



### **Acerca del Autor**

El PhD. Luis Alfonso Pérez Romero, es profesor del MBA desde 1992 de Metodología de la Investigación Científica para la Ciencias Sociales, de la Facultad de Postgrado de Contaduría y Administración de la UNAM, México, profesor de Investigación de Mercados y Análisis Multivariante del pregrado del TEC de Monterrey, profesor e investigador de la Maestría de Mercadotecnia y del MBA de la Escuela de Graduados en Administración y Dirección de Empresas (EGADE) del Instituto Tecnológico y de Estudios Superiores de Monterrey y profesor invitado por la Universidad Javeriana de Colombia, con Métodos Cuantitativos para la Toma de Decisión Empresarial en el MBA, con Tableau, IBM-SPSS, Watson Analytics, AMOS y COGNOS, entre otros. Es autor del concepto de Modelo de Investigación en una Página, publicado en uno de sus libros en el 2004. Actualmente es el CEO de COMEVE ASOCIADOS, SA DE CV, Consultoría en Marketing e Investigación de Mercados y profesor invitado de universidades de América Latina con temas de minería de datos y business analytics.

## Prólogo

*Es muy grato para mi recibir de mi colega Luis Alfonso Pérez Romero PhD, con quien la actividad académica orientada a la formación de posgrado, que realizamos en América latina nos une, el libro “Metodología de la investigación en una página” resultado de su trabajo como investigador y profesor por más de veinte años, el cual presenta de forma fácilmente comprensible en un estilo claro y fácil de entender, mediante una estructura de ocho columnas, empezando con la idea hasta el uso de las herramientas estadísticas, para la realización de una investigación con el método inductivo/deductivo, hipotético y cuantitativo y así reducir la sobrecarga de información propia de estos procesos, enfocándonos en los datos claves, enfatizando la importancia y necesidad de tener información precisa, oportuna e importante, que se procesa mediante el IBM-SPSS y sus análisis estadísticos descriptivos e inferenciales univariados, bivariados y multivariados, que sin duda guiará a neófitos y avezados en la realización exitosa de proyectos de investigación científica.*

**Manuel Alfonso Garzón Castrillón, PhD, Post PhD, Tutor Doctorado en Ciencias de la Administración, Universidad Nacional Autónoma de México, UNAM.**

*Felicito al doctor Luis Alfonso Pérez Romero, por su nuevo logro. Es importante apreciar una obra tan significativa escrita por un experto conocedor quien vivencia la academia desde distintos saberes y culturas. Ha sido muy grato contar con su conocimiento y habilidades docentes desplegadas en sus actividades con los posgrados para los cuales nos ha permitido su acompañamiento. Esta obra será de vital importancia para los estudiantes ansiosos por tener una guía clara y didáctica sobre la metodología de la investigación y sobre el uso de las herramientas propias de la estadística con miras a excelentes análisis de sus resultados. Con total seguridad se constituirá en un texto de consulta permanente para estudiantes y profesores deseosos de acertar con sus escritos y temas de investigación. Auguro al doctor Pérez, el mayor éxito con esta nueva producción.*

**Gladis Rodríguez Muñoz, Directora de Posgrados de la Facultad de Ciencias Económicas y Administrativas Pontificia Universidad Javeriana de Cali, Colombia.**

*En general, realizar una investigación científica es una tarea difícil si no se cuenta con una guía que nos explique claramente cuáles son los pasos que se deben seguir para llevarla a cabo. Precisamente el procedimiento que nos brinda el autor en este libro permitirá que los lectores – estudiantes, profesores e investigadores - cuenten con un marco metodológico adecuado y con la explicación de las herramientas estadísticas, que les sirvan de referencia para desarrollar una investigación en el área de administración y ciencias sociales.*

**Edmundo González Zavaleta, Rector, Universidad Privada de Trujillo, Perú.**

*Con verdadero interés he revisado esta obra que nos ofrece el Dr. Luis Alfonso Pérez Romero, Metodología de la investigación en una página, es una obra que, de forma práctica, pero suficientemente profunda, nos muestra cómo enfrentar el desafío de la investigación. En sus capítulos podrá el lector descubrir: estructura, orden y claridad en las ideas que nos presenta. Es sencillamente una obra que resultará de gran utilidad y relevancia en el quehacer investigativo y profesional.*

*Mediante una narrativa simple, Luis Alfonso nos muestra la forma de estructurar una investigación y comprender cómo y por qué realizarla. Va desde la historia de la ciencia hasta aplicaciones específicas, apoyadas con herramientas contemporáneas como es el software de IBM-SPSS para el levantamiento y procesamiento de información.*

*Reconozco que la obra no podría llegar en un mejor momento; la era de la información. Una época en la que la tecnología nos permite acopiar miles, millones de datos que provienen de dispositivos electrónicos y que forman bases de datos gigantescas, sin embargo, sin su estudio, serían montañas de información desarticulada, sin poder interpretarse. Una época en la que sin algoritmos y métodos estadísticos la llamada Big Data (o Data Science) carecería de todo sentido práctico.*

*Por ello, el tema es de suma actualidad; la metodología de la investigación y el análisis de sus resultados desde múltiples perspectivas. Al revisarle me doy cuenta de que contiene años de experiencia práctica del autor en el uso de estas herramientas, pero que puede ser empleada por diferentes públicos: estudiantes y profesionales. He observado que tiene un enfoque didáctico que permite la comprensión de estudiantes de diversos niveles, pregrado o profesional, o bien posgrado de maestría y doctorado.*

*Estoy seguro de que será de gran ayuda a estudiantes y profesionistas que desean mejorar o perfeccionar sus capacidades de investigación y análisis de los fenómenos que nos ocupan en la actualidad.*

**Dr. Jorge Alfonso González González, Director del Centro Internacional de Casos, Tecnológico de Monterrey, México.**

## **Agradecimientos**

Agradezco a Dios y a mis padres por la vida, a mis hijas Ayari (La esperanza), Manuela (La confianza), Camila (La seguridad) y Paula (La tenacidad), por la prolongación de la vida y a mi esposa Lucia Cristina Lizcano, por ser la primera crítica y usuario de este libro durante más de 18 años en el ejercicio de su profesión de Investigación de Mercados.

Agradezco a las universidades que me invitaron a ser maestro en sus maestrías en Administración, Mercadotecnia, Dirección de Calidad, Planeación Estratégica, Administración de Hospitales: UNAM, EGADE, IBERO, TEC DE MONTERREY, UP, en Colombia la JAVERIANA, EXTERNADO, UNINORTE y la UNIVERSIDAD DE ANTIOQUIA.

Agradezco a las empresas que me han invitado a ser consultor en sus áreas de investigación para la toma de decisiones en México, Colombia, Perú, Bolivia y Panamá.

Agradezco a todo aquel que ha sido mi estudiante, ya que usted ha sido mi verdadero profesor y así se los he externado siempre que he tenido la fortuna de impartir un curso: “He venido a aprender con ustedes y no a enseñarles”.

Agradezco a usted estimado lector que puede encontrar en este escrito puntos de orientación, crítica y búsqueda permanente hacia el conocimiento.

<b>Contenido</b>	<b>Pág.</b>
Introducción	12
<b>Capítulo I</b>	
Metodología de la Investigación	14
• Método inductivo y deductivo	14
• Cientificidad positivista	17
• La ciencia: hechos y conocimientos	18
• Historia de la ciencia	19
Conclusión	27
Palabras clave	28
Preguntas	29
<b>Capítulo II</b>	
Metodología de la Investigación en una Página	29
Ideas y tema	30
• Fuentes de información	30
• Tipos de Investigación	30
• Planteamiento del problema	33
• Objetivos	34
• Hipótesis	34
• Diseño de la investigación	35
• Cronograma	35
Variables	35
Definir las variables	39
Mensurabilidad de las variables	39
Hipótesis	39
Técnicas de Investigación	41
Muestra y trabajo de campo	45
• Fórmulas para la selección del tamaño de la muestra	50
• Determinar el tamaño de la muestra	53
• Fórmulas para la selección del tamaño de la muestra	53
• Determinar el tamaño de la muestra población infinita	53
• Determinar el tamaño de la muestra población finita	55
• Muestreo estratificado desproporcional	58

• Muestreo por conglomerados	59
Método de muestreo probabilístico y no probabilístico	60
• Método Probabilístico	60
○ Simple	60
○ Sistematizado	60
○ Muestreo estratificado proporcional y desproporcional	61
○ Muestreo por conglomerados	61
○ Muestreo por áreas	61
○ Muestreo polietápico	62
• No probabilísticos	62
○ De conveniencia	
○ Con fines especiales	
○ Por cuotas	
○ De juicio	
Trabajo de campo	64
Diseño Estadístico	65
Conclusión	67
Palabras clave	68
Preguntas	69
<b>Capítulo III</b>	
Escalas de Medición e instrumentos de medición	70
• Escala nominal	70
• Escala ordinal	70
• Escala intervalar	71
• Escala de razón	73
Instrumento de medición tipo cuestionario	74
• Prueba de validez	77
○ Prueba de validez de contenido	
○ Prueba de validez de constructo o de concepto	
○ Prueba de validez concurrente	
○ Prueba de validez convergente	
○ Prueba de validez discriminante	
• Prueba de confiabilidad	79
○ Estabilidad	

○ Equivalencia	
○ Pruebas de fiabilidad con Alfa de Cronbach y KR20	81
• Supuesto de normalidad	82
• Datos atípicos o Outliers	83
Conclusión	87
Palabras clave	89
Preguntas	89
<b>Capítulo IV</b>	
Estadística para representatividad de la muestra	90
• Variables asociadas a la distribución normal se utilizará la prueba T de Student.	
• Para variable asociadas a la distribución binomial	93
Conclusión	96
Palabras clave	97
Preguntas	97
<b>Capítulo V</b>	
Estadística descriptiva e inferencial	98
Estadística univariada descriptiva e inferencial	103
• Escala nominal	105
• Escala ordinal	108
• Escala intervalar y de razón	112
• Distribución normal, cuartiles, deciles, percentiles y la probabilidad	113
Estadística Bivariada	125
• Escala nominal prueba de Chi cuadrada	127
• Prueba McNemar	130
• Escala Ordinal Vs Nominal	132
• Escala intervalar y razón	134
Pruebas no paramétricas	138
• Prueba de la Chi Cuadrada	138
• Prueba de Kolmogorov-Smirnov	140
• Prueba U de Mann-Whitney	141
• Prueba de Wilcoxon	143
• Prueba de Kruskal-Wallis	144
• Prueba de la mediana	146

Conclusión	148
Palabras clave	149
Preguntas	149

## Capítulo VI

Estadística Multivariada	150
I. Análisis de la Varianza o ANOVA	155
• Análisis de ANOVA Factorial	160
• Análisis MANOVA o Análisis de Varianza Multivariable	161
• Análisis de Covarianza o ANCOVA	163
• Análisis de Varianza no Métrica	166
II. Análisis de correlación y regresión	167
• Correlación	168
Análisis de regresión	169
• Regresión simple	171
• Regresión múltiple	175
• Regresión discriminante	177
• Regresión Logística Binaria y Multinomial	179
• Modelos Probit y Logit	181
• Procedimiento de Regresión de Cox	182
III. Análisis de factores	184
IV. Análisis discriminante	186
V. Análisis de clúster	191
VI. Ecuaciones estructurales	192
Conclusión	207
Palabras clave	208
Preguntas	208
Referencias bibliográficas.	209
Anexo 1: Ejemplo del modelo de investigación en una página	211
Anexo 2: Tabla distribución normal con valores Z	215

<b>Contenido de gráficas, tablas y diagramas de flujo</b>	<b>Pág.</b>
Gráfica No. 1, Método inductivo/deductivo	14
Gráfica No. 2, Evolución del método inductivo/deductivo	15
Gráfica No. 3, Interacción del método inductivo/deductivo.	16
Gráfica No. 4, Cientificidad	17
Gráfica No. 5, La ciencia: hechos / conocimiento	18
Gráfica No. 6, Historia de la ciencia I	21
Gráfica No. 7, Historia de la ciencia II	21
Gráfica No. 8, Historia de la ciencia III	25
Gráfica No. 9, Historia de la ciencia IV	25
Gráfica No. 10, Historia de la ciencia V	26
Gráfica No. 7, Historia de la ciencia	24
Tabla 1, Investigación en una página	27
Gráfica No. 11, Variables en las investigaciones	37
Gráfica No. 12, Características que afectan la conducta del consumidor	38
Gráfica No. 13, Muestra	47
Gráfica No. 14, Determinación del tamaño de la muestra	48
Gráfica No. 15, Población infinita	53
Gráfica No. 16, Población finita	55
Gráfica No. 17, Muestreo estratificado desproporcional	58
Gráfica No. 18, Muestreo por conglomerados	59
Gráfica No. 19, Método de muestreo probabilístico y no probabilístico	63
Tabla 2, Instrumento de medición - cuestionario	74
Gráfica No. 20, Validez y confiabilidad	80
Gráfica No. 21, Diseño estadístico descriptivo e inferencial	102
Gráfica No. 22, Procedimientos univariados	104
Gráfica No. 23, Procedimientos bivariados	126
Tabla 3, Taller	130
Tabla 4, Pruebas no paramétricas	138
Gráfica No. 24, Proceso del análisis multivariante	151
Gráfica No. 25, Análisis multivariante	153
Gráfica No. 26, Análisis estadístico multivariado	154
Gráfica No. 27, Covarianza	163
Gráfica No. 28, Correlación positiva	167

Gráfica No. 29, Correlación negativa	168
Gráfica No. 30, No correlación	168
Gráfica No. 31, Plano cartesiano	169
Tabla 5, Regresión lineal	170
Diagrama de flujo 1	196
Diagrama de flujo 2 y 3	198
Diagrama de flujo 4	199
Diagrama de Flujo 2.1	200
Diagrama de flujo 3.1	200
Diagrama de flujo 2.2	101
Tabla Nro 6.	204
Diagrama de flujo Nro. 5	205

## INTRODUCCIÓN

Metodología de la investigación en una página es una guía práctica dirigida a todas aquellas personas que deben realizar una investigación cuantitativa desde la idea hasta el desarrollo de técnicas estadísticas con el apoyo del software IBM-SPSS.

Se realiza una narrativa del método de investigación inductivo/deductivo, hipotética y cuantitativa, con su respectiva evolución histórica del pensamiento de la investigación cuantitativa.

Se proporciona el proceso de la investigación en una página en formato de 8 columnas con temas a desarrollar para cada una de ellas y en forma evolutiva desde la columna uno hasta la 8. Lo primero en toda investigación es pasar de la idea a definir las variables dependientes e independientes, definir la población objeto de estudio o la unidad de análisis, definir el tipo de estudio exploratorio, descriptivo o causal, plantear el problema, definir objetivos, hipótesis y diseño de la investigación con su respectivo cronograma de actividades. Los análisis estadísticos a realizar deben ser descriptivos e inferenciales univariado, esto es, variable por variable o bien cada una de las preguntas del cuestionario, para que se pueda comprender el comportamiento de las medidas de centralidad y dispersión con sus respectivos porcentajes de las variables que integran el proyecto de investigación; en el análisis inferencial univariado se trabaja con la hipótesis de investigación para conocer la inferencia de las variables de la muestra hacia la población con el apoyo de estadísticos inferenciales como la T de Student, la Chi Cuadrada y la Kolmogorov-Smirnov entre otras.

El análisis bivariado descriptivo e inferencial, se realiza para encontrar la relación que existe entre dos variables y contrastar la hipótesis entre ellas, por ejemplo, la variable dependiente ventas y la independiente edad, en la que se pretende conocer la relación que existe entre el volumen de ventas en función de la edad de los clientes, para ello se formula la hipótesis nula,  $H_0$ : No hay diferencia significativa entre las ventas y la edad de los clientes; hipótesis alterna  $H_a$ : Si hay diferencia significativa entre el volumen de ventas y la edad de los clientes. En este ejemplo se realiza la estadística descriptiva para encontrar la media y la desviación estándar entre los grupos de edad y las ventas, para ver qué grupo de edad compra más,

mientras que la estadística inferencial permite inferir estos resultados de una muestra hacia el comportamiento de la población, con el apoyo de estadísticos inferenciales como la prueba Z o la prueba T de Student.

El otro apartado del libro lo conforman las pruebas de estadísticas multivariantes, para ello se requiere analizar más de dos variables al mismo tiempo para analizar el comportamiento de las variables dependientes ante la interacción de más de dos variables independientes; como la técnica estadística de ANOVA, ANCOVA, MANOVA y MANCOVA, análisis de regresión simple y múltiple y análisis discriminante. También en este apartado se realizan análisis de agrupación de variables sin diferenciar entre variables dependientes e independientes para encontrar sentido a las agrupaciones de personas para segmentarlas por características comunes dentro de cada segmento, pero claramente diferenciadas entre los segmentos, por ejemplo, se pueden agrupar a los compradores según el perfil demográfico (edad, sexo, escolaridad, ocupación, etc.), análisis de factores y análisis de clúster.

## I. METODOLOGÍA DE LA INVESTIGACIÓN

La investigación es la actividad que permite los avances en las ciencias con el apoyo de la investigación científica aplicando el método científico: proceso dinámico, armónico, sistémico y holístico para la obtención del conocimiento básico y/o aplicado para avanzar en la ciencia (investigación básica) o para resolver problemas concretos (investigación aplicada). El método de investigación viene de la corriente del positivismo, de donde emana un solo método para todas las ciencias, el cual se apoya en el método inductivo, deductivo, hipotético y cuantitativo, que puede partir de los hechos adquiridos a través de la observación, para ir de lo particular a lo general (inductivo), para formular predicciones y explicaciones (deductivo: ir de lo general a lo particular) con la integración de las leyes y teorías. Ver gráfica número 1.

Gráfica No. 1

### MÉTODO INDUCTIVO/DEDUCTIVO



Fuente: Adaptación de Chalmers, Alan, ¿Qué es esa cosa llamada ciencia?, Editorial Siglo XXI, México, pp17

La evolución de los hechos adquiridos a través de la observación y de la experiencia: empirismo e inductivo-deductivo, hipotético y cuantitativo (Positivismo de Aristóteles, Kepler, Galileo, Descartes, Newton, Locke, Kant, Saint Simón,

Comte, Stuart Mill, Hume, Bacon, Popper, Kuhn, Feyerabend y Lakatos: explicación causal de los hechos o galileana), lleva a toda investigación básica para generar más conocimiento o bien aplicada para resolver problemas concretos de una empresa, industria, o sector, a estar muy cerca de la línea entre el método inductivo/deductivo:

- **Método deductivo:** el método científico deductivo, permite usar el pensamiento de la deducción para inferir lo investigado con el soporte o el fundamento teórico necesario, partiendo de lo general a lo particular con el fundamento teórico previo a lo que se va a observar o a encuestar. Con el conocimiento abstracto, se puede afirmar que A es igual a B y B es igual a C, por lo tanto, A y C son iguales, siendo estos los principios y pensamiento científico del silogismo aristotélico para avanzar en la ciencia.
- **Método inductivo:** se refiere a estudios que parten de lo particular a lo general, permitiendo la formulación de nuevas teorías o leyes con el soporte de la observación, trabajos de investigación con cuestionario y la experimentación. Ver gráfica número 2.

Gráfica No. 2

### EVOLUCIÓN DEL MÉTODO INDUCTIVO/DEDUCTIVO



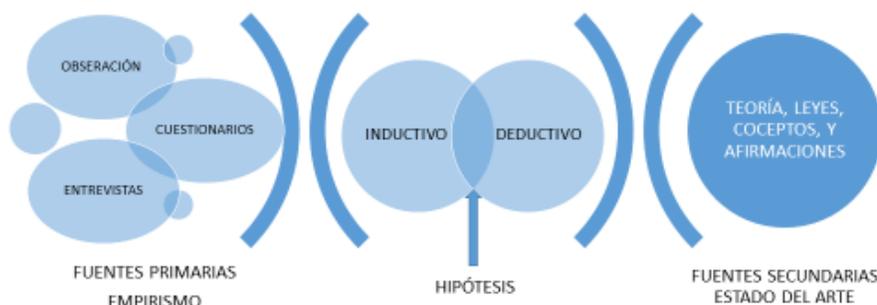
Fuente: aportación del autor

- Interacción entre los métodos inductivo y deductivo: las hipótesis une los dos métodos de investigación, logrando una interacción dinámica entre el empirismo y la teoría. Por el lado inductivo se necesita de las fuentes primarias de investigación como la observación, el cuestionario y/o las entrevistas estructuradas para realizar los análisis estadísticos y matemáticos necesarios para contrastar las hipótesis e identificar patrones que lleven a formular nuevas afirmaciones, conceptos, teorías, leyes o conocimiento; mientras que el método deductivo se apoya de las fuentes secundarias de información como las teorías, leyes, conceptos y afirmaciones para reconfirmar las hipótesis o bien para formular hipótesis sujetas a ser comprobadas con la investigación aplicada a resolver problemas. El método de mayor uso en los proyectos de investigación es el método inductivo/deductivo, hipotético, cuantitativo y fundamentado con los conceptos, teorías, principios o leyes y con técnicas estadísticas y modelos matemáticos para contrastar hipótesis que permitan avanzar en el conocimiento. Ver figura número 3.

Gráfica No. 3



## INTERACCIÓN DEL MÉTODO INDUCTIVO/DEDUCTIVO



Método inductivo/deductivo, hipotético, cuantitativo y fundamentado con teoría, estadística y matemáticas.

Esta corriente del pensamiento científico en las ciencias, lleva a adquirir un monismo metodológico, que permite comparar los estudios que se realizan en varios lugares del mundo, de ahí que se tiene la idea de una doctrina homogénea, generalmente apoyada de los modelos de las ciencias duras, para explicar y predecir los fenómenos en estudio. Ver gráfica número 4.

Gráfica No.4

### CIENTIFICIDAD POSITIVISTA



Fuente: aportación del autor.

La ciencia avanza por el conjunto de hechos observables, registrables, verificables y con el apoyo de fuente secundaria y primaria de información necesarias para la creación del conocimiento (Interacción entre los métodos inductivo y deductivo). Con el método de investigación inductivo/deductivo, seguido del proceso de la investigación con un planteamiento del problema, justificación, objetivos, hipótesis, marco teórico, conceptual y antecedente (Estado del arte), los hechos se registran con las técnicas de investigación de la observación, experimentos y cuestionarios, para describir e inferir patrones e información para contribuir en la creación de nuevas ideas, conceptos, teorías o leyes propias del crecimiento y desarrollo de la ciencia. Es recomendable seguir el proceso de la investigación para avanzar en la

ciencia, pero lo cierto es que muchas investigaciones se han realizado sin seguir un método, pero, es recomendable en primer lugar registrar los hechos (inductivismo), darle un fundamento teórico (deductivismo) a lo registrado con fundamento estadístico y matemático de los modelos que proponen contrastación de hipótesis o bien la generación de nuevo conocimiento, bajo enunciados y verificables, dándole fuerza al marco teórico-conceptual actual y generando un conocimiento más comprensible o nuevo conocimiento. Ver gráfica número 5.

Gráfica No. 5



Fuente: adaptada de Chalmers, Alan, ¿Qué es esa cosa llamada ciencia?, Editorial Siglo XXI, México p20-24

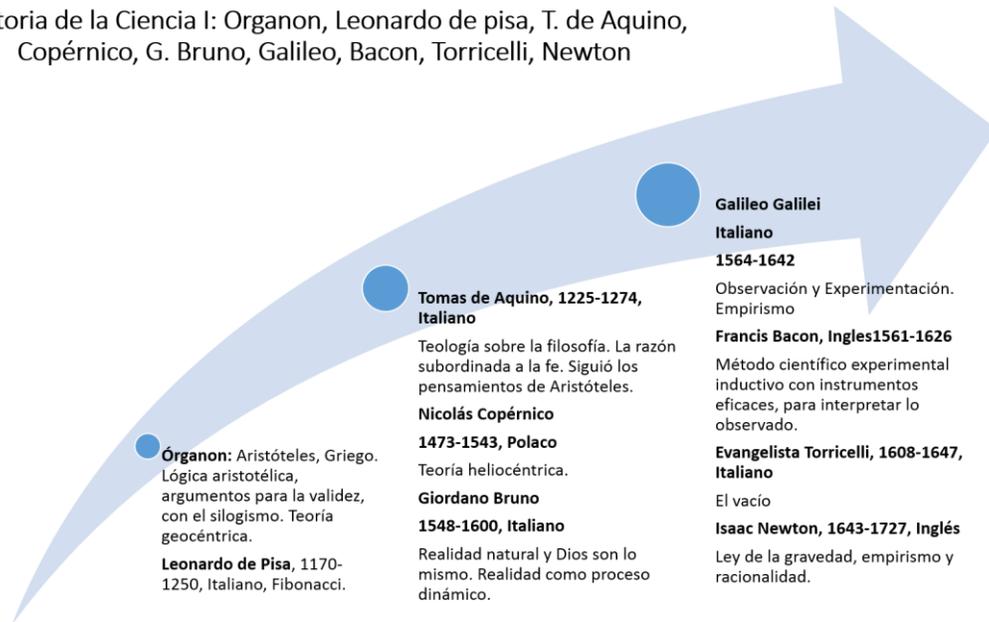
## Historia de la ciencia:

En la historia de la ciencia se puede ver la evolución del pensamiento a partir del escrito de Aristóteles conocido como el Órganon, documento que integra el pensamiento y las ideas del griego Aristóteles (antes de la era cristiana) compartió la lógica del pensamiento aristotélico de observar, registrar y con el silogismo avanzar en la ciencia, de ahí la teoría geocéntrica que perdura por más de mil quinientos años y que sigue siendo la base del pensamiento actual, destacando el pensamiento de que la tierra es el centro del universo (geocentrismo) y que los cuerpos más pesados caen primero que los más livianos; para 1200 después de Cristo, el italiano Leonardo de Pisa, comparte el pensamiento de identificar patrones o secuencias de los eventos de la naturaleza y de la conducta humana con la sucesión Fibonacci o sucesión matemática e infinita de números naturales que siguen una secuencia ordenada de números que se suman partiendo del cero y sumando siempre los dos últimos números: 0, 1, (0+1= 1) 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144... los cuales sirven para identificar series o patrones de comportamiento de la naturaleza o de sucesos sociales, permitiendo de esta manera inferir la futura secuencia o patrón a observar. El italiano Tomás de Aquino en 1250, contribuye al pensamiento científico con la integración de la teología por encima de la filosofía, existe un ser supremo que es el creador de todas las cosas, por ende la razón debe estar siempre subordinada a la fe y siguió los pensamientos de Aristóteles. Para 1500, el polaco Nicolás Copérnico, propone la teoría heliocéntrica y echa abajo el concepto geocéntrico; a finales de 1500, el italiano Giordano Bruno, comparte el concepto de la realidad natural y Dios es la misma cosa integrada en un todo, por lo que la realidad es un proceso dinámico en constante transformación; para 1600, el italiano Galileo Galilei, con sus experimentos y observaciones objetivas reconfirma la teoría heliocéntrica y comprueba que los cuerpos con menos masa caen primeros que los más pesados, por lo que es juzgado a cadena perpetua por ser un hereje y no estar de acuerdo con los pensamientos geocéntricos de Aristóteles. Para el mismo 1600 el italiano Evangelista Torricelli, inventa el concepto del vacío y llega a la conclusión que los cuerpos sin importar la masa y el peso dentro del vacío caen iguales. Para el mismo periodo de tiempo el inglés Francis

Bacon, proporciona el pensamiento hacia el método científico experimental inductivo con instrumentos eficaces para interpretar lo observado y compartir los hallazgos con otros colegas. El matemático inglés Isaac Newton, contribuye a la ciencia con las leyes de Newton y con la ley de la gravedad. Para 1630, el francés René Descartes, contribuye con el método cartesiano en apoyo del modelo matemático en sustitución del silogismo aristotélico y busca remplazar el idealismo subjetivo por el conocimiento abstracto y todo debe girar en el sujeto cognoscente, famoso por la frase: “pienso luego existo”. Para 1660 el holandés, Baruch Spinoza, reafirma el pensamiento racional de Descartes, es idealista y comparte el concepto de la sustancia como Dios y también se centra en la idea de que la fe y la ciencia deben estar separadas para el avance científico; proporcionó los tres elementos clave para la ética y la felicidad: la imaginación, la razón y la intuición. A principio de los 1700, el Alemán Leibniz, aporta el cálculo diferencial en unión con Newton y crea el término de las mónadas como el principio de todas las cosas que dan forma a la materia, fortaleciendo el concepto de sustancia de Baruch Spinoza. A mediados de 1700, el escocés David Hume, fortalece el pensamiento del método de investigación inductivo, afirmando que todo conocimiento emana de la experiencia y se puede inferir los efectos a partir de las causas. Casi en el mismo periodo el matemático inglés Bayes Thomas, crea el concepto de la probabilidad con el teorema de Bayes, fortaleciendo el método de investigación con la inferencia inductiva. A finales de 1700, el Alemán Immanuel Kant, integra el componente de la crítica de la razón pura y considera la ética como práctica central de todo investigador, fortaleciendo el empirismo y racionalismo; la ciencia avanza, gracias al saber con fundamento cognitivo, matemático y físico, generando una conciencia crítica. Ver gráfica número 6 y 7.

Gráfica No. 6

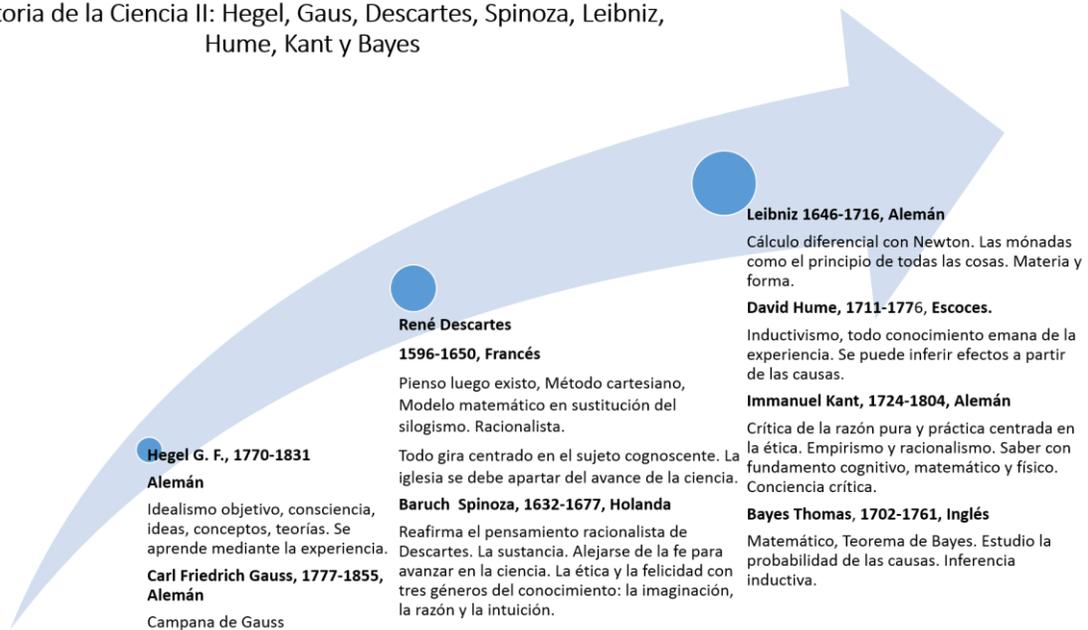
### Historia de la Ciencia I: Organon, Leonardo de pisa, T. de Aquino, Copérnico, G. Bruno, Galileo, Bacon, Torricelli, Newton



Fuente: aportación del autor.

Gráfica No. 7

### Historia de la Ciencia II: Hegel, Gaus, Descartes, Spinoza, Leibniz, Hume, Kant y Bayes



Fuente: aportación del autor.

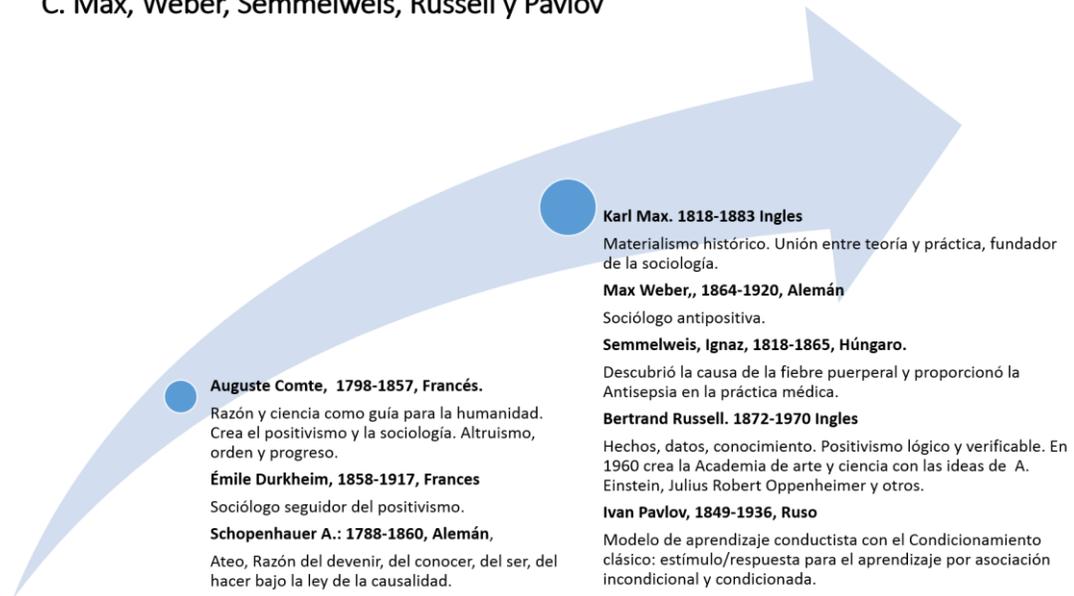
Siguiendo el recorrido cronológico del avance de la ciencia, para 1800, el Alemán Hegel, proporciona el concepto del idealismo objetivo, se debe tomar conciencia de los avances de las ideas, conceptos, teorías, principios y leyes, de ahí que se aprende mediante la experiencia. En el mismo periodo el Alemán Carl Friedrich Gauss, fortalece el conocimiento de la probabilidad con la campana de Gauss y se inicia el pensamiento estadístico de medidas de centralidad y dispersión dentro de una unidad con valores de desviaciones estándar de la agrupación de una serie de datos. El francés Auguste Comte, para mediados de 1800, aporta la razón y ciencia como guía para la humanidad y se da inicio al positivismo y la sociología, con el altruismo, orden y progreso; en ese mismo periodo el alemán Schopenhauer, contribuye con la ley de la causalidad para avanzar en la ciencia, se parte del conocer y del ser, dando libertad a la razón del devenir. Karl Max, a finales de 1800, aporta con su pensamiento del materialismo histórico, dando origen entre la unión de la teoría y la práctica; la ciencia avanza gracias al conocer con el desarrollo del marco teórico y conceptual para aplicarlo con el conocimiento empírico, no se trata de practicar sin saber, se requiere primero del saber para llevarlo a la práctica. En los primeros años de los 1900, el Inglés Bertrand Russell, parte del principio de los hechos que llevan al registro de datos que permiten la generación del conocimiento; se afianza el concepto del positivismo lógico y verificable, apartando el pensamiento idealista y religioso en el proceso de la investigación científica; la ciencia debe avanzar de manera lógica, razonable y verificable, sin ideales fundados en la fe o la fuerza de un ser supremo; en los años de 1950, en compañía de otros investigadores contemporáneos como Albert Einstein, Oppenheimer y otros investigadores crean la Academia de Arte y Ciencia. Para mediados de 1900, el austriaco y nacionalizado británico, Karl Popper, aporta su pensamiento con el positivismo y el falsacionismo como enfoque para avanzar en la ciencia y no por la inducción; la ciencia avanza gracias a las hipótesis que deben ser contrastadas con la realidad y la teoría, dando pauta a la generación de un método universal para avanzar en la ciencia alejado totalmente del poder eclesiástico que estuvo frenando la ciencia en los últimos mil ochocientos años. El pensamiento del austriaco y nacionalizado estadounidense, Paul Karl Feyerabend, contrario al positivismo,

menciona que no hay método para avanzar en la ciencia, ya que al momento de investigar para generar conocimiento “todo vale”, llevando al pensamiento de una metodología anarquista y sugiere separar la ciencia del poder del estado. A finales de 1900, el húngaro Imre Lakatos, fortalece la aportación de Popper, con el falsacionismo y fortalece la idea de que no existe racionalidad instantánea en la ciencia, se avanza en la generación del conocimiento gracias a un método inductivo/deductivo universal, verificable, autocorregible, hipotético y cuantificable. El norteamericano Thomas Kuhn, para 1960, acuña el término paradigma, para definir y enmarcar el estado del arte o conocimiento teórico en el que se debe apoyar toda investigación para conocer los patrones o modelos de los que se parte en el proceso de la investigación, para reconfirmar el paradigma o bien generar un nuevo paradigma, patrón o modelo; la inconmensurabilidad de los paradigmas requieren de un lenguaje teórico común para poder comparar dos teorías o paradigmas y ver cuál de ellas es mejor. Cuando las teorías son inconmensurables no se puede decidir por la mejor de ellas, por ejemplo, la base teórica o paradigmas que apoyan a la medicina alópata es totalmente diferente a las bases teóricas de la medicina homeopática, por lo que es difícil decidir cuál es mejor para tratar ciertas enfermedades. El alemán Martin Heidegger, para mediados de los noventa aporta con su pensamiento del ser y el tiempo, seguido de las aportaciones del polaco Kurt Lewin, quien aporta los lineamientos de la Investigación Acción Participativa, como técnica de investigación para avanzar en las investigaciones cualitativas de cambio social, invitando a la población objeto de estudio a ser parte de la generación de las ideas del cambio social planeado, organizado y con inclusión de todos los actores sociales de una comunidad para generar empoderamiento del crecimiento y desarrollo de las comunidades. El Alemán Gadamer Hans-Geor, aporta a la ciencia con el principio de verdad y método universal, funda la escuela hermenéutica, como base para avanzar en la ciencia, primero se debe tener las habilidades para la interpretación de los textos y aportaciones a la ciencia que han dejado los pensadores durante el paso por esta existencia. La teoría social se ve fortalecida con la aportación del alemán Jurgen Habermas, quien manifestó que la teoría social y de acción comunicativa junto con la democracia deliberativa y ética discursiva

contribuye en la generación del conocimiento y al avance en la ciencia. El Alemán Albert Einstein, con sus pensamientos del tiempo, el espacio, energía y masa, proporciona la ley de la relatividad y se genera nueva era del avance de la ciencia, dejando inconclusa la idea del todo, conocimiento que fue desarrollado por el Inglés Stephen Hawking, con su aportación sobre el espacio-tiempo-energía-masa, llegó a la aportación de la teoría del todo y la radiación de los agujeros negros como fuente de la energía que da la vida. Carl Hempel, Alemán naturalizado estadounidense, nos aportó su pensamiento nomológico deductivo de las ciencias o explicaciones nomológicas deductivas con fundamentos teóricos de leyes o principios universales, que permiten generar hipótesis que pueden ser contrastables con el conocimiento empírico (La observación y la experimentación); fue positivista lógico, toda explicación necesita de una construcción teórica y fundamento estadístico-matemático. Finalmente, en este recorrido histórico se encuentra el argentino Mario Bunge, con su realismo científico y filosofía exacta con el método inductivo/deductivo, sistémico, hipotético y cuantitativo, como método científico para contrastar y generar conocimiento que lleven al avance en la ciencia. Ver gráfica número 8, 9 y 10.

Gráfica No. 8

### Historia de la Ciencia III: Comte, Durkheim, Schopenhauer, C. Max, Weber, Semmelweis, Russell y Pavlov



Fuente: aportación del autor

Gráfica No. 9

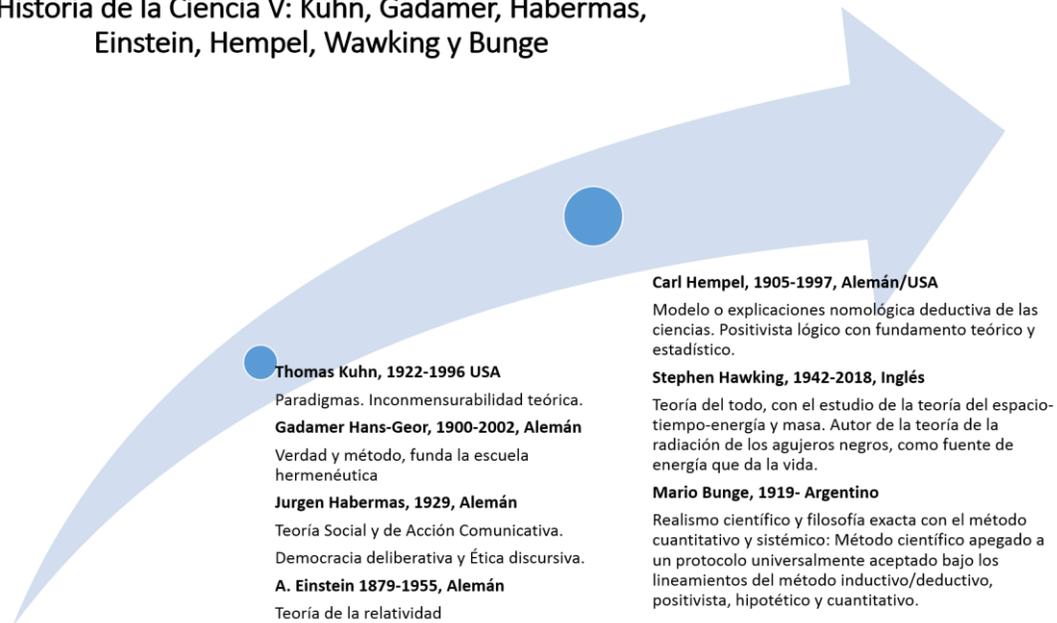
### Historia de la Ciencia IV: Watson, Skinner, Piaget, Vigotsky, Heidegger, Lewin, Popper, Feyerabend y Lakatos



Fuente: aportación del autor.

Gráfica No. 10

## Historia de la Ciencia V: Kuhn, Gadamer, Habermas, Einstein, Hempel, Wawking y Bunge



Fuente: aportación del autor.

## **Conclusión**

El método de investigación más utilizado en las ciencias sociales es precisamente el método inductivo/deductivo, hipotético, cuantitativo y fundamentado con teorías o conceptos, que ha ido evolucionado con el paso del tiempo a partir del pensamiento aristotélico. Inductivo, porque se puede partir de la observación o de lo particular para generalizar a toda una población o bien deductivo que parte de lo general a lo particular gracias a las fuentes secundarias de información que ayudan a la integración de ideas, conceptos, teorías, principios o leyes universales que permiten ir de lo general a lo particular.

La mayoría de los estudios de investigación en estos tiempos integran el método inductivo/deductivo, con enunciados hipotéticos sujetos a comprobación o contrastación con modelos cuantitativos y apoyados con las teorías, leyes o conceptos universalmente aceptados.

Se comparte todo un análisis histórico de como se ha ido desarrollando el pensamiento científico desde Aristóteles hasta Mario Bunge. Evolución del pensamiento desde la parte de la investigación empírica gracias a la observación, el registro y generación de datos que llevan a construir información y conocimiento. Es notorio encontrar el momento histórico en que algunos paradigmas se fueron echando abajo por científicos de la época, como la idea geocéntrica de Aristóteles que duró 1600 años, hasta la teoría Heliocéntrica de Copérnico y Galileo Galilei. A raíz de esas experiencias, la comunidad científica logró quitarle el poder a la iglesia sobre la investigación, ya que la ciencia debe avanzar de manera libre y sin atadura pública, privada o eclesiástica.

Se puede observar también la evolución del pensamiento de la investigación de la parte idealista y subjetiva hasta llegar a una postura objetiva y un monismo que ha permitido el avance de la ciencia apoyada con fundamento teórico, estadístico y matemático. En los últimos años se han visto unas corrientes de la investigación cualitativa que ha estado dando valor a los nuevos pensamientos para el diseño y desarrollo de nuevas técnicas de investigación.

**Palabras clave**

Evolución histórica del pensamiento científico

Método de investigación inductivo

Método de investigación deductivo

Empirismo

Método hipotético y cuantitativo

**Preguntas**

¿Qué tanto ha frenado la Iglesia el desarrollo de la Ciencia?

¿Ha estado subordinado el pensamiento científico por el poder de la Iglesia?

¿Lo científico en estos tiempos está subordinado al poder público?

¿Lo científico está subordinado al poder económico en estos tiempos?

¿Cómo aplicaría el método científico inductivo, deductivo, hipotético y cuantitativo en su proyecto de investigación?

## II. METODOLOGÍA DE LA INVESTIGACIÓN EN UNA PÁGINA

El formato de investigación en una página, fue diseñado para integrar el proceso de la investigación en una página para facilitar el proceso de la investigación y permitir limpieza en todo proyecto de investigación, Pérez (2004) integró el proceso de la investigación en el siguiente formato con el propósito de comprender las variables de investigación, el planteamiento del problema, la formulación de los objetivos, hipótesis, diseño de instrumentos de medición y técnicas estadísticas a realizar para contrastar las diversas hipótesis que se formulan en todo proyecto de investigación. Ver el formato en la tabla número 1.

Tabla No. 1

### INVESTIGACIÓN EN UNA PÁGINA

IDEA Y TEMA 		VARIABLES	DEFINIR VARIABLES	MENSURABILIDAD	HIPÓTESIS	TÉCNICAS DE INVESTIGACIÓN	MUESTRA/CAMPO	DISEÑO ESTADÍSTICO 
Fuentes de información	Primarias Secundarias	Dependientes	Teoría	Escalas	•Ho •Ha	Cuestionario	Marco Muestral	DESCRIBIR:
Tipos de investigación	Cualitativa/Cuantitativa	Independientes	Conceptos	Medición: y ↑ •X <sub>1</sub> → •X <sub>2</sub>	“Tantas hipótesis como variables dependientes e independientes” 	Observación	Selección muestra	Centralidad
	Exploratoria	Extrañas 	Fundamentos 			Nominal	Entrevistas	Procedimiento muestral
Planteamiento del problema	Descriptiva/Causal			Ordinal	“Tantas hipótesis como variables dependientes e independientes” 	Sesiones de grupos	Trabajo de campo	INFERIR:
	Según el tiempo: Transversal/Longitudinal			Intervalar		Técnicas proyectiva	Representatividad de la muestra.	Paramétricas
Objetivos	Unidad de análisis			Razón 	“Tantas hipótesis como variables dependientes e independientes” 	Experimentos		No paramétricas
Hipótesis	Variables: VD, V.I. (s) y extrañas			Confianza 		Paneles		Correlación
Diseño de la investigación	Espacio y tiempo					Validéz		Regresión
Cronograma de actividades/presupuesto						Confabilidad 		Ji cuadrada
<b>CONCLUSIONES</b>								
ANOVA Factor análisis Análisis Discriminante								

Fuente: Pérez, L., Marketing Social, Teoría y Práctica, Pearson Prentice Hall, 2004

## 1. Columna de Idea y Tema:

Este formato e investigación en una página, se integra en ocho columnas e inicia con la idea de la investigación y que se espera evolucione al desarrolla del tema; para ello se requiere consultar las fuentes primarias de investigación con el apoyo de la observación, cuestionarios, entrevistas de profundidad y otras técnicas. Una vez compilada las fuentes de información secundaria o todo lo escrito sobre el tema y publicado en diversos medios como libros, artículos, ensayos, bancos de datos de información de diversas fuentes que deberán ser citadas en cada momento que se utilicen:

### a. Fuentes de información:

- a. Primarias: es la información que se obtiene de manera directa de la población objeto de estudio, también conocida como información de primera mano. Para ello es importante definir el **origen de los datos**, esto es, describir de manera precisa el lugar de la obtención de los datos; **definir el formato para el registro, tabulación y análisis** de los datos.
- b. Secundarias: es la compilación de todo lo escrito y publicado sobre el tema a investigar, se incluye los estados del arte o marcos teóricos, resultados de investigaciones previas, publicaciones internas de las empresas, publicaciones en hemerotecas, revistas, libros y ensayos entre otros.

### b. Tipos de investigación:

- a. Cualitativa: técnicas de investigación de primera mano, con apoyo de técnicas de observación natural, directa, estructurada, no estructurada; entrevistas de profundidad con formatos abiertos o formatos estructurados con preguntas abiertas, sesiones de grupo y la Investigación Acción Participativa entre otras.
- b. Cuantitativa: técnica de investigación que requiere de un formato estructurado para observar o para aplicar un cuestionario para registrar, tabular y codificar la información adquirida de la población objeto de estudio.

- c. Exploratoria: es cuando se realiza la investigación como su palabra lo dice para explorar el terreno de investigación y ver si hay méritos para seguir profundizando en la misma. Se utilizan formatos cualitativos como cuestionario con preguntas abiertas para que el investigado conteste lo que el considere pertinente para la investigación; se realizan entrevistas de profundidad para sacar la información que ayude a formular el planteamiento del problema, las hipótesis y pautas para el diseño de la investigación. En este apartado también se compilan todas las fuentes secundarias de información para darle fundamento a los objetivos e hipótesis de investigación.
- d. Descriptiva: este tipo de investigación es conocida como cuantitativa, porque se apoya de instrumentos de medición estructurados tipo cuestionario con preguntas cerradas y observación con formato estructurado. Se pretende ya no explorar en la investigación, sino el describir el comportamiento de lo observado y registrado para tomar decisiones o bien dar paso al siguiente tipo de investigación conocido como causal. En este tipo de investigación se parte de muestras representativas para realizar las estadísticas descriptivas e inferenciales para contrastar las hipótesis de investigación. La parte de consulta de las fuentes secundarias de información en la fase exploratoria es necesaria en este tipo de investigación para darle fundamento teórico, conceptual al planteamiento del problema, los objetivos y las hipótesis de investigación formuladas.

- e. Causal: si con el tipo de investigación descriptiva no se puede tomar decisiones, se recomienda realizar el tipo de investigación causal o experimental. En este apartado se requiere el diseño de experimentos que permitan tener control de los efectos de la variable dependiente en función de la manipulación de las variables independientes. Ya con la información de las dos fases anteriores, esto es, exploratoria y descriptiva, se cuenta con suficiente información estadística y teórica para realizar el experimento que lleven a encontrar la relación causal. Este tipo de investigación requiere de más recursos de tiempo, económico, técnicos y humanos, entre otros.
- f. Según el tiempo las investigaciones se pueden clasificar en estudios transversales, cuando se realiza un corte en el tiempo con la aplicación de un cuestionario o de un formato de observación estructurada que registra hechos de un momento dado y bien definido, obteniéndose de esta manera información del pasado y no del presente y menos del futuro. La otra investigación en función del tiempo es conocida como estudios de tipo longitudinal o estudios de cohortes, esto es, permite dar seguimiento con el paso del tiempo al comportamiento de las variables estudiadas en la población objeto de estudio. Es útil para proyectar el comportamiento futuro de la variable dependiente en función de una serie de variables independientes.

- c. Planteamiento del problema: **es la pregunta** que moverá todo el proceso de la investigación, se recomienda formularla en forma de pregunta, para ello se requiere definir de manera detallada a la **unidad de análisis o población objeto de estudio**, definir de manera específica las **variables dependientes** (de las que se espera obtener un efecto, las más conocidas son ventas, imagen de la empresa y desempeño, ubicadas en el eje “Y” del plano cartesiano), las **variables independientes** o variable que pueden ser manipuladas por el investigador, como la dosis de un medicamento, el tipo de oferta y precio en marketing, entre otras, se ubican en el eje “X” dentro del plano cartesiano. También es importante en este apartado definir las **variables extrañas** o variables del entorno que pueden influir en el desempeño de variable dependiente, como la situación política, legal, cultural, económica y cultura, entre otras y también se ubican en el eje “X” del plano cartesiano. Todo planteamiento del problema debe contener de manera específica el **espacio** donde se realizará el trabajo de investigación y el **tiempo** en el que se ejecutará el proyecto de investigación.

- d. **Objetivos:** describe de manera clara y sin ambigüedades lo que pretende la investigación, por lo que deben ser medibles, alcanzables y retadores para el proyecto de investigación; se recomienda iniciar con un verbo en infinitivo, por ejemplo, analizar, identificar, formular, medir, mejorar, plantear, probar, verificar, señalar, describir, diagnosticar, explicar, diseñar, comparar, clasificar, buscar y sistematizar entre otros, con alto grado de coherencia con el planteamiento del problema, las hipótesis y lo que se quiere alcanzar como fruto de la investigación. Es recomendable formular objetivo general y algunos específicos, para darle limpieza y orden al proceso de la investigación misma. Son la guía de la investigación que orienta todos los esfuerzos para alcanzarlos. Los objetivos están determinados por los niveles de profundidad de la investigación; en investigaciones exploratorias los objetivos más comunes son diagnosticar, buscar, encontrar, etc. Para investigaciones descriptivas, se utilizan objetivos como relacionar, determinar, describir, mapear, etc. Y en proyectos causales se pretende encontrar las causas de variación de la variable dependiente, por lo que se plantea los objetivos con los verbos determinar, relacionar, explicar, estimar, etc.
- e. **Hipótesis:** son enunciados no verificados donde se integran las variables independientes con la dependiente, generalmente se formulan en término de **hipótesis nula** ( $H_0$ ) en forma de igualdad o en forma de negación, mientras que la **hipótesis alterna** ( $H_a$ ) se plantea en forma de desigualdad o bien como una afirmación, ejemplo,  $H_0$ : No hay diferencia significativa entre las ventas y el género;  $H_a$ : Si hay diferencia significativa entre las ventas y el género. En forma de igualdad la misma hipótesis se plantearía así:  $H_0$ : Las ventas son iguales entre hombres y mujeres:  $H_a$ : Las ventas son diferentes entre hombres y mujeres. ¿Cuántas hipótesis debe tener un proyecto de investigación? Tantas hipótesis como variables dependientes e independientes tengan en el proyecto de investigación, más las hipótesis de investigación del proyecto.

- f. Diseño de la Investigación: en este apartado se describe con detalle cada una de las fases que se realizarán para alcanzar los objetivos y resultados de la investigación. Con fines prácticos casi siempre se integran cinco fases:
1. Fase exploratoria y consulta de las fuentes secundarias de información.
  2. Fase descriptiva y diseño de los instrumentos de medición.
  3. Trabajo de campo, selección de la muestra y aplicación del cuestionario con criterios probabilísticos.
  4. Análisis y resultados univariado, bivariado y multivariado en función de las hipótesis de la investigación.
  5. Conclusiones y propuesta.
- g. Cronograma de actividades y presupuesto: el cronograma de actividades está directamente en función de las fases del proyecto de investigación, para poder definir tiempos y movimientos de cada fase, avances por fases y presupuesto o gasto que se ejecutará en cada una de las fases.

## **2. Columna de variables:**

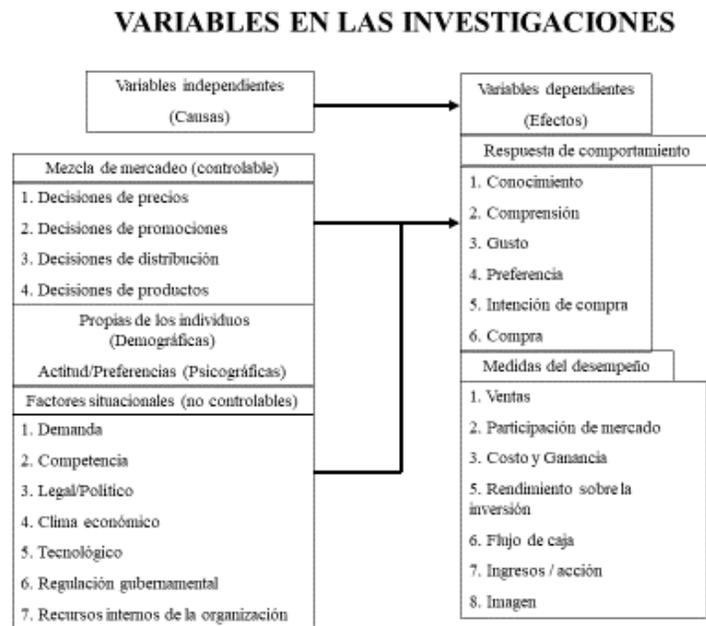
En este apartado se debe definir de manera específica las variables dependientes que recibirán el efecto ocasionado por la interacción con las variables independientes, casi siempre son las variables de desempeño como son las ventas en una empresa, el resultado esperado de una dieta alimenticia, el resultado esperado en función de unos medicamentos suministrados al paciente, entre otras. Estas variables se ubican en el eje Y dentro de un plano cartesiano, ya que se espera medir el efecto en función de las variables independientes y extrañas.

Las variables independientes son las variables que se pueden manipular, como es la dosis de un medicamento, las cuatro P del marketing (Producto, Precio, Plaza y Promoción), el proceso, la presentación y el personal que atiende una empresa de servicio, entre otras. Cada una de esas variables son fácilmente manipulables por los dueños de los negocios y se espera tener un mejor desempeño de la variable dependiente en función de la manipulación de estas variables. Siempre se ubican en el eje X dentro de un plano cartesiano.

Las variables extrañas, también se ubican en el eje X del plano cartesiano, pero con el inconveniente de que no son controlables por el investigador, pero que si pueden afectar el desempeño de la variable dependiente como son: la situación económica, social, política, cultural, demográfica y tecnológicas entre otras.

Es prioritario para toda investigación identificar las variables dependientes e independientes de todo proyecto de investigación, para comprender si realizarán análisis de dependencia, de relación, causales o de interdependencia, de ahí lo relevante el mapear las variables que integran todo proyecto de investigación, tal y como el Dr. Kinneer (1993) lo recomienda en su esquema donde relaciona las variables independientes controlables conocidas como las causas del efecto esperado en la variable dependiente, como es la mezcla del marketing, propias de los individuos y variables independientes no controlables o extrañas como la demanda, la competencia, el entorno económico, político y legal que influyen en la variable dependiente o efectos de respuesta del comportamiento como el grado de conocimiento, gusto y preferencia entre otras, así como las variables dependientes propias del desempeño de las empresas como las ventas, la imagen de la empresa y participación del mercado entre otras. Ver gráfica número 11.

Gráfica No. 11



Fuente: Adaptado de Kinnear Thomas et al, Investigación de Mercados, 4 Edición, McGraw-Hill, 1993, pp. 12-13

Otro grupo de variables de mayor uso en la investigación de mercados son las relacionadas con el comportamiento del consumidor, definiendo como variable dependiente la conducta de compra en función de los factores independientes de la mezcla del marketing como producto, precio, plaza y promoción; factor personales como la percepción del consumidor, su proceso de aprendizaje, factores psicológicos y de personalidad; factores racionales e inconscientes, como el uso de la razón en el momento de la toma de decisión, las sensaciones, las emociones, el proceso de toma de decisión inconsciente (neurociencia y neuromarketing) y la heurística del consumidor; factores externos, como políticos, legales, sociales, económicos, culturales y tecnológicos. Cada uno de estos factores y variables deben ser analizados dentro del esquema de cubos de información multidimensionales tipo OLAP (Online Analytical Processing-procesamiento analítico en línea) ya que todas las variables y datos interactúan casi de manera simultánea en el proceso de la toma de decisión. Ver gráfica número 12.

Gráfica No. 12

## Características que afectan la conducta del consumidor



- **Factores del marketing Mix:** Producto, Precio, Plaza y Promoción
- **Factores personales:** Percepción, Aprendizaje, psicológicos, Personalidad y rol en la compra: usuario, influenciador y comprador.
- **Factores racionales e inconscientes:** Racionales, Sensoriales, Emocionales, Inconscientes y Heurística del consumidor.
- **Factores externos:** Políticos, Legales, Sociales, Económicos, Culturales y Tecnológicos.

Fuente: Schiffman, Wisenblit; Comportamiento del consumidor; 11 Edición, Pearson Educación, México, 2015, pp 20-21

### **3. Columna para definir las variables:**

Una vez seleccionadas las variables que integran el planteamiento de la investigación y por ende el proyecto de investigación, se procede a la definición en función de un exhaustivo marco teórico y conceptual del proyecto de investigación, para evitar ambigüedad y poder proseguir con el proceso de la mensurabilidad de cada una de las variables del proyecto de investigación, en función de las escalas de medición. Es relevante en toda investigación proporcionar el fundamento necesario a cada una de las variables involucradas, para darle validez de contenido a las mismas.

### **4. Columna de mensurabilidad de las variables:**

Cada una de las variables del proyecto de investigación deben ser medidas con la escala de medición correspondiente para obtener la información que se requiere con alto grado de validez y confiabilidad de los instrumentos de medición; para ello existen cuatro escalas de medición, la nominal, ordinal, intervalar y razón, que se describen de manera detalla en el capítulo 3 de este libro, bajo el título escalas de medición e instrumentos de medición.

### **5. Columna de hipótesis:**

Son supuestos o afirmaciones sujetas a contrastación con las técnicas estadísticas correspondientes y marco teórico que fundamente la investigación. Las hipótesis se enuncian de manera nula y alterna, con los símbolos de  $H_0$  o hipótesis nula con enunciados en forma de igualdad o negación y  $H_a$  como hipótesis alterna en forma de desigualdad o afirmación; también se formulan como hipótesis de investigación ( $H_i$ ), altamente relacionadas con la  $H_a$ , esto es, algunos proyectos de investigación solo plantean la hipótesis de investigación ( $H_i$ ) acompañada de un exhaustivo marco teórico o conceptual; casi todo tratamiento estadístico inferencial (inferir de una muestra hacia la población) parte de hipótesis estadísticas, de ahí que los proyectos de investigación puede contener hipótesis nula, alterna, de investigación y estadística. Siendo la hipótesis de investigación la de mayor uso en los trabajos de investigación. En

varios proyectos de investigación se encontrarán los siguientes enunciados de hipótesis en función de las variables a analizar a la vez:

- a. Hipótesis univariada: es contrastar el enunciado de una variable en el proyecto de investigación, por ejemplo, el género y el estado civil entre otras:
  - i. Ho: No hay diferencia en el género de los clientes de la empresa.
  - ii. Ha: Si hay diferencia entre el género de los clientes de la empresa.
- b. Hipótesis bivariada: se da este enunciado hipotético, cuando se relacionan dos variables la vez, por ejemplo, la variable dependiente ventas y la relación con las independientes como edad, género, escolaridad y marketing mix, entre otras:
  - i. Ho: No hay diferencia significativa entre las ventas y el género de los clientes.  
Ha: Si hay diferencia significativa entre las ventas y el género de los clientes.
  - ii. Ho: No hay diferencia significativa entre las ventas y la edad de los clientes.  
Ha: Si hay diferencia significativa entre las ventas y la edad de los clientes.
  - iii. Ho: No hay diferencia significativa entre las ventas y el estado civil de los clientes.  
Ha: Si hay diferencia significativa entre las ventas y el estado civil de los clientes.
- c. Hipótesis multivariada de investigación: se presenta este enunciado hipotético, cuando se relaciona una variable dependiente contra varias independientes a la vez, como la relación que existe entre las variables del marketing mix y las ventas o la relación de las variables demográficas y las ventas, ejemplo:

- i. Ho: No hay diferencia significativa entre las ventas y la mezcla promocional de las mismas (publicidad, anaqueles en el punto de ventas, demostración y promociones en el punto de ventas)
- ii. Ha: Si hay diferencia significativa entre las ventas y la mezcla promocional de las mismas (publicidad, anaqueles en el punto de ventas, demostración y promociones en el punto de venta.

## **6. Columna de Técnicas de Investigación:**

Se describe el diseño de las técnicas de investigación cuantitativa y cualitativa a utilizar en el proyecto de investigación con sus respectivos formatos estructurados o no estructurados para la adquisición de los datos. Se parte del diseño de la investigación donde se describe de manera general las técnicas de investigación a utilizar en el proyecto de investigación. Las técnicas de investigación de mayor uso son:

- a. Cuestionario: documento con formato preestablecido para que los encuestados puedan dar respuesta a cada una de las preguntas del mismo. Pueden tener preguntas abiertas para que el encuestado conteste lo que considere pertinente o bien cerradas para que la respuesta quede condicionada a un formato de respuestas dicotómicas o de opción múltiple. Todo cuestionario debe pasar la prueba de validez y confiabilidad, para estar seguros de que se está midiendo lo que se quiere medir con las preguntas y escalas de medición adecuadas. Más adelante se detallará el proceso para la construcción de un cuestionario.
  - a. Observación directa: aquella que se realiza de manera directa con la unidad de análisis de investigación o población objeto de estudio; generalmente el investigador se integra a la población objeto de estudio como si fuera un miembro más de la comunidad, para poder tener evidencia directa de los hechos observados y registrarlos de primera mano.

- b. Observación indirecta: es aquella en la que el investigador observa desde lejos a la población objeto de estudio sin integrarse a los mismos, un poco parecido como el observar la feria de los toros desde las gradas y no dentro del ruedo taurino.
  - c. Observación estructurada y no estructurada: la estructurada requiere de un formato con los datos precisos a ser observados, medidos y registrados por un periodo de tiempo y espacio previamente seleccionados; mientras que la no estructurada, es cuando el investigador acude a observar la población objeto de estudio sin formato alguno, partiendo del principio de que todo vale y tiene libertad de registrar todo aquello que considere relevante para el proyecto de investigación.
  - d. Observación mecánica y humana: es la que se realiza con el apoyo de cámaras de videos fijos o manipulados por personal que realiza directamente a investigación y humana es cuando se realiza de manera directa sin apoyo de equipos de piso galvanómetros, de filmación y/o grabación.
- b. Entrevista de profundidad: es una herramienta de investigación cualitativa que se realiza cara a cara con preguntas abiertas para que la persona entrevistada pueda contestar con libertad a cada una de las preguntas genéricas y algunas específicas para conocer la opinión y postura del entrevistado. Se sugiere llevar un formato con 3 a 5 preguntas abiertas, dejando las más difíciles para el final de la entrevista. Las primeras preguntas son para romper el hielo e invitar a que el entrevistado en un ambiente relajado pueda compartir sus ideas y pensamientos alrededor de los temas a investigar.

- c. Sesiones de grupos: es también una técnica de investigación cualitativa conocida como entrevistas de profundidad pero en grupos de personas entre 6 a 12 para que se dé la sinergia que normalmente se da cuando se habla en grupo, ya que las ideas de uno de los integrantes de la sesión de grupo puede ayudar a que otros participen bien sea para estar a favor o en contra y llevar al grupo a tratar de llegar a un consenso o bien dejar clara las posturas tomadas durante el desarrollo de la sesión de grupo. Se recomienda desarrollar la guía de la sesión con un número recomendable de 5 a 8 preguntas, debidamente integradas y de forma coherentes para ir desarrollando la sesión durante un periodo recomendable de tiempo de 1 a 2 horas; existen sesiones que duran 24 a 48 horas en algunos lugares de retiro donde se invitan a los participan con alimentación, alojamiento y viáticos. Los pasos para una sesión de grupo más recomendable son: definir el problema y los objetivos, diseñar el formato de la sesión de grupo, diseñar la guía de la sesión con las 5 a 8 preguntas, definir al moderador, seleccionar a los integrantes de la sesión de grupo según la especificación demográfica propia para el proyecto de investigación, seleccionar a la persona que realizará la interpretación y preparación de los análisis y resultados para presentarlos a las personas interesadas.
- d. Técnicas proyectivas: es una técnica de la pregunta de un cuestionario que permite medir de manera indirecta lo que realmente se quiere medir, por ejemplo, la pregunta cuánto gana al mes puede ser una pregunta incómoda y difícil de que den la respuesta correcta, pero si se pregunta de manera indirecta como: Una persona de tu perfil ¿Cuánto cree usted que gane al mes en esa empresa?

- e. Investigación Acción Participativa: es una técnica de investigación cualitativa que integra al objeto de estudio como sujeto pensante y capaz de ser parte del desarrollo de la investigación, pero de manera participativa con otras personas que integran la unidad de análisis o de investigación. Se rompe la relación clásica de la investigación de sujeto-objeto, esto es, el objeto de investigación se integra al equipo de investigadores para ser parte del proceso de la IAP, desde el momento del planteamiento del problema social, fijación de los objetivos y formulación de planes de acción o estrategias a ejecutar para cambiar una realidad de la población objeto de estudio. Como su palabra lo indica Investigación, porque se apoya del método científico y técnicas de investigación universalmente aceptadas; la palabra acción es involucrar a los miembros del colectivo a ser parte de la transformación social de su realidad y participativa porque se debe tomar decisiones en consenso y con la participación de la mayoría de los afectados de una comunidad. A pesar de ser etiquetada en la parte cualitativa se pueden realizar técnicas de investigación cuantitativas para reforzar el conocimiento colectivo y orientar los pasos a seguir en la construcción permanente del conocimiento social en beneficio colectivo. De ahí que varias de las técnicas de investigación como la entrevista de profundidad, entrevista personal, uso de cuestionarios, la observación y técnicas de participación ciudadana para que se incluya en este proceso a todas las personas que formen parte de la solución del problema social.
- f. Criterios de validez y confiabilidad del instrumento de medición: en este tema se desarrollar en profundidad en el capítulo II de este libro.

## 7. Columna de muestra y trabajo de campo:

Seleccionar el tamaño de la muestra que sea representativa de la población y que permita realizar inferencias estadísticas, es todo un proceso que requiere primero identificar la unidad de análisis o elementos a estudiar con su respectivo perfil sociodemográfico e identificar el marco muestral o delimitación geográfica y de georreferenciación delimitada para extraer la muestra. La primera pregunta recomendable hacerse antes calcular el tamaño de la muestra es:

**¿A quién o quienes se van a investigar?** Para ello es importante describir de manera precisa y concreta las características sociodemográficas de las personas que se van a investigar; Y conociendo el perfil de las personas a investigar se procede con la siguiente pregunta: ¿La población es finita o infinita? Se considera población finita cuando son menos de quinientos mil personas e infinita cuando son mayores a quinientos mil. Las fórmulas matemáticas o estadísticas para la selección de la muestra están en función de si la población es de tamaño infinito o finito.

**¿Qué se conoce de la población?**, es otro de los interrogantes clave para la selección de la muestra, ya que cada elemento de medida de la unidad de análisis permite conocer si la escala de medición es numérica (intervalar o razón) o cualitativa (nominal u ordinal). Con estas medidas se puede realizar el cálculo de la media y la varianza o bien el cálculo de las proporciones para escalas cualitativas, para realizar las herramientas estadísticas necesarias de representatividad de la muestra hacia a población. También se debe describir si las variables a analizar en el tamaño de la muestra siguen un comportamiento de la distribución normal o de la binomial, fundamentales para definir las medidas de centralidad o dispersión de la población para la selección de la muestra o la proporción de la población.

**¿Cuántas variables se tienen contempladas en el cuestionario?** Existe una regla práctica y ya escrita que dice que el tamaño de la muestra debe ser de 3 a 5 veces el número de preguntas del cuestionario, esto es: si el cuestionario tiene 50 preguntas, el tamaño de la muestra debe ser de 150 a 250 personas; casi

siempre se escoge el tamaño superior, siendo lo ideal 250 personas. Las razones de este esquema de selección del tamaño de la muestra están fundamentada en los tratamientos estadísticos multivariantes que requieren tamaños de muestras grandes y por el requisito de la normalidad de las variables de la muestra para poder inferir hacia la población.

***¿Cuántos son los recursos para seleccionar la muestra?*** Se tiene la creencia generalizada que una muestra grande es mejor que una pequeña, de ahí que si hay recursos humanos y económicos suficientes se podrá incrementar el tamaño de la muestra para estar más cerca de niveles de representatividad de la muestra hacia la población, por ejemplo: si por fórmula estadística se llega a un tamaño de muestra de 345 personas, pero se cuenta con más recursos para incrementar el tamaño de la muestra se sugiere el incremento de la misma.

***¿Qué nivel de confianza y error quieren manejar?*** Este nivel de confianza y error están en función de la escala de medición de la variable para la selección de la muestra. Si la variable es de escala superior (intervalar o razón) se maneja como estándar un nivel de confianza del 95 por ciento, para valores de desviaciones sigmas de la media conocidos como valores “Z” de 1.96. El valor sigma “Z”, indica con exactitud qué tan alejada se encuentra una persona de la media. El valor 1.96, integra el 95% del total de la población, por lo que se dice: “Esta muestra tiene un nivel de confianza del 95% con un margen de error dado especificado por la dispersión de los valores y el margen de confianza. El error que se desea trabajar en todo tamaño de la muestra siempre se procura que sea pequeño, para variables con medidas de escalas superiores se define en función de la medida de la variable, por ejemplo, en centímetros, kilómetros, gramos, etc., para calcular el margen o intervalo de error en función del valor sigma de  $z = 1.96$  y la medida que permita identificar el error logre integrar en ese 95% el total de la población. Cuando la variable a medir es de proporciones o débiles, se procede a identificar el error en porcentajes, siendo el error del 3 o 5% el más utilizado en todas las fórmulas para seleccionar la muestra y con una Z igual a

1.96, para estar dentro del rango de confianza del 95% Ver gráfica número 13 y 14.

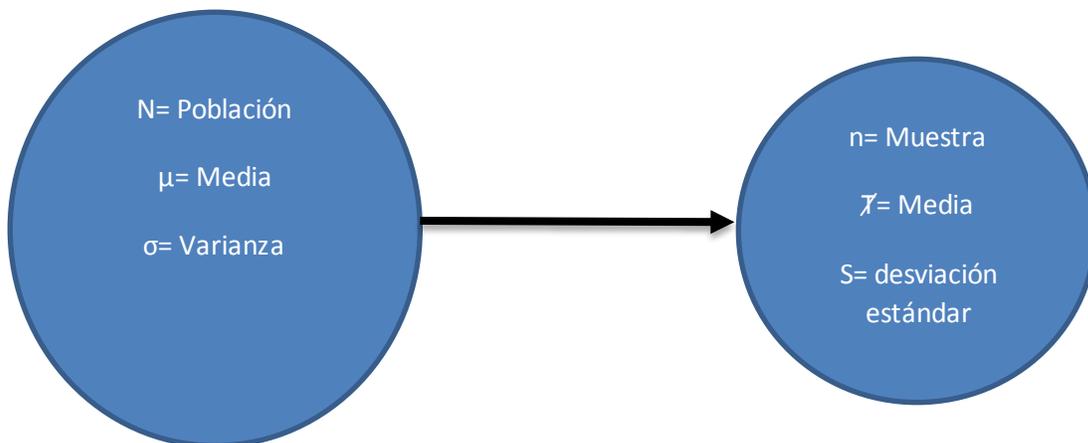
Gráfica No. 13

### MUESTRA

- La población es: {
  - finita
  - infinita
- Qué conocemos de la población:
  - La media y varianza
  - La proporción
- Distribución normal o binomial de la población.
- Cuantas variables tenemos contempladas en nuestro cuestionario.
- Cuántos son nuestros recursos:
  - Humanos
  - Económicos
- Qué nivel de confianza y error queremos manejar.

Fuente: Aportación del autor.

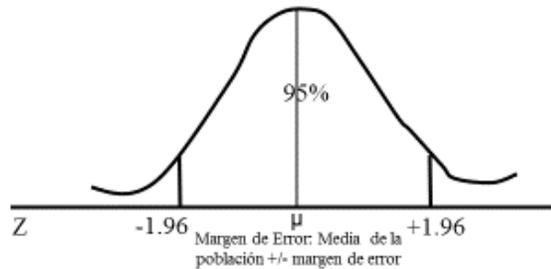
Es importante conocer los símbolos de la población y de la muestra, por ejemplo la población se conoce con el símbolo de  $N$ , la muestra con  $n$ , los símbolos de medidas de centralidad y dispersión de la población son la media identificada con la  $\mu$ , la varianza de la población identificada con el símbolo  $\sigma$ , mientras que en la muestra los símbolos de la desviación estándar es  $S$  y la media es  $\bar{X}$ .



Gráfica No. 14

## DETERMINAR TAMAÑO DE LA MUESTRA

### •MARGEN DE ERROR



$$Z = \frac{X - \mu}{\sigma}$$

$$E = Z \frac{\sigma}{\sqrt{n}}$$

$$E = Z \sqrt{P(1-P)/n}$$

Fuente: aportación del autor.

En la gráfica 10, se puede observar los principios básicos para la creación de las fórmulas estadísticas para la selección de la muestra:

Z = valor sigma en función de la diferencia de un valor de la población (X) menos la media de la población ( $\mu$ ) entre la varianza ( $\sigma$ ), que representa la distancia en relación a la media que ocupa ese valor X. En otras palabras, muestra que tan distante de la media se encuentra el valor X. Ejemplo, en una población se puede tener una media de estatura de 172 centímetros y se quiere saber a qué distancia en relación a la media se encuentra el dato de Luis, con estatura de 190 centímetros ( $X=190$ ), considerando una varianza de 10 centímetros ( $\sigma=10$ ), reemplazando en la fórmula se tiene:

$$Z = X - \mu / \sigma$$

X = Valor de un dato de la población

$\mu$  = La media de la población

$\sigma$  = Varianza

$$Z = 190 - 172 / 10$$

$$Z = +1.8$$

En la tabla de la distribución normal, el valor  $Z = 1.8$ , arroja el 46.41%, por lo que Luis, se encuentra arriba de la media con +1.8 desviación estándar, y a partir de ese dato, el 92.82% ( $46.41+46.41=92.82\%$ ) de los individuos se encuentran por abajo de la estatura de Luis y solo el 7.18% de la población se encontrará arriba de la estatura de Luis.

El valor  $Z$  representa la distancia sigma de cada uno de los datos de la población en relación a la media ( $\mu$ ) u la varianza ( $\sigma$ ) de la población, generando con esta información la desviación estandarizada de la población conocida como varianza o valor que refleja la distancia total de todos los datos en relación a la media poblacional. Un valor  $Z$  de -1.96 en valores de la tabla de la distribución normal arroja 47.72% y +1.96, el otro 47.72, para un valor de 95.44 %, redondeando a 95% del total de los datos se encontrarán dentro de la media y el margen de error para ese valor  $Z$ , por lo que el nivel de confianza para seleccionar la muestra es del 95%.

Con esta fórmula  $Z = \frac{X - \mu}{\sigma}$ , se puede calcular el valor  $X$ , cuando se conoce el valor  $Z$ , solamente se tiene que despejar:  $X = Z\sigma + \mu$  y de esta manera se podrá conocer el valor de uno de los datos que integran la población.

- **Fórmulas para la selección del tamaño de la muestra:**

Las fórmulas estadísticas de mayor uso para seleccionar el tamaño de la muestra (n), están en función de conocer el tamaño de la población (N), la media de la población ( $\mu$ ), la desviación estándar ( $\sigma$ ), el error (E), el nivel de confianza (Z) y/o la proporción de la población (P).

n = Tamaño de la muestra

N = Total de la población.

$\sigma$  = Varianza o desviación estándar de la población

Z = Nivel de confianza, 1.96 (95% de confianza)

E = Error de la muestra en la misma medida que el valor de la varianza o en porcentaje cuando se trabaja con la proporción (P)

P= Proporción de la población que tiene el valor esperado, por ejemplo, porcentaje de personas que tienen sobrepeso, o porcentaje de estudiantes con promedio superior a 9.5, etc.

Las fórmulas para seleccionar una muestra emanan del análisis de la distribución normal, con los datos de la desviación estándar (s) para la muestra o bien con la varianza ( $\sigma$ ) cuando se trata de la población, valores Z para identificar el nivel de confianza deseado, el cual casi siempre es de 1.96 (equivale al 95% de confianza), la media y el margen de error:

$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{E}{\sigma}$$

El error (E) se presenta dentro del marco de las desviaciones estándar de cada muestra que se saque; cada muestra tendrá media y desviación estándar diferente, de ahí que la fórmula  $\sigma_x$  (Desviación estándar de las medias o error estándar de las medias) es igual a la desviación estándar de la muestra (S) entre la raíz cuadrada del total de la población ( $\sigma_x = S/\sqrt{n}$ )

Se eleva al cuadrado para eliminar el efecto de la raíz cuadrada y se despeja para encontrar una de las fórmulas para la selección de la muestra:

$$\sigma_x = \sigma/\sqrt{n}$$

$$E = (\sigma/\sqrt{n})$$

$$E = (\sigma/\sqrt{n}) * z$$

$$(E)^2 = (\sigma/\sqrt{n})^2 * Z^2$$

$$E^2 = S^2/n * z^2$$

$$n = Z^2 * \sigma^2 / E^2$$

Despejando quedaría la primera fórmula para la selección de la muestra:

$$n = Z^2 * \sigma^2 / E^2$$

n = Tamaño de la muestra

$Z^2$  = Nivel de confianza deseado para inferir de la muestra a la población. Se ha tomado como habitual utilizar una Z de 1.96, que representa el 95% de confianza, dentro de una distribución normal.

$\sigma^2$  = Desviación estándar de la población la cual se conoce con el símbolo de  $\sigma$  y representa que tan alejados están todos los datos de un punto medio.

$E^2$  = Error al cuadrado, es esa diferencia entre las medias que estaría dispuesto el investigador a aceptar dentro de un intervalo de confianza del 95%

Para estimar el tamaño de la muestra cuando se conoce la proporción de la variable a medir, se procede de manera similar que cuando se conoce la media y la varianza en variables medidas en escalas superiores o escalares y/o de razón:

$$\sigma_p = \sqrt{P * (1-P)/n}$$

$\sigma_p$  = Error estándar de la proporción. Es un estimado de la desviación dentro de una distribución de muestras de una población con proporciones, como la proporción entre hombres y mujeres.

P = Es la proporción de la variable que se está midiendo. Puede ser sexo, si compraría o no, etc.

n = Es el tamaño de la muestra.

Elevando al cuadrado para quitar el efecto de la raíz cuadrado y despejando quedaría la fórmula para calcular el tamaño de la muestra cuando se conoce la

proporción de la población y se anexa el valor Z al nivel de confianza deseado por el investigador, siendo el de mayor uso el de 95% con un valor  $Z = 1.96$

$$\sigma_p = \sqrt{P * (1-P)/n}$$

$$E = \sqrt{P * (1-P)/n} * Z$$

Se eleva al cuadrado para eliminar el efecto de la raíz cuadrada:

$$E^2 = (\sqrt{P * (1-P) /n})^2 * Z^2$$

Se despeja y queda la fórmula cuando se conoce la proporción:

$$n = Z^2 P (1-P) /E^2$$

Ver gráfica número 15, 16, 17 y 18.

Gráfica No. 15

## DETERMINAR TAMAÑO DE LA MUESTRA

### •POBLACIÓN INFINITA

$$n = \frac{Z^2 \sigma^2}{E^2}$$

$$n = \frac{Z^2 P(1-P)}{E^2}$$

Fuente: Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986 pp.118-125

Como ejemplo para determinar el tamaño de la muestra con población infinita (mayor a quinientos mil personas o elementos de la población), se solicita calcular el tamaño de la muestra de la población de la clase media de la Ciudad de México, con ingresos mensuales de 15 mil pesos, varianza ( $\sigma$ ) de 300 pesos, nivel de confianza del 95%, nivel de error de 15 pesos y con el dato del censo de 18 millones de personas de la clase media que viven en la Ciudad de México:

$$n = Z^2 \times \sigma^2 / E^2$$

$$n = 1.96^2 \times 300^2 / 15^2$$

$$n = 1537$$

Cuando se conoce la proporción de la población se procede a utilizar la fórmula de proporciones, tanto el numerado como el denominador será en porcentajes, de ahí que el error en casi todo proyecto de investigación se fija entre el 3 y 5 por ciento (.03 o .05). Siguiendo el mismo ejemplo anterior con un censo de 18 millones de personas de la clase media que viven en la Ciudad de México, que representa el 60% del total de la población de la Ciudad, con un margen de error del 5% y nivel de confianza del 95%, calcular el tamaño de la muestra.

$$n = Z^2 P (1-P) / E^2$$

$$n = 1.96^2 (.60*.40) / .05^2$$

$$n = 369$$

Se puede ver que el tamaño de la muestra disminuye de manera considerable cuando se utiliza la fórmula de proporciones versus la de media y varianza que arrojan un tamaño de muestra de 1537 personas.

***¿Cuál fórmula se debe utilizar para seleccionar el tamaño de la muestra?***

Queda a criterio del investigador y a los diversos fundamentos estadísticos como el tamaño del error y nivel de confianza a utilizar (valor Z); Si el error disminuye, se incrementará el tamaño de la muestra de manera considerable, por ejemplo, el solo hecho de modificar el error de 5 al 3% en el ejercicio de proporciones, se incrementará el tamaño de la muestra a 1024 personas. Se debe considerar que a mayor tamaño de la muestra se tendrá mayor nivel de confianza y precisión de la muestra para inferir hacia la población total, por lo que es recomendable seleccionar el tamaño de la muestra por cualquiera de las fórmulas que se utilicen, pero si hay más recursos humanos y económicos para incrementar el tamaño de la muestra se sugiere incrementarla.

$$n = 1.96^2 (.60*.40) / .03^2$$

$$n = 1024$$

Cuando no se conoce la proporción de la población o valor P, se procede a utilizar la P de .5, que equivale al 50% de la población que puede tener la característica de la proporción, por ejemplo, se solicita seleccionar el tamaño de la muestra por proporciones de los ciudadanos de la clase media de la ciudad de México que esté de acuerdo con el sistema político mexicano de la actualidad. Al no tener estudios previos para definir la proporción se procede a utilizar un valor de P igual a .5 (50% de la población puede estar de acuerdo y la otra mitad en desacuerdo), ya que es justo con ese porcentaje que el tamaño de la muestra es más grande.

Gráfica No. 16

## DETERMINAR TAMAÑO DE LA MUESTRA

### •POBLACIÓN FINITA

$$n = \frac{\sigma^2}{\frac{E^2}{Z^2} + \frac{\sigma^2}{N}} \qquad n = \frac{P(1-P)}{\frac{E^2}{Z^2} + \frac{P(1-P)}{N}}$$

Fuente: Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986 pp.118-125

Ejemplos para comprender el uso de las fórmulas estadísticas para la selección de la muestra, con población finita (N=igual o menor a quinientos mil personas o elementos de la población). Se requiere calcular el tamaño de la muestra de los 27 mil estudiantes de la Facultad de Contaduría y Administración de la Universidad Nacional Autónoma de la Ciudad de México (UNAM), que vayan a hacer el uso del voto en las próximas elecciones para presidente de México. Como no se tiene la proporción se procede a utilizar un valor de P igual a .5, con un nivel de confianza del 95% y error del 5%.

$$n = P(1-P) / \left\{ \left( \frac{E^2}{Z^2} \right) + \left( \frac{P(1-P)}{N} \right) \right\}$$

$$n = .50(1-.50) / \left\{ \left( \frac{0.05^2}{1.96^2} \right) + \left( \frac{.50(1-.50)}{27000} \right) \right\}$$

$$n = 384$$

En este mismo ejemplo si el error baja al 3%, el tamaño de la muestra quedaría en 1067 estudiantes. No olvidar que la P puede variar en función de estudios previos, por ejemplo, se puede tener un estudio previo que solo el 20% de los estudiantes van hacer uso del voto para elegir Presidente del país, por lo que la P sería de .20; utilizando la misma fórmula se puede ver que el tamaño de la muestra disminuye a 244 estudiantes, solamente modificando el valor de P conservando el error del 5%, el nivel de confianza de 95% y el tamaño de la población de 27 mil estudiantes:

$$n = .20 (1-.20) / \{(0.05^2/1.96^2) + (.20 (1-.20) /27000)\}$$

$$n = 0.16/.00065077 + .16/27000$$

$$n = 0.16/.00065077 + .000005925$$

$$n = .16/.000656695$$

$$n = 244$$

El ejemplo para población finita de una Delegación Municipal de la Ciudad de México, con una población o N igual a 430 mil ciudadanos mayores de edad de clase media que ganan 15 mil pesos al mes, con varianza de 300 y error de 15 pesos:

$$n = \sigma^2 / (E^2 / Z^2 + \sigma^2 / N)$$

$$n = 300^2 / (15^2 / 1.96^2 + 300^2 / 430000)$$

$$n = 1531$$

El muestreo estratificado desproporcional, considera la varianza en cada uno de los estratos a considerar para seleccionar la muestra. Primero se calcula el tamaño de n o de la muestra con una de las fórmulas ya mencionadas para proceder a calcular el tamaño de la muestra en cada uno de los estratos a considerar por el proyecto de investigación, por ejemplo, se desea calcular el tamaño de la muestra tomando en cuenta la varianza y el tamaño de la población en cada uno de los estratos; el estrato A tiene una varianza de 120 pesos de sus ingresos promedio y con una población (N) de 60 mil ciudadanos, el estrato B con una varianza de 80 y población de 120 mil y el estrato C con varianza de 20 y población de 300 mil ciudadanos. El total de la población es de 500 mil ciudadanos (60000+120000+300000= 500000)

Primero se procede a calcular el tamaño de la muestra n por proporciones para población finita por ser igual o menor a 500 mil ciudadanos, con error del 5%, nivel de confianza del 95%, proporción del 50% y población de 500 mil ciudadanos:

$$n = P (1-P) / \{(E^2 / Z^2) + (P (1-P) / N)\}$$

$$n = .50 (1-.50) / \{(0.05^2/1.96^2) + (.50 (1-.50)/500000)\}$$

$$n = 384$$

Ya con todos los valores se procede a calcular el tamaño de la muestra para cada uno de los estratos con el apoyo de la fórmula:

$$n_A = n \cdot N_A \cdot \sigma_A / (N_A \cdot \sigma_A + (N_B \cdot \sigma_B) + (N_C \cdot \sigma_C))$$

$$n_A = 384(60000 \cdot 120) / (60000 \cdot 120 + (120000 \cdot 80) + (300000 \cdot 20))$$

$$n_A = 121$$

$$n_B = 384(120000 \cdot 80) / (60000 \cdot 120 + (120000 \cdot 80) + (300000 \cdot 20))$$

$$n_B = 162$$

$$n_C = 384(300000 \cdot 20) / (60000 \cdot 120 + (120000 \cdot 80) + (300000 \cdot 20))$$

$$n_C = 101$$

El tamaño de la muestra para cada uno los tres estratos socioeconómicos tipificados como A, B y C, suman en total 384 (121+162+101), la diferencia se debe a la varianza y no tanto al tamaño (N) del total de individuos en cada estrato socioeconómico; en C con 101, es suficiente a pesar de tener 300 mil personas; en A es de 121 de tan solo 60 mil ciudadanos. Ver gráfica número 17.

Grafica No. 17

**MUESTRO ESTRATIFICADO  
DESpropORCIONAL**

$$n_A = \frac{n N_A \sigma_A}{(N_A \sigma_A + N_B \sigma_B + N_C \sigma_C + \dots)}$$

$$n_B = \frac{n N_B \sigma_B}{(N_A \sigma_A + N_B \sigma_B + N_C \sigma_C + \dots)}$$

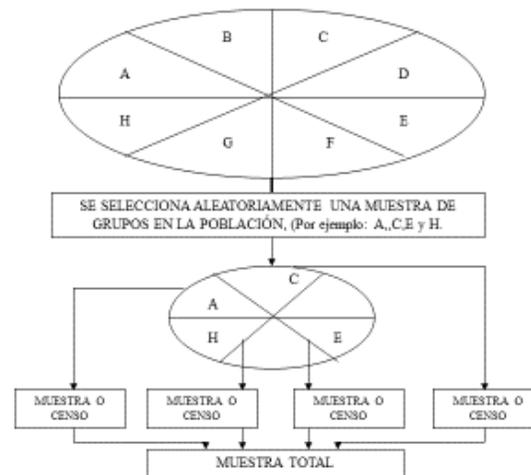
$$n_C = \frac{n N_C \sigma_C}{(N_A \sigma_A + N_B \sigma_B + N_C \sigma_C + \dots)}$$

Fuente: Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986 pp.118-125

El muestreo por conglomerados es útil para dividir los estratos en grupos homogéneos o parecidos, para de ahí seleccionar por el método aleatorio simple una muestra de los conglomerados y de esos sacar la muestra o un censo. Ver gráfica número 18.

Gráfica No. 18

### MUESTREO POR CONGLOMERADOS



A diferencia del muestreo estratificado, en el cual todos los estratos se muestrean, el muestreo por conglomerado supone un hecho con una muestra de los estratos de la población

Fuente: Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986

- **Método de muestreo probabilístico y no probabilístico**

*¿Cuál es el procedimiento para seleccionar la muestra de la población?* Hay dos métodos para seleccionar la muestra, el probabilístico, cuando se aplica la ley de la probabilidad para que todos los elementos de la población puedan tener la misma proporción de ser seleccionados en la muestra y los que no requieren de la probabilidad conocidos como no probabilísticos. Lo probabilísticos son los más preferidos debido a que se puede realizar con mayor confianza las estadísticas inferenciales de la muestra hacia la población y aplicar estadísticos de representatividad de la muestra hacia la población, mientras que los no probabilísticos tienen la limitante de inferir hacia la población por la dificultad de la representatividad de la muestra hacia la población

Los métodos probabilísticos de mayor uso son:

- Muestreo aleatorio simple: es cuando se conoce la probabilidad que tiene cada unidad de análisis de ser parte de la muestra. Por ejemplo, se introducen los 5000 clientes en una tómbola o espacio y de manera aleatoria se procede a sacar uno por uno hasta llegar a los 200. De esta manera la probabilidad de ser seleccionado cada uno de los 5000 datos es del 4% ( $200/5000= 0.04$ )
- Muestreo aleatorio sistemático: es similar al anterior pero con la diferencia que en este proceso se apoya con una lista de los 5000 datos y se procede sacar la muestra con saltos de cada 25 datos ( $5000/200= 25$ ); la persona que ocupe el lugar 25, 50, 75... formarán parte de la muestra hasta llegar a los 200 datos.

- Muestreo estratificado proporcional y desproporcional: es el método donde se divide la población por estratos socioeconómicos para seleccionar la muestra e inferir hacia la población, por ejemplo, sacar una muestra representativa de los estudiantes por nivel de ingresos de una universidad privada, clasificados en estrato social y económico A, B y C: en A hay 50 personas, en B 100 y en C 500, para un total de 650, de los cuales el 8% conforman el estrato A, el 15% lo integran los del estrato B y el 77% del estrato C. La parte proporcional, se refiere a seleccionar la muestra según el peso relativo en porcentaje para cada uno de los tres segmentos; Si la muestra es de 100 personas, se procede seleccionar 8 personas de A, 15 de B y 77 de C (Equivalente al 8%, 15% y 77% de la población en cada estrato), el desproporcional se refiere a considerar la varianza que se presenta en cada uno de los estratos, por lo que el tamaño de la muestra variará según las desviación de los datos, esto es, a mayor variación de los datos se debe sacar un tamaño de muestra mayor y al ser más homogéneos los estratos se requiere una muestra menor en esos estratos.
- Muestreo por conglomerados: es dividir la población en conglomerados homogéneos para de ahí seleccionar aleatoriamente unos cuantos conglomerados y de esos aplicar la muestra para realizar inferencias a todos los conglomerados. Es muy riesgoso, pero se apela a la probabilidad y la estadística para seleccionar la muestra.
- Muestreo por áreas: es parecido al de conglomerados y se refiere más a dividir la población por áreas geográficas y de ahí seleccionar aleatoriamente algunas áreas a las que se seleccionará la muestra.

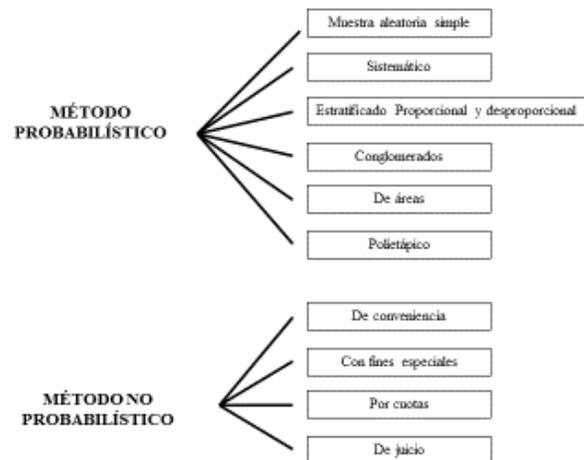
- Muestreo Polietápico: es el método donde se combina áreas, conglomerados, estratos grandes y demarcaciones geográficas pequeñas para inferir hacia la población, por ejemplo, sacar una muestra representativa de los ciudadanos del Estado de Aguascalientes, de México, conformado por 11 Municipios, donde cohabitan 1'200.000 habitantes; se procede a delimitar los 11 Municipios con el total de la población distribuida en cada uno de los Municipios, para seleccionar aleatoriamente solo 6 de los 11 para aplicar el cuestionario o instrumento de medición para inferir hacia toda población del Estado.

Los métodos no probabilísticos: son aquellos que no es requisito conocer la probabilidad de la muestra seleccionada de la población.

- De conveniencia para el investigador o la investigación: es un método que facilita la selección de la muestra según la conveniencia del investigador principal, por ejemplo, si el investigador es un médico, este seleccionará la muestra de sus propios pacientes o si es de un maestro, quien seleccionará la muestra de sus propios estudiantes; en el mundo empresarial se puede seleccionar la muestra de sus propios clientes o bien de sus propios empleados.
- Con fines especiales: se define los atributos de la muestra a ser seleccionada ya que persigue fines especiales, como pacientes de sexo femenino, con antecedentes de cáncer mamario, mayor de 45 años de edad y con antecedentes de haber tenido más de dos hijos.
- Por cuotas: se sacará la muestra según la proporción relativa de cada uno de los segmentos en los que se ha dividido la población, por ejemplo, seleccionar a 100 personas de los siguientes tres segmentos divididos en A, B y C, con los siguientes datos: seleccionar 8 personas de A, 15 de B y 77 de C (Equivalente al 8%, 15% y 77% de la población en cada estrato)
- De juicio: según el juicio del investigador, el decide que la muestra se seleccionará según su propio argumento, el cual no necesariamente debe coincidir con otras posturas de otros investigadores, pero se respeta el juicio y el argumento del investigador. Ver gráfica número 19.

Gráfica No. 19

## MÉTODOS DE MUESTREO PROBABILÍSTICO Y NO PROBABILÍSTICO



Fuente: Adaptada de Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986

**Trabajo de campo: ¿Cómo se realizará el trabajo de campo? ¿Quién lo realizará? ¿Qué nivel de supervisión se debe realizar al trabajo de campo?** El

trabajo de campo se considera como el granito de oro que dará forma al lingote. Es la forma de como adquirir los datos de manera ordenada, ética, transparente, limpia y honesta. Si la data está llena de errores, nunca se podrá tener información confiable para la toma de decisión. Para el campo cuantitativo con la aplicación de un cuestionario se requiere:

- Construir la estructura de cómo se debe realizar el campo: encuestadores con perfil definido para aplicar el cuestionario, entrenamiento del instrumento de medición, capacitación en el trato humano, capacitación técnica del proyecto de investigación, seleccionar al supervisor del campo, prueba piloto para definir el tiempo de duración del cuestionario y de cómo aplicar cada una de las preguntas.
- Gestión del trabajo de campo: se divide el grupo de encuestadores 5 a 8 por cada supervisor en el campo, se les asigna formato con tiempos definidos para aplicar el campo en sitios específicos y se entrega un formato para el control y evaluación del trabajo de campo.
- Supervisión y evaluación del trabajo de campo: se debe realizar monitoria in situ y después de realizar cada prueba esta se debe ver con detalle para ver si las respuestas del cuestionario están todas contestadas y buscar algunas inconsistencias en algunas de las preguntas filtros o preguntas claves para detectar errores de respuesta del mismo. En ese momento se rechazará o se aceptará el cuestionario como bien aplicado y formará parte de la base de datos para codificar y tabular. Los cuestionarios rechazados por el supervisor se deben realizar a otros clientes que tengan el perfil de la muestra deseada y sugeridos por el mismo supervisor del campo.

## **8. Columna de diseño estadístico:**

El diseño estadístico está en función de múltiples factores debidamente interconectados entre sí, como son: comprender las variables dependientes e independientes del proyecto de investigación, las escalas de medición de cada una de las variables, los objetivos, las hipótesis y tomar la decisión de realizar análisis estadístico descriptivo e inferencial univariado, bivariado y multivariado. Este apartado es la esencia del libro y se explicará de manera detallada en el desarrollo de cada técnica estadística. Es importante tener claro en este apartado de la investigación que es recomendable en todo proyecto realizar análisis:

- a. Univariado: analizar cada una de las variables para describir cómo se comporta por sí sola (análisis descriptivo), por ejemplo, en la variable sexo se necesita saber la frecuencia, el porcentajes, gráficas y la razón entre hombres y mujeres. En este mismo análisis univariado se debe realizar las pruebas estadísticas para inferir el comportamiento de la variable analizada hacia la población, conocida como estadística inferencial, por lo que se formularán hipótesis univariadas a ser contrastadas con pruebas estadísticas como la Chi Cuadrada y la T de Student entre otras
- b. Bivariada: análisis de la interacción de dos variables al mismo tiempo que dan origen a las hipótesis bivariadas, por ejemplo, Ho. No hay diferencia significativa entre el sexo y el volumen de compra dentro de una tienda de abarrotes. Para contrastar esta hipótesis se necesita de estadísticos inferenciales como la Chi Cuadrada y la T de Student. Las pruebas descriptivas e inferenciales se realizan dentro de la misma prueba estadística seleccionada, por lo que facilita el proceso del análisis.
- c. Multivariado: análisis de la interacción de varias variables en un mismo momento para encontrar relación entre las interacción de tipo causa y efecto o bien para encontrar agrupaciones o interdependencia de las variables, dando origen a las hipótesis

multivariadas, por ejemplo, la relación de causa y efecto de la variable dependiente ventas con las modificaciones de los medios de promoción y comunicación, diseño del producto, cambios en el precio y servicio al cliente entre otras (variables independientes), dando origen a la hipótesis multivariada de tipo nula y alterna :  $H_0$ . No hay diferencia significativa entre la variable ventas y los cambios en la promoción y comunicación, diseño del producto, cambios en el precio y servicio al cliente;  $H_a$ . Si hay diferencia significativa entre la variable ventas y los cambios en la promoción y comunicación, diseño del producto, cambios en el precio y servicio al cliente. En este tipo de hipótesis se pretende encontrar el orden y el grado de importancia en que incide cada una de esas variables independientes en la variable dependiente ventas.

Tener en mente los tres tipos de análisis estadísticos univariado, bivariado y multivariado en el diseño estadístico dará limpieza y orden desde el inicio del proyecto de investigación. Desde el inicio del proyecto ya se debe identificar las pruebas estadísticas a realizar, las cuales están cien por ciento relacionadas a las hipótesis planteadas y a los objetivos a alcanzar del proyecto de investigación. No se necesita ser experto en estadística para decidir que técnicas o pruebas descriptivas e inferenciales se deben realizar, es necesario el conocimiento de las escalas de medición y las funciones básicas de la estadística descriptiva e inferencial.

## Conclusión

Para una mejor comprensión del proceso de la investigación que ayuda alejarse los errores del proceso y de muestra, este escrito proporciona la metodología de la investigación en una página, de manera dinámica y de sencilla comprensión en un formato con ocho columnas iniciando con la idea, los tipos de información a utilizar en todo proyecto de investigación, los niveles de profundidad de la investigación desde la etapa exploratoria, con mayor apoyo de las fuentes secundarias de información y estudios cualitativos como entrevistas de profundidad, observación no estructurada y los estudios descriptivos con mayor apoyo de cuestionarios y la observación estructurada que permite realizar las pruebas estadísticas necesarias para presentar resultados y causales o experimentales que permiten encontrar las causas de los efectos de la variable dependiente. En función del tiempo las investigaciones se dividen en longitudinales y transversales, siendo las de mayor uso los estudios transversales ya que siempre se apoyan en cuestionarios estructurados que se aplican solo una vez a una población dada. El planteamiento del problema es fundamental en esta primera columna del proceso de la investigación en una página ya que es la pregunta detonadora que dará pauta a los objetivos e hipótesis de la investigación y a los diseños de metodología y estadísticas a utilizar para contrastar las hipótesis planteadas. En la columna 2 se puede visualizar el nombre preciso y sin ambigüedades de cada una de las variables involucradas en el proyecto de investigación, para pasar a la columna 3 donde se describe de manera clara y con el fundamento necesario cada una de las variables considerada en el proyecto. En la columna 4 se puede ver el cómo se van a medir cada una de esas variables en función de las cuatro escalas de medición: nominal, ordinal, intervalar y razón.

En la columna 5 se puede ver los diferentes tipos de hipótesis en función de las variables dependientes e independientes del proyecto y si realizarán análisis inferenciales univariado, bivariado y multivariado.

En la columna 6 se desarrollan las técnicas de investigación con los tipos de observación y cuestionarios.

En la columna 7 se trabaja la muestra y trabajo de campo, desde las fórmulas para la selección de la muestra, hasta técnicas estadísticas para la representatividad de la muestra hacia la población.

En la columna 8 se trabaja los diseños estadísticos a utilizar en función de las variables, sus medidas, los objetivos y las hipótesis planteadas.

**Palabras Clave:**

Metodología de la investigación en una página

Hipótesis

Selección de la muestra

Muestreo probabilístico y no probabilístico

Representatividad de la muestra

**Preguntas:**

Hay dos tipos de errores presentes en toda investigación: el error de sacar una muestra y el error de proceso. ¿Cuál tipo de error se disminuye con el uso de la metodología de la investigación en una página?

¿Qué tipo de hipótesis es recomendable plantear en un proyecto de investigación?

¿Para qué es necesario definir las variables dependientes e independientes?

¿Qué tipo de fórmula estadística es la más recomendable para realizar una investigación a población infinita?

### III. ESCALAS DE MEDICIÓN E INSTRUMENTO DE MEDICIÓN

#### A. Escalas de medición:

Existen cuatro escalas para medir un concepto universalmente aceptadas: nominal, ordinal, intervalar y razón.

- a. La escala nominal: es conocida como escala primitiva, cualitativa, débil, discreta y/o dicotómica. Es una escala mutuamente excluyente entre sí, permite seleccionar un atributo como respuesta de una serie de atributos, por ejemplo, seleccionar si es hombre o mujer (Dicotómico), seleccionar el restaurante donde más le agrada adquirir los alimentos (Politómico, más de dos restaurantes para seleccionar), seleccionar la aerolínea que más le agrada para viajar y seleccionar dentro de que rango de edad se encuentra el encuestado, entre otros ejemplos.

- ¿Cuál es su sexo?                      Masculino                       Femenino
- ¿Cuál restaurante de la ciudad le gusta más para cenar? El rápido   
Familiar     Bohemio     Brisas

Esta escala se conoce como cualitativa o débil, porque solo se puede realizar estadísticos descriptivos de frecuencias, porcentajes y gráficas, y para inferir se puede solicitar la prueba de la Chi Cuadrada. Hay preguntas en el cuestionario que únicamente se pueden medir con esta escala de medición como es la variable sexo.

- b. La escala ordinal: es también conocida como escala cualitativa o débil, jerárquica y discreta que permite la posibilidad de jerarquizar las respuestas de mayor a menor o viceversa; se le pide al encuestado que ordene según su preferencia una serie de atributos a evaluar, por ejemplo, jerarquice de 1 a 4 según su preferencia los siguientes 4 restaurantes, siendo 1 el de menor preferencia y 4 el de mayor preferencia.

- Ordene según su preferencia el restaurante donde le gustaría ir a cenar, siendo uno el de menor agrado y 4 el de mayor agrado:
  - El rápido
  - Familiar
  - Bohemio
  - Brisas

Esta escala permite realizar las mismas técnicas estadísticas de la escala nominal, por esa razón se conocen las dos escalas como escalas débiles estadísticamente hablando.

- c. La escala intervalar: conocida como escala cuantitativa, superior, fuerte, numérica, escalar o intervalar. Creada por el Dr. Rensis Likert, en el año 1960 y ampliamente conocida como la escala Likert, permite conocer que tanto le agrada los atributos que se están midiendo. Esta escala tiene implícita la escala nominal, ya que permite conocer cual atributo le gusta, y también tiene implícita la ordinal, ya que permite la jerarquización de los atributos dándole un orden de mayor a menor y la escala Likert que permite la posibilidad de qué tanto, por ejemplo, evalúe según su preferencia los siguientes 4 restaurantes para ir a cenar con la siguiente escala: 1 Totalmente en desacuerdo, 2 En desacuerdo, 3 Indeciso, 4 De acuerdo y 5 Totalmente de acuerdo.

- Restaurante el Rápido es el mejor restaurante para ir a cenar
- Restaurante Familiar es el mejor restaurante para ir a cenar
- Restaurante Bohemio es el mejor restaurante para ir a cenar
- Restaurante Brisas es el mejor restaurante para ir a cenar

Esta escala Likert tiene la particularidad de que tanto bajo el concepto de la suma algebraica con las posibles respuestas, ya que:

1. Totalmente en desacuerdo, equivale a menos 2
2. En desacuerdo, equivale a menos 1
3. Indeciso, equivale a 0
4. De acuerdo, equivale a más 1
5. Totalmente de acuerdo, equivale a más 2

Las evaluaciones dada a cada uno de los restaurantes corresponde al valor con el signo correspondiente según la escala Likert, por ejemplo, si evaluaron con 5 al restaurante Brisas, con 3 al Bohemio, con 1 al Rápido y con 4 el Familiar, equivale a que la evaluación mayor la tiene el Brisas, ya que 5 equivale a +2, seguido del Familiar con +1, el Brisas con 0 y en el último lugar el Rápido con una evaluación de -2.

La fortaleza de esta escala es que permite realizar todas las estadísticas con las que se cuenta hoy por hoy para analizar los datos. Como tiene implícita las escalas débiles, se podrá realizar las estadísticas propias de esas escalas más las estadísticas superiores.

No olvidar que esta escala tiene un cero arbitrario, como la escala del termómetro, que también tiene un cero arbitrario, donde cero significa punto de congelación del agua y 100 grados centígrados significa punto de ebullición del agua, por ejemplo: En noviembre el clima en New York marca una temperatura de 12 grados centígrados a medio día y en Cancún, marca una temperatura de 24; no se puede concluir que en Cancún el clima es el doble que en New York, ya que el cero es arbitrario, lo único que se puede concluir es que es menos frío en Cancún, ya que el termómetro en Cancún marca 24. Pero, qué sucede con escalas donde el cero es absoluto, como la edad, el peso, la estatura, etc., por ejemplo, la estatura de María es de 120 centímetros y la de Pablo es de 180, por lo que se puede concluir que Pablo es 20 centímetros más alto que María, sin temor a equivocaciones ya que la escala tiene un cero real y absoluto, dando paso a la siguiente escala de medición conocida como de razón.

La escala para medir percepciones de 0 a 100 de cualquier atributo también es de tipo intervalar, por ejemplo, evalúa de 0 a 100 la percepción que tiene usted sobre la calidad del servicio, calidad del producto, calidad en la atención, etc. Siendo 100 la máxima calificación.

- d. Escala de Razón: al igual que la escala Likert, es conocida como escala cuantitativa, superior, fuerte, numérica, escalar o intervalar, con la diferencia que todas las preguntas cuentan con una respuesta que siempre debe ser un número con un cero real y absoluto, por ejemplo, la pregunta de cuánto pesa, cuánto mide, cuánto dinero tiene, cuantos salarios mínimos al mes integra su nivel de ingresos, cuántos hermanos tiene, cuantas ventanas tiene su casa, cuántos carros hay en su casa, cuantas personas viven en su hogar, etc.

Esta escala al igual que la escala Likert, permite realizar todas las estadísticas que hoy por hoy existen.

Es importante mencionar que se recomienda que casi todas las preguntas del cuestionario se midan con estas dos escalas superiores, ya que es posible bajar la escala de categoría superior a escala inferior, pero nunca se podrá pasar la escala inferior a superior, ejemplo, si preguntan la edad con la posibilidad de que la persona de la cifra de años cumplido es de razón, pero si preguntan en rangos de edad, se ha baja la escala de medición a un nivel inferior o nominal. Cuando se pregunta la edad en número y no en rangos, se puede calcular todas las medidas estadísticas de centralidad y dispersión, como las frecuencias, porcentajes, media, mediana, moda, rango y desviación estándar, distribución normal y estadísticas superiores; mientras que en rangos o nominal solo se podrá calcular frecuencias, porcentajes para describir y para inferir la Chi Cuadrada.

## **B. Instrumento de medición tipo cuestionario**

El cuestionario es un documento donde se concentran las preguntas altamente relacionadas con las variables a medir en todo proyecto de investigación. Las preguntas pueden ser abiertas para permitirle al encuestado contestar lo que considere pertinente, cerradas con opciones de sí y no, falso o verdadero, o de opción múltiple, conocido como cuestionario con preguntas estructuradas que solo permite la respuesta de una de las opciones que deben ser seleccionadas por el encuestado. Es recomendable definir la ruta para construir un cuestionario partiendo de la identificación de las variables, definir las variables según la teoría o conceptos que le den el fundamento a cada pregunta, usar la escala de medición para la mensurabilidad que lleve a medir lo que se quiere medir en función de las cuatro escalas de medición conocidas como nominal, ordinal, intervalar y razón. En el apartado de ítems se debe redactar la pregunta según el contexto social y cultural del encuestado, realizar preguntas positivas y evitar el preguntar dos atributos a la vez o que se genere ambigüedad en la respuesta. Relacionar los ítems con cada pregunta del cuestionario, con las hipótesis planteadas.

En evaluación de la viabilidad (Feasibility) con el apoyo de una prueba piloto, se debe ver los criterios de redacción de las preguntas según el contexto social y cultural del encuestado, pregunta de manera positiva y sin ambigüedades, con escala de medición adecuada que permita la codificación óptima y que proporcione limpieza en la interpretación. En el apartado de validez (validity) se busca tener la certeza de que el cuestionario realmente mide lo que se quiere medir y para ello se debe realizar las pruebas de validez de contenido, concurrente de constructo, convergente y discriminante:

- a. Contenido: se debe apoyar con la teoría (marco teórico) o conceptos (marco conceptual) y/o por opinión de experto, para que las preguntas tengan los conceptos a medir; se puede aplicar el análisis de factores exploratorio (cuando se realiza un cuestionario por primera vez o se realiza una traducción del idioma original a otro) o confirmatorio (cuando se prueba un

- cuestionario que ya tiene criterios de validez y confiabilidad- de ahí el nombre de confirmatorio) para identificar los ítems que se agrupan en cada concepto.
- b. Concurrente: identificar la correlación de ítems que miden lo mismo o que están relacionados con indicadores oficiales o con otro fundamento demostrable.
  - c. Constructo: las preguntas deben estar respaldadas por la teoría o el concepto correspondiente y por el análisis de factores que permite identificar la agrupación de los ítems en conceptos que miden lo mismo.
  - d. Convergente: es la correlación de ítems con conceptos altamente relacionados entre sí.
  - e. Discriminante: es la agrupación de ítems que guardan una relación entre sí y que a su vez cada agrupación de ítems son diferentes.

La confiabilidad (reliability) pretende conocer si el cuestionario sirve para medir lo que se quiere medir, ver si el cuestionario es consistente o si tiene consistencia interna con el Alfa de Cronbach del total de las preguntas o por bloques de preguntas. Las pruebas de confiabilidad más recomendadas a realizar son fiabilidad Test-Retest intraobservador y la fiabilidad interobservador:

- a. Fiabilidad Test-Retest, intraobservador: coeficiente de correlación intraclass para preguntas en escala superiores y apoyado con el Alfa de Cronbachs. Para preguntas medidas en escala nominal o para evaluaciones realizadas por expertos se utiliza el índice Kappa de Cohen, para encontrar la concordancia entre dos expertos, examinadores o evaluadores. También se utiliza el método gráfico de Bland y Altman.
- b. Fiabilidad interobservador: mide el grado de acuerdo entre pares evaluadores, con el apoyo del KR20 (Kuder-Richardson) y el índice Kappa de Cohen. las pruebas del Alfa de Cronbach o KR20 (Técnica de Kuder-Richardson) o la de Kappa de Cohen. Ver tabla número 2.

Tabla No. 2

### CUESTIONARIO: CRITERIOS DE VIABILIDAD, VALIDEZ Y CONFIABILIDAD

Variables	Definición	Mensurabilidad	Items	Evaluación de la viabilidad	Validez	Confiabilidad	Teoría y Estadísticos
- Dependientes - Independientes - Intervinientes o extrañas.	Teoría y/o Conceptos con su respectivo pila de páginas. Relacionar las variables con cada pregunta del cuestionario o con cada una de las hipótesis planteadas.	Seleccionar una de las 4 escalas de medición: Nominal Ordinal Intervalar Razón	Redactar las preguntas según el contexto social y cultural del encuestado. Pregunta de manera positiva. Preguntar solo un atributo sin permitir ambigüedades.	<b>Feasibility (Viabilidad):</b> prueba piloto para mejorar la redacción de las preguntas y la escala de medición. Se deben construir preguntas claras con una codificación factible y que proporcione limpieza en la interpretación	<b>Validity:</b> Mide lo que se quiere medir. <b>De contenido:</b> teoría y opinión de expertos. Las preguntas contienen los conceptos a medir. Se puede hacer análisis de factores. <b>Concurrente:</b> correlación de items relacionados con indicadores oficiales o con fundamento. <b>De Constructo:</b> Las preguntas deben estar respaldadas por el concepto y por el análisis de factores. <b>Convergente:</b> correlación de items con conceptos altamente relacionados. <b>Discriminatoria o divergente:</b> agrupación de items que guardan una relación entre sí.	<b>Reliability:</b> ¿El cuestionario sirve para medir lo que se quiere medir? ¿Es consistente? Consistencia Interna con el Alfa de Cronbach del total de las preguntas o bien de manera parcial. <b>Fiabilidad Test-Retest, intraobservador:</b> Coeficiente de correlación intraclass para preguntas en escalas superiores y el índice de Kappa de Cohen para escalas débiles o cualitativas (Concordancia entre dos examinadores o evaluaciones). Método gráfico de Bland y Altman. <b>Fiabilidad interobservador:</b> grado de acuerdo entre pares evaluadoras. KR20 (Kuder-Richardson)	Respaldo teórico. Correlación, factor análisis, discriminante. Kappa de cohen: Superior a .80 se considera buena concordancia. En el spss, analizar, tablas de contingencia, pasar las variables, estadístico kappa. $K = (Co - Ce) / 1 - Ce$ Co= Concordancia observada y Ce= esperada. Alfa de Cronbach: .70-.80 satisfactorio .81-.90 muy bueno Más de .91 excelente. En spss, analizar, escala, fiabilidad, pasar las variables, modelo alfa. KR 20 y 21: en preguntas de respuesta dicotómica y se califica con 1 (acierto) y 0 (No acierto)

Fuente: aportación del autor

El orden de las preguntas dentro del formato del cuestionario debería seguir la siguiente recomendación: primero se debe dar una introducción e invitación a contestar el cuestionario de manera honesta y transparente, la primeras preguntas deben ser de introducción al tema, seguidas de preguntas centrales de la investigación, para dejar al final las preguntas más difícil de contestar y terminar con las preguntas demográficas como edad, sexo, estado civil, escolaridad, nivel de ingresos, estrato socioeconómico y otras variables propias de los individuos que darán respuesta al cuestionario.

- PRUEBAS DE VALIDEZ:

**Prueba de validez de contenido:** se logra cuando se tienen los soportes o fundamentos teóricos o conceptuales que le dan origen a la pregunta del cuestionario, para saber que se está preguntando lo que realmente se quiere medir. También se puede lograr con el apoyo de expertos en los temas para que ellos ayuden en la redacción de la pregunta. Estadísticamente hablando se puede aplicar la prueba de análisis de factores para ver si las preguntas se agrupan en el concepto previamente definido; correlación y análisis discriminante. Para las preguntas cualitativas y/o coincidencias entre los pares evaluadores de un cuestionario, se procede a realizar la prueba Kappa de Cohen, la cual indica el nivel de concordancia entre el par de evaluadores (valores de .80 se consideran como buena concordancia). Para el cálculo del índice de Kappa de Cohen se procede primero a conocer la concordancia observada (Co), la observancia esperada (Ce) para realizar el cálculo:

$K = (Co - Ce) / 1 - Ce$ . Los valores van de 0 a 1. Concordancia menor a .5 se considera mala, de .6 a .7 débil, de .8 a .9 buena y superior a .9 muy buena.

**Prueba de validez de constructo o de concepto:** en varias ocasiones se realizan dos o más preguntas para medir un concepto, por ejemplo, el concepto de calidad en el servicio requiere de por lo menos 15 preguntas que miden este concepto, por lo que es recomendable realizar el análisis de factores para corroborar que las preguntas se agrupen alrededor del concepto en medición.

**Prueba de validez concurrente:** se debe correlacionar la medida obtenida con un valor estándar o una medida aceptada, para ver si hay concurrencia. Otro esquema es medir un concepto con dos escalas de medición diferentes (intervalar y de razón) y se procede también a realizar la correlación entre los resultados de las dos preguntas; si hay alta correlación entre ellas se puede decir que hay concurrencia. Se dice concurrente porque en el mismo momento se aplican las preguntas y se procede a realizar la prueba de correlación para ver su grado de concurrencia, de ahí el nombre de concurrente.

**Prueba de validez convergente:** con el apoyo de la correlación, se pretende medir el grado de correlación que existe entre dos o más variables que miden el mismo concepto. Si hay correlación superior a .6 entre las variables se puede decir que existe correlación convergente. También se puede aplicar la técnica estadística de análisis de factores para ver si las variables se agrupan en factores que miden el concepto preestablecido con el marco teórico conceptual del proyecto de investigación. Lo contrario se conoce como divergente o validez discriminante.

**Prueba de validez discriminante:** se aplica cuando varias preguntas del cuestionario miden varios conceptos y estas se agrupan discriminándose entre grupos, por ejemplo, preguntas que miden el servicio al cliente, y la calidad de producto se pueden agrupar en dos grupos o factores quedando en las variables que miden servicio al cliente en el grupo correspondiente y las de calidad en el producto en la agrupación que mide ese concepto. Estadísticamente se puede realizar el análisis de clúster o de agrupación de variables y también se podría utilizar el análisis de factores.

- PRUEBAS DE CONFIABILIDAD:

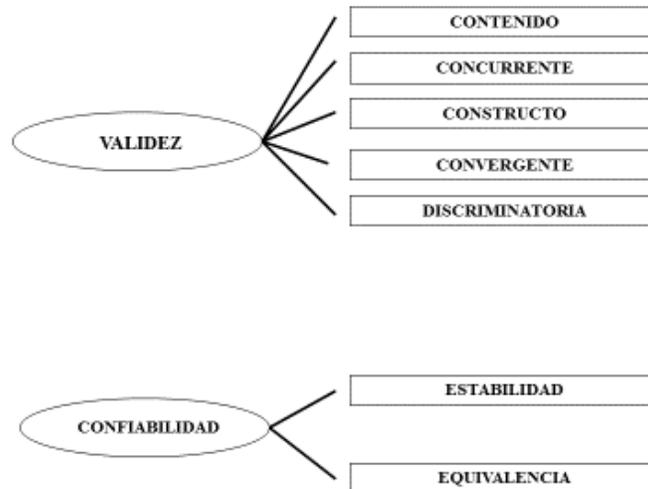
Una vez realizada las pruebas de validez y estar seguros de que cada pregunta del cuestionario mide lo que se quiere medir, se debe proceder a realizar las pruebas de confiabilidad del cuestionario, es muy similar como el aplicar un termómetro para medir la temperatura corporal del ser humano, en estos tiempos nadie duda que el termómetro es válido porque fue diseñado y elaborado para medir la temperatura del cuerpo, pero se puede tener dudas si está bien o está dañado, para ello se procede aplicar el termómetro a una persona normal y si este marca valores arriba o por abajo del estándar de un ser humano normal, que es 37 grados centígrados, se puede decir que no es confiable o simplemente que es válido pero no es confiable o que no sirve. Lo mismo se tiene que hacer con los cuestionarios una vez que pasan la prueba de la validez, se procede aplicar el cuestionario a una muestra que permita realizar las pruebas estadísticas aceptadas para medir la confiabilidad como el Alfa de Cronbach (para preguntas del cuestionario medidas en escalas superiores tipo Likert o de razón) y el KR20 (para preguntas del cuestionario medidas en escala débil o nominal).

Se conoce dos grandes vías para lograr la confiabilidad o fiabilidad del cuestionario o instrumento de medición: la estabilidad y la equivalencia

- Estabilidad: se presenta cuando se aplica el cuestionario en dos tiempos diferentes y se esperaría que los resultados fueran consistentes, ejemplo, se aplica un cuestionario para medir la calidad del servicio en tiempo presente y dentro de 2 meses se vuelve aplicar el mismo cuestionario y se esperaría resultados similares en Alfa de Cronbach entre los dos tiempos de las variables a las que se les está midiendo la confiabilidad para concluir que el cuestionario es estable con el paso del tiempo.
- Equivalencia: es cuando se tiene dos o más de dos preguntas que miden lo mismo y se espera tener una equivalencia de las respuestas o consistencia interna, misma que se mide con el alto nivel de correlación que se debe presentar entre las variables que miden el

mismo constructo. Si las respuestas dadas a las preguntas del cuestionarios se relacionan se podría decir que tienen confiabilidad o fiabilidad de equivalencia Ver gráfica número 20.

Gráfica No. 20



Fuente: Adaptada de Weiers, Ronald, Investigación de Mercados, Prentice-Hall, México, 1986

### **C. Pruebas de confiabilidad con Alfa de Cronbach y KR20 del instrumento de medición o de consistencia interna:**

Se recomienda aplicar el Alfa de Cronbach (Cronbach, L.J., 1951) a todas las preguntas del cuestionario de preferencia medidas en escala intervalar tipo Likert, para conocer la homogeneidad del total de las preguntas o bien aplicar la prueba por bloques de preguntas del cuestionario, para tener valores de confiabilidad por bloques de ítems que miden los mismos conceptos. Es importante saber que esta prueba trabaja mejor con variables del cuestionario medidas en escalas superiores (Intervalar y /o de razón). La prueba KR20, es útil para preguntas del cuestionario medidas en escala nominal o dicotómica.

En el IBM-SPSS, se procede: **Analizar, escala, análisis de fiabilidad, se pasan las variables que se van a analizar, estadísticos: se selecciona correlación, seleccionar el modelo alfa y aceptar.** El índice va de 0 a 1, siendo cero igual a no confiabilidad y uno a perfecta confiabilidad, un alfa menor a .70 se considera inaceptable, de .71 a .79 es adecuado en cuanto al grado de confiabilidad del instrumento de medición, de .80 a .89 es buena confiabilidad y superior a .91 es una confiabilidad muy buena, quedando pendiente la validez del mismo instrumento de medición por los criterios de contenido, constructo, concurrente, convergente y discriminante.

**D. Supuesto de normalidad:** (Kolmogorov-Smirnov-KS y Shapiro-Will)

Esta prueba compara la distribución normal de cada variable superior o medida en escala intervalar o razón, con la distribución normal teórica de cada variable para identificar si se comporta como una distribución normal.

***Para cada variable en IBM-SPSS: analizar, estadístico descriptivo, explorar, lista de variable dependiente (se pasan las numéricas para ver si se comportan normalmente), gráficos con opción de normalidad y aceptar.***

Si el P valor es mayor de .05 se concluye que hay normalidad entre las variables. Se observa el valor arrojado por Kolmogorov-Smirnov (muestras mayores de 50 datos) y/o la prueba de Shapiro-Wilk (para muestras menores a 50 datos), con respectivo nivel de significancia. Se acepta la hipótesis nula (Ho), por lo que se pueden utilizar las variables con datos para estadísticos paramétricos. (Es uno de los pocos momentos en la estadística que se acepta Ho)

***Otro camino con el software de IBM-SPSS es: analizar, cuadro de diálogos antiguos, KS de 1 muestra, seleccionar las variables, marcar normal y aceptar.***

Arroja los resultados de Kolmogorov Smirnov. Si la Significancia es mayor o igual a .05, se concluye que la variable sigue la normalidad.

***Para muestras de antes y después: en el IBM-SPSS: Analizar, pruebas no paramétricas, cuadro de diálogos antiguos, KS de 1 muestra, seleccionar la variable del antes y después, aceptar.***

Si los P valor son mayores se puede concluir que las muestras siguen una distribución normal.

Hay eventos de antes y después, donde se requiere conocer la distribución normal de la variable, pero esta se debe realizar a la diferencia del antes y después. Para ello se debe transformar la variable en el SPSS (***transformar, calcular variable, diferencia, peso antes –peso después***) en la matriz de datos se verá una nueva variable con el nombre de diferencia. Sobre esa variable se le aplicará la normalidad; en el **IBM-SPSS se va analizar variables, cuadro de diálogos**

**antiguos, KS de una muestra, y se traslada la variable diferencia y aceptar.** No olvidar que debe ser mayor o igual a .05 para ver si tiene distribución normal.

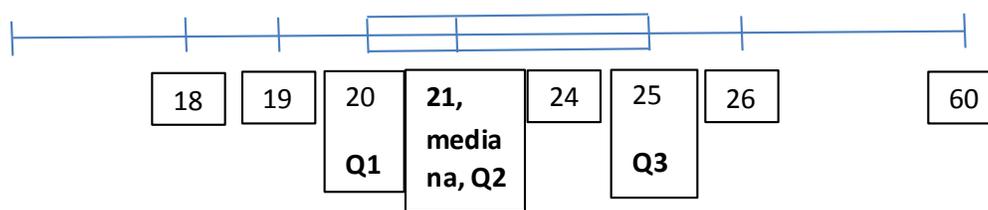
Cuando no hay normalidad, se recomienda analizar la normalidad entre los grupos, en **el IBM-spss: datos, segmentar, seleccionar la variable de agrupación, seleccionar los resultados por grupos y aceptar; de nuevo se realizan los pasos para la normalidad: analizar, cuadro de diálogos antiguos, KS de una muestra, se traslada la variable y aceptar.** Se puede ver la normalidad por grupos y ver si la significancia es mayor o igual a .05, para no rechazar la Ho de igualdad de normalidad.

**La prueba de Wilcoxon**, es una prueba no paramétrica donde compara grupos de antes y después cuando la variable no cumplen con la normalidad.

Ejemplo en **el IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, 2 muestras relacionadas, seleccionar la variable antes y después, Wilcoxon y aceptar.** Si el p valor es menor o igual a .05 se rechazará la Hipótesis nula (Ho) o de igualdad entre las medias de un antes y después.

**E. Identificar datos atípicos (outliers):** es recomendable identificar dentro de un conjunto de datos medidos en escala de razón o variables medidas en escala continua, los datos que estén fuera de un rango o que tengan puntuaciones extremas, por ejemplo, en la variable edad se puede tener una serie de datos con edades muy cercanas para conocer la percepción de la población entre 18 a 30 años; se esperaría no tener en esa unidad de muestra datos menor a 18 o mayor a 30, pero suele suceder que se tenga datos fuera de ese rango. Las pruebas estadísticas para identificar los datos atípicos se apoya en el cálculo de la caja y bigotes, para conocer el valor mínimo, el valor máximo, la mediana, el cuartil 1 (Q1= 25% de los datos), el cuartil 2 (Q2: 50% de los datos) y el cuartil 3 (Q3: 75% de los datos); cálculo del rango intercuartil (RIC) que es la diferencia entre el Q3-Q2. Dato debajo de Q1 o bien por arriba de Q3 más/menos 1.5 por el rango intercuartil es considerado atípico, ejemplo en el siguiente grupo de edad, encontrar el dato atípico:

18, 19, 20, 20, 20, 21, 24, 25, 25, 26, 60. La mediana es el dato 21 ya que divide el conjunto de datos en proporciones iguales o número de datos iguales en los dos lados (5 para el lado izquierdo y 5 para el lado derecho); en la gráfica caja y bigotes se pueden visualizar los datos con la mediana, cuartil 1 ( $Q1=20$ ; 25% de los datos), cuartil 2 ( $Q2=21$ ; 50% de los datos), cuartil 3 ( $Q3=25$ ; 75% de los datos) y el intervalo intercuartil; el lado izquierdo de la caja de cajas y bigote se dibuja a partir del  $Q1$ , que es el valor que a su vez divide a los 5 datos justo en un punto medio hasta la mediana y de la misma manera se calcula el lado derecho. Debido al dato atípico de 60, se ve la caja de bigote inclinada hacia el lado derecho. El Rango Intercuartílico (RIC) es igual a  $Q3-Q1$  ( $25-20= 5$ ) Al eliminar el dato atípico se puede observar mayor homogeneidad de los datos. El dato atípico inferior se calcula con la  $Q1-(1.5 *RIC)$ ;  $20-(1.5*5)= 12.5$  y el dato atípico superior se calcula  $Q3+(1.5*RIC)$  = 28.5. Los datos considerados atípicos estarán debajo de 12.5 años de edad y mayores de 28.5, por lo que el dato de 60 años se considera dato atípico (outlier) y se sugiere eliminarlo del conjunto de datos.



Ejemplo en *el IBM-SPSS: analizar, estadísticos descriptivos, explorar, pasar las variables en lista de variables dependientes, en estadísticos seleccionar descriptivos, valores atípicos, percentiles, continuar, en mostrar seleccionar ambos y aceptar*. Con los mismos datos del ejemplo de las edades el IBM-SPSS, arroja los siguientes resultados:

## Resumen de procesamiento de casos

	Casos					
	Válido		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
¿Cuántos años cumplidos tiene?	11	100.0%	0	0.0%	11	100.0%

## Descriptivos

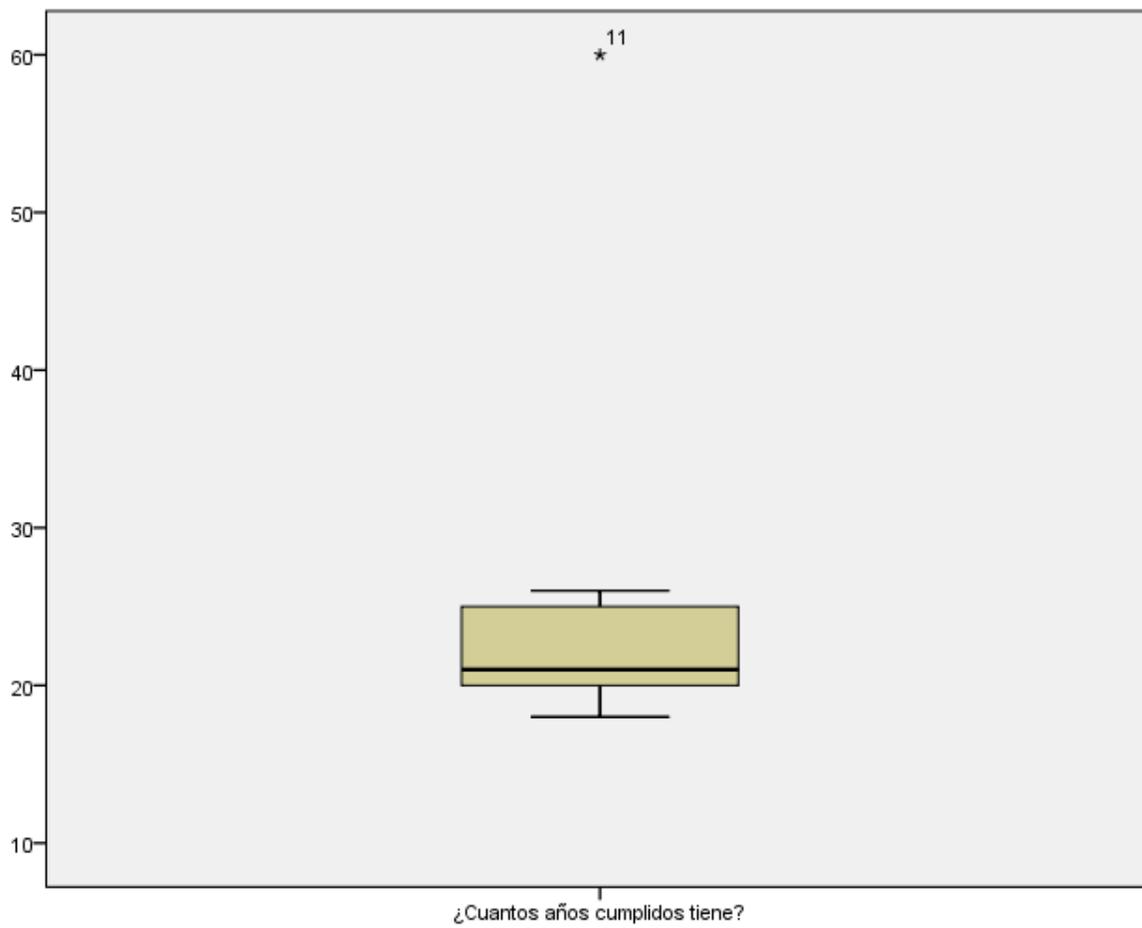
		Estadístico	Error estándar	
¿Cuántos años cumplidos tiene?	Media	25.27	3.570	
	95% de intervalo de confianza para la media	Límite inferior	17.32	
		Límite superior	33.23	
	Media recortada al 5%	23.75		
	Mediana	21.00		
	Varianza	140.218		
	Desviación estándar	11.841		
	Mínimo	18		
	Máximo	60		
	Rango	42		
	Rango intercuartil	5		
	Asimetría	2.998	.661	
	Curtosis	9.460	1.279	

Se puede observar que los datos coinciden ya que la mediana es de 21, el mínimo es 18 y el valor máximo es de 60, el rango intercuartil es de 5, el cuartil 1 está en el percentil 25 con el valor de 20 años de edad, el cuartil 2 está en el percentil 50 con el valor de 21 y el cuartil 3 en el percentil 75 con el valor de 25 años de edad.

En la gráfica se puede apreciar claramente que el valor considerado como atípico o outlier es el 60, registrado en el ibm-spss como el individuo número 11 que tiene 60 años de edad.

## Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado (Definición 1)	¿Cuántos años cumplidos tiene?	18.00	18.20	20.00	21.00	25.00	53.20	.
Bisagras de Tukey	¿Cuántos años cumplidos tiene?			20.00	21.00	25.00		



## Conclusión

Las escalas de medición e instrumentos de medición van de la mano para el desarrollo de la investigación cuantitativa. Existen cuatro escalas para medir un concepto, una idea o una variable. La escala nominal es apropiada para medir la variable sexo, siendo 1 femenino y 2 masculino, con este tipo de escalas solo se puede medir la frecuencia, porcentajes y gráficos, de ahí que es conocida como escala débil. Le sigue la escala ordinal, esta escala incluye la anterior, pero tiene un poco más de medida ya que se pide la jerarquización de mayor a menor de ciertos atributos a evaluar, por ejemplo: ordene según tu preferencia las universidades de tu ciudad por la calidad académica, ocupando el primer lugar aquella que tenga la mejor calidad académica y en el último lugar la que se considera con menor calidad. También esta escala para medir un concepto es conocida como débil.

La escala intervalar, es conocida como superior y sirve para medir las variables o conceptos que permitan conocer el ¡que tanto!; son aquellas que tienen implícitamente las dos medidas anteriores, la más conocida es la inventada por el Dr. Likert, ampliamente conocida como escala Likert, ya que mide que tanto le agrada los atributos a evaluar con una escala que va de 1 a 5, siendo 1 totalmente en desacuerdo, 2 en desacuerdo, 3 indeciso, 4 de acuerdo y 5 totalmente de acuerdo; con esta escala se pueden aplicar todas las técnicas estadísticas que existen para describir e inferir. Es conocida como escala superior y puede tener dimensiones de 1 a 5, de 1 a 7 o de 1 a 9, aunque en los últimos años se está optando por quitarle el número impar y obligar a los encuestados a que se incline a favor o en contra, pero que no se queden el punto intermedio que no ayuda a la investigación. La otra escala superior lleva el nombre de razón, porque se permite realizar razonamiento matemático, por ejemplo, son todas aquellas preguntas del cuestionario donde la respuesta es un número: ¿Cuánto mide? ¿Cuánto pesa? En cuanto al peso si A pesa 90 kilos y B 60, se puede decir que B pesa 30 kilos menos que A, de ahí el razonamiento matemático, situación que no se puede concluir con la escala intervalar tiene un cero arbitrario, como el termómetro el cual cero indica el punto de congelación del agua y 100 el punto de ebullición; Si en estos momentos la temperatura en la Ciudad de México es de 15 grados y en Acapulco de 30 grados,

solo se puede decir que Acapulco es más caliente que México, pero nunca el doble. Todas las escalas de medición son necesarias, se recomienda diseñar las preguntas del cuestionario en el 95% con escalas superiores tipo Likert o de razón y solo el 5% con escalas débiles de tipo nominal u ordinal.

Una vez definidas las escalas de medición se procede a diseñar el instrumento de medición o cuestionario estructurado, para aplicarlo a la prueba piloto que permita realizar las estadísticas de validez y confiabilidad.

Las pruebas de validez y confiabilidad son necesarias en todo instrumento de medición o cuestionario, siendo las de mayor uso, las pruebas de validez de contenido, constructo, concurrente, convergente y discriminatoria, para la prueba de confiabilidad se utiliza las técnicas Alfa de Cronbach para preguntas medidas en escalas superiores y el KR 20 para preguntas del cuestionario en escala nominal; su interpretación va de 0 a 1, siendo 0 ausencia de confiabilidad y 1 de excelente confiabilidad.

**Palabras Clave:**

Escalas de medición

Escala nominal, ordinal, intervalar y de razón

Validez y Confiabilidad

Alfa de Cronbach y KR 20

**Preguntas:**

¿Cuál sería la proporción recomendable de preguntas de un cuestionario con las 4 escalas de medición?

¿Cuáles son los criterios de validez y confiabilidad para un instrumento de medición?

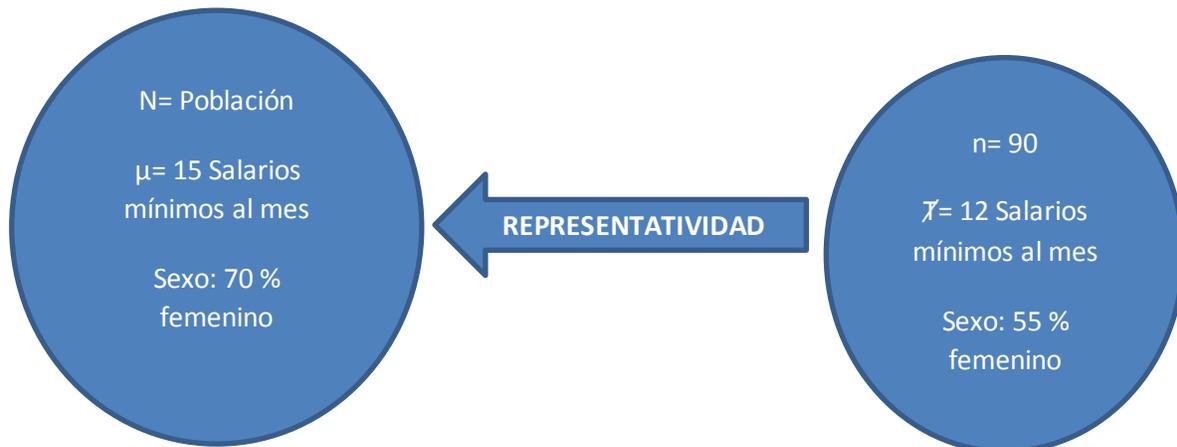
¿Aplicarías un cuestionario que es válido pero que no es confiable?

¿En qué tipo de variables se debe aplicar la técnica de Alfa de Cronbach para medir la confiabilidad del cuestionario?

¿En qué tipo de variables se debe aplicar la técnica KR-20 para medir la confiabilidad del cuestionario?

#### IV. ESTADÍSTICA PARA REPRESENTATIVIDAD DE LA MUESTRA

***La estadística es una ciencia y una rama de la matemática que recopila, organizar y analiza datos e información para la toma de decisión.***



¿Es representativa la muestra de la población? La muestra es representativa de la población cuando los datos de las variables a medir se acercan a los datos de la población y no está sesgada o inclinada a un grupo de atributo de las variable que se miden, por ejemplo, la variable salarios y sexo en la población de 15 salarios mínimos en promedio y la relación de sexo del 70 % femenino marcan la pauta para la selección de la muestra. Se debe buscar en la muestra promedios de ingresos y proporción en sexo cercanos a la población para afirmar con pruebas estadísticas que la muestra es similar o representativa de la población y así realizar la estadística inferencial con mayor precisión.

Para ver si una muestra es representativa de la población se puede utilizar las estadísticas relacionadas a la distribución normal (Variables medidas en escala intervalar o razón) como la T de Student, o las relacionadas a distribución binomial (Variables medidas en escala nominal-Dicotómica-). Se deben aplicar pruebas de representatividad de una muestra en las variables sexo, edad, estado civil, escolaridad y nivel de ingresos entre otras relacionadas a la demografía de la población. Cada una de estas variables se mide en diferentes escalas de medición, por lo que se deben aplicar pruebas estadísticas diferentes según sea el caso.

En las variables asociadas a la distribución normal, como las intervalares y de razón se debe utilizar la prueba T de Student: se requiere conocer la media de la población de la variable que se está midiendo. Como ejemplo se utiliza la variable edad, medida de manera directa sin la facilidad de los rangos, esto es, se pregunta de manera directa en el cuestionario la edad y se deja el espacio para que el entrevistado diga o escriba sus años cumplidos y se compara con la edad promedio de la población.

Ejemplo, la edad promedio de la población es de 25 años. La hipótesis univariada a formular sería:

- Ho: La edad de la muestra es igual al de la población.
- Ha: La edad de la muestra es diferente al de la población.

Ejemplo en el IBM-SPSS: **analizar, comparar medias, prueba T para una muestra, seleccionar la variable numérica edad, en valor de prueba se pasa el valor de la media de edad en la población (en este caso es de 25 años de edad en la población) y aceptar.**

Estadísticos para una muestra

	N	Media	Desviación típ.	Error típ. de la media
Años cumplidos del entrevistado	90	27.43	4.17	.44

Prueba para una muestra

	Valor de prueba = 25					
	t	gl	Sig. (bilateral)	Diferencia de medias	Intervalo de confianza para la diferencia	
					Inferior	Superior
Años cumplidos del entrevistado	5.531	89	.000	2.43	1.56	3.31

El promedio de edad de la muestra de 90 personas es de 27.43 años, siendo diferente al de la población (Definido por investigaciones previas o por fuentes secundarias de información de 25 años). En la siguiente tabla con el nombre "Prueba para una muestra" se observa que la significancia es de 0.000 (**si la significancia es < o = a .05 se rechaza Ho**) con un valor de t de 5.531 y 89 grados de libertad (90 personas menos uno) y con una diferencia de 2.43 entre el valor de la población Vs de la muestra, se concluye que la muestra no es representativa de la población en la variable edad. Cómo el valor es menor a .05 se rechaza Ho y se puede concluir que la muestra no es representativa de la población en la variable edad, por lo que se sugiere incrementar la muestra hasta encontrar representatividad por la variable edad. La representatividad de una muestra está en función de las variables a considerar de la población, por lo que se debe buscar la representatividad por sexo, escolaridad, ingresos, estrato socioeconómico, ocupación y estado civil entre otros. El tamaño de la muestra por formula tiende a ser de tamaño superior a 350 personas, ya que ese número ayuda a tener una representatividad de la muestra hacia todas las variables propias de la población.

**Para variables asociadas a la distribución binomial, nominal u ordinal:** se refiere a todas las variables propias de los individuos que se pueden medir de manera dicotómica, como sexo, preferencia por comprar en “X” lugar, etc. Con el apoyo de la prueba de la binomial se tiene que especificar el valor de la proporción de la población, ejemplo, según estudios de las universidades privadas la proporción de estudiantes de la maestría en género es del 68 % para los hombres y del 32 % para las mujeres. Con este dato se puede concluir que la proporción entre hombres y mujeres es mayor en los hombres con el 68 %. Ver ejemplo:

Hipótesis:

- Ho: La proporción de Hombres es igual en la muestra y la población.
- Ha: La proporción de Hombres es diferente en la muestra y la población.

Ejemplo en el IBM-SPSS: **analizar, pruebas no paramétricas, cuadros de diálogo antiguos, prueba binomial, pasar la variable sexo, proporción de prueba (en este caso es de .68 hombres en la población), en opciones seleccionar estadísticos descriptivos y aceptar.**

Prueba binomial

		Categoría	N	Proporción observada	Prp. de prueba	Sig. asintót. (unilateral)
Sexo de los encuestados	Grupo 1	Masculin o	57	.633333	.68	.202 <sup>a</sup>
	Grupo 2	Femenin o	33	.37		
	Total		90	1.00		

a. La hipótesis alternativa establece que la proporción de casos del primer grupo sea < .68.

b. Basado en la aproximación Z.

Se puede ver que la proporción de hombres en la muestra es del 63.3 % y la proporción de la población es del 68 % con una significancia de .202. Se puede concluir que la muestra es representativa de la población en cuanto a la variable género, pero se tiene que incrementar el número de personas hasta encontrar la

representatividad en edad (según lo visto arriba con la variable edad) Cómo la Significancia es mayor a .05 no se rechaza Ho.

Para ver la representatividad de la variable estrato socioeconómico y género, se recomienda aplicar la prueba de Kruskal-Wallis. En el siguiente ejemplo se cruzaron las variables género como separador de grupos entre mujeres y hombres y la relación entre los niveles de estrato socioeconómico.

- Ho: El estrato socioeconómico es igual entre hombre y mujeres.
- Ha: El estrato socioeconómico es diferente entre hombres y mujeres.

Ejemplo en el IBM-SPSS: **analizar, pruebas no paramétricas, cuadros de diálogo antiguos, K muestras independientes, pasar la variable sexo en variables de agrupación, definir rango como 1 y 2, en lista de variables de prueba se debe pasar la variable estrato socioeconómico, en opciones seleccionar estadísticos descriptivos y aceptar.**

Estadísticos descriptivos

	N	Media	Desviación estándar	Mínimo	Máximo
Estrato socioeconómico de los estudiantes del MBA	102	4.78	.981	2	6
Sexo de los estudiantes del MBA	102	1.68	.470	1	2

Prueba de Kruskal-Wallis, Rangos

	Sexo de los estudiantes del MBA	N	Rango promedio
Estrato socioeconómico de los estudiantes del MBA	Femenino	33	57.05
	Masculino	69	48.85
	Total	102	

Estadísticos de prueba<sup>a,b</sup>

	Estrato socioeconómico de los estudiantes del MBA
Chi-cuadrado	1.881
gl	1
Sig. asintótica	.170

a. Prueba de Kruskal Wallis

b. Variable de agrupación: Sexo de los estudiantes del MBA.

En el cuadro de estadísticos descriptivos se puede observar que la muestra fue de 102 personas y que todos contestaron las preguntas sexo con una media de 1.68, siendo el 68% de la población de sexo masculino y estrato socioeconómico con una media de 4.78, muy cercano a 5, siendo 2 el estrato menor y 6 como el estrato más alto. En la prueba de Kruskal-Wallis, se puede observar el rango promedio del sexo y el estrato socioeconómico con 57.05 para femenino y 48.85 para masculino, por lo el sexo y el estrato socioeconómico es casi similar entre los grupos de femenino y masculino. Los valores de la tabla de estadísticos de prueba Chi cuadrada de 1.881, con un grado de libertad y significancia asintótica de .170, se puede concluir que la muestra es representativa en sexo y estrato socioeconómico, por lo que se acepta la hipótesis nula de igualdad de estrato socioeconómico de la muestra.

## **Conclusión**

Buscar la representatividad de la muestra hacia la población es uno de los puntos clave de cualquier proyecto de investigación, por lo que se debe ser específico en las variables que a considerar para medir la representatividad, por ejemplo si ha considerado medir la representatividad con la variable sexo, ingresos, estados civil y nivel socioeconómico, se debe realizar las estadísticas de representatividad para cada una de ellas aplicando la prueba estadística en relación a la escala de medición de cada variable. Lo que se busca es que el comportamiento de cada variable seleccionada, bien sea la media o la proporción de atributos, se parezcan mucho a los datos de la población para poder realizar las estadísticas inferenciales de la muestra hacia la población.

Las pruebas estadísticas para la representatividad de la muestra están en función de las escalas de medición de la variable, si la variable está medida en escala superior tipo intervalar o razón, se sugiere aplicar la técnica estadística de la T de Student y si la variable está medida en escala nominal u ordinal, se sugiere aplicar la prueba de la binomial y Kruskal-Wallis. Se busca siempre ver si la muestra es representativa de la población en las variables propias de la demografía de la población como edad, escolaridad, estado civil, sexo y estrato socioeconómico entre otras. Puede suceder en varios proyectos de investigación que se tenga la representatividad en género, pero no en nivel de ingresos, por lo que se sugiere aplicar la representatividad en todas las variables definidas como clave dentro del marco de referencia de las variables demográficas de la población.

**Palabras clave**

Representatividad de la muestra hacia la población

Prueba T Student

Prueba Binomial

Prueba Kruskal-Wallis

**Preguntas**

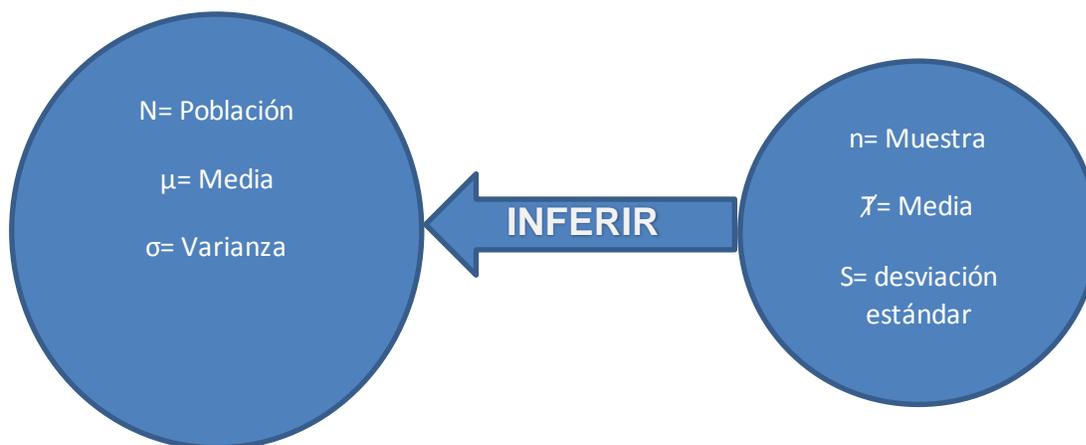
¿Cuáles variables deben ser consideradas para saber si la muestra es representativa de la población?

¿Qué pruebas estadísticas para la representatividad de la muestra hacia la población se deben realizar con escalas débiles de tipo nominal u ordinal?

¿Qué pruebas estadísticas para la representatividad de la muestra hacia la población se deben realizar en variables con escalas superiores de medición como intervalar y/o de razón?

## V. ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

**La estadística se divide en descriptiva para conocer el comportamiento de la variable e inferencial para inferir de la muestra hacia la población.**



La estadística se divide en dos grandes ramas, la descriptiva y la inferencial. La descriptiva permite conocer el comportamiento de la variable en función de la escala de medición, un concepto medido en escala nominal, solo permite calcular la **frecuencia y los porcentajes** de ocurrencia de los atributos de una variable, por ejemplo, en sexo, solo se puede calcular cuántos son hombres y cuantas son mujeres, con su respectivo porcentaje y gráfica; también se puede calcular la **razón (R)** para identificar la comparación entre atributos de una variable expresada en forma de una división donde el cociente o resultado es la razón, por ejemplo, la razón entre hombres y mujeres en un lugar y tiempo específico: si hay 100 hombres y 20 mujeres en un evento social (Razón =  $100/20=5$ ), se podría decir que hay una razón de 5 hombres por cada mujer en el evento. **Las tasas** son útiles para comparar el número de casos reales entre los casos potenciales o el universo por una base de 100, 1000, 10,000 o 100,000 (Los ceros de la base están en función de los ceros del denominador o del universo) casos potenciales por un año o por un periodo de tiempo definido, por lo que se puede calcular el número de nacimientos por cada 1000 mujeres en edad reproductiva durante el 2017, la tasa de divorcio

por cada 1000 matrimonios durante el 2017, etc., por ejemplo, en una localidad donde hay 3000 mujeres en edad fértil, se encontró que nacieron 600 niños en el año 2017, por lo que la tasa de nacimiento estaría expresada,  $600 / 3000 * 1000 = 200$ ; nacieron 200 niños por cada 1000 mujeres en edad fértil de la localidad durante el 2017.

**La tasa de cambio de las variable entre el tiempo T1 y T2**, tomando como base los datos del tiempo T1, nos permite conocer que tanto cambio se presentó durante el tiempo, por ejemplo, en una escuela la población estudiantil en el año 2000 era de 5000 estudiantes y para el 2017 se incrementó a 9000, por lo que tuvo una tasa de cambio positiva del 44.44%

Tasa de Cambio (TC):  $100 \times T2 - T1 / T2$

$TC = 100 * 9000 - 5000 / 9000 = 44.44 \%$

Con las variables o conceptos medidos con la escala ordinal, se calcula también las mismas técnicas estadísticas que se utilizan en la escala nominal, mientras que las variables medidas en escalar o intervalar y de razón, se puede utilizar la estadística de medidas de centralidad como la media, la moda y la mediana y las medidas de dispersión como el rango, la desviación media y la desviación estándar.

La estadística inferencial paramétrica, requiere conocer la distribución normal de las variables métricas, cuantitativas o superiores (intervalar y/o de razón) de la población y de la muestra para inferir, requiere de tamaño de la muestra grande por los menos 5 veces el número de variables contempladas en el cuestionario, ejemplo, si en el cuestionario se tienen 20 preguntas, el tamaño ideal de la muestra para realizar las pruebas paramétricas serían 100; la selección de la muestra debe ser vía probabilística y aleatoria. Las estadísticas más utilizadas en los proyectos de investigación con estadística paramétrica son la T de Student, correlación, regresión simple y múltiple, ANOVA, ANCOVA, MANOVA Y MANCOVA, análisis discriminante, análisis de clúster, análisis de componentes principales con análisis de factores, ecuaciones estructurales y análisis de conjunto entre otros.

La estadística no paramétrica, se utiliza cuando las variables no cumplen con la normalidad, siendo una alternativa estadística para seguir con los análisis. Generalmente se aplican en muestras más pequeñas y con variables cualitativas de tipo nominal u ordinal; toda variable cuantitativa o superior tipo Likert o de razón, se puede bajar de categoría y convertirla en nominal, para poder aplicar las pruebas no paramétricas, por ejemplo, la variable edad con una pregunta concreta de cuántos años tiene, arrojará respuesta concreta según la edad en años cumplidos del encuestado y a partir de esos datos se pueden hacer rangos de edad de 10 en 10 quedando en 5 rangos de la siguiente manera: rango 1 de 20 a 30 años de edad, 2 de 31 a 40, 3 de 41 a 50, 4 de 51 a 60 y 5 mayores de 61. Las técnicas más utilizadas en este apartado son, la Chi cuadrada, Kruskal-Wallis, Kolmogorov-Smirnov, Prueba del Signo, Prueba U de Mann Whitney, Prueba de la Mediana y Prueba de Wilcoxon, entre otras.

Hay tres preguntas que todo investigador se debe hacer para conocer las técnicas estadísticas a utilizar:

1. ¿Cuántas variables se deben analizar al mismo tiempo? En todo proyecto de investigación se recomienda primero realizar un análisis estadístico descriptivo de variable por variable, conocido como análisis univariado, para conocer las medidas estadísticas de centralidad como la media, la moda y mediana, y medidas de dispersión como el rango, la desviación media y la desviación estándar; después se recomienda realizar el análisis estadístico inferencial, para proyectar o predecir el comportamiento de esa variable hacia la población. Según la escala de medición será la técnica estadística a utilizar, por ejemplo, si la variable se mide en escala nominal como la variable sexo, solamente se podrá calcular frecuencias, porcentajes y gráficos para visualizar cuántos fueron mujeres y cuántos hombres. Es importante realizar este tipo de análisis porque ubica al investigador en el comportamiento de cada una de las preguntas del cuestionario y desde ese preciso momento se va visualizando el resultado final de la investigación.

2. ¿Quiere contestar preguntas descriptivas o inferenciales? Es importante desde un inicio del diseño estadístico del proyecto de investigación el esquematizar las diferentes técnicas estadísticas a implementar. Siempre es recomendable aplicar primero las técnicas estadísticas descriptivas para conocer el comportamiento de cada variable para dar paso a las estadísticas inferenciales gracias al planteamiento de las hipótesis de investigación o a las hipótesis estadísticas. Para inferir se parte de una muestra representativa de la población; las inferencias se pueden realizar para cada una de las variables de manera individual, con el cruce de dos variables o análisis bivariados y cuando se agrupan más de dos variables conocido como análisis multivariado. Para aplicar las técnicas estadísticas inferenciales se requiere conocer si las variables cumple el criterio de la normalidad, técnica que se aplica solo a las variables medidas en escalas superiores como Likert y/o de razón; si la variable se comporta como una distribución normal, se podrán realizar las pruebas estadísticas propias para inferir como: análisis de la T de Student, la correlación, análisis de regresión simple y múltiple, el ANOVA, ANCOVA, MANOVA, MANCOVA, regresión simple y múltiple, análisis multivariante, Análisis de Factores, Análisis de Clúster y otros. En variables que no cumplen la normalidad se recomienda aplicar las estadísticas no paramétricas, las cuales no requieren conocer los principios de la normalidad de las variables involucradas y en la mayoría de las veces en que las variables son medidas en escalas débiles nominales u ordinales, las pruebas más comunes y que se verán este libro son: Chi Cuadrada, Kruskal-Wallis, Prueba U de Mann Whitney, Prueba del Signo, Prueba de la binomial, y Kolmogorov Smirnov entre otras.
3. ¿Cuál es el nivel de medición de las preguntas del cuestionario o de las variables según las escalas nominales, ordinales, intervalares y de razón? Las escalas de medición marcan de manera directa el tipo de técnica estadística a utilizar, de ahí la importancia de la comprensión de las escalas para sus respectivos análisis univariado, bivariado y multivariado. Ver gráfica número 21.

Gráfica No. 21

## DISEÑO ESTADÍSTICO DESCRIPTIVO E INFERENCIAL

DESCRIBIR:	<p>MEDIDAS DE CENTRALIDAD: media, mediana y moda</p> <p>MEDIDAS DE DISPERSIÓN: rango, desviación media y desviación estándar.</p>
INFERIR:	<p>PARAMÉTRICA: se necesita conocer la distribución normal de las variables métricas o superiores con muestras aleatorizadas. Las técnicas más utilizadas en proyectos de investigación son: la t de Student, Correlación, ANOVA, ANCOVA, MANOVA y MACOVA, Regresión simple y múltiple, análisis discriminante, análisis de factores y de conjuntos, entre otros.</p> <p>NO PARAMÉTRICA: no se requiere conocer la distribución normal de las variables. Útiles cuando no se cumple el criterio de la normalidad de las variables. Las técnicas estadísticas más utilizadas son: Chi cuadrada, Krustal Wallis, Kolmogorov Smirnow, U de Mann Whitney, de Wilcoxon y del signo entre otras.</p>

Fuente: aportación del autor

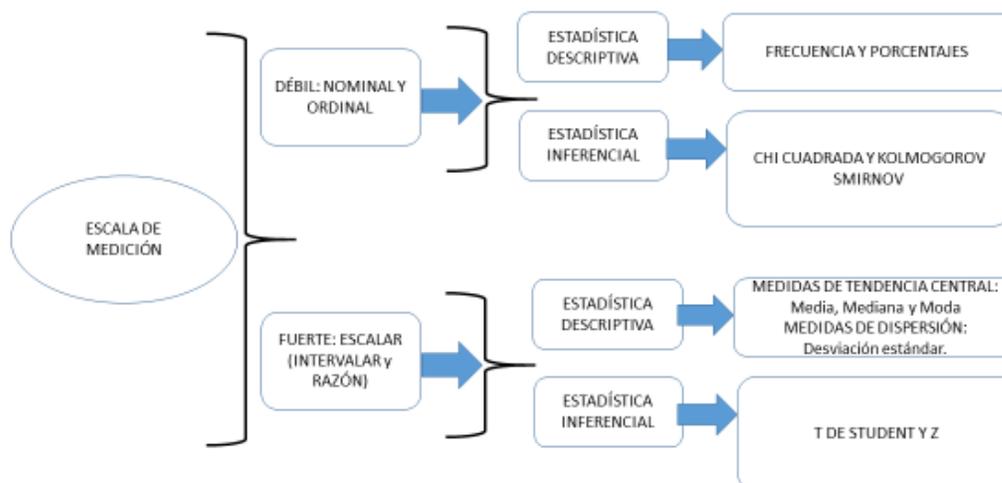
## A. ESTADÍSTICA UNIVARIADA DESCRIPTIVA E INFERENCIAL

Todas las preguntas de un cuestionario se deben analizar univariadamente, esto es variable por variable o pregunta por pregunta, para conocer el comportamiento de cada variable (estadística descriptiva) con sus frecuencias, porcentajes gráficas, medidas de centralidad como la media, mediana y moda, y las medidas de dispersión como el rango, la desviación media y la desviación estándar; se debe realizar estadística inferencial para contrastar las hipótesis univariada de la muestra hacia la población de dónde procede la muestra con técnicas estadísticas paramétricas como la T de Student para muestras menores a 30 y prueba Z para muestra mayores a 30, con variables medidas en escalas cuantitativas o superiores (intervalar o razón); las pruebas no paramétricas como Kolmogorov-Smirnov, para las variables medidas en escala ordinal y la Chi Cuadrada para las variables medias en escala nominal. En el gráfico 18 se puede observar el esquema que guía el camino a seguir para realizar una prueba estadística descriptiva e inferencial en función de las escalas de medición;

El procedimiento univariado para describir variables medidas en escala cualitativa o débil, es el cálculo de las frecuencias y porcentajes; para inferir se recomienda realizar la prueba de la Chi cuadra o Ji cuadrada. Si el camino a seguir en este mismo esquema es el de variables medidas en escala ordinal, se recomienda aplicar igual frecuencia y porcentajes para describir y la prueba de Kolmogorov Smirnov para inferir. Si la variable a seguir es superior, cuantitativa o escalar y/o de razón, para describir se sugieren las medidas de centralidad como la media, la moda y la mediana y de dispersión la desviación estándar y el rango; para inferir se sugiere la prueba Z y T de Student. Ver gráfica número 22.

Gráfica No. 22

## ANÁLISIS ESTADÍSTICO UNIVARIADO



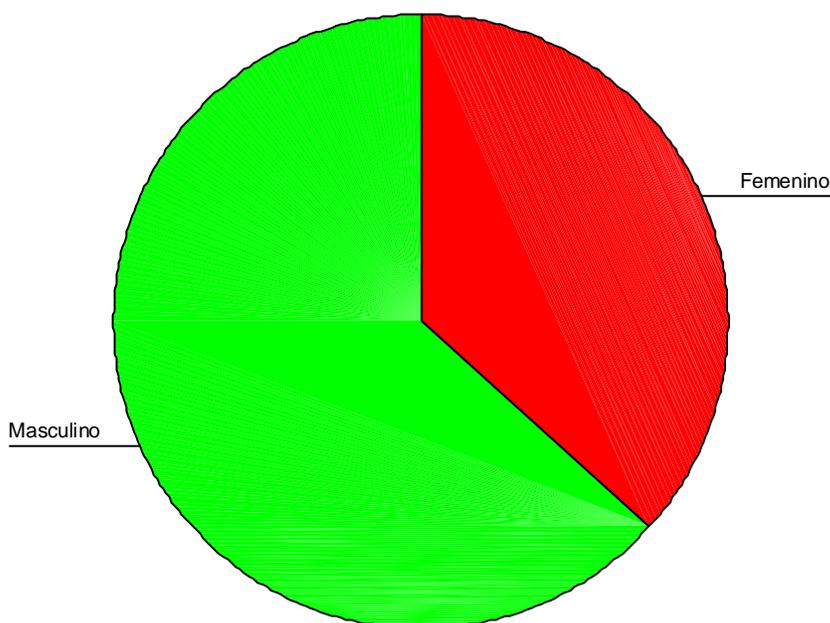
Fuente: aportación del autor.

## 1. UNIVARIADO - ESCALA NOMINAL

Para el análisis univariado se solicitarán las estadísticas descriptivas e inferenciales en función de las escalas de medición, ejemplo: Las variables medidas en escala nominal se le puede realizar la estadística descriptiva de frecuencias, porcentajes, gráficos, la razón y la tasa. Para la estadística inferencial se puede aplicar la técnica de la Chi o Ji cuadrada. Ver el siguiente resultado de una corrida estadística para la variable sexo.

Estadística Descriptiva: Frecuencia, Porcentajes, Porcentaje Acumulado y Gráficos.

Sexo de los encuestados



En el ejemplo de la variable sexo, se puede ver que hay 57 hombres y 33 mujeres, con una razón de 1.72 ( $R=57/33 = 1.72$ ), por lo que hay casi dos hombres por cada mujer, en la estadística Inferencial Univariada para variables medidas en escala nominal se procede con la prueba de la Chi-Cuadrada. Técnica estadística que compara los valores obtenidos de un cuestionario y los compara con la distribución de los datos teóricos, por ejemplo, en la variable sexo con 90 personas, la tabla teórica o de valor esperado ubicaría 45 mujeres y 45 hombres, y los datos obtenidos fueron: 33 femeninos y 57 hombres; se procede a realizar la prueba de la Chi

Cuadrada para ver si esa diferencia entre hombres y mujeres es realmente una diferencia en la población o solo sucede en la muestra (error de muestra o sesgo).

### Sexo de los encuestados

	N observado	N esperado	Residual
Femenino	33	45.0	-12.0
Masculino	57	45.0	12.0
Total	90		

Para el análisis es importante formular la hipótesis nula y la alterna ya que se pretende hacer una inferencia de esta muestra de 90 personas hacia el total de la población:

- Ho. “La proporción de género es igual” o “no hay diferencia entre la proporción de género”.
- Ha. “La proporción de género es diferente” o “si hay diferencia entre la proporción de género”.

Con un valor de Chi-cuadrada de 6.400, un grado de libertad de 1 y una significancia de .011, **(Si la significancia es  $\leq 0.05$ , rechaza Ho)** se rechaza Ho y se acepta Ha; por lo que se puede afirmar que la proporción de género en la población es diferente ya que la mayoría son hombres con el 63.3% y con un nivel de confianza del 95% para inferir de la muestra hacia la población, concluyendo que en la población la mayoría son de sexo masculino.

**Estadísticos de contraste**

	Sexo de los encuestados
Chi-cuadrado <sup>a</sup>	6.400
gl	1
Sig. asintót.	.011

- a. 0 casillas (.0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 45.0.

## 2. UNIVARIADO - ESCALA ORDINAL

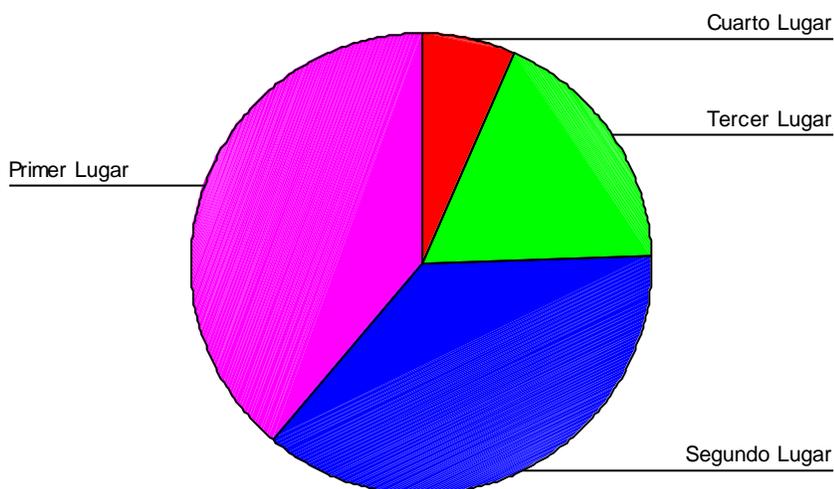
Esta escala nos permite la posibilidad de clasificar el atributo que más le agrada Vs el que menos le agrada al encuestado. Ejemplo: “Jerarquice según sus preferencias el lugar donde le gustaría cenar”.

- a. Estadística descriptiva: Frecuencia, Porcentajes, Porcentaje Acumulado y Moda

**Hogar como preferencia para cenar**

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos Cuarto Lugar	6	6.7	6.7	6.7
Tercer Lugar	16	17.8	17.8	24.4
Segundo Lugar	33	36.7	36.7	61.1
Primer Lugar	35	38.9	38.9	100.0
Total	90	100.0	100.0	

**Hogar como preferencia para cenar**



Como se puede observar en la tabla y en el gráfico de la variable “Hogar como preferencia para cenar”, el primer lugar de preferencia para cenar en casa fue seleccionado por el 38.9 % de los encuestados, el segundo lugar lo prefirieron el 36.7 %, el tercer lugar con 17.8 % y el cuarto lugar con tan solo el 6.7%

b. Prueba univariada para inferir: con la técnica estadística de Kolmogorov Smirnov. Se puede inferir este resultado de la muestra hacia la población:

Ejemplo en el IBM-SPSS: **analizar, pruebas no paramétricas, cuadro de diálogos antiguos, KS de una muestra, se pasa la variable, aceptar.**

Para el análisis de esta prueba se debe formular la hipótesis nula ( $H_0$ ) y la hipótesis alterna ( $H_a$ ):

- $H_0$ . No hay preferencia por cenar en el hogar (Siempre se plantea como una igualdad o negación, en este caso se planteó como una negación)
- $H_a$ . Si hay preferencia por cenar en el hogar (Siempre se plantea al contrario que la  $H_0$ , con una afirmación o bien con una desigualdad)

En el análisis se puede concluir que se rechaza  $H_0$ , con un valor Z de Kolmogorov-Smirnov de 2.203 y con una significancia de .000 (**Si la significancia es  $\leq 0.05$ , rechaza  $H_0$** ), por lo que se tiene la certeza que la mayoría prefiere cenar en el hogar. Esta información se puede corroborar con el análisis de porcentajes y frecuencia en el que muestra al hogar en el primer lugar con un 38.9% y segundo lugar con un 36.7%, por lo que la media es de 1.9222. Esta hipótesis se rechaza con el 95% de confianza. La diferencia absoluta entre el valor esperado y los valores observados es de .232, que arroja una Z de KS de 2.203 (A mayor valor entre la diferencia absoluta se impactará con mayor valor de Z de KS y por consecuencia se tiene una significancia menor de .05- Valor suficiente para rechazar la hipótesis nula-)

La prueba de la Chi-Cuadrada es otra de las técnicas estadísticas no paramétricas para inferir de la muestra hacia la población:

### Prueba de Kolmogorov-Smirnov para una muestra

		Hogar como preferencia para cenar
N		90
Parámetros normales <sup>a,b</sup>	Media	1.9222
	Desviación típica	.9146
Diferencias más extremas	Absoluta	.232
	Positiva	.232
	Negativa	-.157
Z de Kolmogorov-Smirnov		2.203
Sig. asintót. (bilateral)		.000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

### Estadísticos de contraste

	Hogar como preferencia para cenar
Chi-cuadrado <sup>a</sup>	25.822
gl	3
Sig. asintót.	.000

a. 0 casillas (.0%) tienen frecuencias esperadas menores que 5. La frecuencia de casilla esperada mínima es 22.5.

Con la Chi-Cuadrada también se rechaza la misma hipótesis ( $H_0$ ) y se concluye de la misma manera.

La hipótesis por contrastar es la misma que en el análisis anterior:

$H_0$ : No hay preferencia por cenar en el hogar (Formulada en forma de igualdad)

$H_a$ : Si hay preferencia por cenar en el hogar (Formulada en forma de desigualdad)

Con valor de Chi-cuadrada de 25.822, 3 grados de valor y significancia de .000, se rechaza la hipótesis nula ( $H_0$ ) y se acepta la hipótesis alterna ( $H_a$ ). Si hay preferencia por ir a cenar al hogar.

### Hogar como preferencia para cenar

	N observado	N esperado	Residual
Primer Lugar	35	22.5	12.5
Segundo Lugar	33	22.5	10.5
Tercer Lugar	16	22.5	-6.5
Cuarto Lugar	6	22.5	-16.5
Total	90		

### 3. UNIVARIADO - ESCALA INTERVALAR Y DE RAZÓN

A todas las variables medidas en esta escala se le puede realizar la estadística propia de las escalas nominales, ordinales y las propias de esta escala como las medidas de dispersión y centralidad en la parte descriptiva y la prueba T de Student o Z para realizar inferencia estadística.

- a. Estadística descriptiva: Frecuencia, Porcentajes, Porcentaje Acumulado, medidas de centralidad (Media, Mediana y Moda), medidas de dispersión (Desviación Media, Desviación Estándar, Valor Mínimo, Máximo y Rango). Curtosis, asimetría, cuartiles, deciles y percentiles.

#### Años cumplidos del entrevistado

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	21	8	8.9	8.9	8.9
	22	7	7.8	7.8	16.7
	23	8	8.9	8.9	25.6
	24	6	6.7	6.7	32.2
	25	6	6.7	6.7	38.9
	26	3	3.3	3.3	42.2
	27	7	7.8	7.8	50.0
	28	3	3.3	3.3	53.3
	29	5	5.6	5.6	58.9
	30	7	7.8	7.8	66.7
	31	9	10.0	10.0	76.7
	32	13	14.4	14.4	91.1
	33	4	4.4	4.4	95.6
	34	4	4.4	4.4	100.0
	Total	90	100.0	100.0	

## Estadísticos

### Años cumplidos del entrevistado

N	Válidos	90
	Perdidos	0
Media		27.43
Error típ. de la media		.44
Moda		32
Desv. típ.		4.17
Asimetría		-.108
Error típ. de asimetría		.254
Curtosis		-1.413
Error típ. de curtosis		.503
Mínimo		21
Máximo		34
Percentiles	25	23.00
	50	27.50
	75	31.00

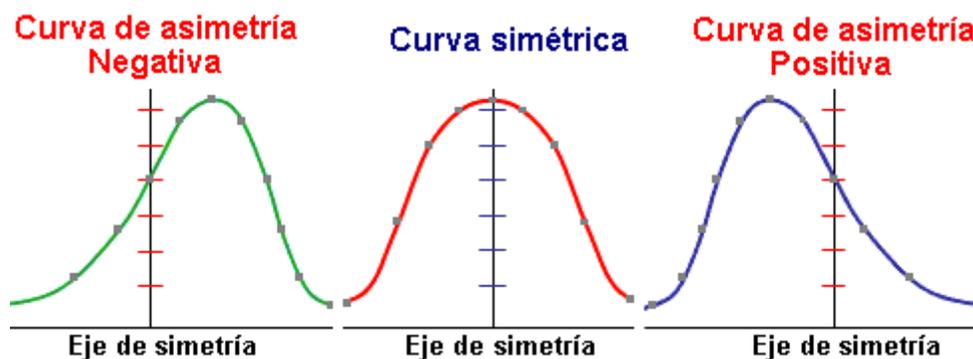
La asimetría y la Curtosis muestra la forma de una distribución de datos en relación a la media, por lo que se puede ver la dispersión o aglomeración de los valores de una variable métrica en relación a la media.

## Estadísticos

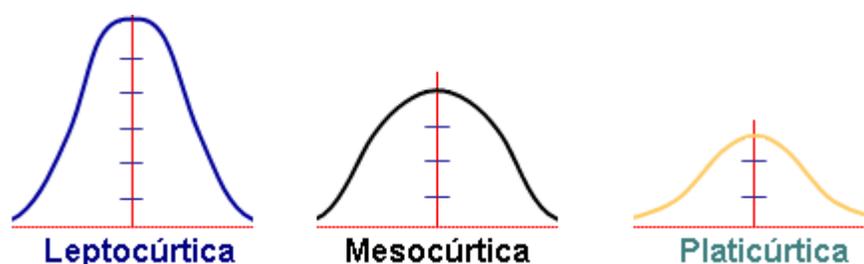
### Años cumplidos del entrevistado

N	Válidos	90
	Perdidos	0
Media		27.43
Error típ. de la media		.44
Moda		32
Desv. típ.		4.17
Asimetría		-.108
Error típ. de asimetría		.254
Curtosis		-1.413
Error típ. de curtosis		.503
Mínimo		21
Máximo		34
Percentiles	25	23.00
	50	27.50
	75	31.00

La distribución simétrica se da cuando los datos se distribuyen de manera uniforme alrededor de la media, esto es, se tiene la misma cantidad de datos en los dos lados de la media de una distribución de datos, con valor 0. La asimetría muestra el grado de dispersión de los datos en relación a la media, el eje de asimetría puede ser positiva, cuando la mayoría de los datos se encuentran por arriba de la media (valor mayor a 1) y negativa cuando la mayoría de los datos se encuentra por debajo de la media (valor menor a 1). Si el signo es negativo muestra que la mayoría de los datos están concentrados hacia la derecha y si son positivos indican que los datos están concentrados hacia la izquierda.



La Curtosis muestra el grado de concentración de los valores de los datos en relación a la media de una distribución. Se puede conocer si hay gran concentración de los valores en la media (Leptocúrtica; valor mayor a 2), concentración normal (Mesocúrtica; valor=2) o baja concentración o plana (Platicúrtica; valor menor a 2). El cero indicaría que la altura tiene la forma de una distribución normal. Si el signo es negativo indicaría que los datos están más dispersos, el signo positivo indica mayor agrupación de datos. Una distribución muy dispersa se conoce como Platicúrtica, si es muy agrupada se conoce como Leptocúrtica y si la altura es normal se conoce como Mesocúrtica. El cero indica que la distribución es Mesocúrtica.



Para calcular la normalidad en el IBMSPSS se procede: **analizar, explorar, gráficos con prueba de normalidad, aceptar. El paquete arrojará una tabla de estadísticos descriptivos y la prueba de normalidad.** Si la Significancia es mayor o igual .05, la variable sigue la distribución normal.

### Distribución normal con cuartiles, deciles, percentiles y la probabilidad:

Cuando los datos siguen el comportamiento de la distribución normal, se puede conocer con exactitud la proporción o el porcentaje de uno de los datos entre el total de los mismo, expresados en cuartiles (Q), deciles (D) o percentiles (P). Los cuartiles se dan cuando se divide el total de los datos en 4 partes iguales, deciles cuando se divide el total de los datos en 10 partes iguales y percentiles cuando se dividen en 100. En los siguientes datos de la estatura de estudiantes de una universidad ordenados de mayor a menor se puede ver la utilidad de los cuartiles, deciles y percentiles:

NOMBRE	ESTATURA	CUARTILES. Kn/4 (Posición)	Deciles. Kn/10 (Posición)	Percentiles Kn/100 (Posición)
1 Miller	188			P1
2 Feibert	185		D1	
3 Norberto	177			
4 Salomón	175			
5 Juancho	174		D2	
6 David	173	Q1		
7 Luis	170		D3	P30
8 Edisney	168			
9 Cristian	167		D4	
10 Ayan	166			
11 Manuela	165			
12 Camila	165	Q2	D5	P50
13 Edison	164			
14 Edis	163		D6	
15 Gloria	159			
16 Martha	158			
17 Patricia	157	Q3		
18 Lucila	156		D8	P80
19 Lucia	155			
20 Paula	152			P90
21 Yolima	151			
22 Elvis	150			
23 Luz	149	Q4	D10	P100

Para el cálculo del primer cuartil (Q1) se utiliza la fórmula  $Kn/4$ , siendo K el cuartil a calcular, n el total de la muestra, en este caso son 23 y 4 es la división de los cuatro cuartiles.  $Q1 = 1 \cdot 23/4 = 5.75$ , aproximando se sube a 6 y se procede a ubicar la

persona que está en la posición 6, en este caso es el estudiante David, con una estatura de 173 centímetros, quiere decir que el 25% de los estudiantes tienen estatura igual o superior a David, siendo solo 6 de los 23 estudiantes.

El cuartil 2 (Q2) se calcula:  $Q2 = 2 \cdot 23 / 4 = 11.5$ , aproximando el valor se sube a la posición 12, que es ocupada por la estudiante Camila, con una estatura de 165 centímetros, siendo ella el 50% de la concentración de estudiantes o datos.

El cuartil 3 (Q3) :  $Q3 = 3 \cdot 23 / 4 = 17.25$ , aproximando se quedaría con el valor 17 y se procede a ver quién está ocupando la posición 17, en este ejemplo, la ocupa la estudiante Patricia con una estatura de 157 centímetros, siendo ella el 75% del total de los estudiantes con una estatura igual o menor a la de ella. Al final la estudiante Luz, ocupa la posición correspondiente al Q4, que representa el cien por ciento de los datos con la estatura de 149 centímetros.

El decil 1 por fórmula se calcula:  $D1 = Kn / 10$ :  $D1 = 1 \cdot 23 / 10 = 2.3$ , aproximando el valor se queda con la posición 2, por lo que se procede a ubicar a la persona que ocupa esta posición, siendo el estudiante Feibert, con una estatura de 185 centímetros, lo que significa que el 10% del total de la población tiene estatura igual o superior a 185 centímetros. El decil 2 lo tiene el estudiante Juancho ( $D2 = 2 \cdot 23 / 10 = 4.6$ ), representa el 20% del total de los 23 estudiantes. El decil 3, lo ocupa el estudiante Luis, con el 30% de los datos con estatura promedio debajo de 170 centímetros. El decil 5 lo ocupa la estudiante Camila, con la estatura de 165, siendo la misma estudiante que tiene el cuartil 2 (Q2) y el percentil 50.

El percentil se calcula con la misma fórmula pero dividiendo entre 100, ejemplo: el percentil 30 (P30) lo ocupa el estudiante Luis, que coincide con el decil 3 (D3), que representa el 30% de la población tiene estatura igual o mayor a 170 centímetros.

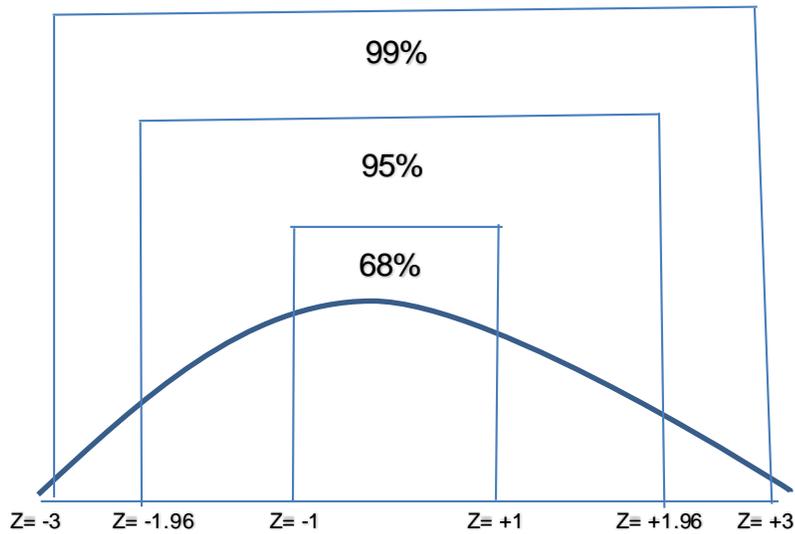
Con los mismos datos de la estatura se calculan los valores Z dentro de una distribución normal para conocer la probabilidad de ocurrencia. Para ello se necesita calcular las medidas de centralidad como la media ( $\bar{X}$ ), la moda ( $M_o$ ) y la mediana ( $M_e$ ) y las medidas de dispersión como el rango (R), la desviación media (DM) y la desviación estándar ( $\sigma$ ) para integrarlas en la fórmula de los valores  $Z = (X - \bar{X}) / \sigma$ ; las medidas de centralidad son:  $\bar{X} = 165$ , Moda = 165 y Mediana = 165; las medidas de dispersión son: R = 39, DM = 8.62 y  $\sigma = 10.50$

NOMBRE	ESTATURA X	$(X - \bar{X})$	Valor Z $Z = (X - \bar{X}) / \sigma$	Valor Tabla A
1 Miller	188	23.35	2.22	0.4868
2 Feibert	185	20.35	1.94	0.4738
3 Norberto	177	12.35	1.18	0.3810
4 Salomon	175	10.35	.99	0.3389
5 Juancho	174	9.35	.89	0.3133
6 David	173	8.35	.80	0.2881
7 Luis	170	5.35	.51	0.1950
8 Edisney	168	3.35	.32	0.1255
9 Cristian	167	2.35	.22	0.0871
10 Ayari	166	1.35	.13	0.0517
11 Manuela	165	.35	.03	0.0120
12 Camila	165	.35	.03	0.0120
13 Edison	164	-.65	-.06	0.0239
14 Edis	163	- 1.65	-.16	0.0636
15 Gloria	159	- 5.65	-.54	0.2054
16 Martha	158	- 6.65	-.63	0.2357
17 Patricia	157	- 7.65	-.73	0.2673
18 Lucila	156	- 8.65	-.82	0.2939
19 Lucia	155	- 9.65	-.92	0.3212
20 Paula	152	- 12.65	-1.20	0.3849
21 Yolima	151	- 13.65	-1.3	0.4032
22 Elvis	150	- 14.65	-1.4	0.4192
23 Luz	149	- 15.65	-1.5	0.4332

Los valores Z se calculan con la fórmula:  $Z = (X - \bar{X}) / \sigma$ , para ir a la tabla A (Anexo 2 de este libro) para conocer la probabilidad que ocupa ese dato, por ejemplo, el estudiante Miller, tiene 0.4868; el 49% de los datos con estatura arriba de la media se encuentran entre 165 a 188 centímetros o bien el 99% de los datos son menores a 188 centímetros. El valor Z de la estudiante Lucia de -0.82, con un valor tabla de 0.3212, equivale al 32% de la población con estatura igual o menor de 156 centímetros y solo el 18% de la población total tiene estatura menor a 156 ( $0.3212 - .5 = 0.1788$ ).

Entre el valor Z del estudiante Norberto ( $Z=1.18$ ) y de Paula ( $Z=-1.20$ ), se encuentra el 77% de la población ( $.3810 + .3849 = 0.7659$ ; redondeando quedaría en 77%).

Para mayor comprensión de los valores Z dentro de una distribución normal se han identificado los valores de  $Z = 1$  y  $Z = -1$  que integran el 68% de la población; valores de  $Z = 1.96$  y  $Z = -1.96$ , que integran el 95% de los casos y valores de  $Z = 3$  y  $Z = -3$ , que integran el 99% de los casos.



Las medidas de centralidad de mayor uso en las escalas intervalares o de razón son la media aritmética ( $\bar{X}$ ), la mediana (Me) y la moda (Mo).

La media se calcula con la sumatoria de todos los datos y divididos entre el total de la población, por ejemplo, con los siguientes 7 datos de salto largo medidos en metros de los concursantes en este deporte se tiene:

Concursante	Salto largo en metros (X)
Miller	10
Feibert	8
Salomón	7
Luis	6
Edisney	6
Cristian	3
Rodrigo	2

$$\bar{X} = \sum X/N$$

$$= 42/7$$

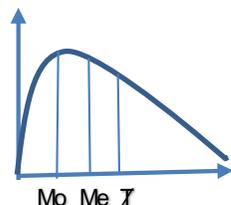
$$= 6$$

La Moda es el valor que más se repite, en este ejemplo es 6. Pueden existir datos con más de dos modas y se conoce como multimodal.

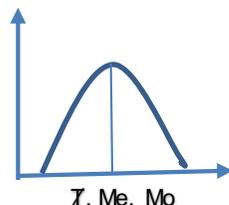
La mediana es el valor que divide a la población en dos grandes grupos, en este ejemplo, el valor es 6, ocupado por el concursante Luis, quien es el que divide al total de la muestra en dos grandes grupos, tres arriba de la media y tres abajo de la media.

En este ejemplo los tres valores de centralidad coinciden, por lo que se cumple el requisito de la normalidad, pero se recomienda tener muestras mayores a 45 datos para tener mayor consistencia en la distribución normal y poder inferir hacia la población.

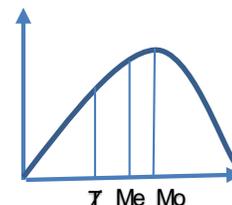
**¿Cuándo es recomendable utilizar una u otra medida de centralidad?** Se recomienda utilizar casi siempre la mediana, ya que este valor divide todos los datos justo en su punto medio, mientras que la media puede mostrar datos muy altos o muy bajos del total de los datos que llevaría a tomar decisiones equivocadas. La mediana (Me) siempre estará en el medio de la media ( $\bar{X}$ ) y la moda (Mo).



Distribución asimétrica positiva



Distribución normal



Distribución asimétrica negativa

**¿Cuáles son las principales medidas de dispersión a utilizar en un proyecto de investigación?** Siguiendo con este mismo ejemplo, ahora se procede a calcular las tres medidas de dispersión: Rango(R), Desviación Media (DM) y la Desviación Estándar ( $\sigma$ ).

Concursante	Salto largo (X)	$X-\bar{X}$	$(X-\bar{X})^2$
Miller	10	4	16
Feibert	8	2	4
Salomón	7	1	1
Luis	6	0	0
Edisney	6	0	0
Cristian	3	-3	9
Rodrigo	2	-4	16
SUMA		14	46

Rango: la diferencia entre el valor más alto y el menor

$$R = 10-2$$

$$R = 8$$

El rango es un número muy alto y de difícil comprensión, pero da orientación que la dispersión de los datos están directamente proporcional al rango, entre más alto sea el rango la variabilidad de los datos será mayor.

La otra medida de dispersión se conoce como la desviación media (DM), que se calcula sumando la diferencia de los valores de X con la media ( $\bar{X}$ ) sin tener en cuenta el valor de los signos entre el total de la población.

$$DM = \sum(X-\bar{X})/N$$

$$DM = 14/7$$

$$DM = 2$$

Este valor es más comprensible y da idea de la dispersión del total de los datos, pero no ha sido de mayor uso en las investigaciones en las ciencias sociales y económicas administrativas, de ahí el avance hacia el cálculo de la desviación estándar.

Desviación estándar ( $\sigma$ ), es una evolución de la desviación media (DM), ya que para su cálculo, se eleva al cuadrado la diferencia entre  $x-\bar{X}$  para la respectiva sumatoria que se elevara al cuadrado entre N y se aplica la raíz cuadrada para regresar los datos a su valor original, ejemplo:

$$\text{Desviación estándar } (\sigma) = \sqrt{\sum(X-\bar{X})^2/N}$$

$$= \sqrt{46/7}$$

$$= 2.56 \text{ metros}$$

Siendo esta cifra más precisa para conocer el grado de dispersión de los datos, entre más pequeño sea este dato refleja poca variabilidad de los mismos. Cuando la desviación estándar es cero, se interpreta que los datos son homogéneos.

Cuando se calcula la razón (R) el valor es 8, cuando se calcula la desviación media (DM) el valor es 2 y con la desviación estándar ( $\sigma$ ) el valor es de 2.56. En la práctica la medida de dispersión entre datos numéricos que más se utiliza en las ciencias sociales y económicas administrativas es la desviación estándar.

Con los mismos datos se puede calcular la desviación estándar con la fórmula de los puntajes crudos o no procesados con la desviación media, ejemplo:

Concursante	Salto largo (X)	X <sup>2</sup>
Miller	10	100
Feibert	8	64
Salomon	7	49
Luis	6	36
Edisney	6	36
Cristian	3	9
Rodrigo	2	4
SUMA		298

$$\text{Desviación estándar } (\sigma) = \sqrt{\sum X^2/N - \bar{X}^2}$$

$$= \sqrt{298/7 - 36}$$

$$= 2.56 \text{ metros}$$

- b. Estadística inferencial: La prueba T de Student para una muestra comparada con un patrón de la población, ejemplo:

### Estadísticos para una muestra

	N	Media	Desviación típ.	Error típ. de la media
Años cumplidos del entrevistado	90	27.43	4.17	.44

### Prueba para una muestra

	Valor de prueba = 24					
	t	gl	Sig. (bilateral)	Diferencia de medias	Intervalo de confianza para la diferencia	
					Inferior	Superior
Años cumplidos del entrevistado	7.805	89	.000	3.43	2.56	4.31

Para realizar el análisis de inferencia estadística es necesario elaborar la  $H_0$  y definir la medida de la población, que generalmente sale de la información de datos secundarios; para este ejercicio se le dijo al paquete que la media de la población de donde se extrajo la muestra es de 24 años (El paquete da la libertad para especificar el valor oficial).

$H_0$ : La edad de la muestra es igual a la edad de la población

$H_a$ : La edad de la muestra es diferente a la edad de la población

Para este ejercicio se rechaza  $H_0$  (**Si la significancia es  $\leq 0.05$ , rechaza  $H_0$** ) y se acepta  $H_a$ , en la que la edad de 27.43 años de la muestra es muy diferente a la media de la población (24 años), con un valor t de 7.905, 89 grados de libertad y la significancia de .000, con una diferencia entre las medias de 3.43.

Nota: A medida que se avanza en las escalas de medición de nominal a intervalar se puede utilizar todas las técnicas estadísticas que se utilizan en nominal y también las de ordinal como la prueba de Kolmogorov-Smirnov y así sucesivamente a

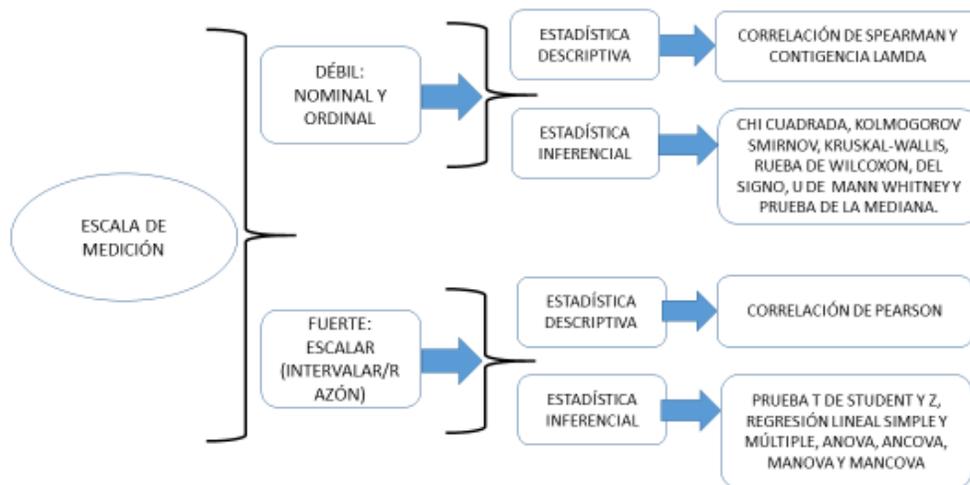
medida que se avanza en la escala de medición. Las escalas intervalares y de razón se pueden bajar de categoría o convertirlas en nominales para aplicar las estadísticas propias de las escalas nominales y ordinales, ejemplo, la escala Likert con 5 posibilidades de respuesta, siendo 1 totalmente en desacuerdo, 2 en desacuerdo, 3 indeciso, 4 de acuerdo y 5 totalmente de acuerdo, se pueden agrupar de manera nominal sumando las respuestas de 1 y 2 con una nueva etiqueta que sería no y se sumarían los datos obtenidos en 4 y 5 con una nueva etiqueta de si, convirtiendo así la escala Likert en nominal con respuestas de sí y no. Las variables medidas en escala de razón también se pueden bajar de categoría, por ejemplo, la variable ingresos, con la pregunta concreta de cuántos salarios mínimos al mes gana usted, se obtiene datos diferentes en cada persona, por lo que se sugiere convertirla en nominal trabajando rangos, ejemplo, dentro de que rango de salarios mínimos al mes están sus ingresos mensuales: 1 de 1 a 3 salarios mínimos al mes, 2 de 4 a 6 salarios mínimos al mes, 3 de 7 a 9 salarios mínimos al mes y 4 con más de 10 salarios mínimos al mes.

## B. ESTADÍSTICA BIVARIADA

En este apartado se cruzarán variables medidas en diferentes escalas de medición. Se selecciona el método asintótico (cuando hay muestras mayores a 20 personas con el 20% de valores menores o iguales a 5 personas por casilla y donde no tenga casillas con valores menores a 1). El método Monte Carlo, se aplica para muestras mayores a 30 sin importar los supuestos anteriores y método Exacto no requiere de nada de lo anterior. De las hipótesis planteadas se procede a definir las variables involucradas son sus respectivas escalas de medición, para ver que técnica estadística sería recomendable realizar; si las dos variables están medidas en escala nominal, se recomienda realizar la estadística descriptiva con el coeficiente de correlación de contingencia lambda y para inferir se recomienda realizar la prueba de la Chi o Ji Cuadrada, observar en el lado derecho de la gráfica 19, por ejemplo, identificar la diferencia significativa entre la variable nominal sexo y la variable nominal volumen de compra codificada en compra mucho, regular y poco, para contrastar la hipótesis:  $H_0$ . No hay diferencia significativa entre el sexo y el volumen de compra. Si las dos variables a analizar están medidas en escala de intervalo y/o razón, para la estadística descriptiva paramétrica que se debe aplicar es el coeficiente de correlación lineal de Pearson y para inferir se recomienda aplicar las pruebas de estadística paramétrica T de Student, regresión lineal simple y múltiple, el ANOVA, MANOVA y ANCOVA. La T de Student es útil para contrastar hipótesis de diferencia entre medias de dos grupos independientes, por ejemplo, la diferencia entre la media de la variable volumen de compra medida en escala de razón y el sexo medida en escala nominal; también la T de Student es útil para medir la diferencia de medias entre una medida antes y después y para medir la media de una muestra comparada con un valor estándar. Cuando se necesita medir la diferencia de medias de 3 o más grupos en relación a una variable dependiente se recomienda el ANOVA, por ejemplo, conocer la diferencia de medias entre los estratos socioeconómicos alto, medio y bajo en relación al volumen de compra medida en escala de razón, con la hipótesis nula:  $H_0$ . No hay diferencia significativa entre los estratos socioeconómicos y el volumen de compra. En la parte central del gráfico se puede observar que las pruebas estadísticas a utilizar para variables medidas en escala ordinal son las pruebas no paramétricas de Wilcoxon, del Signo, U de Mann Whitney, de la mediana, Kolmogorov Smirnov y Kruskal-Wallis, Ver gráfica número 23.

Gráfica No. 23

## ANÁLISIS ESTADÍSTICO BIVARIADO



Fuente: aportación del autor.

## 1. ESCALA NOMINAL: PRUEBA DE CHI CUADRADA

*(En el IBM-SPSS, estadísticos descriptivos, tablas cruzadas, pasar en filas una de las variables y en columnas la otra, en estadísticos seleccionar Chi-Cuadrado y Coeficiente de contingencia, continuar, en casillas seleccionar porcentajes en fila y columna, continuar y en Mostrar los gráficos de barras agrupadas, aceptar: entre la variable Sexo Versus tipo de maestría que están estudiando.*

**Tabla de contingencia Sexo de los encuestados \* Tipo de Maestría que está estudiando**

Recuento

		Tipo de Maestría que está estudiando			Total
		Presencial	Virtual	On Lines	
Sexo de los encuestados	Femenino	9	6	18	33
	Masculino	21	24	12	57
Total		30	30	30	90

### Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	11.196 <sup>a</sup>	2	.004
Razón de verosimilitud	11.232	2	.004
Asociación lineal por lineal	5.749	1	.016
N de casos válidos	90		

a. 0 casillas (.0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 11.00.

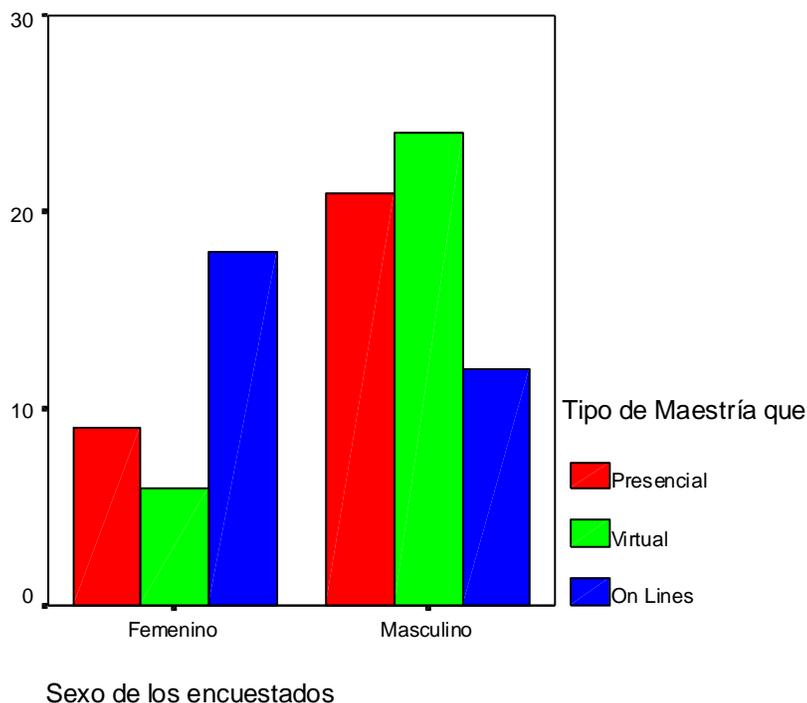
## Medidas simétricas

		Valor	Error típ. asint. <sup>a</sup>	T aproximada <sup>b</sup>	Sig. aproximada
Nominal por nominal	Phi	.353			.004
	V de Cramer	.353			.004
	Coefficiente de contingencia	.333			.004
Intervalo por intervalo	R de Pearson	-.254	.105	-2.465	.016 <sup>c</sup>
Ordinal por ordinal	Correlación de Spearman	-.254	.107	-2.465	.016 <sup>c</sup>
N de casos válidos		90			

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

c. Basada en la aproximación normal.



Para el análisis se requiere formular las hipótesis:

- Ho: El sexo no guarda relación significativa con la preferencia del tipo de maestría
  - Ha: El sexo si guarda relación significativa con la preferencia del tipo de maestría
- Como la significancia es menor de .05 (**Si la significancia es  $\leq 0.05$ , rechaza Ho**) se rechaza Ho y se acepta Ha debido a que los hombres prefieren la maestría

presencial y virtual, mientras que las mujeres prefieren la maestría online. La relación entre estas dos variables es del 35.3%. Esta información se puede visualizar gráficamente.

- a. En variables nominales las pruebas descriptivas e inferenciales bivariadas son el coeficiente de asociación con el Valor de Phi, que indica el grado de asociación entre las variables en tablas de 2x2 (Puede tomar valores entre 0 y la raíz cuadrada de  $q-1$ ;  $q$  es el número de factores o casillas de la variable que tenga menores de ellas) y  $V$  de Kramer (grado de asociación entre dos variables en tablas mayores a 2x2, que va de 0 a 1), o Coeficiente de Contingencia que representa el grado de asociación entre las variables que va de 0 a un valor máximo. El riesgo Odds ratio, indica el tamaño del efecto en tablas de 2x2. Las asociaciones están en función de:

#### ASOCIACIÓN EN ESCALA NOMINAL:

- En correlaciones: arroja los valores de Pearson y Spearman.
- Coeficiente de contingencia: cuando las dos variables son más de 2x2 u ordinales.
- Phi y  $V$  de Cramer: cuando las tablas de comparación son de 2x2 (Phi) y  $V$  de Cramer en tablas mayores de 2x2.
- Lambda: Para tablas mayores de 2x2 (asimétricas y simétricas)
- Coeficiente de incertidumbre: Tablas mayores de 2x2

#### ASOCIACIÓN EN ESCALA ORDINAL:

- Gamma: para dos variables ordinales, es una extensión de Spearman.
- $D$  de Somers: para dos variables ordinales, es una extensión de Spearman.
- Tau-b de Kendall: Cuando las dos variables tienen las mismas medidas, por ejemplo, tablas de 4x4
- Tau-c de Kendall: también para tablas mayores de 2x2, pero sin importar si son similares por ejemplo en tablas de 2x3 u otras.

## ASOCIACIÓN EN NOMINAL CON UNA INTERVALAR

- Eta: es el coeficiente de relación cuando se pretende asociar una medida numérica con una nominal, viene siendo una extensión de la T de Student.
- El estadístico de Cochran y Mantel-Haenszel: Medida de asociación para tres variables, cuando una de ella es variable intermedia o de control.

- b. PRUEBA McNEMAR** Es una prueba no paramétrica que permite medir datos en escala nominal de antes y después, apoyados con la observación o la experimentación y con una tabla de 2x2 para identificar si hay diferencia significativa en la medida de antes y después. *(En el IBM-SPSS: Analizar, estadísticos descriptivos, tablas cruzadas, filas antes y columna después, en estadísticos se selecciona McNemar y en casillas seleccionar porcentaje por filas, total y aceptar)*. Se recomienda en medidas de antes y después (Pre o Pos test), según el siguiente ejemplo de la tabla número 3.

Tabla No. 3

### ANTES Y DESPUES DEL TALLER

ANTES DESPUES	SABE	NO SABE	TOTAL
SABE	3	2	5
NO SABE	90	5	95
TOTAL	93	7	100

Por ejemplo, ver si el taller ayudó para que los estudiantes aprendieran después de la capacitación; se aplica un examen inicial y el mismo examen al final del curso. Si hay significancia se rechazará  $H_0$  y se determina el porcentaje de cambio del grupo de personas entre antes y después:

$H_0$ : No hay diferencia entre los estudiantes que saben antes y después del taller.

$H_a$ : Si hay diferencia entre los estudiantes que saben antes y después del taller.

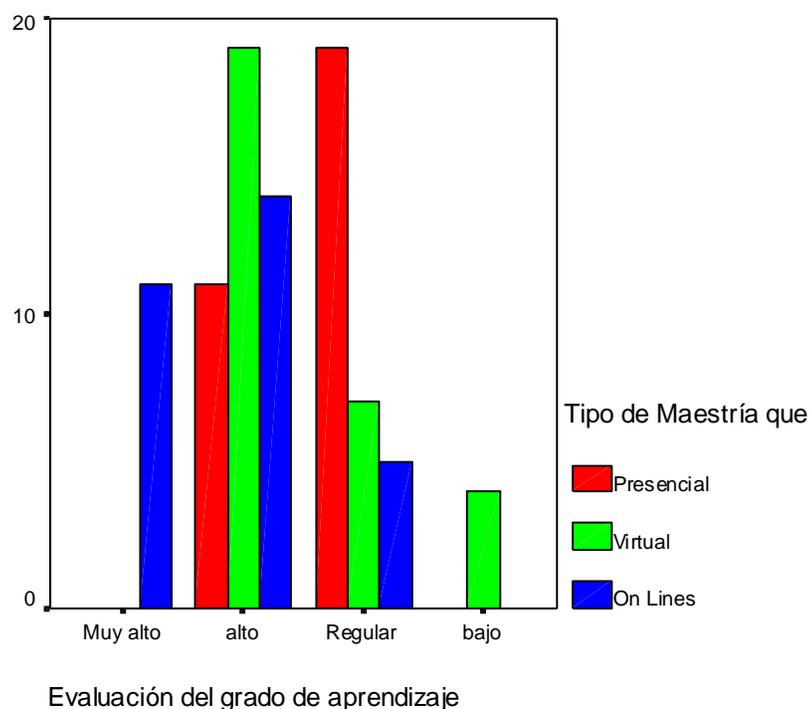
Se puede observar en la tabla de datos número 1, que las personas que sabían antes de la capacitación eran tan solo 3 y después de la capacitación subió a 90 de un total de 95. Por lo que se puede concluir que la capacitación es útil para incrementar el nivel de conocimiento de las personas en un 95 por ciento de las veces.

- c. Kappa de Cohen: Se recomienda cuando se realizan medidas a una unidad con diferentes instrumentos de medición, para identificar el grado de concordancia entre las dos medidas. **Para el IBM-SPSS, se va la ventana analizar, cruce de variables, estadístico en cruce de tablas seleccionar Kappa.** El valor va de 0 a 1 (.00 a .2 Mínima concordancia; .21 a .40 Poca concordancia; .41 a .60 moderada concordancia; .61 a .80 buena concordancia y .81 a 1 mucha concordancia).

## 2. BIVARIADA - ESCALA ORDINAL Vs NOMINAL (Análisis de la varianza de Kruscal Wallis)

**(En IBM-SPSS: Analizar, pruebas no paramétricas, cuadro de diálogo antiguos, k muestras independientes, se pasa la variable categórica en variable de agrupación y selecciona los rangos (en este ejemplo es 1 y 3), y pasa la variable de jerarquización en lista de variables de prueba, se selecciona H. de Kruskal-Wallis y aceptar)**

Se requiere de una escala medida en ordinal como el grado de aprendizaje (Muy alto, alto, regular y bajo) y una escala nominal como es el tipo de aprendizaje presencial, online y virtual:



## Medidas simétricas

		Valor	Error tip. asint. <sup>a</sup>	T aproximada <sup>b</sup>	Sig. aproximada
Ordinal por ordinal	Gamma	-.616	.099	-5.290	.000
N de casos válidos		90			

a. No asumiendo la hipótesis nula.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Se puede ver que hay una relación del 62 % (.616) y que esta relación se presenta en la población con una significancia de .000, por lo que se rechaza la hipótesis nula con un nivel de confianza del 95%.

Ho: La evaluación del grado de aprendizaje es el mismo en cada uno de los tipos de maestría.

Ha: La evaluación del grado de aprendizaje es diferente en los tipos de maestría.

Cómo la significancia es menor a .05, se rechaza Ho y se acepta Ha, en la que se concluye que la evaluación del grado de aprendizaje es muy alto en la maestría Online, alto en la virtual y de regular en el formato presencial, por lo que en online se obtiene la mayor evaluación del aprendizaje para estudiar una maestría.

### 3. BIVARIADA - ESCALA INTERVALAR Y RAZON

- a. Descriptivo e inferencial Bivariada: Las variables deben estar medidas de preferencia en la misma dimensión de la escala. Ejemplo: Escalas Likert con medidas de 1 al 5, del 1 al 7 o del 1 al 9.

**Correlaciones**

		Evaluación de la disciplina de autoestudio	Evaluación del grado de aprendizaje
Evaluación de la disciplina de autoestudio	Correlación de Pearson Sig. (bilateral) N		
Evaluación del grado de aprendizaje	Correlación de Pearson Sig. (bilateral) N	.315** .002 90	

\*\* . La correlación es significativa al nivel 0,01 (bilateral).

Se puede ver que existe una correlación del 31.5 % entre las variables, con una significancia de .002, por lo que se puede inferir que esta relación también se presenta en la población con un nivel de confianza del 95 %; **(Si la significancia es  $\leq 0.05$ , rechaza  $H_0$ )** la hipótesis en este caso es:

$H_0$ : No hay correlación entre las variables evaluación de la disciplina de autoestudio Vs. evaluación del grado de aprendizajes

$H_a$ : Si hay correlación entre las variables evaluación de la disciplina de autoestudio Vs. evaluación del grado de aprendizajes

- b. Prueba T de Student (W.S.Gosset):** Es una prueba para encontrar la diferencia entre las medias de dos grupos o dentro de un mismo grupo. Por ejemplo, se puede encontrar si hay diferencia entre las medias de grupos independientes, como puede ser el sexo y nivel de ingresos; también se puede ver si hay diferencia de antes y después para una misma muestra, como ver si hay diferencia entre la variable numérica después de un tratamiento. Y para ver si hay diferencias entre la media y un estándar oficial. **En el IBM-SPSS se procede: analizar, comparar medias, seleccionar para una muestra, para muestras independientes o para muestras relacionadas y aceptar.**

Prueba de muestras relacionadas

		Diferencias relacionadas				t	gl	Sig. (bilateral)	
		Media	Desviación tip.	Error tip. de la media	Intervalo de confianza para la diferencia				
					Inferior				Superior
Par 1	Evaluación de la disciplina de autoestudio - Evaluación del grado de aprendizaje	.22	1.16	.12	-2.05E-02	.46	1.819	89	.072

En la prueba de la T de Student con variables de muestras relacionadas se puede ver que no se rechaza la  $H_0$ , ya que la significancia es de .072, por lo que se puede concluir que NO SE RECHAZA  $H_0$  y se sugiere incrementar el tamaño de la muestra. **¡No!** Se debe decir que se acepta  $H_0$ , ya que esta conjetura se debe utilizar cuando se tenga un censo de la población. Las hipótesis ante esta prueba son:

- $H_0$ : La evaluación de la disciplina de autoestudio es muy similar a la evaluación del grado de aprendizaje.
- $H_a$ : La evaluación de la disciplina de autoestudio es diferente a la evaluación del grado de aprendizaje.

En el siguiente ejemplo, se puede investigar la relación entre las variables sexo y nivel de integración con los trabajos colaborativos.

## Estadísticos del grupo

Sexo de los encuestados		N	Media	Desviación típ.	Error típ. de la media
Nivel de integración con los trabajos colaborativos	Femenino	33	1.39	.50	8.64E-02
	Masculino	57	3.44	1.18	.16

## Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Intervalo de confianza para la diferencia	
									Inferior	Superior
Nivel de integración con los trabajos colaborativos	Se han asumido varianzas iguales	16.772	.000	-9.460	88	.000	-2.04	.22	-2.47	-1.62
	No se han asumido varianzas iguales			-11.446	82.027	.000	-2.04	.18	-2.40	-1.69

Las hipótesis para este caso son:

- Ho: No hay diferencia significativa entre el sexo y el nivel de integración con los trabajos colaborativos.
- Ha: Si hay diferencia significativa entre el sexo y el nivel de integración con los trabajos colaborativos.

En la tabla se observa que la muestra se compone de 33 mujeres y 57 hombres, con una media para el nivel de integración con los trabajos colaborativos de 1.39 para las mujeres y de 3.44 para los hombres. De entrada, se ve diferencia entre hombres y mujeres, siendo 1 el de mayor integración y 5 el de menor integración. En la prueba de la T de Student para muestras independientes se ve que la Significancia es de .000 (**Si la significancia es  $\leq 0.05$ , rechaza Ho**) como es menor a .05 se rechaza Ho y se acepta Ha, por lo que sí existe diferencia significativa entre el nivel de integración en trabajos colaborativos y el género, siendo la mujer la que se integra de manera rápida ante este nuevo modelo de enseñanza-aprendizaje.

La interpretación sería: Se presenta una tabla con los datos de medias, N, la desviación típica y error típico de la media, la correlación entre las dos variables,

una tabla que muestra la media, desviación típica, error de la media, los límites inferior y superior, el valor de t, los grados de libertad y la significancia. **(Si la significancia es  $\leq 0.05$ , rechaza  $H_0$ )**

*La prueba de Levene muestra si hay diferencia entre las varianzas de las dos muestras (para ver el supuesto de homocedasticidad o igualdad de varianza entre las dos medidas); valores mayores a 0.05 en la significancia indican que la varianza entre los grupos es igual y se procede a interpretar la T de Student.*

#### 4. BIVARIADA – PRUEBAS NO PARAMÉTRICAS

Se utilizan las pruebas no paramétricas cuando no se conoce la distribución normal de las variables para hacer inferencia estadística o cuando no se cumple el requisito de la normalidad de las variables involucradas en el análisis. Están altamente relacionadas y se debe llegar a los mismos resultados utilizando la prueba paramétrica o no paramétrica, como la t de Student que está altamente relacionada con la prueba U de Mann Whitney o de Wilcoxon y la ANOVA con la prueba de Kruskal-Wallis. Ver tabla No. 4

Tabla No. 4

Prueba paramétrica	Prueba no paramétrica
T de Student para muestras independientes	U de Mann Whitney
T de Student para muestras relacionadas	Wilcoxon
ANOVA	Kruskal-Wallis

- a. Prueba de la Chi Cuadrada:** útil para todas las variables medidas en escalas nominal u ordinal, cuando se hace cruces de variables se generan tablas con la agrupación de los datos, por ejemplo, la variable sexo (2 opciones: hombre o mujer) con nivel de escolaridad de primaria, secundaria, preparatoria, pregrado y posgrado (6 niveles de escolaridad), por lo que se estaría en el caso de dos variables nominales con una tabla de  $2 \times 3 = 6$ . Con 6 casillas que conservan valores entre la interacción de las dos variables. En IBM-SPSS, **analizar, estadísticos descriptivos, tablas cruzadas, seleccionar las variables en escala nominal para la columna y para la fila, en estadísticos seleccionar Chi-cuadrado y coeficiente de contingencia, en casillas seleccionar porcentajes en fila y columna, recuento observar, continuar y aceptar.** Las hipótesis están en función de las dos variables a analizar, por ejemplo, si una es sexo y el otro nivel consumo de cerveza con cuatro categorías, la hipótesis sería: “Ho: No hay diferencia significativa entre el sexo y el nivel de consumo de

cerveza por mes; Ha: Si hay diferencia significativa entre el sexo y el nivel de consumo de cerveza por mes”:

Tabla cruzada Sexo de los estudiantes del MBA.\*Cantidad de cerveza que bebe por rangos

			Cantidad de cerveza que bebe por rangos				Total
			1 a 3	4 a 6	7 a 10	Más de 11	
Sexo de los estudiantes del MBA	Femenino	Recuento	43	14	3	3	63
		% dentro de Sexo de los estudiantes del MBA	68.3%	22.2%	4.8%	4.8%	100.0%
	Masculino	Recuento	23	14	11	14	62
		% dentro de Sexo de los estudiantes del MBA	37.1%	22.6%	17.7%	22.6%	100.0%
Total		Recuento	66	28	14	17	125
		% dentro de Sexo de los estudiantes del MBA	52.8%	22.4%	11.2%	13.6%	100.0%

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	17.743 <sup>a</sup>	3	.000
Razón de verosimilitud	18.732	3	.000
Asociación lineal por lineal	16.958	1	.000
N de casos válidos	125		

a. 0 casillas (0.0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 6.94.

Con un valor de Chi-cuadrada de 17.743, 3 grados de libertad y significancia de .000, se rechaza la hipótesis nula y se acepta la hipótesis alterna, concluyendo que si existe diferencia significativa entre el sexo y el nivel de consumo de la cerveza, consumiendo más cerveza por mes los hombres versus las mujeres: el 90 % de las mujeres consumen menos de 6 cervezas al mes y el 63 % de los hombres consumen más de 6 cervezas al mes.

- b. **Prueba de Kolmogorov-Smirnov:** útil para variables en escala ordinal. **En IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, 2 muestras independientes, se pasa la variable de prueba y la de agrupación, aceptar.**

Para comprender esta prueba se sigue con el mismo ejemplo de la prueba de la Chi-cuadrada, con las mismas hipótesis: “Ho: No hay diferencia significativa entre el sexo y el nivel de consumo de cerveza por mes; Ha: Si hay diferencia significativa entre el sexo y el nivel de consumo de cerveza por mes”

		Cantidad de cerveza que bebe al mes.
Máximas diferencias extremas	Absoluta	.312
	Positivo	.312
	Negativo	.000
Z de Kolmogorov-Smirnov		1.742
Sig. asintótica (bilateral)		.005

a. Variable de agrupación: Sexo de los estudiantes del MBA.

Con un valor de Kolmogorov-Smirnov de 1.742 y significancia de .005, se rechaza la hipótesis nula y se acepta la hipótesis alterna; Ha: Si hay diferencia significativa entre el género y la cantidad de cerveza que beben al mes, siendo los hombres los que más cerveza consumen al mes.

- c. **Prueba U de Mann-Whitney:** Esta prueba compara la diferencia entre dos poblaciones de las variables medidas en escala Nominal. Cuando no se puede usar la T de Student para muestras independientes, por no tener un comportamiento normal, se utiliza la U de Mann-Whitney en la comparación de una variable numérica versus nominal; la Hipótesis, sería: No hay diferencia significativa entre la variable independiente nominal versus dependiente numérica. **Ejemplo en el IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, KS dos muestras independientes, se selecciona U de Mann-Whitney, se pasa la variable de agrupación de dos opciones y la variable dependiente o de prueba medida en escala ordinal, intervalar o de razón, aceptar.**

La hipótesis para esta prueba es:

- Ho: Los valores de la integración de los trabajos colaborativos entre sexo son iguales.
- Ha: Los valores de la interacción de los trabajos colaborativos entre sexo son diferentes.

#### Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo
Nivel de integración con los trabajos colaborativos	90	2.69	1.40	1	5
Sexo de los encuestados	90	1.63	.48	1	2

#### Rangos

		N	Rango promedio	Suma de rangos
Nivel de integración con los trabajos colaborativos	Femenino	33	21.97	725.00
	Masculino	57	59.12	3370.00
	Total	90		

### Estadísticos de contraste

	Nivel de integración con los trabajos colaborativos
U de Mann-Whitney	164.000
W de Wilcoxon	725.000
Z	-6.672
Sig. asintót. (bilateral)	.000

#### a. Variable de agrupación: Sexo de los encuestados

En este ejemplo se ve que el nivel de integración con los trabajos colaborativos y el sexo de los encuestados son diferentes. En la primera tabla se ve que las personas que contestaron el total de preguntas fueron 90, la media entre las dos poblaciones, esto es, entre hombres y mujeres para la variable nivel de integración con los trabajos colaborativos es de 2.69, en una escala que va del 1 al 5 siendo 1 el valor más importante y con una desviación típica de 1.43. Para la variable sexo, la media es 1.63, con un valor mínimo de 1 que significa femenino y de 2 que significa masculino; este valor de 1.63 indica que la mayoría de las 90 personas encuestadas son del sexo masculino, situación que se puede corroborar en la siguiente tabla en la que efectivamente el número de hombres es mayor que el de las mujeres en esta muestra (57 hombres y 33 mujeres).

En el cuadro de estadístico de contraste se ve que el valor de U de Mann-Whitney es de 164.000, con una significancia de .000 (**Si la significancia es  $\leq 0.05$ , rechaza  $H_0$** ).

- $H_0$ : No hay diferencia significativa entre el sexo y el nivel de integración con los trabajos colaborativos
- $H_a$ : Si hay diferencia significativa entre el sexo y el nivel de integración con los trabajos colaborativos

Se puede concluir: se rechaza  $H_0$  y se acepta  $H_a$ , por lo que, si existe diferencia significativa entre el sexo y el nivel de integración con los trabajos colaborativos, ya que las mujeres integran más rápidamente este esquema de trabajo colaborativo en el proceso de enseñanza-aprendizaje.

- d. **Prueba de Wilcoxon:** para encontrar diferencia entre las medias de dos variables numéricas relacionadas (medidas antes y después o para registrar que tanto ha variado la variable dependiente antes y después de un estímulo) que no cumplan con la característica de normalidad. Es similar a la T de Student para muestras relacionadas. **Ejemplo en el IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, dos muestras relacionadas, se selecciona variable 1 y 2, Prueba de Wilcoxon, en opciones se selecciona descriptivo, continuar y aceptar.**

Como ejemplo, se puede citar el proyecto de antes y después de regalar un libro sobre el gusto por la lectura. Primero se les aplicó la pregunta a 125 estudiantes de posgrado, con una escala de 1 a 5, siendo 5 el mejor gusto por la lectura, después se les regala dos libros a cada participante y un mes después se les volvió aplicar el mismo cuestionario con la pregunta gusto por la lectura; la hipótesis planteada es,  $H_0$ : No hay diferencia significativa entre el gustos por la lectura y el antes y después de regalar dos libros.

Prueba de Wilcoxon

	N	Media	Desviación estándar	Mínimo	Máximo
El estudiante de postgrado tiene gusto por la lectura	125	3.74	1.165	1	5
El estudiante de posgrado tienen gusto por la lectura despues de recibir como regalo dos libros.	125	4.94	.246	4	5

Estadísticos de prueba <sup>a</sup>

	El estudiante de posgrado tienen gusto por la lectura despues de recibir como regalo dos libros. - El estudiante de postgrado tiene gusto por la lectura
Z	-8.230
Sig. asintótica (bilateral)	.000

<sup>a</sup>. Prueba de rangos con signo de Wilcoxon

Se puede observar en la tabla de descriptivos que los 125 estudiantes de posgrado evaluaron con una media de 3.74 (evaluación de 1 a 5, siendo máximo el mayor valor) el gusto por la lectura y un mes después de haberles regalado un libro, este gusto se incrementó en 4.94; con Z de 8.230 y significancia de .000, se rechaza la hipótesis nula y se acepta la hipótesis alterna, por lo que el efecto del regalo influye de manera positiva en el gusto por la lectura.

- e. **Prueba de Kruskal-Wallis:** es útil cuando la variable categórica o de agrupación es de más de dos opciones, como estado civil, ingresos por categorías, entre otras y se usa en lugar de la ANOVA (por no cumplir la normalidad de la variable dependiente). **Ejemplo en el IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, K muestras independientes, en variables de agrupación se pasa la variable nominal con más de dos categorías y se definen las categorías, se pasan las variables numéricas a contrastar, aceptar.**

Como ejemplo, se cruzarán dos variables que miden el número de cervezas que se consumen al mes en 4 rangos y el nivel de ingreso en salarios mínimos al mes de los estudiantes de posgrado. La hipótesis nula se plantea en forma de negación; Ho: No hay diferencia significativa entre el número de cervezas que se consumen al mes versus el nivel de ingreso de los estudiantes del posgrado.

## Estadísticos descriptivos

	N	Media	Desviación estándar	Mínimo	Máximo
Ingresos de los estudiantes en SMLV	125	9.26	5.321	3	32
Cantidad de cerveza que bebe al mes.	125	1.86	1.083	1	4

## Prueba de Kruskal-Wallis

	Cantidad de cerveza que bebe al mes.	N	Rango promedio
Ingresos de los estudiantes en SMLV	1 a 3	66	73.27
	4 a 6	28	65.13
	7 a 10	14	41.64
	Más de 11	17	37.24
	Total	125	

Estadísticos de prueba<sup>a,b</sup>

	Ingresos de los estudiantes en SMLV
Chi-cuadrado	19.135
gl	3
Sig. asintótica	.000

a. Prueba de Kruskal Wallis

b. Variable de agrupación:  
Cantidad de cerveza que bebe al mes.

En la tabla de estadísticos descriptivos se observa que en la variable Ingresos de los estudiantes en SMLV, contestaron 125 personas con una media de 9.26 salarios mínimos al mes (SMLV), desviación estándar de 5.321, con un valor mínimo de 3 y máximo de 32 salarios mínimos al mes (SMLV), al igual que en la variable cantidad de cerveza que bebe al mes, se puede ver que la media es de 1.86 cervezas al mes, con media de 1.083, mínimo de 1 y máximo de 4. En la tabla de Kruskal-Wallis, se puede ver los 4 niveles de cantidad de cerveza que beben al mes (1 a 3 cervezas, 4 a 6 cervezas, de 7 a 10 cervezas y más de 11 cervezas). En el estadístico de prueba con un valor de Chi-cuadrada de 19.135, 3 grados de libertad y una significancia de .000, se rechaza la hipótesis nula y se

acepta la Ha: Si hay diferencia significativa entre la cantidad de cerveza que se bebe al mes y el nivel de ingresos de los estudiantes del posgrado.

f. **Prueba de la Mediana:** Compara los resultados de la mediana de cada uno de los grupos. Siguiendo con el ejercicio anterior se procede aplicar la prueba de la mediana, para ello se necesita formular las hipótesis:

- Ho: No hay diferencia significativa entre la cantidad de cerveza que se bebe al mes y el nivel de ingresos de los estudiantes del posgrado.
- Ha: Si hay diferencia significativa entre la cantidad de cerveza que se bebe al mes y el nivel de ingresos de los estudiantes del posgrado.

Prueba de la mediana: Frecuencias

		Cantidad de cerveza que bebe al mes.			
		1 a 3	4 a 6	7 a 10	Más de 11
Ingresos de los estudiantes en SMLV	> Mediana	39	11	4	2
	<= Mediana	27	17	10	15

Estadísticos de prueba<sup>a</sup>

	Ingresos de los estudiantes en SMLV
N	125
Mediana	8.00
Chi-cuadrado	14.788
gl	3
Sig. asintótica	.002

a. Variable de agrupación:  
Cantidad de cerveza que bebe al mes.

En la tabla de prueba de la mediana: frecuencias, se pueden ver los datos que se encuentran por arriba y por debajo de la mediana de ingresos. En el rango de 1 a 3 de cantidad de cerveza que se bebe al mes se registran 39 personas con ingresos superiores a la media, en el rango de 4 a 6, se registran 11 personas, en el rango de 7 a 10 se registran 4 personas y en mayores de 11 cervezas al mes solo se registraron 2 personas con ingresos por arriba de la mediana. En estadísticos de

prueba se puede ver N igual a 125 personas que contestaron el cuestionario, con una mediana de 8, Chi-cuadrado de 14.788, 3 grados de libertad y una significancia de .002, por lo que se rechaza la hipótesis nula y se acepta la hipótesis alterna, confirmándose que si hay diferencia significativa entre la cantidad de cerveza que se bebe al mes y los ingresos de los estudiantes de posgrado, siendo el mayor consumo de cerveza en los estudiantes que ganan menos de 8 salarios mínimos al mes. **(Si la significancia es  $\leq 0.05$ , rechaza  $H_0$ )** Como es menor a .05 se rechaza  $H_0$  y se acepta  $H_a$ .

**En el IBM-SPSS: analizar, pruebas no paramétricas, cuadro de diálogos antiguos, K muestras independientes, en variables de agrupación se pasa la variable nominal con más de dos categorías y se definen las categorías, se pasan las variables numéricas a contrastar, Seleccionar mediana y aceptar.**

## Conclusión

Las estadísticas descriptivas e inferenciales de todo proyecto de investigación están en función de los objetivos del proyecto, las hipótesis planteadas y el tipo de variable que se utilizaron en el instrumento de medición o cuestionario. En este capítulo se proporciona un esquema sencillos para comprender la interacción de las variables de medición y el tipo de técnica estadística a implementar, por ejemplo, cuando las dos variables son nominales, la prueba estadística recomendada es la Ji Cuadra o Chi Cuadra, para contrastar la hipótesis de interacción entre las dos variables e inferir los resultados de la muestra hacia la población. Cuando las dos variables son numéricas las estadísticas para describir recomendables son la correlación de Pearson y para inferir se sugiere la prueba Z o la T de Student, la correlación, análisis de regresión simple, ANOVA, MANOVA, ANCOVA y MACOVA. Cuando las variables de medición están en escala ordinal, las pruebas estadísticas para describir es el coeficiente de correlación de Spearman y la correlación Gamma Tao y para inferir se sugiere la prueba de Wilcoxon, U de Mann Whitney, prueba del Signo, Kolmogorov-Smirnov, la prueba de la Mediana y la prueba de Kruskal-Wallis. Las pruebas estadísticas para inferir se dividen en paramétricas y no paramétricas, siendo las paramétricas propias de las variables de escala superior (intervalar y razón) con el requisito de conocer la distribución normal de la variable en la muestra y en la población, tamaño de la muestra grande y representativo de la población para inferir, mientras que las pruebas estadísticas no paramétricas se utilizan en muestras más pequeñas o cuando no se tiene los requisitos de la normalidad de variables involucradas en el análisis. Son más precisas estadísticamente hablando para inferir las pruebas paramétricas que las no paramétricas.

**Palabras clave**

Pruebas paramétricas y no paramétricas

Estadística descriptiva

Estadística inferencial

Estadística univariada

Estadística bivariada

**Preguntas**

¿Por qué es recomendable utilizar pruebas estadísticas paramétricas en los proyectos de investigación?

¿Cuándo es recomendable utilizar pruebas estadísticas no paramétricas en el proyecto de investigación?

¿Cómo se relacionan las escalas de medición con las pruebas estadísticas univariadas?

¿Cómo se relacionan las escalas de medición con las pruebas estadísticas bivariadas?

¿Qué relación hay entre las pruebas paramétricas T de Student y las no paramétricas U de Mann Whitney y Wilcoxon?

¿Qué relación hay entre las pruebas paramétricas de ANOVA y la no paramétrica Kruskal-Wallis?

## VI. ESTADÍSTICA MULTIVARIADA

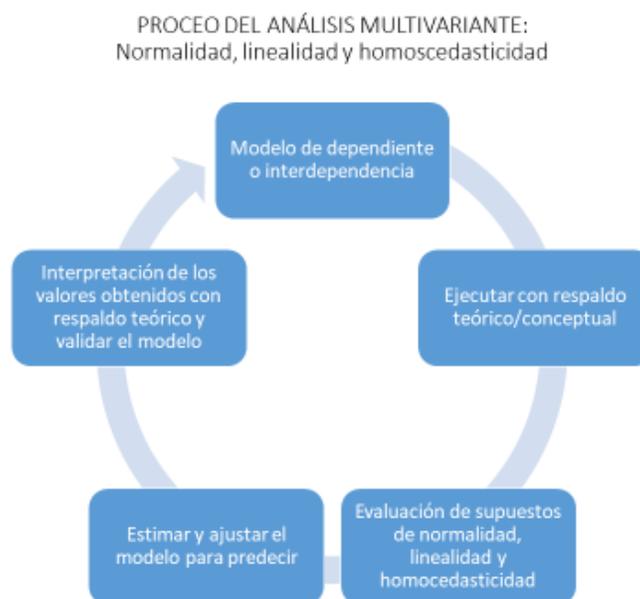
***La estadística multivariable permite el análisis de varias variables en un mismo momento con modelos causales o de interdependencia para predecir el comportamiento de las mismas en una población dada.***

Técnica estadística útil para analizar varias variables a la vez que generan modelos de interacción de variables, teniendo como requisito para cada variable la normalidad, linealidad y homocedasticidad. Estos modelos pueden ser de dependencia o de interdependencia, por ejemplo:

- a. Modelo de interdependencia: cuando se tienen muchas variables que aparentemente miden lo mismo, se recomienda el método multivariable de reducción de variables en factores, como el análisis de componentes principales y de factores (Variables superiores tipo Likert, intervalar o razón), análisis de correspondencia (Variables cualitativas de tipo nominal), etc.
- b. Modelo de características comunes: útil para agrupa las características comunes que lleven a clasificar o segmentar una población dada, como el análisis de clúster, arboles de decisiones y análisis discriminante.
- c. Modelos de dependencia y de causa efecto entre variables cuantitativas o superiores: relación entre variables dependientes e independientes para predecir comportamientos de la población objeto de estudio como la regresión lineal simple, múltiple y regresión logística.
- d. Modelo de relación de la variable dependiente cuantitativa o superior en función de ciertos atributos medidos en escala nominal, como el volumen de ventas (variable dependiente) en función de la posición del producto dentro una tienda o posición del producto en función de los anaqueles de exhibición (variable independiente); para estas situaciones se recomienda el análisis de ANOVA, MANOVA, ANCOVA Y MACOVA
- e. Modelo de agrupación y jerarquización de atributos: útil para diseñar ofertas de valor desde la perspectiva de los clientes con la técnica de análisis de conjuntos.

El proceso del análisis multivariante se inicia con definir el modelo de dependencia o interdependencia de variables con los análisis de normalidad, para ver si cada una de las variables involucradas en el modelo siguen el comportamiento de la distribución normal, la linealidad para ver qué tan relacionadas están las variables entre sí y homocedasticidad de las variables, que mide la homogeneidad de la varianza para poder predecir el comportamiento con mayor nivel de confianza. Se recomienda tener un respaldo teórico y/o conceptual del modelo a utilizar para ajustar el modelo e interpretarlo y validarlo. Se procede a evaluar los supuestos de normalidad, linealidad y homocedasticidad, antes de estimar y ajustar el modelo para predecir y por último se procede a la interpretación de los valores obtenidos con respaldo teórico para su respectiva validación. Ver gráfica número 24.

Gráfica No. 24



Fuente: Aportación del autor

Las técnicas estadísticas multivariantes se pueden clasificar en:

- a. Análisis de reducción de variables o casos, como el análisis de componentes principales para todas las preguntas del cuestionario que sean medidas en escalas superiores como intervalar o razón. También entran en esta

clasificación el análisis de correspondencia con variables nominales y análisis de factores con variables superiores para agrupar y reducir variables que generan nueva información y/o conocimiento.

- b. Análisis de agrupación o reducción, que permite generar grupos homogéneos de individuos, muy útil para segmentar a la población con perfiles propios en cada segmento para diseñar estrategias personalizadas a cada segmento. Una de estas técnicas es el Clúster análisis, que permite agrupar individuos en segmentos similares con variables medidas en escala nominal, ordinal e intervalar; los árboles de decisiones que permite visualizar las características propias de cada agrupación de individuos y el análisis discriminante, que requiere de una variable dependiente medida en escala nominal y las independientes medidas en escalas superiores intervalar o razón, que arroja una ecuación predictora para identificar en cual grupo podría pertenecer un futuro cliente, ejemplo, aplicando la ecuación predictora se podría ver la probabilidad de que una persona caiga al grupo de los enfermos, al grupo de los morosos en el pago de sus créditos, etc.; y, el análisis de conjunto que ayuda a identificar las características de valor medidas en escala inferiores o superiores, para el desarrollo del valor de un nuevo producto o servicio.
- c. Análisis de relación causa/efecto y explicativo, para identificar las variables independientes medidas en escalas cuantitativas o superiores (intervalar y/o razón) que más influyen en el efecto de la variable dependiente. Las técnicas estadísticas de mayor uso en este apartado son el ANOVA, MANOVA, ANCOVA, Y MANCOVA, análisis de ANOVA factorial, las cuales requieren que la variable dependiente estén medida en escala cuantitativa o superior (intervalar y/o Razón) y las independientes en escalas cualitativas o inferiores (nominal y/o ordinal). Otra de las técnicas estadísticas de mayor uso en este apartado son la correlación, regresión lineal simple, múltiple, discriminante, logística, modelo probit y Logit. Ver gráfica número 25 y 26.

Gráfica No. 25

# ANÁLISIS MULTIVARIANTE

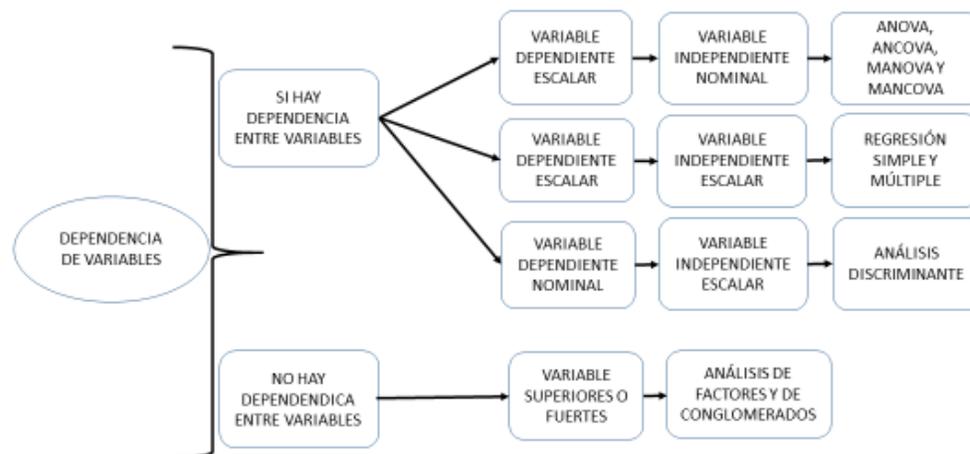
Normalidad, linealidad y homocedasticidad

Análisis de reducción	Análisis de agrupación o clasificación	Análisis de relación causa/efecto y explicativo
<ul style="list-style-type: none"> <li>• Análisis de componentes principales (variables superiores, Likert o escalares)</li> <li>• Análisis de correspondencia (variables nominales)</li> <li>• Análisis de factores para agrupar y reducir variables (variables superiores, Likert o escalares)</li> </ul>	<ul style="list-style-type: none"> <li>• Clúster análisis para agrupar individuos en segmentos similares (variables nominales, ordinales e intervalares)</li> <li>• Árboles de decisiones</li> <li>• Análisis discriminante (Variable dependiente nominal y las independientes de escala superior)</li> <li>• Análisis de conjunto identifica los atributos de valor para el desarrollo de un nuevo producto, etc.</li> </ul>	<ul style="list-style-type: none"> <li>• Análisis de ANOVA, MANOVA ANCOVA Y MANCOVA. (variable dependiente superiores y las independientes nominales)</li> <li>• Regresión lineal simple y múltiple (variables dependiente e independientes medidas en escalas superiores)</li> <li>• Regresión logística (variable dependiente nominal)</li> </ul>

Fuente: Aportación del autor.

Gráfica No. 26

## ANÁLISIS ESTADÍSTICO MULTIVARIADO



## I. ANÁLISIS DE LA VARIANZA

Para el análisis de varianza es importante definir las escalas de las variables dependiente e independiente. La variable de pendiente debe ser cuantitativa o superior (intervalar o de razón) y las independientes deben ser cualitativas o inferiores (Nominal, ordinal y escalar) Ejemplo, se quiere conocer el volumen de ventas de rasuradoras femeninas (Variable dependiente medida en escala de razón) versus los tipos de promociones (Dos por una, regalo de la miel para la piel y regalo de crema para después de la afeitada femenina) y ubicación de los puntos de venta en tiendas de comercio al detalle (Cerca de las cajas registradoras, en blancos y en perfumería)

**Se sugiere primero medir los supuestos de normalidad de la variable dependiente para cada uno de los grupos (en IBM-SPSS, Analizar, estadísticos descriptivos, explorar, pasar la variable de pendiente y las independientes en factores, gráficos con pruebas de normalidad: Shapiro Wilk, para muestras menores de 50 y Kolmogorov-Smirnov para muestras grandes; si la significancia es mayor de .05, se cumple con la normalidad), la independencia de grupos, equivalencia de grupos o tamaños similares en la muestra de los grupos, y la homogeneidad de varianza.**

Primer ejemplo para el IBM-SPSS, se cruzará la variable dependiente ingresos y la independiente niveles de consumo de cerveza al mes, medida en 4 niveles: 1 a 3 cervezas al mes, de 4 a 6, de 7 a 10y más de 11 cervezas al mes. La variable dependiente ingresos se miden en número de salarios mínimos al mes que percibe el encuestado.

***En el IBM-SPSS se va a analizar, comparar medias, ANOVA de un factor, seleccionar la variable dependiente métrica (nivel de ingresos en SMLV) y la variable independiente nominal (Bebe cerveza), en post hoc seleccionar la prueba de Tukey para homogeneidad de varianza, si no hay homogeneidad de varianza se selecciona games-Howel, en opciones seleccionar la prueba de homogeneidad de las varianzas y descriptivos.***

Las hipótesis a contrastar son:

Ho: No hay diferencia significativa entre el número de cervezas que se bebe al mes y el nivel de ingreso de los estudiantes del posgrado.

Ha: Si hay diferencia significativa entre el número de cervezas que se bebe al mes y el nivel de ingreso de los estudiantes del posgrado.

ANOVA: Descriptivos

Ingresos de los estudiantes en SMLV

	N	Media	Desviación estándar	Error estándar	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
1 a 3	66	10.42	5.441	.670	9.09	11.76	4	32
4 a 6	28	9.96	6.191	1.170	7.56	12.37	4	30
7 a 10	14	6.40	2.711	.724	4.83	7.97	3	12
Más de 11	17	5.94	1.784	.433	5.02	6.86	4	9
Total	125	9.26	5.321	.476	8.32	10.20	3	32

Se puede observar en el ANOVA descriptivo, que en el consumo de 1 a 3 hay una media de ingresos en salarios mínimos al mes (SMLV) de 10.42, con una desviación estándar de 5.441, los que beben de 4 a 6 cervezas al mes tienen un promedio de ingresos de 9.96, los que beben de 7 a 10 cervezas, con un ingreso de 6.40 y los que beben más de 11 cervezas al mes, cuentan con un ingreso mensual de 5.94, quedando evidente que consume más cerveza el que tiene menos ingreso mensuales.

ANOVA: Inferencial

Ingresos de los estudiantes en SMLV

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	405.111	3	135.037	5.261	.002
Dentro de grupos	3105.547	121	25.666		
Total	3510.658	124			

En el análisis de ANOVA para inferir, se puede ver que el valor de F es 5.261, con una significancia de .002, (Cuando la significancia es menor o igual a .05 se rechaza la hipótesis nula) por lo que se rechaza la hipótesis nula y se acepta la hipótesis alterna. Si hay diferencia significativa entre el beber cerveza y los ingresos de los estudiantes de posgrado; consumen más cerveza los que ganan menos.

## Ingresos de los estudiantes en SMLV

HSD Tukey

Cantidad de cerveza que bebe al mes.	N	Subconjunto para alfa = .95	
		1	2
Más de 11	17	5.94	
7 a 10	14	6.40	
4 a 6	28		9.96
1 a 3	66		10.42
Sig.		.990	.990

Se visualizan las medias para los grupos en los subconjuntos homogéneos.

- a. Utiliza el tamaño de la muestra de la media armónica = 22.085.
- b. Los tamaños de grupo no son iguales. Se utiliza la media armónica de los tamaños de grupo. Los niveles de error de tipo I no están garantizados.

Con la prueba de Tukey, se puede apreciar la diferencia entre los grupos, quedando dos grandes grupos, el primero conformado por los que beben mucho con niveles de ingresos de 5.94 y 6.40 salarios mínimos al mes. Y el grupo dos conformado por personas que beben poco y tienen salarios superiores que rondan entre 9.96 y 10.42, salarios mínimos al mes. La prueba Tukey en ANOVA, es muy útil para encontrar esos grupos que marcan la diferencia cuando se plantean las diferencias entre los grupos.

El segundo ejemplo para el IBM-SPSS, se cruzan las variables dependiente nivel de integración con los trabajos colaborativos, medida en escala intervalar vs la independiente tipo de maestría medida en escala nominal.

***En el IBM-SPSS se va a analizar, comparar medias, ANOVA de un factor, seleccionar la variable dependiente métrica y la variable independiente nominal, en post hoc seleccionar la prueba de Tukey para homogeneidad de varianza, si no hay homogeneidad de varianza se selecciona games-Howel, en opciones seleccionar la prueba de homogeneidad de las varianzas y descriptivos.***

### Descriptivos

Nivel de integración con los trabajos colaborativos

	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%			Mínimo	Máximo
					Límite inferior	Límite superior			
Presencial	30	3.27	1.55	.28	2.69	3.85	1	5	
Virtual	30	2.93	.78	.14	2.64	3.23	2	4	
On Lines	30	1.87	1.36	.25	1.36	2.37	1	5	
Total	90	2.69	1.40	.15	2.40	2.98	1	5	

### ANOVA

Nivel de integración con los trabajos colaborativos

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	32.089	2	16.044	9.886	.000
Intra-grupos	141.200	87	1.623		
Total	173.289	89			

Se puede observar que el nivel de integración con los trabajos colaborativos es muy diferente en cada uno de los tipos de maestría; en presencial la media es de 3.27, en virtual es de 2.93 y en online es de 1.87, siendo 1 el de mayor integración y 5 el de menor integración con los trabajos colaborativos.

En este Análisis de varianza se observa un valor de F de 9.886 con una significancia de .000, como es menor a .05 se rechaza  $H_0$  y se acepta  $H_a$ , en la que se puede ver una franca diferencia entre la media del nivel de integración con los trabajos colaborativos y el tipo de maestría.

- $H_0$ : El nivel de integración con los trabajos colaborativos es el mismo en los tres tipos de maestría
- $H_a$ : El nivel de integración con los trabajos colaborativos es diferente entre tipos de maestría

La prueba de Duncan muestra claramente en donde está la diferencia y que grupo está marcando verdaderamente la diferencia. Por lo que se puede concluir: se rechaza la  $H_0$  y se acepta la  $H_a$ , ya que la maestría online, es la que presenta

mayor integración con los trabajos colaborativos, seguida del esquema virtual, ocupando el último lugar la presencial. Todas estas afirmaciones con un 95 % de confianza.

Omega cuadrada ( $w^2$ ), indica la proporción de la varianza en la variable dependiente explicada por una variable independiente o factor particular. Índice de Omega cuadrado puede ir de .01 a .15. Alto significa mayor proporción de la variación en la variable dependiente.

$N^2$  indica la proporción de la variación de la variable dependiente por interacción con la independiente

### Nivel de integración con los trabajos colaborativos

Duncan<sup>a</sup>

Tipo de Maestría que está estudiando	N	Subconjunto para alfa = .05	
		1	2
On Lines	30	1.87	
Virtual	30		2.93
Presencial	30		3.27
Sig.		1.000	.314

Se muestran las medias para los grupos en los subconjuntos homogéneos.

- a. Usa tamaño de la muestra de la media armónica = 30.000.

NOTA: Se procede de la misma manera para cuando se tiene dos o más variables independientes.

- **ANÁLISIS DE ANOVA FACTORIAL:** cuando se comparan dos o más variables independientes contra una dependiente.

***En IBM-SPSS, modelo lineal general, univariante, se pasa la variable dependiente, y las variables independientes en el recuadro de factores fijos, en post hoc seleccionar Tukey para la variable con más de dos grupos y pasarla al recuadro de análisis, en opciones seleccionar los descriptivos y pruebas de la homogeneidad de varianzas, en opción de gráficos seleccionar en líneas separadas la variable independiente con menos grupos y en eje horizontal la variable independiente con más de dos grupos-añadir, continuar y aceptar.***

En las primeras tablas del análisis se observa los grupos y el tamaño de la muestra, para ver la similitud de muestra en cada grupo, en la prueba de homogeneidad de Levene se observa si la significancia es mayor a .05, para concluir igualdad de varianzas entre los grupos.

En la tabla de interacción de las variables independientes con la dependiente, se observa la significancia para encontrar en que variable independiente se observa una diferencia de medias de la variable dependiente y se procede a rechazar o no la hipótesis nula y se aceptaría la hipótesis de investigación. Cuando hay efecto en las interacciones se procede a ver las pruebas post hoc, para identificar las diferencias entre las variables y los grupos, observando la significancia; si la significancia es menor o igual a .05, se procede a estudiar las diferencias de medias, las cuales se pueden observar en la tabla de subconjuntos homogéneos. En la interacción entre factores se puede observar no interacción cuando los dos factores son paralelos, positiva cuando la interacción de los dos factores se abre, se da interacción negativa cuando los dos factores se juntan en la parte inferior de la gráfica de factores y son de relación inversa cuando se cruzan los factores.

- ANÁLISIS MANOVA O ANÁLISIS DE VARIANZA MULTIVARIABLE:

Es una extensión del ANOVA, donde se integra más de una variable dependiente debidamente correlacionada entre sí, para ver si existe diferencia entre las medias de los grupos de la variable nominal independiente y el grupo de las variables dependientes. Se formula la Hipótesis Ho: No hay diferencia entre las dos variables dependientes y las independientes que integran el modelo de MANOVA; Ha: Si hay diferencias entre las dos variables dependientes que integran el modelo d MANOVA. **En el IBM-SPSS, se procede primero a ver la prueba de la normalidad: analizar, estadísticos descriptivos, explorar, se pasan las dos variables dependientes y las independientes, en gráfico, se pide prueba de normalidad, aceptar.**

**Para correr el MANOVA se procede: analizar, modelo lineal general, multivariante, seleccionar las variables dependientes e independientes en los cuadros correspondientes; en gráficos seleccionar las dos variables independientes, en post hoc, seleccionar las variables independientes y la prueba Tukey, continuar, en opciones seleccionar prueba de homogeneidad, estimación del tamaño del efecto, en contraste seleccionar la diferencia entre los grupos con contraste simple, categoría de referencia que sea la que dice primero, cambiar y continuar; en gráficos se puede seleccionar grupos y la variable independiente, y aceptar. Se recomienda utilizar la traza de pillar cuando no se cumple con la distribución normal y la raíz mayor de Roy, cuando no se cumpla con el criterio de grupos iguales, pero que si cumpla con la distribución normal.**

Interpretación: con la prueba de Shapiro Wilk, se procede analizar la confirmación de la distribución normal, de las variables dependientes en cada uno de los grupos. Con el test de Box's de igualdad de varianza, se observa la Significancia de igualdad de varianza (Si la significancia es mayor a .05, si existe igualdad de varianza).

En la tabla de contraste multivariante, se puede observar la significancia de cada una de las variables independientes y la interacción de las mismas con el método Trace de Pillai's; si la significancia es menor a .05, se concluye que existe diferencia entre las medias de la dependiente entre los diversos grupos de las variables

independientes analizadas en el modelo. En la tabla de intersujetos, se puede medir la diferencia en las variables dependientes, pero ahora de manera separada con cada uno de los grupos de la variable independiente, para observar en donde hay diferencia significativa de las variables dependientes por separado.

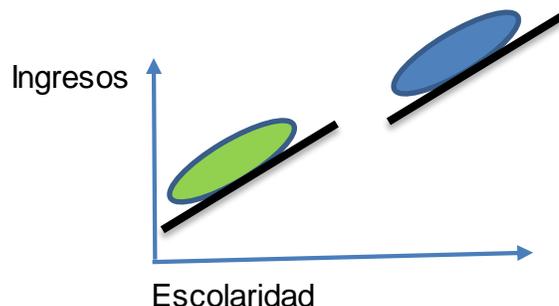
Análisis de contrastes K Matrix, recordando que el level 1 es el grupo que quedó como grupo de referencia.

En gráficos de promedios se puede ver en qué grupo de la variable independiente se encuentra la diferencia de las medias de la variable dependiente.

- ANÁLISIS DE COVARIANZA O ANCOVA:

Útil para identificar el efecto de una variable métrica extraña sobre la variable dependiente. Permite corregir el resultado de la dependiente eliminando la influencia de las variables extrañas o covariables; análisis de colinealidad entre la variable dependiente métrica y la covariable también métrica para quitarle el efecto sobre la variable de agrupación. ¿Cuántas covariables incorporar en el modelo? Están en función de la siguiente consideración  $((.1 \times \text{tamaño de la muestra}) - \text{número de grupos de la variable independiente} - 1)$  por ejemplo: para una muestra de 50 y una variable independiente con 3 grupos serían menos de 3 covariables. Ejemplo de covarianza con las variables sexo y estado civil, como variables independientes, ingresos como variable dependiente y la escolaridad como covariable (¿Qué tanto la escolaridad influye en los ingresos?)  $H_0$ : No hay diferencia significativa entre el sexo y los ingresos (quitando el efecto del grado de escolaridad). Ver gráfica número 27.

Gráfica No.27



Fuente: Aportación del autor

El color azul simboliza a los hombres y el verde a las mujeres, las rectas simbolizan la regresión de cada grupo y permite predecir el nivel de ingreso por sexo eliminando los efectos de la escolaridad. El color azul representa hombres con alto ingresos y con alto grado de escolaridad. Se necesita ver si hay diferencia significativa entre el ingreso y el sexo, quitando el efecto de los años de estudio en escolaridad, por lo que se igualan los grupos con los años de estudio para ver el efecto de los ingresos con el sexo. Se puede ver homogeneidad de regresión entre grupos, esto quiere decir que las rectas son paralelas y que no se cruzan entre sí.

**En el IBM-SPSS se va a gráficos, cuadro de diálogo antiguos, dispersión y puntos, simple, en el cuadro de diálogo seleccionar la variable en el eje Y la variable dependiente (numérica), en el X la covariable (numérica) y estableciendo marcas por la variable independiente (nominal)** Se le pide la recta para visualizar las pendientes de las rectas y ver la homogeneidad de las rectas de regresión. Se procede a ver si hay problemas de multicolinealidad realizando la correlación entre la variable dependiente y la covariable; si la correlación es mayor a .70 se considera que existe multicolinealidad entre las variables.

Primero se realiza el análisis de varianza univariado entre la variable dependiente ingresos y la independiente sexo. Después se realiza la covarianza tomando como variable dependiente ingresos, la covariable nivel de estudios y la de agrupación sexo. Se debe ver el grado de correlación entre la variable dependiente y la covariable, para identificar si hay colinealidad entre ellas (valores de correlación igual o menores a .70 se toman como no colinealidad). Se podrá ver si existe diferencia entre el sexo y los ingresos con el nivel de significancia igual o menor a .05, para concluir quien gana más sin el efecto de escolaridad.

**Pasos en el SPSS:**

- **Analizar, Modelo lineal general, Univariante, inserta la variable dependiente, factores fijos (sexo). Se concluye si hay o no diferencia entre el sexo y el nivel de ingresos.**
- Se repite el paso anterior y en la parte de covariable, se busca primero si hay correlación entre la variable dependiente y covariable. **(Gráfico, cuadro de diálogos antiguos, diagrama de dispersión, simple, se inserta la variable y como dependiente, el eje x como covariable y establecer marcas por sexo)** se puede ver el gráfico de dispersión entre la variable dependiente y covariable, recordar que la correlación menor de .70 es considerada como no colinealidad. Allí sobre la gráfica le das doble clic sobre la gráfica para visualizar las rectas de regresión de las dos variables y se observa el porcentaje de varianza de cada variable y la pendiente de cada variable.

- Se realiza la correlación entre las variables (**en el IBM-SPSS, analizar, correlación, bivariada: pasa la variable dependiente y la covariable, aceptar.**)
- Análisis de covarianza (**En el IBM-SPSS, Analizar, Modelo lineal general, Univariante, inserta la variable dependiente, la covarianza, sexo como factores fijos y en opciones mostrar las medias para sexo, estadística descriptiva, estimación del tamaño del efecto y prueba de homogeneidad, aceptar.**) Se observa según el nivel de significancia si la variable dependiente y la covarianza están asociadas. Se observa si la variable sexo guarda relación significativa con el nivel de ingresos eliminando el efecto de escolaridad.
- Eta al cuadrado parcial muestra el grado del efecto de la varianza sobre sexo y se observa la varianza total explicada del sexo (R cuadrada corregida) y el efecto de la variable escolaridad. Por último, se analiza las medias entre sexo y el ingreso para ver quien marca la diferencia hacia la población de donde se extrajo la muestra eliminando el efecto del nivel de escolaridad.
- Ejemplos para aplicar la covarianza: estimar efecto de la covariable edad sobre los datos del colesterol para tres dietas alimenticias (variable independiente o de tratamiento). Para estudiar el efecto de la covariable ingresos, edad y escolaridad sobre el aumento del azúcar en sangre (variable dependiente) en función de tres dietas alimenticias (variable independiente)

- ANÁLISIS DE VARIANZA NO MÉTRICA:

Útil para identificar la diferencia en las medias de más de dos grupos, cuando la variable dependiente se mide en escala ordinal. El procedimiento es conocido como la prueba de la mediana con k muestras, partiendo del principio de igualdad de mediana entre los grupos; se genera una tabla 2x4 para calcular la estadística de Ji cuadrada. Una prueba más útil es el análisis de varianza unidireccional de Kruskal-Wallis (es una extensión de la prueba de Mann-Whitney), el cual, usa el valor de clasificación de cada caso y no solo su posición en relación a la mediana como se hace en el de K muestras.

## II. ANÁLISIS DE CORRELACIÓN Y REGRESIÓN

### • CORRELACIÓN:

La correlación de Pearson, se ejecuta entre dos variables medidas en escalas superiores o escalares, mide la dirección de la correlación y el grado de la misma. Los gráficos de dispersión de datos entre dos variables medidas en escalas superiores, permite visualizar la dirección de la correlación, la cual puede ser positiva si la relación entre las variables es directamente proporcional, esto es, si aumenta una la otra aumentará directamente proporcional o si disminuye una la otra disminuye directamente proporcional; cuando la dirección es negativa se muestra la relación inversamente proporcional, esto es, si una variable aumenta la otra disminuirá directamente proporcional; el grado de la correlación va de 0 a +1 y de 0 a -1, siendo 0 ausencia de correlación, .5 correlación moderada, menor a .5 correlación débil y arriba de .5 correlación fuerte, tanto positiva como negativa. Cuando no hay correlación entre las variables se puede observar en los diagramas de dispersión una forma curvilínea como puede ser en forma de U. Ver el gráfico número 28, 29 y 30

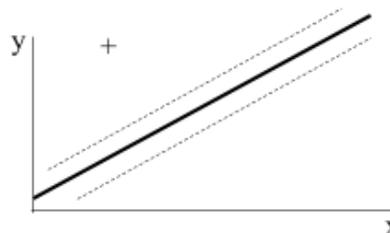
Gráfica No. 28

### CORRELACIÓN POSITIVA

#### GRADO DE ASOCIACIÓN ENTRE DOS VARIABLES QUE MIDE:

1. LA DIRECCIÓN: PUEDE SER POSITIVA O NEGATIVA
2. LA FUERZA MEDIDA DE 0 a + 1 y de 0 a - 1:
 

1. 0	NO HAY CORRELACIÓN
2. + .01 A .49	CORRELACIÓN POSITIVA DÉBIL
3. + .5	CORRELACIÓN POSITIVA MODERADA
4. + .51 A .95	CORRELACIÓN POSITIVA FUERTE
5. + 1	CORRELACIÓN POSITIVA PERFECTA



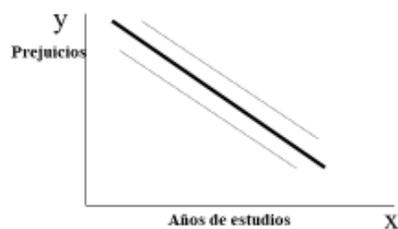
Fuente: Adaptación de Levin J

Gráfica No. 29

## CORRELACIÓN NEGATIVA

### PREJUICIO EN FUNCIÓN DE LA EDUCACIÓN:

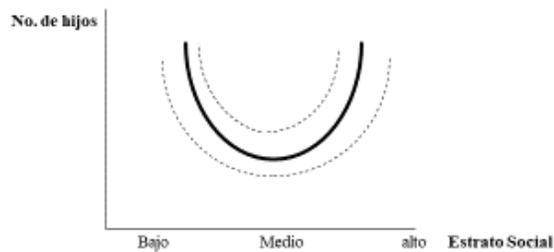
- |               |                               |
|---------------|-------------------------------|
| 1. 0          | NO HAY CORRELACIÓN            |
| 2. -.01 A .49 | CORRELACIÓN NEGATIVA DÉBIL    |
| 3. -.5        | CORRELACIÓN NEGATIVA MODERADA |
| 4. -.51 A .95 | CORRELACIÓN NEGATIVA FUERTE   |
| 5. -1         | CORRELACIÓN NEGATIVA PERFECTA |



Fuente: Adaptación de Levin J.

Gráfica No. 30

## NO CORRELACIÓN

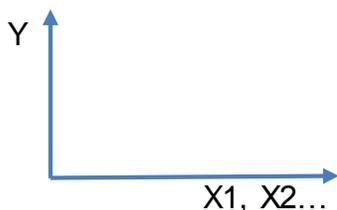


Fuente: Adaptación de Levin J.

- **REGRESIÓN:**

Método estadístico que permite predecir el comportamiento de la variable dependiente en función de una o varias variables independientes, gracias a la representación matemática:  $Y = B_0(\text{constante}) + b_1X_1$  ( $b_1$  es el coeficiente de la variable  $X_1$  o pendiente de la recta y  $X_1$  es el valor que obtiene la variable independiente  $X_1$ ) +  $b_2X_2$  ... Es importante en este apartado el apoyarse en el plano cartesiano para observar gráficamente la interacción entre las variables dependiente o endógena o variable a predecir el comportamiento, casi siempre etiquetada como variable  $Y$ , y las independientes o variables que explicarían el cambio en la variable dependiente, casi siempre etiquetadas en el eje  $X$  como variables independientes controlables, exógenas, intervinientes o extrañas o no controlables. Ver gráfica número 31.

Gráfica No.31



Las variables independientes controlables, son aquellas que se pueden manipular en cantidad, por ejemplo, el número de intensidad de un anuncio publicitario y la dosificación en miligramos de un medicamento, entre otras; también son conocidas como variables exógenas controlables

Las variables intervinientes en un modelo de regresión simple o múltiple, son todas las variables que se ubican en el eje  $X$  del plano cartesiano, sean o no controlables. Las variables extrañas no controlables, son aquellas que afectan el comportamiento de la variable dependiente, pero que están fuera de control por parte del proyecto de investigación, por ejemplo, las condiciones del medio ambiente, las variables del entorno político, social, cultural y económico entre otras, que afectarían la variable dependiente.

La regresión se clasifica en regresión lineal simple o múltiple, cuando las variables están medidas en escalas superiores de medición como son escalas Likert o razón

y en regresión logística o binaria, cuando la variable dependiente está medida en escala inferior nominal o dicotómica (variable dependiente con dos posibilidades de respuesta, por ejemplo compra o no compra, enfermo o no enfermo, sabe o no sabe, etc.) y politómicas en la regresión logística multinomial, cuando la variable dependiente tiene 3 o más opciones de respuestas, como, el estrato socioeconómico, que se divide en números de 1 a 6, el grado de la enfermedad en 4 niveles, Sano, Medio enfermo, enfermo y muy enfermo, etc. Ver tabla número 5.

Tabla No. 5

<b>Regresión lineal:</b>	<b>Regresión logística:</b>
<b>Simple:</b> Una sola variable independiente superior o métrica explicaría el cambio en la variable dependiente superior o métrica	<b>Binaria:</b> La variable dependiente con dos posibilidades de respuesta o dicotómica.
<b>Múltiple:</b> varias variables independientes superiores o métricas explicarían el cambio en la variable dependiente también superior o métrica.	<b>Multinomial</b> y ordinal: cuando la variable dependiente es politómica o con más de 3 posibilidades de respuesta.

- A. **REGRESIÓN SIMPLE:** Permite encontrar la relación que existe entre dos variables métricas conocidas como dependiente e independiente. Gracias a la ecuación matemática de la recta ( $Y=B_0+B_1X_1+\dots$ ) se puede pronosticar el comportamiento futuro entre las dos variables de una población e identificar que tanto la variable independiente explica la variación o efecto de la variable dependiente.

***Pasos a seguir en el IBM-SPSS, analizar, Regresión, Modelación lineal automatizada, selecciona la variable independiente y dependiente, estadística: estimaciones, ajuste del modelo, cambio de R cuadrado, diagnóstico de colinealidad, Durbin-Watson, aceptar.*** R cuadrado, muestra la proporción de la varianza de la variable dependiente que está explicada por la variable independiente, conocido como coeficiente de determinación. El coeficiente de correlación múltiple R, es el valor absoluto de la correlación de Pearson entre las dos variables. R cuadrada corregida, es un poco más pequeña que la R cuadrada y sigue la misma interpretación. El error típico de la estimación debe ser pequeño ya que representa la variabilidad de la variable dependiente que no es explicada por la recta de la regresión. El resultado de Durwin-Watson es útil para ver la independencia de los residuos y va entre valores de 1 a 3 cuando hay independencia, valores por debajo de 1 indican autocorrelación positiva y por arriba de 3 indican autocorrelación negativa, por lo que no se cumple la independencia de errores. La hipótesis a contrastar con Durwin-Watson es  $H_0$ : No existe autocorrelación entre los residuos y la  $H_a$ : Si existe autocorrelación entre los residuos. Se procede a analizar el valor ANOVA F y la significancia que nos muestra si hay relación entre las variables o si las variables están linealmente relacionadas con R mayor a 0; bajo la  $H_0$ : La R es igual a cero y la  $H_a$ : la R es diferente a 0. El modelo es adecuado para predecir la variable dependiente cuando la significancia es menor o igual a .05. El valor t con su grado de significancia indica si la variable aporta al modelo una significancia para predecir el comportamiento de la variable dependiente hacia la población, para ello se requiere crear la  $H_0$ : La variable  $X_1$  no aporta

significancia al modelo y la  $H_a$ : La variable  $X_1$ , si aporta significancia al modelo. Los estadísticos de colinealidad FIV (factor de varianza inflada), indica si hay o no multicolinealidad entre las variables independientes (los valores deben estar cercanos a 1), valores por arriba de 10 indican que hay multicolinealidad. Coeficiente para la ecuación de la recta: constante es el valor que representa el origen de la recta de regresión ( $B_0$ ) + el coeficiente de la variable independiente, que es la pendiente de la recta de regresión ( $b_1$ ) representa el cambio de la variable dependiente por cada unidad de cambio de la variable independiente. La ecuación de regresión es:  
 $Y=B_0+B_1(x_1)$

$Y$ =Variable dependiente

$B_0$ =Intercepción de la línea o constante.

$B_1$ =Pendiente de la línea

$X$ = variable independiente

- Estadísticos:  $R^2$ =Coeficiente de determinación. Va de 0 a 1, mide la fuerza de la asociación entre las variables y representa la proporción de la variación total en  $Y$  por el efecto de  $X$  (variable independiente); La  $R$  es el coeficiente de correlación múltiple de Pearson entre las variables;  $R$  cuadrado corregida normalmente es más baja que el  $R$  cuadrado; el error de la estimación representan una medida de la parte de variación de la variable dependiente que no es explicada por la recta de regresión, tiende a ser bajo cuando mejor es el ajuste de las variables; el valor  $F$  y el grado de significancia mide si la  $R=0$ ; determinan los valores estimados para proyectar en la ecuación matemática; El ANOVA, muestra que el modelo sirve para predecir el comportamiento de la variable dependiente.
- La estadística  $t$  para probar la hipótesis de relación entre las variables,  $B_0$  y  $B_1X_1$ .
- El FIV indica si hay correlación entre las variables independientes del modelo de regresión o ver si se cumple el supuesto de multicolinealidad (ningún valor por arriba de 10. Todos los valores deben estar cercanos a

- 1) Valores superiores a 10 indican que si hay multicolinealidad entre las variables independientes del modelo de regresión múltiple.
- La homocedasticidad o heterocedasticidad: es un procedimiento para conocer la varianza de cada una de las variables independientes que explican la variabilidad de la dependiente; se dice homocedástica cuando la varianza es igual o casi igual dentro de todos los datos de la variable independiente que se está midiendo y heterocedasticidad, cuando se presenta variabilidad de la varianza de todos los datos dentro de la variable independiente. Entre más homogénea sea la varianza de las variables independientes es mejor para predecir el comportamiento de la variable dependiente dentro de un modelo de regresión simple o múltiple; en el IBM-SPSS, se calcula: analizar, regresión lineal (pasar las variables dependientes y las independientes), gráficas (Y=ZRESID y en X=ZPRED), continuar. **La interpretación** es más visual, para ver la nube de puntos cercanos o alejados de un punto medio: ver si sigue el comportamiento lineal, visualizar los valores atípicos y los diferentes niveles de agrupación. Si existe heterocedasticidad, se procede a identificar la causa, si es por valores atípicos (se ubican los casos atípicos y se estudia la posibilidad de eliminarlos con fundamento que no afecte la realidad del comportamiento de la población), los patrones de grupos (puede existir más variabilidad en los datos de hombres versus mujeres). **La solución** se puede dar incorporando al modelo solo los valores de un segmento de la población, por ejemplo, seleccionar el grupo de los que sí están a favor de lo que se está midiendo, mujeres, estratos socioeconómicos bajos, etc. En el IBM-SPSS, se va a analizar y se selecciona variable de selección categórica o nominal y seleccionar en regla la categoría a medir y aceptar; si se observa homocedasticidad en el grupo seleccionado, se procede a realizar la regresión solo para ese grupo (En datos, seleccionar casos, seleccionar el ícono de “si se satisface la condición”, si la opción (pasa la variable y se da la condición Ejemplo, Variable compra=1, siendo 1 de si compra), en la base de datos

se podrá ver con una barra cruzada la eliminación los casos codificados como 2; se procede a realizar la regresión lineal de nuevo en analizar, regresión lineal, se quita la variable de selección y aceptar. En los resultados se puede ver la nueva homocedasticidad y los casos atípicos, para proceder a eliminar los casos atípicos (Primero se identifican dando clic en la gráfica de los casos y seleccionar los puntos alejados para identificar quienes son los que tienen esos valores atípicos). Otro método es el logaritmo neperiano, transformando las variables originales (en el IBM-SPSS, transformar, calcular variable, se le da el nombre de la variable como LnX, se pasa en el cuadro de diálogo de expresión numérica como LnX y aceptar), se procede a realizar de nuevo la regresión, pero insertando la variable con el símbolo de lnX, para ver que tanto ha mejorado la homocedasticidad, quedando pendiente de eliminar en el caso de ser necesario los casos atípicos.

***PASO A PASO: Trazar el diagrama de dispersión, Formular el modelo general, Identificar el coeficiente de regresión, ver la significancia, determinar la fuerza y significancia de la relación, calcular el valor proyectado, verificar la exactitud de lo proyectado, analizar los residuales y realizar validación cruzada del modelo.***

- Procedimiento de mínimos cuadrados para ajustar la línea recta del diagrama de dispersión, conocida como línea de regresión. La distancia vertical entre el punto y la línea de regresión se conoce como el error; se obtiene el cuadrado de las distancias de todos los puntos y se suman para llegar al total de errores cuadrados.
- Durbin-Watson, va de 1 a 3 para aceptar el supuesto de independencia de errores.
- Modelo de regresión de dos variables;  $Y=B_0+B_1(x_1)$
- Coeficiente estándar de regresión, para contrastar la Hipótesis estadística  $H_0= R^2=0$ ;  $H_1=R^2$  diferente a 0. La  $R^2$  va de 0 a 1 e indica el porcentaje de la variación de la variable dependiente por el efecto de las variables independientes. Ejemplo, una  $R^2=.70$ , indica que el 70% de la variación

de la variable dependiente se debe a la interacción de la variable independiente involucrada en el modelo de regresión simple.

- Prueba de significancia para ver si hay relación entre las variables
- Fuerza y significancia de la relación.
- Significancia del coeficiente de determinación con la Prueba F para determinar si hay o no significancia entre B1 con la Ho:  $B1=0$ ; H1: B1 diferente a 0. Si hay significancia en la relación entre X y Y, indica que se pueden proyectar los valores de Y con el uso de la ecuación del modelo de regresión.

**B. REGRESIÓN MÚLTIPLE:** Son los mismos pasos de la simple, pero con dos o más variables independientes. La hipótesis está en función de la variable dependiente versus las independientes involucradas. Regresión lineal múltiple con el **IBM-SPSS: analizar, regresión, lineal, se inserta la variable dependiente y las independientes con el método de regresión jerárquica o múltiple. Para jerárquica, se meten una a una las variables independientes de la siguiente manera: se selecciona una variable independiente, bloque 1 de 1(añadir), siguiente y se mete la siguiente variable y así sucesivamente hasta incluir las variables independientes.**

La ecuación del modelo de regresión es la misma que para la regresión simple, pero con la integración de las variables independientes:

$$Y=B_0+B_1(X_1)+B_2(X_2)+B_3(X_3)\dots \text{ Las B}$$

PASO A PASO:

- Se siguen los mismos pasos que en la regresión simple.

Interpretación:  $R^2$  ajustada, indica la variación de la variable dependiente por la interacción con las independientes. El valor F con su respectiva significancia indica que si hay diferencia en el coeficiente de regresión y que las variables son aptas para integrarlas en el modelo de proyección. La t indica el nivel de significancia del coeficiente para cada una de las variables independientes. El FIV indica si se cumple el supuesto de multicolinealidad (ningún valor por arriba de 10. Todos los valores deben estar cercanos a 1)

- Prueba residual: es la diferencia entre el valor observado de “Y” y el valor proyectado por la ecuación de regresión
- La tabla de coeficientes muestra el valor de la constante y el valor de cada una de las variables independientes para utilizarlos en la fórmula de la regresión para proyectar el comportamiento de la variable dependiente en función de las independientes según la siguiente fórmula:  
$$Y=B_0+B_1(X_1)+B_2(X_2)+B_3(X_3)\dots$$

C. **REGRESIÓN DISCRIMINANTE:** seleccionar entre un gran grupo de variables independientes un subgrupo que afecta en mayor proporción la variación de la variable dependiente. Se insertan en el modelo todas las variables independientes y se van eliminando aquellas con poca incidencia en la  $R^2$  y F. Es importante en esta modalidad de Regresión Múltiple el identificar el grado de multicolinealidad (Estado de intercorrelación entre las variables independientes)

Evaluación de la significancia de las variables de predicción: La prueba F y  $R^2$ ; ver como se cambia el  $R^2$  cuando se van agregando o quitando variables. Se debe analizar la interacción de los valores de coeficiente estandarizados de  $B_1...$  ya que estos toman en cuenta el efecto de la interacción entre las variables independientes.

D. **VALIDACIÓN CRUZADA:** se realiza la corrida estadística con todas las variables y se comparan los datos principales con  $R^2$ , F y T con subgrupos de la misma muestra para ver si hay similitud entre los resultados.

**E. REGRESIÓN LOGÍSTICA:** Es una técnica estadística que permite diseñar un modelo de predicción similar al análisis de regresión lineal, pero en términos de probabilidad de ocurrencia de una variable dependiente cualitativa, categórica en forma dicotómica (la variable tiene dos posibilidades de respuestas, ejemplo Si/No, Femenino/Masculino, Enfermo/No enfermo, etc.) o politómica (Variable con más de dos posibles respuestas, ejemplo, Alto, medio y bajo; segmento A, segmento B y segmento C, etc.). Se mide en término de probabilidad para conocer que tanto una variable independiente cualitativa o nominal, influye en la variable dependiente también cualitativa. Como son variables categóricas, se transforman logarítmicamente (logit) para cumplir el supuesto de linealidad. La verosimilitud-Log, es la suma de probabilidades asociadas con los resultados observados y esperados, es similar a la suma de cuadrados en la regresión simple y mide la información que no se puede explicar después de que el modelo es ajustado.

**SUPUESTOS:** Linealidad con la regresión logística, no multicolinealidad, independencia de errores, información incompleta de variables predictoras y/o datos insuficientes en cada una de las casillas de los datos.

**UTILIDAD:** para predecir con un modelo, la inclinación hacia una variable dependiente nominal dicotómica, en función de un conjunto de variables medidas en escala nominal, ordinal o escalar, expresadas en Odds (porcentajes) de cada una de las variables independientes del modelo, por ejemplo, se puede medir los atributos que afectan el buen desempeño académico, atributos que influyen en un buen estado de salud, atributos que influyen en la drogadicción, etc. Con el diseño del modelo se puede conocer la probabilidad de que suceda un evento en función de la interacción relativa de las variables independientes, por ejemplo, se pueden conocer el peso relativo de variables independientes que ayudan a predecir si una persona puede ser sujeto de crédito o no, si una persona según sus hábitos alimenticios y de estilo de vida afectarán su estado de salud o no, conocer los riesgos de múltiples enfermedades para fumadores versus los no

fumadores, conocer las variables que llevan al alcoholismo y neurosis de una persona, etc.

**Con el apoyo del IBM-SPSS**, se procede: analizar, regresión logística binaria, seleccionar las variables y las especificaciones para el análisis. En el campo Covariables categóricas se introduce la lista de las variables identificadas como categóricas; en opciones seleccionar estadísticos, gráficos, correlaciones de estimación, historial de iteraciones en IC para expB. En guardar se da probabilidad para crear nueva columna de variables dentro del archivo de datos con los valores en forma de probabilidad de ocurrencia o influencia para cada caso.

**RECOMENDACIÓN:** las variables se recomienda codificarlas en el SPSS como dummies, esto es, codificarlas entre 0 y 1, siendo 0 ausencia o equivalente a NO y 1 como SI o presencia de lo que se está midiendo. Las opciones disponibles son: De Cook, Valores de influencia y DfBeta(s). El campo Residuos guarda los residuos. Las opciones disponibles son: No tipificados, Logit, Método de Student, Tipificados y Desviación.

**F. REGRESIÓN LOGÍSTICA BINARIA Y MULTINOMIAL:** la regresión logística binaria se utiliza cuando la variable dependiente tiene solo dos posibilidades o de dos categorías, por ejemplo: si le agrada el producto versus no le agrada; presencia de enfermedad y ausencia de enfermedad; si tiene intención de comprar versus el que no tiene intención de comprar; el que sí es cliente versus el que no es clientes; etc. Cuando la variable dependiente tienen más de 2 opciones de respuesta, se conoce como regresión logística multinomial, por ejemplo, nivel de ingresos alto, medio y bajo; soltero, casado y divorciado; compra alta, media y baja; actitud positiva, media y baja hacia el producto o servicio, intención de compra alta, media y baja, etc. **Con el apoyo del SPSS, se procede: Analizar → Regresión → Logística binaria: seleccione la variable cualitativa dependiente (con dos opciones de respuesta), en factores seleccionar la variable independiente nominales, en variables independientes o covariables seleccionar las escalares o nominales, en guardar seleccionar valores**

***pronosticados/probabilidad, en opciones seleccionar bondad de ajuste de Hosmer-Lemeshow y gráficos de clasificación.***

**Análisis e interpretación:**

En regresión logística multinomial se procede en el SPSS de la siguiente manera: ***analizar*** → ***Regresión*** → ***Logística multinomial: seleccionar la variable cualitativa dependiente (con más de dos opciones de respuesta), en factores o covariables seleccionar las variable independiente nominales, en modelos seleccionar efectos principales*** (proporciona los efectos principales de las covariables y los factores, pero no proporciona efectos de interacción. El modelo factorial, proporciona todos los efectos principales y todas las interacciones de factor por factor, sin interacciones de covariables. Existe el modelo personalizado, donde permite seleccionar los factores (F) fijos y las covariables (C), el ícono modelo es útil para diseñar el modelo en función de los efectos principales y de interacciones que se seleccione, permite incluir o excluir las intersecciones).

***En estadísticos se sugiere seleccionar resumen de procesamiento de casos, Pseudo R cuadrado, resumen de pasos, información de ajuste de los modelos, en parámetros seleccionar intervalo de confianza del 95%, estimaciones, pruebas de la razón de verosimilitud y correlación asintótica, en guardar seleccionar probabilidad de la categoría pronosticada y en opciones seleccionar escala de dispersión Pearson.***

**Análisis e interpretación:**

G. **MODELOS PROBIT Y LOGIT:** el análisis probit mide la explicación y predicción de cada una de las variables independientes en forma de porcentaje sobre la variable dependiente. Mide la fuerza de la relación entre la intensidad de un estímulo y la proporción de casos que presentan respuesta a dicho estímulo. Es útil para las situaciones en las que se dispone de una respuesta dicotómica que puede estar influenciada o causada por los niveles de algunas variables independientes. Este procedimiento le permitirá estimar la intensidad necesaria para que un estímulo llegue a inducir una determinada proporción de respuestas, por ejemplo, conocer la dosis efectiva de un nuevo medicamento para que se dé una respuesta favorable. Es útil para modelos experimentales que lleve a identificar la dosis necesaria para lograr los resultados esperados. También es útil para conocer la proporción de la variable independiente que puede desencadenar una enfermedad, como la hipertensión y la diabetes mellitus entre otras.

**En el IBM-SPSS, se procede: analizar, regresión,** Probit, seleccionar las variables y las especificaciones para el análisis; seleccionar el ícono Factor para indicar qué niveles del factor desea utilizar; en el ícono de covariables se introducen otras variables independientes para el modelo; en modelo se debe elegir probit; en opciones probit y en estadísticos seleccionar frecuencias, potencia relativa de la mediana (potencia relativa de la mediana y prueba de paralelismo se activan cuando se selecciona una variable factor). El ícono de tasa de respuesta natural, y aceptar.

Como estadísticos se obtienen los coeficientes de regresión y errores típicos, intersección y su error típico, Chi-cuadrado de Pearson de la bondad de ajuste, frecuencias observadas y esperadas e intervalos de confianza para los niveles efectivos de la variable o variables independientes. Como diagramas se obtienen los gráficos de respuestas transformadas.

H. **PROCEDIMIENTO REGRESIÓN DE COX:** Este modelo permite incluir en los modelos variables predictoras (covariables), proporcionando estimaciones de los coeficientes para cada una de las covariables-independientes de tipo nominal, escalar o continuas, permitiendo evaluar el impacto de múltiples covariables (Variables independientes) en el mismo modelo. Como estadísticos se obtienen para cada modelo:  $-2LL$ , el estadístico de la razón de verosimilitud y el Chi-cuadrado global. Para las variables dentro del modelo: Estimaciones de los parámetros, errores típicos y estadísticos de Wald. Para variables que no estén en el modelo: estadísticos de puntuación y Chi-cuadrado residual. En cuanto a los datos, la variable de tiempo debería ser cuantitativa y la variable de estado puede ser categórica o continua. Las variables independientes (las covariables) pueden ser continuas o categóricas; si son categóricas, deberán ser auxiliares (dummys) o estar codificadas como indicadores. Las variables de estratos deberían ser categóricas, codificadas como valores enteros o cadenas cortas. En el SPSS, se selecciona el ícono de analizar, supervivencia, regresión de Cox, seleccione las variables del modelo para caracterizar el valor o rango de valores que determinan el evento. Seleccione variables independientes Covariables (existe una opción dentro del procedimiento para recodificar las variables categóricas automáticamente), se puede elegir variables categóricas como covariables; seleccionar opciones para definir distintas características a obtener en la regresión y en gráficos se definen los distintos gráficos a obtener y Aceptar.

### III. ANÁLISIS DE FACTORES

Es un análisis de interdependencia donde no existe variable dependiente e independiente, sino análisis de correlación y agrupación de conjunto de variables con cierto grado de relación. Se requiere que las variables estén medidas en escalas superiores o métricas (intervalar o de razón). Es útil para reducir el número de variables en factores o grupos de variables con un significativo de relativa importancia para cada factor de agrupación, por lo que es muy útil para simplificar en un gran número de datos métricos en factores que posteriormente se pueden integrar en modelos predictores como regresión múltiple y análisis discriminante.

El KMO (Kaiser-Meyer-Olkin) es un indicador de validez que compara los coeficientes de correlación simple con los coeficientes de correlación parcial, sus valores van de 0 a 1 (0.9 a 1 muy bueno, 0.8 a 0.9 bueno, de 0.8 a 0.7 regular, 0.7 a 0.6 bajo, 0.6 a .05 muy bajo y menor a 0.5 inaceptable).

Se debe extraer los factores con el peso relativo de cada variable en cada uno de los factores de la matriz de componentes rotada y darle el nombre correspondiente a cada factor.

Análisis estadístico del análisis de factores:

- En la tabla de estadísticos descriptivos, se pueden ver las medias y las desviaciones estándar, para ver cuál o cuáles de las variables presentan cifras muy altas de desviación estándar.
- Prueba de esfericidad de Bartlett: estudia si las variables están correlacionadas o no en la población, bajo la  $H_0: R=0$  y  $H_a: R$  mayor a 0., se observa el valor KMO y esfericidad de Bartlett, con el valor de significancia correspondiente. Si es menor o igual a .05 se rechaza  $H_0$  y se acepta la  $H_a$ , para concluir que si hay diferencia significativa en la correlación, por lo que se pueden agrupar las variables en los componentes principales. El KMO: es un índice que muestra que tan apropiado es el análisis de factores. Valores arriba de 0.6 indica que el análisis es apropiado.

- Matriz de correlación: muestra la correlación entre todo el grupo de variables. Las variables que más correlacionan entre sí, indica que estas podrían estar en uno de los componentes principales.
- Comunalidad: se observa los valores de extracción que formarán parte de cada componente. Es la cantidad de varianza que una variable comparte con las demás variables dentro de un factor. Es la proporción de la varianza que explica a cada factor donde cargan. Lo ideal es que esté arriba .5
- Varianza total explicada: valor específico representa la varianza total que explica a cada factor. En la columna de componentes y autovalores iniciales, se pueden identificar el número de factores o de componentes principales con el porcentaje individual y acumulativo de la varianza explicada. En la columna de sumas de extracción de cargas al cuadrado, se puede ver el porcentaje de varianza que carga en cada uno de los componentes o factores, con su respectivo valor acumulado a medida que se integra un componente más en el análisis de forma descendente y acumulativa.
- Gráfico de sedimentación: se puede observar gráficamente los componentes principales que integrarán el peso relativo de cada una de las variables que fueron tomadas en cuenta en este análisis. Se toman en cuenta los valores superiores a 1 para seleccionar los componentes.
- Cargas factoriales: son correlaciones entre las variables en cada uno de los factores donde cargan, con valores de .0 a 1, se seleccionarán las variables que mayor carga tenga en cada uno de los componentes.
- Matriz de componente rotado: muestra todas las variables que cargan en cada uno de los componentes, que van de .0 a 1. En esta tabla se procede a identificar cada una de las variables que más carguen en cada uno de los componentes y se procede a etiquetar ese componente con un nombre que tenga relación con el grupo de variables que allí se integran. Este componente pasa a ser una nueva variable que puede ser utilizada

con otros estadísticos como regresión, ANOVA y análisis multivariante entre otros.

- Porcentaje de varianza: muestra el porcentaje de varianza total que se atribuye a cada factor.
- Determinar el número de factores a priori. El número de factores que se extraen deben representar por lo menos el 60% de varianza.

***PASO A PASO EN EL IBM-SPSS: analizar, reducción de dimensiones, factor, seleccionar y pasar las variables; en descriptivos: solución inicial, descriptivo univariados, en matriz de correlaciones, seleccionar coeficientes, determinantes y KMO; en Extracción: en métodos seleccionar el método de componentes principales, Matriz de correlación, visualización, gráfico de sedimentación, en extraer seleccionar autovalores mayor a 1; en Rotación, seleccionar Varimax. Solución rotada y gráfico de saturación; Puntuaciones factoriales para guardar como variable para la correlación; Opción: excluir casos según lista, en formulario de presentación de los coeficientes se debe seleccionar ordenar por tamaño, valor absoluto ideal de 0.4 y aceptar.***

#### IV. ANÁLISIS DISCRIMINANTE

Para realizar esta prueba se necesita que la variable dependiente esté medida en escala nominal tipo dicotómica (también se puede utilizar nominal con más de dos posibilidades, por ejemplo, el nivel socioeconómico clasificado en bajo, medio y alto) y las independientes tienen que estar medidas en escala intervalar o de razón (escalas superiores). Los niveles de la variable dependiente marcarán la diferencia si estamos ante la situación del análisis simple o múltiple, esto es: el análisis se considera simple cuando la variable dependiente tiene solo dos posibilidades, ejemplo, personas que si compran Vs las que no compran y de múltiple podría ser, los que compran mucho, medianamente y poco. Este análisis discrimina de manera objetiva los grupos mostrando las variables independientes que marcan la diferencia en cada uno de los grupos, mismas que formarán parte de la función discriminante de tipo  $Z = B_0 + B_1(X_1) + B_2(X_2) \dots B_n(X_n)$

$B_0$  = Constante

$B_1$  = Valor del coeficiente para la variable ( $X_1$ )

**En el IBM-SPSS, analizar, clasificar, discriminante, variable de agrupación pasar la variable dependiente nominal y definir los rangos o agrupación, pasar las variables independientes numéricas, en descriptivos seleccionar medias y ANOVA, en clasificar seleccionar: calcular según tamaño de grupo, en gráfico seleccionar grupos combinados, grupos separados y mapa territorial, continuar, en guardar se selecciona las tres primeras, continuar y aceptar.**

Los estadísticos más utilizados son:

1. Descriptivos: medias de los grupos de cada una de las variables para ver diferencias de medias entre los grupos. ANOVA, muestra si hay igualdad de medias para cada variable entre los grupos. La prueba de Box, muestra si hay varianza-covarianza igual en los grupos (homocedasticidad). Una significancia menor o igual a .05, rechaza la  $H_0$ : Las varianzas en los grupos son iguales y se acepta la  $H_a$ : Las varianzas en los grupos son diferentes.
2. F de Snedecor: compara la desviación de las medias para cada variable entre los grupos en función de la media total entre la desviación a la media dentro de cada grupo, para identificar si existe diferencia de medias entre los grupos

del modelo discriminante. Entre más grande sea F, indica que las medias de las variables de cada grupo están separadas, por lo tanto, discriminan muy bien los grupos.

3. Correlación Canónica o coeficiente eta:

### Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1.101 <sup>a</sup>	100.0	100.0	.724

- a. Se han empleado las 1 primeras funciones discriminantes canónicas en el análisis.

El 72% de la varianza en la variable dependiente se explica o representa por medio del modelo. Más cercano a 1 indica que el modelo discrimina muy bien, por lo que si hay desviaciones de las puntuaciones discriminantes entre los grupos en relación a la desviación total.

### Lambda de Wilks

Contraste de las funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	.476	64.204	3	.000

4. Lambda de Wilks: útil para rechazar la hipótesis de igualdad entre las medias de los grupos; va de 0 a 1 y confirma si el modelo (función discriminante) es bueno o no, para predecir. Entre más cercano a cero el valor de Lambda de Wilks, indica mejor discriminación de la variable, mientras que entre más se acerque a 1 indica menor discriminación de la variable en el grupo; la F es más pequeña y por ende el alfa o significancia será mayor a .05. (muestra la desviación a la media dentro de cada grupo entre las desviaciones a la media

total). En el ejemplo, con el valor de Lambda de Wilks .476 y Chi-cuadrada de 64.204, con 3 grados de libertad, se procede a rechazar la Hipótesis nula con el nivel de significancia de .000. La hipótesis es:

- Ho: Las medias en la función discriminante son iguales en los dos grupos
- Ha: Las medias en la función discriminante son diferentes en los dos grupos.

Se concluye en este apartado que se rechaza Ho y se acepta Ha, ya que las medias en la función discriminante entre los grupos son diferentes...El Modelo es bueno para predecir con una Lambda de Wilks de .476 (Lo ideal es que sea menor o igual a .45)

### Matriz de estructura

	Función
	1
Evaluación del grado de aprendizaje	.988
Nivel del uso de la tecnología	.416
Nivel de integración con los trabajos colaborativos	.124

Correlaciones intra-grupo combinadas entre las variables discriminantes y las funciones discriminantes canónicas tipificadas

Variables ordenadas por el tamaño de la correlación con la función.

5. Matriz de coeficientes estandarizados para seleccionar las variables con valor absoluto, para integrarlas en la ecuación predictora de agrupación. El coeficiente de la función de clasificación bajo la normalidad va de 0 a 1, indicando la importancia absoluta que tiene cada variable en la función

discriminante. En este ejemplo se observa que la variable evaluación del grado de aprendizaje, con un valor de 0.988 tiene un grado importancia mayor que el nivel del uso de la tecnología con 0.416, siendo el nivel de integración con los trabajos colaborativos el de mayor importancia con .124.

La función discriminante estaría reflejada en la siguiente fórmula:

$$Z=0.988(\text{Evaluación del grado de aprendizaje}) +.416(\text{Nivel del uso de la tecnología})$$

Con esta ecuación podemos predecir las variables que inciden en el grupo de los que si consideran más atractivo estudiar en el TEC y guía a la organización a definir sus estrategias de abordaje para captar y retener estudiantes.

### **Funciones en los centroides de los grupos**

	Función
Estudias en el Tec?	1
Si	-1.037
No	1.037

**Funciones discriminantes canónicas no tipificadas evaluadas en las medias de los grupos**

Se puede observar que el centroide muestra la media de la función discriminante en cada grupo. El símbolo negativo en este caso se interpreta en que las variables involucradas dan medidas muy cercanas al 1 lo que podemos inferir que los sí, evalúan de manera más positiva los tres atributos que se están midiendo (recuerden que 1 significa más Atractividad y 5 menos Atractividad)

### Resultados de la clasificación

			Grupo de pertenencia pronosticado		Total
			Si	No	
Original	Recuento	Si	42	3	45
		No	7	38	45
	%	Si	93.3	6.7	100.0
		No	15.6	84.4	100.0

a. Clasificados correctamente el 88.9% de los casos agrupados originales.

En este último apartado del análisis discriminante se pretende validar el modelo, que puede ser de manera directa, cuando realiza el análisis con todas las personas de la muestra a la vez o bien de manera parcial o discreta, esto es, separa la muestra en dos grandes submuestras y se realiza de manera separada las corridas estadísticas o también se podría hacer de manera cruzada, esto es, que se intercambiarían elementos de la muestra en cada una de las dos submuestras. Podemos ver que se clasificaron de manera correcta el 88.9 % de los casos. Hubo 7 casos incorrectos en los sí y 3 casos incorrectos en los no (personas que en lugar de pertenecer al grupo del si pertenecen al de los no y viceversa)

## V. ANÁLISIS DE CLUSTER

Permite agrupar personas u objetos y no variables (se agrupan con el análisis de factores) en medidas de asociación y/o similitud entre casos analizados, esta agrupación debe ser homogénea en cada grupo y diferente entre los grupos. El algoritmo de agrupación se puede hacer por el método jerárquico y el K media (número grande de casos). Las variables deben estar medidas en escala intervalar o razón.

***En el IBM-SPSS, analizar, clasificar, clúster jerárquico, en variables se pasan todas las escalares, en estadísticos seleccionar historial de conglomerados, matriz de proximidad y solución única para seleccionar el número de clúster (pueden ser 3 o 5 dependiendo de los segmentos o clúster que desee analizar), en gráficos seleccionar dendograma, continuar, en método seleccionar vínculos entre grupos distancia euclídea al cuadrado y en transformación seleccionar puntuaciones Z, continuar y aceptar.***

Analizar: en resumen, de procesamiento de casos, muestra los casos que fueron seleccionados para el análisis. En matriz de proximidades se observa la distancia que hay entre las variables, en la primera columna de la matriz matricial se puede observar el valor menor entre las variables, para identificar entre cuales variables se encuentra la distancia menor y por ende su relación. En la tabla de historial de conglomeración se puede observar la agrupación de las variables en relación con el valor menor de cada una de las columnas de la matriz de aproximación, para mostrar la agrupación entre las variables. En la tabla de clúster de pertenencia se pueden ver las variables y los clústeres a los que apuntan, y en el dendograma se puede ver de manera visual lo que en tablas arrojan las variables.

## VI. ECUACIONES ESTRUCTURALES

Es una técnica estadística multivariable, que inicia con la identificación y agrupación de los ítems de un cuestionario dentro de un concepto dado a priori para integrar las variables independientes (exógenas o predictoras) y dependientes (endógenas, latentes o de criterio) que permite el diseño del flujograma de la relación entre las variables o path análisis y la interacción causa-efecto de las mismas, con una lectura que va de izquierda a derecha. Es una integración de los análisis multivariable de factores (agrupación de ítems en componentes principales), de regresión (causalidad de la interacción de las variables independientes con la dependiente y las dependientes con otras variables dependientes: qué tanto las variables independientes explican el cambio en la variable dependiente), y análisis de covarianza para encontrar que tanto las variables independientes se afectan entre sí, para generar ecuaciones que representen la magnitud de las relaciones con el fundamento teórico del modelo.

Con el modelo de ecuaciones estructurales se pueden analizar de manera simultánea la interacción de dos o más variables dependientes frente a un grupo de variables independientes, para encontrar la relación directa e indirecta de causalidad dentro del flujograma de agrupación-causa-efecto, ya que algunas variables dependientes (endógenas, latentes o de criterio) pueden ser registradas como independientes para ver la relación directa e indirecta con la variable de efecto principal del modelo o dependiente principal.

Es una técnica estadística útil para identificar la relación entre las hipótesis de la investigación, se puede identificar relaciones directas e indirectas dentro del flujograma.

Es recomendable identificar las fases para el diseño de un proyecto de investigación con ecuaciones estructurales para tener orden y limpieza al momento de aplicar las ecuaciones estructurales:

- Fase I. Diseño del modelo: con el apoyo de la teoría o la experiencia se diseña el modelo que intenta determinar las variables independientes que se relacionan y agrupan para explicar el comportamiento de la variable dependiente. Se establecen las variables que se medirán con las preguntas del cuestionario (variable exógenas o independientes) y las variables que se conformarán con la agrupación de las mismas (Variables latentes o endógenas). Modelo que servirá de base para contrastarlo con los resultados de la investigación.

- Fase II. Diseño del cuestionario integrando las variables independientes (exógenas o variables observadas) y las preguntas de las variables endógenas o variable dependiente. Se sugiere que cada una de las preguntas estén medidas en escalas superiores (intervalar y/o de razón).
- Fase III. Análisis y resultados del modelo. Para tener mejores resultados se sugiere tamaños de la muestra superior a 250 personas.

Gracias al programa inicial de Jöreskog, en 1973, “Relación de Estructura Lineal” (LISREL-Linear Structure Relation-, por sus siglas en inglés), se evolucionó en 1986 al LISREL VI (Jöreskog y Sörbom, 1986), dando inicio a la generación de nuevos softwares como el EQS de Bentler (1985) y el Analysis of Moment Structure – AMOS- (Arbuckle, 1997 y Birney, 2001) de fácil interacción con la plataforma IBM-SPSS. Estos modelos estructurales de causalidad con diagramas de rutas reciben el nombre de SEM (Structural Equation Models) o modelos LISREL (Hayduk, L. 1987). Generalmente estos modelos se utilizan para confirmar lo planteado por el marco teórico, de ahí el nombre de “modelos confirmatorios”. En la investigación aplicada a resolver problemas del mercado o de otra índole se pueden usar estos modelos sin previa causalidad, pero si con planteamiento empírico e intuitivo de la relación causal de las variables, por ejemplo: la variable dependiente ventas, imagen de la empresa, posicionamiento, satisfacción e intención de compras entre otras están en función de un número definido de variables independiente que se pueden agrupar en factores que a su vez se integran por un número determinado de variables o ítems, para dibujar el flujograma de la interacción causa-efecto.

Requisitos para el uso del SEM en las investigaciones:

1. Las escalas de medición deben ser superiores tipo Likert y/o de razón (escalar según el cuadro de diálogo en el IBM-SPSS), aunque el modelo permite integrar variables inferiores en escala nominal.
2. Las variables deben seguir la ley de la distribución normal y representatividad de la muestra hacia la población para poder inferir.
3. El tamaño de la muestra debe ser representativo de la población y selección probabilística para poder inferir. Se recomienda un tamaño de la muestra en proporción al número de ítems multiplicado por 10 a 20 personas que forman parte de la muestra, por ejemplo si el cuestionario tiene 50 preguntas se recomienda aplicar el cuestionario a 500 personas, para que el modelo pueda funcionar mejor.
4. Se requiere diseñar el flujograma con estructuras de causa-efecto, entre las variables exógenas (Independientes) y su relación con las endógenas (Dependientes o latentes) que a su vez pueden causar efecto a otras variables endógenas, dando cada interacción la generación de las hipótesis

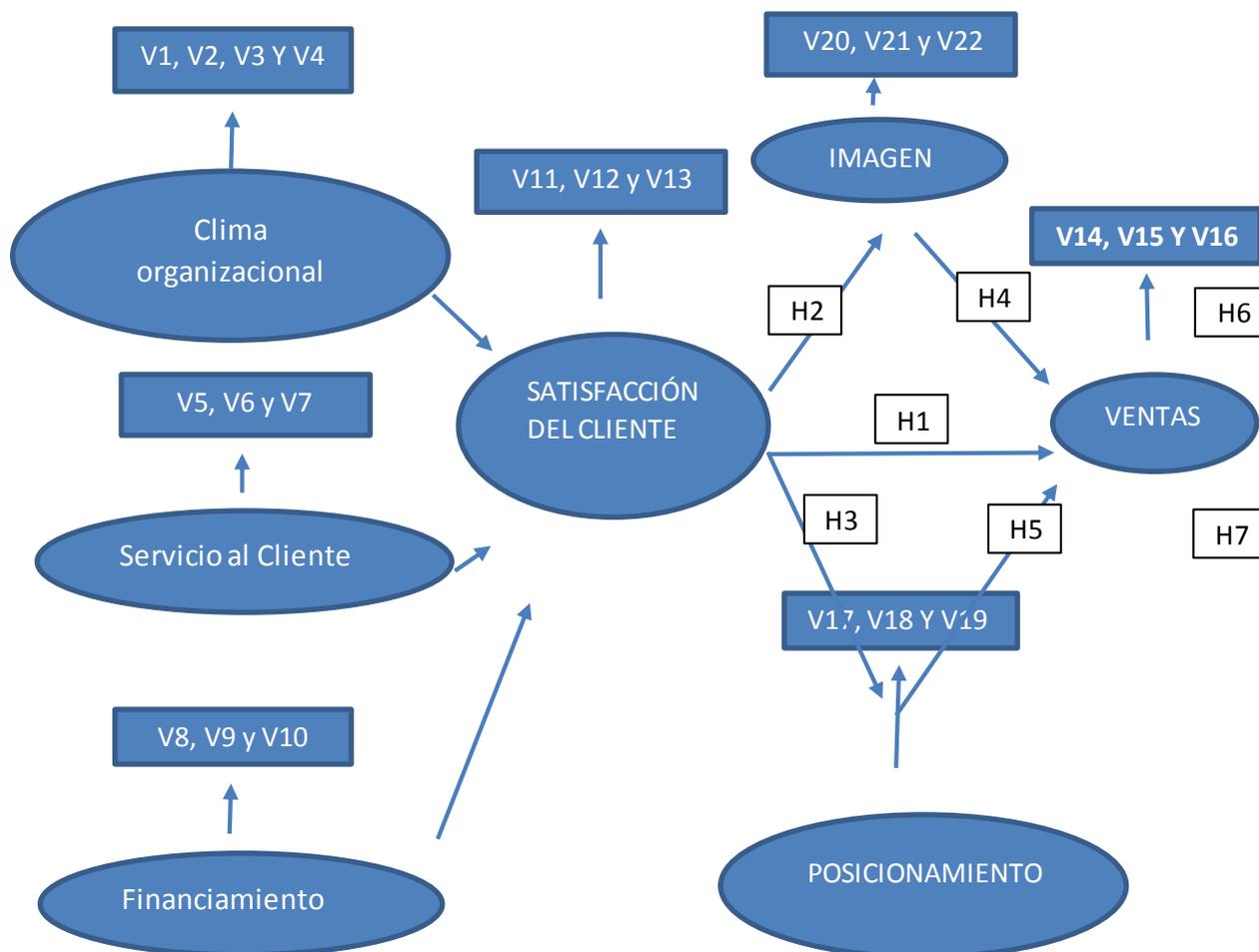
del modelo. La idea es diagramar la relación de causa y efecto entre las variables con la formulación de su respectiva hipótesis. También es conocido como Path Analysis o análisis de rutas (Duncan, 1966).

Es recomendable que los modelos de los diagramas de flujo de agrupación se den desde los ítems (Variables exógenas o independientes) que agrupados generan constructos (Variables endógenas o latente) que permiten identificar la relación de causa-efecto entre variables exógenas agrupadas en un concepto o constructos que a su vez pueden ser variables independientes dentro del modelo para identificar la relación directa e indirecta con la variable dependiente central ; estos flujogramas se deben diseñar en función de la revisión del estado del arte, tomando en cuenta las siguientes reglas estructurales:

- a. Definir los ítems o preguntas del cuestionario debidamente relacionadas con cada una de las variables conocidas como exógenas o predictores (Variables independientes) que afecten de manera directa o indirecta las variable dependientes conocidas como endógenas o de criterio (Variables dependientes).
- b. Las variables o ítems del cuestionario se deben representar en rectángulos.
- c. Los constructos formados por los ítems de cada variable exógena y endógena se representan en figuras tipo ovalo y se recomienda por lo menos tres ítems por variable agrupadas con el análisis factorial confirmatorio.
- d. Las variables extrañas se representan en círculos.
- e. Los errores se representan con el símbolo griego:  $\epsilon$
- f. La relación de Las covarianzas se representan con flechas curvas en las dos direcciones.
- g. La relación del efecto causal se representa con una flecha recta de una vía, saliendo desde la variable independiente (Exógena) y llegando a la variable dependiente (Endógena o latente) con los parámetros estandarizados correspondientes de relación entre las variables.
- h. Los parámetros de relación entre las variables se dan en correlaciones, efectos directos, indirectos, recíprocos y bucles.
- i. El término error se da en los efectos de las variables ( $\epsilon$ ). La variable que recibe el efecto de otra variable dentro del modelo deberá mostrar el error.
- j. Cada variable muestra su varianza y en caso de la variable dependiente, se mostrará la varianza explicada.
- k. Diagramar la relación causa y efecto entre el conjunto de variables de izquierda a derecha. Tener presente que algunas variables juegan el rol de exógenas y de endógenas, por lo que se debe diagramar de manera precisa la relación entre ellas. (En el análisis de regresión múltiple esto no se puede hacer ya que el modelo y la ecuación predictora de la regresión define desde un inicio la variable dependiente y las independientes), por ejemplo: definir la relación que existe entre las variables independientes clima organizacional (medido con los ítems o variables del cuestionario V1, V2, V3 y V4), servicio

al cliente (medido con los ítems o variables del cuestionario V5, V6 y V7), y financiamiento de la empresa hacia los clientes (medido con los ítems o variables del cuestionario V8, V9 y V10), para medir la relación con la satisfacción de los mismos (medida con los ítems o variables del cuestionario V11, V12 y V13), y que tanto se impacta en la imagen de la empresa (medida con los ítems o variables del cuestionario V20, V21 y V22), el posicionamiento (medido con los ítems o variables del cuestionario V17, V18 y V19), y finalmente en las ventas (medida con los ítems o variables del cuestionario V14, V15 y V16), para un total de 20 preguntas o ítems de un cuestionario estructurado que se agrupa en 7 variables debidamente relacionadas dentro del modelo de la ecuación estructural de causa y efecto, tomando como variable dependiente principal las ventas. En este ejercicio se tendrían 7 hipótesis. Ver diagrama de flujo No. 1.

Diagrama de flujo No.1



Las 7 hipótesis de este modelo serían: H1: No hay relación significativa entre la satisfacción y las ventas; H2: No hay relación significativa entre la satisfacción y la imagen; H3: No hay relación significativa entre la satisfacción y el posicionamiento; H4: No hay relación significativa entre la imagen y las ventas; H5: No hay relación significativa entre el posicionamiento y las ventas; H6: No hay relación significativa entre la satisfacción del cliente y la modulación de la imagen sobre las ventas y H7: No hay relación significativa entre la satisfacción del cliente y la modulación del posicionamiento sobre las ventas.

La relación causal entre las variables se puede dar de manera directa, indirecta y recíproca: la relación causal directa se diagrama con una flecha de un solo sentido que muestra la magnitud de relación entre la variable independiente (causa) y la variable dependiente (efecto) y la indirecta también se diagrama con una flecha de

un solo sentido relacionando tres o más variables independientes comprometidas en el modelo estructural para mostrar el efecto relativo de importancia de cada una de las variables independiente sobre la variable dependiente. Estas variables pueden tomar el nombre de variables moduladoras (variables que actúan en una relación indirecta) ya que sirven para medir el efecto relativo sobre la variable dependiente. La relación causal recíproca se refiere a la bidireccionalidad, representada con dos flechas separadas orientadas en sentido contrario (bucle de retroalimentación entre dos variables) que puede ser directa e indirecta implicando otras variables. Siguiendo el ejemplo del diagrama de flujo No. 1, pero solo con la relación entre la satisfacción, posicionamiento y ventas, se puede decir que la satisfacción del cliente (variable independiente X) impacta de manera directa en el posicionamiento (variable dependiente Y1) que se tiene de los productos de la empresa y este de manera directa en las ventas conocida como variable dependiente Y2. (Ver diagrama de flujo No.2); al modificar el modelo según el diagrama de flujo No. 3, se puede observar que existe un efecto indirecto entre la satisfacción del cliente y las ventas, el cual puede ser afectado por la variable moduladora posicionamiento. Se debe considerar que el efecto indirecto entre las variables no elimina la posibilidad de medir el efecto directo entre las variables, en este caso, se sugiere medir el efecto directo entre la satisfacción del cliente y las ventas, para sumarle el efecto indirecto entre la satisfacción del cliente y el posicionamiento más el efecto entre el posicionamiento y las ventas. En el diagrama de flujo No. 4, se puede observar la relación causal recíproca entre la satisfacción del cliente y el posicionamiento, se debe medir antes de cerrar el bucle, para conocer el efecto sobre las ventas; entre mayor sea el posicionamiento y la satisfacción del cliente se puede tener mejores ventas. Ver diagrama de flujo 2, 3 y 4

Diagrama de flujo No. 2 con efecto o relación indirecta.

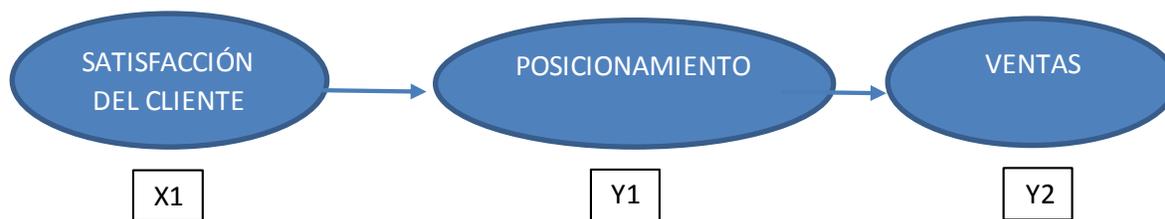


Diagrama de flujo No. 3 con efecto o relación directa e indirecta

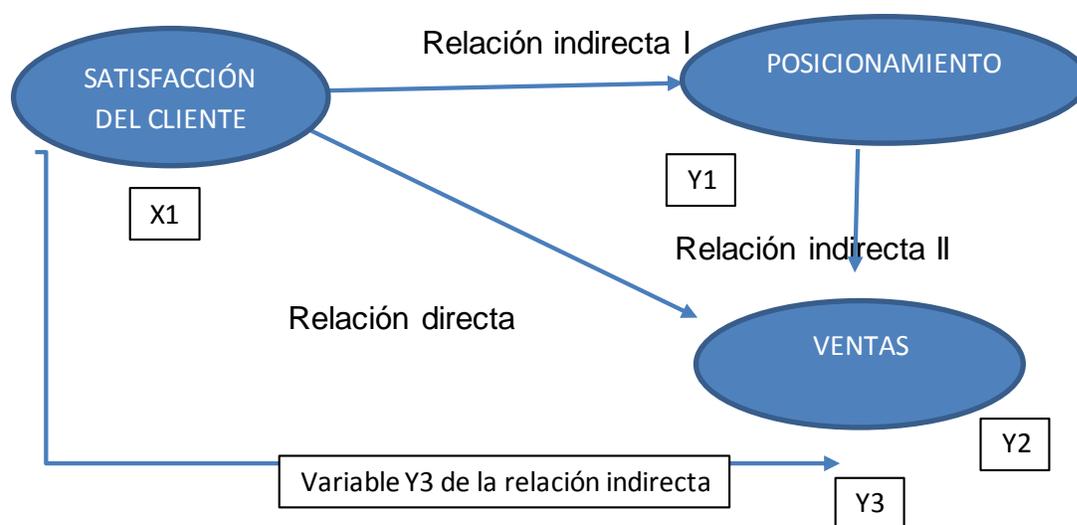
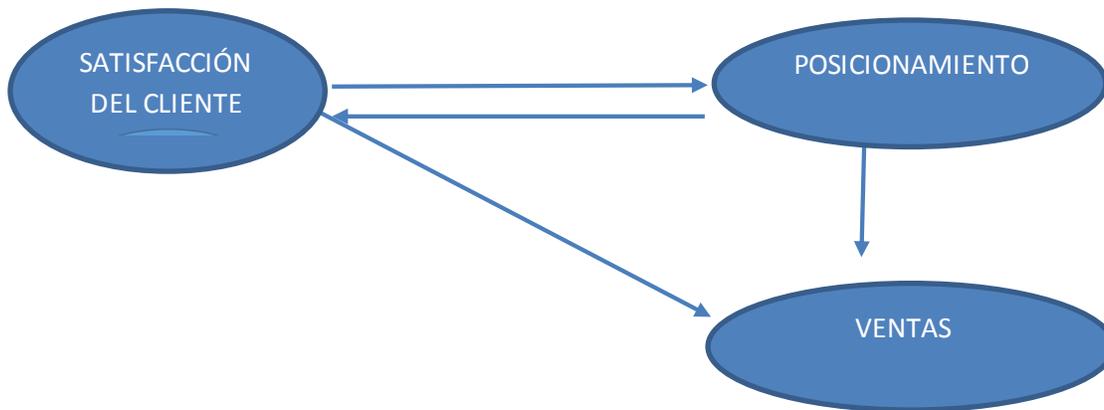


Diagrama de flujo No. 4 con efecto o relación recíproca o bidireccionalidad



- l. Relación Espuria: una relación espuria se da en la interacción entre tres o más variables, comprende la covarianza alta entre dos de tres o más variables independientes, pero puede no mostrar una relación causal con la variable dependiente, por lo que se recomienda ajustar el modelo eliminando una de las variables que correlacionan alto para ver el efecto sobre la relación causal o sobre la dependiente.
- m. Efecto total: el efecto total se mide registrando la suma de los efectos directos e indirectos sobre la variable endógena.
- n. Ajuste del modelo: se debe ajustar las covarianzas (correlaciones altas entre las variables independientes. Superiores a .70)
- o. Parámetros: miden la magnitud de la relación entre la variable independiente (exógena) y la dependiente (endógena). Se representan con letras griegas que se colocan sobre las flechas. La letra griega  $\gamma_{ij}$  (i representa el subíndice numérico de la variable endógena o dependiente y la j representa el subíndice numérico de la variable exógena o independiente). La letra  $\beta_{ij}$  representa el efecto entre dos variables endógenas o dependientes (i hace referencia al subíndice numérico de la variable endógena (efecto) y j hace referencia al subíndice numérico de la variable causa. Siguiendo con el ejemplo del diagrama de flujo No. 2 y 3, se puede observar la posición de los símbolos griegos para comprender mejor los diagramas de causa-efecto.

Diagrama de flujo No. 2.1, con efecto o relación indirecta y la posición de las letras griegas.

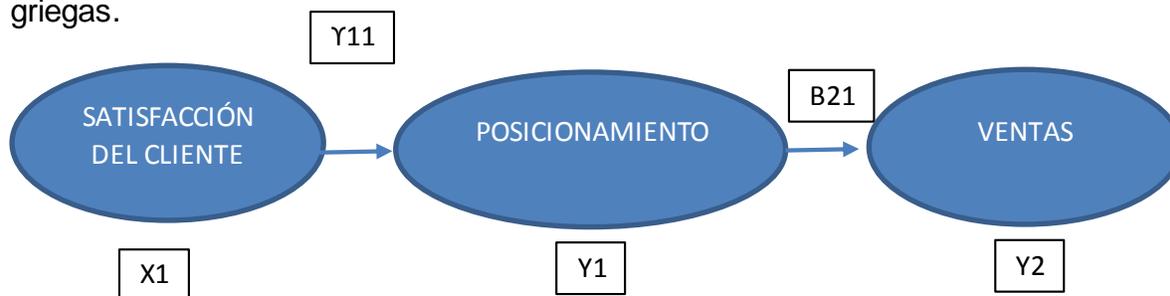
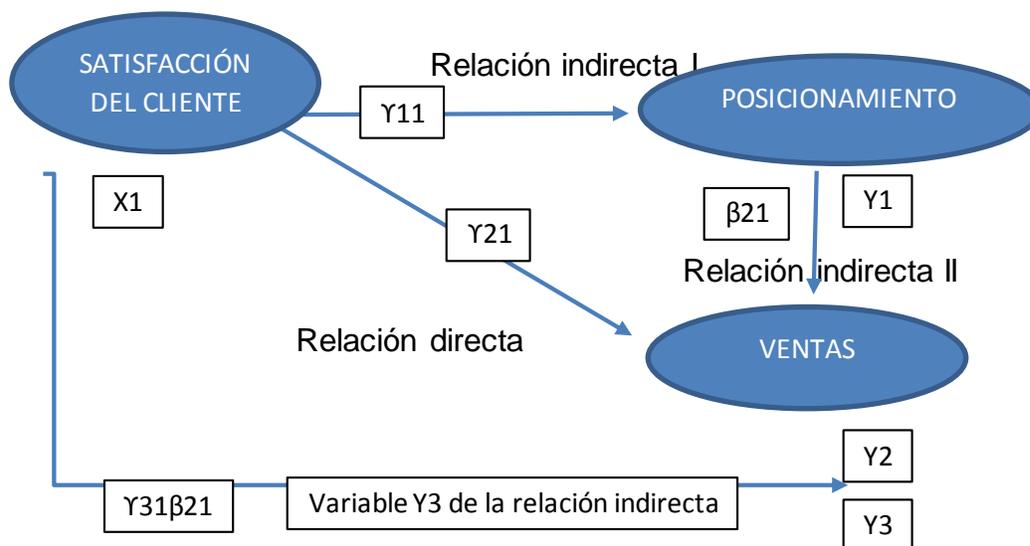


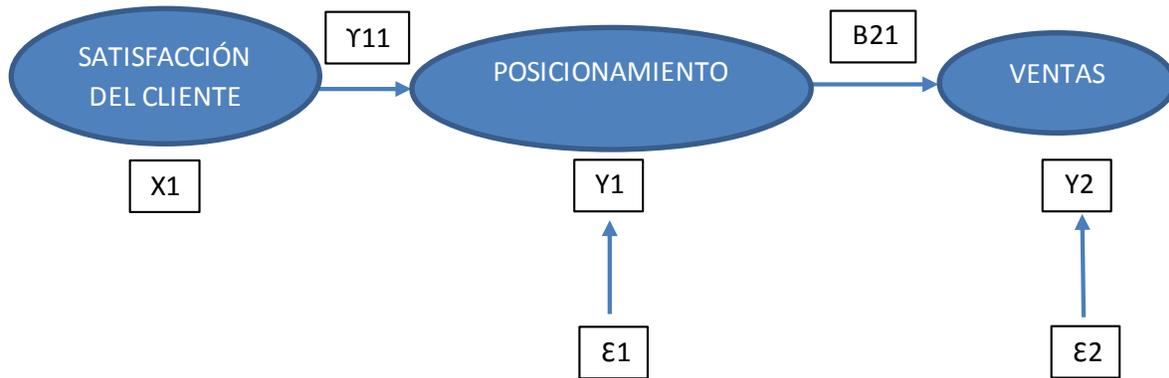
Diagrama de flujo No. 3.1 con efecto o relación directa e indirecta y la posición de las letras griegas



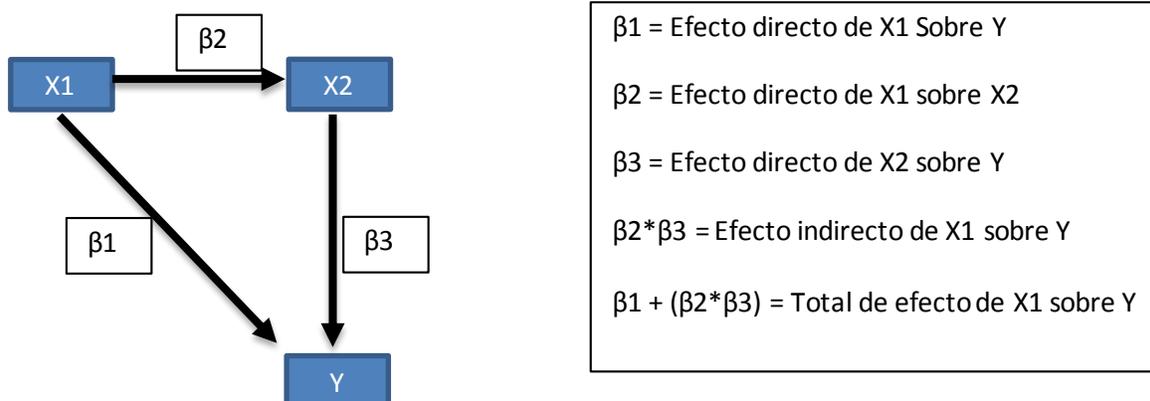
Representación de errores:

Debido a que las variables no se explican en su totalidad, se debe registrar los errores en función de las variables endógenas y se representa con la letra griega  $\epsilon$ , acompañada del mismo subíndice numérico que la variable endógena y unida a la variable endógena con una flecha indicando el efecto del error. Siguiendo con el ejemplo del diagrama de flujo No. 2, se puede ver así:

Diagrama de flujo No. 2.2, con efecto o relación indirecta y la posición de las letras griegas y el error.



Es importante conocer la interacción de los efectos directos e indirectos y a suma total de los efectos dentro del diagrama de interacción de las variables, por ejemplo, al tener dos variables independientes (X1 y X2) y solo una dependiente (Y), de la siguiente manera:



Interpretación y estadísticos de bondad de ajuste:

La ventaja que se tiene con los SEM, es que se analiza el comportamiento de cada una de las variables contra las demás con sus respectivas hipótesis en el mismo tiempo. Ver si las varianzas y covarianzas se ajustan al modelo o a la estructura propuesta. Analizar los coeficientes de relación y los coeficientes de causalidad indicando el porcentaje de variación de cada variable exógena sobre la endógena.

Los estadísticos de bondad de ajuste más utilizados son:

- Estadísticos de ajuste absoluto: la Chi Cuadrada con valores de referencia en significancia mayores de .05; si los valores de Chi Cuadrada/gl son de 2 a 5, el modelo se puede considerar con un ajuste aceptable. Una medida alternativa de la Chi Cuadrada para muestras pequeñas es el parámetro de no centralidad (NCP); si los valores son menores a 2 se considera un modelo aceptable.
- Ajuste comparativo: índice de bondad de ajuste comparativo (CFI), debe ser  $\geq 0.95$  para considerar el modelo como aceptable.
- Índice no normalizado de ajuste (NNFI) o Índice de Tucker-Lewis (TLI), supera las limitaciones del NFI; los valores  $\geq .95$ , consideran el modelo como aceptable.
- El índice de ajuste normalizado (NFI), compara el modelo propuesto y el modelo nulo; valores  $\geq .95$ , son considerados como aceptables.

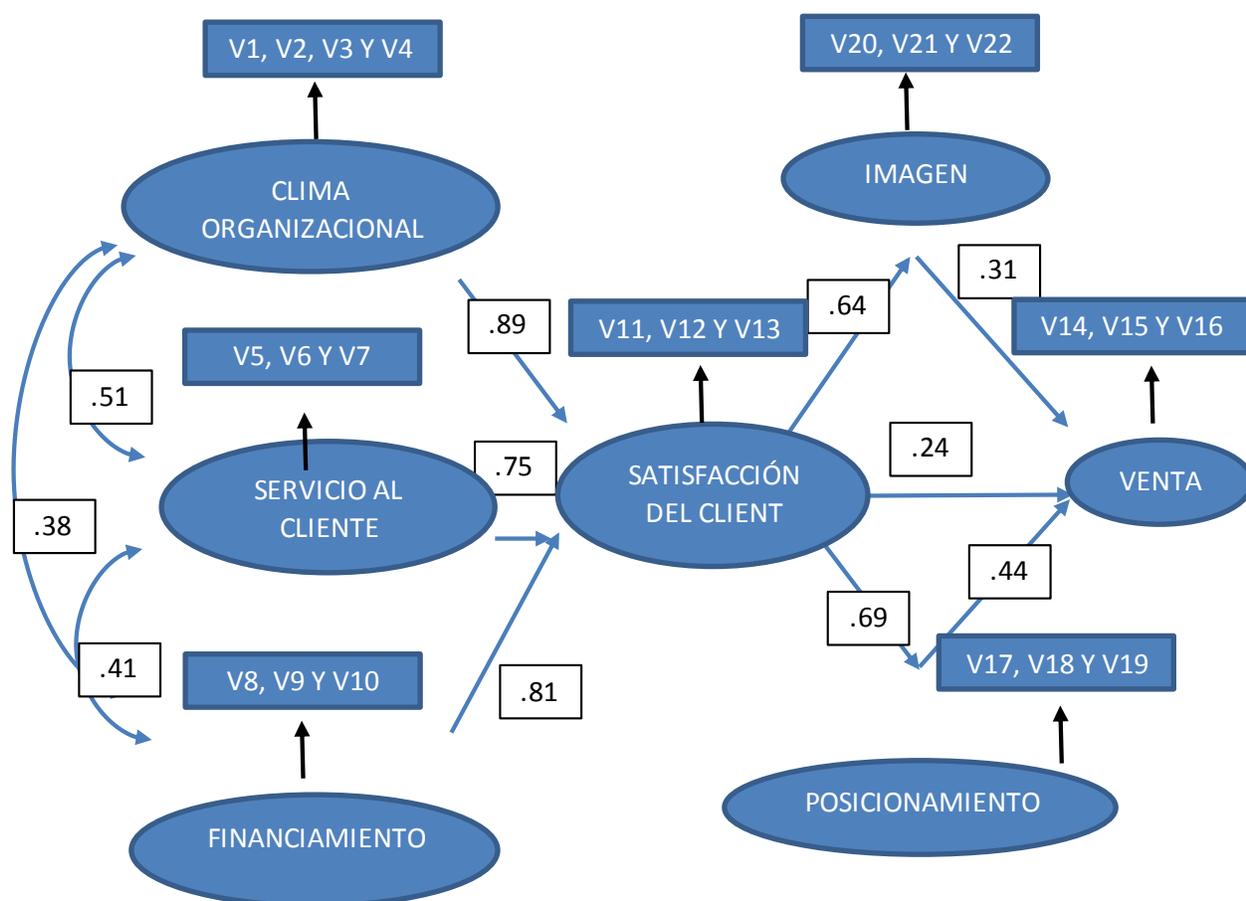
- El índice de validación cruzada esperada (ECVI) representa la correlación entre las variables del modelo, correlaciones cercanas a 1 significan correlación fuerte entre las variables.
- Ajuste parsimonioso, relacionan la calidad del ajuste del modelo con el número de coeficientes necesarios para conseguir el nivel de ajuste, los valores deben ser próximos a 1.
- El Índice de ajuste normado de parsimonia (PNFI) relaciona los constructos con la teoría, entre más cercano a 1 indica mayor relación y aceptabilidad del modelo.
- El criterio de información de Akaike (AIC), es una medida comparativa entre modelos con diferentes números de constructos. Los valores cercanos a cero, indican un mejor ajuste del modelo y mayor parsimonia.
- Índice de bondad de ajuste (GFI), para determinar si el modelo debe ser ajustado. Debe ser  $\geq .95$ . Entre más se acerque a cero se considera un modelo con mal ajuste; Índice de bondad de ajuste corregido (AGFI), es una extensión del GFI, el cual ajusta los grados de libertad entre los modelos y debe ser  $\geq .95$  para considerarse como modelo ajustado; Raíz de residuo cuadrático promedio (RMR), mide las varianzas y covarianzas de la muestra debe estar próximo a cero para considerarse un ajuste óptimo del modelo; y la Raíz del residuo cuadrático promedio de aproximación (RMSEA), el cual debe ser  $\leq 0.05$ , que representa el ajuste anticipado con el valor total de la población o indica un error de aproximación del modelo con la realidad. Ver tabla No. 6

Tabla No. 6

MEDIDAS	MEDIDAS DE AJUSTE ABSOLUTO
Estadístico $\chi^2$ Discrepancia entre $\chi^2$ y grados de libertad (CMIN/DF).	MEDIDAS DE AJUSTE ABSOLUTO: comprobar la significancia del test. Favorable si P-valor es $\geq .055$ . CMIN/DF Entre 1 a 5.
GFI (Goodness of Fit Index-Índice de bondad de ajuste).	MEDIDAS DE AJUSTE ABSOLUTO: valores superiores a .90
RMSEA (Root mean square error of approximation $\chi^2/DF$ -Error cuadrático medio de aproximación).	MEDIDAS DE AJUSTE ABSOLUTO: valore inferiores a .05/.08. Promedio de los residuos entre la correlación/covarianza.
AGFI (Adjusted GFI-Índice de ajuste ponderado).	MEDIDAS INCREMENTALES DE AJUSTE: valores superiores a .90
CFI (Comparative Fit Index-Índice de ajuste comparativo). NFI (Normed Fit Index-Índice de ajuste normalizado). IFI (Incremental Fit Index-Índice de ajuste incremental). RFI	MEDIDAS INCREMENTALES DE AJUSTE: valores mayores a .90
PARSIMONIOSAS: PNFI (Parsimony Normed FI- Parsimonia normalizada del índice de ajuste) PCFI (Parsimony Comparative FI- Parsimonia del índice comparativo)	P > .5

Ejemplo completo para la interpretación:

Diagrama de flujo No.5



Se puede observar que las variables V1, V2, V3 y V4 miden el Constructo de Clima Organizacional, las V5, V6 y V7 miden el constructo Servicio al Cliente y las V8, V9 y V10, miden el constructo Financiamiento. La lectura se recomienda hacer de izquierda a derecha del flujograma con los valores de las covarianzas entre las variables clima organizacional y servicio al cliente de .51, entre el servicio al cliente y financiamiento de .41 y entre el clima organizacional y el financiamiento de .38. El efecto del Clima organizacional sobre la satisfacción es de .89, del servicio al cliente sobre la satisfacción del cliente de .75 y del financiamiento sobre la satisfacción del cliente de .81. Las variables V11, V12 y V13, miden el constructo Satisfacción del Cliente; las variables V20, V21 y V22 miden el constructo imagen; las variables V14, V15 y V16, miden las ventas; las variables V17, V18 y V19, miden el constructo posicionamiento. Las 7 hipótesis se deben analizar con la interacción entre la satisfacción y la imagen (H1), la satisfacción y las ventas (H2), la satisfacción y el

posicionamiento (H3), la imagen y las ventas (H4), el posicionamiento y las ventas (H5) y la interacción indirecta entre la satisfacción modulada por la imagen (H6) y por el posicionamiento (H7).

## **Conclusión**

El análisis multivariante permite analizar varias variables en el mismo momento y a generar modelos de predicción, partiendo de la pregunta si hay que realizar pruebas de dependencia o de causa y efecto como el ANOVA, la regresión múltiple y logística, MANOVA, MANCOVA, análisis discriminante y ecuaciones estructurales. Antes de aplicar en análisis multivariante se debe cumplir con tres requisitos importantes para estas pruebas: el de normalidad para cada una de las variables cuantitativas o superiores involucradas en el modelo, el de linealidad para ver el grado de correlación entre las variables independientes y dependientes y el análisis de homocedasticidad para conocer la homogeneidad de la varianza de cada variable cuantitativa o superior involucrada en el modelo multivariable.

Para dar respuesta a los análisis de agrupación de variables se procede a realizar los análisis de clúster para agrupación de individuos en segmentos similares, análisis discriminante y análisis de conjunto. Para el análisis de reducción de variables se realizan las pruebas de componentes principales, análisis de correspondencia y análisis de factores para agrupar y reducir variables medidas en escalas superiores o cuantitativas. Los análisis descriptivos e inferenciales se realizan de manera simultánea en este tipo de pruebas. Tiene como requisitos al realizar este tipo de pruebas que cada una de las variables cumplan con la distribución normal. En ocasiones la relación lineal simple o múltiple se ve limitada, por lo que se recomienda combinar el análisis de factores con el de regresión, para buscar relaciones multicausales, de ahí que los modelos de ecuaciones estructurales viene a facilitar esas relaciones multicausales para acercarse a la realidad de la mayoría de los comportamientos multicausales de los individuos.

**Palabras clave**

Regresión simple y múltiple

Análisis de factores

Análisis discriminante

ANOVA

MANOVA

ANCOVA

MANCOVA

ECUACIONES ESTRUCTURALES

**Preguntas**

¿Cuáles serían los modelos de dependencia entre variables?

¿Para qué sirven los modelos de dependencia e interdependencia?

¿Cuáles son los modelos multivariantes de interdependencia?

¿En qué escala de medición deben estar las variables para realizar una prueba de regresión simple?

¿Qué escala de medición debe ser la variable dependiente e independiente en la prueba de ANOVA?

¿Cuándo se debe aplicar el ANOVA, MANOVA, ANCOVA y MANCOVA?

¿Cuándo se debe aplicar el análisis discriminante? ¿Cuáles son las escalas de medición en las variables para hacer análisis discriminante?

¿En qué casos se debe aplicar el análisis de factores?

¿Qué tipo de escalas de medición deben ser las variables en las ecuaciones estructurales?

¿En qué casos se debe aplicar el modelo de ecuaciones estructurales?

¿Cuáles serían las variables a considerar en una ecuación estructural?

## REFERENCIA BIBLIOGRÁFICA

- Arbuckle, J. L. (1997). Amos Users' Guide. Version 3.6. Chicago: SmallWaters Corporation.
- Bagozzi, R. O. y Yi, Y. (1994). Advanced Topics in Structural Equation Models. In: R. P. Bagozzi (Ed.). Advanced Methods of Marketing Research. Cambridge: Blackell Publishers.
- Bentler, P. M. (1985). Theory and implementation of EQS: A structural equations program. Los Angeles: BMDP Statistical Software
- Byrne, B. M. (2001). Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming. Mahwah, NJ: Lawrence Erlbaum Associates.
- Castañeda, M. B. (2010). *Procesamiento de datos y análisis estadísticos utilizando SPSS: Un libro práctico para investigadores y administradores educativos*. EDIPUCRS.
- Chalmers, Alan. (2010) *¿Qué es esa cosa llamada ciencia?*, Editorial Siglo XXI, México, pp17
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3), 297-334.
- Duncan, O. D. (1966). Path analysis: Sociological examples. *American Journal of Sociology*, 72: 1-12.
- Elorza H. (1999). *Estadística para las ciencias sociales y del comportamiento*, segunda edición, Oxford University Prees México, ISBN: 970-613-181-7, México.
- Freedman D., Pisani R., Purves R. (1991). *Estadística*, Segunda edición, Antoni Bosch, Editores, ISBN: 84-85855-68-X, Barcelona.
- Hair J., Anderson R., Tatham R., Black W. (1999). *Análisis multivariante*, 5 edición, Prentice Hall, Iberia, ISBN: 84-8322-035-0, Madrid
- Hayduk, L. (1987). *Structural equation modeling with LISREL Essentials and Advances*. Baltimore: The Johns Hopkins University Press.

- Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system, pp. 85-112 in A. S. Goldberger and O. D. Duncan (eds.) Structural Equation Models in the Social Sciences. New York: Seminar.
- Jöreskog, K. G. y Sörbom, D. (1986). LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods. Mooresville, IN: Scientific Software, Inc
- Jöreskog K. G., Sörbom D. (2008). LISREL 9.3. Scientific Software International.
- Jöreskog K. G. (1969). General approach to confirmatory maximum likelihood factor analysis. Psychometrika. 34:183-202
- Kinneer Thomasy. (1993) Investigación de Mercados, 4 Edición, McGraw-Hill, pp. 12-13.
- Levin Jack (1979). Fundamentos de Estadística en la Investigación Social, Editorial Harla, ISBN 0063150123, México, pp. 202-205
- Manual del usuario del sistema básico de IBM SPSS Statistics 25
- Montanero F. J.(2008). Análisis multivariante, Universidad de Extremadura, ISBN: 978-84-691-6343-6, España
- Pérez L. César. (2001). Técnicas estadísticas con SPSS, Editorial Prentice Hall, ISBN: 84-205-3167-7, Madrid.
- Pérez L. Cesar. (2004). Técnicas de análisis multivariante de datos, Pearson Educación, ISBN: 978-84-205-4104-4, España.
- Pérez L. (2004). Marketing Social, Teoría y Práctica, Pearson Educación, ISBN: 970-26-0541-5, México, 2004, pp. 211-234
- RUIZ CAMACHO M., MORCILLO AIXELÁ M.C., GARCÍA GALISTEO J., CASTILLO VÁZQUEZ C. (2000). "Curso de Probabilidad y Estadística", Ed. Universidad de Málaga / Manuales, (Teoría-Problemas).
- SARABIA VIEJO A., MATE JIMÉNEZ C. (1993): "Problemas de Probabilidad y Estadística. Elementos teóricos, cuestiones, aplicaciones con Statgraphics", Ed. CLAGSA,(Problemas).
- Schiffman, Wisenblit. (2015). Comportamiento del consumidor, 11 Edición, Pearson Educación, México.

## ANEXO 1

Ejemplo del modelo de investigación en una página para un proyecto real:

Planteamiento	Variabes	Definir Variables	Mensurabilidad	Hipótesis	Metodología	Muestra	Diseño Estadístico
<p>Idea de la investigación. Conocer la preferencia por consumo de bebidas alcohólicas en los restaurantes.</p> <p>Tema de la investigación. Conocer la preferencia por consumo de bebidas alcohólicas en los restaurantes de nivel medio, medio alto y alto de los municipios de Monterrey y San Pedro Garza García.</p> <p>Planteamiento del problema ¿Cuáles variables de la mezcla de mercadotecnia como producto (whisky, vodka, tequila, ron y brandy), precio, promoción y distribución; así como variables propias de los individuos como edad, sexo, nivel de ingresos, escolaridad y estado civil,</p>	<p>Variabes Dependientes</p> <p>Variable Dependiente (1) Preferencia</p> <p>Variabes independientes</p> <p>1. Mezcla de mercadotecnia.</p> <ul style="list-style-type: none"> <li>• Precio</li> <li>• Producto</li> <li>• Promoción</li> <li>• Distribución</li> </ul> <p>2. Variables propias de los individuos:</p> <ul style="list-style-type: none"> <li>➢ Escolaridad</li> <li>➢ Género</li> <li>➢ Edad</li> <li>➢ Ingresos</li> <li>➢ Estado civil</li> </ul>	<p><b><u>Variabes Dependiente</u></b></p> <p><i>Preferencia</i></p> <p>Conocer las bebidas alcohólicas que el público consume mayormente.</p> <p><b><u>Variabes independentes</u></b></p> <p>1. Mezcla de mercadotecnia.</p> <p>Comparar la marca Bacardí en cuanto a producto, precio, promoción y distribución.</p> <p>2. Variables propias de los individuos:</p> <ul style="list-style-type: none"> <li>➢ escolaridad</li> <li>Grado de estudios del mercado meta.</li> </ul>	<p><b><u>Escalas de medición de las variables dependiente</u></b></p> <p>Preferencia de bebidas en los restaurantes:</p> <p>Escala intervalar</p> <p><b><u>Escalas de medición de las variables independentes</u></b></p> <p>1. Mezcla de mercadotecnia.</p> <p>Escala intervalar</p> <p>2. Variables propias de los individuos:</p> <ul style="list-style-type: none"> <li>➢ escolaridad</li> <li>Escala nominal</li> <li>➢ género</li> <li>Escala nominal</li> </ul>	<p>Ho1= La edad no guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ha= La edad si guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ho2= El sexo no guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ha= El sexo si guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ho3= La evaluación de la preferencia de las bebidas alcohólicas es igual para los diferentes niveles de ingreso.</p>	<p>Entrevistas administradas a los consumidores en restaurantes de clase media, media alta y alta de Monterrey y San Pedro Garza García.</p> <p>Cuestionario:</p> <p>1. ¿Cón qué frecuencia consume algunas de estas bebidas cuando acude a este restaurante (inclusive las bebidas preparadas con estos licores)? (en escala de Licker de 1=nada frecuente a 7= Muy frecuente)</p> <ul style="list-style-type: none"> <li>• Whisky</li> <li>• Vodka</li> <li>• Tequila</li> <li>• Ron</li> <li>• Brandy</li> </ul> <p>2. ¿Tiene usted alguna marca de bebida alcohólica de mayor preferencia?</p> <p>1.- Sí ¿Cuál?</p> <p>2.- No</p> <p>3. Evalúe las siguientes afirmaciones de su marca preferida en una escala</p>	<p>Según los datos de la unidad de análisis en este proyecto de investigación se utilizó la fórmula para seleccionar la muestra de población finita, ya que en total hay 80 restaurantes dirigidos a los estratos medios y altos en los dos Municipios seleccionados de la Ciudad de Monterrey: Monterrey y Garza García, con un total de 480,000 clientes que los visitan al mes.</p> <p><math>n = p(1-p) / E^2/Z^2 + P(1-P)/N</math></p> <p>P= .5</p> <p>E= 0.05</p> <p>Z= 1.96</p>	<p>Se realizará análisis univariado descriptivo e inferencial con Frecuencia, porcentaje, gráficas y razones para las variables nominales, se calculará la media, mediana, moda, desviación estándar y rango en variables escalares o de razón y para inferir, se utilizará la Chi Cuadrada y la T de Student para una sola muestra. Análisis bivariado con la prueba paramétricas como la T de Student, el ANOVA, ANCOVA, MANOVA Y MANCOVA, análisis de correlación y regresión simple. Análisis multivariado como la regresión múltiple, análisis discriminante y análisis de factores.</p>

<p>influyen en la preferencia por bebidas alcohólicas en los restaurantes de nivel medio, medio alto y alto de los municipios de Monterrey y San Pedro Garza García durante el primer trimestre del 2017?</p> <p><b>Unidad de Análisis:</b> personas adultas de 20 a 60 años de edad, guardando la proporción de sexo y la proporción del estrato socioeconómico: 60% de la clase media, 25% de la media alta y 15% de la clase alta, con nivel de escolaridad de preparatoria para arriba, 70% casados y 30% solteros y que ganen más de 5 salarios mínimos al mes.</p> <p><b>Variable dependiente:</b> Preferencia de bebidas alcohólicas</p> <p><b>Variables Independientes:</b> Propias de la mezcla de la mercadotecnia, como producto,</p>		<p>➤ Género Característica propia del individuo.</p> <p>➤ edad Tiempo transcurrido desde el nacimiento hasta la actualidad.</p> <p>➤ Ingresos en salarios mínimos al mes.</p> <p>➤ Estado civil: afiliación jurídica y social clasificada en soltero, casado, divorciado y viudo.</p>	<p>➤ edad Escala intervalar</p> <p>➤ ingresos Escala intervalar</p> <p>Estado civil en escala nominal</p>	<p>Ha= La evaluación de la preferencia de las bebidas alcohólicas es diferente para los diferentes niveles de ingreso.</p> <p>Ho4= La escolaridad no guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ha= La escolaridad si guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ho5= El estado civil no guarda relación significativa con el grado de preferencia por bebidas alcohólicas.</p> <p>Ha= El estado civil si guarda relación significativa con el grado de preferencia por bebidas alcohólicas.</p> <p>Ho6= La promoción no guarda relación significativa con la</p>	<p>de 1 a 5 siendo 1 totalmente en desacuerdo y 5 totalmente de acuerdo.</p> <ul style="list-style-type: none"> <li>• Por su precio</li> <li>• Por su sabor</li> <li>• Porque se combina mejor</li> <li>• Porque la ofrece el establecimiento</li> <li>• Por su publicidad</li> </ul> <p>4. ¿Ha cambiado su preferencia en cuanto a marca de bebida?</p> <ol style="list-style-type: none"> <li>1. Sí ¿Cuál y por qué?</li> <li>2. No</li> </ol> <p>5. ¿Cuál es su edad?</p> <ul style="list-style-type: none"> <li>• 18-30</li> <li>• 31-40</li> <li>• 41-50</li> <li>• 50-60</li> <li>• 60- más</li> </ul> <p>6. ¿Cuál es su grado académico?</p> <ul style="list-style-type: none"> <li>• Preparatoria o menos</li> <li>• Licenciatura</li> <li>• Postgrado</li> </ul> <p>7. Género</p> <p>Femenino</p> <p>Masculino</p> <p>8. ¿Cuál es su ingreso promedio mensual?</p> <ol style="list-style-type: none"> <li>1. 5 salarios mínimos por mes o menos</li> <li>2. De 6 a 10</li> <li>3. De 11 a 15</li> <li>4. Más de 16</li> </ol>	<p>N= 480,000</p> <p>n= 384</p> <p><b>Procedimiento para selección de la muestra:</b> Probabilístico o respetando la proporción en estado civil, sexo e ingresos, que ganen arriba de 5 salarios mínimos al mes y con nivel de escolaridad de preparatoria para arriba.</p> <p><b>Trabajo de campo:</b> Supervisión del cien por ciento de las encuestas aplicadas a los clientes dentro del restaurante bajo la modalidad del proceso probabilístico sistematizado: Uno de cada tres clientes que entren al restaurante durante el horario de las 4 de la tarde</p>	<p>Para la hipótesis nula (Ho):</p> <p>Ho1 hasta la Ho9, se aplicará la Chi Cuadrada para inferir.</p> <p>Multivariada se aplicará el análisis discriminante para la variable dependiente preferencia y todas las variables medidas en escala Likert.</p>
--	--	---	---	---	--	--	---

<p>precio, plaza y promoción.</p> <p>Propias de los individuos como:</p> <ul style="list-style-type: none"> <li>➤ Escolaridad</li> <li>➤ Género</li> <li>➤ Edad</li> <li>➤ Ingresos</li> </ul> <p>Estado civil</p> <p><b>Tiempo y espacio:</b></p> <p>Durante el primer trimestre del 2017 del Municipio de Monterrey y de San Pedro Garza García, de Ciudad de Monterrey, México.</p> <p><b>Según el tiempo:</b></p> <p>Es un estudio transversal con el apoyo de un cuestionario estructurado</p>				<p>preferencia por bebidas alcohólicas.</p> <p>Ha= La promoción si guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ho7= El precio no guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ha= El precio si guarda relación significativa con la preferencia por bebidas alcohólicas.</p> <p>Ho8= El producto no guarda relación significativa con el grado de preferencia por bebidas alcohólicas.</p> <p>Ha= El producto si guarda relación significativa con el grado de preferencia por bebidas alcohólicas.</p> <p>Ho9= La distribución no</p>	<p>9. ¿Cual es su estado civil?</p> <ol style="list-style-type: none"> <li>1. Soltero</li> <li>2. Casado</li> <li>3. Divorciado</li> <li>4. Viudo.</li> </ol>	<p>hasta las 11:30 PM, por 5 días hábiles de la semana.</p>	
---	--	--	--	---	---	---	--

				guarda relación significativa con el grado de preferencia por bebidas alcohólicas.  Ha= La distribución si guarda relación significativa con el grado de preferencia por bebidas alcohólicas.			
--	--	--	--	---	--	--	--

**Conclusión: Se rechazan las hipótesis nulas de la mezcla del marketing por lo que se sugiere tener dinamismo de las 4 Ps del marketing para mejorar la experiencia de compra de sus clientes.**

## ANEXO 2

Tabla A: Distribución normal con valores Z

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Fuente: <http://matepedia-estadistica.blogspot.com/2016/01/tabla-de-distribucion-normal.html>