez de criterio aceptable, respaldada por el acuerdo sustancial entre los expertos.

4.2 Confiabilidad (Fiabilidad): La Constancia de la Medición

Una situación importante para hablar de un instrumento confiable es que diversas mediciones de un mismo sujeto utilizando el mismo instrumento deben dar resultados iguales o muy similares. Existen varios métodos para calcular la confiabilidad. En estos, al igual que en las formas de calcular la validez, el concepto de correlación o asociación está muy ligado: a mayor asociación (correlación) se puede hablar de un mayor coeficiente de confiabilidad de las pruebas (Furr, 2022).

La confiabilidad se refiere a la confianza que se tiene en los datos recolectados, debido a que hay una repetición constante, estable de la medida. La confiabilidad debe obtenerse siempre con los datos de cada muestra para garantizar la medida fiable del constructo en la muestra concreta de investigación (Streiner et al., 2024).

¿Qué es el Error de Medición?

Hace más de cuarenta años, Robert L. Thorndike (en Muñiz, 1998) empezaba su famoso trabajo sobre confiabilidad con estas palabras: "Cuando medimos algo, bien sea en el campo de la física, de la biología, o de las ciencias sociales, esa medición contiene una cierta cantidad de errores aleatorios. La cantidad de errores puede ser grande o pequeña, pero está siempre presente en cierto grado" (p. 6). Su palabra sigue vigente en la actualidad, pues en lo esencial los problemas de la medición cambian poco, aunque los instrumentos de medida vayan y vengan.

Cuando un investigador aplica un test, una escala, un inventario o cualquier otro instrumento de medida a una persona, obtiene una cierta puntuación, que por razones obvias se denomina puntuación empírica. ¿Cómo estar seguros de que esa puntuación obtenida es la que verdaderamente le corresponde a esa persona en esa prueba? En otras palabras, ¿cuánto error afecta a esa puntuación empírica?

Responder a estas preguntas es el objetivo de la confiabilidad. Sin embargo, para algunas personas tales interrogantes parecerían incontestables, pues, al fin y al cabo, el error cometido, sea el que sea, está diluido en la puntuación empírica y no hay manera de separarlo directamente. La teoría de la

confiabilidad nos proporciona los métodos para estimar la magnitud de ese error.

Escalas para Interpretar la Confiabilidad

Diversos autores han propuesto escalas para interpretar los coeficientes de confiabilidad:

De Vellis (2017):

VALOR DE ALFA	INTERPRETACIÓN
Por debajo de 0.60	Inaceptable
0.60 a 0.65	Indeseable
0.65 a 0.70	Mínimamente aceptable
0.70 a 0.80	Respetable
0.80 a 0.90	Muy buena
0.90 a 1.00	Excelente

Murphy y Davishofer (en Hogan, 2004):

VALOR DE ALFA	INTERPRETACIÓN
Alrededor de 0.90	Nivel elevado de confiabilidad
0.80 o superior	Moderada
Alrededor de 0.70	Baja
Inferior a 0.60	Inaceptablemente baja

4.2.1 Estabilidad en el Tiempo: Método Test-Retest

El método Test-Retest como medida de estabilidad de un instrumento consiste en que la misma prueba se aplica dos veces para verificar la confiabilidad de esta (Anastasi & Urbina, 1998). Confiabilidad se refiere al grado en que su aplicación repetida al mismo sujeto u objeto produce resultados iguales. Los valores estarán entre 0 y 1; el resultado no puede ser negativo.

Se refiere a que si el test se aplica hoy o dentro de un tiempo, siga siendo válido y fiable, es decir, que se encuentre una relación entre lo que se obtiene hoy y lo que se obtiene más adelante.

Procedimiento:

El investigador debe aplicar el mismo instrumento dos veces al mismo grupo después de cierto periodo. El grupo debe tener características similares a la muestra. Cuando los datos son variables cuantitativas se utiliza la correlación de Pearson. El coeficiente de correlación de Pearson altamente positivo indica un instrumento confiable. Cuando la variable es cualitativa se utiliza el coeficiente de correlación de Spearman.

Fórmula de correlación de Pearson:

$$\rho = r_{xy} = \frac{n\sum_{i=1}^{n} x_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]}}$$

Donde:

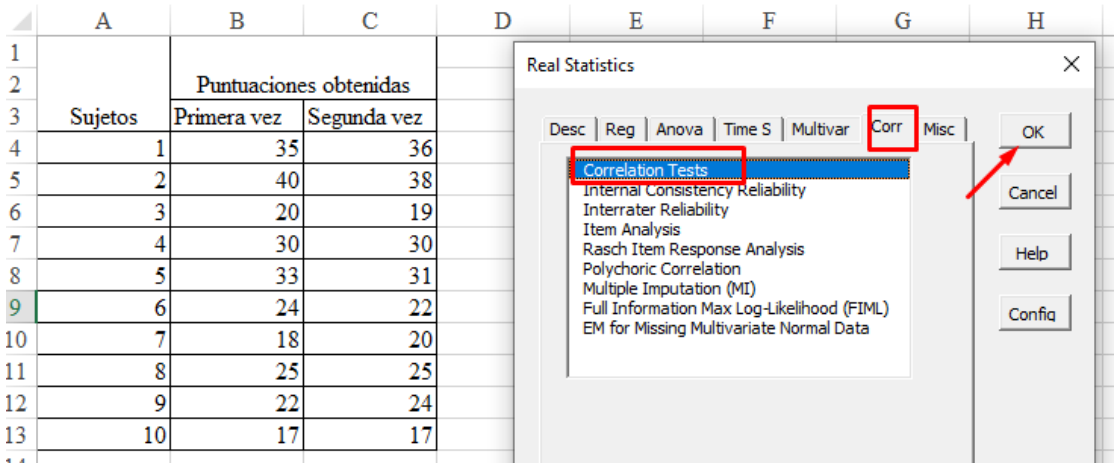
- x: Puntuaciones de la primera aplicación
- y: Puntuaciones de la segunda aplicación
- $\bar{x}$: Media de la primera aplicación
- $\bar{y}$: Media de la segunda aplicación

Ejemplo Práctico: Test-Retest con 10 Estudiantes

Se desea estimar si el aprendizaje de Ciencias Sociales de un grupo de 10 estudiantes es estable. Se somete al grupo a una misma prueba en dos ocasiones, con un intervalo de tiempo, y se obtienen los siguientes resultados:

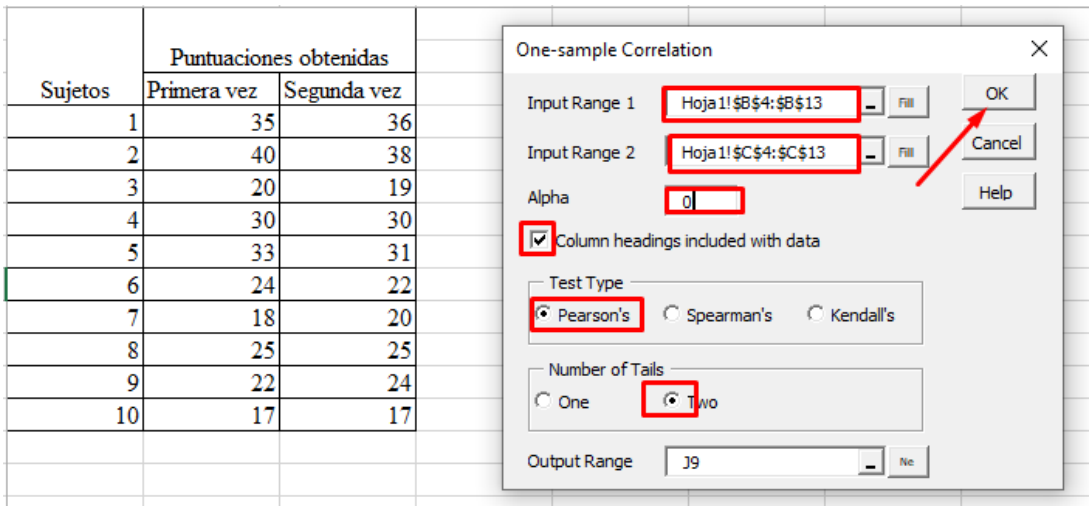
SUJETOS	PRIMERA VEZ (X)	SEGUNDA VEZ (Y)
1	35	36
2	40	38
3	20	19
4	30	30
5	33	31
6	24	22
7	18	20
8	25	25
9	22	24
10	17	17

Como el base de datos se encuentra en EXCEL aprovechamos para procesar la correlación de Pearson activando el complemento REAL STATISTICS dentro de EXCEL para activar dicho complemento primero tiene que descargarse como complemento dentro de EXCEL y se activa CONTROL M. También se desarrolla manualmente y con cualquier programa de estadística.



Sujetos	Puntuaciones obtenidas	
	Primera vez	Segunda vez
1	35	36
2	40	38
3	20	19
4	30	30
5	33	31
6	24	22
7	18	20
8	25	25
9	22	24
10	17	17

Se identifica los rangos y considerar el valor de alpha valor 0



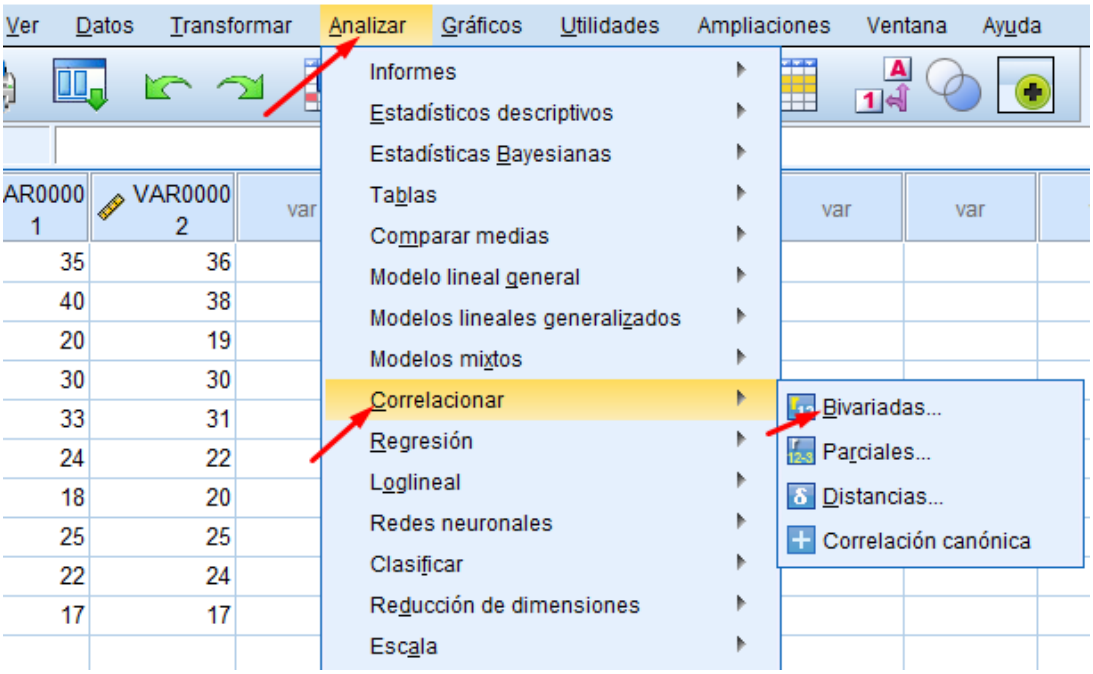
Se obtiene el valor de correlación de Pearson

Pearson	0.98109265

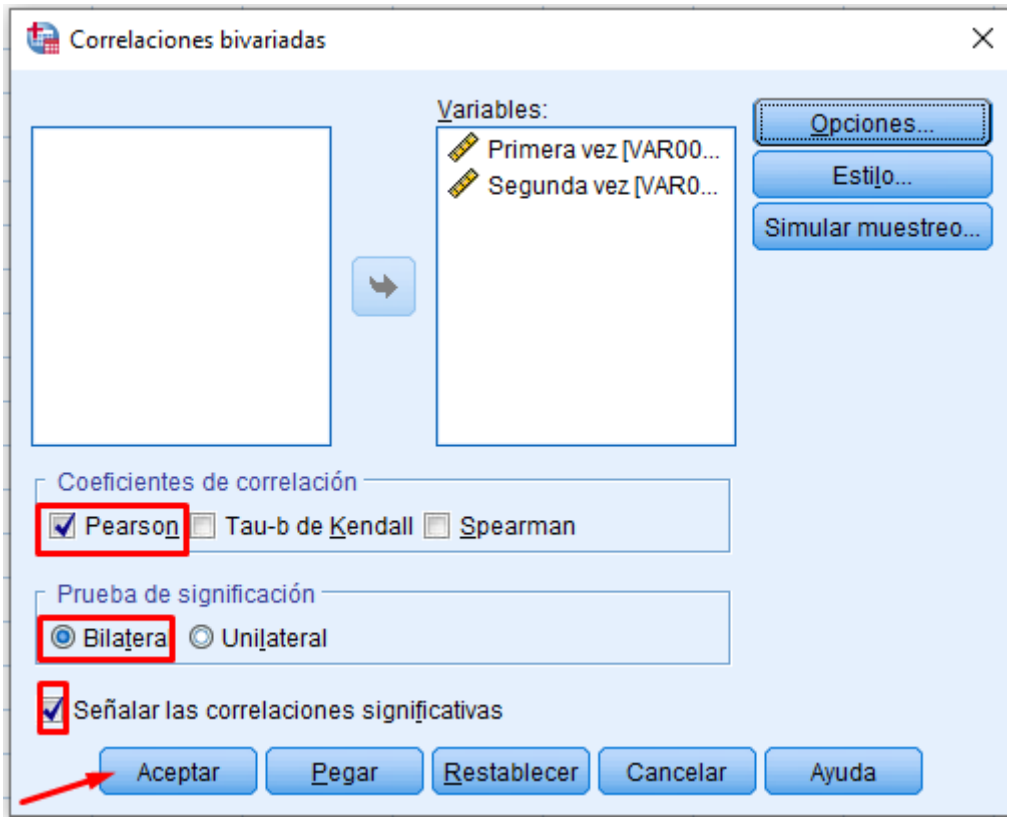
Este resultado indica que existe una correlación muy alta entre las puntuaciones de la primera y la segunda medición, lo cual equivale a decir que el instrumento analizado es altamente confiable, entonces se puede inferir que el grado de estabilidad de los aprendizajes de Ciencias Sociales de estos estudiantes ha permanecido real y efectivamente estable.

Cálculo utilizando SPSS:

1) Ingresar los datos en SPSS: dos variables (primera, segunda).
2) Ir al menú **Analizar → Correlaciones → Bivariadas**.



3) Seleccionar las dos variables y elegir coeficiente de Pearson.
4) Hacer clic en Aceptar.



Resultado:

Correlaciones

		Primera vez	Segunda vez
Primera vez	Correlación de Pearson	1	,981**
	Sig. (bilateral)		,000
	N	10	10
Segunda vez	Correlación de Pearson	,981**	1
	Sig. (bilateral)	,000	
	N	10	10

**. La correlación es significativa en el nivel 0,01 (bilateral).

Es el mismo resultado obtenido con REAL STATISTICS

Interpretación: El coeficiente de correlación de Pearson obtenido es r = 0.978, lo que indica una correlación muy alta entre las puntuaciones de la

primera y la segunda medición. Esto equivale a decir que el instrumento analizado es altamente confiable en términos de estabilidad temporal. Podemos inferir que el grado de estabilidad de los aprendizajes de Ciencias Sociales de estos estudiantes ha permanecido real y efectivamente estable.

4.2.2 Equivalencia: Método de Mitades Partidas (Guttman, Spearman-Brown)

El método de mitades partidas requiere solo una aplicación de la medición al conjunto total de los ítems. El instrumento se divide en dos mitades (por ejemplo, ítems pares e impares, o primera mitad y segunda mitad) y se comparan los resultados o puntuaciones de ambas. Si el resultado de ambas mitades es similar, se concluye que el instrumento es confiable; es decir, están fuertemente correlacionados (Furr, 2022).

Un individuo con baja puntuación en una mitad tenderá a tener también una baja puntuación en la otra mitad.

División por mitades de Guttman: También se denomina coeficiente de consistencia interna. Su fórmula es:

$$r = 2\left[1 - \frac{S_1^2 + S_2^2}{S_t^2}\right]$$

r: Coeficiente de confiabilidad
S_1^2: *Varianza* de las puntuaciones de los ítems pares
S_2^2: *Varianza* de las puntuaciones de los ítems impares
S_t^2: *Varianza* de las puntuaciones del test total

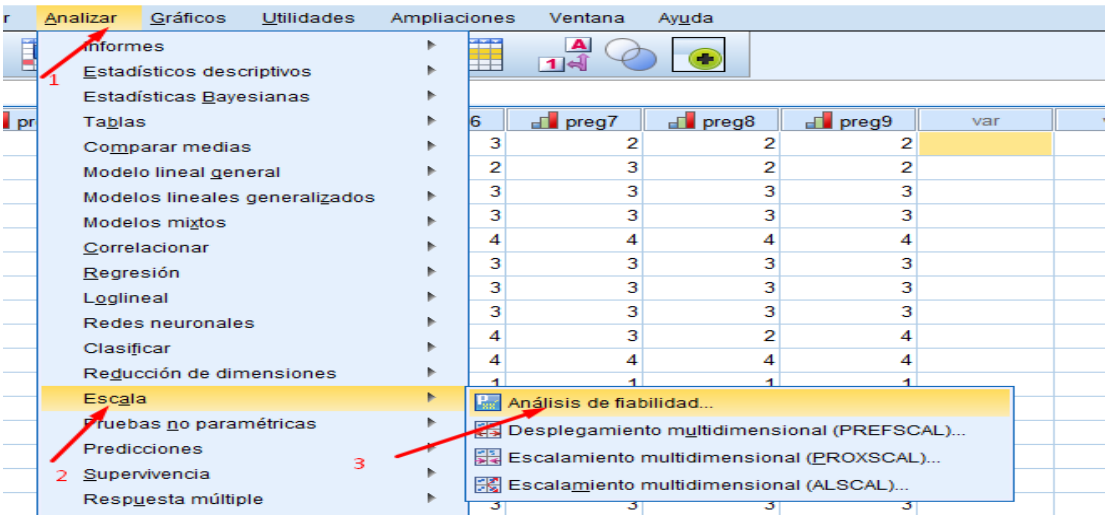
Ejemplo Práctico: Mitades Partidas con SPSS

Utilizaremos la misma base de datos del ejemplo de Validez de Constructo (9 ítems, 78 sujetos cargados en SPSS).

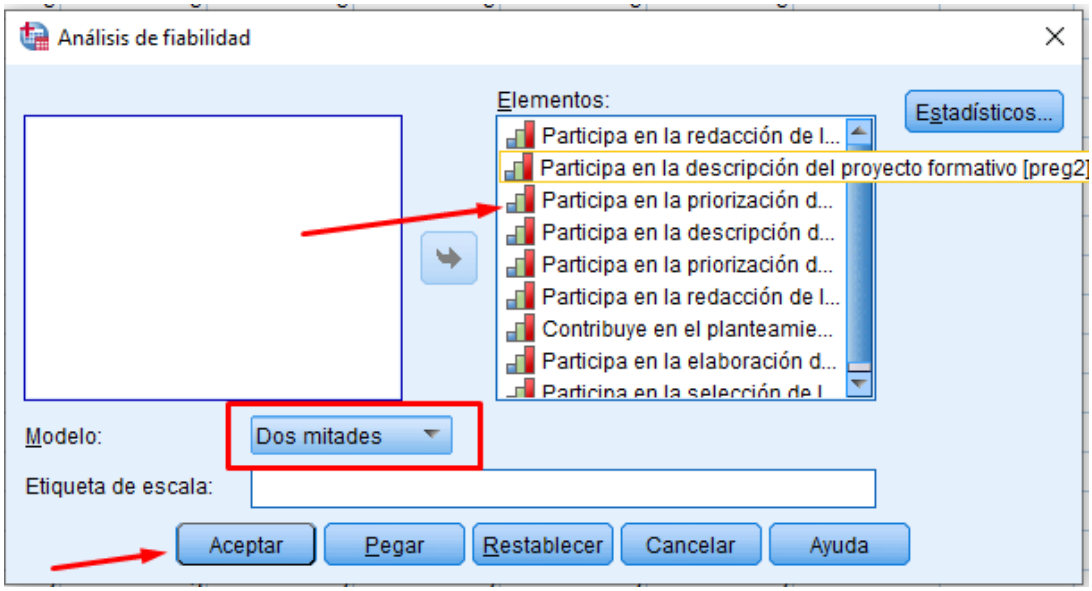
	preg1	preg2	preg3	preg4	preg5	preg6	preg7	preg8	preg9
1	3	4	2	3	3	3	2	2	2
2	2	3	3	3	3	2	3	2	2
3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	3	3
5	4	4	4	4	4	4	4	4	4
6	3	3	4	4	4	3	3	3	3
7	3	3	3	3	3	3	3	3	3
8	3	3	3	3	3	3	3	3	3
9	4	4	4	4	4	4	3	2	4
10	4	4	3	4	3	4	4	4	4
11	2	1	1	1	1	1	1	1	1
12	4	4	4	4	4	4	4	4	4
13	3	4	3	2	3	3	3	3	4
14	3	3	4	3	3	3	3	3	3
15	3	3	4	4	3	4	3	4	3
16	3	3	3	3	3	3	3	3	3
17	1	1	1	3	1	1	1	1	1
18	3	3	3	3	2	3	3	3	3
19	3	3	3	3	3	3	3	3	3
20	3	3	3	3	4	3	3	3	4
21	2	2	3	2	2	2	2	3	4
22	4	3	4	4	3	3	3	3	4
23	3	3	4	3	3	4	3	3	4

Procedimiento en SPSS:

1) Ir al menú **Analizar → Escala → Análisis de fiabilidad**.



2) Seleccionar las 9 variables (ítems) y pasarlas a la lista de "Elementos". Por defecto en el modelo aparece Alfa, como se está tratando el método de mitades partidas, en el desplegable "Modelo", seleccionar Dos mitades (Split-half).
3) Hacer clic en Estadísticos y marcar las opciones deseadas (descriptivos, correlaciones, etc.).
4) Hacer clic en Aceptar.



Resultados obtenidos:

Estadísticas de fiabilidad

Alfa de Cronbach	Parte 1	Valor	,920
		N de elementos	5[a]
	Parte 2	Valor	,941
		N de elementos	4[b]
	N total de elementos		9
Correlación entre formularios			,889
Coeficiente de Spearman-Brown	Longitud igual		,941
	Longitud desigual		,942
Coeficiente de dos mitades de Guttman			,939

a. Los elementos son: Participa en la redacción de los objetivos de aprendizaje, Participa en la descripción del proyecto formativo, Participa en la priorización de las competencias genéricas, Participa en la descripción de los resultados, Participa en la priorización del problema.

b. Los elementos son: Participa en la priorización del problema, Participa en la redacción de la transversalidad, Contribuye en el planteamiento de las características, Participa en la elaboración de la matriz, Participa en la selección de la referencia bibliográfica.

Interpretación:

- La base de datos tiene 78 sujetos y 9 ítems. La Parte 1 tiene 5 ítems (un elemento más por ser número impar), la Parte 2 tiene 4 ítems.
- El coeficiente de Spearman-Brown para longitud igual es 0.941 y para longitud desigual es 0.942.
- El coeficiente de dos mitades de Guttman es 0.939.

Todos estos valores se encuentran alrededor de 0.94, muy por encima del umbral de 0.80 considerado como "muy buena" confiabilidad. Por lo tanto, concluimos que el instrumento es altamente confiable; las puntuaciones de ambas mitades están fuertemente correlacionadas.

4.2.3 Consistencia Interna: El Corazón de la Confiabilidad

La consistencia interna evalúa el grado en que los ítems de un instrumento de medición están correlacionados entre sí. Pone énfasis en las puntuaciones obtenidas y no en el contenido o el formato de los ítems. Consiste en administrar una misma prueba a un grupo de sujetos en una sola oportunidad (L. J. Cronbach, 1951) (García-García et al., 2024).

Existen diferentes procedimientos para estimar la confiabilidad por consistencia interna. Los más conocidos son: el coeficiente KR-20 de Kuder-Richardson (para ítems dicotómicos) y el Alfa de Cronbach (para ítems politómicos o escalas Likert).

Confiabilidad Coeficiente KR-20 (Kuder-Richardson)

El coeficiente **KR-20** permite obtener la confiabilidad a partir de los datos obtenidos en una sola aplicación del test. Es aplicable a instrumentos con ítems de opción dicotómica (sí/no, verdadero/falso, correcto/incorrecto) y cuando existen alternativas dicotómicas con respuestas correctas (1) e incorrectas (0). Es análogo al Alfa de Cronbach, con la diferencia de que el Alfa, se utiliza para medidas continuas no dicotómicas (Kuder & Richardson, 1937).

Fórmula del KR-20:

$$KR_{20} = r_{20} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{n} p\,q}{S_t^2}\right]$$

r_{20} : *Correlación* de la prueba
k:Es rl número de ítems
S_t^2:Variancia total de las puntuaciones
p:Probabilidad de aciertos de los ítems
q=1-p=Probabilidad de los desaciertos del ítems

Ejemplo Práctico: Cálculo de KR-20

Supongamos que tenemos un test de 10 preguntas dicotómicas (1 = correcto, 0 = incorrecto) aplicado a 15 estudiantes.

Matriz de datos:

Individuos	Preguntas										
	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	Totales
1	0	1	1	0	1	1	0	1	1	0	6
2	0	1	0	1	1	1	1	1	1	1	8
3	1	0	1	1	1	1	0	1	1	0	7
4	1	1	1	1	0	1	1	1	1	1	9
5	1	0	0	1	1	0	1	0	1	0	5
6	1	1	1	1	1	1	1	1	1	1	10
7	1	0	0	0	0	1	0	0	0	0	2
8	1	1	1	1	1	1	1	1	1	1	10
9	1	0	1	1	1	1	0	1	1	0	7
10	1	1	1	1	1	1	0	1	1	1	9
11	1	0	0	0	0	1	1	1	0	0	4
12	0	1	1	1	1	1	0	0	1	1	7
13	1	0	0	1	1	0	1	0	1	0	5
14	0	0	1	1	1	1	0	1	0	1	6
15	1	1	1	1	1	1	1	1	1	0	9
Totales	11	8	10	12	12	13	8	11	12	7	
p	0.73	0.53	0.67	0.80	0.80	0.87	0.53	0.73	0.80	0.47	
q	0.27	0.47	0.33	0.20	0.20	0.13	0.47	0.27	0.20	0.53	
(p*q)	0.20	0.25	0.22	0.16	0.16	0.12	0.25	0.20	0.16	0.25	
Sum(p*q)	1.96										
Var	5.35										
K	10										

$$KR_{20} = \frac{10}{10-1}\left[1 - \frac{1,96}{5,35}\right]$$

$$KR_{20} = 1,11[1-0,37] = 0,699 \approx 0,7$$

KR-20	Interpretación
0,9 - 1	EXCELENTE
0,8 - 0,9	BUENA
0,7 - 0,8	ACEPTABLE
0,6 - 0,7	DEBIL
0,5 - 0,6	POBRE
<0,5	INACEPTABLE

Interpretación. De acuerdo con el resultado KR20= 0,7. se concluye que el instrumento en estudio tiene una confiabilidad aceptable.

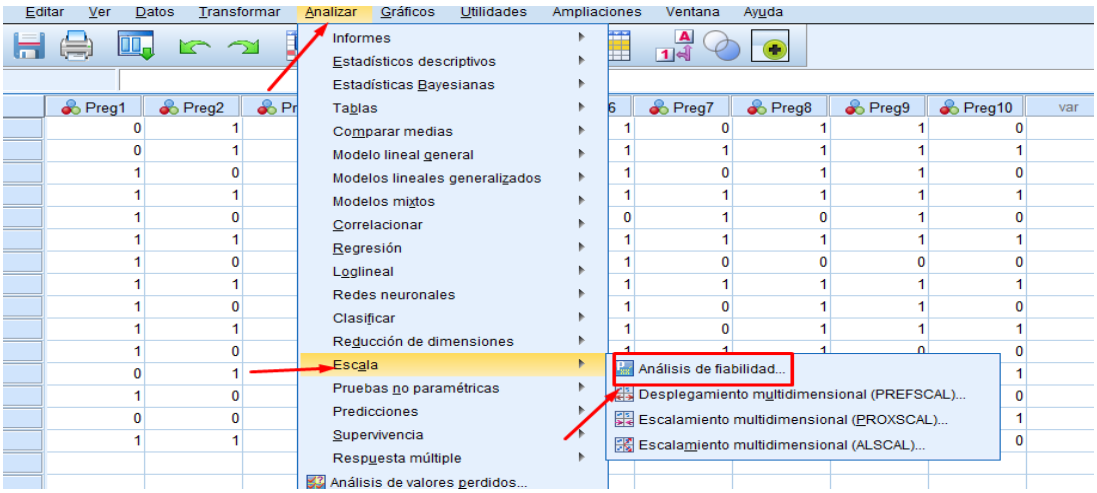
Nota: los palores de p se calcula $p = \frac{preguntas}{total \text{ de individuos}(15)}$

$q = 1 - p$

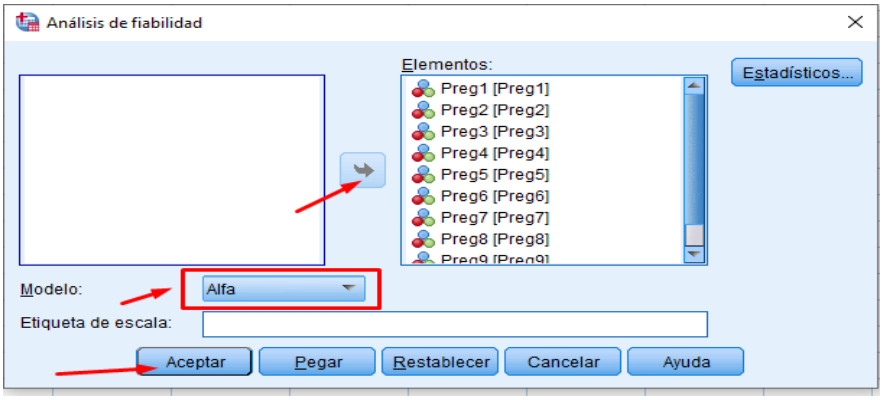
La varianza se obtiene con los totales de cada individuo que han respondido o acertaron de las 10 preguntas es decir con los valores: 6,8,7,9,5,10,2.10,7,9,4,7,5,6,9.

Cálculo de Confiabilidad de KR20 utilizando SPSS:

1) Ingresar los datos en SPSS (cada ítem dicotómico como variable).
2) Ir al menú **Analizar → Escala → Análisis de fiabilidad**.



3) Seleccionar los 10 ítems y pasarlos a la lista de elementos.
4) En el desplegable "Modelo", seleccionar **Alfa** (SPSS no tiene una opción específica para KR-20, pero cuando los datos son dicotómicos, el Alfa de Cronbach es equivalente al KR-20).
5) Hacer clic en Aceptar.



Resultado obtenido:

Alfa de Cronbach = 0.676

Interpretación: El análisis de confiabilidad arroja un valor de 0.676. Redondeando a un decimal, obtenemos 0.7, que es muy similar al valor calculado manualmente (0.702). Concluimos que el instrumento tiene una confiabilidad aceptable.

Confiabilidad del Coeficiente Alfa de Cronbach: El Rey de la Consistencia Interna

Para evaluar la confiabilidad o la homogeneidad de las preguntas o ítems, es común emplear el coeficiente Alfa de Cronbach cuando se trata de alternativas de respuestas politómicas (siempre, a veces, nunca; de acuerdo, ni de acuerdo ni en desacuerdo, en desacuerdo; etc.), como las escalas de Likert (L. J. Cronbach, 1951).

El Alfa de Cronbach es, sin duda, el coeficiente de confiabilidad más utilizado en ciencias sociales. Su popularidad se debe a que requiere una sola administración del instrumento y proporciona una estimación de la consistencia interna basada en el promedio de las correlaciones entre los ítems.

Cálculo del coeficiente Alfa de Cronbach:

A) Mediante la varianza de los ítems y la varianza del puntaje total:

$$\alpha = \frac{k}{k-1}(1-\frac{\sum S_i^2}{S_t^2})$$

α : *Coeficiente* de Cronbach
k:Es el número de pregintas o ítems
S_i^2 : *Es* la suma de varianza de cada ítems
S_t^2 : *Es* la varianza del total de filas (puntaje de los jueces)

B) Mediante la matriz de correlación de los ítems:

$$\alpha = \frac{np}{1+p(n-1)}$$

n : *es* el número de ítems
p:Es el promedio de las correlaciones lineales ente cada uno de los ítems

Ejemplo Práctico: Cálculo de Alfa de Cronbach

Utilizaremos un ejemplo de un cuestionario sobre "Experiencia en la ejecución curricular" aplicado a 78 docentes. El instrumento tiene 13 ítems con escala Likert de 4 puntos (1 = Insatisfactorio, 2 = Poco satisfactorio, 3 = Satisfactorio, 4 = Muy satisfactorio).

Matriz de datos:

	Ítems													Total
Docentes	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	
1	2	3	3	4	4	4	4	3	4	4	4	4	4	47
2	2	2	3	3	2	3	3	3	2	3	3	2	2	33

3	3	3	3	3	3	3	3	3	3	3	3	3	3	39
4	3	3	3	3	3	3	3	3	3	3	3	3	3	39
5	4	4	4	4	4	4	4	4	4	4	4	4	4	52
6	3	3	3	3	3	3	3	3	3	3	4	3	4	41
7	3	3	3	3	3	3	3	3	3	3	3	3	3	39
8	3	3	3	3	3	3	2	3	3	3	3	3	3	38
9	3	3	3	3	3	3	2	4	3	2	4	4	4	41
10	3	3	4	4	4	4	4	4	4	4	4	4	4	50
11	3	3	3	3	3	3	3	2	3	4	4	4	4	42
12	4	4	4	4	4	4	4	4	4	4	4	4	4	52
13	3	4	3	3	3	2	3	4	3	3	3	3	3	40
14	3	3	4	4	3	2	4	3	4	4	3	3	4	44
15	3	3	3	3	3	3	3	4	3	4	3	3	3	41
16	2	2	3	3	3	3	3	3	3	3	3	3	3	37
17	1	1	1	1	1	1	1	1	3	3	3	2	3	22
18	3	3	3	3	3	3	3	2	2	2	3	3	2	35
19	3	3	3	3	3	3	3	3	3	3	3	3	3	39
20	3	2	3	3	3	3	4	4	2	3	3	3	2	38
21	3	3	3	4	3	3	4	4	3	4	4	3	3	44
22	4	4	4	4	4	4	4	4	3	3	4	3	4	49
23	4	3	4	3	3	4	4	4	3	3	3	3	3	44
24	3	3	3	3	3	2	4	4	3	3	3	4	3	41
25	3	3	3	3	3	2	4	3	3	3	3	3	3	39
26	3	3	3	3	3	3	3	1	4	4	3	4	3	40
27	3	2	3	3	3	3	2	3	3	3	3	3	3	37
28	3	3	3	3	3	2	3	3	4	4	4	3	4	42
29	4	3	3	2	3	3	3	3	2	3	3	3	3	38
30	3	2	3	2	2	2	2	2	2	3	3	2	3	31
31	3	3	2	3	3	2	3	2	2	2	3	3	3	34
32	3	3	3	3	3	3	3	3	3	3	3	3	3	39
33	2	3	3	2	3	3	3	3	3	3	3	3	3	37
34	3	3	3	3	3	3	3	3	3	3	3	3	3	39
35	3	2	3	3	3	3	3	3	3	2	3	3	3	37
36	3	3	4	3	4	3	3	3	3	4	3	4	3	43
37	3	3	3	2	2	2	3	3	3	3	3	3	3	36
38	2	2	2	2	2	2	2	2	2	2	2	2	2	26
39	2	2	2	2	2	2	2	2	2	2	2	2	2	26
40	2	2	2	2	2	2	2	2	2	2	2	2	2	26
41	4	4	4	4	4	4	4	4	4	4	4	4	4	52
42	3	3	2	3	2	3	2	3	3	3	2	3	3	35
43	1	1	1	4	4	4	4	4	4	4	4	4	4	43
44	4	4	4	4	4	3	4	4	4	4	4	4	4	51

45	3	3	3	4	4	4	4	4	4	4	4	4	4	49
46	3	3	3	3	3	2	3	1	1	2	1	1	2	28
47	2	3	2	3	3	2	3	2	2	3	2	3	3	33
48	1	1	1	1	1	1	1	1	1	1	1	1	1	13
49	1	1	1	1	1	1	1	1	1	1	1	1	1	13
50	3	3	3	3	3	3	3	3	3	3	3	3	3	39
51	2	3	3	2	3	3	4	3	4	4	4	2	3	40
52	4	3	4	3	4	4	4	4	3	4	4	4	4	49
53	3	3	3	3	3	3	3	3	3	2	3	2	3	37
54	4	4	4	4	3	3	3	3	3	3	4	4	4	46
55	3	3	3	3	3	3	3	2	4	3	3	4	3	40
56	1	1	1	1	1	1	1	1	1	1	1	1	1	13
57	3	3	3	3	3	3	3	3	3	4	3	3	3	40
58	4	4	4	4	4	4	4	4	4	4	4	4	4	52
59	3	4	4	4	4	4	4	4	4	4	4	4	4	51
60	3	3	3	3	3	3	3	3	3	3	3	3	3	39
61	3	4	4	3	4	3	4	3	4	4	3	3	4	46
62	3	3	4	4	4	4	3	3	3	3	4	4	4	46
63	3	3	3	3	3	3	3	3	3	3	3	3	3	39
64	3	3	3	4	3	2	4	4	3	3	4	3	4	43
65	3	3	3	3	3	3	3	3	3	3	3	3	3	39
66	3	3	3	3	3	3	3	3	3	3	3	3	3	39
67	2	2	2	2	2	2	2	2	2	2	2	3	2	27
68	4	4	4	4	4	4	4	4	4	4	4	4	4	52
69	3	3	3	3	3	3	3	3	3	3	3	3	3	39
70	3	2	3	4	3	2	4	3	2	3	4	3	4	40
71	3	3	3	3	3	3	3	3	3	3	3	3	3	39
72	3	2	2	2	3	2	3	2	3	2	2	2	3	31
73	3	4	3	4	3	4	3	4	3	3	3	4	3	44
74	3	3	3	3	3	3	3	3	3	3	3	3	4	40
75	2	3	2	4	3	1	3	3	3	3	4	3	3	37
76	4	4	4	4	4	4	4	4	4	4	4	4	4	52
77	3	3	3	3	2	3	2	3	3	3	3	3	3	37
78	3	3	3	3	3	3	3	3	3	3	3	3	3	39
Total	223	222	229	233	230	219	236	230	228	236	240	235	241	
Var. De cada ítems	0.54	0.57	0.597	0.61	0.55	0.67	0.64	0.74	0.6	0.59	0.61	0.61	0.56	
Suma Var	7.88													
Var. Total	71.5													

K=13

$$\alpha = \frac{13}{13-1}\left[1 - \frac{7,883949}{71,5373}\right]$$

$$\alpha = \frac{13}{12}[1 - 0,11]$$

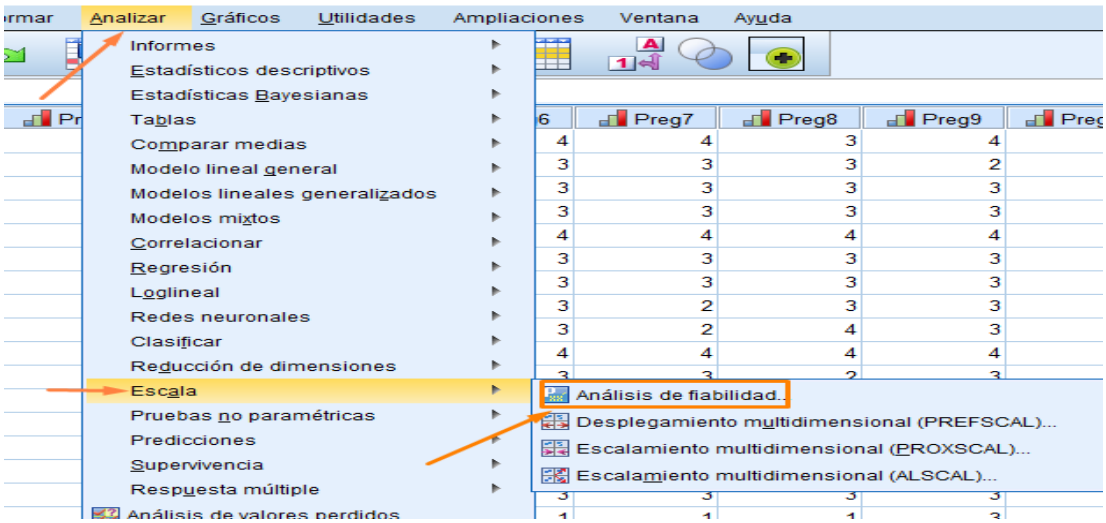
$$\alpha = 0,9612$$

Interpretación: De acuerdo con el resultado (α = 0.9612), se concluye que el instrumento "Experiencia en la ejecución curricular" tiene una confiabilidad excelente (superior a 0.90). Las escalas de valoración de (DeVellis & Thorpe, 2022) y Murphy y Davishofer coinciden en que valores superiores a 0.90 indican un nivel de confiabilidad muy alto.

Cálculo de Alfa de Cronbach utilizando SPSS:

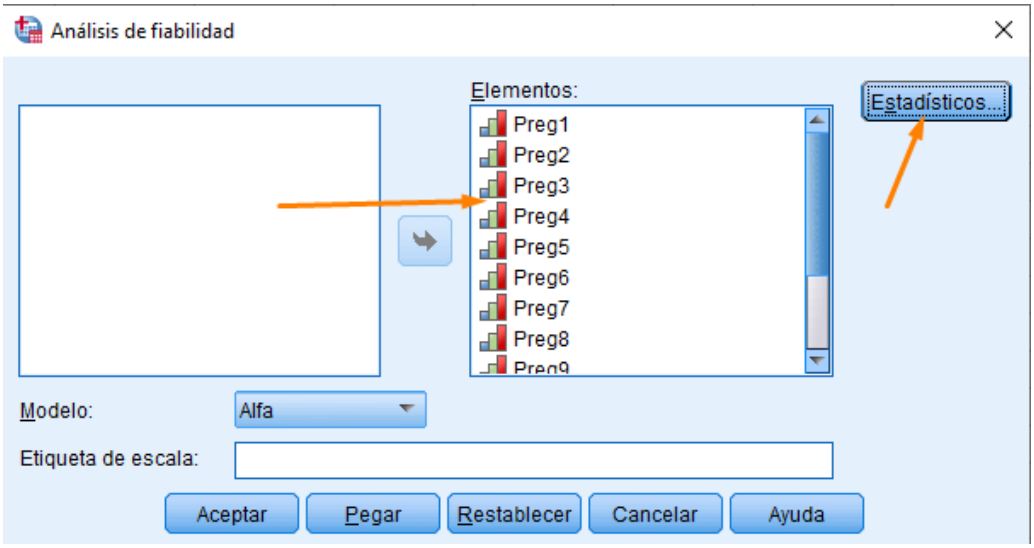
Procedimiento:

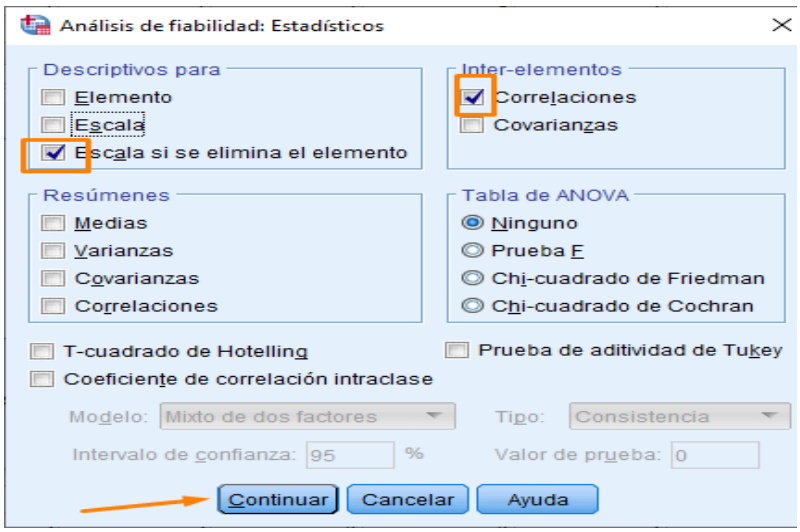
1) Ir al menú **Analizar → Escala → Análisis de fiabilidad**.



2) Seleccionar las 13 variables (P1 a P13) y pasarlas a la lista de "Elementos".
3) En el desplegable "Modelo", asegurarse de que está seleccionado **Alfa**.

4) Hacer clic en **Estadísticos** y marcar:
   - Descriptivos para: Escala si el elemento se ha eliminado
   - Correlaciones entre elementos
   - Correlación elemento-total corregida
5) Hacer clic en Continuar y luego en Aceptar.

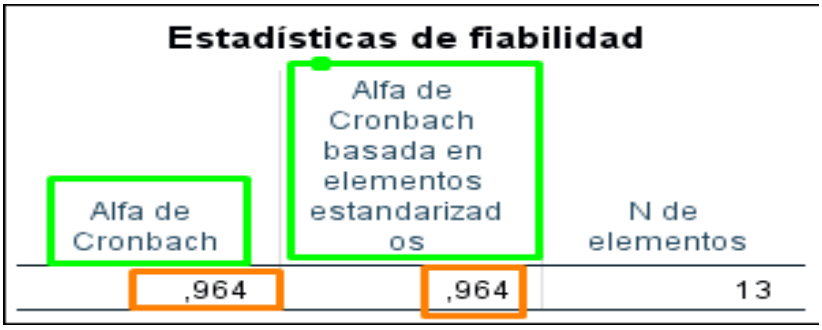




Resultados obtenidos:

```
Estadísticas de fiabilidad
---------------------------
Alfa de Cronbach    N de elementos
     0.964                13

Estadísticas total-elemento
----------------------------
      Media de    Varianza   Correlación   Alfa de
      escala si   de escala  elemento-     Cronbach si
      se elimina  si se      total         el elemento
      el elemento elimina el  corregida     se ha
                  elemento                  eliminado
P1     43.85       65.23        0.52          0.965
P2     43.86       64.89        0.58          0.964
P3     43.79       64.21        0.75          0.961
P4     43.76       63.98        0.78          0.961
P5     43.78       64.56        0.70          0.962
P6     43.85       63.45        0.79          0.960
P7     43.73       64.11        0.76          0.961
P8     43.78       63.45        0.81          0.960
P9     43.81       64.22        0.73          0.962
P10    43.73       64.11        0.77          0.961
P11    43.69       63.89        0.80          0.960
P12    43.74       63.98        0.78          0.961
P13    43.68       64.45        0.74          0.962
```



Estadísticas de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en elementos estandarizados	N de elementos
,964	,964	13

Interpretación:

- El Alfa de Cronbach obtenido (0.964) es prácticamente idéntico al calculado manualmente (la pequeña diferencia se debe a redondeos).
- Las correlaciones elemento-total corregidas son todas superiores a 0.50, lo que indica que cada ítem contribuye positivamente a la consistencia interna.
- Observamos que si eliminásemos el ítem P1, el Alfa aumentaría ligeramente a 0.965, pero la ganancia es mínima (0.001). En general, ningún ítem, si se elimina, mejora sustancialmente el Alfa, lo que sugiere que todos los ítems son valiosos para la escala.

Matriz de correlaciones entre elementos:

	Preg1	Preg2	Preg3	Preg4	Preg5	Preg6	Preg7	Preg8	Preg9	Preg10	Preg11	Preg12	Preg13
Preg1	1,000	,791	,837	,594	,643	,597	,585	,591	,450	,473	,523	,578	,592
Preg2	,791	1,000	,796	,668	,698	,583	,615	,597	,596	,574	,555	,631	,624
Preg3	,837	,796	1,000	,670	,751	,693	,695	,644	,586	,637	,631	,629	,655
Preg4	,594	,668	,670	1,000	,811	,660	,763	,717	,626	,644	,740	,744	,744
Preg5	,643	,698	,751	,811	1,000	,774	,833	,695	,706	,709	,722	,768	,773
Preg6	,597	,583	,693	,660	,774	1,000	,611	,681	,632	,617	,620	,708	,604
Preg7	,585	,615	,695	,763	,833	,611	1,000	,717	,632	,706	,691	,616	,672
Preg8	,591	,597	,644	,717	,695	,681	,717	1,000	,605	,610	,720	,660	,646
Preg9	,450	,596	,586	,626	,706	,632	,632	,605	1,000	,812	,739	,737	,790
Preg10	,473	,574	,637	,644	,709	,617	,706	,610	,812	1,000	,742	,707	,745
Preg11	,523	,555	,631	,740	,722	,620	,691	,720	,739	,742	1,000	,757	,857
Preg12	,578	,631	,629	,744	,768	,708	,616	,660	,737	,707	,757	1,000	,762
Preg13	,592	,624	,655	,744	,773	,604	,672	,646	,790	,745	,857	,762	1,000

Al observar la matriz de correlaciones (no reproducida completamente aquí por brevedad), notamos que el ítem P1 tiene correlaciones más bajas con algunos otros ítems (especialmente con P4, P5, P6, P7, P8, P9, P10, P11, P12 y P13). Esto sugiere que el ítem P1 podría estar midiendo un aspecto ligeramente diferente o podría tener problemas de redacción.

Recomendación práctica: Aunque la escala es excelente en su conjunto, si se desea optimizar aún más, se podría revisar la redacción del ítem P1 o considerar si realmente es necesario mantenerlo. Sin embargo, dado que su eliminación no mejora significativamente el Alfa, y el valor actual ya es excelente, lo más sensato es conservar todos los ítems para mantener la riqueza conceptual del instrumento.

Resumen del Capítulo

En este extenso capítulo hemos explorado los conceptos fundamentales de la validez y la confiabilidad, que son los pilares sobre los que se sostiene la calidad de cualquier instrumento de medición en ciencias sociales. Recordemos las ideas principales:

1) **Validez**: Se refiere a si el instrumento mide realmente lo que pretende medir.
   - Validez de contenido: Evaluada mediante juicio de expertos y cuantificada con el coeficiente V de Aiken.
   - Validez de constructo: Evaluada mediante análisis factorial, que revela la estructura subyacente de los datos.

- Validez de criterio: Evaluada mediante correlación con variables externas, ya sea de forma concurrente o predictiva. El coeficiente Kappa de Fleiss permite cuantificar el acuerdo entre expertos.

2) **Confiabilidad**: Se refiere a la consistencia y estabilidad de las mediciones.
   - Estabilidad temporal: Evaluada mediante el método test-retest y la correlación de Pearson.
   - Equivalencia: Evaluada mediante el método de mitades partidas (Spearman-Brown, Guttman).
   - Consistencia interna: Evaluada mediante el coeficiente KR-20 (para ítems dicotómicos) y el Alfa de Cronbach (para escalas Likert).

3) **Interpretación de coeficientes**:
   - Valores de V de Aiken > 0.70 indican validez de contenido aceptable.
   - Valores de KMO > 0.70 y cargas factoriales > 0.35 indican buena validez de constructo.
   - Valores de Kappa de Fleiss > 0.60 indican acuerdo sustancial entre expertos.
   - Valores de Alfa de Cronbach > 0.70 indican confiabilidad aceptable; > 0.90 indican confiabilidad excelente.
4) **Herramientas de cálculo**: Hemos visto cómo realizar estos análisis tanto manualmente (con fórmulas) como utilizando software estadístico (SPSS y Excel con complemento Real Statistics).

La lección más importante es que validez y confiabilidad no son propiedades fijas del instrumento, sino que deben ser evaluadas en cada contexto y con cada muestra. Un instrumento puede ser muy confiable en

una población y no ser válido en otra, o viceversa. Por eso, todo investigador responsable debe reportar las evidencias de validez y confiabilidad correspondientes a su estudio específico.

En el próximo capítulo (Prueba de Normalidad), exploraremos cómo verificar uno de los supuestos fundamentales para la aplicación de muchas técnicas estadísticas paramétricas.

CAPÍTULO V: PRUEBA DE NORMALIDAD: ¿NUESTROS DATOS SIGUEN UNA DISTRIBUCIÓN NORMAL?

Piensa por un momento en todo lo que hemos caminado juntos hasta aquí. Arrancamos definiendo a esa población que nos quitaba el sueño, aprendimos a elegir una muestra que realmente la representara sin trampas ni atajos, nos rompimos la cabeza construyendo un instrumento que midiera lo que queríamos medir y nos aseguramos de que fuera válido y confiable. Por fin, después de tanto esfuerzo, tenemos los datos en la mano. Llega el momento de la verdad: hay que sentarse y analizarlos.

Y entonces aparece la pregunta del millón: ¿y ahora qué pruebas uso? ¿Por dónde empiezo?

Pues mira, la respuesta tiene menos misterio del que parece, pero exige que mires tus datos con otros ojos. Porque no todas las pruebas valen para todo. Hay un grupo de ellas las más famosas, las que salen en todos los papers, las que todo el mundo quiere usar que tienen un requisito de entrada.

Exigen que los datos se comporten de una determinada manera. Que sigan esa forma de campana que tantas veces has visto en dibujos: la curva normal, la famosa campana de Gauss.

Hablamos de la prueba t de Student, esa que te permite comparar dos grupos; del ANOVA, cuando quieres meterte con tres o más; de la correlación de Pearson, para ver si dos cosas van de la mano. Todas ellas son lo que los expertos llaman pruebas paramétricas. Son potentes, son elegantes, pero son exigentes. Piden que tus datos bailen al ritmo de la normalidad.

Y claro, si tus datos no siguen ese ritmo, si se salen del compás, usar esas pruebas puede llevarte a conclusiones más bien patas arriba. De eso, justamente, vamos a hablar ahora: de cómo saber si tus datos se ajustan a la curva, y de qué hacer cuando no lo hacen. Porque no todo está perdido, hay vida más allá de la campana, y también hay herramientas para cuando los números se rebelan.

Pero ¿qué sucede si nuestros datos no siguen una distribución normal? En ese caso, debemos recurrir a las pruebas no paramétricas, que no asumen una distribución específica. La elección entre uno u otro tipo de prueba no es trivial: usar una prueba paramétrica cuando no se cumple el supuesto de normalidad puede llevar a conclusiones erróneas (sesgadas) (Sheskin, 2020).

En este capítulo, exploraremos la prueba de normalidad, una herramienta fundamental para verificar si las características de calidad que intervienen en el estudio provienen de una población que se distribuye mediante una ley normal. Como señalan (Ciro Martínez, 2012), este es el supuesto básico para realizar el control estadístico de calidad y para aplicar la estadística inferencial paramétrica.

Existen muchas comprobaciones que muestran que en el mundo biológico, sociológico, educativo, ingenieril, económico, etc., se encuentran

poblaciones que, al extraer muestras para una variable específica, presentan una distribución de frecuencias casi superponible a una curva normal. El parecido es tanto mayor cuanto más grande es el tamaño de la muestra. Esto ha hecho que se diseñen una enorme cantidad de pruebas y métodos estadísticos basándose en el comportamiento normal de las variables. Esta corriente de la estadística ha recibido el nombre de Estadística Inferencial o Estadística Paramétrica (Wayne W. Daniel, 2018).

Existe una lista de pruebas estadísticas para verificar la distribución normal de un conjunto de datos. Entre las más populares podemos mencionar: Shapiro-Wilk, Kolmogorov-Smirnov y la prueba Chi-cuadrado de bondad de ajuste. En este capítulo nos centraremos en las dos primeras, que son las más utilizadas en la práctica investigadora.

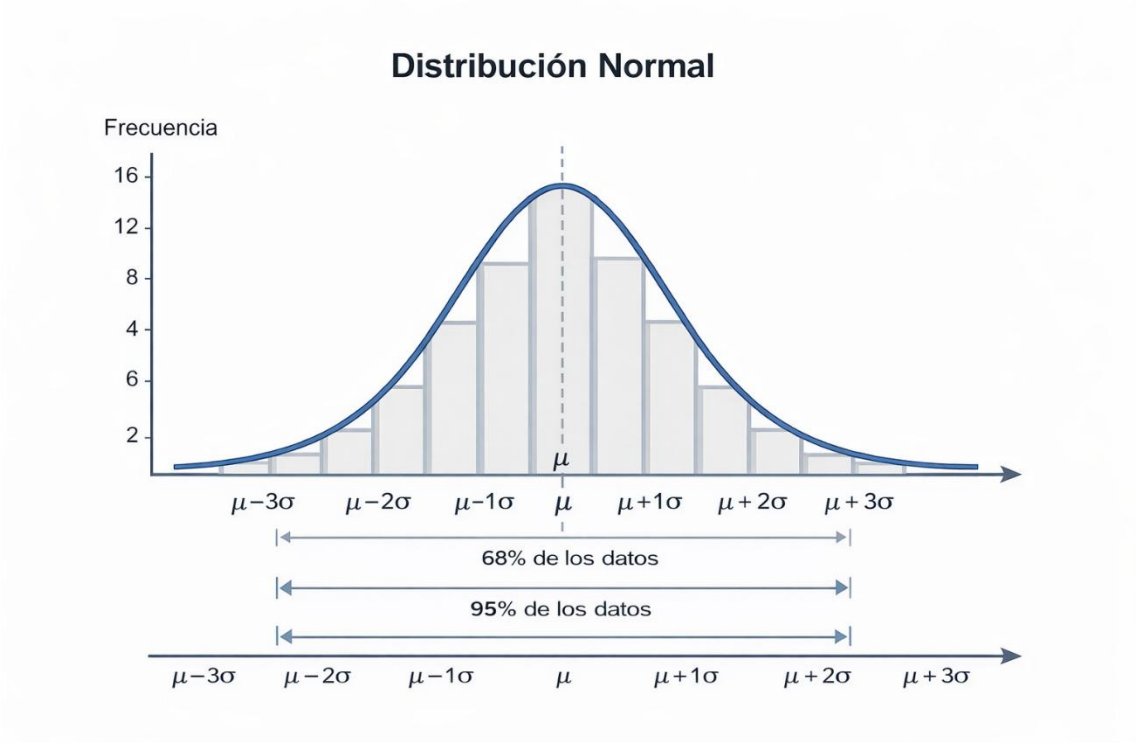
5.1 ¿Qué es la Distribución Normal y Por Qué es Tan Importante?

La distribución normal, también conocida como distribución gaussiana o campana de Gauss, es uno de los conceptos fundamentales de la estadística. Su forma característica de campana, simétrica respecto a la media, describe el comportamiento de innumerables fenómenos naturales y sociales (Altman & Bland, 1995), (Habibzadeh, 2024).

Características de la distribución normal:

1) **Forma acampanada:** Los valores se concentran alrededor de la media y se dispersan simétricamente hacia los extremos.
2) **Simetría:** La media, la mediana y la moda coinciden en el mismo punto.

3) **Asintótica:** Las colas de la distribución se acercan al eje horizontal pero nunca lo tocan.
4) **Áreas conocidas:** Aproximadamente el 68% de los datos se encuentra dentro de una desviación estándar de la media, el 95% dentro de dos desviaciones estándar, y el 99.7% dentro de tres desviaciones estándar (regla empírica).



¿Por qué es tan importante la normalidad?

1) **Base de la estadística paramétrica:** Pruebas como la t de Student, ANOVA, correlación de Pearson y regresión lineal asumen normalidad de los datos o de los residuos.
2) **Teorema Central del Límite:** Establece que, para muestras suficientemente grandes (generalmente n > 30), la distribución de las medias muestrales tiende a una distribución normal, independientemente

de la distribución de la población original. Esto justifica el uso de pruebas paramétricas incluso con datos no normales cuando el tamaño muestral es grande (Lumley et al., 2002).

3) **Interpretación de puntuaciones:** En pruebas estandarizadas (como test de inteligencia), las puntuaciones se interpretan en términos de su posición en la curva normal (percentiles, puntuaciones Z).

5.2 Prueba de Kolmogorov-Smirnov (K-S)

La prueba de Kolmogorov-Smirnov (K-S) es una prueba de bondad de ajuste que compara la distribución acumulada observada de los datos con la distribución acumulada esperada bajo el supuesto de normalidad **Kolmogorov** mencionada en (Stephens, 1992); (N. Smirnov, 1948). Es una de las pruebas más utilizadas, especialmente con muestras grandes.

5.2.1 Características de la Prueba K-S

- **Tipo de variable:** Variables cuantitativas continuas.
- **Tamaño muestral recomendado:** Generalmente se utiliza para muestras grandes ($n > 50$). Algunos autores sugieren que es apropiada para $n > 30$.
- **Sensibilidad:** Es sensible a diferencias en la forma general de la distribución, incluyendo la ubicación y la dispersión.
- **Limitación:** Tiende a ser conservadora (baja potencia) con muestras pequeñas.

5.2.2 Formulación de Hipótesis

- **Hipótesis nula (H_0):** Los datos provienen de una distribución normal.
- **Hipótesis alternativa (H_1):** Los datos no provienen de una distribución normal.

5.2.3 Estadístico de Prueba

El estadístico de Kolmogorov-Smirnov (D) se define como la máxima diferencia absoluta entre la función de distribución acumulada empírica (observada) y la función de distribución acumulada teórica (esperada bajo normalidad):

$$D = \text{máx } |F_o(x) - F_e(x)|$$

Donde:

- $F_o(x)$: Frecuencia acumulada observada
- $F_e(x)$: Frecuencia acumulada esperada bajo distribución normal

5.2.4 Regla de Decisión

Se compara el valor de significancia (p-valor) con el nivel de significación α (generalmente 0.05):

- Si p-valor > α, no se rechaza la hipótesis nula → los datos provienen de una distribución normal.
- Si p-valor < α, se rechaza la hipótesis nula → los datos no provienen de una distribución normal.

5.3 Prueba de Shapiro-Wilk (S-W)

La prueba de Shapiro-Wilk fue desarrollada específicamente para evaluar la normalidad de los datos y es considerada una de las pruebas más potentes, especialmente para muestras pequeñas (Shapiro & Wilk, 1965); (Shapiro et al., 1968).

5.3.1 Características de la Prueba S-W

- **Tipo de variable:** Variables cuantitativas continuas.
- **Tamaño muestral recomendado:** Diseñada originalmente para muestras pequeñas ($n < 50$). Sin embargo, versiones posteriores y la implementación en software estadístico la han extendido hasta $n < 5000$. La mayoría de los textos recomiendan usarla para $n \leq 50$, y Kolmogorov-Smirnov para muestras mayores.
- **Potencia:** Es más potente que la prueba K-S para detectar desviaciones de la normalidad, especialmente en muestras pequeñas (Mohd Razali & Bee Wah, 2011).

5.3.2 Formulación de Hipótesis

- **Hipótesis nula (H_0):** Los datos provienen de una distribución normal.
- **Hipótesis alternativa (H_1):** Los datos no provienen de una distribución normal.

5.3.3 Estadístico de Prueba

El estadístico W de Shapiro-Wilk se calcula como:

$$W = (\sum a_i x_i)^2 / \sum (x_i - \bar{x})^2$$

Donde:

- x_i: Los valores de la muestra ordenados
- a_i: Constantes generadas a partir de las medias, varianzas y covarianzas de los estadísticos de orden de una muestra normal
- $\bar{x}$: Media muestral

5.3.4 Regla de Decisión

Al igual que con K-S, se compara el p-valor con el nivel de significación α:

- Si p-valor > α, no se rechaza H_0 → los datos provienen de una distribución normal.
- Si p-valor < α, se rechaza H_0 → los datos no provienen de una distribución normal.

5.4 Ejemplo Práctico: Prueba de Normalidad con SPSS

Contexto del Ejemplo

Los siguientes datos corresponden al tiempo (en minutos) que han necesitado 30 clientes de un banco para llevar a cabo una transacción bancaria:

20, 42, 62, 32, 28, 20, 38, 26, 30, 18, 42, 54, 16, 32, 42,

36, 18, 56, 41, 16, 14, 42, 34, 14, 24, 51, 49, 24, 18, 56

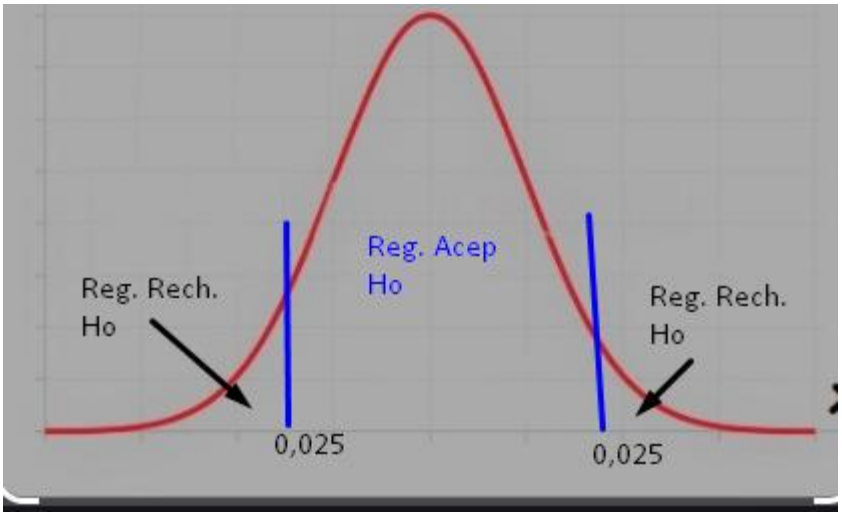
Pregunta de investigación: ¿Los datos se ajustan a una distribución normal?

5.4.1 Formulación de Hipótesis

- **H_0:** La variable "tiempo de transacción" proviene de una distribución normal.
- **H_1:** La variable "tiempo de transacción" no proviene de una distribución normal.

Nivel de significación: $\alpha = 0.05$

Región de aceptación y rechazo de la hipótesis nula.



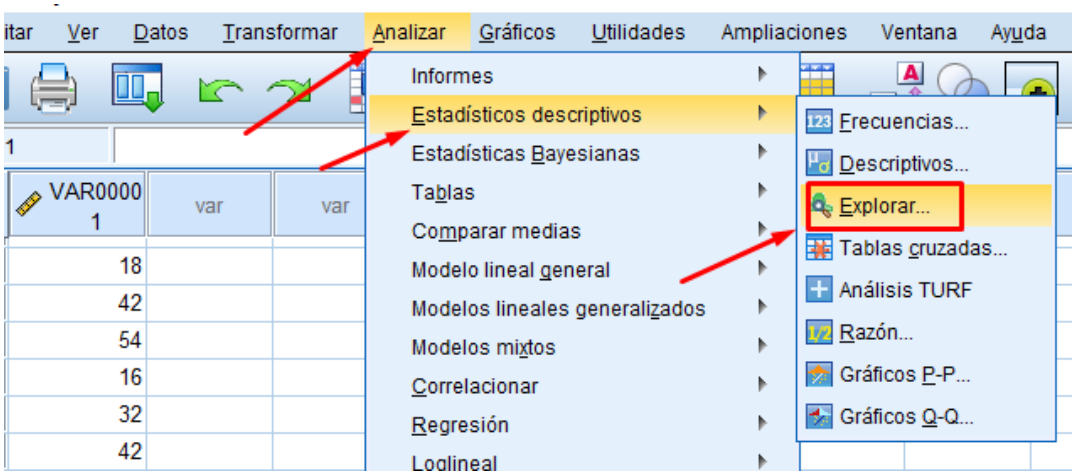
Decisión. Sí el valor de p > α aceptamos la hipótesis nula y sí p < α rechazamos la hipótesis nula.

5.4.2 Procedimiento en SPSS

Paso 1: Ingresar los datos en SPSS. Crear una variable llamada "tiempo" e ingresar los 30 valores.

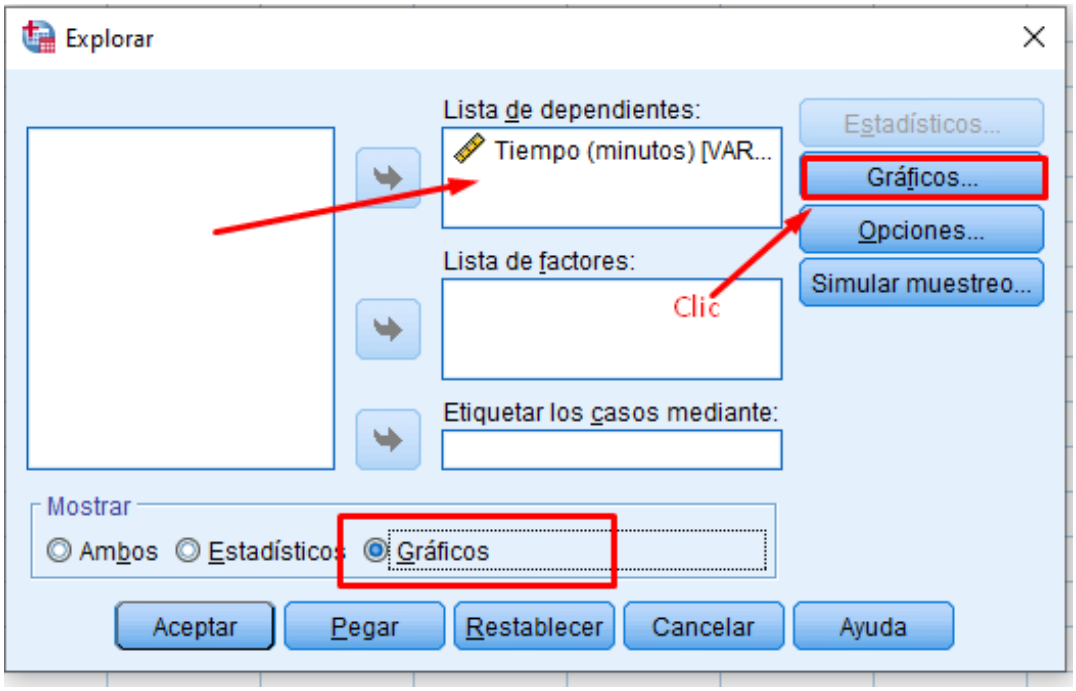
Paso 2: Verificar la normalidad mediante análisis exploratorio:

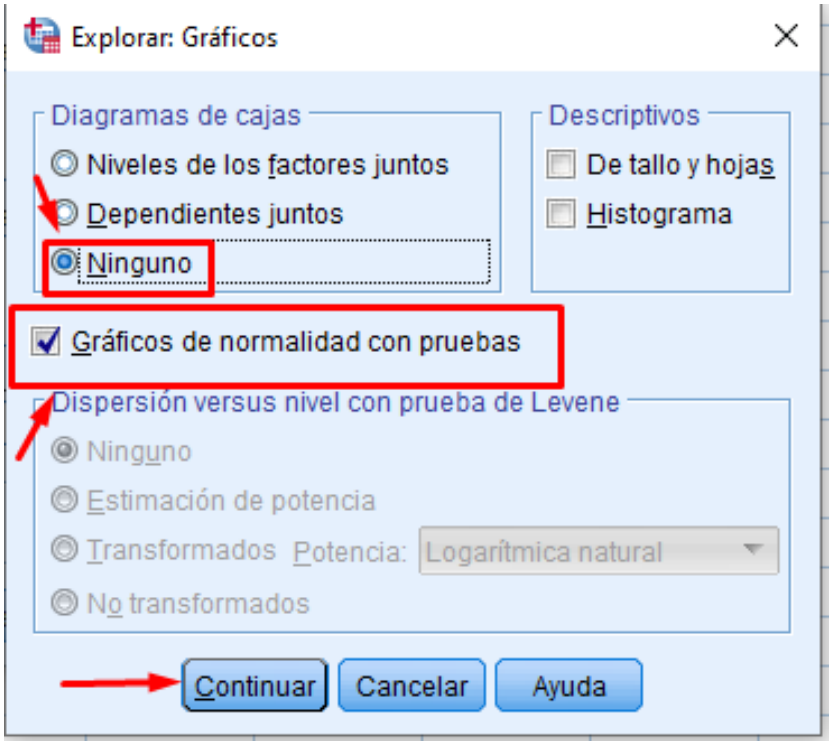
Analizar → Estadísticos descriptivos → Explorar...

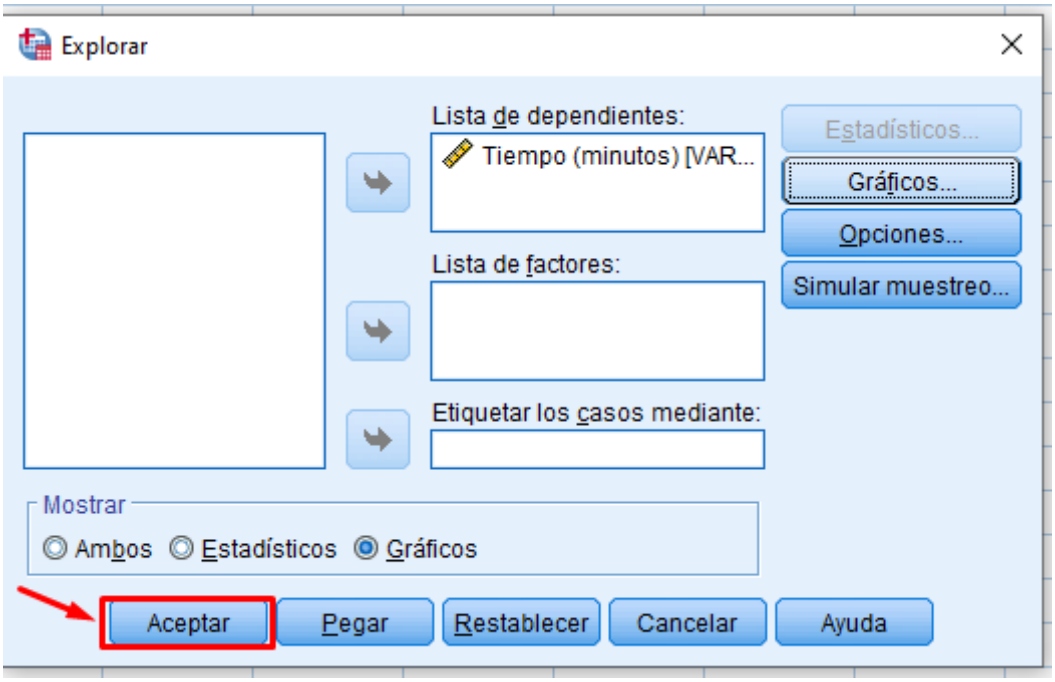


Paso 3: En el cuadro de diálogo:

- Pasar la variable "tiempo" a la lista de "Dependientes".
- Hacer clic en el botón Gráficos.
- Marcar Gráficos con pruebas de normalidad (esto activará tanto Shapiro-Wilk como Kolmogorov-Smirnov).
- También marcar Histograma y Gráficos de normalidad con pruebas (Q-Q plots).
- Hacer clic en Continuar y luego en Aceptar.







Paso 4 (alternativo): También se puede acceder directamente a las pruebas de normalidad mediante:

> Analizar → Pruebas no paramétricas → Cuadros de diálogo antiguos → K-S de una muestra...

Pero el procedimiento "Explorar" es más completo porque ofrece ambas pruebas simultáneamente.

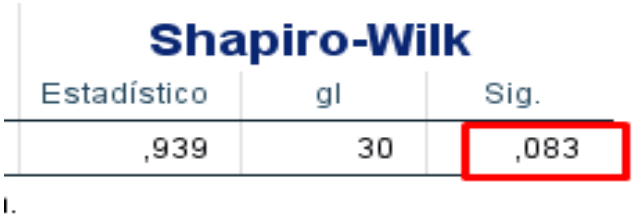
5.4.3 Resultados Obtenidos

Tabla de pruebas de normalidad:

Prueba	Estadístico	gl	p-valor (Sig.)
Kolmogorov-Smirnov[1]	0.112	30	0.200*
Shapiro-Wilk	0.939	30	0.083

[1]Corrección de significación de Lilliefors

*. Esto es un límite inferior de la significación verdadera



El valor estadístico de Shapiro-Wilk es 0,939 y el valor de significancia p = 0.083

Conclusión. Como el valor p = 0,083 > α =0.05 no rechazamos la hipótesis nula en consecuencia la variable tiempo proviene de una distribución normal, por lo tanto, se puede utilizar las pruebas estadísticas paramétricas para su tratamiento.

Histograma con curva normal superpuesta:

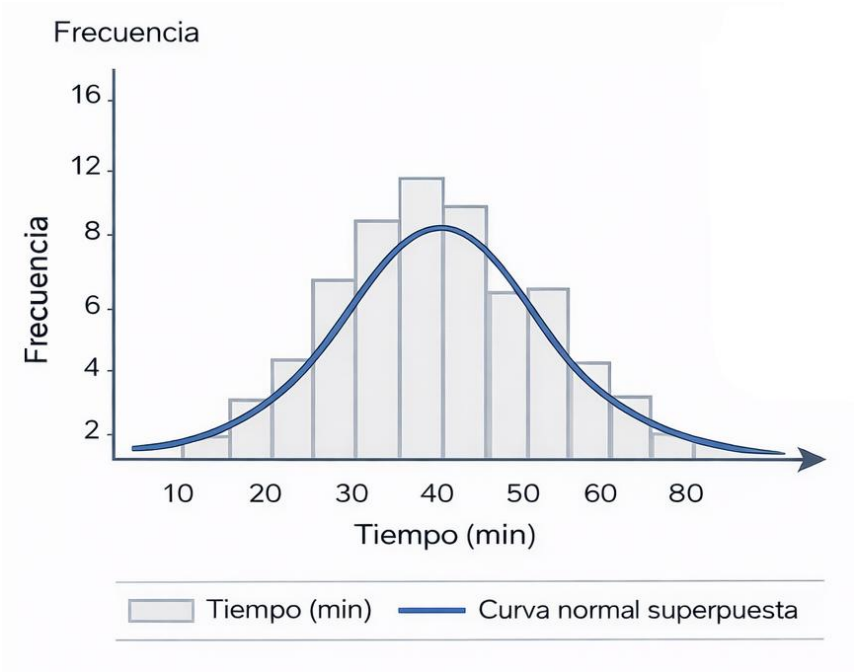
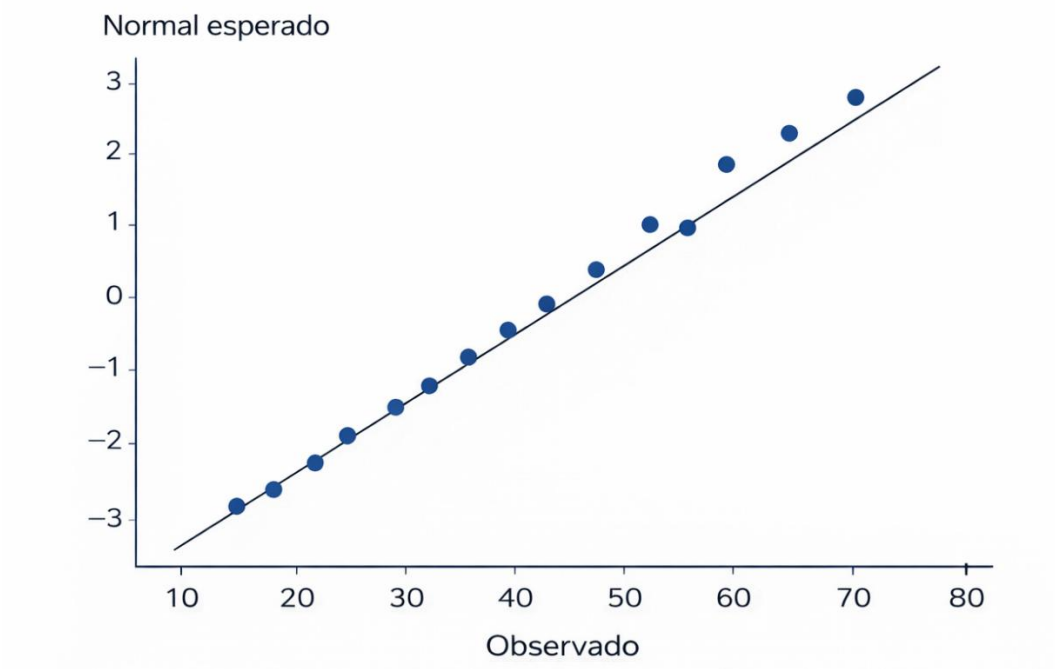


Gráfico Q-Q normal:



5.4.4 Interpretación de Resultados

Interpretación de las pruebas estadísticas:

- **Kolmogorov-Smirnov:** El estadístico es 0.112 con un p-valor de 0.200. Dado que $p = 0.200 > \alpha = 0.05$, no rechazamos la hipótesis nula. Según esta prueba, los datos provienen de una distribución normal.
- **Shapiro-Wilk:** El estadístico es 0.939 con un p-valor de 0.083. Dado que $p = 0.083 > \alpha = 0.05$, tampoco rechazamos la hipótesis nula. Según esta prueba, los datos también provienen de una distribución normal.

Interpretación del gráfico Q-Q:

En el gráfico Q-Q (Quantile-Quantile), los puntos se alinean razonablemente bien a lo largo de la diagonal. Esto indica que los cuantiles observados coinciden con los cuantiles esperados bajo normalidad, lo que respalda visualmente la conclusión de las pruebas estadísticas.

Interpretación del histograma:

El histograma muestra una forma aproximadamente acampanada, aunque con alguna asimetría leve. La curva normal superpuesta se ajusta razonablemente bien a la distribución observada.

5.4.5 Conclusión del Ejemplo

Conclusión: Como el valor p de Shapiro-Wilk = 0.083 > α = 0.05, no rechazamos la hipótesis nula. En consecuencia, la variable "tiempo de transacción" proviene de una distribución normal. Por lo tanto, se pueden utilizar pruebas estadísticas paramétricas (como t de Student, ANOVA, correlación de Pearson) para su tratamiento.

5.5 Comparación entre Pruebas de Normalidad

Característica	Kolmogorov-Smirnov	Shapiro-Wilk	Anderson-Darling
Tipo de prueba	Bondad de ajuste	Bondad de ajuste	Bondad de ajuste
Tamaño muestral óptimo	n > 50	n ≤ 50	n intermedio
Potencia estadística	Baja en muestras pequeñas	Alta en muestras pequeñas	Moderada-alta
Sensibilidad	A toda la distribución	A colas de la distribución	A colas de la distribución
Implementación en SPSS	Sí	Sí	Sí (mediante complementos)

Recomendaciones prácticas:

1) **Para muestras pequeñas (n ≤ 50):** Utilizar preferentemente la prueba de Shapiro-Wilk.

2) **Para muestras grandes (n > 50):** Utilizar la prueba de Kolmogorov-Smirnov (con corrección de Lilliefors).
3) **Siempre complementar con análisis gráfico:** Histograma, Q-Q plot, y boxplot proporcionan información visual valiosa que complementa las pruebas estadísticas.
4) **Considerar el Teorema Central del Límite:** Con muestras grandes (n > 30), incluso si los datos no son normales, las medias muestrales se distribuirán aproximadamente normales, lo que permite usar pruebas paramétricas con cierta robustez.

5.6 ¿Qué Hacer si los Datos No son Normales?

Encontrar que nuestros datos no siguen una distribución normal no es el fin del mundo. Existen varias alternativas:

5.6.1 Transformaciones de Datos

A veces, una simple transformación matemática puede "normalizar" los datos. Las transformaciones más comunes son:

Tipo de datos	Transformación sugerida	Fórmula
Sesgo positivo (cola a la derecha)	Logarítmica	$y' = \log(y)$
Sesgo positivo (moderado)	Raíz cuadrada	$y' = \sqrt{y}$
Datos de conteo	Raíz cuadrada	$y' = \sqrt{(y + 0.5)}$
Proporciones o porcentajes	Arcoseno	$y' = \text{arcsen}(\sqrt{y})$

Tipo de datos	Transformación sugerida	Fórmula
Sesgo negativo (cola a la izquierda)	Cuadrática o potencia	$y' = y^2$

Ejemplo de transformación logarítmica:

Datos originales (sesgados a la derecha):

2, 3, 5, 7, 12, 18, 25, 40, 80, 150

Transformación logarítmica (base 10):

0.30, 0.48, 0.70, 0.85, 1.08, 1.26, 1.40, 1.60, 1.90, 2.18

Después de la transformación, se debe volver a evaluar la normalidad.

5.6.2 Pruebas No Paramétricas

Si los datos no son normales y no pueden ser transformados adecuadamente, se deben utilizar pruebas no paramétricas (también llamadas pruebas de distribución libre). Estas pruebas no asumen normalidad y son válidas con cualquier distribución.

Situación	Prueba paramétrica	Alternativa no paramétrica
Comparar dos grupos independientes	t de Student para muestras independientes	U de Mann-Whitney
Comparar dos grupos relacionados	t de Student para muestras pareadas	Wilcoxon

Situación	Prueba paramétrica	Alternativa no paramétrica
Comparar tres o más grupos independientes	ANOVA de un factor	Kruskal-Wallis
Comparar tres o más grupos relacionados	ANOVA de medidas repetidas	Friedman
Correlación entre dos variables	Correlación de Pearson	Correlación de Spearman

5.6.3 Robustez de las Pruebas Paramétricas

Es importante señalar que muchas pruebas paramétricas son robustas a violaciones moderadas de la normalidad, especialmente cuando:

- Los tamaños muestrales son grandes (n > 30 por grupo).
- Las muestras son aproximadamente iguales en tamaño.
- La violación de normalidad no es extrema (no hay valores atípicos muy extremos).

En estos casos, los resultados de las pruebas paramétricas suelen ser confiables incluso con datos no normales (Norman, 2010).

5.7 Ejemplo Adicional: Datos No Normales

Para ilustrar qué sucede cuando los datos no son normales, consideremos el siguiente conjunto de datos (tiempos de espera en una cola, que típicamente tienen distribución exponencial):

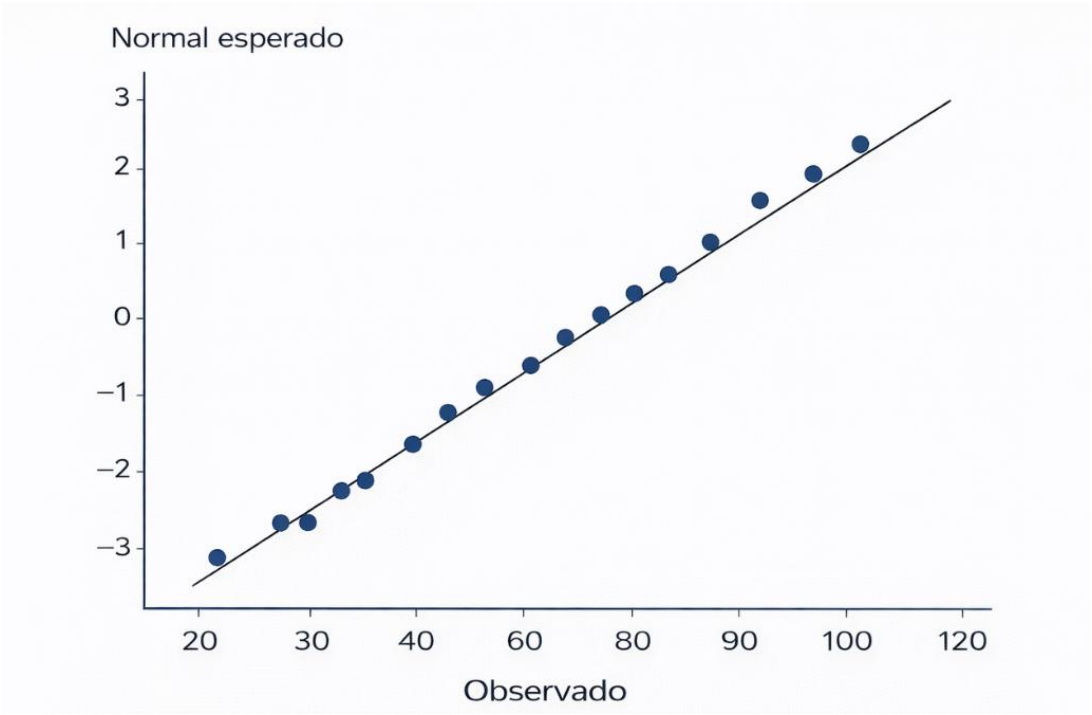
2, 3, 1, 15, 2, 1, 25, 3, 2, 40, 1, 2, 60, 3, 2, 80, 1, 4, 120, 2

Resultados de la prueba de normalidad en SPSS:

Prueba	Estadístico	gl	p-valor
Shapiro-Wilk	0.672	20	0.000

Interpretación: $p = 0.000 < 0.05$, por lo tanto rechazamos H_0. Los datos no provienen de una distribución normal.

Gráfico Q-Q normal:



En este caso, los puntos se desvían claramente de la diagonal, confirmando la falta de normalidad. Para analizar estos datos, deberíamos usar una

transformación (por ejemplo, logarítmica) o recurrir a pruebas no paramétricas.

5.8 La Prueba de Normalidad en el Contexto de la Investigación

5.8.1 ¿Siempre es Necesario Probar la Normalidad?

Sí, siempre que se vayan a utilizar pruebas paramétricas. Los supuestos de las pruebas estadísticas no son meros formalismos académicos; violarlos puede llevar a:

1) **Error Tipo I inflado:** Rechazar la hipótesis nula cuando es verdadera (falso positivo) con más frecuencia de la declarada.
2) **Error Tipo II aumentado:** No rechazar la hipótesis nula cuando es falsa (falso negativo) debido a pérdida de potencia.
3) **Intervalos de confianza incorrectos:** Las estimaciones de precisión pueden ser engañosas.

5.8.2 Reportando Resultados de Normalidad en Artículos Científicos

En un artículo científico, es común reportar los resultados de la prueba de normalidad de la siguiente manera:

"Se evaluó la normalidad de las variables cuantitativas mediante la prueba de Shapiro-Wilk. Los resultados indicaron que todas las variables presentaron una distribución normal ($p > 0.05$), por lo que se utilizaron pruebas paramétricas para los análisis inferenciales."

O, si los datos no son normales:

"La prueba de Kolmogorov-Smirnov reveló que la variable 'tiempo de espera' no se ajustaba a una distribución normal (p = 0.003). Por lo tanto, se utilizó la prueba no paramétrica de U de Mann-Whitney para comparar los grupos."

5.8.3 Tabla Resumen para la Toma de Decisiones

Situación	Prueba de normalidad	Decisión
$n \leq 50$, $p > 0.05$	Shapiro-Wilk no significativo	Datos normales → usar pruebas paramétricas
$n \leq 50$, $p < 0.05$	Shapiro-Wilk significativo	Datos no normales → transformar o usar no paramétricas
$n > 50$, $p > 0.05$	Kolmogorov-Smirnov no significativo	Datos normales → usar pruebas paramétricas
$n > 50$, $p < 0.05$	Kolmogorov-Smirnov significativo	Datos no normales → considerar TCL o usar no paramétricas

5.9 Limitaciones y Consideraciones Adicionales

5.9.1 Tamaño Muestral y Potencia

Con muestras muy pequeñas ($n < 10$), las pruebas de normalidad tienen poca potencia: es difícil detectar desviaciones de la normalidad incluso cuando existen. Con muestras muy grandes ($n > 300$), las pruebas pueden ser

demasiado sensibles, detectando diferencias triviales que no afectan prácticamente los resultados de las pruebas paramétricas (Field, 2024).

En muestras muy grandes, es recomendable complementar las pruebas estadísticas con inspección visual de los gráficos y considerar la robustez de las pruebas paramétricas.

5.9.2 Valores Atípicos (Outliers)

La presencia de valores atípicos puede afectar significativamente las pruebas de normalidad. Antes de evaluar normalidad, es recomendable identificar y decidir qué hacer con los valores atípicos:

- ¿Son errores de medición? → Corregir o eliminar.
- ¿Son casos genuinos pero extremos? → Considerar transformaciones o pruebas robustas.
- ¿Representan una subpoblación diferente? → Analizar por separado.

5.9.3 Normalidad Multivariada

Cuando trabajamos con varias variables simultáneamente (por ejemplo, en MANOVA o regresión múltiple), el supuesto no es solo que cada variable sea normal, sino que existe normalidad multivariada. Esto es más complejo de evaluar y existen pruebas específicas (como la prueba de Mardia) que están fuera del alcance de este libro introductorio.

Resumen del Capítulo

En este capítulo hemos explorado la prueba de normalidad, un paso fundamental antes de decidir qué tipo de análisis estadístico realizar. Recordemos las ideas principales:

1) **Distribución normal:** Es el supuesto base de la estadística paramétrica. Tiene forma de campana, es simétrica y tiene propiedades probabilísticas conocidas.
2) **Pruebas de normalidad:**
   - Kolmogorov-Smirnov: Adecuada para muestras grandes ($n > 50$).
   - Shapiro-Wilk: Más potente para muestras pequeñas ($n \leq 50$).
   - Ambas pruebas tienen como hipótesis nula que los datos provienen de una distribución normal.
3) **Interpretación:** Si $p > 0.05$, no rechazamos H_0 → datos normales. Si $p < 0.05$, rechazamos H_0 → datos no normales.
4) **Complementos visuales:** Histogramas, gráficos Q-Q y boxplots proporcionan información valiosa que complementa las pruebas estadísticas.
5) Qué hacer si los datos no son normales:
   - Aplicar transformaciones (logarítmica, raíz cuadrada, etc.).
   - Utilizar pruebas no paramétricas (U de Mann-Whitney, Wilcoxon, Kruskal-Wallis, etc.).
   - Considerar la robustez de las pruebas paramétricas con muestras grandes.
6) Reporte en investigaciones: Incluir los resultados de las pruebas de normalidad en la sección de metodología o resultados preliminares justifica la elección de las pruebas estadísticas utilizadas posteriormente.

La prueba de normalidad no es un fin en sí misma, sino un medio para garantizar que nuestros análisis estadísticos sean apropiados y nuestras conclusiones, válidas. Como hemos visto a lo largo de este libro, la estadística no es solo una colección de fórmulas, sino un conjunto de herramientas para tomar decisiones informadas basadas en datos. Respetar sus supuestos es parte fundamental de ese proceso.

EPÍLOGO: Y ahora, ¿qué hacemos con todo esto?

Llegamos al final del camino. O mejor dicho, llegamos a un alto en el camino, porque esto no termina aquí. A lo largo de estas páginas hemos ido desgranando, paso a paso, los secretos de ese oficio tan particular que consiste en convertir la realidad en números y los números en comprensión. Arrancamos desde lo más básico ¿qué es eso de la población? ¿y la muestra? y nos fuimos adentrando poco a poco en terrenos más complejos, hasta toparnos con esos conceptos que parecen sacados de un manual de magia: la validez, la confiabilidad, la normalidad. Pero todo este viaje, todo este esfuerzo, tendría poco sentido si se quedara en una colección de definiciones para lucir en conversaciones de café. El verdadero propósito es otro: tomar mejores decisiones. Con los pies en la tierra y los ojos bien abiertos.

Lo que aprendimos (y no deberíamos olvidar)

Cada capítulo de este libro ha sido como una herramienta nueva que hemos ido guardando en la caja. Ahora que la cerramos, conviene repasar lo que hay dentro.

En el **Capítulo I** pusimos los cimientos. Aprendimos que detrás de cualquier estudio hay una población, ese todo que nos interesa, y una muestra, ese trocito que podemos abarcar. Descubrimos que la clave no está en preguntarle a todo el mundo, sino en saber elegir a quién preguntamos. Y nos llevamos una lección que conviene no olvidar: los investigadores rara vez podemos conocer la verdad absoluta, ese parámetro escondido, pero

podemos acercarnos a ella con nuestras estadísticas, siempre que sepamos hasta dónde llega nuestro conocimiento y dónde empieza nuestra ignorancia. La humildad, en esto, es una virtud.

El **Capítulo II** fue el de la toma de decisiones. Allí nos metimos de lleno en el arte de elegir muestras. Vimos que hay dos mundos: el muestreo no probabilístico, rápido y útil para según qué cosas, y el probabilístico, que es el que nos permite ponerle números al error y hablar con propiedad. Aprendimos a calcular tamaños, a sortear personas, a dividir por estratos, a agrupar por conglomerados. Y nos quedó grabado a fuego un principio que parece sencillo pero que mucha gente ignora: una muestra no vale por lo grande que sea, sino por lo bien que represente al conjunto. Una muestra enorme pero sesgada es peor que una pequeña pero bien traída. El tamaño no lo es todo.

En el **Capítulo III** nos enfrentamos al mayor de los desafíos: medir lo que no se ve. Porque la inteligencia no se toca, la satisfacción no pesa, la calidad de vida no tiene color. Y sin embargo, hay que ponerles número. Ahí entró en juego eso de la operacionalización, los niveles de medición, las escalas, especialmente esa reina que es la escala Likert. Aprendimos que un buen instrumento no sale de la ocurrencia del momento, sino de un proceso: definir, preguntar, probar, corregir, volver a probar. Y que los expertos y las pruebas piloto no son un lujo, son una necesidad.

El **Capítulo IV** fue, si me apuras, el corazón de todo. Allí nos hicimos las preguntas incómodas: ¿esto que medimos es lo que creemos que es? (validez). ¿Y si lo medimos mañana, dará lo mismo? (confiabilidad). Descubrimos que la validez tiene caras: de contenido, de constructo, de criterio. Aprendimos a manejar la V de Aiken, a asomarnos sin miedo al

análisis factorial, a calcular la Kappa de Fleiss. Y en el mundo de la confiabilidad, nos hicimos amigos del test-retest, del método de mitades partidas y, cómo no, del famoso Alfa de Cronbach. Ese número que ahora sabemos leer con otros ojos. Y nos quedó clara una cosa: un instrumento puede ser confiable y no válido puede medir siempre lo mismo, pero lo mismo puede ser una tontería, pero si no es confiable, de válido no tiene nada.

Finalmente, el **Capítulo V** nos puso los pies en la tierra. Porque antes de lanzarnos a hacer pruebas estadísticas como quien lanza cohetes, hay que comprobar si los datos se portan bien. Si siguen esa famosa curva normal o si, por el contrario, van cada uno por su lado. Y aprendimos a usar a los guardianes de esa frontera: el Kolmogorov-Smirnov para cuando tenemos muchos datos, el Shapiro-Wilk para cuando son pocos. Ellos nos dicen si podemos usar las herramientas paramétricas, las potentes, o si tenemos que bajar un cambio y recurrir a sus primas no paramétricas.

Y ahora, con todo esto en la mochila, toca salir al mundo. Porque los datos están ahí fuera, esperando. En las noticias, en los informes, en las conversaciones, en los trabajos. Y ahora, cuando los veas, los mirarás de otra manera. Con cariño, con respeto, pero también con ese punto justo de escepticismo que separa al que traga con todo del que realmente entiende. Ese es el verdadero regalo de este viaje.

El Ciclo de la Investigación: Un Proceso, No un Destino

Una de las lecciones más importantes que podemos extraer de todo lo aprendido es que la investigación no es un proceso lineal que termina cuando

obtenemos un resultado. Es, más bien, un ciclo continuo que se retroalimenta a sí mismo:

Y todo esto, ¿para qué sirve en la vida real?

A estas alturas, es posible que haya un pensamiento rondándote la cabeza. Algo así como: "bueno, todo esto está muy bien para señores con bata y papers académicos, pero yo, en mi día a día, ¿qué hago con semejante tinglado?". Pues mira, te sorprenderías. Porque la estadística, cuando la entiendes de verdad, deja de ser una asignatura para convertirse en una especie de superpoder cotidiano. Te pongo ejemplos.

En la calle, como ciudadano de a pie

Vivimos rodeados de encuestas. Cada vez que hay elecciones, cada vez que alguien quiere venderte algo, cada vez que sale un estudio sobre si el desayuno es bueno o malo, hay números de por medio. Pues ahora tú ya no eres el que asiente con la cabeza y se lo cree todo. Ahora eres ese amigo un poco pesado que se pregunta: ¿y a quién preguntaron? ¿Era una muestra decente o preguntaron en la puerta de su oficina? ¿Cuánta gente fue? ¿Las preguntas estaban bien hechas o llevaban trampa? Eso, que parece una manía, es simplemente pensar con cabeza propia. Y en un mundo donde intentan colarte gato por liebre a diario, eso vale oro.

En el trabajo, seas lo que seas

Da igual que te dediques a la salud, a la enseñanza, al marketing, a los recursos humanos o a vender seguros. En todos lados, cada vez más, las decisiones se toman mirando números. Que si los pacientes están satisfechos, que si los alumnos mejoran, que si tal campaña funciona. Saber si esos

números son de fiar, si te están contando la verdad o solo lo que quieres oír, te da una ventaja enorme. Y si encima eres capaz de generar tus propios datos, con criterio, con método, entonces ya no eres uno más. Eres el que sabe.

En la tienda, cuando compras

Porque esos estudios de mercado que dicen que "el 80% de los usuarios prefiere este detergente", ¿tú te los crees a pies juntillas? ¿O ahora te preguntas de dónde sacaron ese 80%, a quién preguntaron, si la muestra era representativa o eran todo empleados de la fábrica? Leer las reseñas de un producto, las opiniones de otros compradores, las estrellitas esas de puntuación... todo eso es estadística aplicada. Y cuanto más sepas de verdad, menos te la cuelan.

Y en lo más hondo, como persona

Vivimos en una época donde la información y la basura se parecen mucho. Donde cualquiera con un gráfico y una mala intención puede hacerte creer lo que quiera. Aprender a leer los datos con ojo crítico no es solo una habilidad profesional. Es una defensa. Una trinchera contra los que quieren manipularte, contra los titulares falsos, contra las verdades a medias. Es, en el fondo, una manera de ser un poco más libres. Y eso, créeme, no tiene precio.

Limitaciones y Advertencias

Ningún texto puede cubrirlo todo, y este no es la excepción. Es importante reconocer las limitaciones de lo que hemos aprendido:

1) **Este es un punto de partida, no un punto de llegada.** La estadística es un campo vasto y en constante evolución. Hay técnicas más avanzadas (modelos de ecuaciones estructurales, teoría de respuesta al ítem, análisis multinivel, etc.) que no hemos abordado y que pueden ser necesarias en investigaciones más complejas.
2) **La práctica es insustituible.** Leer sobre estadística es necesario, pero no suficiente. La verdadera comprensión viene de la práctica: de cargar datos en SPSS, de interpretar resultados, de cometer errores y aprender de ellos. Te animamos a que tomes los ejemplos de este libro, los reproduzcas y, sobre todo, los adaptes a tus propios datos.
3) **El contexto importa.** Las reglas y umbrales que hemos presentado (por ejemplo, "un Alfa de Cronbach superior a 0.70 es aceptable") son guías generales, no leyes inquebrantables. En algunos contextos (investigación exploratoria) pueden aceptarse valores más bajos; en otros (evaluaciones de alto riesgo) se exigen valores más altos. El juicio del investigador, informado por la teoría y el contexto, sigue siendo insustituible.
4) **La ética es fundamental.** Tener las herramientas para manipular datos conlleva la responsabilidad de usarlas éticamente. No se trata de encontrar los resultados que queremos, sino de descubrir la verdad, aunque sea incómoda. La transparencia metodológica (reportar cómo se obtuvieron los datos, qué decisiones se tomaron, qué limitaciones tiene el estudio) es un imperativo ético.

Un Llamado a la Acción: Conviértete en un Consumidor Crítico de Datos

Si este manuscrito ha cumplido su propósito, al cerrar estas páginas no te llevarás únicamente un conjunto de conceptos y fórmulas. Te llevarás, esperamos, una actitud: la actitud de quien no acepta los datos pasivamente, sino que los interpela, los cuestiona, los pone a prueba.

Te invitamos a que, a partir de ahora, cuando escuches una estadística en las noticias, te preguntes: ¿de dónde viene esta cifra? ¿cómo se obtuvo la muestra? ¿el instrumento utilizado era válido y confiable? ¿reportaron el margen de error?

Te invitamos a que, cuando leas un artículo científico, no te detengas en las conclusiones, sino que examines la metodología: ¿qué tipo de muestreo utilizaron? ¿cómo operacionalizaron sus variables? ¿reportaron la confiabilidad del instrumento? ¿verificaron los supuestos estadísticos?

Te invitamos a que, en tu propia práctica profesional o académica, apliques estos principios: que te tomes el tiempo necesario para diseñar bien tu muestra, para construir cuidadosamente tus instrumentos, para evaluar su validez y confiabilidad, para verificar los supuestos antes de aplicar las pruebas estadísticas.

Porque al final del día, los datos no hablan por sí mismos. Somos nosotros quienes les damos voz a través de nuestras decisiones metodológicas. Y la calidad de esas decisiones determina la calidad de nuestras conclusiones y, en última instancia, la calidad de las decisiones que tomamos basadas en esas conclusiones.

Palabras Finales

Quienes hemos escrito estas páginas quisiéramos contarte algo, ya que estamos en confianza. Este libro no ha sido un encargo frío ni un ejercicio académico más. Ha sido, para nosotros, un viaje de redescubrimiento. Mientras explicábamos cada concepto, mientras buscábamos ejemplos que fueran claros sin ser tontos, nos hemos ido topando una y otra vez con la misma pregunta: ¿y por qué demonios nos gusta tanto esto? La respuesta, al final, siempre era la misma.

La estadística nos gusta porque es una manera de asomarnos al mundo con otros ojos. Nos permite encontrarle sentido al caos, ver dibujos donde otros solo ven manchas, hacernos preguntas más inteligentes donde otros se quedan con la primera ocurrencia. Y en un mundo que cada día parece más revuelto, más lleno de ruido y de voces que gritan, tener una brújula así no es un lujo: es casi un salvavidas.

Pero si hay algo que de verdad nos llevamos de este viaje, y que hemos querido contagiarte, es esto: la estadística no es un bicho raro. No es ese monstruo que aparece en las pesadillas de quienes odiaron las matemáticas en el colegio. Es, simplemente, una herramienta. Y como todas las herramientas, está hecha para manos humanas. Para las tuyas. Porque detrás de cada porcentaje, detrás de cada gráfico de barras, detrás de cada prueba de esas que parecen trabalenguas, lo que hay siempre es una pregunta de persona. Alguien que quiere saber algo. Alguien que necesita entender algo. Alguien que busca, a su manera, un poco de luz entre tanta niebla.

Los datos, si lo piensas bien, no son más que el eco de esas preguntas. Y aprender a escuchar ese eco con atención, con cuidado, con esa mezcla justa de confianza y escepticismo es lo que nos convierte en algo más que simples recolectores de números. Nos convierte en buscadores de comprensión.

Así que gracias. Gracias por habernos acompañado hasta aquí. Gracias por haber tenido la curiosidad de asomarte a este mundo con nosotros. Ahora te toca a ti seguir caminando. Y ojalá que los datos que te encuentres por el camino te lleven a preguntas cada vez más interesantes, a respuestas cada vez más sólidas, y a decisiones cada vez más tuyas.

Porque al final, de eso va todo esto. No de acumular números por acumular. No de llenar informes que nadie lee. Sino de usar los datos como lo que son: una herramienta para vivir con un poco más de criterio, para decidir con un poco más de fundamento, para entender el mundo y entenderte con un poco más de claridad**.**

BIBLIOGRAFÍA

Aiken, L. R. (1980). Content Validity and Reliability of Single Items or Questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Altman, D. G., & Bland, j. M. (1995). Statistics notes: The normal distribution. *BMJ*, *310*(6975), 298. https://doi.org/10.1136/bmj.310.6975.298

American Educational Research Association [AERA]. (2024, June). *The Standards for Educational and Psychological Testing*. https://www.apa.org/science/programs/testing/standards

American Psychological Association [APA]. (2020). *Publication Manual of the American Psychological Association, Seventh Edition (2020)*. https://apastyle.apa.org/products/publication-manual-7th-edition

Anastasi, A., & Urbina, S. (1998). *Tests psicológicos*. https://books.google.com.pe/books?hl=es&lr=&id=FV01zgFuk0cC&oi=fnd&pg=PR13&dq=Anastasi,+A.,+%26+Urbina,+S.+(1998).+Tests+psicol%C3%B3gicos+(7a+ed.).+Prentice+Hall.&ots=RC2RT6Dj7J&sig=gn2Q1iiQZ_BuM3BqwnKPSSwFhL0&redir_esc=y#v=onepage&q=Anastasi%2C%20A.%2C%20%26%20Urbina%2C%20S.%20(1998).%20Tests%20psicol%C3%B3gicos%20(7a%20ed.).%20Prentice%20Hall.&f=false

Andrés Gutiérrez. (2016). Estrategias de muestro. Diseño de encuestas y estimación de parámetros. *Ediciones de La U*. https://books.google.com/books/about/Estrategias_de_muestreo.html?hl=es&id=zzOjDwAAQBAJ

Arnab, R. (2017). Survey Sampling Theory and Applications. *Survey Sampling Theory and Applications*, 1–899.

https://books.google.com/books/about/Survey_Sampling_Theory_and_Applications.html?hl=es&id=8_IbDQAAQBAJ

Casella, G., & Berger, R. L. (2024). Statistical Inference, Second Edition. *Statistical Inference, Second Edition*, 1–535. https://doi.org/10.1201/9781003456285

Ciro Martínez. (2012). *Estadistica y muestreo - 13ra edición*. https://books.google.com.pe/books/about/Estadistica_y_muestreo_13ra_edici%C3%B3n.html?id=mfVeDwAAQBAJ&redir_esc=y

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika 1951 16:3*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. (2013). *Five Perspectives on Validity Argument*. 3–17. https://doi.org/10.4324/9780203056905-2

Cronbach, L., & Meehl, P. (1998). *Construct validity in psychological tests*. https://books.google.com/books/about/Personality.html?hl=es&id=fiXyT5MkCLMC

DeVellis, R. F. ., & Thorpe, C. T. . (2022). *Scale development : theory and applications*. 298. https://books.google.com/books/about/Scale_Development.html?hl=es&id=QddDEAAAQBAJ

Escobar-Pérez, J., & Cuervo-Martínez, Á. (2008). *Validez de contenido y juicio de expertos: una aproximación a su utilización*. *6*, 27–36.

Field, A. P. . (2024). *Discovering statistics using IBM SPSS statistics Andy Field*. https://books.google.com/books/about/Discovering_Statistics_Using_IBM_SPSS_St.html?hl=es&id=83L2EAAAQBAJ

Finch, W. Holmes., Hidalgo Montesinos, Dolores., French, B. F. ., & Immekus, J. C. . (2022). *Psicometría aplicada usando SPSS y AMOS*. https://books.google.com/books/about/Applied_Psychometrics_Using_SPSS_and_AMO.html?hl=es&id=4r2CEAAAQBAJ

Fitzner, K. (2007). Reliability and validity: A quick review. *Diabetes Educator*, *33*(5), 775–780. https://doi.org/10.1177/0145721707308172

Furr, R. Michael. (2022). *Psychometrics : an introduction*. 12. https://books.google.com/books/about/Psychometrics.html?hl=es&id=xto9EAAAQBAJ

García-García, J., Ángeles Gil, M., & Asunción Lubiano, M. (2024). On some properties of Cronbach's α coefficient for interval-valued data in questionnaires. *Advances in Data Analysis and Classification 2024 19:3*, *19*(3), 831–854. https://doi.org/10.1007/s11634-024-00601-w

Guttman, L. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review*, *9*(2), 139. https://doi.org/10.2307/2086306

Habibzadeh, F. (2024). Data Distribution: Normal or Abnormal? *Journal of Korean Medical Science*, *39*(3). https://doi.org/10.3346/jkms.2024.39.e35

Hair, J. F., Babin, B. J., Black, W. C., & Anderson, R. E. (2019). Multivariate data analysis. In Pearson Prentice (Ed.), *Pearson Prentice*. Cengage. https://researchdiscovery.drexel.edu/esploro/outputs/book/Multivariate-data-analysis/991019295303204721

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*. https://biblioteca.ucuenca.edu.ec/digital/s/biblioteca-digital/ark:/25654/2140#?c=0&m=0&s=0&cv=0

Israel, Mark. (2015). *Research ethics and integrity for social scientists : beyond regulatory compliance*. 247. https://books.google.com/books/about/Research_Ethics_and_Integrity_for_Social.html?hl=es&id=ZvSICwAAQBAJ

Klinger Angarita, R. (2024). *Muestreo estadístico: métodos básicos*. https://books.google.com.pe/books?hl=es&lr=&id=LCEdEQAAQBAJ&oi=fnd&pg=PA15&dq=Muestreo+estad%C3%ADstico:+Dise%C3%B1o+y+aplicaciones&ots=aCUY8muyPg&sig=0UnlAV5ab7jJlkg9_fP

dj_oMGb8&redir_esc=y#v=onepage&q=Muestreo%20estad%C3%ADstico%3A%20Dise%C3%B1o%20y%20aplicaciones&f=false

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lohr, S. L. (2022). *Sampling Design and Analysis Third Edition*. https://www.crcpress.com/

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(Volume 23, 2002), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Merino Soto, C., & Livia Segovia, J. (2009). *Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken* (Vol. 25). Fidler. http://revistas.um.es/analesps

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*.

Mode, E. B. ., & García Garza, R. . (2021). *Elementos de probabilidad y estadística*. https://books.google.com/books/about/Elementos_de_probabilidad_y_estad%C3%ADstica.html?hl=es&id=uuAbEAAAQBAJ

Mohd Razali, N., & Bee Wah, Y. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 13–14.

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: Segunda edición. In *Psicothema* (Vol. 25, Number 2, pp. 151–157). https://doi.org/10.7334/psicothema2013.24

N. Smirnov. (1948, June). *Table for Estimating the Goodness of Fit of Empirical Distributions on JSTOR*. https://www.jstor.org/stable/2236278

Narayan, K. G., & Sinha, D. K. (2023). Sampling Techniques. *Veterinary Public Health & Epidemiology*, 111–123. https://doi.org/10.1007/978-981-19-7800-5_12

National Council on Measurement in Education. (2018). *National Council on Measurement in Education on JSTOR*. https://www.jstor.org/publisher/ncme

Nencini, P. A. (2022). *Estadística básica para economía y administración*. http://ridaa.unq.edu.ar/handle/20.500.11807/4085

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education 2010 15:5*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Osgood, C. Egerton., Suci, G. J. ., & Tannenbaum, P. H. . (1978). *The measurement of meaning*. 346. https://books.google.com/books/about/The_Measurement_of_Meaning.html?hl=es&id=Qj8GeUrKZdAC

Perloff, R. (1997). Daniel Goleman's Emotional intelligence: Why it can matter more than IQ. *The Psychologist-Manager Journal*, *1*(1), 21–22. https://doi.org/10.1037/h0095822

Rensis Likert. (2017). The method of constructing an attitude scale. *Scaling: A Sourcebook for Behavioral Scientists*, 233–243. https://doi.org/10.4324/9781315128948-23

Roldán, P. L., & Oliva, S. F. (2015). *Metodología de la investigación social cuantitativa*. https://portalrecerca.uab.cat/en/publications/metodolog%C3%ADa-de-la-investigaci%C3%B3n-social-cuantitativa/

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shapiro, S. S., Wilk, M. B., & Chen, H. J. (1968). A Comparative Study of Various Tests for Normality. *Journal of the American Statistical Association*, *63*(324), 1343–1372. https://doi.org/10.1080/01621459.1968.10480932

Sheskin, David. (2020). *Handbook of parametric and nonparametric statistical procedures David J. Sheskin*. https://books.google.com/books/about/Handbook_of_Parametric_and_Nonparametric.html?hl=es&id=nvDqDwAAQBAJ

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*(1–3), 83–117. https://doi.org/10.1023/a:1006985528729

Spoto, A., Nucci, M., Prunetti, E., & Vicovaro, M. (2023). Improving Content Validity Evaluation of Assessment Instruments Through Formal Content Validity Analysis. *Psychological Methods*, *30*(2), 203–222. https://doi.org/10.1037/met0000545

Stephens, M. A. (1992). *Introduction to Kolmogorov (1933) On the Empirical Determination of a Distribution*. 93–105. https://doi.org/10.1007/978-1-4612-4380-9_9

Streiner, D. L. ., Norman, G. R. ., & Cairney, John. (2024). *Health measurement scales : a practical guide to their development and use*. 466. https://books.google.com/books/about/Health_Measurement_Scales.html?hl=es&id=UbXxEAAAQBAJ

Thurstone, L. L. (1928). Attitudes Can Be Measured. *Https://Doi.Org/10.1086/214483*, *33*(4), 529–554. https://doi.org/10.1086/214483

Triola Mario F. (2018). Estadística (12° edición). *Book*, 1–784.

Valderrama, S. (2020). Pasos para Elaborar Proyectos e Investigacion Cientificia. Cuantitativa, Cualitativa y Mixta. *Editorial San Marcos*, 6111. http://www.sancristoballibros.com/libro/pasos-para-elaborar-proyectos-de-investigacion-cientifica_45757

Valderrey Sanz, Pablo. (2010). *SPSS 17 : extracción del conocimiento a partir del análisis de datos*. 463.

Vega, M. R. de la, González, B. M., & Nava, P. B. (2025). Juicio de expertos para la validación de contenido: adaptación de un instrumento de habilidades socioemocionales. *Papeles*, *17*(34), 2025–2154. https://doi.org/10.54104/papeles.v17n34.2154

Véliz Capuñay, C. (2011). *Estadística para la administración y los negocios*. www.FreeLibros.me

Warr, P., Cook, J., & Wall, T. (1979). Scales for the measurement of some work attitudes and aspects of psychological well-being. *Journal of Occupational Psychology*, *52*(2), 129–148. https://doi.org/10.1111/j.2044-8325.1979.tb00448.x

Wayne W. Daniel, C. L. C. (2018). *Biostatistics: A Foundation for Analysis in the Health Sciences*. https://books.google.com.pe/books?hl=es&lr=&id=PON1DwAAQBAJ&oi=fnd&pg=PR7&dq=Daniel,+W.+W.,+%26+Cross,+C.+L.+(2018).+Biostatistics:+A+foundation+for+analysis+in+the+health+sciences+(11th+ed.).+John+Wiley+%26+Sons.&ots=a81ubTslRt&sig=tjHl3hME6i0GJz6ZdVsrS82IjSk&redir_esc=y#v=onepage&q&f=false

Dirección legal: Urb. Paseo del Mar
Nuevo Chimbote, Santa, Ancash
Correo electrónico: ed.honexus@gmail.com
Teléfono: 978653152

ISBN: 978-612-99293-7-8