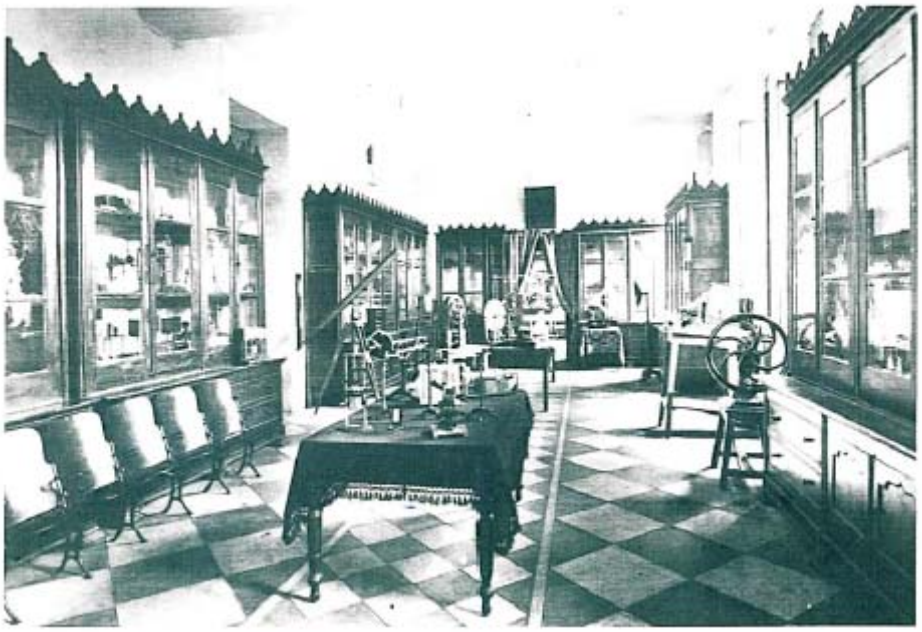
*Subsidios Metodológicos para el Profesor Investigador en Enseñanza de las Ciencias*

# *Métodos Cualitativos y Cuantitativos*



*Marco A. Moreira*
*Paulo R. Rosa*

*Porto Alegre, Brasil*
*2009|2016*

# *Ficha de presentación*

**Tipo de publicación:** Recopilación de trabajos publicados o presentados en congresos sobre el tema *Métodos Cualitativos y Cuantitativos* a fin de subsidiar metodológicamente el profesor investigador, en particular del área de enseñanza de las ciencias.

**Autores:** *Marco A. Moreira* – Instituto de Física, UFRGS, Brasil
*Paulo R. Rosa* – Instituto de Física, UFMS, Brasil
**Fecha y local:** 2009 (1ª edición), 2016 (2ª edición revisada) Porto Alegre, Brasil.

# *Otras publicaciones de la misma serie*

- **Subsidios Teóricos:** *Comportamentalismo, Constructivismo y Humanismo.*
- **Subsidios Teóricos:** *La Teoría del Aprendizaje Significativo.*
- **Subsidios Epistemológicos:** *Epistemologías del Siglo XX.*
- **Subsidios Metodológicos:** *Investigación en Enseñanza – Aspectos Metodológicos.*
- **Subsidios Didácticos:** *Mapas Conceptuales, Diagramas V y Organizadores Avanzados.*

# *Sumario*

**Presentación**....……..……..……..……..…….4

Capítulo 1

**Investigación en enseñanza: métodos cualitativos**….…….……5

Capítulo 2

**Investigación en enseñanza: métodos cuantitativos**……..32

# *Presentación*

Este material bibliográfico está constituido por dos textos de apoyo sobre metodologías de investigación en enseñanza de las ciencias. Fueron escritos para seren usados independientemente uno de otro. Consecuentemente pueden presentar ciertas superposiciones.

El primer capítulo empieza con una breve comparación entre los paradigmas cuantitativo y cualitativo, después concentrase en el enfoque cualitativo y al final aborda el tema de la triangulación metodológica. El segundo intenta rescatar la metodología cuantitativa aplicada a la investigación en enseñanza.

Finalmente, hay que llamar la atención que son tan sólo textos de apoyo, o textos introductorios, que buscan fornecer subsidios metodológicos para profesores que deseen investigar en enseñanza.

Porto Alegre, 2016

Marco Antonio Moreira
Paulo Ricardo S. Rosa

# Capítulo 1

# Investigación En Educación En Ciencias: Métodos Cualitativos[1]

M. A. Moreira

**Resumen**

Después de una breve comparación entre los paradigmas cuantitativo y cualitativo, el texto enfoca sólo el enfoque cualitativo describiendo, con mapas conceptuales, la etnografía, el estudio de casos y la investigación acción. Al final es abordado el tema de la triangulación medotológica.

## Introducción

En este texto la investigación en educación en ciencias está entendida como la producción de conocimientos resultante de la búsqueda de respuestas a preguntas sobre enseñanza, aprendizaje, currículum y contexto educativo en ciencias, así como sobre el profesorado de ciencias y su formación permanente, dentro de un cuadro epistemológico, teórico y metodológico consistente y coherente. Sin embargo, dicho texto se ocupará sólo del domínio metodológico de esa investigación y en ese dominio se enfocará particularmente la *metodología cualitativa*.

La metodología de la investigación en educación en ciencias es la misma de la investigación en educación y ésta ha sido dominada, a lo largo del siglo XX, por dos paradigmas clásicos: uno inspirado en la metodología de las ciencias naturales enfatizando observaciones empíricas cuantificables y adecuadas para tratamientos estadísticos, el otro derivado del área humanística con énfasis en informaciones holísticas y cualitativas y en enfoques interpretativos. El filósofo alemán Wilhelm Dilthey argumentaba (apud Husén, 1988) ya en 1890 que las humanidades tenían su propia lógica de investigación y que la diferencia entre las ciencias naturales y las humanidades era que éstas buscaban comprender mientras que las primeras procuraban explicar (op. cit. p. 17). Dicha distinción nos parece hoy muy simplificada pero nos sirve para señalar que el debate es antiguo.

---

[1] Programa Internacional de Doctorado en Enseñanza de las Ciencias. Universidad de Burgos, España; Universidade Federal do Rio Grande do Sul, Brasil. *Texto de Apoyo n° 14.* Publicado em *Actas del PIDEC,* Vol. *4:*25-55, 2002.

La investigación en educación empieza, (según Landsheere, 1988), alrededor de 1900, bajo el nombre de "pedagogía experimental", con investigadores como Meumann en Alemania, Binet en Francia, Thorndike en Estados Unidos y Claparède en Suiza, poco tiempo después de la "psicología experimental" iniciada por Wundt en Leipzig alrededor de 1880, y fuertemente influenciada por ella. De acuerdo con ese mismo autor (op. cit. p. 11), en las tres primeras décadas del siglo pasado la investigación educativa ha tenido un acentuado énfasis cuantitativo, dirigido hacia el estudio de la eficacia en la enseñanza, particularmente en Estados Unidos. Posteriormente, en los años 30 a 50, la crisis económica y la guerra llevaron a una gran reducción en la actividad de investigación en educación, en especial en Europa. Sin embargo, en esa misma época aparecen como campo de interés de los investigadores los estudios de naturaleza sociológica cuestionando la escuela como mecanismo de reproducción de distinciones sociales y prácticas discriminatorias (ibid. p. 13). En las décadas de 60 y 70 otra vez hubo un período de mucho apoyo financiero a la investigación educativa, particularmente aquella volcada hacia el desarrollo curricular en ciencias y matemáticas. Fue también una época de predominio del enfoque cuantitativo, no obstante, la reacción a esa "tradicción positivista" empezaba a ser cada vez más fuerte en el contexto de la investigación educativa a nivel internacional. A tal punto que en los años 80 y 90 hubo un claro predominio del abordaje cualitativo en la investigación en educación en general y en ciencias en particular.

Esta pequeña y poco rigurosa reseña histórica fue hecha solamente para reforzar la aserción de que los dos paradigmas clásicos – el cuantitativo y el cualitativo – han dominado la investigación educativa en el siglo XX, con una cierta alternancia.

Desde el punto de vista epistemológico, en todo ese tiempo fueron planteadas tesis de incompatibilidad paradigmática kuhneana (Smith, 1983; Smith y Heshusius, 1986; Marshal, 1986), de compatibilidad práctica, funcional, pragmática (e.g., Shulman, 1981; Miles y Huberman, 1984), de conciliación y triangulación metodológicas (e.g., Eisner, 1981; Firestone, 1987) o integradoras como la de Keeves (1988) y la de Bericat (1998). Esos planteamientos están discutidos en otro texto anterior y complementario a este (Moreira, 2000).

En este texto nos quedaremos en la perspectiva integradora, pero este tema quedará para el final. Por el momento seguiremos en la distinción entre los dos paradigmas. Aunque esta distinción e incluso la idea de paradigma pueden ser objeto de críticas por parte de investigadores (e.g., Walker y Evers, 1988) continuaremos en ella, por algún tiempo, por razones didácticas. Por dichas razones, en el apartado siguiente distinguiremos entre los dos paradigmas clásicos como si constituyeran una dicotomía. Una vez establecida esa distinción, el texto enfocará sucesivamente metodologías cualitativas como la etnografia, el estudio de caso, la investigación acción y otras. En la conclusión abandonaremos la visión dictómica y defenderemos una postura integradora.

## Los dos paradigmas clásicos

En la tabla n° 1 se establece una comparación dicotómica entre los paradigmas cuantitativo y cualitativo en términos de presupuestos, objetivos, métodos, papel del investigador y retórica de presentación del conocimiento producido. Por ser autoexplicativa dicha tabla no será comentada.

Tabla n° 1. Un paralelo entre los paradigmas cuantitativo y cualitativo en la investigación educativa (M.A. Moreira, 2000)

| | ***Paradigma cuantitativo realista/racionalista*** | ***Paradigma cualitativo idealista/naturalista*** |
|---|---|---|
| ***Presupuestos*** | Realidad objetiva, independiente de creencias, con existencia propia. Investigar no afecta a lo que se está investigando. Los instrumentos son una manera de alcanzar mediciones precisas de objetos y eventos con existencia propia; instrumentos válidos son los que producen representaciones exactas de la realidad. Si el investigador deja de estudiar algo, ese algo continuará existiendo y permanecerá ligado a otras cosas de la misma manera. Dualismo sujeto-objeto. Verdad es una cuestión de correspondencia con la realidad (Smith, 83). | Realidad socialmente construida; no hay realidad independiente de los esfuerzos mentales de crear y moldear; lo que existe depende de la mente humana. Lo que se investiga no es independiente del proceso de investigación. Los instrumentos no tienen lugar independientemente de aquello a lo que se destinan para medir, son extensiones de los investigadores en su intento de construir o de dar forma a la realidad. La realidad no tiene existencia previa la investigación y dejará de existir si la investigación se abandonara. No hay dualismo sujeto-objeto. Verdad es cuestión de concordancia en un contexto (Smith, 83). |
| ***Objetivos*** | Procuran explicar causas de cambios en hechos sociales, principalmente a través de medición objetiva y análisis cuantitativo (Firestone, 87). Enfocan comportamientos de grupos o individuos (Eisner, 81). Buscan la predicción y control de eventos, algoritmos, verdades, universales abstractos a los que se llega a través de generalizaciones estadísticas de muestras para poblaciones (Erickson, 86). | Buscan la comprensión del fenómeno social según la perspectiva de los actores a través de participación en sus vidas (Firestone, 87). Enfocan significados y experiencias; acción en vez de comportamiento (Eisner, 81). Procuran la explicación interpretativa; heurísticas en vez de algoritmos; universales concretos alcanzados a través del estudio detallado de un caso y de la comparación con otros estudiados con igual detalle (Erickson, 86). |
| ***Métodos*** | Toman prestado el modelo de las ciencias físicas para investigar el mundo social y humano. Se ocupan de diseños experimentales, cuasi-experimentales y correlacionales; tests de hipótesis; instrumentos válidos y fidedignos; tests de significancia; muestreo; inferencia estadística; generalización. Siguen un modelo hipotético-deductivo. | Usan técnicas etnográficas, estudios de caso, antropología educativa. Se ocupan de observación participativa; significados individuales y contextuales; interpretación; desarrollo de hipótesis; indicadores de baja inferencia; casos, grupos o individuos específicos; particularización. Pueden hacer uso de estadística descriptiva. Son más bien inductivos. |
| ***Papel del investigador*** | Distante para evitar viés (Firestone, 87); objetivo. Se limita a lo que es. Cuantifica registros de eventos. Usa medios científicos. Busca fiabilidad y validez. | Inmerso en el fenomeno de interés (Firestone, 87); participante. Anota, oye, observa, registra, documenta, busca significados, interpreta. Procura credibilidad. |
| ***Retórica*** | Patronizada, estadística, objetiva. Extenso uso de tablas, gráficos, coeficientes. Procura neutralizar la personalidad del investigador. Fría, científica, buscando convencer al lector de que el análisis hecho es neutro, impersonal (Firestone, 87). | Persuasiva, descriptiva, detallada. Extenso uso de transcripciones, viñetas, documentos, ejemplos, comentarios interpretativos. Usa el lenguaje cotidiano con suficiente detalle para evidenciar que son válidas las interpretaciones de los significados tenidos por los actores (Erickson, 86). |

## *La investigación cualitativa (interpretativa)*

En la figura n° 1 se presenta un mapa conceptual para el enfoque cualitativo en la investigación educativa. Toda vez que los mapas conceptuales no son autoexplicativos lo comentaremos brevemente. En el tope aparece el concepto de *investigación cualitativa* al cual están asociados atributos como *interpretativa, holística, naturalista, participativa, interaccionista simbólica, constructivista, etnográfica, fenomenológica* y *antropológica.* El

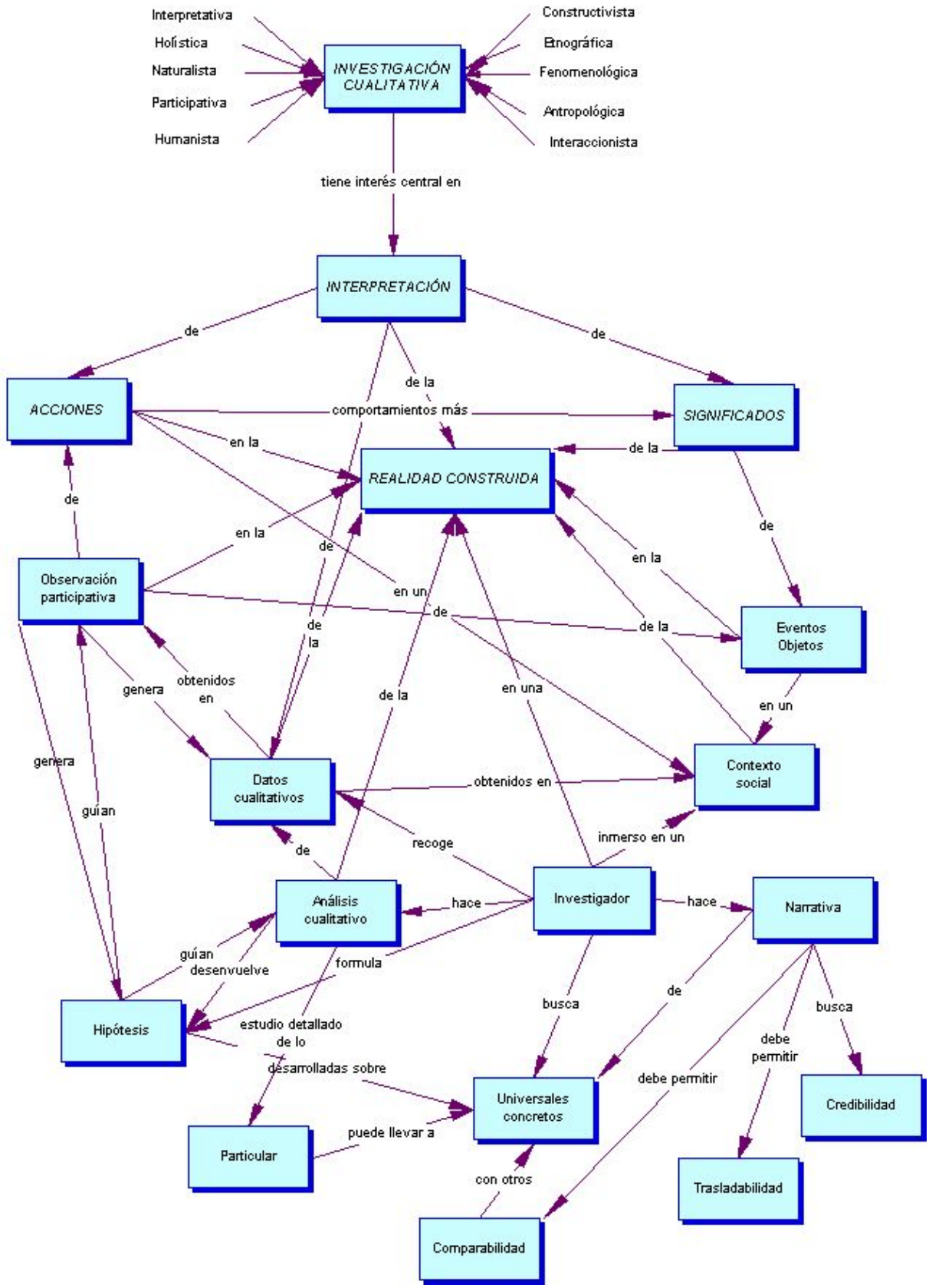



Figura 1. Un mapa conceptual para la investigación educativa cualitativa (M.A. Moreira, 2000)

interés central de esa investigación está en una *interpretación de los significados* atribuidos por los sujetos a sus *acciones* en una *realidad socialmente construida*, a través de *observación participativa*, es decir, el *investigador* queda *inmerso* en el fenómeno de interés. Los *datos* obtenidos por medio de esa participación activa son de *naturaleza cualitativa* y analizados correspondientemente. Las *hipótesis* son *generadas* durante el proceso investigativo. El investigador busca *universales concretos* alcanzados a través del estudio profundo de *casos particulares* y de la comparación de ese caso con otros estudiados también con gran profundidad. A través de una *narrativa detallada* el investigador busca *credibilidad* para sus modelos interpretativos.

La investigación cualitativa es llamada también *naturalista* porque no involucra manipulación de variables, ni tratamiento experimental (es el estudio del fenómeno en su acontecer natural); *fenomenológica* porque enfatiza los aspectos subjetivos del comportamiento humano, el mundo del sujeto, sus experiencias cotidianas, sus interacciones sociales y los significados que da a esas experiencias e interacciones; *interaccionista simbólica* porque toma como presupuesto que la experiencia humana es mediada por la interpretación, la cual no se da de forma autónoma sino que en la medida que el individuo interactúa con otro, es por medio de interacciones sociales es como van siendo construidas las *interpretaciones,* los significados, la visión de realidad del sujeto (André, 1998, p. 17-18).

Erickson (1986, p. 119), un muy conocido investigador en educación, prefiere el término *investigación interpretativa* para referirse a toda una familia de enfoques de investigación participativa observacional, en vez de investigación cualitativa, por ser más incluyente, por evitar la idea de que sea esencialmente no cuantitativa y por apuntar al interés central de esa investigación que es el significado humano en un contexto social y su dilucidación y exposición por el investigador. Para él (op. cit., p. 121), la investigación interpretativa involucra a) intensa y larga participación en el contexto investigado, b) cuidadosos registros de lo que ocurre en dicho contexto juntamente con otras fuentes de evidencia (e.g., apuntes, documentos, ejemplos de cosas hechas por los sujetos, grabaciones en audio o en video) y c) análisis reflexivo de todos esos registros y evidencias así como descripción detallada (i.e., utilizando la narrativa y transcripciones literales de verbalizaciones de los sujetos).

Para Erickson (op. cit., p. 129) la tarea de la investigación interpretativa es la de descubrir maneras específicas a través de las cuales formas locales y no locales de organización social y cultural se relacionan con actividades de personas específicas en sus elecciones y acciones sociales conjuntas. Para la investigación en el aula eso significa descubrir cómo las elecciones y acciones de todos los actores constituyen un currículum prescrito — un ambiente de aprendizaje. Profesores y alumnos juntos e interactuando adquieren, comparten y crean significados no sólo a través de los sistemas lingüístico y matemático sino también por medio de otros sistemas como la ideología política, los presupuestos de las subculturas étnicas y sociales respecto al papel de mujeres y hombres, a relaciones adecuadas entre adultos y niños, etc., es decir por aculturación.

Una vez presentadas algunas características generales de la investigación cualitativa pasaremos a enfocar, sucesivamente, tres metodologías principales dentro de ese enfoque: la *etnografía*, el *estudio de caso* y la *investigación acción*.

## *La etnografía*

La etnografía es una metodología, una herramienta para estudiar y comprender una cultura, la manera de vida de un grupo de personas, es decir, sus ideas, creencias, valores y presupuestos, sus comportamientos y las cosas que hacen. (Ogbu et al., 1988, p. 48). En otras palabras, *la etnografía es un intento de describrir una cultura* (André, 1998, p. 19). La investigación etnográfica consta esencialmente de una descripción de eventos que ocurren en el cotidiano de la vida de un grupo con especial atención a las estructuras sociales y conductas de individuos respecto a su status de pertenencia o membresía al grupo, y una interpretación de lo que significa todo eso para la cultura del grupo. (Taft, 1988, p. 71).

En la etnografía el investigador participa, lo más que puede, de la vida normal del grupo investigado, de la cultura investigada. La investigación es conducida en el escenario natural de los eventos, en el contexto en el cual ocurren los acontecimientos, a través de observación participativa. Para llegar a una *comprensión descriptiva contextualizada* de la cultura, el investigador tiene que meterse en dicha cultura, aprender el *"lenguaje nativo"*, como dijo el célebre antropólogo Malinowski, *interactuar* con los miembros de esa cultura, desarrollar una *comprensión empática* de la vida de las personas tal como ellas la perciben, así como una *perspectiva holística* del grupo. Todo eso, por cierto, implica un largo "tiempo de residencia" en esa cultura. Es decir, el investigador debe permanecer "inmerso" en la cultura investigada durante un periodo de tiempo "suficientemente grande" para, de acuerdo con Malinowski, contextualizar los datos en un "account" holístico y coherente y describir "la vida tal como es vivida" (Ogbu et al., 1988, p. 50).

El investigador etnográfico tiene, consecuentemente, un doble papel: participante y observador. Por un lado, él tiene que envolverse con el grupo, "aculturarse" a ello. Por otro lado, debe ser capaz de observar, interpretar, discernir, desarrollar perspectiva holística. Siendo al mismo tiempo observadores y participantes, los investigadores etnográficos no son desprendidos del fenómeno de interés; ellos influencian sus datos y son influenciados por éstos en todas las etapas de observación, interpretación y descripción (Taft, 1988, p. 72). La gran ventaja de ser observador participante parece ser al mismo tiempo la principal dificultad que debe enfrentar el investigador participante. Al mismo tiempo que intenta "pertenecer" a la cultura investigada, él o ella debe también ser capaz de "mirarla desde fuera", interpretarla, describirla.

La observación participativa es la principal técnica de investigación etnográfica. Sin embargo, las entrevistas son también muy utilizadas. Los datos generados por esas dos técnicas son frecuentemente complementados por otros como documentos, narrativas, historias de vida, artefactos, diagramas, producidos por en el grupo investigado. En general, el investigador etnográfico busca recoger toda la información que puede, no sólo a través de observación participativa y entrevistas, para interpretarla inductivamente y construir una realidad social que es su comprensión descriptiva contextualizada de la cultura investigada.

La metodología etnográfica es cualitativa y holística, haciendo uso de la intuición, empatía y otras habilidades del investigador para interpretar descriptivamente una cultura. Su interés está en descubrir (en el sentido de construir una descripción comprensiva contextualizada) y no en verificar. Sin embargo, eso no implica no tener ninguna hipótesis o teoría inicial. El investigador etnográfico no empieza un trabajo de campo "sin tener nada en la cabeza". Eso no existe. Él o ella siempre tendrán conocimientos teóricos previos que de algúna manera van a orientar sus pasos iniciales, pero no deben tener hipótesis y teorías que serán verificadas o rechazadas en el estudio. Es decir, el investigador etnográfico no debe

tener ideas preconcebidas, tal como ha recomendado Malinowski (apud Taft, 1988, p. 74). Las hipótesis son formuladas recursivamente durante el proceso, durante el desarrollo de la investigación. Gradualmente puede emerger una base teórica para comprensión de los procesos grupales. Esa base teórica es conocida como *teoría fundamentada*, es decir, fundamentada en el propio proceso de investigación (ibid.), o fundamentada en los datos. Dicha teoría que fue desarrollada de manera inductiva probablemente generará hipótesis útiles para guiar, inicialmente, nuevas observaciones participativas. Sin embargo, no se está hablando aquí del inductivismo científico ingenuo tan criticado epistemológicamente, ni de hipótesis que serán "comprobadas" en estudios "más rigurosos". (El tema de la credibilidad de los "resultados" de la investigación cualitativa será discutido más adelante en otro apartado.)

La etnografía es una metodología de investigación en antropología que ha llegado a la investigación en educación no hace mucho, en los años 60 del siglo pasado. Una gran diversidad de etnografías educativas han sido desarrolladas desde esa época, sin embargo, el concepto de cultura ha permanecido como constructo unificador. Tres orientaciones principales pueden ser identificadas (Ogbu et al., 1988, p. 50-51) a partir de distintos niveles de análisis y diferentes énfasis en sus definiciones de cultura: *etnografía holística* (también conocida como etnografía tradicional, vieja etnografía o macroetnografía), la *etnociencia* (también llamada nueva etnografía o antropología cognitiva) y la *microetnografía* (etnografía de la comunicación).

La *etnografía holística*, de la cual hemos hablado hasta ahora, intenta describir la cultura, o el grupo, como un todo mientras que la etnociencia y la microetnografía focalizan unidades mucho más pequeñas como palabras, individuos o escenas (ibid.).

La *microetnografía* es una etnografía enfocada, es decir, una etnografía que se ocupa de mirar repetidas veces y de analizar detalladamente registros audiovisuales de interacciones humanas en escenas-clave, en situaciones-clave de interacción social, acompañadas de observación participativa del contexto más amplio en el cual dichas escenas ocurren (op. cit., p. 51). Es una etnografía de la comunicación, enfocando sujetos individuales y su discurso en ciertos escenarios.

La *etnociencia* se aleja de la etnografía holística tradicional al definir cultura primariamente en términos de cogniciones de las personas. Sus presupuestos básicos son que el contenido de los datos culturales consta de reglas, códigos y un ordenamiento ideativo de la sociedad que está organizado en distintos dominios culturales de conocimiento. Las experiencias son codificadas en "lexemes" o palabras; por tanto, el lenguaje es la principal fuente de datos culturales y técnicas de estudio del lenguaje pueden ser aplicadas al estudio de la cultura ideativa o cognición. Consecuentemente, hay menos énfasis en la observación participativa y más énfasis en la recolección de vocabularios sobre eventos particulares, así como en los esquemas clasificatorios (op.cit., p. 52). Relacionada con la etnociencia está la *etnometodología* que según André (1998, p. 18) no es exactamente una metodología sino un campo de investigación: es el estudio de cómo los individuos comprenden y estructuran su cotidiano, es decir, es el intento de descubrir "los métodos" que las personas utilizan en su día-a-día para entender y construir la realidad que las rodea. En consecuencia, sus principales focos de interés son los conocimientos tácitos, las formas de comprensión del sentido común, las prácticas cotidianas y las actividades rutinarias que moldean las conductas de los actores sociales (ibid.).

Independiente de esos aparentemente distintos tipos de etnografía, podemos caracterizarla de manera general como el intento de descripción de una cultura. La principal preocupación en la etnografía se refiere al significado que tienen acciones y eventos para las personas o grupos estudiados (op. cit., p. 19). La etnografía es un esquema de investigación desarrollado por los antropólogos para estudiar una cultura y una sociedad. Etimologicamente, etnografia significa "descripción cultural" (ibid., p. 27). En educación, rigurosamente hablando, lo que se hace son estudios etnográficos, es decir, una adaptación de la etnografía a la educación, una vez que el fenómeno de interés de la investigación educativa es, en último análisis, el proceso educativo, no una cultura o un grupo social en sí mismos. Dichos estudios etnográficos han incluido, por ejemplo, un aula en particular, un pequeño grupo en un aula o en una escuela, escenas o diálogos en el aula, relaciones escuela-comunidad, etc.

A modo de conclusión de este apartado correspondiente a la etnografía se presenta en la figura n° 2 un mapa conceptual para la etnografía. En el tope aparece como concepto más abarcador el propio concepto de *etnografía* que puede ser *holística* (la etnografía tradicional o "vieja" etnografía), *microetnografía* (la etnografía de la comunicación) o *etnociencia* (la antropología cognitiva o la "nueva" etnografía). Sin embargo, la etnografía es siempre un intento de describir una *cultura* (o una microcultura) que es caracterizada principalmente por *significados*, construidos y compartidos por el grupo social, es decir, por el "punto de vista nativo" (ideas, creencias, valores, presupuestos), según Ogbu et al. (1988, p. 50). La descripción de una cultura requiere *observación participante, trabajo de campo* (durante un *tiempo suficiente*) e *interacción personal* (intersubjetividad, empatía) en un *contexto natural*. La etnografía busca *descripción* y utiliza *inducción* para llegar a una *realidad construida*; las *hipótesis* son desarrolladas a lo largo del proceso y las teorías emergen de los datos, es decir, son *teorías fundamentadas* (en ese sentido, los métodos cualitativos son inductivos). El resultado de todo el proceso es una *comprensión descriptiva contextualizada*, de un grupo social, de unas escenas, de un discurso, de unas cogniciones o, en términos más abarcadores y originales, de una cultura.

## El estudio de casos

De acuerdo con Sturman (1988, p. 61), estudio de caso es un término genérico para la investigación de un individuo, un grupo o un fenómeno. Mientras las técnicas usadas en esa investigación pueden variar e incluir tanto enfoques cualitativos como cuantitativos, la característica que más distingue el estudio de caso es la creencia de que los sistemas humanos desarrollan una completud e integración, es decir, no son simplemente un conjunto de partes o de trazos. Consecuentemente, el estudio de caso encaja en una tradición holística de investigación según la cual las características de una parte son determinadas grandemente por el todo al cual pertenece. La comprensión de las partes requiere la comprensión de sus interrelaciones en el todo. Es una visión sistémica que presupone que los elementos de un evento educativo, por ejemplo, son interdependientes e inseparables y un cambio en un elemento implica un cambio en todo lo demás.

Por lo tanto, hacer una investigación del tipo estudio de caso, es decir, para entender un caso, para comprender y descubrir cómo las cosas ocurren y por qué ocurren, para quizás predecir algo a partir de un único ejemplo o para obtener indicadores que puedan ser usados en otros estudios (quizás cuantitativos) es necesario un profundo análisis de las interdependencias de las partes y de los patrones que emergen. Lo que se requiere es un estudio de patrones, no de variables aisladas (ibid.). Para todo eso, las técnicas de investigación cualitativa son frecuentemente las más adecuadas.

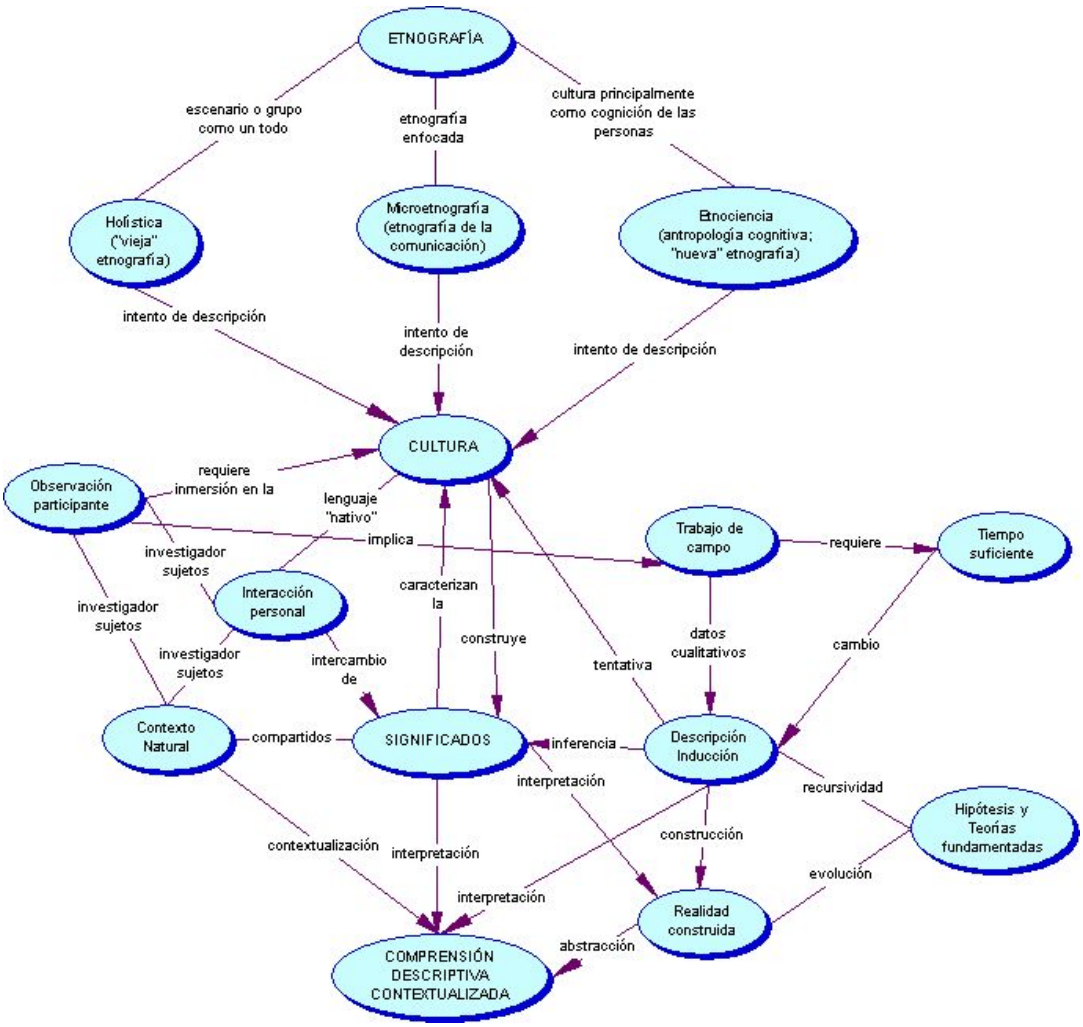



Figura n° 2. Un mapa conceptual para la etnografía (Moreira, 2002).

El estudio de casos se utiliza hace mucho tiempo en campos como el derecho, la medicina, la psicología y la administración. Sin embargo, en la investigación educativa el uso de esa metodología es más reciente y dentro de una concepción bastante restringida, o sea, el estudio descriptivo de una unidad que puede ser una escuela, un profesor, un alumno o un aula (André, 1998, p. 21). Puede también, por ejemplo, ser el estudio de un currículum o de un sistema escolar.

¿Cuál sería la diferencia entre estudio de caso y etnografía? Según André (ibid.) se puede hacer un estudio de caso etnográfico, es decir, aplicar el enfoque etnográfico al estudio de un caso. Por otro lado, no todos los tipos de estudios de caso encajan en la perspectiva etnográfica ni todos los estudios etnográficos son estudios de caso.

¿ Cuáles serían entonces los distintos tipos de estudios de caso?

En primer lugar hay que distinguir entre estudio de caso y conceptos afines. De acuerdo con Serrano (1998, p. 85), el *trabajo con casos* denota los procedimientos correctivos, remediales, de desarrollo o ajuste que siguen al diagnóstico de las causas de desajuste; el *método de casos* es una estrategia didáctica en la cual los elementos principales del estudio de

casos se presentan a los estudiantes con propósitos ilustrativos, sin necesidad de darles una visión completa de los hechos (el propósito es más bien el de establecer un marco de discusión y debate); la *historia de casos* es la búsqueda del pasado de una persona, grupo o institución; el *estudio de casos,* a su vez, puede definirse como una descripción intensiva, holística y un análisis profundo de una entidad singular, un fenómeno o unidad social.

Antes de hablar de tipos de estudios de casos es interesante también destacar sus rasgos esenciales. Según esa misma autora (op. cit., p. 91), las propiedades esenciales de un estudio de casos cualitativo son la *particularización* (se centran en una situación, evento, programa o fenómeno particular), la *descripción* (el producto final es una descripción rica y densa del objeto de estudio), la *heurística* (iluminan la comprensión del lector respecto al objeto de estudio) y la *inducción* (se basan en el razonamiento inductivo; las teorías, los conceptos o las hipótesis surgen de un examen de los datos fundados en el contexto mismo).

Respecto a tipos de estudios de caso, Serrano (ibid., p. 97) argumenta que pueden clasificarse por la naturaleza del informe final, independientemente de su orientación disciplinaria o área de interés, en *descriptivos, interpretativos* y *evaluativos.*

*Estudios de caso descriptivos* (ibid.) se caracterizan por un informe detallado de un fenómeno objeto de estudio sin fundamentación teórica previa; son enteramente descriptivos, no se guían por generalizaciones establecidas o hipotéticas, ni desean formular hipótesis o teorías.

*Estudios de casos interpretativos* (op. cit., p. 98): contienen descripciones ricas y densas; sin embargo, los datos descritos los utilizan para desarrollar categorías conceptuales o para ilustrar, defender o desafiar presupuestos teóricos difundidos antes del estudio. El investigador debe reunir tanta información sobre el objeto de estudio como le sea posible, con la pretensión de interpretar o teorizar sobre el fenómeno.

*Estudios de casos evaluativos* (ibid.): implican descripción, explicación y juicio; sobretodo, este tipo de estudio de casos sopesa la información para emitir un juicio; la emisión de juicios es el acto final y esencial de la evaluación.

No obstante, la misma autora plantea que aunque se pueda establecer esta clasificación y aunque algúnos estudios de casos puedan ser puramente descriptivos, en educación la mayoría de los estudios de casos son una combinación de descripción y evaluación o de descripción e interpretación.

El estudio de casos interpretativo nos remite otra vez al tema de la *teoría fundamentada* referido en la etnografía. Este tipo de estudio de casos parece ser una metodología ideal para fundamentar una teoría, es decir, para inducir una teoría a partir de datos descriptivos muy ricos. Sin embargo, no se trata de una teoría formal en sentido usado en las ciencias naturales, tampoco del "método inductivista". Son más bien categorías, hipótesis comprensivas.

Otra clasificación de tipos de estudios de caso es aportada por Stenhouse (1985, apud Sturman, 1988, p. 63):

*Estudio de caso etnográfico*, del cual ya hablamos, que involucra el estudio profundo de una entidad singular generalmente a través de observación participante y entrevistas.

*Estudio de caso investigación acción* en el cual el foco está en generar un cambio en el caso bajo estudio.

*Estudio de caso evaluativo* que involucra la evaluación de programas y en el cual un trabajo de campo más condensado muchas veces reemplaza el enfoque etnográfico más demorado.

*Estudio de caso educativo* que está diseñado para mejorar la comprensión de la acción educativa.

Como se puede percibir de las clasificaciones de Serrano y Stenhouse, es difícil separar el estudio de casos de otros tipos de investigación cualitativa como la etnografía y la investigación acción. Podemos caracterizar bien el estudio de casos argumentando que su preocupación central es la comprensión de una instancia singular, lo que significa que el objeto estudiado es caracterizado como único, como una representación singular de la realidad que es multidimensional e históricamente ubicada (André, 1998, p 21). Sin embargo, como hemos visto, un estudio de caso puede hacerse a través de una etnografia o de una investigación acción, por ejemplo.

Las características esenciales de los estudios de caso, así como sus distintos tipos están mapeados conceptualmente en la figura n° 3. Los dos conceptos claves son *estudio de casos* e *instancia singular*. En la parte superior del mapa, encima del concepto de estudio de casos están sus propiedades esenciales (*inducción, particularización, heurística* y *descripción*); en la parte inferior, debajo del concepto de instancia singular aparecen instancias de dicho concepto. En el eje central del mapa están los diferentes tipos de estudios de caso identificados por Serrano (1998) y Stenhouse (1985). Los conectores intentan explicitar las relaciones entre los conceptos y las flechas sugieren ciertas convergencias.

## *La investigación-acción*

El objetivo fundamental de la investigación-acción consiste en mejorar la práctica en vez de generar conocimientos. La producción y utilización del conocimiento se subordina a este objetivo y está condicionado por él (Eliott, 1993, p. 67). La mejora en la práctica consiste en implantar aquellos valores que constituyen sus fines, por ejemplo, la educación en la enseñanza (ibid.). Sin embargo, el concepto de educación como fin de la enseñanza transciende la conocida distinción entre proceso y producto. La mejora de la práctica supone tener en cuenta a la vez los resultados y los procesos.

Según Kemmis y McTaggart (1988; apud Kemmis, 1988, p. 174), la investigación-acción es definida como una forma de investigación *colectiva* auto-reflexiva emprendida por participantes de situaciones sociales para mejorar la productividad, racionalidad y justicia de sus propias prácticas sociales o educativas, así como su comprensión respecto a dichas prácticas y respecto a las situaciones en que ocurren. Los participantes pueden ser profesores, alumnos, directores, padres y otros miembros de la comunidad, es decir, cualquier grupo que comparta una preocupación, un objetivo. Es una *investigación colaborativa*; sin embargo, es importante enfatizar que esta acción colaborativa depende de que cada individuo examine críticamente sus proprias acciones (ibid.).

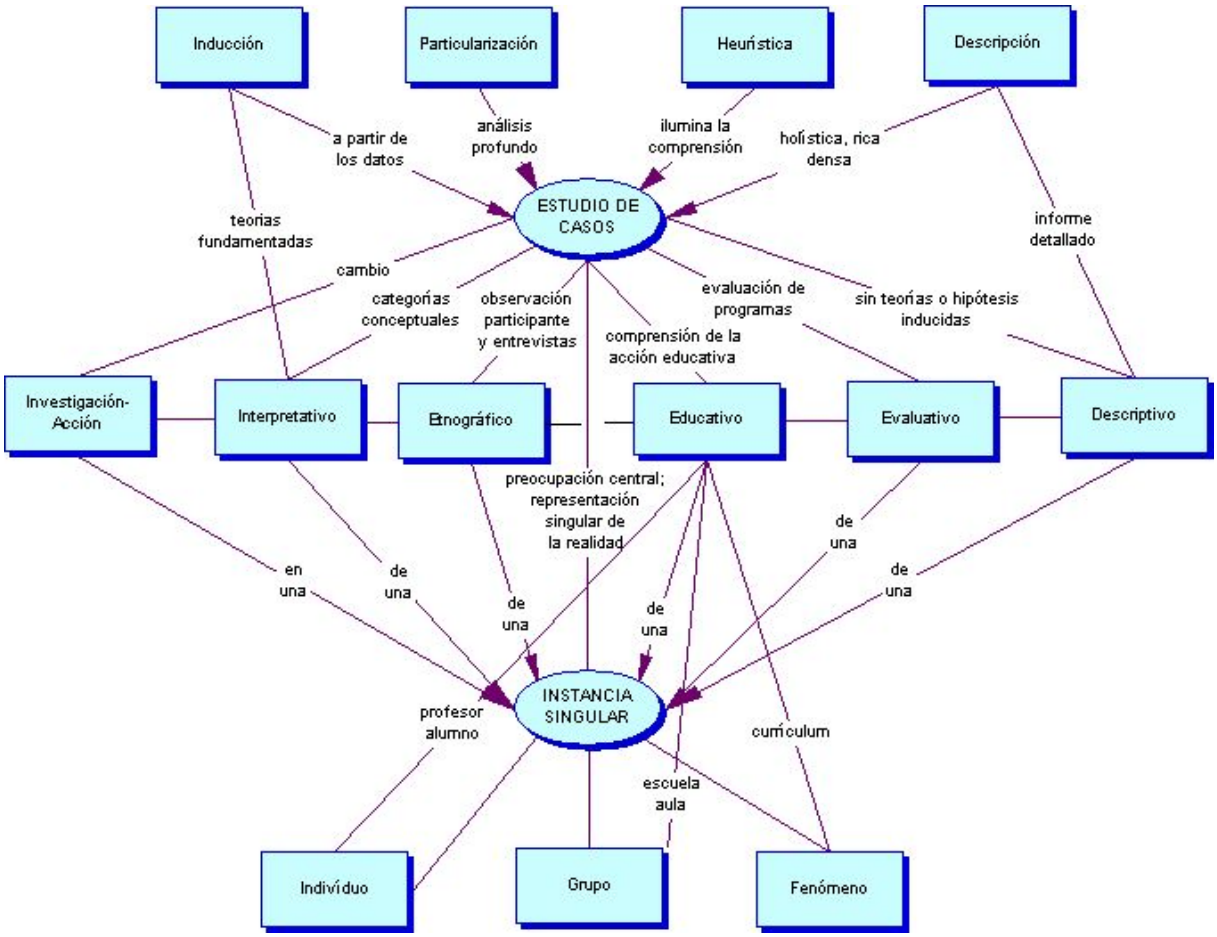



Figura n° 3. Un mapa conceptual para estudio de casos (Moreira, 2002).

En educación, cuando se pretende mejorar la práctica, hay que considerar conjuntamente los procesos y los productos. Este tipo de reflexión simultánea sobre la relación entre procesos y productos constituye, según Eliott (op. cit., p. 68), una característica fundamental de la investigación-acción. Es una *práctica reflexiva* que aspira a mejorar la concreción de los valores del proceso, muy diferente del razonamiento técnico que versa sobre los medios para conseguir un fin. Para Elliot (op. cit., p. 69), es a la vez ético y filosófico. En la medida en que la reflexión trata de la elección de un curso de acción en un determinado conjunto de circunstancias para llevar a la práctica los propios valores, reviste carácter ético. Pero como la elección ética supone la interpretación de los valores que han de traducirse a la práctica – la reflexión sobre los medios no puede separarse de la reflexión sobre los fines – la reflexión ética tiene una dimensión filosófica (ibid.).

En la investigación-acción, los profesores son incentivados a cuestionar sus propias ideas y teorías educativas, sus propias prácticas y sus propios contextos como objetos de análisis y crítica (Kemmis, 1988, p. 174). Desde una reflexión cuidadosa los profesores pueden desvelar ideas o suposiciones teóricas que resultan injustificadas y los dejan perdidos en su tarea docente; por ejemplo, si tienen suposiciones muy rígidas respecto a la naturaleza de habilidades innatas de los estudiantes (ibid.). Análogamente, los docentes, a través de la reflexión crítica, pueden concluir que prácticas antiguas moldeadas por hábito y tradición son inútiles o irrelevantes en los tiempos actuales; por ejemplo, prácticas disciplinares que

funcionaban antes hoy ya no son aceptables o son contraproductivas (ibid.). Respecto al contexto, ellos pueden llegar a la conclusión de que su estructura es inadecuada y obstaculiza el alcance de metas educativas; por ejemplo, la estructura física del aula puede dificultar el trabajo en grupos, la interacción personal, la enseñanza centrada en el alumno.

No obstante, no hay que olvidarse de que la investigación-acción es una investigación colectiva, colaborativa. La reflexión personal es importante, pero el verdadero cambio viene de la auto-reflexión colectiva. Que los participantes sientan la necesidad de iniciar cambios, de innovar, es condición necesaria antecedente de la investigación-acción, pero no suficiente.

El proceso de investigación-acción, según Kemmis y McTaggart (1988) y Elliot (1993, p. 88), se caracteriza por una espiral de ciclos de reconocimiento (descubrimiento de hechos): reconocimiento de una situación que se quiere cambiar; planificación general de la acción objetivando el cambio; *desarrollo, implementación y evaluación* de esa acción; reflexión a la luz de la evidencia recogida en la implementación; revisión del plan general; planificación de nueva acción; implementación, evaluación, reflexión, revisión del plan; planificación e implementación de una tercera acción...

Naturalmente, este carácter cíclico no significa un proceso lineal, automático, mecánico. Tal como se ha dicho en el comienzo de este apartado, la investigación-acción, a través de esa espiral de ciclos, tiene por objetivo la *mejora* de las prácticas y comprensiones de situaciones, y el *envolvimiento* de tantos cuanto sea posible de todos los afectados íntimamente por las acciones en todas las fases del proceso investigativo. La investigación-acción es un proceso colaborativo, auto-reflexivo en el cual el envolvimiento directo de los profesores y otros implicados, en la recolección de datos, análisis, crítica, reflexión, crea inmediatamente un sentido de responsabilidad respecto a la mejora de la práctica (Kemmis, 1988, p. 174).

La investigación-acción unifica procesos considerados a menudo independientes; por ejemplo: la enseñanza, el desarrollo del currículum, la evaluación, la investigación-educativa y el desarrollo profesional (Eliott, 1993, p. 72). La enseñanza, por ejemplo, en el marco de la investigación acción se concibe como una forma de investigación encaminada a comprender cómo traducir los valores educativos a formas concretas de práctica. El desarrollo del currículum no es un proceso antecedente a la enseñanza; el desarrollo de programas curriculares se produce a través de la práctica reflexiva de la enseñanza (ibid.). La investigación-acción no refuerza la postura de los profesores en cuanto conjunto de individuos que operan de forma independiente y autónoma, que no comparten sus reflexiones con los demás.

De una manera general, se puede decir que la investigación-acción siempre implica un plan de acción basado en objetivos de cambio (mejora), la implementación y control de ese plan a través de fases de acción, así como la descripción concomitante del proceso cíclico resultante. Sin embargo, Kemmis y McTaggart (1988, apud Kemmis, 1988) identifican varias características básicas de la investigación-acción que ayudan a distinguirla de otros tipos de investigación cualitativa. Según ellos, la investigación-acción:

- es un enfoque para mejorar la educación a través de cambios y para aprender desde las consecuencias de los cambios;
- se desarrolla a través de una espiral auto-reflexiva de ciclos de planificación, acción, observación sistemática, reflexión, replanificación, nueva acción, observación y reflexión;
- es participatoria, las personas trabajan para mejorar sus propias prácticas;

- es colaborativa, crea grupos auto-críticos que participan y colaboran en todas las fases del proceso investigativo;
- involucra los participantes en un proceso de teorización sobre sus prácticas, cuestionando circunstancias, acciones y consecuencias de esas prácticas;
- requiere que las personas pongan en cheque sus ideas y suposiciones respecto a instituciones;
- es abierta respecto a lo que cuenta como evidencia, o datos, pero siempre implica mantener y analizar registros de las consecuencias de las acciones implementadas;
- permite que los participantes al mismo tiempo mantengan registros de sus propios cambios personales y analicen críticamente las consecuencias de esos cambios;
- empieza pequeña; normalmente con pequeños cambios que un pequeño grupo, o quizás una sola persona, pueda intentar, pero se desplaza, gradualmente, hacia cambios más extensivos;
- requiere que los participantes analisen criticamente las situaciones (aulas, escuelas, sistemas educativos) en los cuales trabajan;
- es un proceso político porque involucra cambios en las acciones e interacciones que constituyen y estructuran prácticas sociales; dichos cambios típicamente afectan las expectativas e intereses de otros más allá de los participantes inmediatos en esas acciones e interacciones.

Tal como se ha hecho en las secciones anteriores, ésta finaliza con un mapa conceptual respecto al tema enfocado. La figura n° 4 presenta un mapa conceptual para investigación-acción. El concepto central es *cambio*: la investigación-acción tiene como meta mejorar la *práctica* a través del cambio. Es también central la tríada *(re) planificación* ↔ *acción* ↔ *(auto) reflexión* que carateriza el *proceso cíclico* de la investigación-acción. Por otro lado, la investigación-acción es un proceso *participativo, colectivo, colaborativo, político, auto-reflexivo, auto-crítico, auto-evaluativo* que requiere el envolvimiento de los *participantes* en todas las fases y en todos los aspectos característicos de ese proceso.

## Otros tipos de investigación cualitativa

Hemos empezado este texto comparando los enfoques cualitativo y cuantitativo en téminos bastante dicotómicos; después enfocamos características generales de la investigación cualitativa y pasamos en seguida a describrir la etnografía, el estudio de casos y la investigación-acción que consideramos las tres metodologías principales de ese enfoque. Sin embargo, hay otras de las cuales presentaremos algúnas, sucintamente, a continuación. Una vez concluidas esas breves presentaciones volveremos a temas generales, en el sentido de que se aplican a varias metodologías cualitativas, focalizando tópicos como la presentación y análisis de datos, la fiabilidad y validez de los estudios cualitativos, posibilidades de generalización, triangulación y otros. Al final, retomaremos la cuestión de los paradigmas desde una mirada integradora.

## La fenomenografía

La fenomenografía es el estudio empírico de los distintos modos a través de los cuales las personas vivencian, perciben, aprehenden, comprenden, o conceptualizan varios fenómenos en él, y aspectos del, mundo en su entorno. Las palabras vivencia, percepción, comprensión o conceptualización son usadas de manera intercambiable. Sin embargo, eso no significa que no hay diferencias en sus significados, sino más bien que el número limitado de

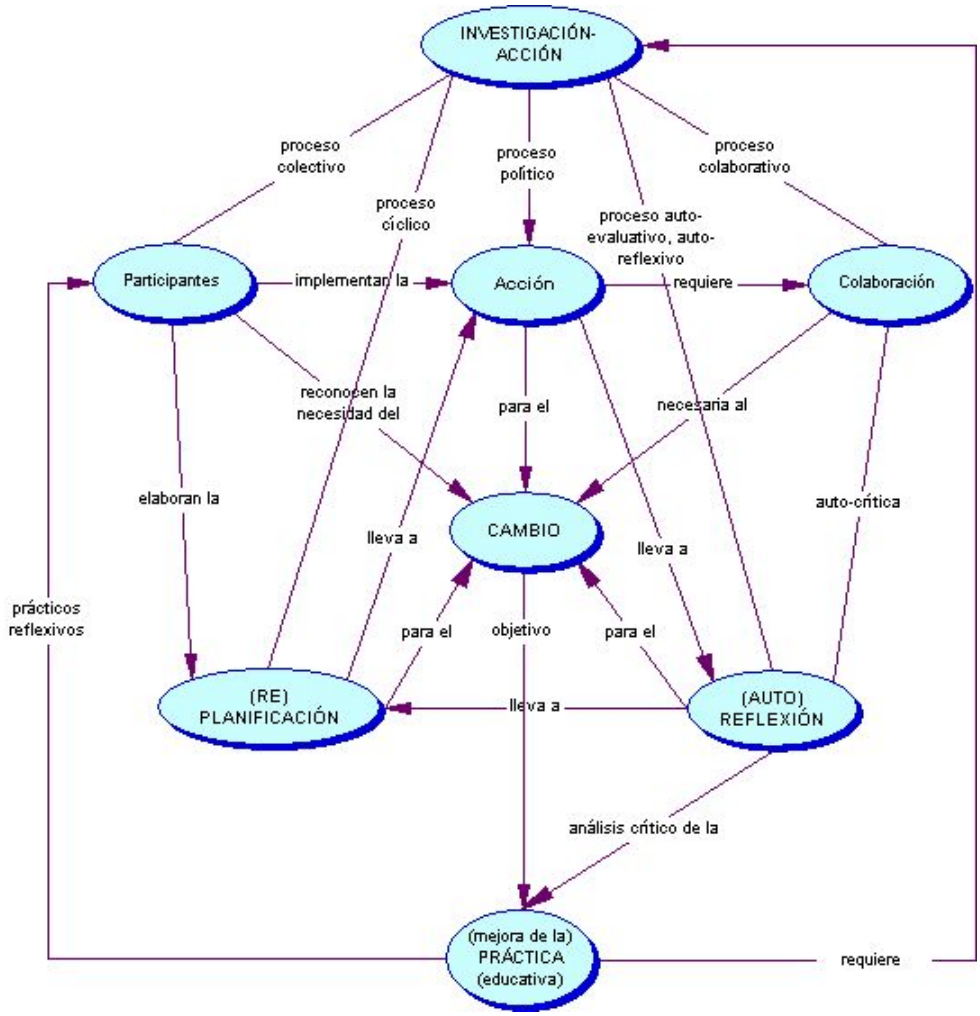



Figura n° 4. Un mapa conceptual para investigación-acción (Moreira, 2002).

maneras a través de las cuales un cierto fenómeno es interpretado por las personas puede ser identificado, por ejemplo, independiente de si están embebidos en la experiencia inmediata del fenómeno o en un reflexión sobre el mismo fenómeno (Marton, 1988, p. 95-97). Las diferentes experiencias, comprensiones, percepciones, etc., son caracterizadas en términos de "categorías de descripción" logicamente relacionadas y jerarquizadas. Dichas categorías representan distintas capacidades de lidiar con (o entender) un fenómeno. Como algúnas maneras de experienciar un fenómeno son más eficientes que otras respecto a algún criterio, es posible establecer una jerarquía de categorías de descripción (ibid.). La fenomenografía se aplica, por ejemplo, al estudio de la resolución de problemas o en investigaciones que buscan diferencias críticas en los significados atribuidos a ciertos fenómenos, conceptos o princípios claves en un cierto campo de conocimientos.

## *La hermenéutica*

La hermenéutica fue originalmente definida como el arte, o la ciencia, de la interpretación, en particular de la Biblia. Sin embargo, contemporáneamente se la define como la teoría y la práctica de la interpretación y comprensión en distintos tipos de contextos humanos (religiosos, seculares y cotidianos). Es decir, la hermenéutica no ya se refiere sólo a la exégesis y interpretación de textos, sino que considera la comprensión y la interpretación como una marca definitiva de la existencia humana y de la vida social (Ödman y Kerdeman,

1988, p. 185). Así como la hermenéutica, la fenomenología también se ocupa de la estructura de la comprensión; sin embargo, la fenomenología construye la comprensión primeramente en términos de constructos y funciones cognitivos, mientras que para la hermenéutica la comprensión no es solamente una función cognitiva, es también la condición ontológica de la existencia humana (op. cit., p. 186). Así como la teoría crítica, la hermenéutica mantiene que comprensión y significado son constituyentes de la vida social; no obstante, la hermenéutica mantiene que una vez que el sujeto está siempre envolvido en algún proceso de comprensión es imposible captar en cualquier forma final o definitiva los significados embebidos en una tradición. La hermenéutica, por lo tanto, evita (*eschews*) el intento de fundamentar la comprensión en un marco teórico o en un método y se concentra en interpretar culturas desde dentro de ciertas situaciones y contextos (ibid.). En la investigación educativa, la hermenéutica puede profundizar la comprensión del fenómeno educativo focalizando los significados que subyacen a estrategias y prácticas educativas específicas. Por ejemplo, ¿cómo deben ser interpretadas ciertas prácticas administrativas? Cuáles son los significados subyacentes? Análogamente, la hermenéutica puede profundizar la comprensión de los significados y propósitos subyacentes a un currículum.

## *La investigación participatoria*

La investigación participatória es descrita por Hall (1988, p. 198), de una manera general, como un proceso que combina tres actividades: investigación, educación y acción. Es una acción social sesgada en favor de los dominados, explotados, pobres, excluidos. La preocupación por poder y democracia y sus interacciones es central en la investigación participatoria. Es también crítica la atención a género, raza, etnicidad, orientación sexual, habilidades físicas y mentales, y otros factores sociales (ibid.). La investigación participatoria está diseñada para contribuir a los procesos de cambio de poder o democratización en una variedad de contextos. En la investigación participatoria no hay recetas ni ortodoxias metodológicas: las cuestiones y los métodos de trabajo deben fluir de los sujetos involucrados y de su contexto.

(No se debe confundir investigación participatoria con observación participante que es una mezcla de varias técnicas, un estilo o estrategia de investigación, en la cual, como ya hemos visto, el principal instrumento es el propio envestigador que debe quedar inmerso en el escenario para oir, ver y empezar a experienciar la vida como los sujetos la viven; Ball, 1988, p. 310).

## *Historia oral*

Como una técnica de investigación la historia oral va más allá de lo que el entrevistado contesta a las preguntas del entrevistador. Todos los matices de su testimonio son significativos: la no-respuesta, el silencio, la vacilación, todo cuenta como evidencia. Las fuentes de historia oral son, portanto, más que una cuestión de evidencia hablada y más que simplemente registrar factos (Hyams, 1988, p. 91).

Como un enfoque a la investigación la historia oral permite dar voz a los "sin voz", a los anónimos. Es como que construir la historia desde abajo hacia arriba. Aunque sufra restricciones entre los historiadores (por ejemplo, sólo los sobrevivientes son entrevistados; no se puede generalizar) este enfoque es usado en otros campos, como el de la educación. Por ejemplo, las actitudes de los profesores respecto al sistema educativo o a las reformas

educativas pueden ser obtenidas de manera más expansiva, más abierta. Las reminiscencias de los profesores pueden permitir una mejor comprensión de la implementación de las políticas educativas. La historia oral tiene potencial para explicar interpretaciones no-oficiales de problemas educativos.

El carácter único del testimonio individual es considerado importante para construir una historia. Sin embargo, como fue dicho antes, hay que tener en cuenta mucho más que la información factual provista. Se trata de buscar información que no está en los registros escritos, en las biografías.

## *Fiabilidad, generalización y validez*

La *fiabilidad* se refiere al grado en el que se pueden replicar las medidas y los estudios. En un enfoque cuantitativo la fiabilidad de las mediciones y de los instrumentos es un requisito básico, sin embargo, en un enfoque cualitativo dicho concepto no tiene sentido o debe tener otro significado, pues, en gran medida, el investigador es el principal instrumento o, en otras palabras, el instrumento es una extensión del investigador.

Respecto a los estudios, la fiabilidad exige que un investigador que utilice los mismos métodos que otro, llegue a idénticos resultados (Goetz y Lecompte, 1988, p. 214). Ello plantea un enorme problema en las investigaciones sobre el comportamiento natural o de los fenómenos únicos. El establecimiento de la fiabilidad de un diseño interpretativo se complica aún más por la naturaleza de los datos y del proceso de investigación, por los usos de la presentación de los resultados y por la visión de mundo de los investigadores en ese campo. (ibid.)

De acuerdo con esos autores (op. cit., p. 215), comparados con los diseños experimentales de laboratorios, estrictamente controlados, o con los experimentos de campo, los diseños de la investigación naturalística parecen resistirse a todo intento de replicación. Por ejemplo, los problemas de unicidad e idiosincrasia pueden llevar a afirmar que es imposible replicar un estudio etnográfico o un estudio de caso. Además, puesto que el comportamiento humano nunca es estático, ningún estudio, independientemente de sus métodos y diseños, puede ser replicado con exactitud (ibid.).

Sin embargo, se puede contrargumentar que dichas unicidades e indiosincrasias no son tan extremas que no tengan algún grado de "representatividad", algúnos trazos semejantes, a otras. Es decir, los grupos o fenómenos sociales o individuos investigados pueden no ser tan únicos e idiosincrásicos que no tengan nada que ver con otros grupos, fenómenos o individuos. O si, de hecho, lo son, no son de interés de la investigación educativa. ¿Para qué sirve una etnografía de un aula de ciencias que no tiene absolutamente nada que ver con otras aulas de ciencias o que no genere una comprensión contextualizada con algún valor para ellas? ¿Para qué sirve estudiar un profesor de ciencias tan único, que no tiene nada que ver con otros profesores de ciencias?

Eso nos remite al tema de la *generalización:* ¿se deben buscar generalizaciones en estudios cualitativos, a través de, por ejemplo, estudios de "casos representativos"?

Según André (1998, p. 58), la generalización en el sentido de leyes que se aplican universalmente no es un objetivo de la investigación cualitativa. Sin embargo, la idea de generalización es bastante aceptada en ese enfoque en el sentido de que los datos de un estudio pueden ser útiles para comprender datos de otros estudios (ibid.). Por eso, la

descripción densa es considerada vital cuando se pretende hacer comparaciones o transferencias de una situación a otra; el análisis de similitudes y diferencias torna posible juzgar en qué medida las comprensiones construidas en un estudio pueden ser consideradas hipótesis sobre lo que puede o no ocurrir en otras situaciones. Esta posición es compartida por Ogbu et al. (1988, p. 53) al decir que la generalización se torna posible porque el conocimiento construido a través del enfoque cualitativo es profundo y contextualizado; las detalladas descripciones comprensivas permiten a los lectores hacer comparaciones y tomar decisiones bien fundamentadas respecto a la generalización.

Posición semejante es tomada por Taft (1988, p. 74) al argumentar que a fin de generalizar de un caso individual a otros es necesario alcanzar una comprensión suficientemente detallada sobre la significatividad de los eventos respecto al contexto en que ocurren para poder extender interpretaciones a otros contextos y grupos. Cuando un investigador intenta comprender un grupo, él o ella es ayudado por conocer otros grupos; las generalizaciones son hechas a través de la capacidad que tenga el investigador de mediar entre un grupo y otros. Por tanto, la descripción etnográfica de una escuela, por ejemplo, deriva su valor en gran parte del hecho de que el investigador – así como los lectores – tienen familiaridad con otras escuelas, y con escuelas en general (ibid.).

La postura de Erickson (1986, p. 130), como hemos visto en el comienzo de este texto, es la de que en la investigación interpretativa la búsqueda no es de "universales abstractos" alcanzados a través de inferencias estadísticas de muestras para poblaciones, sino de "universales concretos" a los cuales se llega estudiando un caso con mucho detalle y comparándolo con otros casos estudiados con igual detalle.

No obstante, ninguna generalización de esa naturaleza debe ser considerada final, sólo como hipótesis de trabajo para otros estudios interpretativos o como subsidio para encuestas, entrevistas o tests (Taft, 1988, p. 74).

El tema de la generalización tiene que ver con el de la *validez externa* de los estudios, es decir, ¿en qué medida los constructos y universales concretos creados por los investigadores son aplicables a más de un grupo? Por otro lado hay que considerar también la cuestión de la *validez interna*, o sea, ¿los investigadores están interpretando lo que creen interpretar? (Desde el punto de vista cuantitativo, le pregunta sería si los investigadores están midiendo lo que creen medir.)

La validez puede ser pensada como una cualidad de las conclusiones y de los procesos a través de los cuales son alcanzadas, pero su significado exacto depende del criterio de verdad que se está utilizando. Considerando que en el enfoque cualitativo verdad es una cuestión de concordancia en contexto, el mejor significado de validez en ese enfoque parece ser el de *credibilidad* como sugiere Taft (op. cit., p. 73). La credibilidad depende del convencimiento de la comunidad de investigadores y lectores respecto a las evidencias presentadas y a los procesos utilizados. Sturman (1988, p. 65) propone las siguientes estrategias para alcanzar credibilidad:

- los procedimientos de recolección de datos deben ser explicados;
- los datos recogidos deben ser presentados y estar listos para reanálisis;
- instancias negativas deben ser relatadas;
- sesgos deben ser reconocidos;
- análisis de trabajos de campo deben ser documentados;
- la relación entre aserción y evidencia debe ser aclarada;

- evidencias primarias deben ser distinguidas de las secundarias, así como las descripciones de las interpretaciones;
- diarios o logs deben dar cuenta de lo que fue hecho durante las distintas fases del estudio;
- técnicas debe ser diseñadas para "chequear" la calidad de los datos.

Erickson (1986, p. 140), se reporta al mismo tema, desde otra perspectiva, al indicar cinco tipos de inadecuaciones de las evidencias presentadas:

1. *cantidad inadecuada de evidencias;* el investigador tiene poca evidencia para garantizar ciertas aserciones claves;
2. *diversidad inadecuada de tipos de evidencias;* el investigador no tiene distintas fuentes de datos (por ejemplo, observaciones, entrevistas, documentos); no busca triangulación de datos;
3. *interpretaciones incorrectas de las evidencias*; el investigador no comprende bien ciertos aspectos claves de la complejidad de la acción o de los significados atribuidos por los actores en el contexto;
4. *inadecuadas evidencias desconfirmadoras;* el investigador no tiene datos que pudieran desconfirmar una aserción clave; o, más importante, no presenta evidencia de que una búsqueda deliberada de datos potencialmente desconfirmadores fue conducida;
5. *análisis inadecuado de casos discrepantes;* el investigador no escuadriña las instancias desconfirmadoras y las compara con las confirmadoras para determinar qué aspectos de los casos desconfirmadores eran iguales o diferentes de los aspectos análogos de los casos confirmadores.

El mismo autor (op. cit., p. 145) recomienda que el informe de estudios etnográfios contengan nueve elementos principales:

1. *aserciones empíricas* (una tarea básica del análisis de datos es generar dichas afirmaciones en gran medida a través de inducción);
2. *viñetas narrativas analíticas* (una viñeta narrativa es una representación vívida de un evento en lo cotidiano del caso o grupo investigado; la viñeta intenta persuadir al lector de que las cosas en el contexto eran como el autor dice que eran);
3. *citas de las notas de campo* (se pueden citar directamente las notas en el informe, indicando la fecha en que fueron tomadas; una serie de extractos de notas de campo pueden servir como evidencia de que el modo particular en que ocurrió un cierto evento fue típico);
4. *citas de entrevistas* (las palabras de los entrevistados son un medio de transmitir a los lectores los puntos de vista de los sujetos del estudio);
5. *informes sinópticos de los datos* (mapas, tablas de frecuencias, diagramas);
6. *comentarios interpretativos encuadrando una cierta descripción;*
7. *comentarios interpretativos encuadrando la descripción general* (la descripción general tiene como principal objetivo establecer la posibilidad de generalizar los patrones que fueron ilustrados en las descripciones particulares);
8. *discusión teórica* (comentarios interpretativos respecto al significado más amplio de los patrones que han emergido de los datos);
9. *informe sobre la historia natural de la indagación en el estudio* (es decir, una discusión/descripción respecto a cómo ciertos conceptos claves en el análisis evolucionaron o cómo patrones no esperados fueron encontrados durante el trabajo de campo y en la reflexión subsecuente).

Los comentarios interpretativos que encuadran las descripciones particular y general pueden ser de tres tipos: los que preceden y siguen una descripción particular en el texto, la discusión teórica que apunta hacia la significatividad más amplia de los patrones identificados en los eventos mencionados, y una reseña de los cambios que han ocurrido en el punto de vista del investigador durante el transcurso de la indagación (op. cit., p. 152).

Para Erickson (ibid, p. 145) cada uno de esos nueve elementos, separadamente y en conjunto, permiten al lector tres cosas: en primer lugar, le posibilitan experimentar de forma vicaria el escenario descrito y confrontar instancias de aserciones claves y constructos analíticos; en segundo, le permiten examinar todo el espectro de evidencias en el cual está basada la interpretación del investigador; y en tercero, dejan que el lector considere los fundamentos teóricos y personales de la perspectiva del autor tal como ha cambiado a lo largo del estudio.

Esta sección de este texto fue dedicada al tema de la fiabilidad, generalización y validez de los estudios cualitativos. Aunque se pudiera argumentar que son conceptos típicos de estudios cuantitativos, toda la sección fue desarrollada con el objetivo de mostrar que dichos conceptos tienen sentido en el contexto de una investigación interpretativa una vez que se le dé el significado apropiado. Tal significado, como hemos visto, parece ser el de credibilidad, el cual nos remite a otro concepto importante en la metodología de la investigación interpretativa: la triangulación, una estrategia central para alcanzar credibilidad (Sturman, 1988, p. 65).

## *Triangulación*

La triangulación puede involucrar el uso de distintas fuentes de datos, diferentes perspectivas o teorías, diferentes investigadores o diferentes métodos; es una respuesta holística a la cuestión de la fiabilidad y validez de los estudios interpretativos (ibid.). Para Denzin (1988, p. 318), la triangulación es el empleo y combinación de varias metodologías de investigación en el estudio de un mismo fenómeno. No es una estrategia típica de la investigación cualitativa, tampoco es una estrategia nueva: el uso de múltiples mediciones y métodos de modo que se supere las debilidades inherentes al uso de un único método o un único instrumento tiene una larga historia en las ciencias naturales y sociales; en la investigación cuantitativa, la triangulación, es usada, tradicionalmente, como una estrategia de validación de observaciones (ibid.).

No obstante, según Denzin (op. cit.) la apropiación de ese concepto por los investigadores interpretativos y su aplicación a problemas típicos de la investigación cualitativa es más reciente y representa un compromiso con un sofisticado rigor metodológico por parte de los investigadores, en el sentido de que están comprometidos a tornar sus esquemas empíricos e interpretativos lo más público posible. De acuerdo con ese autor (p. 319), hay cinco tipos básicos de triangulación:

1. *triangulación de datos*, involucrando tiempo, espacio y personas;
2. *triangulación de investigadores,* la cual consiste en el uso de múltiples observadores en vez de uno sólo;
3. *triangulación de teorías,* que consiste en utilizar más de un esquema teórico en la interpretación del fenómeno investigado;
4. *triangulación metodológica* que involucra el uso de más de un método y puede consistir en estrategias intra métodos o entre métodos;

5. *triangulación de verificación por sujetos* en la cual los investigados examinan y confirman o desconfirman lo que ha sido escrito sobre ellos.

Existe también la triangulación múltiple en la cual el investigador combina en una investigación múltiplos observadores, perspectivas teóricas, fuentes de datos y metodologías.

Por otro lado, aunque Denzin asocie la triangulación con un compromiso con el rigor metodológico, el uso de esa estrategia en la investigación cualitativa no está libre de críticas: los argumentos son, por ejemplo, que la triangulación de datos tiene un viés positivista, que dos investigadores nunca observan el mismo fenómeno de la misma manera, que distintos métodos generan distintas imágenes y recortes de la realidad y que la triangulación de teorías no tiene sentido epistemológicamente.

Sin embargo, el mismo Denzin (op. cit, p. 321) contra-argumenta diciendo que la triangulación en los estudios cualitativos no debe ser comparada con el análisis de correlación en las investigaciones cuantitativas y que nunca debe ser una estrategia ecléctica. En la triangulación de investigadores no se espera que observen exactamente de la misma manera y que uno corrobore lo que el otro observa, sino que sus distintas observaciones expandan la base interpretativa del estudio y que revelen aspectos del fenómeno investigado que no serían necesariamente observados por un único investigador. Respecto a la triangulación metodológica, lo importante es justamente que distintas imágenes puedan emerger. En relación a la triangulación de teorías, Denzin (ibid.) dice que ella, en vez de requerir que las interpretaciones fueran consistentes con dos o más teóricos, simplemente requiere que el investigador sea consciente de las distintas maneras a través de las cuales el fenómeno puede ser interpretado.

El tema de la triangulación, en particular las triangulaciones teóricas y metodológicas, está muy vinculado a la cuestión de los paradigmas, enfocada desde una perspectiva dicotómica en el comienzo de este texto. En esa oportunidad hemos dicho que la mirada dicotómica estaba siendo usada por razones didácticas y que al final retomaríamos el asunto de los paradigmas bajo una visión integradora. Es llegado, entonces el momento de volver a los paradigmas.

## Los paradigmas de investigación en educación: hacia la acomodación

La dicotomía establecida anteriormente entre los "dos paradigmas clásicos" es simplificadora. La metodología de la investigación en las ciencias sociales no puede ser pensada simplemente como "no-positivista" en contraposición a una supuesta tradición positivista de la investigación en las ciencias naturales. Si así fuera, no tendría sentido hablar de acomodación de paradigmas. Sin embargo, considerando paradigma como "un conjunto básico de creencias que orienta la acción" (Guba, 1990, apud Alves-Mazzotti, 1996, p. 17), es decir, una concepción de mundo que guía al investigador, no sólo en elecciones de método, sino también en sus posiciones ontológicas y epistemológicas, se pueden distinguir por lo menos tres paradigmas como sucesores del positivismo: el *postpositivismo*, la *teoría crítica* y el *naturalismo/constructivismo* (ibid.). El primero sería una versión modificada del positivismo revisando puntos insostenibles (e.g., la realidad se supone ahora que existe pero que nunca será totalmente aprehendida por la investigación); la *teoría crítica* es ideológicamente orientada una vez que rechaza la neutralidad: el proceso de investigación es mediado por el investigador, y el término crítica se refiere tanto a la crítica interna que resulta del cuestionamento analítico de la argumentación y del método como al análisis de las condiciones de regulación social, desigualdad y poder; al *naturalismo/constructivismo* subjace

a la idea de que los resultados de cualquier investigación son siempre influenciados por la interacción investigador/investigado, de suerte que el conocimiento es siempre producto de la actividad humana y, por lo tanto, nunca puede ser visto como algo definitivo, sino como algo que está siempre modificándose (Alves-Mazzotti, 1996, pp. 17-20).

El paradigma naturalista/constructivista es el que hemos enfatizado en este texto: la realidad es socialmente construida (lo que implica que hay siempre múltiples realidades; las ideas, los valores y la interacción investigador/investigado influye en la configuración de los "hechos" (lo que implica que la teoría es subdeterminada). Estas características, a su vez, implican un relativismo que es problemático para los otros dos paradigmas: *si alguién se propone comprender los significados atribuidos por los actores a las situaciones y eventos de los cuales participan, si intenta entender la "cultura" de un grupo u organización en el cual coexisten diferentes visiones correspondientes a los subgrupos que los componen, entonces el relativismo no contituye un problema; pero si nos proponemos la construcción de teorías (post-positivismo) o la transformación social (teoría crítica), lo que exige acuerdo alrededor de decisiones o principios que posibiliten la acción conjunta, el relativismo pasa a ser un problema* (op. cit., p. 21).

Con la identificación de estos tres paradigmas "pos-clásicos", y posiblemente otros, ya se percibe que la cuestión de los paradigmas en ciencias sociales no es dicotómica, que las ciencias sociales son multiparadigmáticas y que la acomodación o integración de paradigmas no debe ser considerada imposible o, por lo menos, es una cuestión en abierto, como dijo Alves-Mazzotti (op.cit., p. 22).

Un buen ejemplo de argumento en favor de la integración paradigmática es dado por Bericat (1998). Este autor considera que existen tres razones fundamentales que pueden motivar el diseño multimétodo en una investigación social: *complementación, combinación* y *triangulación* (p. 37).

La *complementación* existe cuando, *en el marco de un mismo estudio, se obtienen dos imágenes, una procedente de métodos de orientación cualitativa y otra de métodos de orientación cuantitativa* (ibid.), resultando así un doble y diferenciado conjunto de aserciones de conocimiento sobre el fenómeno de interés. Lo que se obtiene son dos perspectivas diferentes sin pretensión alguna de solapamiento, o convergencia. Las aserciones de conocimiento son presentadas con dos partes bien diferenciadas, cada una de las cuales expone resultados alcanzados por la aplicación del respectivo método. Según Bericat (ibid.) *en la complementación el grado de integración metodológica es mínimo, y su legitimidad se soporta sobre la creencia de que cada orientación es capaz de revelar diferentes zonas de la realidad social, así como que es necesario contar con esa doble visión para un mejor entendimiento del fenómeno* (ibid.).

En la *combinación* la estrategia es la de integrar subsidiariamente una metodología, sea la cualitativa o la cuantitativa, en la otra con el objeto de fortalecer la validez de ésta compensando sus debilidades mediante la incorporación de informaciones que proceden de la aplicación de la otra metodología. Lo que se busca no es la convergencia de resultados, que finalmente procederán de una sola metodología, sino más bien una adecuada combinación metodológica (op. cit., p. 39).

Finalmente, en la *triangulación* lo que se pretende es un solapamiento o convergencia de resultados. No se trata de complementar la visión de realidad con dos miradas, sino de utilizar dos metodologías para el estudio de un mismo e idéntico aspecto de una realidad

social. Las metodologías, tal como en la complementación, son implementadas de forma independiente pero se enfocan hacia un mismo objeto de estudio buscando resultados convergentes. *La legitimidad de esta estrategia depende de si creemos que ambas metodologías realmente pueden captar idéntico aspecto de la realidad, esto es, si el solapamiento es posible. En la medida en que pensemos que conducen a visiones inconmesurables de la realidad, entonces estaríamos en el caso de la complementación* (op. cit., p. 38).

En la figura n° 5 estas tres estrategias de integración están esquematizadas en un mapa conceptual. Se trata siempre de captar aspectos de una realidad social. Sin embargo, eso se puede intentar con sólo una metodología o integrándolas de modo complementario, triangular o combinatorio. La triangulación y combinación sólo son posibles en la medida en que se acepta por lo menos un cierto grado de conmensurabilidad paradigmática. En caso contrario, la única alternativa es la complementación.

## Conclusión

En la figura n° 6 se presenta, a modo de conclusión de este texto un diagrama V. Este tipo de diagrama, también conocido por Ve epistemológico, fue diseñado por D.B. Gowin (1981) para esquematizar la estructura del proceso de producción de conocimiento. Aquí lo usamos para reflejar dicha producción en el marco del paradigma cualitativo, particularmente en educación. Este diagrama pretende ser una especie de resumen de todo el texto. Como tal, es preciso tener en cuenta que ninguno de los apartados que aparecen en el diagrama V está completo. Son dados sólo ejemplos de itens que podrían integrar cada apartado. Es ésta la razón de los tres puntos que aparecen al final de ellos.

## Referencias

Bericat, E. (1988). *La integración de los métodos cuantitativo y cualitativo en la investigación social.* Barcelona, Editorial Ariel.

Eisner, E.W. (1981). On the differences between scientific and artistic approaches to qualitative research. *Educational Researcher, 10*(4): 5-9.

Firestone, W.A. (1987). Meaning in method: the rethoric of quantitative and qualitative research. *Educational Researcher, 16*(7): 16-21.

Husén, T. (1988). Research paradigms in education. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 16-21.

Landsheere, G. de (1988). History of educational research. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 8-16.

Marshal, C. (1984). The wrong time for mechanistics in qualitative research. *Educational Researcher, 13*(9): 26-28.

Miles, M.B. and Huberman, A.M. (1984) Drawing valid meaning from qualitative data: toward a shared craft. *Educational Researcher, 13*(5): 20-30.

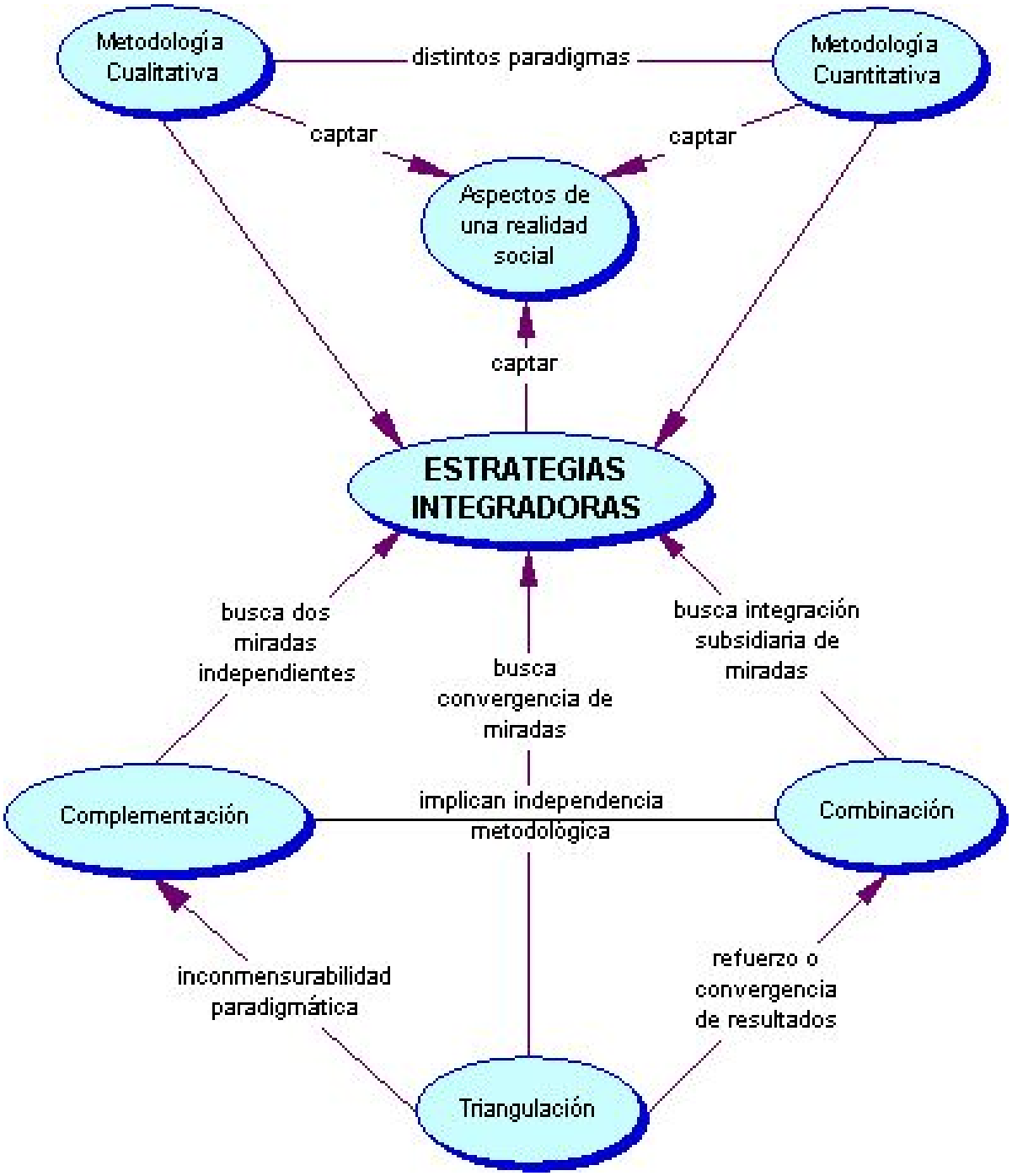



Figura n° 5: Un mapa conceptual para estrategias integradoras entre las metodologías cualitativa y cuantitativa (Moreira, 2002)


Figura n° 6. Un diagrama V para la investigación cualitativa, particularmente en educación (M.A. Moreira, 2002)

Shulman, L.S. (1981). Disciplines of inquiry in education: an overview. *Educational Researcher, 10*(6): 5-12.

Smith, J.K. (1983). Quantitative versus qualitative research: na attempt to clarify the issue. *Educational Researcher, 12*(3): 6-13.

Smith, J.K. and Heshusius, L. (1986). Closing down the conversation: the end of the quantitative-qualitative debate among educational inquirers. *Educational Researcher, 15*(1): 4-13.

Keeves, J.P. (1988). Towards a unified view of educational research. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 1-7.

Moreira, M.A. (2000). Investigación en enseñanza: aspectos metodológicos. In *Actas de la I Escuela de Verano sobre Investigación en Enseñanza de las Ciencias.* Burgos, Servicio de Publicaciones de la Universidad de Burgos. pp. 13-51.

Erickson, F. (1986). Qualitative methods in research on teaching. In Wittrock, M.C. (Ed.). *Handbook of research on teaching.* New York: Macmillan Publishing Co. p. 119-161. Traducción al español: Erickson, F. (1989) Métodos cualitativos de investigación sobre la enseñanza. In Wittrock, M.C. (Comp.). La investigación en la enseñanza, II. Barcelona, Paidós. pp. 195-301.

André, M.E.D.A. (1998). *Etnografia da prática escolar.* 2ª ed. São Paulo, Papirus Editora.

Ogbu, J.U., Sato, N.E. and Kim, E.Y. (1988). Anthropological inquiry. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 48-54.

Taft, R. (1988). Ethnographic research methods. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 71-75.

Sturman, A. (1988) Case study methods. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 61-66.

Serrano, G.P. (1998). Investigación cualitativa. Retos e interrogantes. I. Métodos. Madrid, La Muralla S.A.

Stenhouse, L. (1985). Case study methods. In Husén, T. & Postlethwaite, T.N. (Eds.). *International Encyclopedia of Education.* Oxford, Pergamon Press.

Eliott, J. (1993). *El cambio educativo desde la investigación-acción.* Madrid, Ediciones Morata.

Kemmis, S. (1988). Action research. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 173-179.

Kemmis, S. and McTaggart, R. (Eds.). (1988). *The action research reader.* 3rd ed. Geelong, Deakin University Press.

Goetz, J.P. y Lecompte, M.D. (1988). *Etnografía y diseño cualitativo en investigación educativa.* Madrid, Ediciones Morata.

Denzin, N.K. (1988). Triangulation in educational research. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 318-322.

Marton, F. (1988). Phenomenography. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 95-101.

Ödman, P.J. & Kederman, D. (1988). Hermeneutics. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 185-192.

Hall, B.L. (1988). Participatory research. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 198-204.

Ball, S.J. (1988). Participant observation. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 310-314.

Walker, J.C. & Evers, C.W. (1988). Research in education: epistemological issues. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press. pp. 22-31.

Alves-Mazzotti, A.J. (1996). O debate atual sobre os paradigmas de pesquisa em educação. *Cadernos de Pesquisa,* São Paulo, n. 96: 15-23.

Guba, E.G. (1990). The alternative paradigm dialog. In: Guba, E.G. (Ed.) *The paradigm dialog.* London, Sage. Apud: Alves-Mazzotti (1996). O debate atual sobre os paradigmas de pesquisa em educação. *Cadernos de Pesquisa,* São Paulo, n. 96: 15-23.

Hyams, B.K. (1988). *Oral history*. In Keeves, J.P. (Ed). *Educational research, methodology, and measurement. An international handbook.* Oxford, Pergamon Press.

# Capítulo 2
# Investigación en Enseñanza: Métodos Cuantitativos[1]

M.A. Moreira
P.R.S. Rosa

**Resumen**

La finalidad de este texto es rescatar la metodología cuantitativa aplicada a la investigación en la enseñanza. No hay en él nada nuevo. Todos los asuntos abordados se encuentran mucho más elaborados en libros de metodología de la investigación o de Estadística. Sin embargo, llamar la atención para cuestiones como tratamiento de variables, diseño de investigación, muestreo, significancia estadística, fidedignidad y validez, análisis de varianza, tests no paramétricos, puede ser una contribución para la actual investigación en la enseñanza que, hace mucho tiempo, prácticamente sólo usa métodos cualitativos y, así, acaba sesgada.

## Introducción

D. B. Gowin (1981, 2005) propuso un dispositivo heurístico – que hoy es conocido como diagrama V (Moreira, 2006) – para desvelar la estructura del proceso de producción del conocimiento. La Figura 1 presenta ese dispositivo aplicado a la investigación cuantitativa en la educación. El centro de ese diagrama es el *dominio de interés* de la investigación (por ejemplo, la enseñanza y el aprendizaje) y las preguntas e hipótesis que se hacen dentro de ese dominio. Para responderlas, confirmarlas o refutarlas es necesario hacer registros de un evento que se provoca o que acontece naturalmente. El lado izquierdo del diagrama es el *dominio conceptual* (o teórico) de la investigación; en él están los conceptos, principios, modelos, teorías y filosofías que fundamentan teórica y epistemológicamente la investigación y que interactúan con los registros, transformaciones y afirmaciones que constituyen el *dominio metodológico* que aparece en el lado derecho del diagrama.

Este texto se centrará en el dominio metodológico. Como se ve en la Figura 1, el lado metodológico empieza con los registros. Sin registros no se hace investigación empírica. A partir de ahí, un paso fundamental de ese tipo de investigación es la conversión de esos registros en índices numéricos.

Por ejemplo, en el caso de que los registros sean mapas conceptuales, es necesario definir criterios, como por ejemplo, tantos puntos para la jerarquía, tantos para conectivos, etc., para llegar a un resultado para cada mapa. O definir categorías como muy bueno, bueno, regular, equivocado, y atribuir puntos a cada categoría.

Es cierto que también se puede trabajar cuantitativamente con índices no numéricos como, por ejemplo, variables dicotómicas como sí o no, femenino o masculino, pero en la investigación empírica predomina el uso de índices numéricos.

---

[1] *Texto de Apoyo n° 29*, Programa Internacional de Doctorado en Enseñanza de las Ciencias, Universidad de Burgos, España; Universidad Federal de Río Grande del Sur, Brasil. Publicado en *Actas del PIDEC,* 2007, Vol. 9: 03-56.

Es igualmente correcto que, en cualquier tipo de investigación, lo más importante es la pregunta de la investigación. La búsqueda de respuesta para tal pregunta es lo que genera conocimiento. El conocimiento humano es construido, reconstruido, refutado, modificado, por la búsqueda, muchas veces obstinada, de respuestas a preguntas sobre determinados fenómenos de interés.

La identificación de una cuestión de investigación que valga la pena investigar, que pueda generar conocimientos, es la parte más difícil de la investigación. Sin embargo, es también muy importante un diseño de investigación que permita hacer registros relevantes que, a su vez, originen datos (típicamente índices numéricos, en la investigación cuantitativa) de máxima relevancia para la pregunta de la investigación.

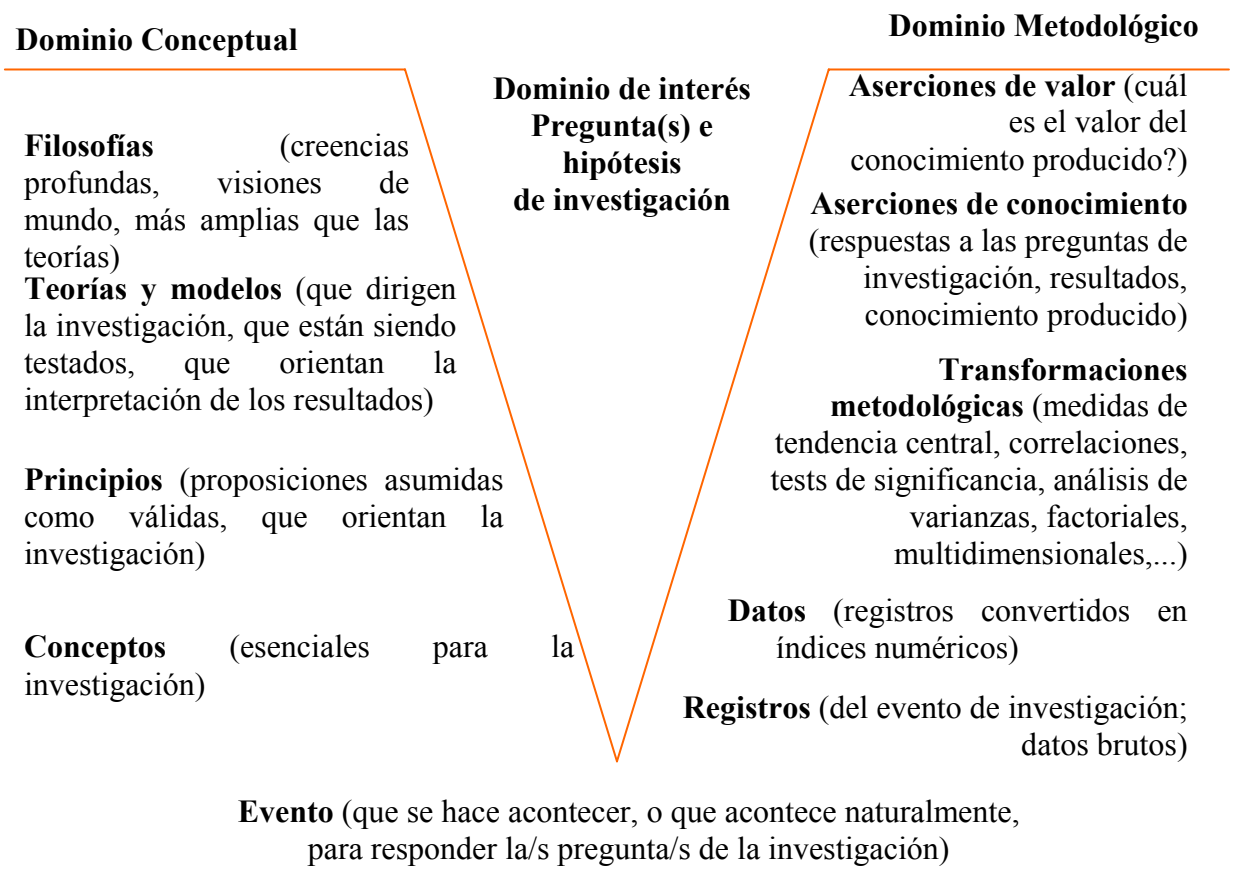



Figura 1 - El diagrama V aplicado a la investigación cuantitativa en educación.

## *Datos y tratamiento estadístico*

Típicamente, en la investigación educacional empírica, el investigador se pregunta cuál es la evidencia que lo apoyará con relación a ciertas hipótesis de la investigación. Entonces, hace registros, los convierte en números y los trata estadísticamente para ver si sirven como evidencia.

Se suele decir que buenos datos hablan por sí mismos. Eso significa que si los datos son de buena calidad, los procedimientos estadísticos son inmediatos. Por otro lado, si los datos son malos no hay tratamiento estadístico que los transforme en buenos.

Lo importante es la calidad de los datos no las manipulaciones estadísticas. La relevancia de las conclusiones estadísticas nunca será mayor que la adecuación de los datos numéricos trabajados estadísticamente. La interpretación de los resultados estadísticos depende de lo que está por detrás de los datos. Los números a ser analizados no son entidades sagradas representando puras abstracciones. Al contrario, su utilidad en la investigación empírica reside en el hecho de que tienen referentes en el contexto de la investigación, que significan algo en el mundo real (Millman, 1970).

Una buena estrategia para abordar mejor la cuestión del análisis y calidad de los datos es considerarla antes, o sea, pensar en el análisis de los datos antes de recogerlos evitando, así, que los números obtenidos no tengan mucho que ver con las cuestiones investigadas.

Otro aspecto a considerar en esa cuestión es el de la fidedignidad y validez de los instrumentos. Sin instrumentos fidedignos y válidos, los números que de ellos resulten no serán confiables y de nada servirá tratarlos estadísticamente.

En lo que se refiere a instrumentos, es conveniente testarlos previamente, preguntándole a algunos sujetos cómo interpretan determinados ítems, o hacer un análisis del tipo "¿tiene sentido atribuir el mismo número a una no respuesta y a una respuesta neutra?", es decir, "¿una pregunta en blanco puede ser considerado como respuesta neutra?", "¿sin opinión es lo mismo que opinión neutra?" (op.cit.).

## *Significancia estadística y significancia práctica*

En el análisis de los datos es importante tener en cuenta que el nivel de significancia estadística ($p < 0,5$, por ejemplo) no es una medida de la importancia o de la significancia práctica de un resultado, pues ese nivel depende del número de casos y de la eficiencia del diseño de investigación.

Cuando un resultado es estadísticamente significativo, el investigador debe analizar la magnitud de los efectos, o sea, ¿qué importancia tiene, en el contexto de la investigación, una diferencia de medias, estadísticamente significativa, por ejemplo, entre 7,5 y 7,2? o ¿cuál es la relevancia de una correlación, estadísticamente significativa al nivel 0,05, de un coeficiente de correlación de 0,23 entre dos variables?

Otros aspectos a considerar en la cuestión de la significancia estadística y significancia práctica son el tamaño y la variabilidad de la muestra. Aun cuando se obtengan diferencias, correlaciones o factores estadísticamente significativos para muestras de 12 ó 13 sujetos, hay que preguntarse cuál es la significancia práctica de esos resultados. Por otro lado, muestras pequeñas e, incluso, muestras mayores de 30 ó 40 sujetos, por ejemplo, pueden presentar gran variabilidad. En una muestra de 10 sujetos, si dos tienen el máximo de diez puntos, dos obtienen cero puntos y los demás alcanzan cinco o seis puntos, la media será cinco o más puntos, pero ¿qué significa esa media en la práctica? Es claro que, en ese caso, es mejor trabajar con la moda, pero así mismo es necesario preguntarse cuál es la significancia práctica de ese resultado. Análogamente, hay que hacer ese cuestionamiento cuando es grande la variabilidad de los resultados en muestras más grandes.

## La estadística apropiada

En la elección de la técnica estadística adecuada para el tratamiento de los índices numéricos es conveniente considerar en primer lugar si lo que se quiere es *describir* características de un conjunto de números o si lo que se pretende es *estimar* valores de la población. En el primer caso, se debe usar la estadística descriptiva, en el segundo, la inferencial.

También es conveniente volver a la pregunta de la investigación y, teniendo en cuenta su naturaleza, considerar si para responderla, de hecho, son necesarios análisis factoriales o multidimensionales, medidas de tendencia central, correlaciones, tests de significancia para diferencias de medias, análisis de varianza.

## El papel del ordenador

Hay algoritmos de ordenador para realizar cualesquier análisis estadísticos de datos numéricos. Basta inyectar los datos en esos algoritmos para que rápidamente salgan medias, desvíos estándar, coeficientes de correlación, varianzas, factores, etc. El ordenador atiende comandos y sus procedimientos actúan sobre los datos que le son suministrados. Y volvemos al comienzo del asunto de análisis de datos: si los datos son malos, no hay ordenador que dé buenos resultados. De nada vale darle a los comandos para que dé más y más coeficientes, tablas, factores. ¡Todo basura!

Por otro lado, si los datos son buenos es necesario saber interpretar los productos de los procedimientos estadísticos a los cuales son sometidos. No es necesario usar lápiz y papel, o calculadora, para, tediosamente, calcular estadísticas (valores de la muestra), parámetros (valores de la población), coeficientes, correlaciones, factores. El ordenador lo hace mucho más rápidamente y sin errores. Pero es fundamental saber interpretar los resultados (op.cit.).

El análisis de los datos es, como se dijo al principio, una etapa fundamental de la investigación empírica. Sin duda. Pero más importante es la pregunta de la investigación y los registros que se hace de los eventos usados para responder a esa cuestión. Son esos registros que serán convertidos en índices numéricos que, a su vez, serán analizados estadísticamente. Los procedimientos estadísticos están disponibles en profusión y el ordenador los ejecuta rápidamente. Pero el análisis, en sí, tiene que ser realizado por el investigador.

Pasemos ahora a describir y tratar de aspectos y técnicas de los métodos cuantitativos de investigación en educación.

## Variables

Las condiciones que pueden ser variadas o seleccionadas por el investigador son llamadas **variables independientes**. Las medidas de las respuestas realizadas durante el experimento constituyen las **variables dependientes**.

Los **niveles** de una variable se refieren a categorías de la variable. Sexo, por ejemplo, tiene dos niveles. El número de niveles de una variable como edad puede ser arbitrariamente determinado y puede variar de dos (por ejemplo, por encima de *35* o abajo o debajo o igual a *35* años) al infinito (cuando edad es tratada como una variable continua).

La palabra *factor* es frecuentemente usada como sinónimo de la expresión variable independiente. Una **variable independiente manipulada** es una condición que está bajo el control directo del experimentador. **Variables de tratamiento** son, normalmente, variables independientes manipuladas por el investigador, cuyos efectos está queriendo observar. Una **variable independiente normativa** es aquélla en la cual el investigador no está libre para producir la condición en sí misma, aunque sea libre para decidir qué niveles de la variable serán incluidos en la investigación[2]. Sexo y edad son ejemplos de variables de este tipo.

Medidas iniciales, o sea, obtenidas antes del inicio de la investigación, que son usadas para formar grupos homogéneos (con relación a la variable dependiente) son llamadas **variables de agrupación**. Observaciones suplementarias a las observaciones antes del tratamiento, con relación a posibles diferencias, son llamadas **observaciones concomitantes** o **covariables**. Una observación concomitante puede ser usada como una alternativa a la agrupación en delimitaciones experimentales[3] o, en algunas delimitaciones casi-experimentales, puede ser usada como un esfuerzo para superar las deficiencias del diseño.

Otro tipo de caracterización de variables concierne a la naturaleza del proceso de medición de la variable. Podemos definir cuatro grupos de variables, con respecto a la escala usada para medir las variables:

1. **Variables nominales** - estamos manejando ese tipo de variable cuando apuntamos sólo si la grandeza medida por la variable está presente o no. Así, por ejemplo, una persona puede ser profesor o no. Normalmente, se atribuye valor 1 si el atributo medido por la variable está presente y se atribuye el valor 0 si el atributo no está presente. O sea, en este caso, profesor (1) - no profesor (0).
2. **Variables ordinales** - son aquéllas donde los datos presentan algún tipo de ordenación. A cada sujeto se le atribuye un grado, conforme alguna escala, a medida que presenta el atributo que se está midiendo. Por ejemplo, una test clasificatorio de suficiencia en matemática, donde los sujetos son ordenados de acuerdo con las notas obtenidas en el test.
3. **Variables intervalares** - este tipo de variable se caracteriza por valores que están distribuidos en una escala con una diferencia constante entre dos valores consecutivos. Edad puede ser un ejemplo de este tipo de variable, si contamos solamente los años efectivamente cumplidos.
4. **Variables racionales** - son aquéllas donde la escala de medida está compuesta por números racionales y, además, existe un cero que define la ausencia de la propiedad medida por la variable. Por ejemplo, notas en un test son variables de ese tipo si atribuimos valores fraccionarios a las respuestas[4].

## Control de variables

Variables que no son de interés directo del investigador pueden ser eliminadas o tienen su influencia minimizada por varios métodos:

---

[2] Este tipo de variable también se llama parámetro.
[3] Aquellos en los cuales la muestra es aleatoria; p. 37.
[4] Si, por otro lado, solamente consideramos como cierto (*1* punto) o errado (*0* puntos) entonces tendremos una escala intervalar.

1. **Remoción de variables** - algunas variables pueden ser eliminadas seleccionando casos con características uniformes (por ejemplo, usando sólo mujeres para eliminar la influencia de la variable sexo).
2. **Emparejamiento de casos** - seleccionando parejas o conjuntos de individuos con características idénticas (o casi idénticas) y distribuyéndolos en el grupo experimental y en el de control. Sin embargo, el emparejamiento no es considerado satisfactorio a menos que los sujetos de las parejas o conjuntos sean distribuidos aleatoriamente en el grupo experimental o en el de control. Una limitación de ese método es la dificultad para emparejar sujetos usando dos o más variables.
3. **Balance de casos** - se distribuyen los sujetos en el grupo experimental y en el de control de tal modo que las medias y varianzas de los grupos sean semejantes, lo más posible. Este método también presenta una dificultad similar a la observada en el emparejamiento de casos: la dificultad de formar grupos con base en más de una variable.
4. **Análisis de covarianza** - este método permite al experimentador eliminar diferencias iniciales en varias variables entre los grupos experimental y de control por métodos estadísticos. Usando resultados de pre-tests como co-variables, este método es considerado preferible al convencional emparejamiento.
5. **Aleatoriedad** - la aleatoriedad puede ser obtenida a través de la selección de los sujetos al azar, entre la población que se quiere estudiar, que van formar parte de los grupos de control y experimental. La aleatoriedad nos da un método efectivo de eliminar errores sistemáticos y de minimizar el efecto de variables externas. El principio de la aleatoriedad se basa en la hipótesis de que a través de la selección aleatoria cualquier diferencia entre los grupos es simplemente debido al muestreo o al azar. Esas diferencias son conocidas como **errores de muestreo** o **errores de varianza** y su intensidad puede ser estimada por el investigador. En un experimento, diferencias en la variable dependiente que pueden ser atribuidas al efecto de la variable independiente son conocidas como **varianza experimental**. La significancia de un experimento puede ser testada comparando la varianza experimental con el error de varianza. Si al término del experimento las diferencias entre los grupos experimental y de control son muy grandes para ser atribuidas al error de varianza, se puede presumir que estas diferencias son atribuibles a la varianza experimental. La aleatoriedad es el método más efectivo de formar grupos y controlar variables externas y debe ser usado siempre que las circunstancias lo permitan (Best, 1970).

## *Diseños de investigación*

Se entiende por **diseño** de una investigación el conjunto compuesto por el plan de trabajo del investigador, la manera como éste selecciona sus muestras y analiza sus datos. Se puede decir que de nada valen la observación cuidadosa y el exhaustivo y detallado análisis estadístico si esto es realizado para un plan de investigación inadecuado para la situación en estudio. Conviene recordar lo que ya se desatacó en este texto: **no** es una buena estadística lo que hace que una investigación sea buena.

El asunto del diseño experimental se encuentra muy bien desarrollado y presentado en la obra de Campbell y Stanley (1963, 1991), libro que, además, es lectura obligatoria para el investigador interesado en métodos cuantitativos. Aquí sólo se hará un resumen de la clasificación de Campbell y Stanley.

Se seguirá aquí a Campbell y Stanley en lo que se refiere a la designación de observaciones y tratamientos. Designaremos por la letra O una observación. Un subíndice en la letra O indica una observación particular de una serie, *no necesariamente en orden cronológico*. El índice funciona solamente como un rótulo para una determinada observación. Designaremos por la letra *X* un tratamiento. Por ejemplo, la siguiente secuencia:

$$O_1 \;\; X \;\; O_2$$

indica que fue realizada una observación ($O_1$), enseguida fue aplicado un tratamiento *X* y después fue realizada una segunda observación ($O_2$). Cuando la letra *A* esté delante de una secuencia de observaciones y tratamientos, esto significa que las muestras fueron seleccionadas aleatoriamente. Cada secuencia que se encuentra en una línea concierne a un mismo grupo de sujetos (la muestra). Así la secuencia:

$$\begin{array}{llll} A & O_1 & X & O_2 \\ A & O_3 & & O_4 \end{array}$$

se lee: se observa un grupo de sujetos una vez ($O_1$), se somete entonces el grupo de sujetos a un determinado tratamiento *X* y se observa el mismo grupo de sujetos una segunda vez ($O_2$). Este grupo de sujetos es llamado *grupo experimental*. La segunda línea significa que se observa un segundo grupo una vez ($O_3$), no se aplica el tratamiento[5] y entonces se observa el grupo de sujetos una segunda vez ($O_4$). A ese segundo grupo se le da el nombre de *grupo de control*. La letra *A* indica que los dos grupos, de control y experimental, fueron escogidos aleatoriamente. Otra convención adoptada es que cuando dos letras se encuentran en la misma columna significa que los eventos acontecieron simultáneamente en el tiempo. Así, en el ejemplo, las observaciones $O_1$ y $O_3$ tuvieron lugar en el *mismo instante de tiempo*[6].

Siguiendo la clasificación de Campbell y Stanley los diseños de investigación pueden ser divididos según tres clases:

- **Diseños no experimentales.**
- **Diseños experimentales.**
- **Diseños cuasi-experimentales.**

Por *experimental*, que es el adjetivo común a todas las clases citadas, entendemos condiciones *controladas* de investigación. O sea, el investigador debe ser capaz de controlar o de tener en cuenta de forma apropiada todas las variables pertinentes a un determinado estudio. Los factores de validez de cada experimento pueden ser clasificados como factores de validez interna, los cuales se refieren a las variables que, si no se controlan, tornan sin significancia cualquier afirmación de conocimiento sobre los resultados del experimento, o factores de validez externa, que si no se controlan, invalidan la generalización de los resultados del experimento para una determinada población. Un determinado experimento sólo es verdaderamente experimental si controla la totalidad de las variables que influyen en los factores de validez, interna y/o externa. Obviamente, cuando se trata de investigación en Ciencias Sociales, no siempre es posible controlar apropiadamente todas las variables

[5] De hecho esa denominación es arbitraria, ya que un no-tratamiento también es un tratamiento.

[6] Claro que esta afirmación debe ser entendida como aproximadamente en el mismo instante de tiempo, pudiendo haber un cierto intervalo de tiempo entre las observaciones.

involucradas en determinada situación. Por tanto, el estudio estará más próximo de un experimento verdadero cuanto más control el experimentador tenga sobre estas variables. La línea básica de raciocinio para clasificar determinado diseño en una o en otra categoría es cuánto control ofrece el diseño de los factores de validez interna y externa.

Diseños que ofrecen poco o ningún control de las variables pertinentes son llamados *Diseños no experimentales* o *pre-experimentales*. Por otro lado, diseños que ofrecen alto grado de control son llamados *Diseños experimentales*. Finalmente, diseños que ofrecen grado de control en nivel medio, pero no ofrecen control en los niveles de la categoría anterior, son llamados *Diseños cuasi-experimentales*.

En la óptica cuantitativa, el investigador **siempre** debe procurar un diseño experimental para su trabajo. Si es imposible un tratamiento de este tipo, es aceptable un diseño cuasi-experimental. Un tratamiento no experimental **nunca es aceptable**.

## *Diseños no experimentales o pre-experimentales*

Dentro de la clasificación de Campbell y Stanley, los diseños no experimentales o pre-experimentales son de tres tipos:

**Diseño de tipo 1** - En este diseño se observa sólo un grupo bajo la acción del tratamiento *X*. Ese diseño es esquematizado como:

$$X \quad O_1$$

o sea, en ese diseño, sólo se observa el grupo que experimentó el tratamiento *X*, y solamente una vez.

Las desventajas de ese diseño son evidentes. Tal vez la mayor de ellas sea el hecho de no tener ningún control sobre las variables externas que actúan concomitantemente con *X*. Variables tales como *historia, maduración, interacción del experimentador con el tratamiento*, etc. no son controladas. No hay razón plausible para la utilización de ese diseño, debiendo evitarlo a todo costo. Un ejemplo de ese tipo de diseño es aquella situación en que el profesor aplica un nuevo método de enseñanza y un examen. Ninguna consecuencia que se saque del resultado del examen es válida debido a las deficiencias de ese diseño.

**Diseño de tipo 2** - Un diseño muy usado en la investigación en enseñanza, pero que en verdad es un diseño pre-experimental, es el siguiente:

$$O_1 \quad X \quad O_2$$

En ese diseño, se aplica un pre-test $O_1$ a un grupo, se somete el grupo a un tratamiento *X* y se aplica, entonces, un post-test $O_2$. $O_1$ y $O_2$ significan que el mismo grupo es observado antes y después del tratamiento que puede ser, por ejemplo, un nuevo método de enseñanza o un recurso didáctico alternativo. Diferencias entre $O_1$ y $O_2$ (que pueden ser simples tests de conocimiento) evidenciarían la eficacia o ineficacia del tratamiento *X*. El problema de ese diseño es que no controla otras variables, además de *X*, que podrían explicar las diferencias entre $O_1$ y $O_2$. Por ejemplo, los alumnos podrían tener mejores resultados en el post-test porque haya ocurrido algún evento entre la aplicación del pre-test y del post-test (variable historia) y no porque el tratamiento *X* haya sido eficiente.

**Diseño de tipo 3** - se debe tener cuidado con no confundir este diseño con uno de los diseños experimentales que van a ser descritos más adelante. En ese diseño, tenemos dos grupos, experimental y de control, pero la selección de los sujetos que pertenecen a los dos grupos **no es aleatoria**. De esa forma, ese diseño no controla la variable *selección*. Ese diseño tiene la forma:

$$\begin{array}{ll} X & O_1 \\ & O_2 \end{array}$$

donde la ausencia de la letra *A* a la izquierda de X significa que no hubo aleatoriedad en el proceso de selección. En ese tipo de diseño no hay evidencia alguna de la equivalencia entre los dos grupos antes del inicio del experimento. Un ejemplo de ese tipo de diseño ocurre cuando se seleccionan dos grupos de sujetos de determinada escuela, para pertenecer a los grupos experimental y de control, por el simple hecho de pertenecer al mismo grupo. Si los grupos son formados por alumnos que en el año anterior fueron buenos alumnos en Matemáticas o malos alumnos en esta disciplina, entonces la variable selección con seguridad influirá cualquier tratamiento alternativo sobre la enseñanza de Matemáticas que sea aplicado a un u otro grupo.

## Diseños experimentales

Los diseños experimentales son aquéllos donde se consigue controlar la mayor parte, si no todas, las fuentes de invalidez interna y externa. Siguiendo la tradición de Campbell y Stanley, estos diseños pueden ser clasificados en tres categorías.

**Diseño de tipo 4** - Un diseño experimental muy usado es el siguiente:

$$\begin{array}{llll} A & O_1 & X & O_2 \\ A & O_3 & & O_4 \end{array}$$

En este diseño, se trabaja con dos grupos y los sujetos de la investigación son designados *aleatoriamente* a uno de ellos (éste es el significado de *A*). Se observan los dos grupos antes de la aplicación del tratamiento *X*, por ejemplo, aplicando un pre-test a los dos grupos ($O_1 = O_3$). Uno de los grupos (grupo experimental) es entonces sometido al tratamiento *X*, mientras el otro (grupo de control) no recibe el tratamiento. Después, se observan los grupos, aplicando, por ejemplo, un post-test ($O_2 = O_4$) a los dos grupos. En la práctica, los pre-test y los post-test pueden ser iguales.

Un error común en el uso de diseños de ese tipo es analizar el resultado para determinar la eficacia del tratamiento del siguiente modo: se toman las diferencias entre los resultados del pre-test y del post-test en los dos grupos ($O_2 - O_1$ y $O_4 - O_3$), aplicando a continuación un test estadístico. Si la diferencia entre las medias del grupo experimental antes y después de la aplicación del tratamiento *X*, es estadísticamente significativa y la diferencia entre las medias del grupo de control no es significativa, se considera entonces el tratamiento como eficaz. Ésta es una forma equivocada de analizar la eficacia del tratamiento y no ofrece ninguna evidencia sobre el efecto del tratamiento *X*. La forma correcta de proceder es comparar el resultado *final* (las medias finales en un test de conocimiento, por ejemplo) entre los grupos experimental y de control entre sí.

Este diseño controla variables en la medida en que éstas influirán igualmente en los dos grupos, excepto *X*, obviamente, y, por tanto, los efectos de esas variables no pesarán en la comparación de las diferencias $O_2 - O_1$ y $O_4 - O_3$.

Además, la aleatoriedad de la designación de los sujetos a uno de los grupos, aunque no garantice equivalencia entre los grupos en 100%, reduce al mínimo la probabilidad de que sean diferentes. Según Kerlinger (1980, p. 102):

> *"Casualización es la designación de objetos (sujetos, tratamientos, grupos) de un universo a subconjuntos del universo de tal manera que, para cualquier designación dada a un subconjunto, todo miembro del universo tiene igual probabilidad de ser escogido para la designación. No hay total garantía de que la casualización 'igualará' los grupos, pero la probabilidad de igualar es relativamente alta. Hay otra forma de expresar esa idea: [...] ya que en procedimientos aleatorios todo miembro de una población tiene igual probabilidad de ser escogido, miembros con ciertas características distintas - hombre o mujer, alto o bajo grado de inteligencia, dogmático o no dogmático, y así sucesivamente - si son seleccionados, probablemente serán contrabalanceados a largo plazo por la selección de otros miembros de la población con la cantidad o calidad 'opuestas' de la característica."*

**Diseño tipo 5**[7] - La aleatoriedad de la designación de sujetos a los grupos de control y experimental es, por tanto, la más adecuada seguridad de que no existen diferencias iniciales o tendenciosidad entre los grupos. En ese caso, el pre-test no es condición esencial para que un diseño sea verdaderamente experimental. Así, el diseño anteriormente presentado podría ser simplemente:

| | | |
|---|---|---|
| *A* | *X* | $O_1$ |
| *A* | | $O_2$ |

De hecho, ese diseño no sólo puede ser usado en lugar del anterior sino que también es más adecuado, pues elimina toda influencia del pre-test en el experimento. Sin embargo, tal vez por razones psicológicas, muchos investigadores no renuncian a saber "con seguridad" si los grupos experimental y de control eran iguales en el inicio del experimento, de modo que el cuarto ejemplo de diseño aquí presentado es probablemente más usado que el quinto, aunque menos apropiado lógicamente.

**Diseño tipo 6 (Diseño de cuatro grupos de Solomon)** - Este diseño es la suma de las ventajas de los diseños cuatro y cinco. Su esquema es:

| | | | |
|---|---|---|---|
| *A* | $O_1$ | *X* | $O_2$ |
| *A* | $O_3$ | | $O_4$ |
| *A* | | *X* | $O_5$ |
| *A* | | | $O_6$ |

Este tipo de diseño controla variables como *interacción del pre-test con el tratamiento, maduración e historia*. La desventaja de ese tipo de diseño es la dificultad de obtener tantos grupos para participar de la investigación.

[7] Aquí se está haciendo una inversión entre la denominación dada por Campbell y Stanley a los diseños 5 y 6. Para Campbell y Stanley, lo que se está llamando diseño 5 es el diseño 6 y viceversa.

## *Diseños casi-experimentales*

Un tercer grupo de diseños identificado por Campbell y Stanley es el de los diseños cuasi-experimentales, o sea, aquéllos en los que le falta al investigador *"el pleno control de la aplicación de los estímulos experimentales - cuándo y quién exponer y la capacidad de casualizar exposiciones"* (op. cit., p. 61). Todos los diseños pertenecientes a ese grupo carecen del rigor y control existentes en los diseños pertenecientes al grupo de los diseños experimentales, pero pueden ser usados cuando la situación no permita el uso de diseños verdaderamente experimentales.

**Diseño tipo 7 (Serie temporal) -** El diseño "serie temporal" ejemplariza esa situación:

$$O_1\ O_2\ O_3\ O_4\quad X\quad O_5\ O_6\ O_7\ O_8$$

En este diseño, los sujetos son observados varias veces antes de aplicar el tratamiento *X* y varias veces después de la aplicación. Suponiendo que antes del tratamiento las observaciones fuesen casi homogéneas, sin variaciones, presentando una calidad bien definida y que hubiese un salto en los resultados de las observaciones realizadas después del tratamiento y que, a partir de ahí, hubiese una nueva estabilización en los resultados de las observaciones, con la presentación de otro nivel, ese salto cuantitativo en la serie temporal sería tomado como evidencia del efecto *X*.

Hay que observar que ese diseño es semejante al primero presentado como ejemplo, sin embargo implica muchas observaciones más, lo que minimiza, aunque no excluya, las deficiencias del primero. También hay que observar que implica la existencia de un único grupo, que, en la práctica, es una ventaja, pues muchas veces es difícil obtener dos grupos de sujetos.

Un ejemplo simple de aplicación de ese diseño sería aquél en que el profesor observa cuidadosamente sus alumnos durante algunas semanas, haciendo varias mediciones (que pueden ser testes de aprovechamiento o de actitud) antes de hacer uso de una nueva estrategia de enseñanza. De ese modo, después del uso de la estrategia, vuelve a observar a sus alumnos, durante algún tiempo, haciendo nuevos registros. Diferencias, cualitativas o cuantitativas, en el desempeño de los alumnos después del uso de la estrategia, y que se mantienen a lo largo del tiempo, pueden ser tomadas como evidencia del efecto de la estrategia sobre el aprendizaje cognitivo o afectivo de los alumnos.

**Diseño de tipo 8 (Muestras temporales equivalentes) -** Este diseño es, de hecho, una variación del diseño anterior. En este diseño se introduce la variable experimental (el tratamiento *X*) alternadamente y se observa el grupo. Su esquema es el siguiente:

$$O_1\ X\ O_2\ X_0\ O_3\ X\ O_4\ X_0\ O_5\ X\ O_6\ X_0\ O_7\ X\ O_8$$

Como se puede ver en ese diseño, el mismo grupo de sujetos es observado alternadamente en la presencia del tratamiento y sin la presencia del tratamiento (aquí simbolizada por el símbolo $X_0$). El análisis es realizado a partir de la comparación de los valores medios del grupo con y sin tratamiento experimental. En ese punto, ese diseño se asemeja a un diseño con dos grupos.

**Diseño de tipo 9 (Grupo de control no equivalente) -** Este diseño tiene la siguiente estructura:

$O_1$ $X$ $O_2$
$O_3$ $O_4$

En este caso, el grupo de control y el grupo experimental no poseen equivalencia muestral, pues no fue usada la aleatoriedad en la selección de las muestras. En este tipo de diseño, los grupos constituyen colectivos reunidos naturalmente, tales como clases escolares ya compuestas previamente a la acción del investigador. El control del investigador consiste únicamente en la decisión sobre cuál de los grupos va a recibir el tratamiento y cuándo.

Seguramente, en este tipo de diseño habrá problemas serios derivados del factor *selección* y de su interacción con otros factores importantes tales como *historia, maduración*, etc.

Campell y Stanley proponen otros cinco diseños casi- experimentales que no serán presentados aquí.

En esta sección se dio énfasis bastante grande al diseño porque ésta es una cuestión crucial en la realización de una investigación cuantitativa en enseñanza. Así como el investigador debe formular una pregunta de investigación clara, orientadora y relevante, también debe investigarla usando un diseño adecuado.

Un mal diseño puede invalidar las afirmaciones de conocimiento (resultados) y de valor de una investigación, echando por tierra todo el trabajo realizado, sea por no controlar las fuentes de invalidez interna sea por no controlar las de invalidez externa.

Siempre que sea posible, se debe utilizar uno de los diseños experimentales. Cuando eso no es posible, la alternativa es el uso de uno de los diseños casi-experimentales aquí expuestos (y discutidos con mayor extensión en Campbell y Stanley, 1963, 1991) teniendo en mente las deficiencias que estos diseños ofrecen. La investigación no debe dejar de ser realizada si la situación no permite el uso de un diseño puramente experimental, pero el investigador debe dejar claro para sí y para los demás investigadores la limitación de objetivo, en lo que se refiere a la validez de sus resultados, así como cuáles son los puntos donde se deben de realizar nuevos trabajos de modo a estudiar la interferencia de factores que no pudieron ser controlados en aquel experimento específico. Hay que recordar que no es una única investigación lo que construye el cuerpo de conocimientos de un área, sino un conjunto de ellas.

## Un poco de estadística

El objetivo de esta sección y de las próximas es discutir los principales tópicos relacionados con el análisis de experimentos cuantitativos en enseñanza. Como se dijo anteriormente, la principal herramienta para ese tipo de análisis es la Estadística. Se justifica, por tanto, dedicar algún espacio a ella.

### Muestreo aleatorio

El término técnico **aleatorio** indica que la muestra es seleccionada de tal modo que cada elemento de la población tiene una posibilidad igual de entrar en la muestra. El investigador debe tener una lista completa de todos los elementos de la población y entonces

seleccionar su muestra de modo que ningún elemento de la población sea privilegiado por el procedimiento de la elección.

El propósito de la aleatoriedad no es garantizar que los dos grupos se portarán igualmente bien en la ausencia del tratamiento. La aleatoriedad no garantiza igualdad. La aleatoriedad permite evitar aquel tipo de resultado que podría ser atribuido a la variabilidad de la muestra. Aleatoriedad es un procedimiento para seleccionar muestras y no una característica de la muestra. Ella también no asegura representatividad, ni da indicativo de cómo se portará la muestra.

**Error de muestreo** es aquél tipo de error que se comete al seleccionar muestras aleatorias para representar la población. En virtud de este tipo de error, es prácticamente imposible para un grupo pequeño ser exactamente representativo de otro mucho mayor. Ese error de muestreo está presente cada vez que se seleccionan muestras, no importa el cuidado con el que se realice la selección aleatoria. A continuación son definidos algunos tipos de procedimientos de selección de muestras:

- **Muestreo aleatorio simple** - es el proceso de seleccionar observaciones de un grupo mayor de tal modo que cada sujeto en la población de donde se está seleccionando la muestra tenga una probabilidad igual e independiente de ser seleccionado.
- **Muestreo estratificado** - es, algunas veces, un modo recomendado de proceder al escoger muestras. Se divide la población en grupos menores homogéneos de modo que se obtenga una mejor representación. Con cada subgrupo, se puede usar algún proceso de selección aleatoria. Este proceso le da al investigador un muestreo más significativo del que se obtendría directamente de la comunidad entera. Para tener una representación más fiel de la población como un todo, se le puede dar pesos al número de sujetos pertenecientes a los diferentes grupos de manera que se tenga una representación proporcional a la distribución en la población.
- **Muestreo estratificado proporcional** - este tipo de proceso de muestreo tiene lugar cuando se toma un porcentaje de cada grupo en el proceso de composición de la muestra.
- **Muestreo sistemático** - se tiene ese tipo de muestreo cuando una población está listada y entonces se hace algún tipo de selección, según algún criterio, como por ejemplo, tomar el *enésimo* elemento de la lista.
- **Muestreo por agrupación** - es una variación del muestreo aleatorio simple, particularmente apropiado cuando la población es grande o cuando la distribución geográfica de la población es diseminada. A cada agrupación se le da un número y se seleccionan grupos escogidos aleatoriamente. El uso del muestreo por agrupación es generalmente escogido por razones económicas y aspectos administrativos.

En cualquier tipo de muestreo, la característica de la muestra inevitablemente diferirá en algún grado, aunque pequeño, de la característica de la población. Pero, cuando se usa el muestreo aleatorio, las posibilidades de que el error de muestreo influya en la variable dependiente en una dirección particular son las mismas que existen de influir la misma variable en otra dirección cualquiera. *Muestreo aleatorio es la única forma de muestreo por la cual alguna cantidad de error específica puede ser estimada. Se puede decir que es el muestreo aleatorio lo que diferencia las investigaciones experimentales de las investigaciones no experimentales*.

## *Medidas de tendencia central*

Se entiende por **distribución de frecuencias** una tabulación (listado) de los resultados obtenidos en una cierta muestra con el número de veces en que esos resultados aparecen en la muestra. Así, un listado con las notas obtenidas por los alumnos en determinada evaluación con el número de veces que aparece cada nota es una distribución de frecuencias.

Cuando el número de sujetos pertenecientes a la(s) muestra(s) es pequeño, se puede tener una idea de cómo se comportan los valores. Sin embargo, si tenemos muestras muy grandes, es difícil tener una idea del comportamiento de la muestra y, entonces, tenemos que recurrir a transformaciones sobre los valores. Una de las formas de hacer esto es usando las llamadas **medidas de tendencia central**.

Se define una medida de tendencia central como un número alrededor del cual se distribuyen los valores de la distribución de frecuencias. En Estadística, se usan básicamente tres tipos de medidas de tendencia central: la media, la moda y la mediana.

- **La Moda** de una distribución de frecuencias es definida como el valor que tiene la más alta frecuencia.
- **La Mediana** es definida como el valor que marca el punto medio del conjunto de datos, o sea, aquél valor para el cual se tiene 50 % de los resultados con valores superiores y 50 % de los resultados con valores inferiores.
- **La Media aritmética simple** de una distribución de frecuencias es definida como el número obtenido a partir de la división de la suma total de todos los resultados obtenidos por el número de elementos en la distribución de frecuencia. Matemáticamente:

$$\overline{X} = \frac{\sum X_i n_i}{N}$$

donde las $X_i$ son los resultados obtenidos, $n_i$ es el número de veces que aparece cada resultado y $N$ es el número de elementos en la distribución de frecuencias. El símbolo $\sum$ significa que se está sumando.

*Ejemplo*: sea la distribución de frecuencias que constan en la Tabla 1.

*Tabla 1 - Distribución de frecuencias para un examen de una disciplina hipotética.*

| | |
|---|---|
| *Clase 1* | 9,8 |
| *Clase 2* | 7,4 |
| *Clase 3* | 6,2 |
| *Clase 4* | 6,0 |
| *Clase 5* | 5,9 |
| *Clase 6* | 4,5 |
| *Clase 7* | 3,4 |
| *Clase 8* | 3,4 |
| *Clase 9* | 3,4 |

| | |
|---|---|
| *Clase 10* | 1,0 |
| *Clase 11* | 0,5 |

La **moda** de esta distribución es dada por el valor *3,4*, pues es el que aparece el mayor número de veces en la distribución (3 veces). La **mediana** de la distribución es el valor *4,5*, pues, para ese valor, *50* % (5 resultados) son mayores que él y *50* % son menores. Por fin, la **media aritmética simple** de la distribución es dada por:

$$\overline{X} = \frac{9{,}8 + 7{,}4 + 6{,}2 + 6{,}0 + 5{,}9 + 4{,}5 + 3 \times 3{,}4 + 1{,}0 + 0{,}5}{11}$$

$$\overline{X} = 4{,}7$$

Otros tipos de media pueden ser definidos como, por ejemplo, la **media aritmética ponderada** y la **media geométrica**.

La media aritmética ponderada es usada cuando se desea tener una medida de tendencia central de cierta distribución de frecuencias donde los resultados contribuyen con pesos diferentes. Matemáticamente es definida por:

$$\overline{X_p} = \frac{\sum pn_i X_i}{\sum p}$$

donde $\overline{X}_p$ es el valor de la media ponderada, $p$ es el peso atribuido a cada resultado $X_i$ y $n_i$ es el número de veces que aparece el resultado en la distribución de frecuencias.

La **media geométrica** es definida por:

$$\overline{X_g} = \sqrt[N]{\prod n_i X_i}$$

donde $\overline{X}_g$ es la media geométrica, el símbolo $\Pi$ significa que debemos multiplicar los elementos que vienen a continuación y $X_i$ y $n_i$ fueron definidos anteriormente. Ese tipo de media es usado cuando los valores de la distribución se obtienen unos de los otros a partir de un factor multiplicativo.

## *Medidas de variabilidad*

La media, sea del que tipo sea, no dice todo a respecto de una distribución de frecuencias. Aquí cabe un comentario a respecto de los procesos estadísticos y la pérdida de información que necesariamente ocurre cuando se usan números, tales como la media, para que representen distribuciones de frecuencias. En el proceso de mediación tiene lugar una pérdida de información una vez que se sustituye la información total, o sea, la distribución, por algo que pretende ser representativo de esa distribución. Al hacerlo, se pierde la estructura detallada de información dada por el conjunto completo de valores. Así, por ejemplo, consideremos las hipotéticas Tablas 2 y 3:

| *Tabla 2 - Notas en la Clase 1.* | |
|---|---|
| *alumno 1* | *5,1* |
| *alumno 2* | *5,9* |
| *alumno 3* | *7,2* |
| *alumno 4* | *5,9* |
| *alumno 5* | *5,9* |

| *Tabla 3 - Notas en la Clase 2.* | |
|---|---|
| *alumno 1* | *9,0* |
| *alumno 2* | *4,0* |
| *alumno 3* | *6,0* |
| *alumno 4* | *7,0* |
| *alumno 5* | *2,0* |

Si se calcula la media aritmética simple para esas dos distribuciones de frecuencia, se obtendrá para ambas la media *6,0*. Sin embargo, si esas distribuciones representan notas de alumnos en dos clases diferentes, en una asignatura cuya media de aprobación sea *6,0*, se observa, entonces, que en la clase *1* hay *1* sujeto aprobado mientras que en la clase *2* hay *3*. Observando solamente la media de las dos clases, esa información se perdería.

Una forma de minimizar esa pérdida de información, como consecuencia del uso de una medida de tendencia central, es usando las **medidas de variabilidad**. Una medida de variabilidad indica lo diseminados que están los valores en la distribución. O sea, una medida de variabilidad es una forma de tener una idea de cuánto se alejan los valores de la medida de tendencia central que se está utilizando.

La más simple de las medidas de variabilidad es el **intervalo (I)**. El intervalo dice entre qué valores se distribuyen los resultados de la distribución que se está analizado. Así, en el ejemplo, el intervalo para la distribución *1* es dado por *[7,2; 5,1]* mientras que en la distribución *2* el intervalo es dado por *[9,0; 2,0]*.

Otra medida de variabilidad es el **desvío de la media (dm)**. Esa cantidad dice cuánto se desvían los valores de la media. En el ejemplo dado, el valor *9,0* de la Tabla *3* posee un desvío de la media de *3,0 (9-6=3).*

Una medida de variabilidad de las más utilizadas en análisis de distribuciones de frecuencia es la **desviación estándar *(dp)***. La desviación estándar indica lo diseminada que es una distribución. La desviación estándar tiene una interpretación muy simple, originada de la ecuación que define la *distribución normal* (que será analizada en la próxima sección). Tomando un intervalo definido por $[\overline{X} - dp;\ \overline{X} + dp]$ se tiene dentro de ese intervalo alrededor de *68* % de los valores de la distribución. Haciendo una nueva medida, hay una *probabilidad* de *68* % de que esa nueva medida pertenezca a ese intervalo.

Considerando intervalos definidos por múltiplos de la desviación estándar se estará englobando un número cada vez mayor de valores dentro del intervalo en cuestión.

Matemáticamente, la desviación estándar es dada por:

$$dp = \sqrt{\frac{\sum_1^N (X_j - \overline{X})^2}{N}}$$

donde $N$ es el número de valores, $X_j$ significa el *j-ésimo* elemento de la distribución de valores, $\overline{X}$ es el valor medio de la distribución.

## Curva normal

Se define como **curva de la distribución** la representación gráfica de una frecuencia de distribución de resultados, donde los valores de los resultados son indicados en el eje horizontal y los valores de las frecuencias de los resultados particulares indicados en el eje vertical. Curvas de distribución pueden venir en diferentes formas y tamaños. Sin embargo, muchas frecuencias de distribución tienden a seguir un cierto modelo llamado **distribución normal**, especialmente cuando existen muchos resultados. La forma de la curva que se puede trazar de estas distribuciones es llamada **curva normal** (Figura 1). Una de las características de la curva normal es la *simetría*; otra característica importante es que la media, la mediana y la moda son *idénticas*.

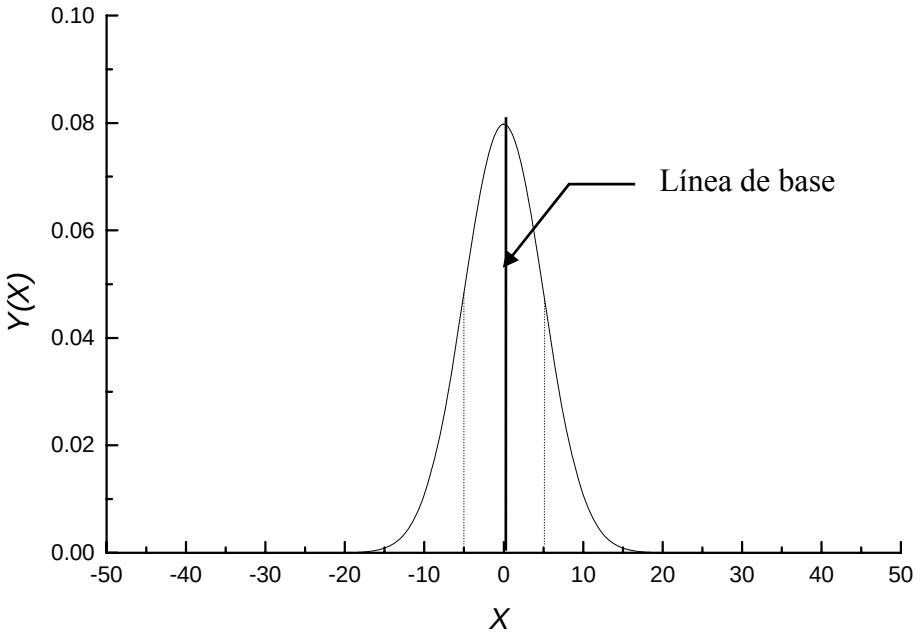



*Figura 1 - La curva normal*

En la curva normal, la distancia de la línea de base hasta la primera línea punteada es la **desviación estándar** de la distribución, por definición[8].

La ecuación que define la curva normal es dada por:

$$Y(X) = \frac{1}{dp\sqrt{2\pi}} \exp\left[ -\frac{1}{2} \frac{(X - \overline{X})^2}{dp^2} \right]$$

donde aquí *dp* es la desviación estándar, $\overline{X}$ es el valor medio y *X* es el valor para el cual se quiere calcular *Y*.

[8] En el presente ejemplo, la curva fue realizada con una desviación estándar asumida de 5.

## Intervalos de confianza

Cuando se dice que un resultado aleatoriamente seleccionado caerá dentro de un intervalo específico de los valores de los resultados obtenidos, es necesario hacerlo con algún grado de confianza, o sea, sabiendo lo probable que es que sea correcto. El intervalo de confianza de 95 % es llamado **intervalo de confianza** porque, si la distribución es normal, se puede estar seguro de que 95 % de las veces que se escoja aleatoriamente un valor de la variable en cuestión, éste estará dentro de ese intervalo. En términos de probabilidades, el intervalo de confianza de 95 % designa los dos valores entre los cuales existe la probabilidad *p=0,95* de que un resultado seleccionado aleatoriamente pertenezca a ese intervalo. El intervalo de confianza de 95 % es dado por:

$$I_{95\%} = \left[\overline{X} - 1{,}96dp; \overline{X} + 1{,}96dp\right]$$

donde $dp$ es la desviación estándar y $\overline{X}$ la media calculadas para la distribución.

El intervalo de confianza de 99 % es otro intervalo normalmente utilizado. Los límites de ese intervalo están entre:

$$I_{99\%} = \left[\overline{X} - 2{,}58dp; \overline{X} + 2{,}58dp\right]$$

## Distribución de medias muestrales

Supongamos que son seleccionadas de una población hipotética un gran número de muestras, cada una de las cuales, por ejemplo, con *50* sujetos, y que se calcule el resultado medio para cada grupo en alguna variable de interés. Se puede entonces trazar una curva representando la distribución de esos resultados. En esa distribución de medias muestrales, la media **de todas las medias** es la media de la población y las medias muestrales son distribuidas alrededor de la media de la población siguiendo la distribución normal.

Cuando se trata de distribuciones de medias, la desviación estándar es llamada **desviación estándar de la media (*dpm*)**. La interpretación de esa cantidad es semejante a la de la desviación estándar de medida: tomando otra muestra de la población, la probabilidad de que la **media** obtenida por esa nueva muestra esté dentro del intervalo de confianza de 95 % es dada por:

$$I_{95\%} = \left[\overline{X} - 1{,}96dpm; \overline{X} + 1{,}96dpm\right]$$

## Inferencias con respecto al valor de la media de la población

Cuando se sabe la media de una muestra y se toma una estimativa de la desviación estándar de la media, no es posible inferir el valor de la media de la población de la media que se tiene para aquella muestra, pero se puede hacer hipótesis sobre el valor para la media de la población y, usando la estimativa de la desviación estándar de la media, determinar la probabilidad de obtener una media muestral que difiera de la media hipotética de la población tanto cuanto se quiera.

Supongamos que una media muestral sea *97* y que la desviación estándar de la media estimada sea de *dpm=2*. Supongamos también que la hipótesis realizada fue que la media de la población es *100*. Se puede, entonces, determinar la probabilidad de obtener nuestra media de la muestra de *97*.

La Figura *2* indica que el intervalo de confianza de 95 % varía de *96,08* a *103,92*. Esto dice que si la media de la población es *100*, entonces, la probabilidad de seleccionar una muestra cuya media esté dentro del intervalo de confianza considerado es *p=0,95*. Con otras palabras, la probabilidad de obtener una media muestral menor que *98,08* o mayor que *103,92* es *p=0,05*. Por tanto, se puede aceptar la hipótesis de que la muestra, cuya media es *97*, viene de una distribución de medias muestrales, retiradas de una población cuya media es *100*.

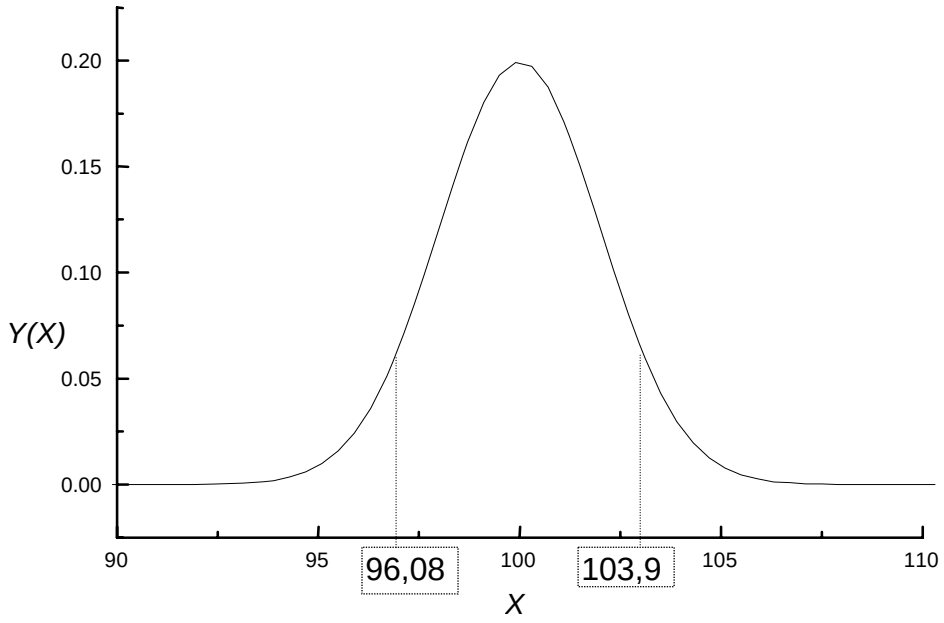



*Figura 2 - Curva normal indicando intervalo de confianza de 95 %.*

## Comparaciones entre múltiples muestras

En situaciones de investigación, comúnmente se desea comparar dos o más muestras. Por ejemplo, se puede querer determinar si existe una diferencia en adquisición de conocimientos entre estudiantes que son enseñados por un método *A* y los que son enseñados por un método *B.* La cuestión es: *¿cuál es la probabilidad de que las diferencias entre las dos medias de las muestras se deban simplemente a error de muestreo?* En otras palabras, ¿se *puede atribuir la diferencia de medias entre las dos muestras al error aleatorio en las muestras, o los alumnos enseñados por un método realmente aprenden más que los enseñados por el otro método? ¿Tenemos de hecho dos muestras pertenecientes a poblaciones diferentes, representadas por dos distribuciones normales de resultados?*

Al testar la **hipótesis nula** se está asumiendo que la diferencia entre las dos muestras se debe simplemente al error de muestreo.

## *Tipos de hipótesis*

Cuando se quiere decidir si determinado procedimiento es mejor que otro, se formula la hipótesis de que **no existe diferencia** entre los procedimientos (es decir, cualquier diferencia observada es meramente debida a oscilaciones al tomar dos muestras **de la misma** población). Este tipo de hipótesis es la llamada **Hipótesis Nula** denotada por $H_0$. La hipótesis alternativa a la $H_0$, o sea, de que la diferencia observada **no se debe** meramente al muestreo, es denotada por $H_1$.

Al decidir si se rechaza o no la hipótesis nula, puede haber dos tipos de errores. El primer tipo de error tiene lugar cuando se rechaza la hipótesis nula con base en datos de muestras que de hecho vienen de la misma población, se dice que se comete un error ***Tipo I.*** Por otro lado, se acepta la hipótesis nula cuando, de hecho, las muestras vienen de poblaciones diferentes, se dice que se cometió un error del ***Tipo II***[9].

En la práctica, según el caso, un tipo de error puede ser más serio que el otro y, así, una solución de compromiso debe ser obtenida en favor de una limitación del error que, en aquella situación, sea considerado más serio. La única forma de disminuir la influencia de esos dos tipos de error es aumentar el tamaño de la muestra, lo que no siempre es posible.

Al testar la hipótesis nula, la probabilidad máxima con la cual el investigador acepta correr el riesgo de cometer un error de *tipo I* es llamada **nivel de significancia estadística** de la investigación. Cuando el investigador decide el nivel de probabilidad que usará al rechazar la hipótesis nula, estará dando la probabilidad con la cual se arriesga a equivocarse en su decisión. Si selecciona el nivel de significancia como *0,05*, está diciendo que hay una probabilidad de *0,05* de que esté equivocado. Si no quiere correr un riesgo de error tan significativo, puede escoger para nivel de significancia *p=0,01*. En este nivel, es menos probable que esté cometiendo un error del *tipo I*, sin embargo estará incrementando la probabilidad de cometer un error del *tipo II*.

En la práctica de la investigación en educación, los niveles de significancia de *0,05 y* 0,01 son usuales, aunque se utilicen también otros niveles de significancia.

Una hipótesis que no indica la dirección de la diferencia esperada, sino que meramente establece que existe una diferencia, es llamada **hipótesis bilateral** (*two-tailed).* Ese tipo de hipótesis es así designada porque está preocupada con las dos "colas" de la distribución normal de las diferencias entre medias muestrales.

Una hipótesis que afirma qué tratamiento es mejor que el otro se llama **hipótesis *unilateral*** *(one-tailed*) porque está preocupada solamente con uno de los lados de la distribución de diferencias entre medias muestrales.

## *Distribución de medias de pequeñas muestras*

Hasta aquí fueron consideradas muestras que contienen un gran número de sujetos (treinta o más). Las propiedades de la distribución normal son válidas para grandes muestras,

[9] Técnicamente, el investigador no debería de aceptar la hipótesis nula, "mas sim, falhar" en rechazar la hipótesis nula.

pero no cuando hay un número pequeño de sujetos en cada muestra. La distribución tiende a ser achatada cuando, en cada muestra, el número de sujetos es pequeño.

Para fines estadísticos, eso significa que, para datos donde las muestras son pequeñas, no es posible usar las propiedades de la curva normal para decidir a favor o contra la aceptación de la hipótesis nula. En lugar de eso, se deben usar valores que reflejen ese achatamiento de la curva normal. Esos valores son llamados valores **t** para los cuales también fueron calculados valores para los niveles de significancia *p=0,05* y *p=0,01* para muestras de cualquier tamaño. Existen tablas estadísticas preparadas (una de las cuales se presenta en la Tabla 4) para estos valores **t** para todos los tamaños de muestras siendo comparadas, de modo que sabiéndose cuántos sujetos están en cada muestra que se está comparando, podremos fácilmente determinar el valor de **t** necesario para el nivel de significancia escogido (normalmente *0,05* o *0,01*).

Si un valor **t** indica diferencias dentro del intervalo de confianza de 95 %, su valor normalmente **no es comunicado**. Al revés, el investigador afirma que el valor **t** no es significativo. En este caso, acepta la hipótesis nula y atribuye la diferencia observada entre sus muestras al simple error de muestreo.

Este tipo de test estadístico es llamado **test t** y es utilizado para comparación entre medias de muestras pequeñas **cuando, por hipótesis, las muestras fueron escogidas aleatoriamente y los resultados proceden de poblaciones distribuidas según la distribución normal.** Otros tests estadísticos son disponibles si no se puede realizar la hipótesis de normalidad.

Muestras con número de sujetos menor de 30 son llamadas pequeñas muestras. Un estudio estadístico de distribuciones muestrales, en el cual las muestras son pequeñas, es llamado *Teoría de Pequeñas Muestras*. Sin embargo, un nombre más apropiado sería *Teoría Exacta del Muestreo*, ya que los resultados obtenidos se mantienen tanto para pequeñas como para grandes muestras. Una distribución importante es la distribución **t de Student**[10]. Esta distribución es dada, matemáticamente, por:

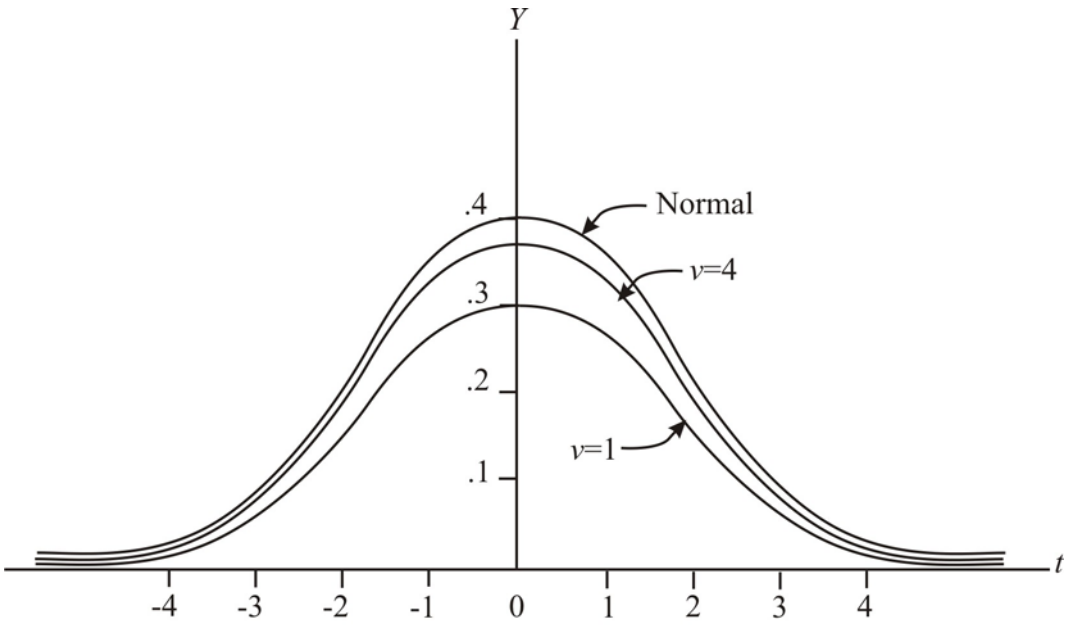
$$Y = \frac{Y_0}{\left(1 + \frac{t^2}{N-1}\right)^{N/2}}$$

$$Y = \frac{Y_0}{\left(1 + \frac{t^2}{gl}\right)^{(gl+1)/2}}$$

donde $Y_0$ es una constante que depende de $N$ de tal modo que el área total sobre la curva es 1[11]. La cantidad N-1 es llamada *grado de libertad* (*gl*). Para grandes valores de *gl* o *N* (*N 30*), las curvas de la figura se aproximan a la curva normal estándar. La F*igura 3* muestra varias curvas de esta distribución para varios valores de grados de libertad *gl.*

*gl=N-1*

[10] Se mantiene aquí el nombre en inglés por ser consagrado en la literatura y por el uso en el área.
[11] A esto se llama condición de normalización.




Student's ***t*** distributions for various values of ***v***.

*Figura 3 - La distribución* ***t*** *de Student para distintos grados de libertad.*

Para fines de cálculo entre dos muestras, con $N_1$ y $N_2$ sujetos, de medias y desvíos estándar dados por $\overline{X}_1$, $dp_1$, $\overline{X}_2$ y $dp_2$ respectivamente, el valor de **t** es dado por:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

donde la cantidad σ es definida por:

$$\sigma = \sqrt{\frac{N_1 dp_1^2 + N_2 dp_2^2}{N_1 + N_2 - 2}}$$

Valores de **t** fueron calculados, para niveles de significancia de *0,05* y 0,01, para cualquier tamaño de muestra. Los estadísticos prepararon tablas estadísticas de estos valores **t** para todos los tamaños de las muestras siendo comparadas, de modo que, si sabemos cuántos sujetos existen en cada muestra, podemos compararlas fácilmente y determinar el valor **t** necesario para el nivel de significancia deseado (*0,05* ó *0,01*). Tabla de ese tipo está ejemplificada en la Tabla 4.

*Tabla 4 - Tabla t para hipótesis unilaterales y bilaterales.*

| | *Nivel de significancia para hipótesis unilateral* | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **gl** | .40 | .25 | .10 | .05 | .025 | .01 | .005 | .0025 | .001 | .0005 |
| | *Nivel de significancia para hipótesis bilateral* | | | | | | | | | |
| | .80 | .50 | .20 | .10 | .05 | .02 | .01 | .005 | .002 | .001 |
| **1** | 0.325 | 1.000 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 127.32 | 318.31 | 636.62 |
| **2** | .289 | 0.816 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 14.089 | 22.326 | 31.598 |
| **3** | .277 | .765 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 7.453 | 10.213 | 12.924 |
| **4** | .271 | .741 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 5.598 | 7.173 | 8.610 |
| **5** | 0.267 | 0.727 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 4.773 | 5.893 | 6.869 |
| **6** | .265 | .718 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 4.317 | 5.208 | 5.959 |
| **7** | .263 | .711 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.029 | 4.785 | 5.408 |
| **8** | .262 | .706 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 3.833 | 4.501 | 5.041 |
| **9** | .261 | .703 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 3.690 | 4.297 | 4.781 |
| **10** | 0.260 | 0.700 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 3.581 | 4.144 | 4.587 |
| **11** | .260 | .697 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 3.497 | 4.025 | 4.437 |
| **12** | .259 | .695 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.428 | 3.930 | 4.318 |
| **13** | .259 | .694 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.372 | 3.852 | 4.221 |
| **14** | .258 | .692 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.326 | 3.787 | 4.140 |
| **15** | 0.258 | 0.691 | 1.341 | 1.753 | 2.133 | 2.602 | 2.947 | 3.286 | 3.733 | 4.073 |
| **16** | .258 | .690 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.252 | 3.686 | 4.015 |
| **17** | .257 | .689 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.222 | 3.646 | 3.965 |
| **18** | .257 | .688 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.197 | 3.610 | 3.922 |
| **19** | .257 | .688 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 |
| **20** | 0.257 | 0.687 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 |
| **21** | .257 | .686 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 |
| **22** | .256 | .686 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.119 | 3.505 | 3.792 |
| **23** | .256 | .685 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.104 | 3.485 | 3.767 |
| **24** | .256 | .685 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.091 | 3.467 | 3.745 |
| **25** | 0.256 | 0.684 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.078 | 3.450 | 3.725 |
| **26** | .256 | .684 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.067 | 3.435 | 3.707 |
| **27** | .256 | .684 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.057 | 3.421 | 3.690 |
| **28** | .256 | .683 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.047 | 3.408 | 3.674 |
| **29** | .256 | .683 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.038 | 3.396 | 3.659 |
| **30** | 0.256 | 0.683 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.030 | 3.385 | 3.646 |
| **40** | .255 | .681 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 2.971 | 3.307 | 3.551 |
| **60** | .254 | .679 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 2.915 | 3.232 | 3.460 |
| **120** | .254 | .677 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 2.860 | 3.160 | 3.373 |
| ∞ | .253 | .674 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 2.807 | 3.090 | 3.291 |

## Fidedignidad y validez

Antes que se pueda hacer cualquier tipo de análisis de los datos, es necesario que el investigador se pregunte: *¿la forma de obtención de los datos me da seguridad sobre su validez*?

En secciones anteriores fue discutida la estructura de un experimento analizando los varios tipos de diseños posibles (destacando lo que diferencia un diseño del tipo experimental de los que no poseen esta propiedad), así como aspectos relacionados con las características de un buen test. Sin embargo, de nada vale ser cuidadoso en la elección del diseño de la

investigación si los instrumentos de recogida de datos son inadecuados. Inadecuación significa decir que el instrumento no mide correctamente lo que se propone a medir (cuestión relacionada con la **fidedignidad** del instrumento) o mide otra cosa diferente de la que se propone a medir (cuestión relacionada con la **validez** del instrumento). Para usar una analogía común cuando se habla de fidedignidad y validez, se puede imaginar la siguiente situación: un tirador de dardos acierta repetidamente el mismo lugar del objetivo. En este caso, se dice que hay fidedignidad, pues en varias repeticiones el tirador acierta siempre en el mismo lugar o, en otras palabras, consigue reproducir el mismo resultado (posición) con el mismo instrumento. Sin embargo, si la posición acertada no es el centro del objetivo, decimos que no hay validez, pues el objetivo del juego es acertar el centro. Si la posición acertada es siempre el centro, entonces decimos que hay fidedignidad y validez.

La herramienta básica para el análisis de fidedignidad es la correlación estadística entre variables. Veamos cómo se calcula esta cantidad y cuál es su significado.

## ¿Qué es la correlación entre dos variables?

Conceptualmente la *correlación* o el *coeficiente de correlación* (*r*) indica cómo se comportan dos o más variables, unas con relación a las otras. Cuando tenemos una correlación alta, esto indica que el crecimiento de una variable es acompañado por el crecimiento (en el caso de un coeficiente de correlación cerca de *+1*) o por la disminución (en el caso de un coeficiente de correlación cerca de *-1*) de la otra variable. Conviene recordar aquí que el hecho de que dos variables sean correlacionadas (tanto positiva como negativamente) *no* implica una relación causal entre las dos variables. Para el establecimiento de una relación causal entre ellas se debe recurrir a otras herramientas de análisis, o sea, debemos buscar en la teoría las razones de esa dependencia y los factores de comprobación de esa dependencia.

El grado de correlación es indicado por el valor del coeficiente de correlación, el cual es denotado por *R.* El coeficiente para una correlación perfectamente positiva es mostrado en la *Figura 4.a* y tiene valor *+1*. El coeficiente para una correlación perfectamente negativa es mostrado en la *Figura 4.b* y tiene valor de *-1*. Estos dos valores son los valores máximos para *R.* El coeficiente *r=0* indica la inexistencia de correlación. En este caso, el comportamiento de una variable no es relacionado de cualquier modo al comportamiento de la otra variable, como lo muestra la *Figura 4.c.*

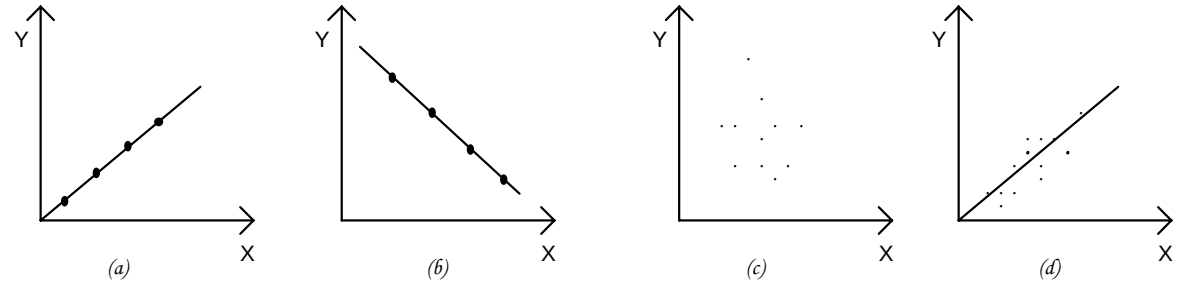



*Figura 4 - Representación gráfica de los varios tipos de correlación posibles entre variables.*

La *Figura 4.d* nos muestra un caso donde la correlación existe, pero es imperfecta. Por la simple observación de esa figura, podemos ver que los puntos tienden a alinearse a lo largo de una dirección específica, aunque no estén todos sobre una misma recta. Esto nos indica que la correlación es positiva, aunque no sea perfecta.

Tal como ocurre con los resultados medios de muestras, los coeficientes de correlación, calculados a partir de datos muestrales, son afectados por errores de muestreo. Así, como se hizo con las medias muestrales, se impone una pregunta: *¿cuál es la probabilidad de que el coeficiente de correlación obtenido a partir de los datos muestrales no sea fruto del error de muestreo y refleje una verdadera relación existente en la población*? Es razonable si toma por hipótesis que, como en toda inferencia hecha a partir de un proceso de muestreo, exista un error debido al propio proceso de muestreo. Sin embargo, a ejemplo de lo que sucede para otros tipos de tests estadísticos, existen tablas para varios valores de tamaño de muestra, a cualquier nivel de significancia estadística deseado[12].

## Cálculo del coeficiente de correlación

La forma de cálculo del coeficiente de correlación es función del tipo de variable con la cual estamos tratando. Como ya fue discutido, éstas pueden ser divididas en cuatro grupos: nominales, ordinales, intervalares o racionales. Para cada pareja de variables la forma de cálculo del coeficiente de correlación es diferente, debiendo tener en cuenta los tipos de variables involucrados. La fórmula de cálculo del coeficiente de correlación, definida a continuación, es válida solamente cuando las variables involucradas son (ambas) de los tipos *intervalares* o *racionales*. Para otros tipos de variables, se sugiere consultar Glass y Stanley (1970).

El término *correlación, tal como está siendo usado aquí, significa correlación lineal*. En ese caso, el coeficiente de correlación entre dos variables *X* y *Y* es dado por:

$$r_{xy} = \frac{\sum xy}{\sqrt{\left(\sum x^2\right)\left(\sum y^2\right)}}$$

donde $x = X - \overline{X}$ e $y = Y - \overline{Y}$[13].

Esta ecuación es llamada *fórmula producto-momento*[14]. Hay que observar la simetría entre las variables que componen esa ecuación: si cambiamos *x* con *y*, el resultado es el mismo. Con eso $r_{xy} = r_{yx}$ o sea, es lo mismo calcular el coeficiente de correlación llamando una de las variables *X* y la otra *Y* que calcular denominando inversamente las variables.

## Validez y fidedignidad de tests

Tests de conocimiento intentan medir lo que un individuo aprehendió en una cierta área, su nivel actual de dominio de un cierto contenido o su desempeño. Muchos tests usados en escuelas son tests de conocimiento. Frecuentemente, resultados en tests de conocimiento son usados en la evaluación de cursos, profesores, métodos de enseñanza y otros factores considerados significativos en la práctica educacional. Son usados en la clasificación, promoción o retención de estudiantes en niveles particulares de enseñanza. Son usados como

[12] Usualmente los niveles *0,01* y *0,05*.
[13] Los valores con barra, siguiendo la convención anterior, son las medias en X e Y.
[14] Hay que observar que esa ecuación suministra automáticamente la señal correcta para el coeficiente de correlación.

herramientas para diagnosticar puntos débiles y fuertes de los estudiantes y como una base para premios, recompensas, etc. dentro del ambiente escolar.

Tests de aptitud intentan predecir el grado de conocimiento que se puede esperar de individuos en una actividad particular. Estos tests intentan prever la capacidad de un individuo en particular para adquirir un mejor desempeño a partir de entrenamiento adicional. De hecho, capacidad (o aptitud) no se puede medir directamente. Aptitud solamente puede ser inferida con base en el desempeño presentado. Tests de aptitud pueden ser usados para dividir los estudiantes en grupos relativamente homogéneos con fines instruccionales de modo a identificar estudiantes para becas escolares o para elaborar guías individuales dentro de áreas donde ellas más probablemente tendrán éxito.

En la investigación, tests son instrumentos de medida usados por los investigadores para, de un modo general, recoger datos. De este modo, si la investigación es de calidad, se supone que los procedimientos usados para la recogida de datos (y su análisis) deben ser buenos. La calidad de la investigación no puede ser mejor que la calidad de los procedimientos usados para recoger y analizar los datos.

Como ya se dijo, esos instrumentos deben tener dos características indispensables:

1. **Fidedignidad**: un test es fidedigno si sus medidas son precisas y consistentes cuando se aplica en tiempos diferentes.
2. **Validez**: en general, un test posee validez si mide lo que se propone a medir.

Un test puede ser fidedigno, aunque no sea válido. Un test válido siempre es fidedigno. No existe una única forma de validez o fidedignidad de un instrumento. Existen muchos tipos de validez y fidedignidad. En general, fidedignidad está relacionada con consistencia y validez está relacionada con la interpretación del test. Un test puede ser válido para un objetivo y no ser válido para otro.

## *Fidedignidad*[15]

Reiterando, por fidedignidad entendemos cuál es la precisión de los datos, en el sentido de su estabilidad o reproducibilidad. Un instrumento de recogida de datos fidedignamente perfecto es aquél que si fuese aplicado dos veces bajo las mismas circunstancias nos suministraría los mismos resultados. Como se vio, la *correlación* es la herramienta estadística básica en el análisis de fidedignidad de tests. Una correlación de *1,00* indicaría perfecta fidedignidad, mientras que correlación *0,00* indicaría ninguna fidedignidad. Correlaciones intermedias indicarían diferentes niveles de fidedignidad.

Son varias las formas de medir la fidedignidad de un test:

**1. Test-retest** - es exactamente lo que el nombre implica; producimos el primer conjunto de datos a través de la aplicación del test en un determinado instante de tiempo y, después de algún intervalo de tiempo lo suficientemente largo como para olvidar el test, pero lo suficientemente corto como para que no se produzcan alteraciones significativas en las personas que responden, un segundo conjunto de datos es obtenido por la aplicación del mismo test al mismo conjunto de personas (el *re-test*). Después de la segunda aplicación los

[15] Parte de esa sección está basada en *The Research Process in Education*, por D.J. Fox, 1969.

dos conjuntos de datos son correlacionados y la correlación obtenida estima la fidedignidad del test.

El mayor problema en esta estrategia se refiere al intervalo de tiempo. ¿Cómo definir lo que se entiende por intervalo de tiempo lo suficientemente grande para olvidar el test pero lo suficientemente corto de modo que no se produzcan alteraciones en las personas que responden? La respuesta a esta cuestión depende del tipo de investigación que se está realizando.

El procedimiento de *test-retest* es más aplicable a las situaciones en que la aptitud, habilidad o conocimiento que se está evaluando posee estabilidad, cambiando en escalas de tiempo largas frente al período entre el *test* y el *retest*. Ésta, obviamente es una limitación de orden práctica siendo, bajo el punto de vista técnico, muy difícil garantizar la integridad de las personas que responden entre el *test* y el *retest*. Otro aspecto peligroso de la estrategia del *test-retest* es que estamos calculando la correlación entre datos provenientes de dos exposiciones de los sujetos al mismo contenido. De ese modo, si existe un direccionamiento del *test* en alguna dirección será, desde luego, un direccionamiento consistente en cada una de las exposiciones al instrumento.

**2. Forma alternada**: en esta forma de medición de fidedignidad, el investigador debe desarrollar dos formas paralelas o equivalentes de su instrumento, digamos formas *A* y *B*, administrar ambas a las mismas personas y correlacionar los dos conjuntos de datos obtenidos. La administración de las dos formas puede ser con un intervalo de tiempo entre las dos aplicaciones o de una única vez, aplicando las dos formas consecutivamente. En los dos casos, es aconsejable alternar el orden de aplicación entre las dos formas del test, es decir, la mitad responde a la forma A del test y después a la forma *B* y la otra mitad responde primero a la forma *B* y después a la forma A del test. Esta manera de aplicar el test elimina efectos que podrían enmascarar los resultados obtenidos tales como cansancio o enfado durante la aplicación de la segunda forma del test.

Este tipo de procedimiento posee las siguientes dificultades:

- En muchos casos, es difícil desarrollar una forma del instrumento, más difícil todavía desarrollar dos.
- Hay situaciones en que es difícil, si no imposible, preguntarse lo mismo dos veces, de forma equivalente o análoga.
- Si el procedimiento implica cierto intervalo de tiempo entre la aplicación del test y de su forma alternada, de la misma forma que ocurrió para la forma test-retest, puede surgir el problema de reagrupar el mismo grupo de sujetos para responder el test por segunda vez.
- Administrar consecutivamente dos formas del mismo test exige dos veces más tiempo y exige que las personas que responden mantengan el interés suficiente, ya que trabajarán, básicamente, sobre los mismos contenidos.

**3. Método de la mitad o par-impar**: un tercer procedimiento llamado *la mitad* o *par-impar* soluciona cada uno de los problemas destacados anteriormente, cuando se discutieron las otras formas de cálculo de fidedignidad. Implica sólo la administración de un instrumento en apenas una única forma. El instrumento es aplicado una única vez a un grupo de personas, sin embargo, es analizado de forma que separe los valores de cada persona en dos mitades. Estos dos conjuntos son entonces correlacionados. De ese modo, si un instrumento posee *100* ítems, para obtener la estimativa de fidedignidad en la forma *par-*

*impar*, se obtendrían los valores de cada persona en las preguntas impares y un valor separado para las otras *50* preguntas pares. Es posible, pero no lo más aconsejable, usar el procedimiento de separación en *mitades*, es decir, obtener un valor para cada persona basado en la primera mitad del test y otro basado en la segunda mitad del test.

La forma *par-impar* es la preferible debido a las siguientes ventajas:

- Normalmente, un instrumento de medida cubre diferentes áreas del conocimiento en diferentes secciones, las cuales generalmente son bien diferenciadas.
- Factores tales como fatiga o pérdida de interés podrían causar omisión por parte del entrevistado en las últimas preguntas del test.

Sin embargo, no importa cuál sea el procedimiento usado, estimativas de fidedignidad obtenidas a partir del uso de mitades de un instrumento también presentan sus problemas, principalmente relacionados con el hecho de que la fidedignidad está relacionada al número de ítems de un instrumento.

El siguiente procedimiento fue desarrollado para calcular la fidedignidad de un instrumento como un todo a partir del cálculo de la fidedignidad usando mitades del test. Es la llamada **fórmula de Spearman-Brown**, que recibe ese nombre en homenaje a los investigadores que, de forma independiente, la desarrollaron:

$$\alpha_{SB} = \frac{2\times|\alpha|}{1+|\alpha|}$$

donde $\alpha_{SB}$ es la llamada *estimativa de fidedignidad de Spearman-Brown*, $\alpha$ es la correlación entre las dos mitades del test. Lo que nos da esta fórmula es solamente una predicción o estimativa de la fidedignidad que el investigador podría esperar para el instrumento como un todo a partir de los valores de fidedignidad obtenidos para cada mitad del test.

La mayor ventaja de usar correlación entre mitades como una estimativa de fidedignidad del test total es de naturaleza práctica: se necesita sólo un test y una sesión para testarlo por parte de los entrevistados. Sus desventajas son las mismas.

Las expectativas para la fidedignidad de un instrumento diferirán dependiendo de la naturaleza de la información que se está buscando. Si lo que se está buscando es una información tipo demográfica, como, por ejemplo, lugar de nacimiento, escolaridad y experiencia profesional, podríamos esperar alta fidedignidad de un instrumento. En términos de correlaciones, esto implica coeficientes de correlación del orden o superiores a *0,90*. Por otro lado, si lo que se busca es sobre conocimiento y habilidades, los cuales son informaciones de carácter no tan fijo como el ejemplo anterior, la expectativa de fidedignidad será menor y un coeficiente de fidedignidad de *0,85* es aceptable. Por fin, si el tipo de información buscada es más movediza, como por ejemplo actitudes e intereses, un coeficiente del orden de *0,70* será aceptable.

Una pregunta que surge naturalmente en ese punto es la siguiente*: ¿cómo mejorar la fidedignidad de un test*? Como se comentó anteriormente, la fidedignidad de un test está directamente relacionada con su extensión. Por tanto, un procedimiento obvio para mejorar la fidedignidad de un test es alterar su extensión. *Pero ¿cuánto podría ser mejorada la fidedignidad añadiendo una o más preguntas al test*? La respuesta a esa pregunta se encuentra en la fórmula de cálculo del coeficiente de Spearman-Brown, la cual suministra la

fidedignidad máxima que se podría obtener con el aumento del tamaño del test. Esa estimativa es dada por la siguiente ecuación:

$$r_k = \frac{kr}{1+(k-1)r}$$

donde *k* es el número de veces que el nuevo test es mayor que el anterior, *r* es la fidedignidad presentada por la forma actual del test y $r_k$ es la estimativa de fidedignidad de la nueva forma del test.

*Ejemplo*: supongamos que la fidedignidad de un test es estimada en *0,50*. Este test tiene su extensión duplicada por el aumento de ítems. ¿Cuál será la estimativa de la nueva fidedignidad?

$$r_k = \frac{kr}{1+(k-1)r}$$
$$r_k = \frac{2 \times 0{,}5}{1+(2-1) \times 0{,}5}$$
$$r_k = 0{,}67$$

## Un algoritmo para el cálculo del coeficiente de fidedignidad de tests - Análisis de Consistencia Interna

Una de las aplicaciones del coeficiente de correlación es en el **Análisis de la Consistencia Interna** de tests. Es muy común en la práctica docente que el profesor sume los valores de ítems aislados de tests componiendo, así, un valor bruto, que es usado para análisis e inferencias. Sin embargo, ese procedimiento solamente es aceptable cuando todos los ítems del test se refieren a un mismo conjunto de conceptos y/o habilidades. El análisis de consistencia interna de un test tiene el objetivo de verificar cuánto existe de verdad en esa hipótesis a respecto de determinado test. La idea general es comparar el desempeño en cada ítem de los entrevistados con el desempeño de los entrevistados como un todo. Si el desempeño de los entrevistados en un ítem no se correlaciona con el desempeño de los entrevistados como un todo, esto significa que aquel ítem en particular no está evaluando las mismas características de las demás cuestiones que componen el test y, por tanto, debe ser desechado (o modificado). El Análisis de Consistencia Interna es parte indispensable del proceso de investigación. El investigador no puede usar un test sin verificar su consistencia interna. Sin esa etapa, la suma de los valores atribuidos a ítems particulares no puede ser realizada y toda inferencia obtenida a partir de ese valor total no tendrá significado.

Se presenta a continuación un guión[16] para la realización de análisis de consistencia interna. Los datos utilizados en el ejemplo son retirados de la Tabla 5:

[16] Este guión fue elaborado por el profesor Fernando Lang da Silveira del Instituto de Física de la UFRGS.

*Tabla 5. - Datos para el ejemplo de análisis de consistencia interna.*

| *ÍTEM →* <br> *INDIVIDUO* <br> *↓* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *Total* |
|---|---|---|---|---|---|---|---|---|---|
| *1.* | *5* | *5* | *5* | *5* | *5* | *5* | *5* | *2* | *37* |
| *2.* | *5* | *4* | *4* | *5* | *3* | *5* | *5* | *3* | *34* |
| *3.* | *5* | *5* | *4* | *5* | *3* | *4* | *5* | *3* | *34* |
| *4.* | *4* | *4* | *5* | *4* | *4* | *5* | *5* | *2* | *33* |
| *5.* | *5* | *4* | *5* | *5* | *2* | *4* | *4* | *3* | *32* |
| *6.* | *4* | *5* | *4* | *4* | *3* | *5* | *5* | *2* | *32* |
| *7.* | *4* | *4* | *5* | *5* | *5* | *5* | *3* | *1* | *32* |
| *8.* | *4* | *5* | *5* | *4* | *4* | *4* | *5* | *1* | *32* |
| *9.* | *3* | *4* | *5* | *4* | *3* | *5* | *5* | *2* | *31* |
| *10.* | *4* | *3* | *4* | *5* | *2* | *5* | *5* | *3* | *31* |
| *11.* | *5* | *5* | *5* | *4* | *1* | *4* | *3* | *4* | *31* |
| *12.* | *4* | *4* | *3* | *4* | *5* | *4* | *4* | *3* | *31* |
| *13.* | *4* | *4* | *4* | *3* | *3* | *3* | *4* | *4* | *29* |
| *14.* | *4* | *4* | *4* | *4* | *2* | *4* | *3* | *3* | *28* |
| *15.* | *3* | *4* | *4* | *4* | *3* | *4* | *4* | *2* | *28* |
| *16.* | *3* | *4* | *5* | *3* | *2* | *4* | *5* | *2* | *28* |
| *17.* | *4* | *3* | *3* | *5* | *4* | *3* | *2* | *4* | *28* |
| *18.* | *4* | *4* | *4* | *3* | *3* | *5* | *4* | *1* | *28* |
| *19.* | *3* | *3* | *4* | *4* | *4* | *4* | *3* | *2* | *27* |
| *20.* | *4* | *4* | *3* | *3* | *2* | *3* | *4* | *4* | *27* |
| *21.* | *4* | *3* | *3* | *3* | *4* | *3* | *4* | *3* | *27* |
| *22.* | *3* | *4* | *3* | *3* | *4* | *4* | *4* | *2* | *27* |
| *23.* | *3* | *4* | *3* | *3* | *1* | *3* | *4* | *5* | *26* |
| *24.* | *3* | *3* | *3* | *4* | *3* | *4* | *3* | *3* | *26* |
| *25.* | *2* | *3* | *3* | *3* | *5* | *4* | *3* | *1* | *24* |
| *26.* | *3* | *2* | *3* | *3* | *4* | *3* | *4* | *2* | *24* |
| *27.* | *1* | *2* | *3* | *2* | *5* | *3* | *3* | *5* | *24* |
| *28.* | *4* | *3* | *3* | *3* | *4* | *1* | *1* | *4* | *23* |
| *29.* | *2* | *2* | *2* | *1* | *4* | *3* | *3* | *4* | *21* |
| *30.* | *3* | *3* | *1* | *2* | *3* | *2* | *2* | *2* | *18* |
| *31.* | *1* | *1* | *2* | *1* | *3* | *1* | *2* | *4* | *15* |
| *F(1)* | *2* | *1* | *1* | *2* | *2* | *2* | *1* | *4* | |
| *F(2)* | *2* | *3* | *2* | *2* | *5* | *1* | *3* | *10* | |
| *F(3)* | *9* | *8* | *11* | *10* | *10* | *8* | *8* | *8* | |
| *F(5)* | *13* | *14* | *9* | *10* | *9* | *12* | *10* | *7* | |
| *F(6)* | *5* | *5* | *8* | *7* | *5* | *8* | *9* | *2* | |

**1**. Transformar la respuesta de cada individuo a cada ítem en un valor.
**2**. Calcular el valor total de cada individuo sumando los valores obtenidos en cada ítem por aquel individuo.
**3**. Ordenar los individuos, en *orden decreciente*, por el valor total.
**4**. Construir la matriz de los resultados (ver la *tabla 5*).

**5**. Determinar la frecuencia (número de veces que el valor aparece) de cada valor en cada ítem (ver las últimas cinco líneas de la *tabla 5)*.
**6**. Calcular la media del valor total. Esto se hace sumando los resultados totales de cada uno de los sujetos y dividiendo por el número de sujetos (*N*). Así, en el ejemplo, es la suma de la última columna de la tabla 5 y se divide por 31 (número de sujetos):

$$\overline{T} = \frac{37 + 34 + \ldots + 15}{31} = \frac{868}{31} = 28{,}00$$

**7**. Calcular la varianza del resultado total. En el ejemplo:

$$V_t = \frac{\sum T^2}{N} - (\overline{T})^2$$
$$V_t = \frac{24986}{31} - (28{,}00)^2$$
$$V_t = 22{,}00$$

**8**. Calcular la media y la varianza de cada ítem. Por ejemplo, para el *ítem 1* se tiene:

$I=(1x2)+(2x2)+(3x9)+(4x13)+(5x5)=110$ (suma de los resultados)
$I^2 =(1^2x2)+(2^2 x2)+(3^2 x9)+(4^2 x13)+(5^2 x5)=424$ (suma de los cuadrados de los resultados)
$\overline{I} =\sum I/N=110/31=3{,}548$ (media de los resultados atribuidos al ítem)
$V_I = 424/31-(3{,}548)^2=1{,}089$ (varianza en el ítem).

La *Tabla 6* trae los valores de la media y de la varianza para cada ítem de nuestro ejemplo:

*Tabla 6 - Cálculos para el primer sujeto de la tabla 5.*

| *ÍTEM* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|
| *MEDIA* | *3,548* | | | | | | | |
| *VARIANZA* | *1,089* | *0,946* | | | | | | *1,273* |

**9**. Calcular la totalidad de la varianza de los ítems. Para los datos del ejemplo (ver la *tabla 5*):

$V_I =1{,}089+0{,}946+\ldots+1{,}273=9{,}144$

**10.** Calcular el **coeficiente de fidedignidad** (**coeficiente α de Cronbach**):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_I}{V_T}\right)$$
$$\alpha = \frac{8}{7}\left(1 - \frac{9{,}144}{22{,}00}\right)$$
$$\alpha = 0{,}668$$

donde *k* es el número de ítems del test.

Como se dijo, el coeficiente de fidedignidad es un número entre *-1 y 1*. Cuanto más próximo de *1* (en módulo) mejor es el test. En el ejemplo, el coeficiente de fidedignidad

obtenido está bastante lejos de un valor aceptable. Para mejorarlo, es necesario descubrir qué ítems del test están llevando a ese valor bajo. Para eso, se debe calcular el *índice de discriminación* de cada ítem, descartando los ítems que posean bajo índice de discriminación. La lógica de ese procedimiento es la siguiente: si una cuestión del test tiene, aproximadamente, la misma frecuencia de valores en dos grupos con resultados totales diferentes (por ejemplo los 25% con resultados superiores y los 25% con resultados inferiores) entonces ése no es un buen ítem del test. Ésa es la próxima fase.

**11**. Dividir los sujetos en tres grupos (superior, medio e inferior) a través del resultado total. Los grupos superior e inferior deben sumar, aproximadamente, *25*% de los individuos.

**12**. Calcular la media de cada ítem en el grupo superior y en el grupo inferior.

En el ejemplo, media del ítem 1 en el grupo superior:

$$\overline{I_s} = (4x5+4x4)/8 = 4,5$$

Media del ítem 1 en el grupo inferior:

$$\overline{I_i} = (2x3+4+2x2+2x1)/7 = 2,286$$

La *Tabla 7* trae los valores calculados para cada uno de los ítems del ejemplo.

*Tabla 7 - Valores medios para cada ítem relativos a los grupos superior $G_s$ e inferior $G_i$ para los datos del ejemplo.*

| *1. ÍTEM* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|
| *2. MEDIA $G_S$* | *4,5* | | | | | | | |
| *3. MEDIA $G_I$* | *2,286* | | | | | | | |

**13**. Calcular el **índice de discriminación** de cada ítem. El índice de discriminación es definido como la diferencia entre los valores medios de los resultados del ítem en los grupos superior e inferior:

$$D_i = \overline{I_s} - \overline{I_i}$$
$$D_i = 4,5 - 2,286 = 2,214$$

en el ejemplo, para el ítem 1.

La *Tabla 8* presenta los valores del índice de discriminación para todos los ítems de nuestro ejemplo.

*Tabla 8 - Valores del índice de discriminación para cada ítem del ejemplo.*

| *ÍTEM* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|
| *$D_i$* | *2,214* | | | | | | | |

**14**. Calcular a continuación el **coeficiente de correlación ítem-total**. Ese coeficiente es el que indicará cuales son los ítems que están perjudicando la fidedignidad del test y que deben ser eliminados. El **coeficiente de correlación ítem-total** es definido por:

$$r_{IT} = \frac{D_i}{K \times \sqrt{V_I}}$$

La constante *K* que aparece en esta ecuación depende de la extensión de los grupos extremos; para los datos del ejemplo, su valor es de 2,57. Valores de esa constante para varias combinaciones de grupos se encuentran tabulados en libros de Estadística.

Para el *ítem 1* del ejemplo, el coeficiente de correlación ítem-total es dado por:

$$r_{IT}=2,214 / [2,57 \times \sqrt{(1,089)}]=0,826$$

Para los demás ítems, se organizaría una *tabla como la Tabla 9:*

*Tabla 9 - Valores del coeficiente de correlación ítem-total para los datos del ejemplo.*

| *ITEM* | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* |
|---|---|---|---|---|---|---|---|---|
| $r_{IT}$ | *0,826* | | | | | | | |

**15**. Verificar se existen ítems con coeficiente de correlación ítem-total próximo de cero o negativo. Si existen, deben ser eliminados. En el ejemplo, los ítems *5* deben ser eliminados.
**16**. Recalcular el resultado total de cada individuo eliminando los ítems deficientes. Hay que notar que ahora los ítems que hay que tener en cuenta son seis y no ocho. Así, por ejemplo, el resultado total del primer sujeto pasará a ser *30*, del segundo *28* y así sucesivamente.
**17**. Repetir las etapas de *6* a *9* con el objetivo de encontrar el nuevo coeficiente de fidedignidad. Para los datos del ejemplo, después de la eliminación de los ítems *5* y de la realización del nuevo cálculo, el coeficiente de fidedignidad pasará a ser *0,904.*

En la práctica, ese procedimiento se hace en segundos, con un programa de ordenador apropiado. El ejemplo detallado, paso a paso, fue dado con la intención de facilitar la comprensión que significa "calcular el coeficiente de fidedignidad".

## *Validez*[17]

Mientras la fidedignidad es prerrequisito básico para cualquier instrumento de investigación, la validez es la característica más importante que debe poseer. Pues la validez se refiere a la relación entre los datos obtenidos y el propósito para el cual ellos fueron recogidos. De este modo, validez es definida como el grado con que el procedimiento realmente mide lo que se propone a medir.

La fidedignidad es una etapa que precede a la etapa de la validez, siendo esencial para ésta, y pone un límite superior a la validez del instrumento. Así, un instrumento con fidedignidad *0,00* no puede tener validez; en el otro extremo, un instrumento con validez *1,00* puede *posiblemente* ser perfectamente válido. Para los grados intermedios de fidedignidad, la validez máxima puede ser estimada a partir de la raíz cuadrada del coeficiente de fidedignidad. Sin embargo, la fidedignidad pone límites a la validez, pero no la garantiza. Por

[17] Adaptado de Fox, 1969.

ejemplo, puede acontecer que un instrumento con una fidedignidad de *0,60* tenga una validez mucho menor que $0{,}77\left(\sqrt{0{,}60} = 0{,}77\right)$ y, de hecho, puede que no tenga ninguna validez.

Por tanto, la fidedignidad es prerrequisito para que un instrumento sea válido, garantizando que mida de forma correcta, pero esto no es garantía de que mida lo que tiene que medir. La validez debe ser estimada separadamente una vez que la fidedignidad del instrumento ya haya sido establecida y que los valores encontrados sean satisfactorios.

Son varios los tipos de análisis que pueden ser realizados para establecer la validez del instrumento. Sin embargo, no hay, como para la fidedignidad, una forma matemática de establecer la validez de un instrumento. A continuación se presentan los varios tipos de análisis posibles que llevan al establecimiento de la validez de un instrumento:

1. **Validez de apariencia**: este tipo de validez es establecido a partir del análisis superficial de la naturaleza del instrumento, es decir, por la presentación del instrumento. Obviamente, éste es el tipo de análisis de validez menos consistente.
2. **Validez de contenido**: más apropiada que la anterior, esta técnica de análisis de validez es, muchas veces, la forma más adecuada disponible para el investigador para analizar ciertos tipos de instrumento, tales como cuestionarios y entrevistas. Sin embargo, es una técnica que depende del juicio de quien hace el análisis de validez que, sin duda, es su punto flaco. Esta técnica verifica si el instrumento está midiendo lo que se propone a medir a través del análisis, realizado por especialistas en el contenido del instrumento, de la existencia de razones racionales para la elección de los ítems del instrumento o de una base, lógica o empírica, para esta elección.
3. **Validez de constructo**: esta técnica es definida como la habilidad del instrumento de distinguir grupos que se sabe previamente que se portan de forma diferente en la variable o constructo en estudio. A nivel de procedimiento, para determinar la validez de constructo hay dos estadios. El primero consiste en la definición de un *criterio* para identificar los grupos que difieren en el constructo que el nuevo instrumento se propone a medir. El segundo estadio consiste en administrar el instrumento a estos grupos y determinar si difieren significativamente también en el nuevo instrumento. Si así es, entonces hay argumentos para defender la validez del nuevo instrumento.
4. **Validez convergente o discriminante**: el investigador que ofrece validez discriminante o convergente presenta datos de naturaleza correlacional, mostrando que el desempeño en su nuevo instrumento se correlaciona con el desempeño de algún instrumento de medida de la variable ya existente y aceptado como válido. Si el criterio que está siendo empleado por el nuevo instrumento es de la misma familia que el instrumento antiguo, se habla de validez convergente. Si, por otro lado, el criterio empleado es de naturaleza diferente entre los dos instrumentos se habla de validez discriminante.
   La esencia de la validez discriminante y convergente está en la relevancia y validez del criterio.
5. **Validez predictiva**: existe validez predictiva cuando el investigador puede prever comportamientos de los entrevistados, en el área de interés de la investigación, a partir de los datos obtenidos por el instrumento. Este tipo de proceso de validación implica que el investigador debe esperar algún tiempo para saber si las predicciones hechas se concretaron o no y en qué extensión. Estos datos se pueden presentar de varias formas, entre ellas la correlación entre los resultados previstos y el resultado real, porcentajes de predicciones correctas, etc.

## Análisis de la Varianza

Como ya se vio en las secciones sobre Estadística, la tendencia central (la media, por ejemplo) no es suficiente para distinguir distribuciones de frecuencia. La media aritmética de dos muestras, por ejemplo, puede ser prácticamente la misma, pero la variabilidad de los resultados puede ser muy diferente.

## El test F - análisis de la varianza (anova)

Cuando se quiere determinar si los resultados en una muestra son más variables que los resultados en otra muestra, se puede usar la técnica llamada *test F.* Usando el *test F,* podemos determinar si la variabilidad en un conjunto de datos es significativamente mayor que la variabilidad en otro conjunto de datos. Al conducir un *test* F, se usa una medida de la variabilidad *llamada varianza,* en lugar del desvío estándar. **La varianza, en términos simples, es el cuadrado del desvío estándar**.

Para ejecutar un *test F* entre dos varianzas, simplemente se divide la mayor varianza por la menor. Esto da lo que se llama *razón F* entre las dos varianzas. La cuestión es si la varianza obtenida de una muestra difiere significativamente de la varianza obtenida en otra muestra. En este caso, la hipótesis nula es que no existe diferencia entre la variabilidad de los resultados en una muestra con relación a la variabilidad de los resultados de la otra muestra. Se usa entonces una tabla de valores *F*, existente en libros de Estadística, para determinar si se rechaza o no la hipótesis nula, en el nivel de significancia escogido.

El *test F* también puede ser usado para analizar la variabilidad entre medias de resultados de tres o más muestras cuando se puede asumir que las muestras se obtuvieron a través de selección aleatoria y a partir de una población distribuida normalmente[18]. El *test F* usado para comparar varias medias de resultados es llamado **Análisis de la Varianza (ANOVA)** y consiste en la comparación de dos varianzas estimadas.

Lo que se pretende es comparar una estimativa de la varianza de la población obtenida a partir de los resultados dentro de cada muestra con una estimativa de la varianza obtenida de los resultados medios de las varias muestras. Una de las varianzas estimadas se obtiene comparando la varianza estimada para cada una de las muestras separadamente y después se combinan para obtener una estimativa única llamada **estimativa de varianza dentro de los grupos**.

La otra varianza estimada se calcula a partir de los resultados medios para cada una de las muestras y se calcula la varianza estimada usando estos resultados medios y el tamaño de la muestra en el cálculo. Esta varianza es llamada **estimativa de varianza entre los grupos**.

El objetivo es determinar si la estimativa de varianza entre grupos es significativamente mayor que la estimativa de varianza dentro de los grupos. Si la estimativa de varianza entre los grupos es significativamente mayor que la estimativa de varianza dentro de los grupos, podremos rechazar la hipótesis nula y decir que las muestras no provienen de la misma población. Para aplicar el análisis de la varianza, se calcula una *razón F* entre las dos varianzas estimadas: usando la varianza entre grupos estimada como numerador y la varianza

[18] De hecho, podríamos usar el *test F* para comparar dos muestras, pero en ese caso los resultados serían idénticos a los del *test t*.

dentro de los grupos como denominador. A partir de ahí, usando una tabla de valores *F*, podemos determinar, para cualquier tamaño de las muestras, la *razón F* necesaria para rechazar la hipótesis nula, en el nivel de significancia especificado.

*Ejemplo:* supongamos que queremos determinar si los niveles de iluminación afectan la productividad en el trabajo en una empresa de productos electrónicos. Para estudiar esto se seleccionan, aleatoriamente, cuatro muestras de cuarenta empleados cada una y se distribuyen en diferentes niveles de iluminación. Se mide entonces, la productividad de cada grupo y se obtienen los datos de la *Tabla 10.*

Es evidente que la productividad media de las cuatro muestras es diferente. Pero, es necesario saber si la variabilidad entre las medias muestrales (es decir, la diferencia entre las medias) ocurrió como resultado de un error de muestra o si esa variabilidad puede ser atribuida a la cantidad de iluminación.

*Tabla 10 - Datos para el ejemplo 1.*

| *Nivel* | *Valor medio* |
|---|---|
| *I* | *40* |
| *II* | *38* |
| *III* | *27* |
| *IV* | *26* |

En este ejemplo, la hipótesis nula que será testada es que *no hay diferencia en la productividad de los empleados como resultado de los diferentes niveles de iluminación*. Para determinar esto, o sea, si se puede o no despreciar la hipótesis nula, se examina este conjunto de datos usando el análisis de la varianza. La interpretación de los resultados del análisis de la varianza se hace del mismo modo que la del test ***T.*** Supongamos que la *razón F* en el ejemplo dado sea significativa al nivel *0,01*. Entonces, se podría rechazar la hipótesis nula y concluir que el nivel de iluminación está relacionado a la productividad.

La inspección de las medias de las muestras dice que la productividad entre los niveles *I* e *II* difiere solamente en dos puntos. Del mismo modo, la diferencia entre los niveles *III* y *IV* es solamente de un punto. La diferencia mayor aparece entre los niveles *II* e *III*. El análisis de la varianza solamente nos dice que existe una diferencia general entre las cuatro medias, sin embargo, no informa cuál de los grupos es el responsable de la diferencia significativa. Como se verá más adelante, el análisis de cuál es la causa de la diferencia observada se hace a través del **Análisis Factorial de la Varianza**. La técnica de análisis de varianzas puede ser usada para el análisis de diferencias entre cualquier número de muestras y es aplicable también para el análisis de diferencias entre grupos dentro de muestras, tales como masculino - femenino o agrupaciones por edad. En el ejemplo, las muestras se podrían dividir en grupos de empleadas y grupos de empleados y, además, sería posible agruparlos también por edad. Naturalmente, en este caso serían necesarios mucho más empleados en la muestra. Usando el análisis de varianzas se podrían haber examinado diferencias de productividad entre sexos, de acuerdo con la edad de los empleados y entre diferentes niveles de iluminación. Esto se podría llamar análisis de varianzas trilateral, ya que se podría analizar la productividad en función de la edad, del sexo y de la cantidad de iluminación.

## El método del análisis de la varianza

Para describir el método de cálculo de la *razón F,* será utilizado otro ejemplo. Supongamos que un investigador educacional esté interesado en la eficacia relativa de dos métodos de enseñanza, $A_1$ y $A_2$. Después de seleccionar diez estudiantes, el investigador los divide en dos grupos, aleatoriamente, uno experimental y otro de control. Después de algún tiempo, mide el aprendizaje de los sujetos de los dos grupos, usando algún tipo de test. Los resultados se encuentran en la *Tabla 11*:

*Tabla 11 - Dos conjuntos de datos experimentales para un ejemplo hipotético.*

| | $A_1$ | $x$ | $x^2$ | $A_2$ | $x$ | $x^2$ | |
|---|---|---|---|---|---|---|---|
| | 4 | 0 | 0 | 3 | 0 | 0 | |
| | 5 | 1 | 1 | 1 | -2 | 4 | |
| | 3 | -1 | 1 | 5 | 2 | 4 | |
| | 2 | -2 | 4 | 2 | -1 | 1 | |
| | 6 | 2 | 4 | 4 | 1 | 1 | |
| $\sum X$ | 20 | | | 15 | | | $\sum X_t=35$ |
| $\sum x^2$ | | | 10 | | | 10 | |
| $M$ | 4 | | | 3 | | | $M_t=3,5$ |

El trabajo que se ha de realizar con esos datos es localizar y computar las diferentes varianzas que componen la varianza total. Para calcular la *varianza total,* se usa la fórmula:

$$V_t = \frac{\sum x^2}{N-1}$$

donde $\sum x^2$ es la suma de cuadrados (*sq*), $x=X-\overline{X}$ es el desvío de los resultados medios y $N$ es el número de casos en la muestra como un todo. De este modo, reagrupando los datos de la *tabla 6.2*, sin preocuparse del grupo al que pertenece el sujeto, se obtiene para la varianza total el valor de $V_t$=2,5 (ver la *Tabla 12*).

*Tabla 12 - Cálculos de $V_t$ para los datos del ejemplo.*

| | $X$ | $x$ | $x^2$ |
|---|---|---|---|
| | 4 | 0,5 | 0,25 |
| | 5 | 1,5 | 2,25 |
| | 3 | -0,5 | 0,25 |
| | 2 | -1,5 | 2,25 |
| | 6 | 2,5 | 6,25 |
| | 3 | -0,5 | 0,25 |
| | 1 | -2,5 | 6,25 |
| | 5 | 1,5 | 2,25 |
| | 2 | -1,5 | 2,25 |
| | 4 | 0,5 | 0,25 |
| $\sum X$ | 35 | | |
| $M$ | 3,5 | | |
| $\sum x^2$ | | | 22,5 |

Hay que observar que existe también una varianza entre los grupos, la cual, presumiblemente, se debe a las manipulaciones experimentales. Es decir, el experimentador hizo algo para un grupo y algo diferente para el otro. Estos tratamientos diferentes podrían ocasionar diferencias entre los grupos, expresadas por las medias diferentes. De este modo, si los grupos son diferentes habrá una *varianza entre los grupos*:

$$V_b = \frac{\sum x_b^2}{k-1}$$

donde $k$ es el número de grupos y $x_b$ es la diferencia entre la media del grupo y la media entre los grupos. El cálculo de la varianza entre los grupos, $V_b$ se encuentra en la *Tabla 13*.

*Tabla 13 - Cálculos de $V_b$ para los datos del ejemplo 1.*

| | $X$ | $x$ | $x^2$ |
|---|---|---|---|
| | 4 | 0,5 | 0,25 |
| | 3 | 0,5 | 0,25 |
| $\sum X$ | 7 | | |
| $M$ | 3,5 | | |
| $\sum x^2$ | | | 0,50 |

Existe también una fuente de varianza que se debe al error de muestreo. Se tiene en cuenta ese tipo de error cuando se calcula la *varianza dentro de cada grupo* separadamente y entonces se toma la media de esos valores, obteniendo $V_W$. Para el ejemplo:

$$\frac{\sum x_{A1}}{n_{A!}-1} = 10/4 = 2{,}5$$

$$\frac{\sum x_{A2}}{n_{A2}-1} = 10/4 = 2{,}5$$

La media de esos valores da $V_w = 2{,}5$.

Antes se dijo que la varianza total está compuesta por fuentes separadas de varianza: la varianza entre grupos ($V_b$) y la varianza dentro de los grupos ($V_w$). Lógicamente, esas varianzas deberían sumarse de modo a suministrar la varianza total ($V_t$). La ecuación teórica sería entonces: $V_t = V_b + V_w$. Como se puede ver en los datos anteriores, no es eso lo que pasa. La razón es que en los cálculos realizados se usaron los grados de libertad como denominadores, en lugar de $N$, $n$ y $K$. El cálculo de las varianzas usando $N$, $n$ y $k$ es matemáticamente correcto, pero estadísticamente incorrecto. Otro aspecto importante del análisis es la estimativa de los valores para la población. Se puede mostrar que usando los grados de libertad en los denominadores, la fórmula de la varianza dará estimativas imparciales de los valores de la población, un tópico de gran preocupación estadística. Por tanto, hay que recordar que varianzas, como se usan en el análisis de varianzas, no son aditivas. Por otro lado, sumas de cuadrados son siempre aditivas. Las sumas de cuadrados también son, desde luego, una medida de la variabilidad. Excepto en la fase final del análisis, las sumas de cuadrados (*sq*) son calculadas, estudiadas y analizadas.

Comparando el análisis de varianza con el test **t**, se puede decir que la forma de encarar el problema es conceptualmente análoga, con diferencias en la metodología. El método es general: mientras que diferencias de más de dos grupos pueden ser testadas en términos de significancia estadística usando el análisis de varianza, el test **t** se aplica solamente a dos grupos. El análisis de varianzas para dos grupos suministra los mismos resultados que el test **T.**

La razón formada por la división de la varianza entre grupos ($V_b$) por la varianza dentro de los grupos ($V_w$) es llamada de razón *F*:

$$F = \frac{V_b}{V_w}$$

Los valores *F* de los datos experimentales son calculados y comparados con una tabla de valores *F*. Si los valores obtenidos son mayores o mucho mayores que los valores tabulados, en aquel nivel de significancia estadística y grados de libertad, las diferencias expresadas por $V_b$ reflejan diferencias significativas. En este caso, la hipótesis nula, de que no hay diferencia alguna entre las medias es rechazada en aquel nivel de significancia.

En la práctica podemos calcular la *razón F* a partir de los resultados brutos. En el ejemplo precedente, se usaron varianzas estándar para mostrar los aspectos fundamentales del método. Naturalmente, todo ese cálculo se puede realizar rápidamente con un aplicativo apropiado.

## Análisis factorial de la varianza

En el análisis factorial de la varianza, dos o más variables varían independientemente o interactúan una con a otra para producir variaciones en la variable dependiente. El análisis factorial de la varianza es el método estadístico que analiza los efectos independientes e interactivos de dos o más variables independientes en una variable dependiente.

En el pasado, muchos investigadores (y esto también es válido para muchos investigadores en los días de hoy) creían que el método de investigación más efectivo era permitir que una variable independiente evolucionase, mientras las demás variables independientes eran controladas al máximo. El análisis factorial de la varianza cambió ese cuadro, permitiendo que se pueda analizar el efecto de varias variables independientes al mismo tiempo.

El análisis factorial sirve para varios propósitos. Primero, el diseño factorial y el análisis factorial de la varianza le permiten al investigador manipular y controlar dos o más variables simultáneamente[19]. Por ejemplo, podemos no solamente estudiar los efectos de determinado método de enseñanza sobre el aprendizaje, sino también analizar los efectos, digamos, de tipos de refuerzo sobre las respuestas. Además podemos controlar variables como sexo, inteligencia y clase social.

[19] Aunque sea posible controlar más de tres variables, estos diseños son poco prácticos debido a la dificultad de obtener un número suficiente de sujetos de modo que ocupe todas las células. La forma más simple de un análisis factorial de varianza es $2 \times 2 \times 2$.

Variables que no son manipuladas pueden ser controladas. En lugar del procedimiento diseminado de "parear" los sujetos para tests sobre inteligencia o actitudes, se pueden construir estas variables (y muchas otras) en diseños con un carácter factorial. No solamente se controlan estas variables, sino también se obtiene información adicional de gran valor y significancia.

## *Interacción y Estadística*

En cuanto a la interacción estadística, la hipótesis nula es que **no hay interacción entre las variables independientes**, o sea, que no existe influencia de la combinación de variables. Esta hipótesis se podría llamar *hipótesis de la diferencia constante o hipótesis de los resultados relativos constantes*. Lo que se quiere decir con esto es que las diferencias entre celdas de líneas diferentes permanecerán constantes de columna para columna; equivalentemente se podría decir que las diferencias entre celdas de columnas diferentes permanecerán constantes de línea para línea.

Como un ejemplo, se considere la *Tabla 14*, una hipotética tabla de medias, de 12 celdas.

*Tabla 14 - Datos de un ejemplo donde **no hay** interacción.*
*Columnas Líneas*

| | | *Columnas* | | | |
|---|---|---|---|---|---|
| *L* | | ***1*** | ***2*** | ***3*** | ***4*** |
| *i* | ***1*** | *3* | *6* | *5* | *7* |
| *n* | ***2*** | *2* | *5* | *4* | *6* |
| *e* | ***3*** | *5* | *8* | *7* | *9* |
| *a* | | | | | |
| *s* | | | | | |

Éste es un ejemplo típico de falta de interacción. Hay que observar que la diferencia de las medias en celdas en la segunda línea con relación a la primera línea es constante e igual a *1,* mientras que la diferencia entre los valores en la primera línea y los de la tercera también es constante e igual a *2*. Siendo constante la diferencia entre líneas, las diferencias entre columnas también lo serán. La *Figura 5* muestra gráficamente lo que significa que no hay interacción entre las variables.

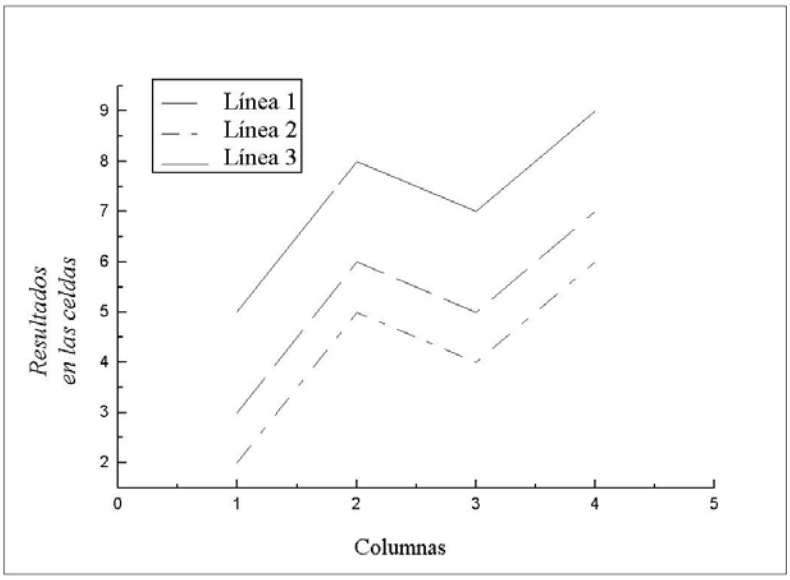



*Figura 5 - Un ejemplo donde no hay interacción.*

Como ya se comentó anteriormente, gráficamente la interacción aparece como líneas paralelas en un gráfico donde se representa los valores medios de las celdas en cada línea. Para cada línea de la tabla, se marcan en el gráfico los valores medios y se relacionan por una línea. Si no hay interacción (*hipótesis nula*), entonces las líneas obtenidas para cada línea de la tabla no se cruzarán, o más precisamente, serán paralelas. En el caso de que haya algún tipo de interacción, entonces las líneas ya no serán paralelas, no habiendo necesidad de que se crucen. Un ejemplo de interacción (ficticio) es la *Figura 6, que* es una representación gráfica de los datos de la Tabla 15:

*Tabla 15 - Datos de un ejemplo donde **existe** interacción.*

| | *Columnas* | | | |
|---|---|---|---|---|
| | ***1*** | ***2*** | ***3*** | ***4*** |
| *Líneas* | ***1*** | *4* | *5* | *7* | *5* |
| | ***2*** | *3* | *1* | *4* | *4* |

Como un último comentario de esta sección, conviene discutir lo que se entiende por análisis de varianza unilateral y lo que se entiende por análisis de varianza bilateral. No es el **número de variables** que se está utilizando lo que define el tipo de análisis de varianza, sino el **número de clases** de variables utilizadas. Así, por ejemplo, en un determinado experimento se analiza la influencia de tres métodos de enseñanza diferentes. Entonces, en ese caso, se hace un análisis de varianzas unilateral, ya que las variables analizadas (**métodos de enseñanza**) pertenecen a una única clase (**métodos**). Por otro lado, si además de métodos de enseñanza se están analizando diferentes tipos de motivación, el caso será de un análisis de varianzas bilateral, pues, en ese caso, habrá **dos** clases: **métodos de enseñanza y tipos de motivación**. Ese ejemplo será discutido en la próxima sección donde se discutirá el método propiamente dicho.

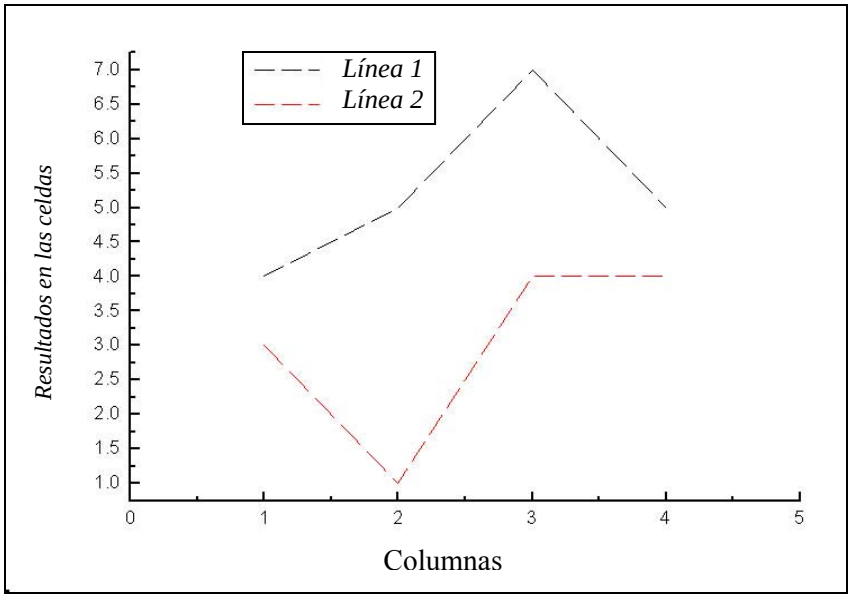



*Figura 6 - Un ejemplo donde hay interacción.*

## El método del análisis factorial de la varianza

Con el fin de estudiar el método usado en el análisis factorial de la varianza, se presenta a continuación un ejemplo hipotético[20].

Un investigador está interesado en el estudio de la influencia de dos factores en el aprendizaje. El primero de esos factores es el método de enseñanza, que puede ser uno de los dos métodos que serán estudiados, y que serán expresados por $A_1$ y $A_2$ y el segundo, factores de motivación, los cuales podrán ser uno de dos posibles, y serán expresados por $B_1$ y $B_2$. De ese modo, los sujetos son distribuidos en celdas para el estudio donde interactúan un método de enseñanza y un factor de motivación como, por ejemplo $A_1$ $B_2$ que nos indica que los sujetos serán sometidos al método de enseñanza $A_1$ y al factor de motivación $B_2$. Las posibilidades están dispuestas en la *Tabla 16.* Nuestra muestra hipotética está compuesta por ocho sujetos, distribuidos en la forma de dos sujetos por celda.

*Tabla 16 - Diseño factorial para variables* ***método de enseñanza y motivación.***

| | ***Métodos*** | | |
|---|---|---|---|
| | | $A_1$ | $A_2$ |
| ***Motivación*** | $B_1$ | $A_1$ $B_1$ | $A_2$ $B_1$ |
| | $B_2$ | $A_1$ $B_2$ | $A_2$ $B_2$ |

[20] Extraído de Kerlinger, 1964.

## *Tests estadísticos paramétricos y no paramétricos*

Un tópico central en la moderna teoría estadística es la *Estadística Inferencial*. La estadística inferencial está preocupada en resolver dos tipos de problemas: la estimativa de los parámetros de la población y tests de hipótesis. En la inferencia estadística, la preocupación es como sacar conclusiones sobre un gran número de eventos con base en observaciones de una porción de ellos. La Estadística da herramientas con las cuales se formalizan y se estandarizan los procedimientos para que podamos tomar decisiones.

Un problema común en estadística inferencial es determinar, en términos de probabilidades, si las diferencias observadas entre dos muestras significan que las poblaciones a partir de las cuales se retiraron las muestras son realmente diferentes. Las diferencias pueden ocurrir solamente debido al azar durante el proceso de muestreo.

En el desarrollo de los métodos estadísticos, las primeras técnicas estadísticas de inferencia que aparecieron fueron las que hacían muchas hipótesis sobre la naturaleza de la población de la cual eran retirados los valores. Ya que los valores de la población son parámetros, estas técnicas estadísticas fueron llamadas *paramétricas*[21]. Por ejemplo, la técnica de inferencia se puede basar en la hipótesis de que los valores fueron retirados de una población cuyos valores siguen la distribución normal. O la técnica se puede basar en la hipótesis de que los conjuntos de valores fueron retirados de poblaciones que tienen la misma varianza o el mismo distanciamiento de los valores.

Más tarde surgió un gran número de técnicas estadísticas de inferencia que no hacen hipótesis demasiado numerosas o restrictivas sobre los parámetros de la población. Esas técnicas, que son *independientes de distribución* o *no paramétricas*, llevan a conclusiones que presentan pocas limitaciones. Algunas técnicas no paramétricas son llamadas *testes de ordenamiento*. Este nombre tiene origen en el hecho de que mientras las técnicas paramétricas se centran en las diferencias de medias y varianzas, las técnicas no paramétricas se centran en el *ordenamiento* de los valores y no en sus valores numéricos.

Mientras un parámetro es un valor de una determinada población, una estadística es una medida calculada de una muestra. *Un test estadístico no paramétrico es un test cuyo modelo no especifica condiciones sobre los parámetros de la población de la cual fueron retiradas las muestras*.

Cuando se hace alguna afirmativa a respecto de la naturaleza de la población y sobre el proceso de muestreo, se está estableciendo un *modelo estadístico*. Asociado con *todo* test estadístico existe un modelo y una prescripción de medida; el test estadístico en cuestión es válido sobre ciertas condiciones, y el modelo y la prescripción de medida especifican estas condiciones. Algunas veces, es posible testar si las condiciones de un modelo estadístico particular se encuentran presentes pero, muchas veces, se tiene que tomar por hipótesis la presencia de esas condiciones. De este modo, las condiciones de validez de un test, el modelo estadístico (o sea, las hipótesis hechas en el tiempo de la construcción del test), son muchas veces llamadas de *hipótesis* del test.

Es obvio que cuanto menos o más débiles son las hipótesis por detrás del modelo subyacente a un test estadístico, menos restricciones hay que hacer sobre las conclusiones

[21] La media, el desvío estándar y la varianza de una población, o cualquier otra medida de la población, son *parámetros*.

obtenidas por el test estadístico asociado al modelo. O sea, cuanto menor o menos restringentes son las hipótesis por detrás del modelo, más *generales* las conclusiones y/o resultados obtenidos.

Los tests más poderosos son justamente los que tienen las hipótesis más fuertes o condiciones más restrictivas. Los tests paramétricos, por ejemplo, el test **t** o el *test F*, tienen una variedad de hipótesis fuertes fundamentando su uso. Cuando aquellas hipótesis son válidas, estos tests son los que más probablemente rechazarán la hipótesis $H_0$ cuando esta hipótesis sea falsa.

Las condiciones que deben ser satisfechas para hacer el test **t,** el test más poderoso, necesarias para poder tener confianza en cualquier inferencia hecha a partir de resultados obtenidos con el test, son:

1. **Las observaciones deben ser independientes** - con esto se quiere decir que la selección de cualquier caso de la población para inclusión en la muestra no debe influenciar la probabilidad de inclusión de cualquier otro caso. De esta forma, el valor atribuido a un caso no debe influenciar el valor atribuido a otro caso.
2. **Las observaciones deben ser retiradas de una población normal** - Como se vio anteriormente, una de las hipótesis del test **t** es que la población sigue la distribución normal.
3. **Las varianzas de las poblaciones de donde son retiradas las muestras deben ser iguales**.
4. **Las escalas deben ser intervalares o racionales** - eso es necesario para que se puedan ejecutar operaciones de carácter aritmético sobre los números representativos de los resultados.
5. **Aditividad (condición para validez del test F) -** las medias de las distribuciones normales deben ser combinaciones lineales de los efectos debidos a columnas y/o líneas. Es decir, los efectos deben ser aditivos.

De este modo, un test estadístico paramétrico es un test cuyo modelo especifica ciertas condiciones sobre los parámetros de la población de la cual fue retirada la muestra. Ya que esas condiciones normalmente no son testadas, son asumidas como verdaderas. A significancia y validez de un test paramétrico depende del grado de acierto al suponer correctas estas hipótesis.

Un test estadístico no paramétrico es un test cuyo modelo no especifica condiciones sobre los parámetros de la población de la cual fueron retiradas las muestras.

Varios criterios podrían ser considerados en la elección de un test estadístico para uso cuando está en curso un proceso de toma de decisión sobre la hipótesis de investigación. Estos criterios son:

1. **El poder del test: Poder = 1 - probabilidad de un error tipo II.**
   Es decir, la probabilidad de aceptar $H_0$ cuando de hecho es falsa.
2. **La aplicabilidad a los datos de la investigación del modelo estadístico en el que se basa el test**.
3. **Poder-eficiencia** - El concepto de poder - eficiencia es relativo a la cantidad de aumento en el tamaño de la muestra, la cual es necesaria para hacer un test *B* tan poderoso como un test *A.* El poder-eficiencia del test *B* con relación al test *A* es definido por:

$$P_B = 100 \times \frac{N_A}{N_B}$$

donde $N_A$ y $N_B$ son, respectivamente, los tamaños de las muestras sometidas a los tests *A* y *B*.

**4. El nivel de la medida obtenido en la pesquisa**.

## *Ventajas de tests estadísticos no paramétricos*

1. Afirmaciones probabilísticas obtenidas a partir de tests estadísticos no paramétricos son **probabilidades exactas**.
2. Si los tamaños de las muestras son pequeños $(N \cong 6)$ no existe alternativa al uso de tests estadísticos no paramétricos a menos que se conozcan **exactamente** las características de la distribución seguida por la población.
3. Existen tests estadísticos apropiados para tratamiento de muestras que provienen de varias poblaciones diferentes. Ninguno de los tests paramétricos puede manipular datos de este tipo sin exigir que se crea en hipótesis irreales.
4. Hay tests estadísticos no paramétricos para tratar datos que son inherentemente *ordinales,* es decir, el investigador es capaz solamente de decir que un sujeto presenta más o menos determinada característica, pero no sabe decir *cuánto* a más o a menos con relación a otro sujeto.
5. Los tests no paramétricos son capaces de tratar datos que son simplemente clasificatorios, siguiendo una escala tipo *nominal*. Ninguno de los tests paramétricos es capaz de manejar datos con esas características.
6. Por fin, los tests no paramétricos son más fáciles de aprender y aplicar que los tests paramétricos.

Sin embargo, si todas las hipótesis del modelo estadístico paramétrico se encuentran de hecho en los datos y si las medidas son del tipo exigido por el test, entonces tests estadísticos no paramétricos son innecesarios. Una medida de la necesidad del uso de un test no paramétrico en una situación de ese tipo es dada por el *poder-eficiencia* del test no paramétrico. Supongamos que el poder-eficiencia del test no paramétrico sea de *90*%. Esto significa que un test no paramétrico en la misma situación exigiría solamente *10* % de la muestra para ser tan efectivo como el test paramétrico.

## *Tests no paramétricos - el caso de una muestra*

Esta situación tiene lugar cuando, aleatoriamente, se retira una muestra de determinada población y se testa la hipótesis de que esa muestra viene de una población con una distribución específica.

Una técnica paramétrica común en el caso de una muestra es usar el test **t** para la diferencia entre las medias observada (la de la muestra) y esperada (la de la población). Sin embargo, existen muchos tipos de datos para los cuales no se puede aplicar el test **t**. Esos factores de inaplicabilidad del test **t** pueden tener varias fuentes, entre las cuales cabe destacar:

1. Las hipótesis y exigencias para la aplicabilidad del test no son satisfechas para los datos del experimento particular.

2. Frente a determinada situación, puede ser preferible evitar las hipótesis sobre las cuales el test **t** es construido y ganar en generalidad de las afirmaciones resultantes de la investigación.
3. Los datos de la investigación son, inherentemente, de los tipos nominal u ordinal y, por tanto, no son pasibles de análisis por el test **t.**
4. El investigador no está de hecho interesado solamente en diferencias en la localización, sino en cualquier tipo de diferencia no importando su origen.

En este caso el experimentador puede escoger usar uno de los siguientes tests estadísticos no paramétricos:

1. **El test binomial**: la distribución binomial es la distribución de muestreo de proporciones que podemos observar en muestras retiradas de una población compuesta por dos clases. Es decir, suministra los varios valores que pueden ocurrir bajo $H_0$. Por tanto, cuando los valores de la investigación están en dos clases, la distribución binomial puede ser utilizada para testar $H_0$.
2. **El test $\chi^2$ para una muestra**: este test es utilizable cuando queremos analizar datos que caen en más de dos categorías. Este test es útil para decidir si es significativa la diferencia observada entre el número de objetos que recaen en cada categoría y el número esperado con base en la hipótesis nula.
3. **El test de Kolmogorov-Smirnov para una muestra**: este test está preocupado con el grado de concordancia entre la distribución de un conjunto de valores observados y alguna previsión teórica específica. Determina si los valores en la muestra pueden razonablemente ser pensados como procedentes de una población que obedece a la distribución prevista teóricamente.
4. **Test del período para una muestra**: en este caso, se está interesado en saber si los datos bajo análisis provienen de una distribución aleatoria. Está basado en la secuencia en la que los valores aparecen originalmente, es decir, el número de períodos que presenta la muestra.

## *Tests no paramétricos - el caso de dos muestras relacionadas*

Tests estadísticos de dos muestras son usados cuando el investigador desea establecer si dos tratamientos son diferentes o si un tratamiento es '*mejor*' que otro. En este tipo de comparación, algunas veces se observan diferencias significativas, que no son consecuencia del tratamiento. Una manera de superar la dificultad impuesta por diferencias extrañas entre los grupos es usar dos muestras relacionadas en el experimento. Es decir, podemos emparejar o relacionar de alguna otra manera las dos muestras estudiadas. Este emparejamiento puede ser alcanzado usando cada sujeto como su propio control, o por el emparejamiento de sujetos y distribuyendo los miembros de la pareja aleatoriamente a los grupos de control y experimental. Cuando el sujeto sirve como su propio control, es expuesto a los dos tratamientos en tiempos diferentes. Cuando se usa el método del emparejamiento, el esfuerzo es para seleccionar para cada pareja de sujetos individuos que sean lo más iguales posible en la(s) variable(s) extraña(s) que puede(n) influir en el experimento.

Cuando sea posible, el método de usar cada sujeto como su propio control es preferible al método de emparejamiento. La razón para esto es que es difícil emparejar personas por desconocimiento de variables relevantes, que determinan el comportamiento. El diseño de emparejamiento es una buena opción sólo cuando el investigador sea hábil en emparejar los sujetos, aunque esta habilidad, frecuentemente, bastante limitada.

La técnica paramétrica más usada para comparaciones entre dos muestras correlacionadas es el test **t.** Una diferencia de valores puede ser obtenida de dos valores provenientes de cada uno de los miembros de los sujetos emparejados o de los valores de cada sujeto bajo las dos condiciones. El test **t** asume que estas diferencias en los valores normalmente son distribuidas en la población de la cual fue retirada la muestra.

En muchos casos, el test **t** es inaplicable. En estas situaciones, el investigador puede escoger uno de los siguientes tests estadísticos no paramétricos:

1. **El test de McNemar para la significancia de variaciones**: este test es particularmente aplicable para los diseños de *"antes"* y *"después"* en los que cada persona es usada como su propio control y en medidas en las que los valores se obtienen usando variables nominales u ordinales. De este modo, por ejemplo, el test puede ser usado para testar la efectividad de un tratamiento particular (encuentro, editorial de un periódico, propaganda vía correo, visita personal, etc.) en las preferencias de votos entre varios candidatos. O puede ser usado para averiguar la influencia de las migraciones del campo para la ciudad en la preferencia política, etc.
2. **El test del Signo**: el test del signo recibe su nombre del hecho de que, en lugar de las medidas cuantitativas, usa más los signos de más y menos. Este test es particularmente útil para la investigación en la que son imposibles o impracticables las medidas cuantitativas, pero es posible ordenar los miembros de las parejas, uno con respecto al otro.
3. **El test de Walsh**: si el experimentador puede asumir que la diferencia en los valores, en dos muestras relacionadas, proviene de poblaciones que obedecen a distribuciones simétricas, puede usar un test bastante poderoso desarrollado por Walsh. Hay que observar que la hipótesis no es que las diferencias en los valores son de poblaciones normales (para las cuales se puede usar el test paramétrico **t**) y que las diferencias tampoco provienen de la misma población. Lo que hace el test es asumir que las poblaciones son simétricas, de modo que la media es una adecuada medida de tendencia central y es igual a la mediana.

## *Tests no paramétricos - el caso de dos muestras no relacionadas*

Cuando el uso de dos muestras relacionadas es impracticable o inapropiado, se puede hacer uso de dos muestras independientes. En este diseño, las dos muestras pueden ser obtenidas por uno de los dos métodos:

1. las muestras pueden ser retiradas aleatoriamente de dos poblaciones,
2. las muestras pueden originarse debido a la atribución aleatoria de dos tratamientos a miembros de una misma muestra, cuyos orígenes son arbitrarios.

En los dos casos, no es necesario que las dos muestras tengan el mismo tamaño.

La técnica paramétrica normal para averiguar la diferencia entre dos muestras no relacionadas es aplicar el test **t** en las medias de las muestras. En el caso de la no aplicabilidad del test paramétrico (por ejemplo, en el caso de no tener seguridad de que las distribuciones son normales), el investigador, para analizar sus datos, puede escoger uno entre los varios tests no paramétricos, los cuales se presentan a continuación.

## *Test de la probabilidad exacta de Fisher*

Ésta es una técnica extremadamente útil para analizar datos discretos cuando las dos muestras independientes tienen tamaños pequeños. Este test es usado cuando los valores de las dos muestras independientes, escogidas aleatoriamente, caen en una de dos categorías mutuamente exclusivas. En otras palabras, todo sujeto en los dos grupos obtiene uno de dos valores posibles. Los valores son representados por frecuencias en una tabla de contingencia *2×2*, como la Tabla 17.

*Tabla 17 - Tabla de contingencia para el test de Fischer.*

| | – | + | |
|---|---|---|---|
| *Grupo I* | *A* | *B* | *A+B* |
| *Grupo II* | *C* | *D* | *C+D* |
| *Total* | *A+C* | *B+D* | *N* |

Para los datos en esa tabla (donde *A, B*, *C* y *D* son frecuencias) podría ser determinado si el *Grupo I* y el *Grupo II* difieren significativamente en la proporción de más o menos atribuidos a ellos.

La probabilidad exacta de observar un particular conjunto de frecuencias en una tabla 2□2, cuando los totales marginales se mantienen fijos, es dada por la *distribución hipergeométrica:*

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$$

$$p = \frac{(A+C)!}{A!C!}\frac{(B+D)!}{B!D!} \times \frac{(A+B)!(C+D)!}{N!}$$

De este modo:

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

O sea, la probabilidad exacta de que ocurra de la manera observada es obtenida tomando la razón entre el producto de factoriales de cuatro totales marginales y el producto de los factoriales de los valores encontrados en cada celda por factorial de *N*, el número total de observaciones independientes[22].

Ejemplo: considérense los datos de la *Tabla 18*:

[22] El factorial de un número *a*, expresado por *a*! (se lee a factorial), es obtenido por el producto de los enteros, empezando en 1, hasta el número a. Así, por ejemplo: *3!=1x2x3=6*. O *0*! es igual a *1*, por definición.

| *Tabla 18 - Datos del Ejemplo.* | | | |
|---|---|---|---|
| | - | + | |
| *Grupo I* | *10* | *0* | *10* |
| *Grupo II* | *4* | *5* | *9* |
| *Total* | *14* | *5* | *19* |

En esta tabla, *A=10, B=0, C=4* y *D=5*. Los totales marginales son *A+B=10*, *C+D=9*, *A+C=14* y *B+D=5*. *N*, el número total de observaciones independientes, es *19*. La probabilidad exacta de que estos *19* casos recaigan en las celdas como aparece en el ejemplo es dada por:

$$p = \frac{10!9!14!5!}{19!10!0!4!5!} = 0{,}0108$$

Por tanto, determinamos que la probabilidad de obtener esa distribución de los valores, bajo $H_0$, es *p=0,0108*.

## El test $\chi^2$ para dos muestras independientes

Cuando los datos consisten en categorías discretas, ese test puede ser usado para determinar la significancia estadística de diferencias entre dos grupos independientes. Las medidas pueden ser incluso las de una escala nominal. Por ejemplo, si se desea saber si dos grupos de profesores, de Física y de Química, difieren en cuanto a su opinión con relación a una cierta estrategia de enseñanza, se puede medir esa opinión con un simple "a favor" o "contra", calcular las frecuencias y aplicar el test $\chi^2$.

## El test de la Mediana

Éste es un procedimiento para testar si dos muestras independientes difieren en tendencia central. Suministrará información de lo probable que es que las dos muestras independientes (no necesariamente del mismo tamaño) hayan sido retiradas de poblaciones con la misma mediana.

## El test U de Mann-Whitney

Cuando se dispone de por lo menos una medición ordinal, ese test puede ser usado para verificar si dos muestras independientes fueron tiradas de la misma población. Es uno de los más potentes tests no paramétricos y es una de las mejores alternativas para el test paramétrico **t** cuando el investigador quiere evitar suposiciones subyacentes al test **t** o cuando la medición realizada es más débil que una escala intervalar (Siegel, 1956, p. 116).

## El test de dos muestras de Kolmogorov-Smirnov

Éste es también un test para verificar si dos muestras independientes fueron retiradas de la misma población. La forma bilateral del test es sensible solamente a cualquier tipo de

diferencia en la distribución de la que fueron retiradas las dos muestras: diferencias en localización (tendencia central), en dispersión, en simetría, etc. El test unilateral es utilizado para testar si los valores de un grupo experimental serán *"mejores"* que los del grupo de control.

## El test de Wald-Wolfowitz

Si queremos testar la hipótesis nula de que dos muestras independientes hayan sido seleccionadas de la misma población contra la hipótesis alternativa de que los dos grupos difieren completamente, podremos utilizar este test. Es decir, con muestras suficientemente grandes, este test puede rechazar $H_0$ si las dos poblaciones difieren en cualquier aspecto: inclinación central, variabilidad, simetría o alguno otro factor. De este modo, este test puede ser usado en una extensa clase de hipótesis alternativas. Mientras que muchos tests son destinados a tipos específicos de diferencias entre dos grupos, el test de Wald-Wolfowitz analiza cualquier tipo de diferencia.

## El test de Moisés de reacciones extremas

En ciencias del comportamiento, algunas veces se espera que una condición experimental cause en algunos sujetos el aparecimiento de comportamientos extremos en una determinada dirección mientras que en otros sujetos, el comportamiento será extremado en la dirección opuesta. De este modo, se puede pensar que depresión económica e inestabilidad política provocarán en algunas personas reacciones extremadamente conservadoras, mientras que otras reaccionarán de una forma extremadamente progresista, en términos de opiniones políticas.

El test de Moisés es específicamente proyectado para ser usado con datos recogidos para testar ese tipo de hipótesis. Podría ser usado cuando se espera que la condición experimental afecte algunos sujetos de un modo y otros de manera opuesta.

## El test de la aleatoriedad para dos muestras independientes

Ésta es una técnica no paramétrica poderosa y útil para testar la significancia de la diferencia entre las medias de dos muestras independientes cuando $N_1$ y $N_2$ son pequeños. Con el test de la aleatoriedad, podemos determinar la probabilidad exacta, bajo $H_0$, asociada a nuestras observaciones y podemos hacerlo sin asumir la distribución normal u homogeneidad de la varianza en las poblaciones en cuestión (las cuales deben ser asumidas si se usa el test paramétrico equivalente, el test **t**).

## Discusión

Todos los tests no paramétricos para dos muestras independientes testan si es probable que las dos vengan de la misma población. Sin embargo, los varios tests presentados son más o menos sensibles a los diferentes tipos de diferencias entre las dos muestras. Por ejemplo, cuando se quiere testar si dos muestras representan poblaciones que difieren en localización (tendencia central), existen tests que son más sensibles a este tipo de diferencia y por tanto

podrían ser escogidos: el test de la mediana, el test de Fisher (para *N* pequeña), el test U de Mann-Whitney, el test de Kolmogorov-Smirnov (para dos muestras, unilateral) y el test de la aleatoriedad. Por otro lado, si fuese deseo del investigador determinar si sus dos muestras provienen de poblaciones que difieren en cualquier aspecto en general, es decir, localización o dispersión o simetría, etc., podría escoger uno de los siguientes tests: el $\chi^2$, el test de Kolmogorov-Smirnov (bilateral) o el test de Wald-Wolfowitz. La técnica restante, el test de Moisés, sirve únicamente para testar si un grupo experimental está exhibiendo reacciones extremas, sean extremistas o defensivas, en comparación con las reacciones exhibidas por un grupo de control independiente.

La elección entre los tests que son sensibles a diferencias en localización son determinadas por el tipo de medida obtenida en la investigación y por el tamaño de las muestras. El test más poderoso en términos de localización es el test de la aleatoriedad. Sin embargo, este test puede ser usado solamente cuando los tamaños de las muestras son pequeños y cuando tengamos confianza en la naturaleza numérica de la medida obtenida. Con grandes muestras o medidas débiles (medidas ordinales), la alternativa sugerida es el test U de Mann-Whitney el cual, casi siempre, es más poderoso que el test de la aleatoriedad. Si las muestras son muy pequeñas, el test de Kolmogorov-Smirnov es levemente más eficiente que el test U. Si la medida es de un tipo tal que es significativo solamente dividir las observaciones por encima o por debajo de la mediana, entonces el test de la mediana es aplicable. Este test no es tan poderoso como el test U de Mann-Whitney en términos de protección contra diferencias de localización, pero es más apropiado cuando los datos de las observaciones no pueden ser completamente jerarquizados. Si el tamaño de las muestras en consideración es muy pequeño, cuando se aplica el test de la mediana, el investigador podría hacer uso del test de Fisher.

La elección entre los tests que son sensibles a todas las diferencias es decidida a partir de la intensidad de las medidas obtenidas, el tamaño de las muestras y el poder relativo de los tests disponibles. El test $\chi^2$ es apropiado para datos que son medidos en escalas nominales o más fuertes. Cuando N es pequeña y los datos están en una tabla de contingencia 2x2, podríamos usar el test de Fisher en lugar del test $\chi^2$. En muchos casos, el test $\chi^2$ puede no hacer uso eficiente de toda la información contenida en los datos. Si los valores de las poblaciones son continuamente distribuidos, podemos escoger el test de Kolmogorov-Smirnov (bilateral) o el test de Wald-Wolfwitz, en lugar del test $\chi^2$. De todos los tests para cualquier tipo de diferencia, el test de Kolmogorov-Smirnov es el más poderoso. Si es usado con datos que no asumen la hipótesis de continuidad, es aceptable, pero opera más conservadoramente, es decir, los valores de *p* obtenidos serán levemente mayores. Si la hipótesis nula es rechazada a partir de esos datos, se puede seguramente tener confianza en la decisión. El test de Wald-Wolfwitz también protege contra todos los tipos de diferencias, pero no es tan poderoso como el anterior.

Dos puntos deben ser enfatizados sobre el uso de tests del segundo grupo. Primero, cuando se está interesado en testar hipótesis alternativas de que los grupos difieren en tendencia central, es decir, de que un grupo tiene media mayor que el otro, entonces se debe usar un test específicamente proyectado para capturar diferencias en localización, uno de los tests del primer grupo citado anteriormente. Segundo, cuando se rechaza la hipótesis nula con base en un test que capta cualquier tipo de diferencia (uno de los tests del segundo grupo), se debe asegurar que los dos grupos provienen de poblaciones diferentes, sin embargo no se puede decir en qué difieren.

## *Bibliografía*

BEST, J. W. (1970). *Research in education*. 2. ed. Englewood Cliffs: Prentice Hall.

CAMPBELL, D. R.; STANLEY, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In Gage, N. L. *Handbook of research on teaching*, Chap. 5. Chicago: Rand Mc Nally.

CAMPBELL, D. R.; STANLEY, J. C. (1991). *Diseños experimentales y cuasiexperimentales en la investigación social*. Buenos Aires: Amorrortu Editores.

DARLINGTON, R. B. (1975). *Radicals and squares*. Ithaca, N. Y.: Logan Hill Press.

ELSEY, F. F. (1967). *A first reader in statistics*. Belmont, CA: Brooks/Cole Publishing Co.

FOX, D. J. (1969). *The research process in education*. New York: Holt, Rinehart and Winston.

GLASS, G. V.; STANLEY, J. C. (1970). *Statistical methods in education*. Englewood Cliffs, N. J.: Prentice Hall.

GOWIN, D. B. (1981). *Educating.* Ithaca, N.Y.: Cornell University Press.

GOWIN, D. B.; ALVAREZ, M. (2005). *The art of educating with V diagrams.* New York: Cambridge University Press.

KERLINGER, F. N. (1964). *Foundations of behavioral research*. New York: Holt, Rinehart and Winston.

KERLINGER, F. N. (1980). *Metodologia da pesquisa em ciências sociais*. São Paulo: E.P.U., EDUSP, INEP.

MILLMAN, J. (1970). *Data analysis*. Conferência convidada proferida no Simpósio Nacional de Professores de Pesquisa Educacional, St. Louis, USA.

MOREIRA, M. A. (2006). *Mapas conceituais e diagramas V*. Porto Alegre: Editora do Autor.

RUNKEL, P. J.; MC GRATH, J. E. (1972). *Research on human behavior*. New York: Holt, Rinehart and Winston.

SIEGEL, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill Book Co.

SPIEGEL, M. R. (1973). *Statistics*. New York: Schaum Publishing Co.