compAS
Grupo de capacitación e investigación pedagógica

Introducción a la
Psicometría y Estadística

Héctor Salomón Mullo Guaminga
Jessica Alexandra Marcatoma Tixi

Introducción a la
Psicometría y Estadística

Héctor Salomón Mullo Guaminga
Jessica Alexandra Marcatoma Tixi

©Héctor Salomón Mullo Guaminga
Escuela Superior Politécnica de Chimborazo (ESPOCH)

Jessica Alexandra Marcatoma Tixi
Universidad Nacional de Chimborazo (UNACH)

Título del libro
Introducción a la Psicometría y Estadística

ISBN: 978-9942-33-590-6

Publicado 2022 por acuerdo con los autores.

Guayaquil-Ecuador

Cita.
Mullo, H., Marcatoma, J. (2022) Introducción a la Psicometría y Estadística. Editorial Grupo Compás.

Estudio in vitro realizadas en raíces distales con conductos ovalados de molares inferiores. Editorial Grupo Compás.

@grupocompas.ec
compasacademico@icloud.com

compÁs
Grupo de capacitación e investigación pedagógica

Contenido

Capítulo 1 Introducción a la Estadística

En este capítulo, se muestra una breve reseña histórica de la estadística y se expone la importancia de la estadística. Luego, se presentan los elementos básicos de la estadística como población, muestra, variable, dato y escalas de medida. Más adelante, se presenta una clasificación de la estadística junto con una presentación moderna de la estadística descriptiva. Al final, se exponen algunas actividades de aprendizaje sobre estadística descriptiva.

1.1. Reseña Histórica

La estadística como ciencia transversal del conocimiento se fundamenta en el manejo y análisis de grandes cantidades de información y su uso se remonta desde épocas inmemorables definidas por varias etapas de evolución. Sin embargo, la historia intelectual de la psicometría (que puede entenderse como una estadística aplicada a la psicología) se remonta solo a dos siglos atrás, y su existencia como disciplina identificable ha sido clara durante siete décadas.

1.1.1. Edad Antigua

Esta etapa inicia en 2.238 a.C. bajo el mandato de Yao, emperador de China quien dictaminó la elaboración de un censo que permita recolectar información sobre actividades agrícolas, industriales y comerciales, así también Grecia y Roma realizaron censos para contabilizar la distribución y posesión de tierras, junto con ello se crearon los primeros registros de la cantidad de fincas, personal de servicio y esclavos, mencionados registros permitieron facilitar las actividades tributarias y discernir la cantidad de personas que pudieran colaborar con el ejército. Egipto por su parte realizó recuentos de sus habitantes y sus riquezas con el fin de planificar la construcción de sus pirámides.

1.1.2. Edad Media

Aproximadamente entre 476 y 1453 d.C. la estadística no experimentó grandes avances, pero los países continuaban desarrollando censos de población con el fin de organizar a sus colectivos para distintas actividades, entre los censos más destacables se puede mencionar el solicitado por Carlomagno en 762 cuyo objetivo fue conocer las extensiones de tierra que le pertenecían a la iglesia. Otra actividad para destacar fue la obra "Originum sive Etymologiarum" publicada por Isidro Sevilla tras la recopilación y clasificación de datos biológicos.

1.1.3. Edad Moderna

Entre 1.454 y 1.789 los censos no se detuvieron y ganaron protagonismo a nivel de los diferentes gobiernos de todo el mundo, por ejemplo, en España se desarrolló el Censo de Pecheros de los Obispos. Uno de los protagonistas de la época fue John Graunt al publicar en 1662 la obra titulada "Natural and Political Observations" cuyo contenido presentaba proyecciones de la cantidad de defunciones de la época, sus patrones históricos de información fueron el número de defunciones semanales resultantes de la epidemia de peste en Inglaterra y el número de nacimientos por sexo que se mantenían en los colectivos.En esta etapa la estadística empieza a alinearse con fines políticos y de salud debido a las decisiones que se tomaron en su momento en base a los estudios generados.

1.1.4. Edad Contemporánea

Pasado 1.789 la estadística evoluciona a pasos agigantados puesto que se perfeccionaron los censos de población y se generaron estudios demográficos, económicos y sociales, tanto si tenían fines políticos como si no. El avance de la Matemática y de otras ciencias permitieron construir técnicas analíticas para establecer relaciones entre variables.

El uso del muestreo y la aplicación de técnicas de inferencia permitió que los científicos como Laplace, Gauss y Legendre desarrollaron la teoría sobre los errores en la observación, y el método de los mínimos cuadrados. Así también Galton y Pearson introdujeron conceptos de correlación y curva de regresión.

Por otro lado, los primeros trabajos relacionados con la psicometría fueron realizados por Brown y Thompson (1921) y Guilford (1936) entre otros, quienes hicieron importantes contribuciones y mostraron que es importante la teoría estadística matemática para la investigación psicométrica.

1.2. Importancia de la estadística

Los conceptos y argumentos de la estadística se utilizan en la actualidad en un gran número de ocupaciones. Las técnicas estadísticas constituyen una parte integral de las actividades de investigación en distintas áreas del saber humano. Esta ciencia día con día gana terreno de aplicación en toda actividad humana por simple que ésta sea.

A continuación, se citan algunos ejemplos de la utilidad de la estadística:

1. Los métodos estadísticos son ampliamente utilizados en psicometría. La psicometría utiliza varios tipos de encuestas y pruebas para obtener los datos primarios y un amplio espectro de métodos estadísticos para analizar los datos.
2. En las agencias gubernamentales, tanto federales, estatales o municipales utilizan la estadística para realizar planes y programas para el futuro.
3. En el campo de la ingeniería se aplica en muchas de sus actividades tales como:
    - La planeación de la producción.
    - El control de calidad.
    - Las ventas.

- El almacén.

4. En el campo económico su uso es fundamental para informar el desarrollo económico de una empresa o de un país que da a conocer los índices económicos relativos a la producción, a la mano de obra, índices de precios para el consumidor, las fluctuaciones del mercado bursátil, las tasas de interés, el índice de inflación, el costo de la vida, etc. Todos estos aspectos que se estudian se reportan e informan, no solamente describen el estado actual de la economía, sino que trazan y predicen el camino de las futuras tendencias. Así mismo, sirve a los encargados de las agencias, para tomar decisiones acertadas en sus operaciones.
5. En el campo demográfico la estadística se aplica en los registros de los hechos de la vida diaria, tales como:
   - Nacimientos.
   - Defunciones.
   - Matrimonios.
   - Divorcios.
   - Adopción.
   - Migración.
6. En materia de población los datos aportan una buena ayuda para fijar la política de estímulos al control de la natalidad, dirigir la inmigración o emigración, establecer los planes de lucha contra las enfermedades epidémicas o plagas que azotan los campos, etc.
7. En el campo educativo la Estadística contribuye al conocimiento de las condiciones fisiológicas, psicológicas y sociales de los alumnos y de los profesores. Al perfeccionamiento de los métodos de enseñanza, de evaluación, a la efectividad de programas de tutorías, la necesidad de reformas curriculares en función de los requerimientos sociales reales, etc.
8. En la industria la utilizan para el control de calidad, la implementación de incentivos a la producción, entre otros.

9. En la agricultura, se emplea en actividades como experimentos sobre la reproducción de plantas y animales entre otras cosas. También se usa la estadística para determinar los efectos de clases de semillas, insecticidas y fertilizantes en el campo.
10. En la Biología se emplean métodos estadísticos para estudiar las reacciones de las plantas y los animales en diferentes períodos ambientales y para investigar la herencia. Las leyes de Mendel sobre la herencia en donde los factores hereditarios se atribuyen a unidades llamadas genes y al estudio sistemático de los cruzamientos entre individuos portadores de genes diferentes, lo que ha permitido precisar de qué manera los genes se separan o se reúnen en las generaciones sucesivas. La verificación de las hipótesis formuladas por Mendel y sus continuadores necesitó el empleo de la estadística.
11. En la medicina, los resultados que se obtienen sobre la efectividad de fármacos se analizan por medio de métodos estadísticos. Los médicos investigadores se ayudan del análisis estadístico para evaluar la efectividad de tratamientos aplicados. La Estadística también se aplica en el establecimiento y evaluación de los procedimientos de medida o clasificación de individuos con el propósito de establecer la especificidad y sensibilidad a las enfermedades. En el Sector Salud, los técnicos de la salud la utilizan para planear la localización y el tamaño de los hospitales y de otras dependencias de sanidad. También se aplica en la investigación sobre las características de los habitantes de una localidad, sobre el diagnóstico y la posible fuente de un caso de enfermedad transmisible; sobre la proporción de personas enfermas en un momento determinado, de ciertos padecimientos de una localidad, sobre la proporción de enfermos de influenza

en dos grupos, uno vacunado contra el padecimiento y el otro no.

12. En los negocios se pueden predecir los volúmenes de venta, medir las reacciones de los consumidores ante los nuevos productos, probar la efectividad de una campaña publicitaria, etc.
13. En el deporte se ocupa para determinar el impacto de una nueva dieta alimenticia en el rendimiento de atletas o someter a prueba la efectividad de dos o más técnicas de ejercitación y práctica de un deporte.
14. El Mundo Político, todo intento de buen gobierno exige, dejando a un lado los presupuestos ideológicos, algo tan simple y complejo a la vez como es el conocer sobre qué realidad se gobierna; exige el estar perfectamente informado de las posiciones objetivas de partida para desde ellas, tomar las medidas adecuadas a fin de dirigir la sociedad a esa meta. Es claro que cuanto más, correcto y veraz sea este conocimiento de la realidad, las medidas de gobierno serán también más correctas. El conocimiento de la realidad para los fines del buen gobierno pasa por su cuantificación, o que es equivalente, por la obtención de estadísticas.

1.3. Elementos básicos

1.3.1. Población y muestra

Un **Población o Colectivo** es un conjunto de unidades estadísticas con una característica en común, de acuerdo con la posibilidad de tabulación se los puede agrupar en colectivos finitos e infinitos. Por ejemplo: computadoras, papayas hawaianas, etc.

La **Unidad de Observación** o **Unidad Estadística** es el elemento mínimo al que se observa dentro de un estudio, puede estar representado por entidades simples (persona, animal, cosa o fenómeno natural) o compuestas (familia,

institución educativa, ciudad, etc.). Por ejemplo: estudiante, papaya, computador, conejo, etc.

Una **Población Estadística** o **Colectivo Estadístico** es un conjunto de unidades estadísticas que guardan una característica en común pero delimitadas en tiempo y en espacio. Algunos ejemplos se muestran a continuación:

- Estudiantes de cuarto semestre de la carrera de Estadística del semestre número 1 del año 2022.
- Papayas hawaianas producidas en Guayaquil-Ecuador en el primer bimestre del año.
- Computadoras de marca Dell de la Facultad de Ciencias de la Escuela Superior Politécnica de Chimborazo durante el año 2022.

Una **Muestra** es un subconjunto de una población. En general es deseable que este subconjunto sea representativo de la población. La representatividad se puede encontrar a través de técnicas de muestreo.

1.3.2. Variables y datos

Una **variable** es una cualidad o característica de los elementos de interés de la población. Esta variable debe ser medible. Por ejemplo: preguntas de una prueba, edad, dimensión de la calzada, etc.

Los **datos** son las mediciones de las características de los elementos de interés de toda la población. Por ejemplo: la variable edad puede tomar los siguientes valores 2, 13, 23, 69, etc.

1.3.3. Clasificación de las variables

Las **Variables Cualitativas** son aquellas cuyas características son categóricas, es decir, indican cualidades o

atributos. También se las conoce como **Variables Categóricas**.

Según el número de caracteres, las variables cualitativas se clasifican en:

- **Dicotómicas:** El dominio de la variable admite únicamente dos datos a elegir. Por ejemplo: para la variable estado de un artefacto, las categorías pueden ser estado Bueno o Malo.
- **Politómicas:** El dominio de la variable admite tres o más datos a elegir. Por ejemplo: para el sector de nacimiento, las categorías pueden ser sector sur, centro o norte.

Las **Variables Cuantitativas** son aquellas cuyos datos son de tipo numérico y simbolizan una cantidad.

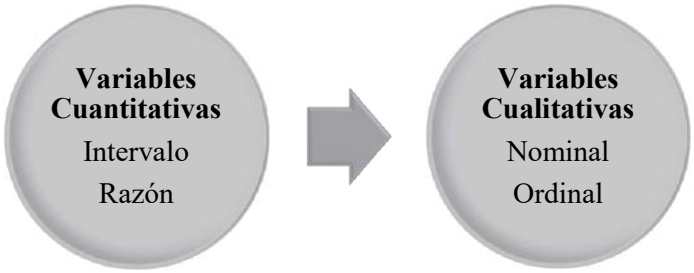
Las variables cuantitativas se clasifican en:

- **Discretas:** En su dominio solo admite valores enteros. Por ejemplo: la variable número de hermanos pueden tomar los siguientes valores 0, 1, 3, 4, 8, etc.
- **Continuas:** En su dominio se admite cualquier valor numérico ya sea entero, fraccionario o incluso irracional. Teóricamente, se cubren todos los posibles valores en un intervalo. Este tipo de variable se obtiene principalmente a través de mediciones y está sujeta a la precisión de instrumentos de medición. Por ejemplo: la variable calificación en una prueba puede tomar valores en el intervalo $[0, 20]$.

1.3.4. Escalas de medida

Es importante diferenciar la escala de medida de los valores observados en las variables. A continuación, se describen las cuatro escalas de medida según el tipo de variable (ver Figura 1.1):

Figura 1.1: Clasificación de variables y escalas de medición.



Fuente: Elaboración propia.

Se dice que una variable cualitativa se mide mediante una **Escala Nominal**, o es de tipo nominal, si sus valores son etiquetas o atributos y no existe un orden entre ellos. Cada uno de los caracteres agrupa a un grupo mutuamente excluyente y la única relación implicada es la de equivalencia (=). Por ejemplo: la variable estrato social se mide en escala nominal, ya que tiene las características de estrato bajo, mediano o alto.

Una variable cualitativa se mide mediante una **Escala Ordinal**, o es de tipo ordinal, si sus valores son etiquetas o atributos, pero existe un cierto orden entre ellos. Cada uno de los caracteres agrupa a un grupo mutuamente excluyente y la relación implicada es la de equivalencia (=) dentro de cada grupo y la de mayor que (>) entre grupos. Por ejemplo: la variable nivel de instrucción es de tipo ordinal, ya que tiene las características (en el Ecuador) de Primaria, Secundaria, Pregrado, Maestría, Doctor, Posdoctorado.

Una variable cuantitativa se mide mediante una **Escala de Intervalo** si existe una noción de distancia entre los valores de la variable, aunque no se pueden realizar operaciones numéricas y el cero en el dominio de la variable es relativo. Por ejemplo: la variable temperatura medida en grados Celsius está en escala de intervalo, debido a que cero grados Celsius no representa la ausencia de temperatura.

Se dice que una variable cuantitativa se mide mediante una **Escala de Razón** si los valores de la variable tienen un sentido físico y existe el cero absoluto. Por ejemplo: la variable distancia recorrida por una bicicleta está en escala de razón, porqué el valor de cero representa ausencia de desplazamiento.

Es posible transformar una variable cuantitativa en cualitativa, creando un conjunto de intervalos contiguos que abarcan el rango de la variable de interés, este proceso se conoce como **Desratización**.

1.3.5. Parámetro y estadístico

Un **parámetro** es una medida estadística que resume la información de una población. Esta medida se calcula utilizando todos los datos recolectados en la población de interés. Es un valor fijo para la población de interés y generalmente se representa con letras griegas.

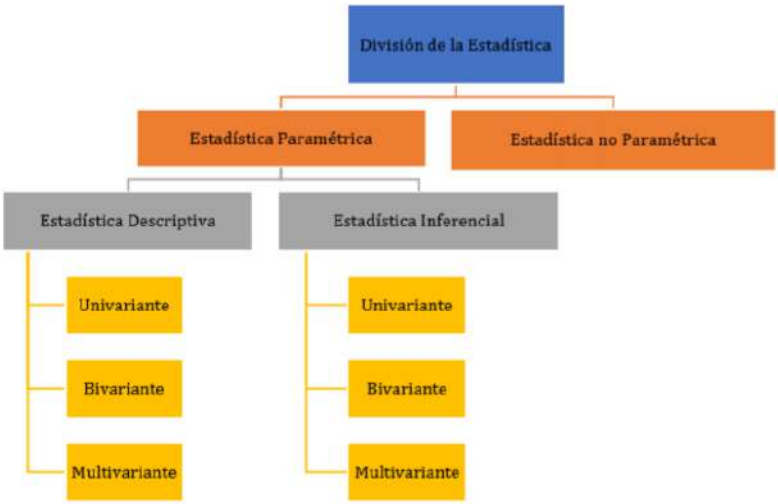
Un **estadístico** es una medida estadística que resume la información de una muestra. Esta medida se calcula utilizando los datos de la muestra. Se utiliza para estimar los parámetros. Es un valor que cambia en función de la muestra extraída de la misma población y generalmente se representa con letras latinas.

1.4.La estadística y su clasificación

La **Estadística** es el estudio científico relativo al conjunto de métodos y técnicas encaminados al análisis de fenómenos conocidos e inciertos a través de la obtención, representación y análisis de observaciones numéricas o categóricas, así como inferir generalizaciones acerca de las características para los colectivos de interés y tomar las decisiones más acertadas en el campo de su aplicación.

La clasificación de la estadística se presenta en la Figura 1.2 siguiente:

Figura 1.2: Clasificación de la estadística.



Fuente: Elaboración propia.

Estadística Paramétrica: Estudia modelos específicos de distribución donde deben cumplirse ciertos supuestos acerca de los parámetros (medidas estadísticas calculadas mediante información de todo el colectivo) de la población en función de una muestra investigada, supuestos obligatorios a cumplirse ya que la validez de los resultados de una investigación que utiliza técnicas paramétricas depende de su comprobación. Este grupo prioritario de la estadística se subdivide en:

Estadística Descriptiva: Ciencia que recopila, organiza e interpreta la información numérica o cualitativa. Tiene como propósito presentar resúmenes de un conjunto de datos y poner de manifiesto sus características principales, mediante representaciones tabulares o gráficas y complementándose con medidas descriptivas. El interés se centra en describir el conjunto dado de datos y no se plantea el extender las conclusiones a otros datos diferentes o bien, a una población.

Estadística Inferencial: Conjunto de técnicas que se utiliza para obtener conclusiones que sobrepasan los límites del conocimiento aportado por los datos, busca obtener información de un colectivo mediante un procedimiento del

manejo de datos de la muestra. Se pude realizar estudios de tipo univariados, bivariados y multivariados haciendo referencia a la manipulación de una, dos y tres o más variables respectivamente.

Estadística no Paramétrica: Se encarga del estudio de distribuciones no específicas y no requiere de la comprobación de supuestos sobre los parámetros de la población; sin embargo, previo a la aplicación de técnicas no paramétricas se comprueba la existencia de aleatoriedad de las observaciones captadas en una muestra.

1.5. Estadística Descriptiva

Los datos recolectados a partir de una muestra se pueden resumir en función de una serie de estadísticas descriptivas. En este apartado mostramos las estadísticas más importantes para la tendencia central, dispersión y forma de la distribución de los datos recolectados. Como se dijo anteriormente, las letras latinas se utilizan normalmente para nombrar estadísticas de datos de muestra. Si las estadísticas se calculan para la población, se denominan parámetros en lugar de estadísticas. Los parámetros se representan con letras griegas.

Ejemplo 1.1 En la Tabla 1.1 siguiente se muestran algunos parámetros y estadísticos muy utilizados:

Tabla 1.1: Ejemplo de parámetro y estadística.

Medida de resumen	Parámetro	Estadístico
Media aritmética	μ	$\bar{X}$
Varianza	σ^2	s^2
Desviación estándar	σ	s
Proporción	π	p

Fuente: Elaboración propia.

Las estadísticas que se pueden calcular dependen de la naturaleza de los datos. Si queremos hablar de la ubicación de los datos nominales, la información numérica se limita a las frecuencias (la mayor frecuencia, en particular). Para datos ordinales, podemos tener en cuenta el orden en los datos. Por lo tanto, puede tener sentido hablar, por ejemplo, del elemento medio de una muestra. Para las variables medidas en una escala de intervalo o una escala de razón, la media aritmética juega un papel importante. Una descripción general de las estadísticas que abordamos en este libro, según la clasificación de variables y escala de medición se muestra a continuación:

Variables cuantitativas en escala de intervalo o razón:

- Medidas de tendencia central:
  - Moda.
  - Mediana.
  - Cuartiles.
  - Media Aritmética.
- Medidas de dispersión:
  - Rango.
  - Rango intercuartil.
  - Desviación media absoluta.
  - Variación y desviación estándar.
  - Coeficiente de variación.
- Medidas de forma:
  - Coeficiente de asimetría de Pearson.
  - Coeficiente de curtosis.
- Momentos:
  - Momentos centrales
  - Momentos no centrales
- Medidas de correlación y asociación:
  - Covarianza.
  - Coeficiente de correlación.
  - Coeficiente de correlación de rango.

Variables cualitativas en escala ordinal:

- Medidas de tendencia central:
  - Moda.
  - Mediana.
  - Cuartiles.
- Medidas de dispersión:
  - Rango.
  - Rango intercuartil.
  - Índice de dispersión ordinal.
- Medidas de correlación y asociación:
  - Coeficiente de correlación de rango.

Variables cualitativas en escala nominal:

- Medidas de tendencia central:
  - Moda.
- Medidas de dispersión:
  - Índice de dispersión nominal.

Es importante tener en cuenta que las estadísticas que se definen para una escala de medición particular generalmente también se pueden usar para datos en una escala de medición más alta, así se podría calcular el índice de dispersión nominal en variables cuantitativas en escala de razón luego de un proceso de discretización.

El enfoque en este capítulo está principalmente en estadísticas descriptivas univariadas para variables cuantitativas. Esto debido a que, se trabajará con datos psicométricos de tipo cuantitativo a lo largo de este Libro.

Se denota el número de observaciones en una muestra o el tamaño de la muestra por n. Siempre que se trabaje con una sola variable cuantitativa esta es, x, y los valores observados de esa variable son, $x_1, \dots, x_n$. Ahora, cuando se consideran dos

variables, estas son x e y. Las observaciones son $x_1, \dots, x_n$ e $y_1, \dots, y_n$.

1.5.1. Medidas de ubicación relativa y de dispersión

Una medida de ubicación relativa indica la posición de una observación en comparación con las demás observaciones.

1.5.1.1. Estadístico de orden

El estadístico de $i-$ésimo orden $x_{(i)}$ en una muestra de n observaciones es la $i-$ésimo observación después de ordenar las observaciones de menor a mayor.

El estadístico de primer orden es el mínimo, mientras que el estadístico de último orden es el máximo. Llamamos a estos dos valores x_{min} y x_{max} .

Ejemplo 1.7 Encontrar el estadístico de primer orden y de noveno orden de la edad de 20 señores, cuyos valores son:

Edad	
$x_1 = 30$	$x_{11} = 40$
$x_2 = 30$	$x_{12} = 42$
$x_3 = 32$	$x_{13} = 43$
$x_4 = 33$	$x_{14} = 44$
$x_5 = 35$	$x_{15} = 45$
$x_6 = 36$	$x_{16} = 45$
$x_7 = 36$	$x_{17} = 46$
$x_8 = 36$	$x_{18} = 47$
$x_9 = 38$	$x_{19} = 59$
$x_{10} = 39$	

El estadístico de primer orden es $x_{(1)} = x_{min} = 30$ y el de noveno orden es $x_{(9)} = 38$.

1.5.1.2. Percentiles

El $(100 * p) - ésimo$ percentil o cuantil c_p de una muestra, donde $0 < p < 1$, es un número real mayor que (alrededor de) $100 * p\%$ de las observaciones, y menor que (alrededor de) $100 * (1 - p)\%$ de las observaciones.

Hay varios métodos ligeramente diferentes para calcular los percentiles. Los diferentes enfoques conducen a diferencias notables en conjuntos de datos pequeños, pero para conjuntos de datos grandes, prácticamente no hay diferencia entre los métodos de cálculo. A continuación, mostramos una forma de calcular estos percentiles:

- Para el cálculo de percentiles, en primer lugar, se ordenan las n observaciones de menor a mayor en las posiciones $1, 2, \dots, n$.
- Se calcula la posición del $(100 * p) - ésimo$ percentil como $q = p(n + 1)$.
- Si q es un número entero, entonces $x_{(q)}$ es el $(100 * p) - ésimo$ percentil o cuantil c_p de la muestra.
- Si q no es un número entero:
    - Primero, determine el entero más grande que es menor que q, y llame a ese entero a.
    - Luego, determine la diferencia entre q y a, y llame a esa diferencia f;
    - Entonces, el $(100 * p) - ésimo$ percentil o cuantil c_p de la muestra es

$$c_p = (1 - f).x_{(a)} + f.x_{(a+1)}$$

$$= x_{(a)} + f.\left(x_{(a+1)} - x_{(a)}\right).$$

Ejemplo 1.8 Considere los siguientes datos sobre las temperaturas en grados Celsius en el Riobamba-Ecuador de los primeros 31 días del año 2022, tomados a las 11:38 horas todos los días en el centro de la ciudad. Estos datos se muestran a continuación:

Temperaturas	
$x_1 = 17$	$x_{21} = 20$
$x_2 = 18$	$x_{22} = 18$
$x_3 = 19$	$x_{23} = 20$
$x_4 = 21$	$x_{24} = 22$
$x_5 = 22$	$x_{25} = 18$
$x_6 = 20$	$x_{26} = 20$
$x_7 = 19$	$x_{27} = 22$
$x_8 = 15$	$x_{28} = 16$
$x_9 = 16$	$x_{29} = 20$
$x_{10} = 16$	$x_{30} = 16$
$x_{11} = 20$	$x_{31} = 17$
$x_{12} = 18$	$x_{32} = 18$
$x_{13} = 15$	$x_{33} = 21$
$x_{14} = 18$	$x_{34} = 20$
$x_{15} = 21$	$x_{35} = 21$
$x_{16} = 17$	$x_{36} = 16$
$x_{17} = 18$	$x_{37} = 20$
$x_{18} = 19$	$x_{38} = 18$
$x_{19} = 18$	$x_{39} = 16$
$x_{20} = 16$	$x_{40} = 21$

A partir de estos calcular los percentiles 33 y 67.
Para dar respuesta a esta pregunta, en primer lugar, ordenamos los datos de menor a mayor. Los datos ordenados se presentan en la siguiente tabla:

Temperaturas	
$x_1 = 15$	$x_{21} = 18$
$x_2 = 15$	$x_{22} = 19$
$x_3 = 16$	$x_{23} = 19$
$x_4 = 16$	$x_{24} = 19$

$x_5 = 16$	$x_{25} = 20$
$x_6 = 16$	$x_{26} = 20$
$x_7 = 16$	$x_{27} = 20$
$x_8 = 16$	$x_{28} = 20$
$x_9 = 16$	$x_{29} = 20$
$x_{10} = 17$	$x_{30} = 20$
$x_{11} = 17$	$x_{31} = 20$
$x_{12} = 17$	$x_{32} = 20$
$x_{13} = 18$	$x_{33} = 21$
$x_{14} = 18$	$x_{34} = 21$
$x_{15} = 18$	$x_{35} = 21$
$x_{16} = 18$	$x_{36} = 21$
$x_{17} = 18$	$x_{37} = 21$
$x_{18} = 18$	$x_{38} = 22$
$x_{19} = 18$	$x_{39} = 22$
$x_{20} = 18$	$x_{40} = 22$

Luego para el percentil 33, calculamos $q = p(n + 1)$, para $p = 0.33$ y $n = 40$. Esto nos da $q = 0.33(40 + 1) = 13.53$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 13$ y $f = 0.53$. En consecuencia, $c_{0.33} = (1 - 0.53)x_{(13)} + 0.53 * x_{(14)} = 0.47 * 18 + 0.53 * 18 = 18$.

Para el caso del percentil 67, calculamos $q = p(n + 1)$, para $p = 0.67$ y $n = 40$. Esto nos da $q = 0.67(40 + 1) = 27.47$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 27$ y $f = 0.47$. En consecuencia, $c_{0.67} = (1 - 0.47)x_{(27)} + 0.47 * x_{(28)} = 0.53 * 20 + 0.47 * 20 = 20$.

Si el producto $(100 \times p)$ es un múltiplo de 10, el percentil correspondiente a veces es llamado el $(100 \times p) - \acute{e}simo$ decil. El quinto decil, es decir, el percentil 50 o cuantil $c_{0.5}$, es siempre igual a la mediana.

El primer (Q_1), segundo (Q_2) y tercer (Q_3) cuartil son los percentiles $25, 50$ y 75 respectivamente. En otras palabras, $Q_1 = c_{0.25}$, $Q_2 = c_{0.5}$ y$Q_3 = c_{0.75}$. Es evidente que el segundo cuartil es la mediana.

Entre otras cosas, los cuantiles se utilizan para probar si una muestra de datos podría originarse de una densidad de probabilidad dada. Esto se puede hacer mediante la construcción de los llamados diagramas de cuantiles.

1.5.1.3. Rango

Las estadísticas más conocidas de variación o dispersión son para datos cuantitativos. Estas estadísticas miden la variación o dispersión alrededor de un valor central. Datos con la misma media o la mediana aún pueden diferir mucho en sus dispersiones. También existen medidas menos conocidas de variación para variables nominales y ordinales. Estas medidas se irán describiendo más adelante.La medida más fácil de variación de un conjunto de datos es su rango. Para determinar el rango, se necesita al menos una escala ordinal.

El **rango** R de un conjunto de observaciones es la diferencia entre el valor de la observación más grande y más pequeña:

$$R = x_{max} - x_{min}.$$

La mayor ventaja de la definición del rango es su simplicidad. El mayor inconveniente es que solo se utilizan dos observaciones en el cálculo. Toda observación intermedia no tiene influencia. Está claro que el rango es particularmente sensible a valores extremos. En la industria, la moda se utiliza a menudo en el control estadístico de procesos.

1.5.1.4. Rango intercuartílico

Se obtiene una mejor imagen de la variación o dispersión de los datos de la muestra usando la distancia entre el primer y el tercer cuartil:

El **rango intercuartílico** Q se define como la diferencia entre el tercer y primer cuartil:

$$Q = Q_3 - Q_1.$$

Dado que la mitad de los datos están entre Q_1 y Q_3, el rango intercuartílico es una medida que se extiende por la mitad del conjunto de datos. Esta medida de dispersión es insensible a los valores extremos, siempre que menos del 25% de los valores de los datos sean extremadamente pequeños y menos del 25% sean extremadamente largos.

1.5.1.5. Desviación absoluta media

Una medida de variación de una muestra de datos cuantitativos alrededor de la media aritmética es la desviación absoluta media (DAM). Al igual que la media aritmética, la desviación absoluta media es sensible a los valores extremos.

La **desviación media absoluta** es la media aritmética de la desviación absoluta de cada dato con respecto a la media aritmética del conjunto de datos:

$$DAM = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

1.5.1.6. Varianza

Para datos de muestra medidos en una escala de razón o una escala de intervalo, se usa la varianza más a menudo como una medida de dispersión.

La varianza muestral s^2 de un conjunto de observaciones $x_1, \dots, x_n$ es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

La varianza muestral es la media de las desviaciones al cuadrado de cada dato con respecto a la media aritmética del conjunto de datos, esto dividido para $n-1$ en lugar de n. La varianza de la muestra también se puede calcular con las fórmulas alternativas:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right\}.$$

Sabiendo que $\sum_{i=1}^{n} x_i = n\bar{x}$, se puede mostrar que la ecuación anterior se cumple.

Si calculamos la varianza de todos los N elementos de una población o un proceso con N elementos, entonces hablamos de una varianza de población o una varianza de proceso. Para estas varianzas, se usa el símbolo σ^2. Se calculan en base a la siguiente ecuación:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N},$$

Donde μ representa la media poblacional. Para muestras grandes, la varianza muestral y la varianza de la población son similares.

Dado que la varianza de la población se calcula como la media aritmética de un conjunto de desviaciones cuadradas, es más intuitivo que la varianza muestral, donde dividimos por $n-1$ en lugar de n. La varianza muestral no se puede calcular cuando el conjunto de datos contiene sólo una observación, porque necesitamos dividir por $n-1$. Teniendo en cuenta el significado de la varianza de la muestra, esto tiene sentido, ya que una sola observación no contiene ninguna información sobre la dispersión.

El denominador $n-1$ en la definición de la varianza muestral se denomina número de grados de libertad. Cada grado de libertad corresponde a una unidad de información. En una muestra de n observaciones, tenemos n unidades de información. Para calcular la varianza muestral, necesitamos calcular primero la media muestral (media aritmética). Calculado este valor nos muestra una unidad de información. Entonces, solo quedan $n-1$ unidades de información o grados de libertad para calcular la varianza.

Al igual que la media aritmética, la varianza muestral se puede calcular con la ayuda de frecuencias si los datos están agrupados. La fórmula requerida es:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k} f_i(x_i - \bar{x})^2,$$

Donde x_i es el centro de la $i-ésima$ clase, f_i la frecuencia de la $i-ésima$ clase, n el número de observaciones, y k el número de clases.

Vale la pena señalar que una varianza siempre es no negativa, y cero si y sólo si todas las observaciones en la muestra tienen el mismo valor, es decir, si y sólo si las observaciones no varían. Una varianza siempre se expresa en una unidad que es el cuadrado de la unidad original medida. Por ejemplo, si los datos se miden en segundos, entonces la varianza es medida en segundos al cuadrado.

La varianza de una transformación lineal $y_1 = ax_1 + b, y_2 = ax_2 + b, \dots, y_n = ax_n + b$ de las observaciones $x_1, x_2, \dots, x_n$, donde a y b son constantes, es igual a la varianza de los datos originales multiplicados por a^2, es decir, $a^2 * s_x^2$.

1.5.1.7. Desviación estándar

La desviación estándar de la muestra es la raíz cuadrada (positiva) de la varianza de la muestra:

$$s = \sqrt{s^2}.$$

Análogamente, la desviación estándar de la población es la raíz cuadrada (positiva) de la varianza de la población:

$$\sigma = \sqrt{\sigma^2}.$$

Las desviaciones estándar se expresan en la misma unidad que los datos originales. En general, la desviación estándar de la muestra da una mejor imagen de la distribución del conjunto de datos que el rango. Sin embargo, si solo tenemos dos observaciones, entonces la desviación estándar de la muestra y el rango contienen la misma información.

Ejemplo 1.9 Considere los siguientes datos de las calificaciones sobre 10 puntos de estudiantes universitarios de la asignatura de Álgebra Lineal en una evaluación sorpresa:

Calificaciones	
$x_1 = 8$	$x_{21} = 14$
$x_2 = 15$	$x_{22} = 10$
$x_3 = 13$	$x_{23} = 17$
$x_4 = 7$	$x_{24} = 7$
$x_5 = 18$	$x_{25} = 10$
$x_6 = 18$	$x_{26} = 12$
$x_7 = 7$	$x_{27} = 7$
$x_8 = 17$	$x_{28} = 10$
$x_9 = 12$	$x_{29} = 18$
$x_{10} = 13$	$x_{30} = 7$
$x_{11} = 13$	$x_{31} = 5$
$x_{12} = 8$	$x_{32} = 7$
$x_{13} = 16$	$x_{33} = 16$
$x_{14} = 8$	$x_{34} = 17$
$x_{15} = 11$	$x_{35} = 12$
$x_{16} = 17$	$x_{36} = 14$
$x_{17} = 19$	$x_{37} = 17$
$x_{18} = 8$	$x_{38} = 9$
$x_{19} = 11$	$x_{39} = 14$
$x_{20} = 6$	$x_{40} = 7$

A partir de estos calcular la desviación absoluta media, varianza y la desviación estándar.
La desviación absoluta media de los datos es:

$$DAM = \frac{|8 - 11.875| + |15 - 11.875| + \cdots + |7 - 11.875|}{40},$$

$$= 3.631.$$

La varianza muestral de los datos es:

$$s^2 = \frac{(8-11.875)^2 + (15-11.875)^2 + \cdots + (7-11.875)^2}{40-1},$$
$$= 17.548.$$

La desviación estándar muestral es:

$$s = \sqrt{17.548} = 4.189.$$

1.5.1.8. Coeficiente de variación

Aunque la varianza y la desviación estándar juegan un papel extremadamente importante en estadística, a veces no son las mejores opciones cuando se necesita comparar la variación entre dos conjuntos de datos.

El **coeficiente de variación** CV se define como la relación entre la desviación muestral s y la media aritmética $\bar{x}$:

$$CV = \frac{s}{\bar{x}}.$$

Ejemplo 1.10 Sea una muestra de datos con media aritmética de 11.875 y una desviación estándar muestral de 4.189. Calcular el coeficiente de variación:

$$CV = \frac{s}{\bar{x}} = \frac{4.189}{11.875} = 0.353\ .$$

El coeficiente de variación no es confiable cuando $\bar{x}$ es muy pequeño y es sensible a valores atípicos. El coeficiente de variación es útil para comparar la dispersión de datos con diferentes medias e indispensable cuando se compara la variación de datos con diferentes dimensiones, es decir, datos expresados en diferentes unidades de medida.

1.5.2. Medidas de asimetría

Los histogramas y los diagramas de tallo y hoja de datos de muestra pueden ser simétricos o asimétricos. Un histograma que no es simétrico se llama sesgado. En un histograma que está sesgado a la izquierda (o negativamente sesgado), la cola de la izquierda es más larga que la cola de la derecha. En un histograma que está sesgado a la derecha (o positivamente sesgado), la cola de la derecha es más larga que la cola izquierda.

En un histograma unimodal, la asimetría se puede determinar en función de las posiciones de la media aritmética, la mediana y la moda. En un histograma perfectamente simétrico, las tres estadísticas son idénticas. En un histograma sesgado a la izquierda, la media es menor que la mediana, que a su vez es menor que la moda. Cuando un histograma está sesgado a la derecha, la moda es más pequeña que la mediana, que a su vez es más pequeña que la media aritmética. La razón de esto es que la media aritmética es más sensible a valores extremadamente grandes o pequeños que la mediana. Basado en esta observación, Pearson introdujo una medida de asimetría y se detalla a continuación:

El **coeficiente de asimetría de Pearson** se define como:

$$S_p = \frac{3(\bar{x} - \tilde{x})}{s}.$$

En la definición del coeficiente de asimetría de Pearson, se divide por la desviación estándar de la muestra para obtener una medida que sea independiente de la unidad de medida. Sin esta división, $\bar{x} - \tilde{x}$ podría hacerse de manera fácil artificialmente grande o pequeño, simplemente cambiando la unidad de medida. El factor 3 en la definición del coeficiente de asimetría asegura que siempre esté entre -3 y $+3$. En un sesgo a la derecha o positivamente distribución sesgada, $\bar{x} > \tilde{x}$ y en consecuencia $S_p > 0$. En una distribución sesgada a la izquierda o negativamente en una distribución sesgada, lo contrario es cierto.

Ejemplo 1.11 Considere los siguientes datos sobre las temperaturas en grados Celsius en el Riobamba-Ecuador de los primeros 31 días del año 2022, tomados a las 11:38 horas todos los días en el centro de la ciudad. Estos datos se muestran a continuación:

Temperaturas	
$x_1 = 15$	$x_{21} = 18$
$x_2 = 15$	$x_{22} = 19$
$x_3 = 16$	$x_{23} = 19$
$x_4 = 16$	$x_{24} = 19$
$x_5 = 16$	$x_{25} = 20$
$x_6 = 16$	$x_{26} = 20$
$x_7 = 16$	$x_{27} = 20$
$x_8 = 16$	$x_{28} = 20$
$x_9 = 16$	$x_{29} = 20$
$x_{10} = 17$	$x_{30} = 20$
$x_{11} = 17$	$x_{31} = 20$
$x_{12} = 17$	$x_{32} = 20$
$x_{13} = 18$	$x_{33} = 21$
$x_{14} = 18$	$x_{34} = 21$
$x_{15} = 18$	$x_{35} = 21$
$x_{16} = 18$	$x_{36} = 21$
$x_{17} = 18$	$x_{37} = 21$
$x_{18} = 18$	$x_{38} = 22$

$x_{19} = 18$	$x_{39} = 22$
$x_{20} = 18$	$x_{40} = 22$

Calcular el coeficiente de asimetría de Pearson.

La media aritmética de los datos es:

$$\bar{x} = \frac{15 + 15 + \cdots + 22}{40},$$
$$= 18.575.$$

La mediana, es decir, el percentil 50 se obtiene calculando $q = p(n + 1)$, para $p = 0.50$ y $n = 40$. Esto nos da $q = 0.50(40 + 1) = 20.50$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 20$ y $f = 0.50$. En consecuencia, $c_{0.50} = (1 - 0.50)x_{(20)} + 0.50 * x_{(21)} = 0.50 * 18 + 0.50 * 18 = 18$.

La varianza muestral de los datos es:

$$s^2 = \frac{(15 - 18.575)^2 + (15 - 18.575)^2 + \cdots + (22 - 18.575)^2}{40 - 1},$$
$$= 4.199.$$

La desviación estándar muestral es:

$$s = \sqrt{4.199} = 2.049.$$

El coeficiente de asimetría de Pearson es:

$$S_p = \frac{3 * (18.575 - 18)}{2.049},$$
$$= 0.842.$$

Esto indica un sesgo ligeramente positivo, que corresponde a que la distribución de los datos tiene una cola derecha larga.

Una segunda medida de asimetría se deriva del tercer momento central de los datos de la muestra. Generalmente, el $k - ésimo$ momento central de una muestra de tamaño n se define de la siguiente manera:

El $\boldsymbol{k - ésimo}$ **momento central** de una muestra es la media de las $\boldsymbol{k - ésimo}$ potencias de las desviaciones de la media muestral:

$$m_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^k}{n}.$$

La **asimetría de Fisher**, quizás la medida de asimetría más utilizada se basa en el tercer momento central, es decir, m_3 y se calcula como $\frac{m_3}{s^3}$, o una función del mismo. Una posibilidad es calcular la asimetría de Fisher como

$$S_F = \frac{n^2}{(n-1)(n-2)} \frac{m_3}{s^3} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^3.$$

La medida de asimetría es adimensional y toma el valor de cero para un histograma simétrico ($S_F = 0$), positivo para un

histograma sesgado a la derecha ($S_F > 0$), y negativo para un histograma sesgado a la izquierda ($S_F < 0$).

Ejemplo 1.12 Sea un conjunto de datos sobre la cantidad de sustentantes de las 24 provincias del Ecuador en la evaluación Ser Bachiller del año 2021 siguiente:

Cantidad de sustentantes	
$x_1 = 10031$	$x_{13} = 11024$
$x_2 = 10187$	$x_{14} = 10068$
$x_3 = 10223$	$x_{15} = 11114$
$x_4 = 10249$	$x_{16} = 11155$
$x_5 = 10258$	$x_{17} = 11204$
$x_6 = 10336$	$x_{18} = 11209$
$x_7 = 10456$	$x_{19} = 11262$
$x_8 = 10559$	$x_{20} = 11407$
$x_9 = 10589$	$x_{21} = 11615$
$x_{10} = 10655$	$x_{22} = 11659$
$x_{11} = 10770$	$x_{23} = 11736$
$x_{12} = 11011$	$x_{24} = 11795$

Calcular el coeficiente de asimetría de Fisher.

La media aritmética de los datos es

$$\bar{x} = \frac{10031 + 10187 + \cdots + 11795}{24},$$

$$= 10898.79.$$

La varianza muestral de los datos es

$$s^2 = \frac{(10031 - 10898.79)^2 + (10187 - 10898.79)^2 + \cdots + (11795 - 10898.79)^2}{40 - 1},$$

$$= 290986.3.$$

La desviación estándar muestral es:

$$s = \sqrt{290986.3} = 539.432.$$

El coeficiente de asimetría de Fisher es:

$$S_F = \frac{40}{(39)(38)}\left[\left(\frac{10031 - 10898.79}{539.432}\right)^3 + \left(\frac{10187 - 10898.79}{539.432}\right)^3 + \cdots + \left(\frac{11795 - 10898.79}{539.432}\right)^3\right].$$

$$= 0.030.$$

Por lo tanto, los datos están ligeramente sesgados a la derecha.

1.5.3. Medidas de tendencia central

Entre las medidas características de una distribución destacan las llamadas medidas de centralización, que indican el valor promedio de los datos, o en torno a qué valor se distribuyen estos. Sea el conjunto de datos $x_1, \ldots, x_n$, las medidas de tendencia central se calculan según las fórmulas siguientes:

1.5.3.1. Media aritmética

La **media aritmética** $\bar{x}$, es la suma de los datos de la variable dividida para su el tamaño de muestra. Su función es:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

La media aritmética es una medida adimensional y representa el punto central del dominio de la variable cuantitativa. No es una medida robusta ante la presencia de valores atípicos, es decir, se ve influenciado en su valor cuando existen datos atípicos o fuera de lo común en el conjunto de datos.

Algunas propiedades de la media aritmética de una muestra son:

- La suma de todas las observaciones es igual a la media aritmética multiplicada por el tamaño de la muestra n, es decir, $\sum_{i=1}^{n} x_i = n\bar{x}$.
- La suma de las desviaciones de las observaciones de la media es cero, es decir, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
- La suma de las desviaciones al cuadrado de las observaciones de una constante c, es decir, $\sum_{i=1}^{n}(x_i - c)^2$ es mínimo si $c = \bar{x}$.
- La media aritmética de una muestra de valores constantes a, … , a es igual a la constante, es decir, $\bar{a} = a$.
- La media aritmética de un número de observaciones $ax_1 + b, \ldots, ax_n + b$ (donde a y b son constantes), obtenidas transformando linealmente un conjunto original de observaciones, $x_1, \ldots, x_n$ se puede obtener aplicando la transformación lineal a la media original, es decir, $a\bar{x} + b$.

Ejemplo 1.2 Calcular la calificación promedio de un curso de niños de 4to de Educación General Básica de Ecuador, donde se evaluó la asignatura de Matemática Básica con 20 ejercicios

(conocidos comúnmente como ítems). Las calificaciones fueron 19, 17, 16, 6 y 15 para los estudiantes numerados como 1, 2, 3, 4 y 5 respectivamente.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}, \\ = \frac{19 + 17 + 16 + 6 + 15}{5}, \\ = 14.6.$$

La calificación promedio en la asignatura de Matemática Básica de un curso de 4to de Educación General Básica de Ecuador es de 14.6.

1.5.3.2. Media geométrica

La **media geométrica** G de un conjunto de observaciones $x_1, \dots, x_n$ es:

$$G = \sqrt[n]{x_1 * \dots * x_n} = (x_1 * \dots * x_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}.$$

Esta definición implica que la media geométrica sólo se puede calcular para estrictamente observaciones positivas. La media geométrica es siempre menor o igual que la media aritmética. Ambos son iguales sólo si todas las observaciones en un conjunto de datos son iguales.

Ejemplo 1.3 Calcular la media geométrica de diez personas que ganan en dólares: 170; 172; 168; 165; 173; 178; 180; 165; 167; 172

$$G = \sqrt[n]{170 * 172 * 168 * 165 * 173 * 178 * 180 * 165 * 167 * 172} = 170.93$$

1.5.3.3. Mediana

La **mediana**, es el valor central del dominio de la variable cuantitativa y separa al conjunto de datos ordenados en dos partes iguales. Es una medida robusta ante la presencia de valores atípicos y su cálculo depende si la colección de datos es par o impar. Sea el conjunto de datos $x_1 + \cdots + x_n$, en primer lugar, se ordena en forma ascendente, esto es $x_1 \leq x_2 \leq \cdots \leq x_n$ incluyendo las repeticiones para proceder al cálculo de la medida del siguiente modo:

- Si el tamaño de muestra n es par la mediana se calcula como $\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$. Donde $x_{\frac{n}{2}}$ y $x_{\frac{n}{2}+1}$ son los datos ordenados que se ubica en la posición $\frac{n}{2}$ y $\frac{n}{2}+1$ respectivamente.
- Si el tamaño de muestra n es impar la mediana se calcula como $\tilde{x} = x_{\frac{n-1}{2}+1}$. Donde $x_{\frac{n-1}{2}+1}$ es el dato ordenado que se ubica en la posición $\frac{n-1}{2} + 1$.

Ejemplo 1.4 Calcular la mediana de la calificación de un curso de niños de 4to de Educación General Básica de Ecuador, donde se evaluó la asignatura de Matemática Básica con 20 ejercicios (conocidos comúnmente como ítems). Las calificaciones fueron 17, 16, 6 y 15.
En primer lugar, se ordena la información:

Calificaciones	Parámetro
$x_1 = 6$	$x_3 = 16$
$x_2 = 15$	$x_4 = 17$

Ahora, como $n = 4$ y claramente es par, el valor de la mediana es el promedio de los valores en las posiciones $\frac{n}{2} = \frac{4}{2} = 2$ y $\frac{n}{2} + 1 = \frac{4}{2} + 1 = 3$, es decir, la mediana es

$$\tilde{x} = x_3 = 16.$$

Ejemplo 1.5 Calcular la mediana de la calificación de un curso de niños de 4to de Educación General Básica de Ecuador, donde se evaluó la asignatura de Matemática Básica con 20 ejercicios (conocidos comúnmente como ítems). Las calificaciones fueron 19, 17, 16, 6 y 15.
En primer lugar, se ordena la información:

Calificaciones	**Parámetro**
$x_1 = 6$	$x_4 = 17$
$x_2 = 15$	$x_5 = 19$
$x_3 = 16$	

Ahora, como $n = 5$ y claramente es impar, el valor de la mediana es el dato en la posición:

$$\frac{n-1}{2} + 1 = \frac{5-1}{2} + 1,$$
$$= 3,$$

Es decir, la mediana es $\tilde{x} = x_3 = 16$.
Algunas propiedades de la mediana son:

- Alrededor del 50% de las observaciones están por debajo o por encima de la mediana.
- La mediana no se ve afectada por unas pocas observaciones extremadamente grandes o pequeñas.
- La suma de las desviaciones absolutas de las observaciones x_i con respecto a una constante c. Esto en ecuación es $\sum_{i=1}^{n} |x_i - c|$, la expresión anterior es mínima para un $c = \tilde{x}$.

- La mediana es la media de un conjunto de datos truncados.

1.5.3.4. Moda

Para datos nominales, la única información numérica es la frecuencia de las diferentes clases o categorías. Aparte de determinar frecuencias, no podemos realizar ningún cálculo con los datos.

Para este tipo de datos, la estadística más utilizada es la clase con mayor frecuencia en la muestra.

La **moda** $\boldsymbol{M_0}$ de una muestra es la observación con mayor frecuencia.

Ejemplo 1.6 Encontrar la moda de la estatura de 19 señoras, cuyas medidas son:

Estaturas	
$x_1 = 1.45$	$x_{11} = 1.50$
$x_2 = 1.45$	$x_{12} = 1.53$
$x_3 = 1.48$	$x_{13} = 1.53$
$x_4 = 1.48$	$x_{14} = 1.53$
$x_5 = 1.48$	$x_{15} = 1.53$
$x_6 = 1.48$	$x_{16} = 1.53$
$x_7 = 1.50$	$x_{17} = 1.53$
$x_8 = 1.50$	$x_{18} = 1.53$
$x_9 = 1.50$	$x_{19} = 1.53$
$x_{10} = 1.50$	

Para responder a esta pregunta es necesario realizar una tabla de frecuencias, esta tabla considera todos los valores que puede tomar la variable (x_i) y la frecuencia relativa (n_i) de las mismas, además presenta las frecuencias acumuladas (N_i) que es más que la suma de las frecuencias relativas.

Estatura (x_i)	n_i	N_i
1,45	2	2
1,48	4	6
1,50	5	11
1,53	8	19
Total	19	

Como la frecuencia de la puntuación 1.53 es de 8 y es la mayor frecuencia, entonces la moda de este conjunto de datos es 1.53. La moda no solo se puede determinar para datos en escala nominal, ordinal, escala y de razón. Sin embargo, la moda rara vez se usa para variables cuantitativas continuas porque hay mejores estadísticas para este tipo de datos. Para datos agrupados, él depende en gran medida de qué intervalos se utilizan para agrupar los datos (es decir, qué se eligen intervalos para construir un histograma o para calcular frecuencias), lo que no hace atractivo su uso. En el caso de no existir un dato con frecuencia mayor al resto, la variable en análisis es amodal.

La moda puede existir y ser única, sin embargo, pueden existir dos o más modas, es decir, pueden existir dos o más valores que aparecen con la misma frecuencia máxima en el conjunto de datos. En este caso se dice que la variable es bimodal o multimodal, según sea el caso.

Los histogramas multimodales a menudo resultan del uso de un número demasiado grande de clases, o de una muestra basada en datos de más de una población, o un proceso que puede operar en más de una forma.

1.5.4. Medidas de curtosis

La medida en que un histograma tiene un pico pronunciado se cuantifica mediante un número denominado curtosis. La curtosis puede verse como una medida de inclinación.

La **curtosis** de una muestra de datos es

$$g = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

Al igual que la asimetría, la curtosis es adimensional. La curtosis es cero para datos normalmente distribuidos. Un valor positivo para la curtosis indica un pico más prominente que en los datos normalmente distribuidos, mientras que un valor negativo indica un pico más plano.

Ejemplo 1.13 Sea un conjunto de datos sobre el número de llegadas por día de clientes a un local comercial de calzado deportivo para hombres, contabilizado en todo el mes de febrero del año 2022 siguiente:

Cantidad de clientes	
$x_1 = 92$	$x_{15} = 105$
$x_2 = 100$	$x_{16} = 103$
$x_3 = 119$	$x_{17} = 100$
$x_4 = 103$	$x_{18} = 101$
$x_5 = 96$	$x_{19} = 92$
$x_6 = 95$	$x_{20} = 94$
$x_7 = 88$	$x_{21} = 99$
$x_8 = 101$	$x_{22} = 100$
$x_9 = 112$	$x_{23} = 108$
$x_{10} = 83$	$x_{24} = 85$
$x_{11} = 114$	$x_{25} = 110$
$x_{12} = 88$	$x_{26} = 86$
$x_{13} = 96$	$x_{27} = 91$
$x_{14} = 98$	$x_{28} = 97$

Calcular la curtosis del conjunto de datos.

La media aritmética de los datos es

$$\bar{x} = \frac{92 + 100 + \cdots + 97}{28},$$
$$= 98.429.$$

La varianza muestral de los datos es

$$s^2 = \frac{(92 - 98.429)^2 + (100 - 98.429)^2 + \cdots + (97 - 98.429)^2}{28 - 1},$$
$$= 79.810.$$

La desviación estándar muestral es

$$s = \sqrt{79.810} = 9.934.$$

La curtosis del conjunto de datos es

$$g = \frac{28 * 29}{27 * 26 * 25}\left[\left(\frac{92 - 98.429}{9.934}\right)^4 + \cdots + \left(\frac{97 - 98.429}{9.934}\right)^4\right] - \frac{3(27)^2}{26 * 25},$$
$$= -1.252$$

Por lo tanto, la distribución de los datos es menos apuntada, es decir, más plana que la distribución normal.

1.5.5. Transformación y estandarización de datos

Anteriormente se indicó como la media y la varianza de un conjunto de datos, x_i, se ven afectadas por una transformación lineal, es decir, $y_i = ax_i + b$, donde a y b son constantes. La media de los datos transformados linealmente se puede encontrar aplicando la transformación lineal a la media de los datos originales, es decir, $\bar{y} = a\bar{x} + b$. También, la mediana de los datos transformados linealmente es igual a la transformación lineal de la mediana original. La varianza de los valores de y_i se puede calcular como $s_y^2 = a^2 s_x^2$, es decir, a^2 veces la varianza original. La desviación estándar de los valores de y_i es igual a $s_y = |a|s_x$. Los coeficientes de asimetría de los datos transformados son idénticos a los de los datos originales si a es positivo y opuestos si a es negativo. La curtosis permanece sin cambios bajo una transformación lineal de los datos.

Para transformaciones no lineales, no existen fórmulas simples para calcular estadísticas de resumen. Por ejemplo, si $y_i = x_i^2$, no es cierto que $\bar{y} = \bar{x}^2$. Si $y_i = \ln(x_i)$, entonces $\bar{y} \neq \ln(\bar{x})$.

La transformación lineal más utilizada en estadística se denomina **estandarización**. Para este tipo de transformación, a y b se establecen en $\frac{1}{s}$ y $-\frac{\bar{x}}{s}$, respectivamente. La nueva variable obtenida mediante esta transformación se denota con la letra z. Un valor estandarizado

$$z_i = \frac{x_i - \bar{x}}{s},$$

Expresa a cuántas desviaciones estándar se encuentra una observación x_i de la media de la muestra, ya que $x_i = \bar{x} + z_i s$. Para cualquier variable estandarizada z, la media es cero y la varianza es uno. Tenga en cuenta que las variables estandarizadas se utilizan en el cálculo de la asimetría de Fisher y de la curtosis.

1.5.6. Diagramas de caja

Una representación gráfica de uso frecuente de datos ordinales o cuantitativos univariados es el llamado diagrama de caja. Hay muchas versiones diferentes de diagramas de caja que se pueden encontrar en la literatura estadística. La parte central de los datos suele representarse mediante un recuadro. La caja está delimitada por el primer y el tercer cuartil. Por lo general, la mediana se representa con una línea entre estos dos cuartiles.

Además, de estas estadísticas, un diagrama de caja indica valores extremadamente grandes y pequeños usando puntos. Para ello, se utilizan reglas generales. Una de esas reglas generales establece que una observación x_i es extrema si

$$x_i < Q_1 - 1.5 * Q \quad o \quad x_i > Q_3 + 1.5 * Q.$$

En esta expresión, Q es el rango intercuartil. Algunos diagramas de caja también contienen líneas o bigotes que llegan hasta el más pequeño y hasta el valor de muestra más grande que no se consideran valores extremos.

Ejemplo 1.14 Sea un conjunto de datos sobre el número de llegadas por día de clientes a un local comercial de calzado deportivo para hombres, contabilizado en todo el mes de febrero del año 2022 siguiente:

Cantidad de clientes por días	
$x_1 = 35$	$x_{15} = 105$
$x_2 = 38$	$x_{16} = 103$
$x_3 = 20$	$x_{17} = 100$
$x_4 = 103$	$x_{18} = 101$
$x_5 = 96$	$x_{19} = 92$
$x_6 = 95$	$x_{20} = 94$
$x_7 = 88$	$x_{21} = 99$
$x_8 = 101$	$x_{22} = 100$
$x_9 = 112$	$x_{23} = 108$
$x_{10} = 83$	$x_{24} = 85$
$x_{11} = 114$	$x_{25} = 110$
$x_{12} = 88$	$x_{26} = 86$
$x_{13} = 96$	$x_{27} = 91$
$x_{14} = 98$	$x_{28} = 97$

Dibujar un diagrama de caja. Observe que los tres primeros datos son demasiado pequeños en comparación con las otras observaciones.En primer lugar, se ordena los datos y quedan de la siguiente manera:

Cantidad de clientes por días	
$x_1 = 20$	$x_{15} = 97$
$x_2 = 35$	$x_{16} = 98$
$x_3 = 38$	$x_{17} = 99$
$x_4 = 83$	$x_{18} = 100$
$x_5 = 85$	$x_{19} = 100$
$x_6 = 86$	$x_{20} = 101$
$x_7 = 88$	$x_{21} = 101$
$x_8 = 88$	$x_{22} = 103$
$x_9 = 91$	$x_{23} = 103$
$x_{10} = 92$	$x_{24} = 105$
$x_{11} = 94$	$x_{25} = 108$
$x_{12} = 95$	$x_{26} = 110$
$x_{13} = 96$	$x_{27} = 112$
$x_{14} = 96$	$x_{28} = 114$

Luego, para el cálculo del percentil 25, calculamos $q = p(n + 1)$, para $p = 0.25$ y $n = 28$. Esto nos da $q = 0.25(28 + 1) = 7.25$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 7$ y $f = 0.25$. En consecuencia, $Q_1 = c_{0.25} = (1 - 0.25)x_{(7)} + 0.25 * x_{(8)} = 0.75 * 88 + 0.25 * 88 = 88$.

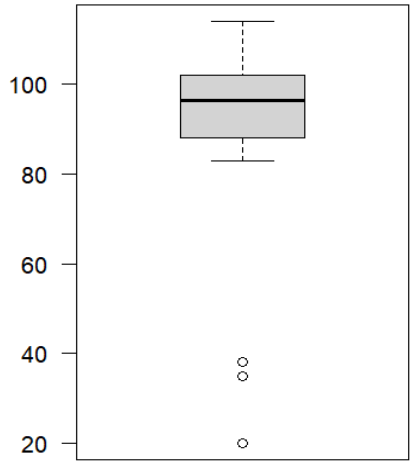
La mediana, es decir, el percentil 50 se obtiene calculando $q = p(n + 1)$, para $p = 0.50$ y $n = 28$. Esto nos da $q = 0.50(28 + 1) = 14.50$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 14$ y $f = 0.50$. En consecuencia, $c_{0.50} = (1 - 0.50)x_{(14)} + 0.50 * x_{(15)} = 0.50 * 96 + 0.50 * 97 = 96.50$.

Para el cálculo del percentil 75, calculamos $q = p(n + 1)$, para $p = 0.75$ y $n = 28$. Esto nos da $q = 0.75(28 + 1) = 21.75$. Ahora, como p no es un entero, tenemos los siguientes valores $a = 21$ y $f = 0.75$. En consecuencia, $Q_3 = c_{0.75} = (1 - 0.75)x_{(21)} + 0.75 * x_{(22)} = 0.25 * 101 + 0.75 * 103 = 102.5$.

Ahora, el rango Inter cuartil es $Q = Q_3 - Q_1 = 102.5 - 88 = 14.50$.

Los bigotes del diagrama de caja son $Q_1 - 1.5 * Q = 88 - 1.5 * 14.50 = 66.25$ y $Q_3 + 1.5 * Q = 102.5 + 1.5 * 14.50 = 124.25$. Por lo tanto, el diagrama de caja es el siguiente:

Figura 1.3: Diagrama de caja de la cantidad de clientes por días.



Fuente: Elaboración propia.

La Figura 1.3 contiene el diagrama de caja de la cantidad de clientes por días de un local comercial de calzado deportivo para hombres. Los diagramas de caja en la figura indican que hay tres valores atípicos dentro de los 28 datos, a saber, en tres días el número de clientes fue bastante inferior que lo normal. Estos tres puntos de datos corresponden a los primeros tres días del mes de febrero del año 2022.

En el mismo sentido, no hay observaciones extremadamente grandes en este conjunto de datos. Los bigotes de los diagramas de caja se extienden hasta el valor máximo de 124.25 por un lado, y hasta el valor más pequeño 66.25. La mediana es de 96.50, mientras que el primer y el tercer cuartil son 88 y 102.50 respectivamente. El rango intercuartílico es de 14.50. La media aritmética es 90.64.

1.5.7. Distribución estadística de frecuencia

La distribución estadística de frecuencias es usada para resumir la información de una variable cuantitativa continua, a través de clases o intervalos. La construcción de la **distribución estadística de frecuencia** es el siguiente procedimiento:

En primer lugar, se calcula el rango de los datos, es decir, se calcula $R = x_{max} - x_{min}$. Luego se determina el número de intervalos o clases, del siguiente modo: $k = \sqrt{n}$. Siguiente, se calcula la amplitud de la clase mediante $A = \frac{R}{k}$.

El límite inferior de la primera clase es igual al dato mínimo de la variable y el límite superior de la última clase debe ser igual al dato máximo de la variable. Esta observación garantiza que todos los datos de la variable se encuentren formando parte del conteo de frecuencias. Esto se representa a continuación:

$$L_{inferior\ de\ la\ clase\ 1} = X_{mín}\ \ y\ \ L_{superior\ de\ la\ última\ clase} = X_{máx}.$$

El cuadro general que se forma se presenta a continuación (ver Tabla 1.2):

Tabla 1.2: Distribución estadística de frecuencia.

Variable Cuantitativa		**Frecuencia Absoluta**	**Frecuencia Relativa**	**Frecuencia Absoluta Acumulada**	**Frecuencia Relativa Acumulada**
Li	**Ls**	o_i	f_i	O_i	F_i
Li_1	Ls_1	o_1	f_1	$O_1 = o_1$	$F_1 = f_1$
Li_2	Ls_2	o_2	f_2	$O_2 = O_1 + o_2$	$F_2 = F_1 + f_2$
⋮	⋮	⋮	⋮	⋮	⋮
Li_k	Ls_k	o_k	f_k	$O_k = O_{k-1} + o_n$	$F_k = F_{k-1} + f_k$
	Total	n	1		

Fuente: Elaboración propia.

Ejemplo 1.15 Construir la distribución estadística de frecuencia del tiempo en segundo de simulación de los ordenadores para generar 1000000 de datos de una distribución uniforme. Los datos se presentan a continuación:

Tiempo en segundos	
$x_1 = 140$	$x_{26} = 148$
$x_2 = 140$	$x_{27} = 149$
$x_3 = 140$	$x_{28} = 150$
$x_4 = 140$	$x_{29} = 151$
$x_5 = 141$	$x_{30} = 152$
$x_6 = 141$	$x_{31} = 152$
$x_7 = 142$	$x_{32} = 152$
$x_8 = 142$	$x_{33} = 152$
$x_9 = 143$	$x_{34} = 152$
$x_{10} = 143$	$x_{35} = 153$
$x_{11} = 144$	$x_{36} = 153$
$x_{12} = 145$	$x_{37} = 153$
$x_{13} = 145$	$x_{38} = 154$
$x_{14} = 146$	$x_{39} = 154$
$x_{15} = 146$	$x_{40} = 155$
$x_{16} = 147$	$x_{41} = 155$
$x_{17} = 147$	$x_{42} = 155$
$x_{18} = 147$	$x_{43} = 155$
$x_{19} = 147$	$x_{44} = 156$
$x_{20} = 148$	$x_{45} = 157$
$x_{21} = 148$	$x_{46} = 157$
$x_{22} = 148$	$x_{47} = 157$
$x_{23} = 148$	$x_{48} = 158$
$x_{24} = 148$	$x_{49} = 159$
$x_{25} = 148$	$x_{50} = 161$

En primer lugar, se calcula el rango, esto es $R = 161 - 140 = 21$. El número de intervalos es igual a $\mathrm{k} = \sqrt{50} \cong 7$ y cada uno de ellos tendrá una amplitud de $\mathrm{A} = \frac{21}{7} = 3$. Luego de conocer esta información se presenta la distribución estadística de frecuencias:

Tiempos de Simulación Li − Ls		o_i	f_i	O_i	F_i
[140	143)	8	0,16	8	0,16
[143	146)	5	0,10	13	0,26
[146	149)	13	0,26	26	0,52
[149	152)	3	0,06	29	0,58
[152	155)	10	0,20	39	0,78
[155	158)	8	0,16	47	0,94
[158	161]	3	0,06	50	1
	Total	50	1		

En líneas de interpretar la información más relevante de la tabla se puede indicar que el 26% de los ordenadores simulan 1000000 de datos de una distribución uniforme en un tiempo dentro del intervalo [146,149).

1.5.8. Datos bivariados

Ahora, se estudian situaciones en las que se desea trabajar con dos variables simultáneamente con las variables x e y. Si se tiene n observaciones de la primera variable y n observaciones correspondientes de la segunda variable. La estructura de los datos se muestra en la Tabla 1.3. Por ejemplo, podríamos tener una tabla de **datos bivariados** cuando registramos la altura (x) y el peso (y) de un número de personas. Una tabla de datos con más de dos variables se denomina **tabla de datos multivariados**.

Tabla 1.3: Datos bivariados de las variables x e y.

Observación	x	y
1	x_1	y_1
2	x_2	y_2

3	x_3	y_3
⋮	⋮	⋮
n	x_n	y_n

Fuente: Elaboración propia.

Medir e interpretar la relación o asociación entre dos (o más) variables es una de las principales tareas de la estadística. La covarianza y la correlación se utilizan con frecuencia para medir la fuerza de una relación lineal entre dos variables cuantitativas.

1.5.8.1. Covarianza

Suponga que tiene un conjunto de datos con n observaciones de dos variables cuantitativas x e y. La **covarianza muestral** entre las variables x e y se define como

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

La covarianza puede ser positiva o negativa. Un término $(x_i - \bar{x})(y_i - \bar{y})$ en la definición de la covarianza muestral es positiva si la observación i tiene:

- Un valor de x menor que $\bar{x}$, y un valor de y menor que $\bar{y}$.
- Un valor de x mayor que $\bar{x}$, y un valor de y mayor que $\bar{y}$.

Un término $(x_i - \bar{x})(y_i - \bar{y})$ en la definición de la covarianza muestral es negativo si la observación i tiene:

- Un valor de x menor que $\bar{x}$, y un valor de y mayor que $\bar{y}$.
- Un valor de x mayor que $\bar{x}$, y un valor de y menor que $\bar{y}$.

Si el número de términos negativos es dominante, esto da como resultado una covarianza negativa. En el otro caso, la covarianza es positiva.

Ejemplo 1.16 El Instituto Nacional de Evaluación Educativa (INEVAL) del Ecuador desarrolla la evaluación Ser Estudiante en sustentantes de 4to, 7mo y 10mo de Educación General Básica (EGB) y en sustentantes de 3ero de Bachillerato General Unificado (BGU). Al respecto, se muestra a continuación las variables de índice socioeconómico (el rango de valores de $]-3{,}3[$) y calificación (sobre 100 puntos) obtenida de 14 sustentantes de 7to de EGB:

Índice socioeconómico (y)	**Calificación** (x)
$x_1 = 1.34$	$y_1 = 80.00$
$x_2 = 2.61$	$y_2 = 104.31$
$x_3 = 1.77$	$y_3 = 78.72$
$x_4 = 1.50$	$y_4 = 78.05$
$x_5 = 2.60$	$y_5 = 85.19$
$x_6 = -0.33$	$y_6 = 48.57$
$x_7 = -0.23$	$y_7 = 74.31$
$x_8 = 2.77$	$y_8 = 81.56$
$x_9 = 2.21$	$y_9 = 74.18$
$x_{10} = -0.45$	$y_{10} = 59.96$
$x_{11} = -0.93$	$y_{11} = 62.20$
$x_{12} = 1.14$	$y_{12} = 95.76$
$x_{13} = 2.61$	$y_{13} = 78.41$
$x_{14} = -0.34$	$y_{14} = 72.94$

Calcular la covarianza muestral entre las variables índice socioeconómico y calificación.

Las medias muestrales son:

$$\bar{x} = \frac{1}{14}(1.34 + 2.61 + \cdots - 0.34),$$

$$= 1.16.$$

$$\bar{y} = \frac{1}{14}(80.00 + 104.31 + \cdots + 72.94),$$

$$= 76.73.$$

La covarianza muestral es:

$$s_{XY} = \frac{1}{13}[(1.34 - 1.16)(80 - 76.73) + \cdots + (-0.34 - 1.16)(72.94 - 76.73)]$$

$$= 13.07.$$

El concepto de covarianza también se puede aplicar a una población finita de N elementos, la covarianza se define de la siguiente manera:

La **covarianza poblacional** entre las variables x e y como:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y),$$

donde μ_X es la media poblacional de la variable x y μ_Y es la media poblacional de la variable y.

Está claro que la varianza muestral es un caso especial de la covarianza muestral, y la varianza poblacional un caso especial de la covarianza poblacional. Una varianza es simplemente la covarianza de una variable consigo misma. Por consiguiente, se tiene que $s_X^2 = s_{XX}$ y $\sigma_X^2 = \sigma_{XX}$.

Una desventaja de la covarianza como estadístico de asociación entre dos variables es que el resultado depende de la unidad de medida utilizada para cada una de las variables. Siempre que, por ejemplo, una variable que originalmente estaba expresada en metros se vuelve a expresar en centímetros, la covarianza de esa variable con cualquier otra variable dada se multiplicará por 100.

Debido a su sensibilidad a las unidades de medida, la magnitud de una covarianza es difícil de interpretar. El coeficiente de correlación, que no sufre de este problema, es una medida más popular de la asociación entre dos variables cuantitativas.

1.5.8.2. Correlación

El coeficiente de correlación muestral, también conocido como **coeficiente de correlación de Pearson**, de las observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ Se define como

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

El **coeficiente de correlación poblacional** entre dos variables es

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Los coeficientes de correlación están acotados entre -1 y 1. Las variables con una correlación de 1 están perfectamente correlacionadas positivamente. Las variables con una correlación de -1 están perfectamente correlacionadas negativamente. En cada uno de estos casos, existe una relación lineal de la forma $y = ax + b$ entre las dos variables en estudio. Un coeficiente de correlación siempre tiene el mismo signo que la covarianza correspondiente. Las variables con una correlación de 0 se llaman no correlacionadas.

Una observación importante sobre los coeficientes de correlación es que solo indican hasta qué punto existe una relación lineal entre dos variables. Una correlación de (casi) cero solo indica que no existe una relación lineal entre las dos variables. Sin embargo, puede haber una relación cuadrática, cúbica o logarítmica entre las variables.

Ejemplo 1.17 En un hospital público se desea estudiar la correlación entre las variables edad (x) y peso en libras 120,34, (y) de los niños que fueron atendidos el día 31 de enero del año 2022. Esta información se muestra en la tabla siguiente:

Edad (y)	**Peso** (x)
$x_1 = 11$	$y_1 = 120$
$x_2 = 4$	$y_2 = 35$
$x_3 = 7$	$y_3 = 50$
$x_4 = 6$	$y_4 = 40$

$x_5 = 7$	$y_5 = 43$
$x_6 = 10$	$y_6 = 60$
$x_7 = 4$	$y_7 = 30$
$x_8 = 11$	$y_8 = 130$
$x_9 = 7$	$y_9 = 52$
$x_{10} = 8$	$y_{10} = 55$
$x_{11} = 5$	$y_{11} = 44$
$x_{12} = 5$	$y_{12} = 43$
$x_{13} = 10$	$y_{13} = 60$
$x_{14} = 6$	$y_{14} = 55$

Calcular el coeficiente de correlación de Pearson entre las variables edad y peso.

Las medias muestrales son:

$$\bar{x} = \frac{1}{14}(11 + 4 + \cdots + 6),$$

$$= 7.21.$$

$$\bar{y} = \frac{1}{14}(120 + 35 + \cdots + 55),$$

$$= 58.36.$$

El coeficiente de correlación de Pearson es:

$$s_{XY} = \frac{1}{13}[(11 - 7.21)(120 - 58.36) + \cdots + (6 - 7.21)(55 - 58.36)],$$

$$= 59.69.$$

Las desviaciones estándar de las variables x e y son:

$$S_X = \sqrt{\frac{1}{13}[(11 - 7.21)^2 + \cdots + (6 - 7.21)^2]},$$

$$= 2.46.$$

$$S_Y = \sqrt{\frac{1}{13}[(120 - 58.36)^2 + \cdots + (55 - 58.36)^2]},$$

$$= 29.66.$$

El coeficiente de correlación de Pearson es:

$$r_{XY} = \frac{59.69}{2.46 * 29.66},$$

$$= 0.818.$$

Por lo tanto, la correlación lineal de Pearson es de 0.818, es decir, la correlación es fuerte y positiva.

1.6.Actividades de Aprendizaje

1.6.1. Población, muestra, variables y datos

1. Formule colectivos a partir de las unidades estadísticas citadas.
   - Institución educativa .
   - Mujer.
   - Fruta.
   - Vehículo.

2. Construya colectivos estadísticos a partir de las siguientes unidades estadísticas.
   - Paciente.
   - Hombre.
   - Futbolista.
   - Celular.

3. Son finitos o infinitos los siguientes colectivos.
   - Constelaciones.
   - Embutidos La Europea.
   - Niños.
   - Empleados Públicos.

4. Proponga características cualitativas y cuantitativas para las siguientes unidades estadísticas:

Unidad estadística	Variable cuantitativa	Variable Cualitativa
Institución educativa		
Mujer		
Fruta		
Vehículo		
Paciente		
Hombre		
Futbolista		

5. Visite el enlace https://www.eumed.net/rev/caribe/2018/10/mujeres-diagnostico-cervicitis.html y lea el resumen del artículo e identifique la unidad estadística, colectivo o colectivo estadístico, muestra, variable cualitativa y cuantitativa.

1.6.2. Clasificación de variables y tipos de datos

1. Del siguiente grupo de variables identificar el tipo de variable, escala de medida y la unidad de medida:

Variable	Tipo de Variable	Escala de Medida	Unidad de medida
• Rendimiento de un sistema (horas) • Tipos de software (libre, pagado) • Sistemas Operativos (Linux, Windows, Redhat) • Memoria RAM de ordenadores (GB) • Tipos de Comunicación (transaccional, personal y seriada) • Herramientas Tecnológicas (Zoom, Teams, Virtual Message) • Marcas de computadores (Dell, Hp, Asus) • Precios de impresoras ($) • Temperatura de ordenadores (°C) • Tamaño de una aplicación (GB) • Número de programas instalados en un ordenador • Velocidad del internet (GB/seg) • Proveedor de Internet (Netlife, Cnt, Fast Net) • Dispositivos tecnológicos (tablets, laptops, celulares) • Metodologías para desarrollo de aplicaciones móviles (Cascada, Lineal, Espiral, Prototipado)			

1.6.3. Resúmenes de información estadística

1. Se solicitó a un grupo de estudiantes seleccionen el color favorito para un computador y los resultados fueron:

Color favorito				
negro	azul	blanco	rojo	azul
azul	rojo	negro	blanco	rojo
rojo	blanco	blanco	azul	rojo

   Indique, ¿Cuál fue el color preferido?

2. Indique cuál es la categoría que predomina en las siguientes variables:

   a. Proveedor de internet: Netlife, Cnt, Fast net, Netlife, Cnt, Fast net, Netlife, Netlife, Netlife, Netlife, Cnt, Cnt, Movistar, Movistar.
   b. Consumo de redes sociales más de 8 horas al día: si, si, no, no, no, si, si, si, si, si, si, si, si, si, si, no, no.

1.6.4. Distribución de frecuencias y estadísticas descriptivas

1. Aplicada una prueba de medición de la inteligencia a un grupo de 40 alumnos, las puntuaciones obtenidas sobre 100 puntos son las que aquí se presentan:

Puntuaciones
45 56 78 80 100 87 75 64 89 90
46 89 100 100 69 98 87 76 45 39
77 85 45 68 88 99 75 98 65 40
59 48 99 93 96 11 74 10 100 66

Resumir la información mediante una tabla de frecuencias sin clases.

2. En un curso de 40 alumnos, se desea estudiar el comportamiento de la variable estatura, registrándose los siguientes valores:

Puntuaciones									
1.52	1.64	1.54	1.64	1.73	1.55	1.56	1.57	1.58	1.58
1.59	1.53	1.60	1.60	1.61	1.61	1.65	1.63	1.79	1.63
1.62	1.60	1.64	1.54	1.65	1.62	1.66	1.76	1.70	1.69
1.71	1.72	1.72	1.55	1.73	1.73	1.75	1.67	1.78	1.63

Resumir la información mediante una tabla de frecuencias con clases.

3. Calcule e interprete las medidas de tendencia central y dispersión de los ejercicios 1 y 2.

1.6.5. Datos bivariados

1. El Instituto Nacional de Evaluación Educativa del Ecuador desarrolla la evaluación Ser Estudiante en sustentantes de 4to, 7mo y 10mo de Educación General Básica (EGB) y en sustentantes de 3ero de Bachillerato General Unificado (BGU). Al respecto, se muestra a continuación las variables de índice socioeconómico (el rango de valores de $]-3,3[$) y calificación (sobre 100 puntos) obtenida de 25 sustentantes de 4to de EGB:

Índice socioeconómico	Calificación
2.97	73.76
-1.31	39.52
-1.47	38.24
-1.32	39.44
0.91	57.28
-1.23	40.16
-0.44	46.48
-1.52	37.84
-2.19	32.48
-2.33	31.36
0.76	56.08
-0.8	43.6
-2.35	31.2
-0.57	45.44
0.25	52
2.38	69.04
2.75	72
-1.88	34.96
0.65	55.2
0.34	52.72
-0.72	44.24
-0.5	46
2.97	73.76
-2.65	28.8
1.91	65.28

Calcular e interpretar la covarianza y coeficiente de correlación.

Capítulo 2 : Introducción a la Psicometría

En este capítulo se estudia una aproximación a la psicometría mediante la presentación de una breve historia de la psicometría. Después, se describe la medición psicológica y se formula la Teoría Clásica de la Prueba y Teoría de Respuesta al Ítem. En el mismo apartado, se describen los supuestos de estas teorías y se muestra la utilización práctica de los parámetros de los ítems a partir del dictamen de aceptación y rechazo de ítems. Al final, se expone el vínculo entre la estadística y la psicología.

2.1.Introducción

El término psicometría en el diccionario de la Real Academia Española (RAE). Significa "Medida de los fenómenos psíquicos". Sin embargo. se necesita una mejor definición. Al respecto. Ramsay (2021) considera como psicométricos la mayoría de los modelos cuantitativos para el comportamiento de humanos y animales que es controlado por el sistema nervioso central. Por su parte, Word Net define la psicometría como "cualquier rama de la psicología relacionada con las mediciones psicológicas" (Fellbaum y Christiane, 2005). En este sentido un enfoque importante de la psicometría es la medición y el análisis de las habilidades. los intereses. las creencias y los valores humanos.

La psicología cognitiva se ocupa de cuestiones de memoria. resolución de problemas formación de conceptos. razonamiento y en general. de todos los aspectos del procesamiento de la información dado que la teoría estadística puede considerarse como un modelo para procesar información de este tipo, es evidente que la psicología cognitiva es relevante para la estadística. Por lo tanto, para el propósito del presente libro se considera los métodos psicométricos y los

temas relacionados con la psicología cognitiva bajo el término único psicometría.

2.2. Breve historia de la psicometría

Uno de los primeros tratados sobre psicofísica. correlación y pruebas de capacidad es el de Brown y Thompson (1921). quienes hicieron importantes contribuciones al desarrollo de la teoría de las pruebas y el análisis factorial. respectivamente. Otro pionero fue Guilford (1936) quién publicó el libro "Psychometrika Methods". Casi al mismo tiempo la Sociedad Psicométrica de la Universidad de Chicago publicó el primer ejemplar de la revista Psychometika en 1936. Estas publicaciones hacían un fuerte llamado a reconocer una base estadística matemática para la investigación psicológica.

La revista Educational and Psychological Measurement apareció en 1941. En 1946, la American Psychological Association (APA) creó la Division of Evaluation. Measurement. and Statistics. *La Society for Mathematical Psychology* comenzó a reunirse en 1968 y dio lugar a el *Journal of Mathematical Psychology,* esta revista publica artículos sobre psicometría aplicada a la medición en entornos educativos. Desde este punto de partida, la psicometría estableció su territorio intelectual y se ha venido desarrollando a lo largo de los años.

2.3. Medición psicológica

La medición es la asignación de números a objetos o eventos de acuerdo con reglas (Stevens, 1951). Cuando la medición hace uso de datos psicológicos. se denomina escalamiento

psicológico. porque la asignación de numerales coloca los objetos o eventos en una escala. Para hacer uso de datos psicológicos las reglas para la asignación de números generalmente se basan en modelos matemáticos o estadísticos para esos datos. Los métodos de escalamiento de Thurstone se generaron directamente a partir del empleo de la distribución gaussiana. primero para repetir las respuestas de un individuo dado. luego con una extensión a la distribución de respuestas para muchos individuos. Estos métodos estimularon desarrollos posteriores en el escalamiento multidimensional y en la teoría de respuesta al ítem.

2.3.1. Teoría de la Prueba

2.3.1.1. Teoría Clásica del Test (TCT)

La teoría de las pruebas es el paradigma de un problema exclusivo de la investigación psicológica que requiere una solución estadística. Considerando la puntuación de una prueba desde un punto de vista estadístico. es muy deseable derivar una declaración auxiliar de su precisión. En el enfoque más básico de la teoría de la puntuación verdadera. La puntuación de la prueba U se considera la suma de una puntuación verdadera V y un error aleatorio E.

$$U = V + E.$$

La desviación estándar de los errores E es una declaración de la falta de precisión. o error estándar, de la puntuación de la prueba. Ese error estándar se puede estimar utilizando una estimación de la confiabilidad del puntaje U. donde la confiabilidad es la correlación al cuadrado entre los puntajes observados U y los puntajes verdaderos V. Algunos indicadores

de TCT son la dificultad, la discriminación y la fiabilidad de los reactivos.
Son importantes, algunos indicadores de TCT como es la dificultad, discriminación y fiabilidad de los reactivos. La dificultad del ítem se define como

$$dif = \frac{c}{n},$$

Donde dif es la dificultad del ítem, c el número de sustantivos que respondieron correctamente el ítem y n el total de sustentas.

La discriminación del ítem se obtiene así:

$$\rho = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu}_X)(Y_i - \hat{\mu}_Y)}{(n-1)\hat{\sigma}_X\hat{\sigma}_Y},$$

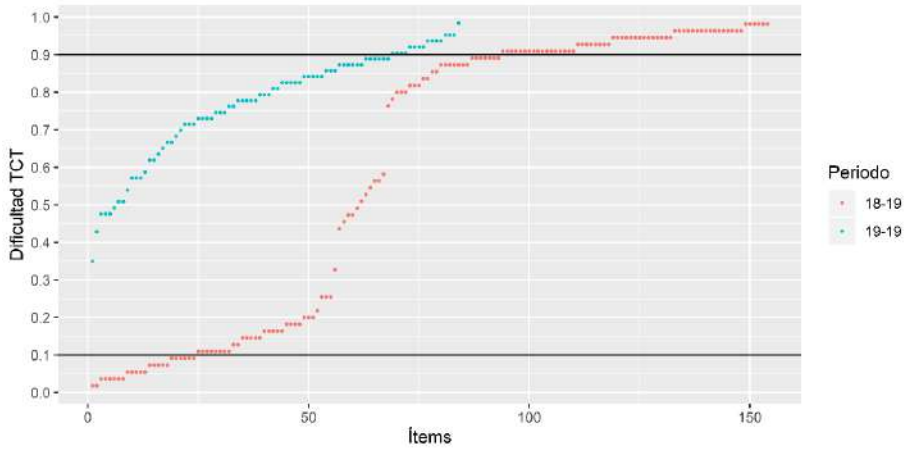
donde ρ es la correlación de Pearson (discriminación) y captura la capacidad del ítem para distinguir entre sustentantes de alto y de bajo rendimiento, n representa el total de sustentantes, X es la variable aleatoria de respuesta al ítem, codificadas con 1 si la respuesta es correcta y 0 si no, Y es la variable aleatoria del número de respuestas correctas en la prueba, $\hat{\mu}_.$ y $\hat{\sigma}_.$ son la media aritmética y desviación estándar de las variables aleatorias X y Y.

La fiabilidad del instrumento de evaluación se calcula del siguiente modo:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right),$$

donde $\hat{\alpha}$ es la fiabilidad o coeficiente alfa de Cronbach (ver Figura 2.3) y estudia la tendencia a la consistencia entre las mediciones repetidas de una prueba si un sustentante no recordara sus respuestas anteriores, k es el número de reactivos en la prueba, $\hat{\sigma}_i^2$ la varianza del puntaje en el reactivo i y $\hat{\sigma}_X^2$ es la varianza del puntaje total.

Figura 2.1: Dificultad de pruebas aplicadas en los periodos 2018-2019 y 2019-2019 a un conjunto de sustentantes de 4to EGB del Sistema Nacional de Educación del Ecuador. Las líneas horizontales representan los dictámenes TCT.



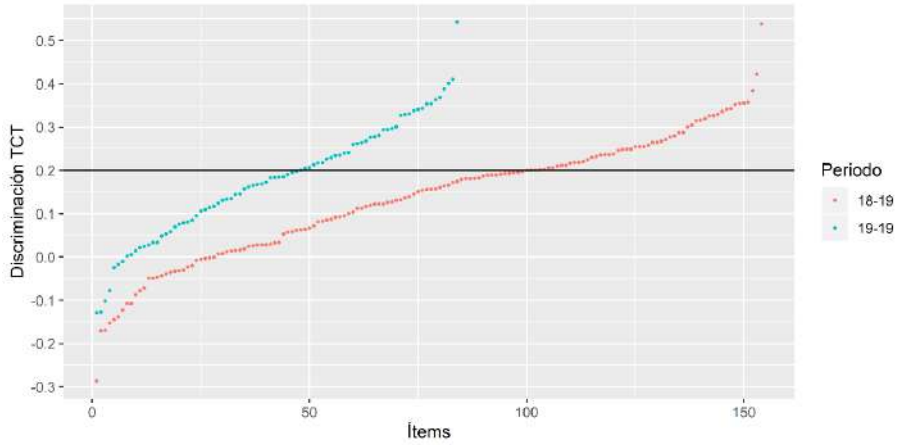
Fuente: Elaboración propia.

En la Figura 2.1 se puede observar la distribución de las dificultades de los ítems ordenados de menor a mayor dificultad, junto con los dictámenes[1] como líneas horizontales. En el eje de las abscisas se observa la cantidad de ítems y en el eje de las ordenadas el porcentaje de dificultad TCT que va de 0 a 100, siendo 0 muy difícil y 100 muy fácil.

Figura 2.2: Discriminación de pruebas aplicadas en los periodos 2018-2019 y 2019-2019 a un conjunto de sustentantes de 4to EGB del Sistema

[1] El dictamen es la evaluación de un ítem en dos categorías "aceptado" o "no aceptado". Se acepta un ítem cuando los valores de los parámetros del ítem caen dentro de los intervalos permitidos. En la gráfica los ítems aceptados caen entre las líneas horizontales de color negro.

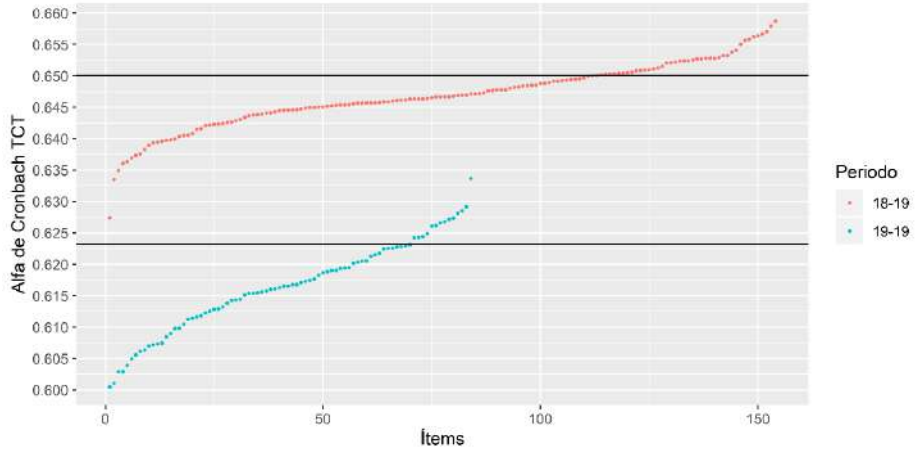
Nacional de Educación del Ecuador. La línea horizontal representa el dictamen TCT.



Fuente: Elaboración propia.

La Figura 2.2 muestra la discriminación de los ítems junto con el dictamen como línea horizontal (los valores que caen por encima de la línea horizontal son aceptados). La Figura 2.3 indica la tendencia a la consistencia entre las mediciones repetidas de una prueba si un sustentante no recordará sus respuestas anteriores.

Figura 2.3: Alfa de Cronbach de pruebas aplicadas en los periodos 2018-2019 y 2019-2019 a un conjunto de sustentantes de 4to EGB del Sistema Nacional de Educación del Ecuador. Las líneas horizontales representan los coeficientes de alfa Cronbach de las pruebas.



Fuente: Elaboración propia.

Las primeras estimaciones de confiabilidad de consistencia interna ampliamente utilizadas provienen del artículo de Kuder y Richardson (1937). Hoyt (1941) proporcionó una declaración más general de confiabilidad de consistencia interna. que fue popularizada por Cronbach (1951). Cronbach y Meehl (1955) presentaron la primera exposición detallada de la validez de construcción.

La Teoría Clásica de la Prueba, parte de cuatro supuestos principales a la luz de estos se puede tener una idea del alcance y también ciertas limitaciones de la teoría, así como también las "reglas principales" sobre las cuales se desarrolla esta teoría. A continuación, se enuncian estos supuestos:

1. El primer supuesto se puede enunciar de dos maneras equivalentes. De acuerdo a la primera manera de enunciar, se dice que la puntuación verdadera es la esperanza matemática de la puntuación observada, en donde el espacio muestral está formado por un gran conjunto de tomas sucesivas de evaluaciones equivalentes. Es decir, la puntuación verdadera sería el "promedio" de las puntuaciones observadas de un sustentante que rinde evaluaciones equivalentes un gran número de veces. Si se enuncia de esta manera, resulta que la esperanza matemática del error es nula. La otra manera de enunciar es partir del supuesto de que la esperanza matemática del error es nula, y, por tanto, esto resulta en que la puntuación verdadera es la esperanza matemática de la puntuación observada.
2. La correlación entre el error y la puntuación verdadera es cero. Es decir, que no existe ningún tipo de relación lineal entre el valor del error y el valor de la puntuación verdadera. Podemos escribir esto matemáticamente como: $cor(U, E) = 0$. Esto quiere decir que una puntuación verdadera mayor no implica necesariamente un error mayor ni viceversa; una

puntuación verdadera mayor tampoco implica necesariamente un error menor ni viceversa. Una puntuación verdadera menor no implica necesariamente un mayor error ni viceversa; y una puntuación verdadera menor no implica necesariamente un menor error ni viceversa.

3. Sean A y B dos pruebas administradas a los mismos estudiantes. Este supuesto indica que la correlación entre la puntuación verdadera en la Prueba A (U_A) y el error en la Prueba B (E) es cero. Es decir, no existe ningún tipo de relación lineal entre la puntuación verdadera en A y el error en B. Escribimos esto matemáticamente como: $cor(U_A, E_B) = 0$. Esto quiere decir que una puntuación verdadera mayor en A no implica necesariamente un error mayor en B ni viceversa; una puntuación verdadera mayor en A tampoco implica necesariamente un error menor en B ni viceversa. Una puntuación verdadera menor en A no implica necesariamente un mayor error en B ni viceversa; y una puntuación verdadera menor en A no implica necesariamente un menor error en B ni viceversa.
4. El último supuesto, indica que si A y B son dos Pruebas administradas a los mismos sustantivos la correlación entre los errores de la Prueba A (E_A) y LA Prueba B (E_B) es cero. Es decir, no existe ningún tipo de relación lineal entre el error en A y el error en B. Escribimos esto matemáticamente como: $cor(E_A, E_B) = 0$. Dicho en otras palabras: un error mayor en A no implica necesariamente un error mayor en B ni viceversa; además un error mayor en A tampoco implica necesariamente un error menor en B ni viceversa. Un error menor en A no implica necesariamente un error mayor en B ni viceversa; y un error menor en A no implica necesariamente un error menor en B ni viceversa.

2.3.1.2. Teoría de respuesta al ítem (TRI)

TRI enfoca el análisis estadístico de los datos de la prueba en las respuestas a los ítems. en lugar de en la puntuación total sumada de la prueba. que era la base de la Teoría Clásica del Test. La mayoría de los modelos y métodos que caen dentro de la TRI asumen que una o más variables no observadas (o latentes) subyacen a las respuestas a los ítems de prueba en el sentido de que la variación entre individuos en esas variables latentes explica la covariación observada entre las respuestas a los ítems (Green, 1954). En los modelos TRI. la relación entre la posición de los individuos en la(s) variable(s) latente(s) y las respuestas al ítem se describen mediante modelos estadísticos que describen la probabilidad de una respuesta al ítem en función de la(s) variable(s) latente(s).

Lord (1953) dio cuenta de las conjeturas en lo que se llama en la actualidad el modelo logístico de tres parámetros (3PL). y abogó por el uso de funciones de información en el desarrollo de pruebas.

Los trabajos de Lord (1952) y Lazarsfeld (1950) aclararon la naturaleza de los modelos TRI y enfatizaron la idea de que eran modelos de "variable latente".

El volumen de Lord y Novick (1968) colocó el modelo TRI de ojiva normal sobre una base teórica sólida como una integración de una variable latente que representa las diferencias individuales con el modelo. Por otro lado, Rasch (1960) desarrolló un modelo de respuesta al ítem conocido como modelo logístico de un parámetro de Rasch, llamado así por su único parámetro de ítem que refleja dificultad. El modelo de Rasch está estrechamente relacionado con el modelo logístico de dos parámetros de Birnbaum (1968). Seguidamente Rasch (1961, 1966, 1977) desarrolló una base filosófica por su modelo que estableció la familia de modelos

Rasch. En general. después de 1968. La investigación y el desarrollo de la teoría de la respuesta al ítem aumentaron extraordinariamente rápido y hasta la fecha TRI sigue siendo una de las áreas de investigación más activas en psicometría.

El modelo matemático de la **Teoría de la respuesta al ítem** plantea la presencia de una relación entre las puntuaciones en la variable latente, es decir, la habilidad de los sustantivos y la probabilidad de responder correctamente el ítem. Esta relación se describe en tres funciones matemáticas las cuales se conocen como Curvas Características del Ítem (CCI) (ver Figura 2.4). La ecuación siguiente justamente representa el modelo logístico de tres parámetros2:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1+e^{-a_i(\theta - b_i)}} \qquad i = 1, 2, \dots, k$$

donde $P_i(\theta)$ es la probabilidad de que un examinado con habilidad θ responda correctamente el ítem i, b_i es la dificultad, a_i la discriminación, c_i es la probabilidad de responder correctamente el ítem i por azar o pseudo-adivinación y e la constante neperiana. Además, tanto θ, a y b toman valores entre $-\infty$ y $+\infty$. El modelo de dos parámetros se forma cuando c_i es igual a 0 y el modelo de un parámetro cuando a_i es igual a 1 y c_i es 0.

Son de mucha importancia ciertas transformaciones de la escala de habilidad, por ejemplo, la Curva Característica de la Prueba (CCP), esta estima la cantidad de respuestas correctas que se espera cuando un sustentante tiene habilidad θ. En el mismo sentido, la Función de Información del Ítem (FII) es utilizada generalmente en la selección de reactivos para el desarrollo de una versión de prueba, confrontar pruebas y estudiar la precisión en la estimación de una habilidad específica, para estos mismos fines sirve la Función de Información de la Prueba (FIP).

[2] Siguiendo a Meneses et al., (2013) y Chávez y Saade (2009).

Como se mencionó anteriormente, una transformación importante de la escala de habilidad es la Curva Característica de la Prueba y estima el número de respuestas correctas esperado, para un sustentante con habilidad θ (ver Figura 2.5) y se define

$$\tau(\theta) = \sum_{i=1}^{k} P_i(\theta),$$

donde $\tau(\theta)$ es el puntaje verdadero, $P_i(\theta)$ la probabilidad de que un examinado con habilidad θ responda correctamente el reactivo i y k el número de reactivos.

Por otra parte, uno de los problemas típicos de las pruebas es la existencia de reactivos inadecuados para medir a determinados sustentantes, en este sentido, un reactivo proporciona más información sobre los sustentantes cuyo nivel de habilidad esté cerca de su dificultad, que de los sustentantes con una habilidad alejada de este punto. En este aspecto, la Función de Información del Ítem (ver Figura 2.6) permite, entre otras cosas, seleccionar reactivos para el ensamble de versiones, comparar pruebas y definir la precisión en la estimación de una habilidad específica, y se define

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad i = 1, 2, \dots, k,$$

Donde $I_i(\theta)$ es la información suministrada por el reactivo i en el nivel de habilidad θ, $P_i(\theta)$ la probabilidad de que un examinado con habilidad θ responda correctamente el reactivo, $Q_i(\theta) = 1 - P_i(\theta)$ y $P_i'(\theta)$ es la derivada de $P_i(\theta)$.

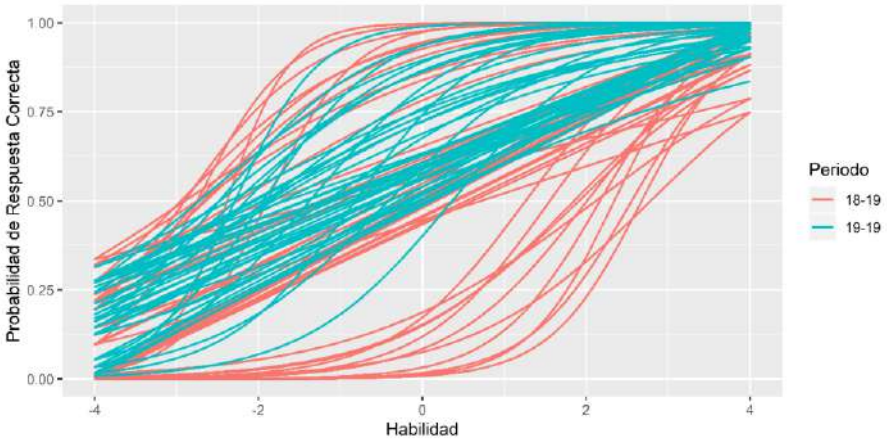
Es sustancial también en TRI, mostrar la precisión con la que se está midiendo la prueba en el rango de habilidad, para este

caso se usa la Función Información de la Prueba (ver Figura 2.7) que se define

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta),$$

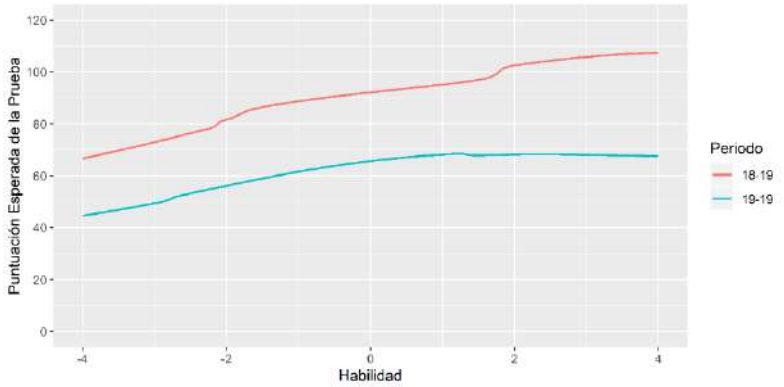
donde $I(\theta)$ es la cantidad de información de la prueba en el nivel de habilidad θ, $I_i(\theta)$ la información suministrada por el reactivo i en el nivel de habilidad θ y k es el número de reactivos en la prueba.

Figura 2.4: Curva Característica del Ítem: Probabilidad de Respuesta Correcta del Ítem según habilidad, periodos 2018-2019 y 2019-2019.



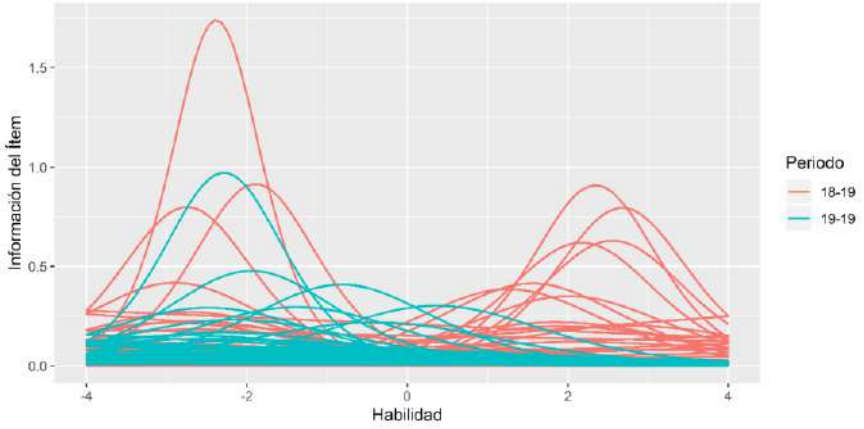
Fuente: Elaboración propia.

Figura 2.5: Curva Característica de la Prueba: Puntuación Esperada de la Prueba según habilidad, periodos 2018-2019 y 2019-2019.



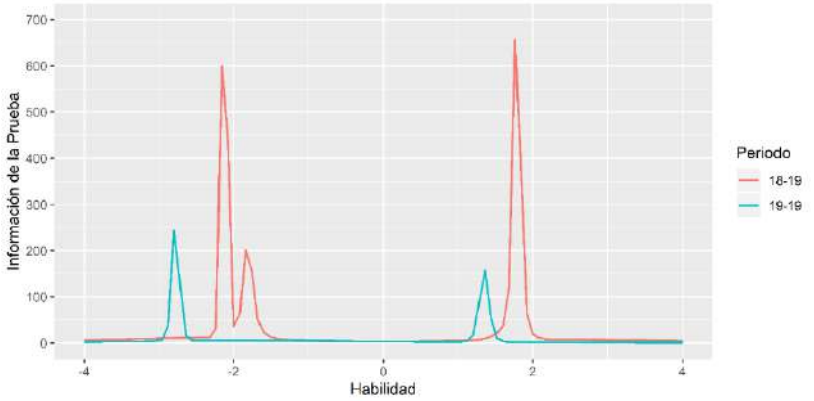
Fuente: Elaboración propia.

Figura 2.6: Función de Información del Ítem según habilidad, periodos 2018-2019 y 2019-2019.



Fuente: Elaboración propia.

Figura 2.7: Función de Información de la Prueba según habilidad, periodos 2018-2019 y 2019-2019.



Fuente: Elaboración propia.

El ajuste de los datos al modelo logístico se estudia generalmente mediante dos indicadores de ajuste: i) uno externo (outfit) y ii) otro interno (infit). El primero estudia la sensibilidad en relación con la cercanía del nivel de habilidad de la persona (Tristán-López, 1998) y, el segundo, estudia la sensibilidad de la conducta inesperada que afecta a los ítems que tienen un nivel de dificultad alejado de la habilidad de un sustentante. Los valores de outfit e infit entre 0.5 y 1.5 muestran que el ajuste de los datos al modelo logístico es óptimo (Linacre, 2003).

Para la Teoría de Respuesta al Ítem, se tiene tres supuestos, los cuales se enuncian a continuación:

1. El primer supuesto es que la probabilidad de que el sustentante acierte a un ítem depende de una llamada habilidad o rasgo latente θ y que esta no es observable directamente pero que se pretende medir con el ítem. La probabilidad mencionada anteriormente se describe mediante una función, $P(\theta)$, que depende de θ y de unos parámetros (mediciones de las características del ítem) que se desean tomar en cuenta. Dependiendo de la forma que tome la función P, se tienen los diferentes modelos resultantes. Las funciones más utilizadas son la función de distribución logística y la función de distribución normal.

2. Al segundo supuesto se lo denomina independencia local, lo cual consiste en que los ítems son independientes unos de otros. Es decir, la probabilidad de acertar correctamente a un ítem no depende de la probabilidad de contestar correctamente a otro de los ítems.
3. El tercer supuesto, denominado indeterminación de escala, indica que los parámetros de los ítems y las medidas de habilidades son únicos salvo una transformación lineal. Es decir, si se calculan los parámetros en base a dos poblaciones evaluadas distintas, los parámetros obtenidos y las mediadas de habilidades de ambos cálculos están relacionados mediante una transformación lineal.

2.3.1.3. Dictamen TCT y TRI de aceptación de ítems

Después de obtener la estimación de las dificultades, discriminación y pseudo-adivinación de los ítems. Se desarrolla una selección de ítems y se clasifican como "aceptados" o "descartados" dependiendo del valor de la dificultad y discriminación. Cuando los valores están dentro de los intervalos que se muestran en la Tabla 2.1. La confiabilidad generalmente se calcula a partir del coeficiente alfa de Cronbach en TCT y para el ajuste al modelo de dos parámetros en TRI se ocupan los indicadores infit y outfit.

Tabla 2.1: Dictamen TCT y TRI de aceptación de ítems.

TCT		TRI	
Dificultad	Discriminación	Dificultad	Discriminación
$(0.3, 0.7)$	$(0.1, 1)$	$(-3, +3)$	$(0.1, 2.8)$

Fuente: Elaboración propia.

2.3.2. Estadística y psicológica

En general desde el inicio de la medición psicológica los psicometristas han hecho contribuciones a la estadística

aplicada. Los avances en el escalamiento psicológico y la teoría de las pruebas tienen el requisito del conocimiento de los conceptos fundamentales de estadística. El trabajo diario de los psicometristas en los departamentos académicos de psicología ha incluido hasta el día de hoy la instrucción en cursos de estadística aplicada y se han producido varios libros adicionales para estudiantes universitarios.

Los psicometristas han estado especialmente interesados en proporcionar métodos computacionales para procedimientos estadísticos que son ampliamente útiles en la investigación psicológica. Appelbaum (1986) indica que quizás el tema de más larga duración en la historia de la psicometría involucró el cálculo de la correlación que forma la base de muchos enfoques para el análisis de ítems en la teoría de pruebas. A saber, la investigación en psicología produce datos multivariados en muchos contextos experimentales y de observación. en este sentido, se han desarrollado algoritmos basados en matrices para una serie de procedimientos estadísticos lineales multivariantes. Un trabajo estrechamente relacionado del grupo Gifi (1990) amplía el análisis computacionalmente intensivo de datos multivariados utilizando transformaciones no lineales de datos nominales, ordinales o continuos para proporcionar análogos no lineales de técnicas multivariadas como la regresión y la correlación canónica, y para el análisis de conglomerados. Existen otros temas de investigación que han captado el interés de los psicometristas en el transcurso de los años. Por lo tanto, la psicometría no ha sido simplemente un consumidor de las herramientas de la estadística, sino que ha sido uno de los campos que contribuyeron a la estadística aplicada a lo largo de su larga historia.

La psicometría o psicología cuantitativa. es el hogar disciplinario de un conjunto de modelos y métodos estadísticos que se han desarrollado principalmente para resumir. describir y extraer inferencias de datos empíricos recopilados en la investigación psicológica.

Capítulo 3 : Análisis Psicométrico de Pruebas Educativas

La evaluación educativa es la labor constante de estimar, apreciar y emitir juicios sobre procesos pedagógicos, administrativos o de sus resultados con la intención de mejorar continuamente (Ministerio de Educación Nacional, 1997). Ahora, para evaluar la educación es necesario contar con buenos instrumentos de medición. Chávez y Saade (2009) muestran que la evaluación educativa debe abarcar las fases de:

- Diseño.
- Construcción.
- Verificación.
- Ensamble.
- Aplicación y calificación.
- Mantenimiento de la base de datos de ítems.

Lord (1977), Baker y Kim (2017), Chávez y Saade (2009), CENEVAL (2012) y Lord (2012) muestran que la revisión cuantitativa o análisis estadístico (fase de verificación) de los ítems es importante, ya que es una forma de control de calidad en la elaboración del instrumento de evaluación. Además, abre una comunicación entre quienes elaboran los ítems, diseñan la prueba (es decir, eligen los ítems de la prueba) y quienes desarrollan el análisis estadístico de los datos que surgen de aplicar la prueba a los sustentantes. En este sentido, se muestra en este capítulo todas las fases necesarias para desarrollar una buena evaluación educativa.

3.1. Diseño y construcción de pruebas

Diseño y construcción de pruebas: Con el inicio de un proyecto de evaluación, se debe en primer lugar crear los insumos base de evaluación, tales como: marco de referencia, estructuras del proyecto y fichas técnicas, las cuales dan paso a la elaboración de ítems. Los ítems tienen que ser redactados de forma técnica y responder completamente al marco de referencia, estructura y fichas técnicas.

Un proyecto de evaluación puede tener más de una estructura de acuerdo con característica de la población, por ejemplo, si la evaluación está dirigida para personas con discapacidad, la estructura se adapta para este segmento de la población. Además, existen evaluaciones en las que es necesario contar con una estructura de acuerdo con el perfil del sustentante. De esta manera, se debe conocer las estructuras a evaluar, la cantidad de definiciones operacionales por estructura, y de ser el caso la cantidad de ítems por definición operacional de cada estructura. Por lo tanto, la **Estructura de Evaluación** es la base para la construcción del instrumento de evaluación, consta de elementos interrelacionados con una lógica jerárquica y en niveles que describen los contenidos a evaluar. Se compone de: campo, grupo temático, tópico, definición operacional, de ser necesario, acotamiento y la cantidad de ítems por definición operacional.

Por otro lado, una **Ficha Técnica** es el documento donde consta toda la información correspondiente al instrumento, campos a evaluar, número de ítems por campo, el largo de la prueba es decir de cuántos ítems consta el instrumento, el tiempo de duración el cual está dado por minutos de acuerdo con cada fase de verificación de los procesos de evaluación, contenidos temáticos, entre otros.

3.2. Verificación

Verificación: Después que se elaboran los ítems que formarán parte del instrumento de evaluación, estos deben

pasar por una revisión de forma o cualitativa y posteriormente por una revisión cuantitativa (ver Tabla 3.1) y desarrollar luego el proceso de armado de un instrumento de evaluación. Para la revisión cuantitativa se necesita estimar los parámetros de los ítems, por lo tanto, se debe seleccionar una muestra de sustentes y aplicar la prueba a este grupo este proceso se conoce como **Piloteo de reactivos**. Luego, se estiman los parámetros de los ítems en un proceso computacional que se conoce como **calibración**, para más tarde desarrollar el **dictamen** de reactivos. A continuación, se muestra cómo se lleva a cabo el proceso de calibración en la práctica.

Tabla 3.1: Actividades para la verificación de la calidad de los reactivos.

Cualitativa	Cuantitativa
Revisión técnica de reactivos	Piloteo de reactivos
Validación con especialistas	Calibración
Revisión de estilo	Dictamen

Fuente: Elaboración propia.

3.2.1. Calibración de reactivos

La calibración de reactivos se lleva a cabo para valorar al ítem mediante sus características psicométricas las cuales permiten seleccionar los reactivos que conformarán la prueba operativa. La calibración de los reactivos se realiza con base en la Teoría Clásica del Test y la Teoría de la Respuesta al Ítem con el modelo logístico de uno, dos o tres parámetros. En el análisis de los reactivos, con base en la TCT, se atienden dos indicadores:

- **Grado de dificultad:** Se refiere al porcentaje de personas que responden correctamente un reactivo de una prueba. Entre mayor sea esta proporción, menor será su dificultad. Los valores que se consideran aceptables en este parámetro son aquellos que están dentro del intervalo $(0.3, 0.7)$. Por otro lado, la dificultad es una información básica al ensamblar el instrumento

de evaluación (elección de ítems que formarán la prueba), ya que debe estar balanceado en cuanto a la dificultad global y en cada una de sus áreas o secciones.

- **Correlación punto biserial o coeficiente de discriminación:** Se calcula para determinar el grado en que cada reactivo aporta información certera en cuanto a las habilidades que mide de manera particular, y en conjunto con los demás ítems que conforman la prueba. El índice de correlación permite relacionar la tendencia de respuesta de cada ítem con respecto a la escala de la cual forma parte. Se supone que un sustentante con una puntuación alta en toda la prueba tiene mayores probabilidades de contestar correctamente un reactivo. También se debe esperar lo contrario, es decir, que quien tuvo bajas puntuaciones en la prueba, tenga pocas probabilidades de contestar correctamente el reactivo. Así, un buen reactivo debe discriminar entre aquellos que obtuvieron buenas calificaciones en la prueba y aquellos que obtuvieron bajas calificaciones. Generalmente, basta con que la correlación punto biserial de un reactivo sea positiva para aceptarla como medida de "discriminación" del reactivo. Sin embargo, es deseable que el valor de la discriminación del ítem este dentro del intervalo (0.1,1) para considerar que el ítem es útil.

En el mismo sentido, se calibran los reactivos mediante TRI, la diferencia entre la TCT y la TRI reside en que esta última se centra en las propiedades de los reactivos individuales y no en las propiedades globales de la prueba. La herramienta de la TRI que se utiliza para valorar la calidad de los reactivos es la Curva Característica del Ítem, la cual resulta del ajuste de una función matemática al comportamiento del reactivo, y representa la probabilidad de que un sujeto, con una determinada habilidad, responda correctamente el reactivo.

Para el análisis de los reactivos de una prueba se puede ocupar los modelos de uno, dos o tres parámetros, en este documento describimos a continuación los elementos del modelo de tres parámetros. Los parámetros de este modelo hacen alusión a los índices de discriminación, dificultad y pseudo-adivinación de los reactivos:

- **Índice de discriminación** ($\boldsymbol{a}$)**:** Indica la cualidad que tiene el reactivo de diferenciar a los sustentantes que dominan el conocimiento de aquellos que no. Para considerar aceptables a los reactivos la discriminación debe encontrarse en el intervalo $(0.1,2.8)$.
- **Índice de dificultad** ($\boldsymbol{b}$)**:** Indica la posición del ítem en la escala de aptitud. Cuanto más grande es el valor de dificultad, mayor es la aptitud requerida para que el examinado tenga una probabilidad alta de resolver correctamente el ítem. Lo ideal es que los reactivos se encuentren en el intervalo $(-3, 3)$.
- **Índice de pseudo-adivinación** ($\boldsymbol{c}$)**:** Permite conocer la probabilidad de que cada reactivo sea contestado correctamente al azar.

Después de la calibración es posible elegir los reactivos con parámetros óptimos que conformarán las versiones operativas finales de la prueba, es decir, es factible realizar el ensamble del instrumento de evaluación. Los ítems no escogidos en el diseño y sólo aquellos que cuentan con buenas propiedades psicométricas se deben mantener en el banco de reactivos, para posteriores aplicaciones o para presentarse en materiales de apoyo a los sustentantes.

Ahora, es importante indicar que una **Base de ítems calibrados** son el conjunto de ítems calibrados con información estadística de dificultad, discriminación y/o pseudo-adivinación de los ítems que ya han sido aplicados. Esta base se utiliza en la realización del ensamble cuantitativo para distinguir los ítems con buenos parámetros psicométricos y los que no. De esta manera se escogen los ítems cuyos

parámetros son aceptables y los mejores para que formen parte del instrumento.

3.3. Ensamble

El ensamble de la prueba debe pasar por dos etapas, la primera establece la teoría psicométrica y las medidas objetivo de los parámetros psicométricos que debe cumplir los ítems y la prueba, mientras que la segunda etapa es donde se selecciona los ítems de la prueba operativa final y se verifica que las dificultades este bien repartidas en su rango de variación. En detalle se describe a continuación los pasos de las dos etapas del diseño de una prueba:

La primera etapa considera:

- La determinación de la teoría psicométrica a utilizar, esto es la elección de TCT o TRI.
- Selección de ítems elegibles para la construcción de la prueba en base a los dictámenes TCT o TRI según sea el caso.
- Establecimiento de una distribución equilibrada del número de ítems según campo (por ejemplo: matemática, lengua y literatura, etc.) y tipo de ítem (por ejemplo: longitudinal, sesión, individual y día) conforme a la Tabla 3.2, donde n representa el número de ítems correspondiente a la tabla de contingencia y $n_{\cdot}$, $n_{\cdot}$ y $n_{\cdot\cdot}$ son los totales por columna, fila y suma general según corresponda.

Tabla 3.2: Distribución del número de ítems según Campo (C) y Tipo (T), TCT y TRI.

Campo	$\boldsymbol{T_1}$	$\cdots$	$\boldsymbol{T_q}$	**Total**
$\boldsymbol{C_1}$	$n_{1,1}$	$\cdots$	$n_{1,q}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$

$\boldsymbol{C_p}$	$n_{p,1}$	$\cdots$	$n_{p,q}$	$n_{p.}$
Total	$n_{.1}$	$\cdots$	$n_{.q}$	$n_{..}$

Fuente: Elaboración propia.

- Tomado como base el punto de corte (θ^e), se realiza la definición de una distribución equilibrada de la dificultada de los ítems, es decir, se establece el valor de la media aritmética y desviación estándar de la dificultad cuando se trabaja con TCT (ver Tabla 3.3). Por otro lado, si se utiliza TRI se determina la distribución según campo y tipo de reactivo, del promedio geométrico de la probabilidad de responder un ítem cuando el sustentante tiene habilidad $\theta = \theta^e$ (ver Tabla 3.4). Siguiendo con TRI, se determina la distribución según habilidad de la curva de información de la Prueba (ver Tabla 3.5).

Tabla 3.3: Distribución de la dificultad media según Campo (C) y Tipo (T), TCT.

Campo	$\boldsymbol{T_1}$	$\cdots$	$\boldsymbol{T_q}$	**Promedio**
$\boldsymbol{C_1}$	$\lvert\theta^e - (-3)\rvert/6$	$\cdots$	$\lvert\theta^e - (-3)\rvert/6$	$\lvert\theta^e - (-3)\rvert/6$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$\boldsymbol{C_p}$	$\lvert\theta^e - (-3)\rvert/6$	$\cdots$	$\lvert\theta^e - (-3)\rvert/6$	$\lvert\theta^e - (-3)\rvert/6$
Promedio	$\lvert\theta^e - (-3)\rvert/6$	$\cdots$	$\lvert\theta^e - (-3)\rvert/6$	$\lvert\theta^e - (-3)\rvert/6$

Fuente: Elaboración propia.

Tabla 3.4: Distribución de probabilidad media con habilidad $\theta = \theta^e$, según dificultad y discriminación, TRI.

Campo	$\boldsymbol{T_1}$	$\cdots$	$\boldsymbol{T_q}$	**Promedio**
$\boldsymbol{C_1}$	$\lvert\theta^e - (-3)\rvert/6$	$\cdots$	$\lvert\theta^e - (-3)\rvert/6$	$\lvert\theta^e - (-3)\rvert/6$

⋮	⋮	⋱	⋮	⋮
$\boldsymbol{C_p}$	$\lvert\theta^e-(-3)\rvert/6$	⋯	$\lvert\theta^e-(-3)\rvert/6$	$\lvert\theta^e-(-3)\rvert/6$
Promedio	$\lvert\theta^e-(-3)\rvert/6$	⋯	$\lvert\theta^e-(-3)\rvert/6$	$\lvert\theta^e-(-3)\rvert/6$

Fuente: Elaboración propia.

Tabla 3.5: Distribución de Información según habilidad, TRI.

Habilidad	**Información**
⋮	⋮
$\boldsymbol{\theta^e-0.2}$	$I_{\theta^e-0.2}$
$\boldsymbol{\theta^e-0.1}$	$I_{\theta^e-0.1}$
$\boldsymbol{\theta^e}$	$max\left(I_{(-\infty,+\infty)}\right) = I_{\theta^e}$
$\boldsymbol{\theta^e+0.1}$	$I_{\theta^e+0.1}$
$\boldsymbol{\theta^e+0.2}$	$I_{\theta^e+0.2}$
⋮	⋮

Fuente: Elaboración propia.

La segunda etapa considera:

- La selección cuidadosa de ítems para la prueba operativa final, de forma que se cumpla en el caso de utilizar TCT, las distribuciones del número de ítems y la distribución de dificultad. Para el caso de TRI debe cumplirse la distribución del número de ítems y forma de la curva de información o la distribución de la probabilidad de responder un ítem con habilidad cero según convenga.
- En el caso TCT, la comprobación de que haya presencia de ítems con dificultades que varíen en todo el rango de valores de la dificultad (de 0.3 a 0.7). Del mismo modo en TRI, se comprueba la presencia de ítems que tengan

dificultades cercanas al punto de corte definido para el instrumento de evaluación.

La primera etapa del ensamble de una prueba responde a elecciones arbitrarias o técnicas que solo depende de los administradores de la evaluación educativa, no así la segunda, donde la selección cuidadosa de ítems cambia dependiendo si se está trabajando con TCT o TRI. En el caso de TRI también depende de si estamos enfocados en la probabilidad de responder un ítem con habilidad cero o en la curva de información de la prueba. En síntesis, si estamos trabajando con TCT o TRI, y cuando en TRI nos importa la probabilidad de responder un ítem con habilidad cero se elige los ítems de modo que cumplan con la distribución según campo y tipo de reactivo previamente especificado. En cambio, en TRI cuando nos importa la curva de información de la prueba siguiendo a Lord (1977), el procedimiento de selección de ítems para una prueba, sugerido por primera vez por Birnbaum (1968), es mediante la función de información del ítem. Específicamente se persiguen los siguientes pasos:

1. Decidir la forma deseada para la función de información de la prueba, recordando que esta función de información es inversamente proporcional a la longitud al cuadrado del intervalo de confianza asintótico para estimar la habilidad a partir del puntaje de la prueba. ¿Qué precisión de estimación de habilidad se requiere de la prueba en cada nivel de habilidad? La curva deseada es la curva de información objetivo.
2. Seleccionar ítems con curvas de información que colmen las áreas difíciles de llenar bajo la curva de información objetivo.
3. Sumar acumulativamente las curvas de información de los ítems, obteniendo en todo momento la curva de información para la prueba parcial compuesta por los ítems ya seleccionados.

4. Continuar (retroceder si es necesario) hasta que el área bajo la curva de información objetivo se llene hasta una aproximación satisfactoria.

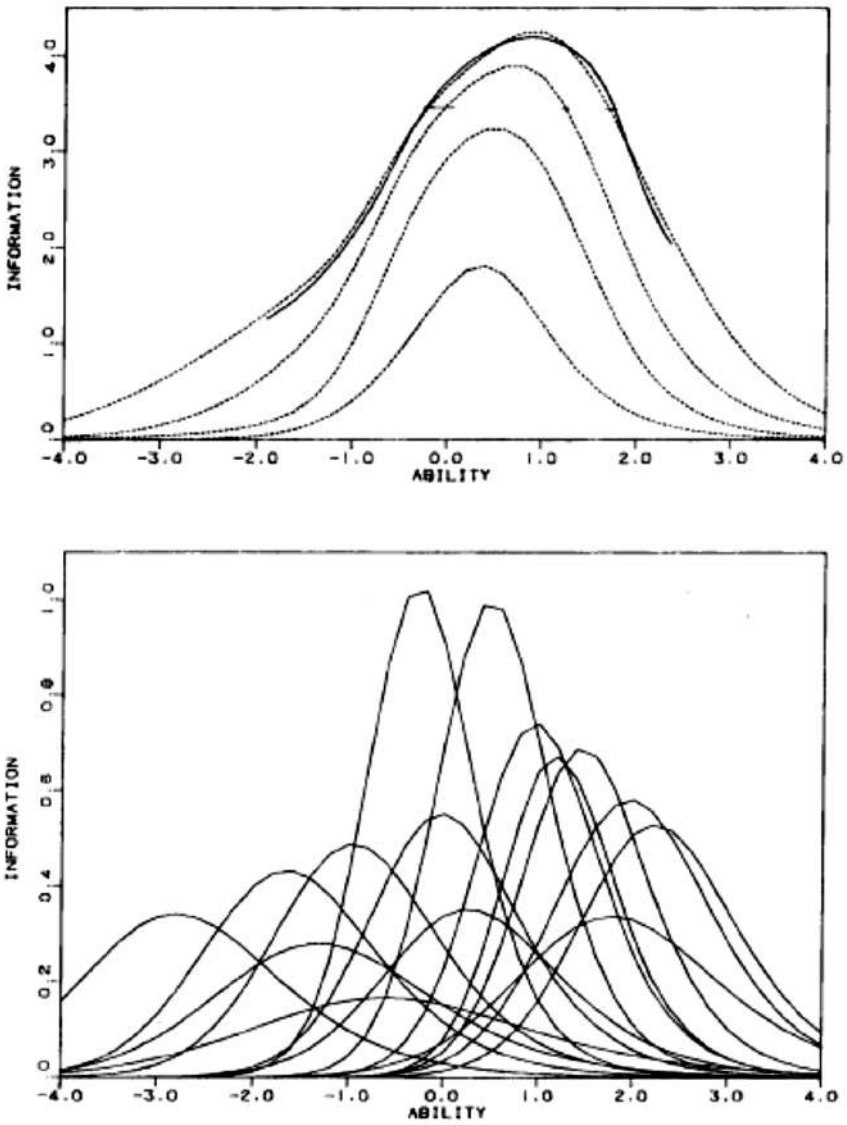
La Figura 3.1, tomada de Lord (1977) ilustra cómo se puede construir una prueba para aproximar una curva de información objetivo hipotético (línea gruesa). Las curvas de información de quince ítems se muestran en la parte inferior. Los tres elementos intermedios se seleccionan primero. Después de esto, los elementos se agregan de a poco, procediendo hacia afuera desde el medio. Las curvas de información de la prueba parcial de muestra para 3, 7 y 11 ítems. La curva de información final para la prueba de 15 ítems se aproxima a la curva de información objetivo.

3.3.1. Importancia de utilizar la función de información

La función de información es importante en el diseño de una prueba, ya que es el recíproco de la varianza con la que se puede estimar un parámetro de habilidad. Por lo tanto, si pudiera estimar un parámetro con precisión (es decir, una variabilidad menor), sabría más sobre el valor del parámetro que si lo hubiera estimado con menos precisión (es decir, una variabilidad mayor). Estadísticamente, la magnitud de precisión con la que se estima un parámetro está inversamente relacionada con el tamaño de la variabilidad de las estimaciones en torno al valor del parámetro. De esta manera la función de información de la prueba es una característica extremadamente útil de la teoría de respuesta al ítem. Básicamente le dice qué tan bien le está yendo a la prueba al estimar la habilidad en todo el rango de puntajes de habilidad. Si bien la función de información de prueba ideal a menudo puede ser una línea horizontal, puede no ser la mejor para un propósito específico. Por ejemplo, si estaba interesado en

construir una prueba para otorgar becas, este ideal podría no ser óptimo. En esta situación, le gustaría medir la habilidad con una precisión considerable en niveles de habilidad cercanos a la habilidad utilizada para separar a aquellos que recibirán la beca de aquellos que no. La mejor función de información de prueba en este caso tendría un pico en la puntuación de corte. Otros usos especializados de las pruebas pueden requerir otras formas de la función de información de la prueba. Es así como, considerando algunos de los objetivos de prueba típicos puede haber pruebas de detección, de amplio rango y máximas que se detallan a continuación:

Figura 3.1: Arriba: Curva de información objetivo (sólido) y curvas de información de subprueba (n = 3, 7, 11, 15). Abajo: Curvas de información para 15 ítems utilizados para aproximar la curva objetivo.



Fuente: Lord, 1977, p. 121.

- **Pruebas de detección:** Las pruebas utilizadas con fines de detección tienen la capacidad de distinguir con bastante claridad entre los examinados cuyas habilidades están justo por debajo de un nivel de habilidad dado y aquellos que están en o por encima de ese nivel. Dichas pruebas se utilizan para asignar becas

y para asignar estudiantes a programas de instrucción específicos, como remediación o colocación avanzada.

- **Pruebas de amplio rango:** Estas pruebas se utilizan para medir la habilidad en un amplio rango de escala de habilidad subyacente. El objetivo principal es poder hacer una declaración sobre la capacidad de un examinado y hacer comparaciones entre los examinados. Las pruebas que miden la lectura o las matemáticas suelen ser pruebas de amplio rango.
- **Pruebas máximas:** Dichas pruebas están diseñadas para medir la habilidad bastante bien en una región de la escala de habilidad donde se ubicará la mayoría de las habilidades de los examinados, y menos bien fuera de esta región. Cuando uno crea deliberadamente una prueba de pico, es medir bien la habilidad en un rango de habilidad que es más amplio que el de una prueba de detección, pero no tan amplio como el de una prueba de rango amplio.

3.3.2. Consideraciones al usar la función de información

Al utilizar la función de información en la construcción de una prueba es importante tener en cuenta las siguientes cosas:

1. El nivel general de la función de información de prueba depende de:
   a) El número de elementos en la prueba.
   b) El valor promedio de los parámetros de discriminación de los ítems de prueba.
   c) Los dos anteriores se mantienen para los tres modelos de curvas características[3].

[3] Modelo logístico de un parámetro (Modelo de Rasch), Modelo logístico de dos parámetros y de tres parámetros. Donde los parámetros son: la dificultad (b), discriminación (a) y pseudo-adivinación (c) de los ítems.

2. La forma de la función de información de prueba depende de:
   a) La distribución de las dificultades del ítem sobre la escala de habilidad.
   b) La distribución y el valor promedio de los parámetros de discriminación de los ítems de prueba.
3. Cuando las dificultades del ítem se agrupan en torno a un valor dado, la función de información de prueba alcanza su punto máximo en ese punto de la escala de habilidad. La cantidad máxima de información depende de los valores de los parámetros de discriminación.
4. Cuando las dificultades del ítem se distribuyen ampliamente en la escala de habilidad, la función de información de la prueba tiende a ser más plana que cuando las dificultades están estrechamente agrupadas.
5. Los valores de ($a < 1.0$) dan como resultado un nivel general bajo de la cantidad de información de prueba. Los valores de ($a > 1.7$) dan como resultado un alto nivel general de la cantidad de información de prueba.
6. Según el modelo de tres parámetros, los valores del parámetro de adivinanza *c* mayor que cero disminuyen la cantidad de información de prueba en los niveles de habilidad bajos. Además, los valores grandes de c reducen el nivel general de la cantidad de información de prueba.
7. Es difícil aproximar una función de información de prueba horizontal. Para hacerlo, los valores de *b* deben extenderse ampliamente sobre la escala de habilidad y los valores de *a* deben estar en el rango moderado a bajo y tener una distribución en forma de U.
8. No debe existir dos ítems en el banco de ítems que posean exactamente la misma combinación de valores de parámetros de elementos.
9. Los valores de los parámetros del artículo están sujetos a las siguientes restricciones:

$-3.0 <= b <= +3.0,$

$0.1 <= a <= +2.8,$

$0 <= c <= 0.35,$

donde:

b: parámetro de dificultad del ítem.

a: parámetro de discriminación del ítem.

c: parámetro de pseudo-adivinación.

Los valores del parámetro de discriminación se han restringido para reflejar el rango de valores que generalmente se ven en grupos de ítems bien mantenidos.

En este punto es relevante plantear algunas cosas para tener en cuenta, según el tipo de prueba a desarrollar:

1. **Pruebas de detección:**
   a) La curva característica de prueba deseada tiene el puntaje verdadero medio en el nivel de habilidad de corte especificado. La curva debe ser lo más empinada posible en ese nivel de habilidad.
   b) La función de información de prueba debe alcanzar su máximo con el nivel de habilidad de corte.
   c) Los valores de los parámetros de dificultad del elemento deben agruparse lo más cerca posible de la capacidad de corte de interés. El caso óptimo es cuando todos los valores de los parámetros de dificultad del elemento están en el punto de corte y los valores de los parámetros de discriminación del elemento son grandes. Sin embargo, esto no es realista porque un grupo de elementos rara vez contiene suficientes elementos con valores de

dificultad comunes. Si se debe elegir entre los elementos, seleccione los elementos que produzcan la cantidad máxima de información en el punto de corte.

2. Pruebas de amplio rango

a) La curva característica de prueba deseada tiene su puntaje verdadero medio en un nivel de habilidad correspondiente al punto medio del rango de habilidad de interés. Muy a menudo, este es un nivel de habilidad de cero. La curva característica de prueba debe ser lineal para la mayor parte de su rango.
b) La función de información de prueba deseada es horizontal en el rango más amplio posible. La cantidad máxima de información debe ser lo más grande posible.
c) Los valores de los parámetros de dificultad del ítem deben distribuirse uniformemente sobre la escala de habilidades y tan ampliamente como sea práctico. Existe un conflicto entre los objetivos de una cantidad máxima de información y una función de información de prueba horizontal. Para lograr una función de información de prueba horizontal, se necesitan elementos con parámetros de discriminación de elementos bajos a moderados que tengan una distribución en forma de U de los parámetros de dificultad del elemento. Sin embargo, tales elementos producen una cantidad general bastante baja de información y la precisión general será baja.

3. Pruebas máximas

a) La curva característica de prueba deseada tiene su puntaje verdadero medio en un nivel de habilidad en el medio del rango de interés de la habilidad. La

curva debe tener una pendiente moderada a ese nivel de habilidad.

b) La función de información de prueba deseada debe tener su máximo en el mismo nivel de habilidad que el puntaje verdadero medio de la curva característica de prueba. La función de información de prueba debe redondearse en apariencia sobre el rango de habilidades de mayor interés.
c) Los parámetros de dificultad del ítem deben agruparse alrededor del punto medio del rango de habilidad de interés, pero no tan estrictamente como en el caso de una prueba de detección. Los valores de los parámetros de discriminación deben ser tan grandes como sea práctico. Los ítems cuyos valores de los parámetros de dificultad del ítem están dentro del rango de habilidad de interés deberían tener valores mayores de los parámetros de discriminación de ítem que los ítems cuyos valores de los parámetros de dificultad del ítem están fueran de este.

Es recomendable que se muestre los datos psicométricos generales de las pruebas piloto, según la Tabla 3.6 para la posterior elección de las medidas estadísticas deseadas de las versiones operativas.

Tabla 3.6: Datos psicométricos importantes de los reactivos pilotados.

Estadísticas	Medidas
Reactivos elaborados	
Reactivos validados	
Reactivos pilotados	
Dificultad media de TCT	
Correlación biserial media de TCT	
Coeficiente de discriminación de TRI	
Coeficiente de dificultad de TRI	
Reactivos con parámetros óptimos	

Fuente: Elaboración propia.

3.3.3. Establecimiento del punto de corte para pruebas de detección

El **Punto de Corte** es un valor de la escala de habilidad, $\theta \in (-\infty, +\infty)$ y se refiere a la puntuación más baja posible en una prueba estandarizada que un estudiante debe tener para ser considerado como apto, es decir, sirve para clasificar a los sujetos en dos categorías que suponen diferentes niveles de competencia en relación con el dominio, normalmente se clasifica en apto y no apto.

El hecho de que el resultado obtenido por un sustentante se considere insuficiente, elemental, bueno o excelente depende en gran parte de las expectativas que tengamos. Por eso la determinación de los puntos de corte es un proceso delicado y sujeto siempre a discusión. La mayor parte de los procedimientos consisten en acuerdos entre expertos que deben justificar sus decisiones. Una vez justificadas esas decisiones es posible hacer un juicio que se base en evidencia empírica acerca de la congruencia de las decisiones adoptadas. Observando la Tabla 3.7 y si aprendiéramos que nuestras estimaciones tengan mayor precisión para los sujetos de capacidad relativamente alta, es decir, los que se encuentran en los niveles 'Bueno' y 'Excelente' es importantes distribuir las dificultades de los reactivos, de modo que se incluyan más reactivos que midan adecuadamente en la parte superior de la distribución para poder captar adecuadamente los cambios que puedan darse, sin embargo, si en la población evaluada la mayor parte de los individuos se encuentra en los niveles 'Insuficiente' y 'Elemental' no se los podría medir de manera adecuada. Por eso es mejor escoger reactivos que tengan su nivel de dificultad en los niveles 'Elemental' y 'Bueno'.

Tabla 3.7: Rangos de habilidad generales.

Nivel de Dominio	Puntos de corte (habilidad)
Insuficiente	$(-\infty, -1.5]$
Elemental	$(-1.5, 0.0]$
Bueno	$(0.0, 1.5]$
Excelente	$(1.5, \infty)$

Fuente: Elaboración propia.

Dentro de los métodos para el establecimiento del punto de corte, se tiene el método centrado en el test y el centrado en las personas. El método centrado en el test se basa en opiniones de los jueces sobre los ítems del test. Por ejemplo, si se quiere evaluar la habilidad de los sustentantes en psicometría, los jueces responderían a la pregunta ¿una persona que sea mínimamente competente en psicometría qué conocimientos demostrará? El método centrado en los sustentantes se basa en el rendimiento de los sustentantes. A continuación, se describe de manera más amplia estos métodos:

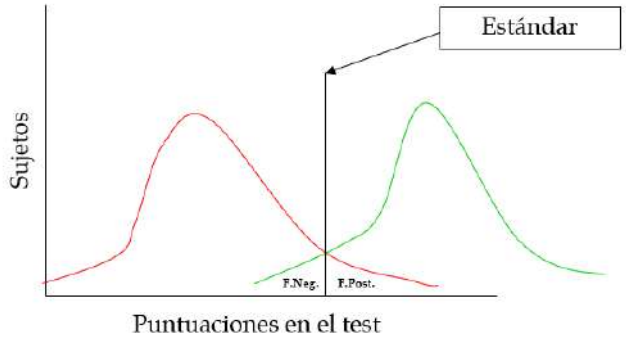
Método centrado en el test: Método de Angoff (1971) es el método más utilizado, investigado y recomendado que consta de los siguientes pasos:

- Identificar una población de jueces y seleccionar una muestra.
- Cada juez debe definir qué significa para él la competencia mínima; y consensuar con el resto de los jueces.
- Considerar cada ítem del test y decidir para cada uno de ellos la probabilidad de que un examinado mínimamente competente responda al ítem correctamente (estimación a-priori de la dificultad).
- Para obtener el punto de corte se suman todas las probabilidades y se promedian para todos los jueces.

Método centrado en las personas: Método de los grupos de contraste (Zieky y Livingston, 1977), en este los juicios se basan en el rendimiento de los sujetos examinados.

- Identificar una población de jueces y seleccionar una muestra. Es imprescindible que sean capaces de juzgar el nivel de rendimiento de los sujetos por las puntuaciones del test.
- Se pide a los jueces que definan tres categorías: competente; límite; e inadecuado o incompetente.
- Los jueces evalúan a los examinados, y basándose en otras informaciones calificarían al grupo de los "límites".
- Los sujetos realizan el test; y los estándares se establecen en función del rendimiento de "competentes" e "incompetentes".
- Se establece como punto de corte, la puntuación que mejor discrimina entre los dos grupos.

Figura 3.2: Punto de corte según método de los grupos de contraste.



Fuente: Chacón y Sanduvete, s.f, p.21.

Observando la Figura 3.2 se tiene que el punto de corte viene dado por la intersección entre ambas distribuciones, es decir, la de los sujetos competentes e incompetentes. Ya que iguala los dos tipos de errores, es decir aptos |verdaderos que no pasan el test, y no aptos que sí pasan el test. Desplazar el punto

de corte a la derecha implica primar la presencia de falsos negativos y desplazar a la izquierda la de falsos positivos. Si, por ejemplo, el test pretende seleccionar a los sujetos encargados de apretar el botón nuclear qué duda cabe que habrá que ser lo más conservador posible y desplazar el punto de corte lo más a la derecha posible para asegurarnos de que no haya ningún falso positivo, es decir, un sujeto que habiendo pasado el punto de corte no sea competente.

3.3.4. Ensamble de la Prueba

En esta fase se integra la versión operativa de la prueba con los reactivos que resultaron ser los mejores en las revisiones cualitativas, cuantitativas y aquellos que fueron seleccionados en el diseño de la prueba. La validez de las interpretaciones de la calificación final depende en gran parte de que el ensamble se realice de manera adecuada. La disposición de los reactivos debe obedecer al plan de prueba que se estableció desde el inicio; así, las versiones son congruentes con los contenidos y niveles taxonómicos necesarios para contar con una muestra representativa del dominio de conocimiento que se va a evaluar.

El ensamble permite que un instrumento mantenga consistencia en sus contenidos a lo largo del tiempo manteniendo una métrica supervisada y controlada a través de parámetros probados y sustentados matemáticamente. Para ello se utiliza la Teoría Clásica del Test y la Teoría de Respuesta al Ítem.

Otras cosas que han de tomarse en cuenta para el ensamble son los aspectos de la administración de la prueba, como el número de reactivos, contenidos temáticos, número de sesiones, tiempo de aplicación.

En este sentido, se muestra a continuación la forma de realizar un ensamble de una prueba (selección de reactivos) de forma práctica. Los ítems que se utilizan en este punto son únicamente los ítems que fueron aceptados en el proceso de validación cualitativa. La versión piloto de la prueba es importante, para que la evaluación operativa final se conforme con reactivos que han probado su efectividad para brindar información en cuanto a las habilidades que posee la población objetivo.

3.3.5. Desarrollo de un ensamble de una Prueba

Cuando pretendemos realizar un ensamble de una prueba la primera información importante es conocer los campos de evaluación. En este sentido, en las líneas que vienen se describe el desarrollo de un ensamble de prueba cuando se pretende evaluar los campos de i) Razonamiento verbal, ii) Razonamiento numérico y iii) Razonamiento abstracto. El número de ítems de la prueba es de 31, donde 16 son ítems ancla, es decir, estos permanecerán constantes en las diferentes formas de la prueba. Una forma de prueba es una acomodación de ítems, donde los ítems catalogados como ancla son fijos, mientras que los ítems restantes (llamados operativos) pueden ser reemplazados por otros ítems. Para este ensamble de prueba, se cuenta con un banco de 60 ítems calibrados y el objetivo es que la probabilidad geométrica utilizando TRI de un parámetro sea de 0.50, es decir, se necesita que un sustentante tomado al azar con una habilidad, $\theta = 0$, tenga una esperanza matemática de responder correctamente la mitad de los ítems. Por lo tanto, el diseño requerido se desarrolló mediante la teoría de respuesta al ítem.

Ahora, a continuación, se muestra: i) la distribución de ítems por campo y tipo de ítem; ii) la media geométrica de la probabilidad de responder un ítem cuando el sustentante tiene

una habilidad, $\theta = 0$; iii) curva característica de la prueba; iv) función de información de la prueba; v) esperanza de aciertos según la habilidad de los sustentantes y vi) la información según habilidad.

Tabla 3.8: Distribución de Ítems.

Campo	Ancla	Operativo	Total Ítems
1. Razonamiento verbal	5	5	10
2. Razonamiento numérico	6	6	12
3. Razonamiento abstracto	5	4	9
Total	16	15	31

Fuente: Elaboración propia.

Tabla 3.9: Media geométrica, probabilidad TRI (1 parámetro).

Campo	Ancla	Operativo	Total Ítems
1. Razonamiento verbal	0.49	0.50	0.50
2. Razonamiento numérico	0.50	0.50	0.50
3. Razonamiento abstracto	0.51	0.50	0.50
Total	0.50	0.50	0.50

Fuente: Elaboración propia.

Tabla 3.10: Media aritmética, dificultad TCT.

Campo	Ancla	Operativo	Total Ítems
1. Razonamiento verbal	0.50	0.49	0.50
2. Razonamiento numérico	0.51	0.50	0.51
3. Razonamiento abstracto	0.49	0.52	0.51
Total	0.50	0.50	0.51

Fuente: Elaboración propia.

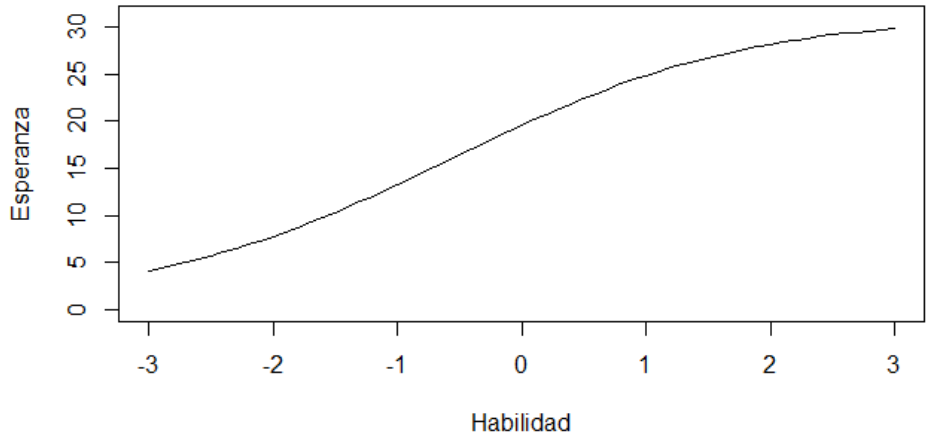
Las propiedades del ensamble se presentan en las Tablas 3.8, 3.9 y 3.10, la primera tabla muestra la distribución de ítems según campo y tipo de ítem. La segunda tabla está enfocada en TRI y muestra la probabilidad media de responder un ítem cuando el sustentante tiene habilidad cero, por ejemplo, para los ítems ancla del campo razonamiento verbal existe una probabilidad media de 0.49 de responder cualquier ítem, en contraste la probabilidad para los ítems operativos del campo

razonamiento numérico es de 0.50. En general, la probabilidad media de responder un ítem en la prueba es de 0.50, es decir, un sustentante con habilidad cero, tiene una esperanza aproximada de responder correctamente 16 ítems de los 31 de la prueba. La tercera tabla, enfocada en TCT indica el porcentaje medio de respuestas correctas según campo y tipo de ítem, de esta manera, se espera que el porcentaje medio de respuestas correctas sea de 51.00% para el instrumento de evaluación.

Retomando el enfoque TRI, se expone a continuación en las Figuras 3.3 y 3.4, la curva característica de la prueba y la función de información de la prueba respectivamente. La primera gráfica junto con la Tabla 3.11 muestran la relación entre habilidad y el número esperado de ítems resueltos de forma correcta en la prueba, donde, para una habilidad cero existe una esperanza aproximada de responder 16 ítems, por otro lado, observado la curva vemos que es creciente, es decir, conforme aumenta la habilidad, aumenta el número esperado de respuestas correctas, además la curva muestra una buena discriminación (pendiente alta) entre aquellos sustentantes que tienen habilidades relativamente altas (mayores que cero) y otros con habilidades relativamente bajas (menores que cero). Esto garantiza una buena selección de ítems del instrumento de evaluación. La segunda Figura unido con la Tabla 3.12 muestran la función de información de la prueba, la cual es el recíproco de la varianza con la que se puede estimar un parámetro de habilidad, por lo tanto, si pudiera estimar un parámetro con precisión (es decir, una variabilidad menor), sabría más sobre el valor del parámetro que si lo hubiera estimado con menos precisión (es decir, una variabilidad mayor). Estadísticamente, la magnitud de precisión con la que se estima un parámetro está inversamente relacionada con el tamaño de la variabilidad de las estimaciones en torno al valor del parámetro. De esta manera la función de información de la prueba es una característica extremadamente útil de la teoría de respuesta al ítem. Básicamente le dice qué tan bien le está

yendo a la prueba al estimar la habilidad en todo el rango de puntajes de habilidad. Así, para el presente diseño en el intervalo de -0.5 a 0.5 del rango de habilidad las estimaciones son más precisas en comparación del resto del rango de habilidad. Esto va en sintonía con la naturaleza de la prueba planteada, donde, se pretende medir la habilidad con una precisión considerable en niveles de habilidad cercanos a cero (habilidad utilizada para separar a aquellos aptos y no aptos), por lo tanto, la función de información de prueba en este caso tiene un pico en la puntuación cercana al corte.

Figura 3.3: Curva Característica de la Prueba.



Fuente: Elaboración propia.

Tabla 3.11: Esperanza de aciertos según Habilidad.

Habilidad	Esperanza
-0.5	12.41
-0.4	13.05
-0.3	13.68
-0.2	14.32
-0.1	14.94
0	15.56

0.1	16.16
0.2	20.75
0.3	25.33
0.4	25.90
0.5	26.44

Fuente: Elaboración propia.

Figura 3.4: Función de Información de la Prueba.

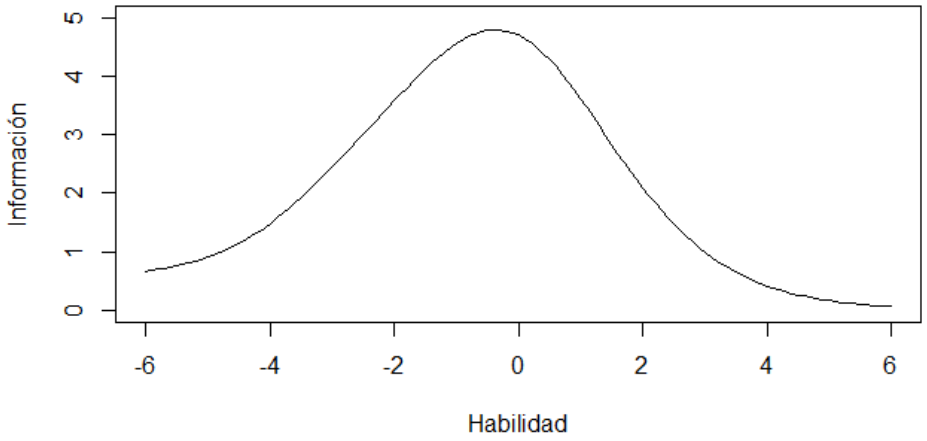


Figura 3.5: Función de Información de la Prueba.
Fuente: Elaboración propia.

Tabla 3.12: Información según Habilidad.

Habilidad	Información
-0,5	4.79
-0,4	4.80
-0,3	4.80
-0,2	4.78
-0,1	4.75
0	4.71
0,1	4.65
0,2	4.58
0,3	4.50
0,4	4.40
0,5	4.29

Fuente: Elaboración propia.

3.3.6. Conclusión de las fases de verificación y ensamble

La etapa de calibración de reactivos es importante para el diseño de la prueba en el sentido que nos permite conocer las propiedades psicométricas de los ítems tanto en TCT y TRI con uno, dos y tres parámetros, y de esta manera, garantizar el mejor ensamble de la prueba.

El ensamble de la prueba mediante TCT, es básicamente en primer lugar, establecer la teoría psicométrica y las medidas objetivo de los parámetros psicométricos que debe cumplir los ítems y la prueba. Luego, en segundo lugar, es esencialmente seleccionar los ítems de la prueba operativa final y verificar que las dificultades estén bien repartidas en su rango de variación. Por otro lado, el diseño mediante TRI varía de TCT, en la segunda etapa, solo cuando el enfoque de diseño se centra en la función de información de la prueba. Este diseño se elabora siguiendo a Lord (1977) y Birnbaum (1968), utilizando como información el punto de corte de la evaluación y el tipo de prueba, que puede ser: prueba de detección, de amplio rango o máximas. El punto de corte, luego que se conoce el puntaje de un sustentante permite clasificar al sustentante como insuficiente, elemental, bueno o excelente. La elección del punto de corte es un proceso delicado y sujeto siempre a discusión y la mayor parte de los procedimientos consisten en acuerdos entre expertos que deben justificar sus decisiones. El tipo de prueba por su parte determina la forma de la función de información de la prueba, cuando se quiere una prueba de detección la forma de la curva es leptocúrtica, para una prueba de amplio rango la curva toma una forma mesocúrtica, mientras que cuando la prueba es de amplio rango la curva es menos apuntada que una distribución mesocúrtica.

Finalmente, se destaca la importancia de la función de información en la teoría de respuesta al ítem, dado que esta

curva informa, qué tan bien le está yendo a la prueba al estimar la habilidad en todo el rango de puntajes de habilidad.

3.3.7. Cantidad de formas

3.3.7.1. Introducción

Como se explicó anteriormente, con el inicio de una evaluación educativa se crean los insumos base de la evaluación, tales como: i) marco de referencia, ii) estructuras del proyecto y iii) fichas técnicas, las cuales dan paso a la elaboración de los ítems. Posteriormente, cuando ya se cuentan con los ítems que formarán parte del instrumento de evaluación estos son revisados de forma cualitativa y cuantitativa. Luego, se realiza el proceso de armado de un instrumento de evaluación, es decir, el ensamble del instrumento.

Ahora que se cuenta con el ensamble del instrumento de evaluación, en ocasiones es necesario crear nuevos ensambles tomando como base el ensamble original, esto es, se puede reemplazar los ítems operativos (o los considerados como "ítems no anclas") por ítems espejos (ítems que tienen características psicométricas similares a los ítems originales). Estos nuevos ensambles se denominan formas de una prueba. Las formas de la prueba, se generar en función de las particularidades de la evaluación y sus distintos factores tales como: longitud del instrumento de evaluación, número de ensambles realizados, cantidad de ítems disponibles[4], cantidad de ítems ancla, ítems espejo/réplica, ítems nuevos sin calibración, días de aplicación, número de sesiones diarias, tamaño de la población y sus características.

[4] Ítems con buenos parámetros psicométricos que no se consideraron en la generación de versiones.

3.3.7.2. Longitud de la evaluación, versiones, ítem ancla, espejo o nuevo

En general dentro de la estructura de la evaluación educativa se define la longitud del instrumento, que se refiere a la cantidad de ítems que formarán parte del instrumento de evaluación en cada una de las formas.

En una evaluación educativa se crean uno o varios ensambles, estas son únicas y a partir de ellas se pueden crear posibles combinaciones con ítems disponibles, como son los ítems espejo/réplica o ítems nuevos sin calibración psicométrica.

En los diferentes proyectos de evaluación deben existir ítems ancla, que permitan lograr una medición antes de la aplicación del instrumento, es decir, estimar una esperanza del instrumento a través de la calibración de los ítems que se decida incluir, de forma técnica se debe procurar que la proporción de ítems ancla no sea menor a 30% ni mayor a 50% del total de reactivos del instrumento de evaluación; estos ítems se mantendrán de manera transversal en cada forma, y deben quedar distribuidos a lo largo de todo el instrumento y en toda la escala de dificultad. Estos reactivos deben tener, necesariamente, características cuantitativas y/o cualitativos ideales, según el modelo de medición seleccionado (INEE, 2018).

De acuerdo con el cronograma de aplicación de cada instrumento de evaluación, se planifica los días programados para rendir los las pruebas, sesiones en los que se fracciona cada día de acuerdo a la densidad poblacional, hora de cada sesión, características de la población y cantidad de sustentantes, esta información ayuda a decidir el número de ensambles adecuadas, en concordancia con la cantidad de ítems cuyos parámetros sean ideales para elaborar el diseño, y

a partir de estos crear las formas para el instrumento de evaluación.

Es necesario realizar una distinción entre las formas a aplicarse, en los diferentes días programados, ya que, si existe una alta similitud entre las formas de las sesiones de un mismo día, a día seguido, aumenta la probabilidad de divulgación del contenido del instrumento, lo cual perturba la correcta aplicación y medición de los resultados de los instrumentos de evaluación.

3.3.7.3. Características de la población

Los fundamentos de una evaluación educativa a gran escala poseen características que deben ser consideradas en el proceso de generación de formas de una prueba. Estas características son:

- Discapacidad, por ejemplo, visual, auditivo, intelectual, etc.
- Extranjeros.
- Auto identificación étnica.
- Zonas geográficas de vulnerabilidad.
- Población escolar y no escolar.

El instrumento debe estar armado de manera que cubra las particularidades de los segmentos poblacionales, tales como tiempo para responder el cuestionario, carga cognitiva, nivel de dificultad de las preguntas y sus opciones, contexto y habilidades lingüísticas y tecnológicas.
Para la población con ciertas discapacidades, se debe tomar en cuenta las adaptaciones de los ítems, de acuerdo con su propia estructura.

En la aplicación del instrumento en zonas de vulnerabilidad se debe crear una versión específica, procurando que no existan

combinaciones ni réplicas en las formas ensambladas para el resto de las zonas.

3.3.7.4. Cantidad de formas

De acuerdo con las características de la evaluación educativa la manera de calcular la cantidad de formas a aplicar puede variar. Existen evaluaciones con un número reducido de sustentantes por lo que se elabora una forma y esta es suficiente para cubrir la población, en el caso de las evaluaciones con un amplio número de sustentantes, será necesario determinar la cantidad idónea de formas.

La cantidad de formas dependerá principalmente de la producción de ítems, la cual está ligado a las características del instrumento de evaluación, los ítems disponibles se utilizarán de manera que sea posible crear una versión de ensamble por cada día de aplicación del instrumento, considerando el porcentaje de cambio que se requiere en las formas por versión en cada sesión.

Se considerará como particularidades adicionales las zonas geográficas de vulnerabilidad, que son aquellas en las que se registra un riesgo de filtración y de deshonestidad académica, si es necesario se debe crear una versión adicional para estas zonas. Para poblaciones con discapacidades se puede crear una versión de ensamble y de ella diferentes combinaciones de posiciones, sujeto a la estructura.

Considerando lo expuesto anteriormente en este apartado, en esta sección se determina el número de formas posibles y la cantidad de formas de aplicación. En el primer caso, el número de formas posibles se determina en función de:

- l: La longitud del instrumento de evaluación.
- v: Número de versiones generadas.

- ia: Número de ítems anclas.
- id: Cantidad de ítems disponibles.
- ie: Cantidad de ítems espejo/réplica.
- in: Número de ítems nuevos sin calibración utilizables.

En el segundo caso, la cantidad de formas de aplicación responden a:

- Número de días de aplicación.
- Número de sesiones diarias de aplicación.
- Cantidad de sustentantes.
- Zonas de vulnerabilidad o de alta probabilidad de filtración de ítems (zonas con potencial deshonestidad académica).

3.3.7.5. Número de formas posibles

A continuación, se muestra la forma de calcular el número de formas posibles considerando los factores identificados en dos escenarios. Considerando que una forma se genera al reemplazar por lo menos un ítem que no es ancla de una versión, mediante un ítem disponible, espejo/réplica o un ítem nuevo sin calibración.

Escenario 1: Cuando se cuenta con suficientes ítems disponibles con buenos parámetros psicométricos que no se consideraron en la generación de versiones. Se distribuye estos ítems en las posiciones de los ítems que no son ancla cuidando que evalúen la misma habilidad. En este sentido, el número de formas posibles (m^p) se obtiene como:

$$m^p = m_1^p + m_2^p + \ldots + m_v^p,$$

con $\quad m_j^p = \left(id_{DO_1} + 1\right).\left(id_{DO_2} + 1\right)\ldots\left(id_{DO_{l-ia}} + 1\right) - 1; \quad j = 1,2,\ldots,v,$

donde:

m^p: número de formas posibles.

m_j^p: número de formas posibles en la versión $j - ésima$.

id_{DO_i}: Cantidad de ítems disponibles de la definición operacional $i - ésima$.

$l - ia$: Número de definiciones operaciones que no son anclas.

$v:$ Número de versiones generadas.

Le sumamos una unidad a cada id_{DO_i} por la consideración para la generación de una forma y le restamos uno ya que corresponde a la combinación de no incluir ningún ítem a la versión, es decir, es la forma original.

Escenario 2: En situaciones cuando el número de ítems disponibles con buenos parámetros psicométricos que no se consideraron en la generación de versiones no es suficiente, se echa mano de los ítems espejo/réplica e ítems nuevos sin calibración que son utilizables. Luego se distribuye estos ítems en las posiciones de los ítems que no son ancla cuidando que evalúen la misma habilidad. En este sentido, el número de formas posibles se obtiene de manera similar al escenario uno:

$$m^p = m_1^p + m_2^p + \ldots + m_v^p,$$

con $m_j^p = \left(ien_{DO_1} + 1\right).\left(ien_{DO_2} + 1\right)\ldots\left(ien_{DO_{l-ia}} + 1\right) - 1;\ j = 1,2,\ldots,v$ y $ien = ie + in,$

donde:

m^p: número de formas posibles.

m_j^p: número de formas posibles en la versión $j-$ é$sima$.

ien_{DO_1}: Cantidad de ítems espejo/réplicas e ítems nuevos de la definición operacional $i-$ é$sima$.

$l-ia$: Número de definiciones operaciones que no son anclas.

v: Número de versiones generadas.

Tanto en el escenario uno y dos podemos notar que existen formas que solo se diferencian en pocos ítems (uno, dos, etc.) Respecto a la versión original, en este punto es importante el porcentaje de diversificación de ítems que se quiere que tengan las formas, al respecto se construye la Tabla 3.13 resumen según porcentaje de diversificación y cantidad de formas posibles para ayudar en esta decisión.

Cantidad de ítems reemplazados	**Diversificación**	**Cantidad de formas**
1	$\frac{1}{(l-ia)}$	$m_{d_1}^p$
2	$\frac{2}{(l-ia)}$	$m_{d_2}^p$
⋮	⋮	⋮
$(\boldsymbol{l-ia})-\mathbf{1}$	$1-\frac{1}{(l-ia)}$	$m_{d_{(l-ia)-1}}^p$
$(\boldsymbol{l-ia})$	1	$m_{d_{(l-ia)}}^p$

* $m_{d_k}^p$: Es la cantidad de formas posibles cuando se diversifica k ítems de la prueba.

Tabla 3.13: Cantidad de formas ideales según porcentajes de diversificación.
Fuente: Elaboración propia.

Los valores de $m_{d_k}^{p}$ con $k = 1,2,\ldots, l - ia$ se determinan considerando que ir_{DO_i} puede representar la cantidad de ítems disponibles o la cantidad de ítems espejo/réplicas e ítems nuevos de la definición operacional $i - ésima$ con los que se cuenta para diversificar las formas, entonces tenemos que:

$$m_{d_1}^{p} = ir_{DO_1} + ir_{DO_2} + \ldots + ir_{DO_{l-ia}},$$

$m_{d_2}^{p}$ se obtiene multiplicando todas las parejas posibles sin que importe el orden del conjunto de valores $\{ir_{DO_1}; ir_{DO_2}; \ldots; ir_{DO_{l-ia}}\}$, es decir, se debe multiplicar $\binom{l-ia}{2} = \frac{(l-ia)!}{2!(l-ia-2)!}$ parejas de valores del conjunto y al final sumarlos del siguiente modo:

$$m_{d_2}^{p} = \left(ir_{DO_1}\right).\left(ir_{DO_2}\right) + \left(ir_{DO_1}\right).\left(ir_{DO_3}\right) + \ldots + \left(ir_{DO_{(l-ia)-1}}\right).\left(ir_{DO_{l-ia}}\right)$$

$m_{d_3}^{p}$ se obtiene multiplicando todas las ternas posibles sin que importe el orden del conjunto de valores $\{ir_{DO_1}; ir_{DO_2}; \ldots; ir_{DO_{l-ia}}\}$, es decir, se debe multiplicar $\binom{l-ia}{3} = \frac{(l-ia)!}{3!(l-ia-3)!}$ ternas de valores del conjunto y al final sumarlos del siguiente modo:

$$
\begin{aligned}
& m_{d_3}^{p} \\
& = \left(ir_{DO_1}\right).\left(ir_{DO_2}\right).\left(ir_{DO_3}\right) \\
& + \left(ir_{DO_1}\right).\left(ir_{DO_2}\right).\left(ir_{DO_4}\right) + \ldots + \left(ir_{DO_{(l-ia)-2}}\right).\left(ir_{DO_{(l-ia)-1}}\right).\left(ir_{DO_{l-ia}}\right) \\
& \vdots
\end{aligned}
$$

Finalmente, $m_{d_{l-ia}}^{p}$ se obtiene multiplicando todos los arreglos posibles de tamaño $l - ia$ sin que importe el orden del conjunto de valores $\{ir_{DO_1}; ir_{DO_2}; \ldots; ir_{DO_{l-ia}}\}$, es decir, se debe multiplicar, $\binom{l-ia}{l-ia} = 1$, vez los todos los valores del conjunto así:

$$m_{d_{l-ia}}^{p} = \left(ir_{DO_1}\right).\left(ir_{DO_2}\right)\ldots\left(ir_{DO_{l-ia}}\right),$$

Luego de la construcción de la Tabla 3.13 se elige entre los diferentes porcentajes de diversificación de ítems considerando que:

- Del 80% al 100% es una diversificación Altamente Satisfactoria;
- Del 60% al 79% es una diversificación Muy Satisfactoria.
- Del 40% al 59% es una diversificación Satisfactoria.
- Del 20% al 49% es una diversificación Poco Satisfactoria.
- Del 0% al 19% es una diversificación Nada Satisfactoria.

Esto determina la cantidad de formas que se debe generar, sin embargo, es importante destacar que la cantidad de formas ideales descritas en la Tabla 3.13 disminuye conforme la

revisión cualitativa detecta ítems que no cumplen con los parámetros de redacción.

3.3.7.6. Cantidad de formas de aplicación

La cantidad de formas de aplicación debe ser calculado en función del número de días de aplicación y el número de sesiones diarias en función de la cantidad de sustentantes y de las zonas de vulnerabilidad o de alta probabilidad de filtraciones. El cálculo no es más que la multiplicación de los valores posibles de las características a considerar. Lógicamente, este número no puede ser mayor que la cantidad de formas posibles para que se puede asignar las formas generadas.

3.4. Aplicación y calificación

Después de haber pasado por las fases de diseño, construcción, verificación y ensamble de la evaluación educativa, es hora de aplicar y calificar la prueba. La aplicación de la prueba se realiza según el i) número de días; ii) número de sesiones diarias; iii) cantidad de sustentantes y iv) zonas de vulnerabilidad o de alta probabilidad de filtración de ítems (zonas con potencial deshonestidad académica). Luego el procedimiento de la calificación se describe a continuación:

1. Se obtienen los resultados de la evaluación educativa, es decir, una matriz (denominada matriz binaria) donde en sus filas están los sustentantes y en las columnas los ítems aplicados. La información de la matriz son valores de 0 (ítem contestado correctamente) o 1 (ítem contestado incorrectamente).
2. Es práctica común descartar de la matriz binaria a aquellos sustentantes que registraban más de treinta por ciento de

ítems no contestados, por considerarse que dichas cadenas podrían alterar la calibración de los ítems.
3. Es importante desarrollar la calibración por constructo cuando la evaluación educativa mide más de un constructo.
4. El índice de dificultad y discriminación en teoría clásica de los test, son calculados mediante promedio y correlación respectivamente, lo cual se lo realiza mediante el software estadístico R.
5. Los índices de dificultad, discriminación y pseudo-adivinación en teoría de respuesta al ítem requirieron del uso de la función tam.mml.2pl del paquete TAM asociado al software estadístico R.
6. Una vez calculados todos los parámetros, se acepta o descarta ítems utilizando las tablas de dictámenes anteriormente expuestas.
7. Finalmente, se generan tablas con las calibraciones y dictámenes.

3.5. Mantenimiento de la Base de Datos de Ítems

Luego del proceso de aplicación y calificación de la prueba de evaluación se debe desarrollar una revisión técnica de validación y aprobación de ítems con el objetivo de que formen parte del banco de ítems, para ocuparlos en posteriores aplicaciones de la evaluación educativa. Como ya se cuenta con las características psicométricas de los ítems, ahora se debe realizar una revisión exhaustiva de los parámetros de calibración TCT y TRI, en compañía de un especialista por campo de evaluación.

Esta revisión exhaustiva tendrá los siguientes parámetros:

- Código único del ítem.
- Parámetros cuantitativos obtenidos del proceso de calibración.

- Redacción del ítem.
- El ítem evalúa efectivamente la habilidad requerida.
- Respuesta correcta mal marcada.
- No cuenta con respuesta correcta.
- Para resolver el ítem se necesita un conocimiento adicional fuera de lo común.
- Opciones de respuesta plausibles.
- El ítem trata de un tema sensible.

En función de los parámetros descritos en la lista anterior se debe dictaminar si el ítem ingresa o no al banco de ítems.

Capítulo 4 : Fiabilidad, Validez y Ajuste de una Prueba Educativa

En el presente capítulo, en primer lugar, se describe la fiabilidad y validez de contenido que son características psicométricas que pueden ser evaluadas mediante ciertos modelos, estadísticos o procedimientos empíricos. En segundo lugar, se muestra una forma de estudiar el ajuste de los datos del modelo a los ítems.

4.1.Fiabilidad

La fiabilidad o confiabilidad es el grado en que una herramienta de evaluación produce resultados estables y consistentes. Los instrumentos de evaluación educativas deben ser confiables, es decir, no debe haber ninguna diferencia si un estudiante toma la evaluación por la mañana o por la tarde; un día o el siguiente, etc. La Tabla 4.1 describe tres medidas de confiabilidad comunes.

Tabla 4.1: Medidas de confiabilidad comunes.

Tipo de [illegible]	Cómo medir
Estabilidad o prueba-repetición	Se debe realizar la misma evaluación dos veces, separadas por días, semanas o meses. La confiabilidad se establece como la correlación entre las puntuaciones en el Tiempo 1 y el Tiempo 2.
Forma alternativa	Se crean dos formas de la misma prueba. La confiabilidad se establece como correlación entre los puntajes de la Prueba 1 y la Prueba 2.
Consistencia interna (Alfa, *a*)	Se compara una mitad de la prueba con la otra mitad. O se utiliza el Alfa de Cronbach.

Fuente: Elaboración propia.

Todos los tipos de confiabilidad se pueden evaluar mediante la recopilación y el análisis de datos. En el mismo sentido, este capítulo evalúa la confiabilidad en la práctica de una prueba educativa que evalúa los campos de i) razonamiento verbal, ii) razonamiento numérico y iii) razonamiento abstracto. La evaluación se realiza mediante el tipo de confiabilidad alternativa descrito en la Tabla 4.1, comparando los parámetros psicométricos de los ítems antes de la evaluación con los mismos después de la evaluación. La prueba educativa de interés consta de 25 ítems distribuidos según campo como se muestra en la Tabla 4.2.

Tabla 4.2: Distribución de ítems según campo y forma.

Campo	Forma 1 y 2
1. ***Razonamiento verbal***	8
2. ***Razonamiento numérico***	8
3. **Razonamiento abstracto**	9
Total	25

Fuente: Elaboración propia.

En la forma uno de aplicación de la evaluación el diseño se desarrolló mediante 19 ítems y en la forma dos se trabajó con 18 ítems. Entre las dos formas comparten 17 ítems, entonces (ver Tabla 4.3) en total se consideraron 20 ítems únicos para el ensamble de la evaluación educativa. Un conjunto de reactivos no se consideró para el diseño debido a que eran ítems nuevos o anteriores sin parámetros psicométricos.

Tabla 4.3: Distribución de ítems según campo y forma.

Campo	Forma 1	Forma 2	Ítems en común forma 1 y 2
1. ***Razonamiento***	8	8	8
2. ***Razonamiento***	5	5	4
3. **Razonamiento abstracto**	6	5	5
Total	19	18	17

Fuente: Elaboración propia.

El ensamble de la evaluación educativa se desarrolló mediante la Teoría de Respuesta al Ítem con dos parámetros psicométricos (dificultad y discriminación) y más específicamente con la probabilidad de respuesta al ítem cuando la habilidad de un sustentante es de cero. En este contexto, se presenta a continuación (ver Tabla 4.4) la comparación de los promedios geométricos de ítems del ensamble antes y después de la evaluación.

Tabla 4.4: Comparación de promedios geométricos de ítems de diseño antes y después de la evaluación.

	Forma 1		Forma 2		Total	
Campo	Antes	Después	Antes	Después	Antes	Después
1. ***Razonamiento***	61.90%	72.54%	61.90%	72.54%	61.90%	72.54%
2. ***Razonamiento***	49.76%	58.11%	47.93%	57.21%	48.85%	57.66%

3.	**Razonamiento abstracto**	44.49%	51.63%	47.43%	56.17%	45.80%	53.64%
	Total	52.66%	61.46%	53.54%	63.25%	53.09%	62.32%

Fuente: Elaboración propia.

En general en la tabla anterior se observa discrepancias que varían en el rango de 7.14% a 10.64%, la "distancia" menor entre las puntuaciones promedio antes y después se obtuvo para el campo Razonamiento abstracto, mientras que la mayor discrepancia que se encontró fue en el campo Razonamiento verbal. La confiabilidad de la forma alternativa del diseño en la evaluación educativa se obtiene en este caso relacionado con la probabilidad de respuesta al ítem (con dos parámetros) cuando la habilidad de un sustentante es de cero. Para esto primero se presenta la tabla con esta información más la discrepancia y descripción del ítem en las dos formas (ver Tabla 4.5).

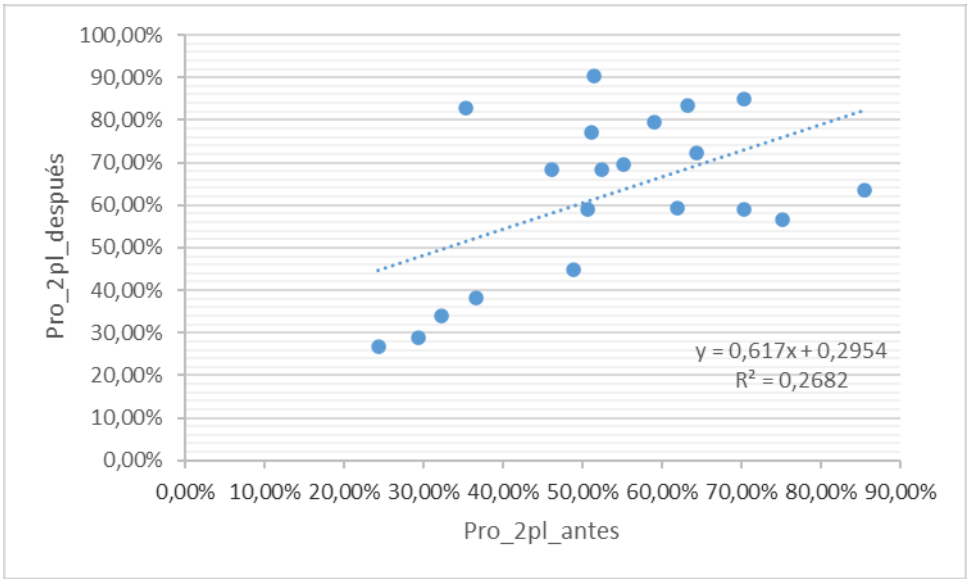
Tabla 4.5: Probabilidad de respuesta al ítem con habilidad cero antes y después de la aplicación de la evaluación.

Ítem	Forma	Prob_2pl_antes	Prob_2pl_despues	Discrepancia
1	F001 y F002	85,54%	63,65%	-21,89%
2	F001 y F002	51,18%	77,04%	25,86%
3	F001 y F002	75,20%	56,66%	-18,54%
4	F001 y F002	35,33%	82,85%	47,52%
5	F001 y F002	70,44%	59,13%	-11,32%

6	F001 y F002	63,29%	83,50%	20,21%
7	F001 y F002	59,02%	79,47%	20,45%
8	F001 y F002	70,39%	84,89%	14,49%
9	F001 y F002	51,40%	90,42%	39,02%
10	F001 y F002	61,90%	59,37%	-2,53%
11	F001 y F002	64,35%	72,30%	7,96%
12	F001	29,46%	28,89%	-0,57%
13	F001 y F002	50,73%	59,10%	8,36%
14	F001 y F002	36,71%	38,31%	1,60%
15	F001	32,29%	33,87%	1,57%
16	F001 y F002	46,21%	68,37%	22,16%
17	F001 y F002	52,39%	68,47%	16,08%
18	F001 y F002	55,18%	69,70%	14,52%
19	F001 y F002	48,95%	44,72%	-4,23%
20	F002	24,34%	26,70%	2,36%

Fuente: Elaboración propia.

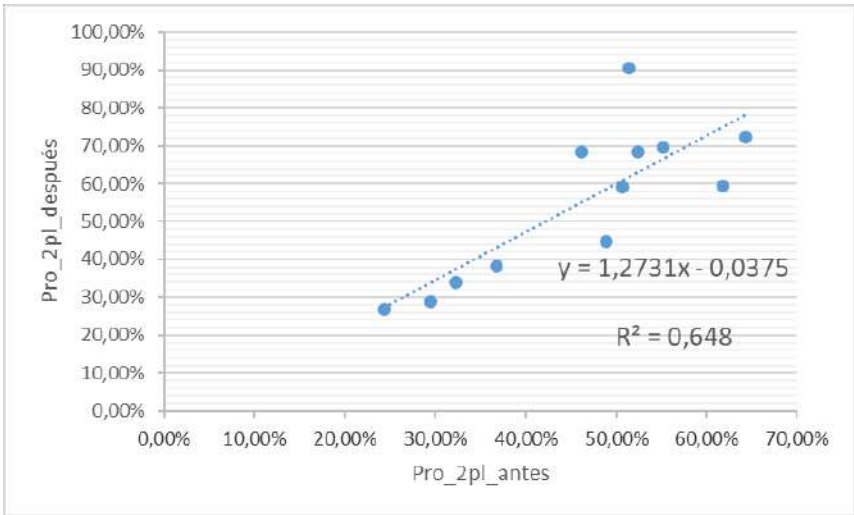
Figura 4.1: Confiabilidad de la forma alternativa del diseño en la evaluación.



Fuente: Elaboración propia.

En general la confiabilidad del diseño es pobre ya que el valor de R-cuadrado es de 0.2682 (ver Figura 4.1), sin embargo, si observamos con cuidado la anterior ilustración, se vira que la consistencia es buena para aquellos ítems que tienen Pro_2pl_después menor que 64%, mientras que los otros ítems distorsionan la relación entre las probabilidades. De hecho, los ítems 4, 2, 7, 6 y 8 presentan las mayores discrepancias absolutas entre las probabilidades antes y después. Estos ítems pertenecen al campo Razonamiento verbal que en su mayoría presentan un aumento en la probabilidad, por lo tanto, en realidad existe una buena confiabilidad ya que el valor de R-cuadrado es de 0.6482 (ver Figura 4.2). Naturalmente, también se pudo calcular el coeficiente alfa de Cronbach.

Figura 4.2: Confiabilidad de la forma alternativa del diseño en la evaluación Razonamiento Intercultural Bilingüe, campos Razonamiento abstracto y numérico.



Fuente: Elaboración propia.

4.2. Validez

En este apartado se muestran los resultados del estudio de validez de contenido considerando los criterios de evaluación de Relevancia, Pertinencia, Claridad, Alcance y Focalización de Contenido, en un instrumento de evaluación de interés. Al respecto, se muestra el estudio de validez de contenido siguiendo a Tristán (2008) según los criterios de validación medidos en escala Likert:

- Relevancia: El parámetro es esencial para medir el criterio de desempeño.
- Pertinencia: El parámetro es adecuado para la población objetivo.
- Claridad: El parámetro planteado carece de ambigüedades.
- Alcance: El parámetro planteado cubre el o los tópicos de evaluación que se pretenden medir.

- Focalización de Contenido: El parámetro mide únicamente el contenido y habilidad implícitos en los criterios de desempeño.

Para esto, se calcula en primer lugar el número de ítems catalogados como Esenciales (niveles 4 y 5 en la escala de Likert), Útiles (nivel 3 en la escala de Likert) y No necesarios (niveles 1 y 2 en la escala de Likert). Luego se obtiene la razón de validación de contenido (RVC) e índice de validación de contenido (IVC) utilizando las siguientes expresiones:

$$RVC_i = \frac{n_i}{N},$$

Donde:

RVC_i: es la razón de validación de contenido del $i - ésimo$.

n_i: es el número de expertos que coincidieron en dar una calificación de Esencial al Parámetros $i - ésimo$.

N: es el número de expertos.

$$IVC_{aceptable} = \frac{\sum_{i=1}^{P} RVC_i}{P},$$

Donde:

$IVC_{aceptable}$: Índice de validez de contenido tomado en cuenta sólo los parámetros considerados como aceptables[5].

[5] Un parámetro es considerado como aceptable cuando la RVC es superior a 0.58 (Tristán, 2008).

RVC_i: es la razón de validación de contenido del $i-ésimo$ parámetro.

P: es el número de parámetros considerados como aceptables en el estudio.

$$IVC_{total} = \frac{\sum_{i=1}^{T} RVC_i}{T},$$

Donde:

IVC_{total}: Índice de validez de contenido tomado en cuenta todos los parámetros del estudio.

RVC_i: es la razón de validación de contenido del $i-ésimo$ parámetro.

T: es el número total de parámetros considerados en el estudio.

Un valor del índice de validación de contenido superior a 0.58 indica que el ítem ($IVC_{aceptable}$) o toda la prueba (IVC_{total}) tiene una validez de contenido aceptable.

Luego de mostrar la teoría que soporta el estudio de validez de contenido, a continuación, se muestra un resumen por criterios de la evaluación de los ítems estudiados:

4.2.1. Relevancia

Tabla 4.6: Relevancia: Validez de contenido del instrumento de evaluación.

Ítems	Esencial	Útil	No Necesario	RVC	Aceptable	Datos Faltantes
1	6	0	0	1	1	0

2	5	1	0	0.83	1	0
3	6	0	0	1	1	0
4	6	0	0	1	1	0
5	6	0	0	1	1	0
6	5	0	1	0.83	1	0
7	6	0	0	1	1	0
8	6	0	0	1	1	0
9	6	0	0	1	1	0
10	6	0	0	1	1	0
11	4	0	2	0.67	1	0
12	5	0	1	0.83	1	0
13	6	0	0	1	1	0
14	6	0	0	1	1	0
15	6	0	0	1	1	0
16	6	0	0	1	1	0
17	6	0	0	1	1	0

Fuente: Elaboración propia.

Por lo tanto, el índice de validación de contenido es $IVC_{aceptable} = IVC_{total} = 0.95$.

4.2.2.Pertinencia

Tabla 4.7: Pertinencia: Validez de contenido del instrumento de evaluación.

Ítems	Esencial	Útil	No Necesario	RVC	Aceptable	Datos Faltantes
1	5	1	0	0.83	1	0
2	6	0	0	1	1	0
3	6	0	0	1	1	0
4	6	0	0	1	1	0
5	6	0	0	1	1	0
6	6	0	0	1	1	0
7	6	0	0	1	1	0
8	6	0	0	1	1	0
9	6	0	0	1	1	0
10	6	0	0	1	1	0
11	4	0	2	0.67	1	0
12	6	0	0	1	1	0
13	6	0	0	1	1	0
14	6	0	0	1	1	0
15	6	0	0	1	1	0
16	6	0	0	1	1	0
17	6	0	0	1	1	0

Fuente: Elaboración propia.

Entonces, el índice de validación de contenido es $IVC_{aceptable} = IVC_{total} = 0.97$.

4.2.3.Claridad

Tabla 4.8: Claridad: Validez de contenido del instrumento de evaluación.

Ítems	Esencial	Útil	No Necesario	RVC	Aceptable	Datos Faltantes
1	4	0	2	0.67	1	0
2	5	0	1	0.83	1	0
3	4	1	1	0.67	1	0
4	5	0	1	0.83	1	0
5	5	0	1	0.83	1	0
6	5	0	1	0.83	1	0
7	5	0	1	0.83	1	0
8	5	0	1	0.83	1	0
9	5	0	1	0.83	1	0
10	5	0	1	0.83	1	0
11	4	0	2	0.67	1	0
12	5	0	1	0.83	1	0
13	5	0	1	0.83	1	0
14	5	0	1	0.83	1	0
15	5	0	1	0.83	1	0
16	5	0	1	0.83	1	0
17	5	0	1	0.83	1	0

Fuente: Elaboración propia.

El índice de validación de contenido es $IVC_{aceptable} = IVC_{total} = 0.80$.

4.2.4. Alcance

Tabla 4.9: Alcance: Validez de contenido del instrumento de evaluación.

Ítems	Esencial	Útil	No Necesario	RVC	Aceptable	Datos Faltantes
1	4	1	1	0.67	1	0
2	6	0	0	1	1	0
3	6	0	0	1	1	0
4	6	0	0	1	1	0
5	5	0	1	0.83	1	0
6	6	0	0	1	1	0
7	6	0	0	1	1	0
8	6	0	0	1	1	0
9	6	0	0	1	1	0
10	5	0	1	0.83	1	0
11	4	0	2	0.67	1	0
12	5	0	0	0.83	1	1
13	6	0	0	1	1	0
14	6	0	0	1	1	0
15	6	0	0	1	1	0
16	6	0	0	1	1	0
17	6	0	0	1	1	0

Fuente: Elaboración propia.

El índice de validación de contenido es $IVC_{aceptable} = IVC_{total} = 0.93$.

4.2.5. Focalización de contenido

Tabla 4.10: Focalización del Contenido: Validez de contenido del instrumento de evaluación.

Ítems	Esencial	Útil	No Necesario	RVC	Aceptable	Datos Faltantes
1	6	0	0	1	1	0
2	6	0	0	1	1	0
3	6	0	0	1	1	0
4	6	0	0	1	1	0
5	5	0	1	0.83	1	0
6	6	0	0	1	1	0
7	6	0	0	1	1	0
8	6	0	0	1	1	0
9	6	0	0	1	1	0
10	4	0	2	0.67	1	0
11	5	0	1	0.83	1	0
12	6	0	0	1	1	0
13	6	0	0	1	1	0
14	5	0	1	0.83	1	0
15	6	0	0	1	1	0
16	6	0	0	1	1	0
17	6	0	0	1	1	0

Fuente: Elaboración propia.

Por lo tanto, el índice de validación de contenido es $IVC_{aceptable} = IVC_{total} = 0.95$.

Observando los resultados de la Tabla 4.6 a 4.10, con respecto a los criterios de validación de Relevancia, Pertinencia, Claridad, Alcance y Focalización de Contenido se tiene que los 17 parámetros son aceptables ya que el valor de la razón de validación de contenido de Tristán es superior a 0.58. Estos parámetros son aptos para integrar el instrumento o banco de parámetros. En el mismo sentido, se estudió la validez de los

17 parámetros en conjunto mediante el cálculo del índice de validación de contenido dando un valor de 0.95, 0.97, 0.80, 0.93, 0.95 para Relevancia, Pertinencia, Claridad, Alcance y Focalización de Contenido, es decir, el conjunto de parámetros es aceptable considerando todos los parámetros. Por lo tanto, los parámetros individuales como en conjunto según la RVC e IVC de Tristán (2008) tiene validez de contenido en los cinco criterios de validación del Instrumento de evaluación.

4.3. Ajuste

Generalmente, después del desarrollo de la calibración en evaluaciones educativas se realiza un análisis de residuales para examinar el ajuste de los datos del modelo a los ítems, esto es importante ya que, si los datos no se ajustan a un modelo de teoría de respuesta al ítem, esto puede conducir a una pérdida de la propiedad de invarianza o invariante (Wells y Hambleton, 2016). Al respecto, dentro de las ventajas de la Teoría de Respuesta al Ítem están i) la invarianza de los parámetros de los ítems respecto a la muestra que se calcula, es decir, que los parámetros de ítem no cambian, aunque las personas que contesten sean distintas y ii) invarianza del parámetro del rasgo del sujeto respecto al instrumento utilizado para estimarlo, es decir, que el nivel de habilidad de la persona no depende del test.

4.3.1. Método para evaluar el ajuste de los datos a un modelo

Rasch (1980) describió una teoría de medición basada en los requisitos de objetividad específica que respalda la medición invariante. La medición invariable proporciona puntajes significativos que mantienen su significado en diferentes

contextos cuando se obtiene un ajuste adecuado del modelo y los datos. Para darse cuenta de las propiedades deseables de la teoría de medición de Rasch, se requiere un nivel adecuado de ajuste modelo-datos.

El modelo dicotómico de Rasch se puede escribir de la siguiente manera:

$$\phi_{ni1} = \frac{exp\ (\theta_n - \delta_i)}{1 + exp\ (\theta_n - \delta_i)},$$

Dónde ϕ_{ni1}, representa la probabilidad de que el sustentante n tenga una ubicación de θ_n en una variable latente que da una respuesta positiva al ítem i con una ubicación de δ_i en la variable latente.

Una vez que se obtienen las estimaciones de las ubicaciones de los sustentantes y los ítems, se puede usar la ecuación de, ϕ_{ni1}, para obtener las probabilidades esperadas (predichas), P_{ni1}, de la siguiente manera:

$$P_{ni1} = \frac{exp\ (\hat{\theta}_n - \hat{\delta}_i)}{1 + exp\ (\hat{\theta}_n - \hat{\delta}_i)}.$$

Con base en el modelo dicotómico de Rasch en la ecuación de, ϕ_{ni1}, se puede obtener una estimación de las varianzas de respuesta (información estadística), Q_{ni}, de la siguiente manera:

$$Q_{ni} = P_{ni}(1 - P_{ni}).$$

Los residuos de puntuación, Y_{ni}, se pueden definir como

$$Y_{ni} = X_{ni} - P_{ni},$$

Donde X_{ni} son las respuestas dicotómicas observadas. Un conjunto de residuos estandarizados, Z_{ni}, se puede definir de la siguiente manera:

$$Z_{ni} = \frac{Y_{ni}}{Q_{ni}^{1/2}}.$$

En la teoría de medición de Rasch se utilizan normalmente varios índices de ajuste basados en residuos: estadísticas de error cuadrático medio (MSE) de Infit y Outfit. Estas estadísticas MSE de Infit y Outfit se basan en residuos, y los residuos pueden resumirse e informarse tanto para ítems como para sustentantes. Por ejemplo, la estadística Outfit MSE para el sustentante n se puede definir como

$$Outfit\ MSEn = \frac{\sum_i^L Z_{ni}^2}{L},$$

Donde Z_{ni}^2 son los residuos estandarizados al cuadrado y L es el número de ítems. Las estadísticas de $Outfit\ MSE$ son promedios no ponderados y son sensibles a valores atípicos y residuales inesperados extremos. Las estadísticas Infit MSE proporcionan una versión ponderada de información de la estadística MSE. Se pueden definir para el sustentante n como:

$$Infit\ MSEn = \frac{\sum_i^L Y_{ni}^2}{\sum_i^L Q_{ni}},$$

donde Y_{ni}^2 son los residuos al cuadrado, Q_{ni} es la varianza de las probabilidades de respuesta esperadas, y L es el número de ítems. Debido a que las estadísticas Infit MSE son promedios ponderados por información, estas estadísticas son menos sensibles a valores atípicos extremos. Los residuos se pueden resumir de manera similar sobre los ítems para producir Infit y Outfit MSE para ítems. Cuando los datos se ajustan al modelo de Rasch, el valor esperado de las estadísticas MSE de Infit y Outfit es 1.00, y las estadísticas MSE pueden variar de 0 a infinito positivo. Los valores de MSE superiores a 1.00 indican patrones de respuesta con más variación de la esperada, mientras que los valores inferiores a 1,00 tienden a reflejar una variación menor de la esperada según el modelo de medición.

Es recomendable utilizar en la práctica las siguientes pautas descriptivas adaptadas de Wright y Linacre (1994) para etiquetar e interpretar las estadísticas de MSE (ver Tabla 4.11):

Tabla 4.11: Pautas descriptivas para MSE.

MSE	Interpretación	Categoría
$0.50 \leq MSE \leq 1.50$	Productiva para medición.	A
$MSE < 0.50$	Menos productivo para la medición, pero no distorsiona las medidas.	B
$1.50 \leq MSE \leq 2.00$	Improductivo para la medición, pero no distorsionador de las medidas.	C
$MSE \geq 2.00$	Improductivo para la medición, distorsionando las medidas.	D

Fuente: Elaboración propia.

Bibliografía

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (Ed.), Educationalmeasurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Appelbaum, M. I. (1986). Statistics, data analysis, and Psychometrika: Major developments. Psychometrika51.

Birnbaum, A. (1968). Classification by ability levels. Statistical theories of mental test scores, 436-452.

Chacón, S., & Sanduvete, S. (s.f). Baremación, estandarización y equiparación de puntuaciones. Consultado el 21 de junio de 2022. https://personal.us.es/vmanzano/docencia/psicometria/ppt/tema7.pdf

Chávez, C., & Saade, A. (2009). Procedimientos básicos para el análisis de reactivos Cuaderno técnico 8. CENEVAL.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika16.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin52.

Fellbaum, C. (2005). Word Net y redes de palabras. Encyclopedia of Language and Linguistics (2a ed.). Elsevier.

Gifi, A. (1990). Nonlinear Multivariate Analysis. Wiley.

Green, B. F. (1954). Attitude measurement. Addison-Wesley.

Guilford, J. P. (1936). Psychometric Methods. McGraw-Hill.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika6.

INEE, (2018). Guía para la elaboración de instrumentos de evaluación. INEE.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika2.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. Wiley.

Linacre, M. (2003). A user's guide to Winsteps. Mesa Press.

Lord, F. M. (1952). A Theory of Test Scores. Whole.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika18.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 117-138.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Addison-Wesley.

Meneses, J., Barrios, M., Bonillo, A., Cosculluela, A., Lozano, L., Turbany, J., & Valero, S. (2013). Psicometría. Universitat Oberta de Catalunya.

Ramsay, J. O. (2001). Psychometrics. International Encyclopedia of the Social & Behavioral Sciences, 12416-12422. https://doi.org/10.1016/B0-08-043076-7/00650-1

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Denmarks Paedagogiske Institut, Copenhagen. Republished in 1980 by the University of Chicago Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 321–333.

Rasch, G. (1966). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology19.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In: Blegvad, M. (Ed.),The Danish Yearbook of Philosophy. Munksgaard, Copen-hagen.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (Expandeded.). Copenhagen, Denmark: Danish Institute for Educational Research. (Original work published in 1960).

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. Wiley.

Tristán, A. (1998). Análisis de Rasch para todos. CENEVAL.

Tristán-López, A. (2008). Modificación al modelo de Lawshe para el dictamen cuantitativo de la validez de contenido de un instrumento objetivo. Avances en medición, 6(1), 37-48.

Wells, C. S., & Hambleton, R. K. (2016). Model fit with residual analyses. In W. J. van derLinden (Ed.), Handbook of item response theory: Vol. 2. Statistical tools(pp. 395-413). Boca Raton, FL: CRC Press.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370.

Zieky, M. J., & Livingston, S. A. (1977). Basic Skills Assessment. Manual for Setting Standards on the Basic Skills Assessment Tests.

Héctor Mullo Guaminga
Estudió la carrera de estadística en la Facultad de Ciencias de la ESPOCH. Realizó estudios de maestría (Máster Universitario en Estadística Aplicada) y doctorado (Estadística Matemática y Aplicada) en la Universidad de Granada, España. Es profesor en la Facultad de Ciencias de la ESPOCH desde 2014. Su trabajo docente lo ha dirigido principalmente en asignaturas relacionadas a la estadística teórica y aplicada. Además, ha realizado trabajos de investigación en muestreo, psicometría y en general en estadística aplicada.

Jessica Alexandra Marcatoma
Estudió la carrera de estadística en la Facultad de Ciencias de la ESPOCH. Realizó estudios de maestría (Máster Universitario en Estadística Aplicada) en la Universidad de Granada, España. Fue profesora en la Facultad de Ciencias de la ESPOCH desde 2014 a 2019 y es profesora en la Facultad de Ingeniería de la UNACH desde 2019. Su trabajo docente lo ha dirigido principalmente en asignaturas relacionadas a la estadística teórica y aplicada. Además, ha realizado trabajos de investigación en estadística aplicada.

ISBN: 978-9942-33-590-6

compAs
Grupo de capacitación e investigación pedagógica

@grupocompas.ec
compasacademico@icloud.com