UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

HERRAMIENTAS ESTADÍSTICAS APLICADAS A LA INVESTIGACIÓN DE MERCADOS PARA FORTALECER LA TOMA DE DECISIONES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

LILIA CRUZ HERNÁNDEZ

DR. JUAN GONZÁLEZ HERNÁNDEZ

2008

A mis padres,
Angela y Mauro

A mis guías
Que siempre están a mi lado

Agradecimientos

A mi Profe, asesor y amigo Juan,
principalmente por su gran paciencia

A mi familia y amigos,
por su apoyo incondicional

A mis sinodales,
Ruth, Francisco, Luis y Lupita,
por su tiempo, comentarios, sugerencias
(cualquier error en este trabajo
es responsabilidad mía)

Índice

Introducción.

La finalidad de un profesional de la Actuaría es estudiar, planear, formular y aplicar modelos de contenido matemático acerca de fenómenos que involucran riesgos para proveer información que nos permita planear, prever y tomar decisiones.

Desde finales del milenio pasado, el desarrollo del sistema empresarial se ha ido modificando como consecuencia de las transformaciones experimentadas en el entorno, lo cual ha creado un nuevo enfoque a lo que en mercadotecnia de refiere con el fin de lograr una supervivencia satisfaciendo las necesidades del consumidor.

Al igual que ha sucedido en el ámbito empresarial, el desarrollo académico se va adecuando a las necesidades del entorno. El propósito de este trabajo es dar un enfoque general de algunos métodos estadísticos que se emplean en la obtención y análisis de la información de mercados como una de las aplicaciones de la Actuaría, que, como inicialmente se mencionó, su finalidad es analizar eventos contingentes y cuantificar sus consecuencias financieras, sociales y económicas mediante (en este caso) herramientas estadísticas.

En las condiciones actuales de mercados abiertos e intercambios para el fortalecimiento de la toma de decisiones para mejorar la competitividad, es importante establecer algunos conceptos para asentar las bases a una mejor comprensión de la importancia y necesidad de las herramientas estadísticas enfocadas a la Investigación de mercados en el México de hoy. Entre los puntos importantes podemos mencionar los mercados competitivos, la importancia de la publicidad, el consumidor hoy y la Estadística.

1. Marco general.

1.1 Mercados competitivos.

En un determinado momento hemos sido testigos de como en una categoría donde participaban tres o cuatro marcas, repentinamente hemos tenido otro tanto de importación, independientemente de si competían o no en precio y sobretodo en calidad, la oferta existía y el consumidor podía comprar. Así, hoy en día, la comparación y la elección es significativamente mayor y esta situación llegó para quedarse.

1.2 Publicidad.

De forma simple, podríamos clasificar dos grandes áreas:

- Lo relacionado con los medios.
- Lo relacionado con los mensajes y la creatividad.

En lo referente a los medios, no se ha dado una significativa mejora en su calidad sino una muy importante implicación y diversidad. Hoy por hoy, los consumidores nos vemos literalmente bombardeados de anuncios publicitarios de todo tipo de productos a cualquier hora y en cualquier lugar.

Por su parte, los mensajes y creatividad también han pasado por una turbulenta etapa, en donde podemos disfrutar desde verdaderas joyas hasta pésimas tropicalizaciones de anuncios que ni siquiera se toman la molestia de realizar los doblajes correctos.

1.3 El consumidor hoy.

En términos generales, el consumidor de hoy es más informado, sus motivaciones obedecen hoy más a juicios que a impulsos, según señalan diversas investigaciones.

"La clave para el desarrollo del hombre de publicidad es escudriñar el interior de la naturaleza humana", dice Bill Bernback. En la medida en que el hombre tiene una gran diversidad de deseos y necesidades de acuerdo con su edad, ingreso, sexo y posición social y económica, tiene también una serie de opciones para elegir satisfactores a esas necesidades, su elección es personal y es libre de elegir entre las opciones del mercado y de acuerdo a sus condiciones particulares. Elegir la opción más adecuada depende, en gran medida, del grado de información que tenga sobre las ventajas, desventajas y diferenciación de cada opción. La información existe aunque no siempre se hace un uso y búsqueda adecuados de la misma.

1.4 Estadística.

Antes de llegar a una definición de Estadística formal, analicemos las definiciones irónicas, ya que nos alerta del mal uso de ésta. Si pocas son las personas que tienen una definición apropiada de Estadística, son muchas las que conocen o intuyen alguna de estas definiciones que A. Piatier denominó humorísticas. Casi todas las definiciones, más que humorísticas, sarcásticas, tienen por núcleo central establecer una relación entre estadística y mentira, a menudo lo que se observa es hacer creer que el promedio se va a usar siempre como estimador de nuestra población objetivo o muestra, evidentemente primero se tiene que establecer que es lo que se quiere medir y en base a

esto usar el estadístico adecuado, que no siempre es la media. Pero la Estadística no es la equivocada, sino la persona misma que la emplea mal y que sin objetivos precisos mete los datos a un paquete de análisis estadístico sin que éstos se relacionen a los métodos estadísticos que se están usando.

Un análisis detallado de estas definiciones puede tener un efecto positivo para aproximarnos al concepto de Estadística. El riesgo es ayudar a difundir opiniones de personas que no tienen "ningún conocimiento del método ni de sus aplicaciones y éxitos", A. Piatier. Se dice, por ejemplo, que si una persona gana un millón y otra nada "la Estadística" establece que las dos han ganado medio millón, se dice también que si una persona pone la cabeza en el congelador y los pies en el horno su temperatura media será correcta o que la Estadística pronostica un acierto para el caso de un soldado que dispara sobre un blanco una vez medio metro a la derecha y otra medio metro a la izquierda. Estos argumentos si no se analizan, parecen suficientes para desvirtuar una disciplina, puesto que si no sabe hacer algo tan sencillo ¿cómo es posible que sea capaz de resolver problemas más complejos? Ejemplos parecidos han proliferado por doquier. Si desgranamos estos, en apariencia sencillos ejemplos, encontraremos algunas de las características que diferencian a la Estadística de su incorrecta aplicación y que nos permiten aproximarnos a una definición de ésta.

¿Qué es la Estadística?

"Ciencia que se ocupa del estudio de fenómenos de tipo genérico, normalmente complejos y enmarcados en un universo variable, mediante el empleo de modelos de reducción de la información y de análisis de validación de los resultados en términos de representatividad". La información puede ser numérica, alfabética o simbólica. El proceso estadístico consiste en las fases de recolección de información, de análisis y de presentación e interpretación de los resultados y elaboración de métodos de Inferencia Estadística. El término Estadística se emplea para referirse a cualquiera de estas fases.
Estadístico: Es el valor de un atributo que se obtiene de una muestra y mediante el cual se infiere o se estima el parámetro poblacional.
Estadístico: También se aplica a la persona que desarrolla o aplica esta ciencia.

Definida así, se establece su carácter genérico y su campo de acción en el estudio de fenómenos complejos ubicados en un universo amplio y variable. Con esta afirmación de complejidad, se introduce el factor de incertidumbre que acompaña a los fenómenos aleatorios pero sin limitar el campo de la Estadística de forma que puede aplicarse también a fenómenos determinísticos. Con la referencia al universo se expresa la relación descrita por D.S. Moore acerca de que los datos estadísticos lo son en un contexto. La definición continúa estableciendo los procedimientos que utiliza, que tienen en común reducir la información. Modelos de este tipo comprenden desde el cálculo de la media aritmética hasta la determinación de complicados modelos de Correlación canónica. El último aspecto que consideramos importante es el de la referencia a los análisis de validez de los resultados en términos de representatividad. Con esta especificación podemos diferenciar lo que es una simple operación aritmética de lo que es una cifra o un estudio estadístico. Como regla general podríamos establecer que un estudio será estadístico cuando a los modelos de reducción empleados le acompañe, o sea posible realizar, un análisis de validez de los resultados obtenidos en términos de representatividad. En cuanto al tipo de información, los datos pueden ser cuantitativos, cualitativos o incluso existe una rama de la Estadística que se ocupa de lo

que se denomina datos simbólicos (por ejemplo, en los accidentes de coche determinar el entorno mediante valoraciones: visibilidad, existencia de árboles en el entorno, curvas, lluvia, velocidad,...etcétera). El resto de la definición aborda cuestiones relacionadas con el uso de la palabra Estadística en el lenguaje.

La Estadística es un medio, no una revelación divina. Utilizada correctamente, ayuda a hacer del mundo un lugar mejor, si se emplea de forma equivocada, hará un lugar peor. Y saber que se utiliza de forma correcta o incorrecta no es cuestión de tener talento para las cifras, sino de poseer inteligencia para hacer un juicio sólido, de manera que, el enfoque pertinente para la Estadística se concentra en el criterio y reduce al mínimo los cálculos.

En la actualidad, una ventaja competitiva la establecen aquellas empresas que brindan productos y servicios competitivos, quienes mediante una buena publicidad llegan a permanecer en la mente del consumidor y para ello es necesario contar con herramientas estadísticas que nos guíen en el camino del nicho/segmento al cual nos queremos dirigir, algunas de las cuales se mencionarán brevemente a continuación.

1.5 Herramientas estadísticas.

Mencionamos brevemente algunas de las herramientas estadísticas de que podemos disponer, las cuales posteriormente desarrollaremos y que nos ayudarán a fortalecer la toma de decisiones**:**

1.5.1 Tabulación cruzada.

Técnica Estadística que describe dos o más variables simultáneamente y produce cuadros en que se muestra la distribución conjunta de dos o más variables que tienen un número limitado de categorías o valores distintos.

1.5.2 Regresión múltiple.

El análisis de Regresión múltiple plantea, en escala de intervalos o nominal, la dependencia de una variable en función de un conjunto de variables independientes.

La Regresión en la Investigación de mercados se puede ver afectada si se limita el uso a variables independientes que se representen en escalas de intervalos. Es aquí cuando se hace uso de la Regresión logística donde se pueden utilizar las variables independientes nominales en términos de Regresión. Este tipo de Regresión convierte las variables nominales en variables binarias que se codifican cero-uno (dummy). Por ejemplo, en el caso del lanzamiento de un nuevo producto se podría medir la respuesta de los agentes con un cero para aquellos que no usaron el producto y un uno para los que si lo usaron.

1.5.3 Coeficientes de Correlación.

Es un índice de asociación entre dos variables, en esencia, es una medida de como se mueven las variables juntas. Podemos mencionar Phi, contingencia, rangos de Spearman y Pearson como típicos ejemplos de coeficientes de Correlación.

Un coeficiente de Correlación mide la fuerza de las relaciones entre dos variables aleatorias. Para variables ordinales o métricas, especifica si la Correlación es positiva o negativa. Para datos métricos provee una medida de la suma de la varianza compartida. Podría sugerir pero no implicar necesariamente un enlace causal entre las variables.

1.5.4 Análisis Discriminante.

Muchas situaciones en Investigación de mercados giran en torno a dos diferentes grupos de consumidores. Por ejemplo, con frecuencia existe preocupación con las diferencias entre los usuarios y los no usuarios de un artículo o una marca dada. En tales situaciones, regularmente existe interés en identificar las características (es decir, edad, ingresos, educación, etc.) de usuarios contra no usuarios del producto. Una técnica para analizar qué carácterísticas "discriminan" a los miembros de los dos grupos y su importancia relativa es el análisis Discriminante.

1.5.5 Análisis Conjunto.

Su principal aplicación en la mercadotecnia es medir los intercambios que los consumidores realizan en los atributos de los productos. El análisis Conjunto principia con un orden de rangos de las preferencias de productos y luego calcula los valores de utilidad para las características centrales que describen al producto. Lo que pretende es encontrar un conjunto de utilidades que expliquen el orden en que se clasificaron los productos.

La aplicación más común es en el lanzamiento de nuevos productos complejos mientras que mantiene el contexto muy realista en cuanto a las decisiones para el que responde.

1.5.6 Análisis de Factores (Factor analysis).

El análisis de Factores incluye variaciones tales como análisis de componentes y análisis de factores comunes, es una herramienta estadística que pude ser usada para analizar interrelaciones a través de un gran número de variables en términos de sus dimensiones en común (factores). Este análisis trata de encontrar una manera de condensar la información contenida en un número de variables originales dentro de un pequeño conjunto de dimensiones (factores) con una mínima pérdida de información.

Entre sus aplicaciones se encuentran la reducción de datos, la identificación de estructuras y la transformación de datos. Ha sido usado en la mercadotecnia para el desarrollo de escalas de personalidad, identificación de atributos de productos clave, segmentos de mercado basados en datos psicográficos y similitud entre productos.

1.5.7 Análisis de Correspondencia.

Es una técnica gráfica para la exploración y búsqueda de afinidad de datos en tablas de contingencia y categorías de datos multivariados, esto a través de la reducción de dimensiones en un contexto de tablas de contingencia representando gráficamente el patrón de asociación entre dos variables o más variables con escala nominal. Se inicia a partir de una tabla de contingencia donde cada renglón y cada columna definen un perfil. Compara los perfiles de los renglones y las columnas por separado para

colocarlos en un punto de la gráfica de tal manera que perfiles semejantes estén asociados a puntos cercanos.

Se utiliza para:

- Crear mapas perceptuales, posicionamiento y seguimiento del producto, así como en estudios de imagen.
- Evaluar publicidad.
- Probar conceptos y nombres de productos.
- Desarrolla un sistema de segmentación (Análisis de Correspondencia Múltiple).

1.5.8 Análisis de conglomerados (Cluster analysis).

El análisis de conglomerados es una técnica que puede ser usada para generar subgrupos significativos de individuos u objetos. Específicamente, el objetivo es clasificar una muestra de entidades (individuos u objetos) en un pequeño número de grupos mutuamente excluyentes basándose en las similitudes entre las entidades. En el análisis de conglomerados los grupos no están predeterminados o definidos a priori, en lugar de ello, esta técnica es utilizada para identificar los grupos.

1.5.9 Escalas multidimensionales.

El objetivo de este análisis es transformar los juicios de similitudes o preferencias de los agentes (por ejemplo, preferencias sobre productos) en distancias representadas en el espacio multidimensional. Si los objetos A y B son juzgados por los que responden como los que son más similares comparados con todos los otros posibles pares de objetos A y B de tal manera que la distancia entre ellos en el espacio multidimensional será menor que la distancia entre otro par de objetos.

Entre las aplicaciones de esta técnica destacan:

- ✓ La identificación de los atributos notables de un producto, percibidos por el agente.
- ✓ Los productos que se consideran como sustitutos y aquellos que se diferencian entre sí.
- ✓ Los segmentos viables que existen en el mercado.
- ✓ Investigación sobre el cambio de marcas.

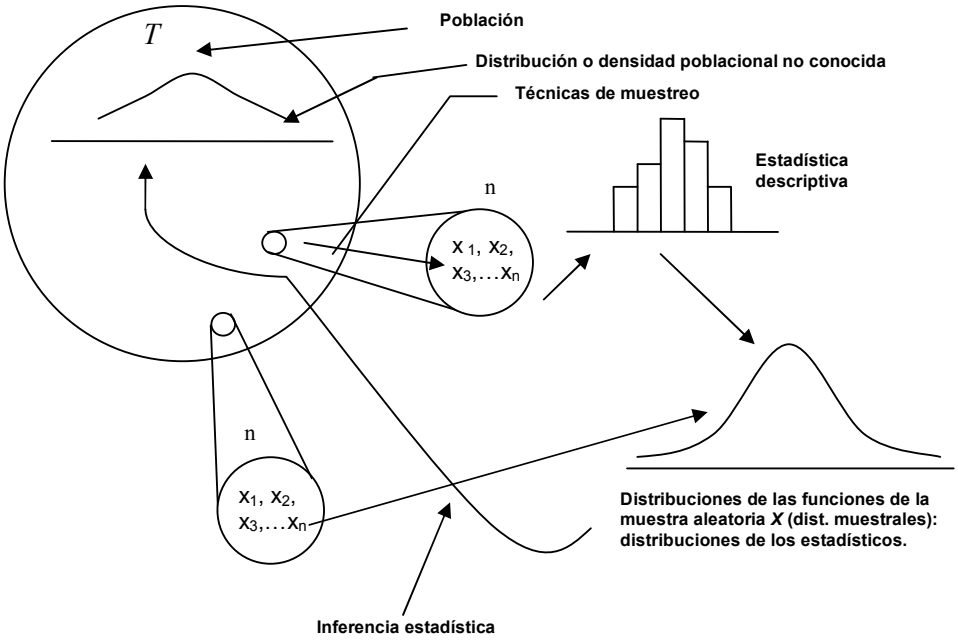
Finalizamos esta introducción con la invitación al lector a una continua actualización y mejora en nuestro cambiante mundo de la mercadotecnia, pero sobre todo, crear conciencia de la importancia de aplicar herramientas estadísticas para fortalecer la toma de decisiones.

2. Conceptos de Estadística.

El aspecto más importante de las Herramientas estadísticas es la obtención de conclusiones basadas en los datos experimentales, este proceso se conoce como Inferencia estadística. Nuestro objetivo es la interpretación de los resultados en base a estimar y predecir algunas características de la población, con base en la información contenida en una muestra para fortalecer la toma de decisión.

Para comprender la naturaleza de la Inferencia estadística es necesario entender los siguientes conceptos:

- **Población**: es un conjunto de individuos sobre los que se estudian una o varias características que son, de alguna forma, observables.
- **Muestra**: es un subconjunto de la población. El número de elementos de la muestra se denomina tamaño muestral.
- **Marco**: Es el listado o instrumento que contiene todas las unidades que integran la población estudiada, el cual sirve de soporte para la extracción de la muestra.
- **Parámetro**: es cualquier característica medible de la función de distribución de la variable en estudio (media, varianza,… etc.).
- **Estadístico**: es una función de la muestra *T* (una muestra aleatoria dentro de la población), se puede conocer su comportamiento con la ayuda de funciones. A las funciones de la muestra aleatoria se les conoce como estadísticos. $Estadístico = F(x)$ Y estos estadísticos, por ser funciones de la variable aleatoria, se comportan a su vez como variables aleatorias; por lo tanto tienen cada uno una distribución llamada distribución muestral, una media, una varianza, etc.



Obtenida la muestra, la información más completa respecto a posibles sucesos está contenida en una distribución probabilística completa o función de densidad. Puesto que esto puede ser muy general, se calculan medidas que resumen la información contenida en la distribución. El primer grupo de medidas es el descriptivo.

2.1 Estadística descriptiva.

Lo primero que hacemos al contar con los datos, es ordenarlos. La primera información (inmediata) lo constituyen las estadísticas descriptivas de los datos, en particular estas son valiosas cuando se tienen a la mano grandes conjuntos de datos.

2.1.1 Frecuencia.

Una descripción informativa de cualquier conjunto de datos está dada por la frecuencia de repetición o arreglo distribucional de las observaciones de un conjunto.

Ejemplo. Supongamos que se hizo la siguiente pregunta a 500 amas de casa:

¿Qué porcentaje de las compras de su hogar realiza personalmente?
_____ Ninguna
_____ Menos de la mitad
_____ Aproximadamente la mitad
_____ La mayor parte de ellas
_____ Todas las compras

Los resultados tabulados toman la forma de la tabla (*Tabla 2.1*):

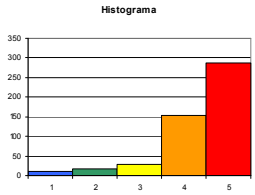
Respuesta (Clases)	***Frecuencia***	***Frecuencia Relativa***
1. Ninguna	8	1.6
2. Menos de la mitad	18	3.6
3. Aproximadamente la mitad	29	5.8
4. La mayor parte	154	30.8
5. Toda	287	57.4
6. Sin respuesta	4	0.8
Total	**500**	**100**

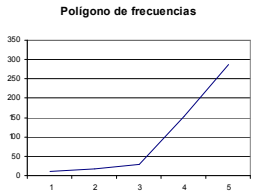
La agrupación de observaciones que no se superpongan entre sí recibe el nombre de *clase*, el número de observaciones en una clase recibe el nombre de *frecuencia de clase*, mientras que el cociente de una frecuencia de clase con respecto al número total de observaciones se conoce como la *frecuencia relativa* de la clase. Si existe una cantidad sustancial de datos, el número de clases deberá encontrarse entre ocho y doce y generalmente no existirán más de quince clases, debe considerarse un número que no sea demasiado pequeño que desdibuje la información, ni demasiado alto que no justifique la agrupación de los datos. Una buena práctica es la creación de clases que tengan una longitud igual, aunque existen algunas excepciones.

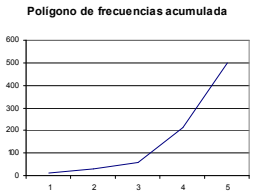
2.1.2 Gráficos.

Los datos se pueden representar gráficamente por medio de histogramas, polígonos de frecuencias, de manera circular, etc.

Por lo general las frecuencias se representan por medio de histogramas, y las frecuencias relativas y frecuencias relativas acumuladas por medio de polígonos.







2.2 Medidas de tendencia central.

Estas medidas localizan el centro de la distribución de mediciones. A continuación se citan las más conocidas:

2.2.1 Moda.

La moda de un conjunto de observaciones es el valor de la observación que ocurre con mayor frecuencia en el conjunto.

En la tabla 2.1, la moda corresponde a la respuesta No. 5 ya que ahí tenemos el mayor número de observaciones (287 de 500).

2.2.2 Mediana.

La mediana de un conjunto de observaciones es el valor para el cual, cuando todas las observaciones se ordenan de manera creciente, la mitad de éstas es menor que este valor y la otra mitad mayor.

Si el número de observaciones en el conjunto es impar, la mediana es el valor de la observación que se encuentra a la mitad del conjunto ordenado. Si el número es par se considera la mediana como el promedio aritmético de los valores de las observaciones que se encuentran a la mitad del conjunto ordenado.

En el ejemplo, la mediana también es 5.

2.2.3 Media.

La media de las observaciones $x_1, x_2, ..., x_n$ es el promedio aritmético de éstas y se denota por

$$\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$$

La media es una medida apropiada de tendencia central para muchos conjuntos de datos. Sin embargo, no hay que olvidar las críticas por usar este estadístico, dado que cualquier observación en el conjunto se emplea para su cálculo, el valor de la media puede afectarse de manera desproporcionada por la existencia de algunos valores extremos.

Para el ejemplo, la media se calcula utilizando el valor codificado para cada respuesta y ponderando la frecuencia con la cual se da la respuesta.

Para los casos de "No respuesta" podemos:

- Transformar. En este caso, podríamos suponer que la falta de contestación representaría una forma de decir que la pregunta se consideró irrelevante, por consiguiente, se podría agregar a la categoría de "Ninguna".
- Excluir. Podemos simplemente excluir estos casos. Codificando, optando por excluir la no respuesta, tendríamos:

Código	*Frecuencia*	*Código por frecuencia*
1	8	8
2	18	36
3	29	87
4	154	616
5	287	1435
Total	**500**	**2182**

En este caso estamos suponiendo que la distribución de los que no contestaron y de los que si, es la misma para ambos casos el cálculo de la media sería:

1. $\bar{x} = {12(8+4)+36+87+616+1435}/{500} =$ 4.372, clasificando los casos de "No respuesta" con código 1.

2. $\bar{x} = {8+36+87+616+1435}/{500} =$4.364, excluyendo los casos de "No respuesta".

En general, convertir la falta de contestación es preferible que pasarla por alto.

2.3 Medidas de variabilidad.

Son medidas que indican el grado de variabilidad de los datos, permiten identificar que tan dispersos o concentrados se encuentran los datos respecto a una medida de tendencia central. Existen varias medidas de dispersión de las cuales se mencionan las siguientes:

2.3.1 Rango.

El rango mide la dispersión de los datos. Es la diferencia entre el mayor y menor valor de la muestra.

Rango intercuartil es la diferencia entre el percentil 75 y el 25.

2.3.2 Varianza.

La varianza de las observaciones $x_1, x_2, ..., x_n$ es en esencia, el promedio del cuadrado de las distancias entre cada observación y la media del conjunto de observaciones. La varianza se denota por:

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

Que también es igual a:

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

La varianza es una medida razonablemente buena de la variabilidad debido a que si muchas de las diferencias son grandes (o pequeñas) entonces el valor de la varianza s^2 será grande (o pequeño). El valor de la varianza puede sufrir un cambio muy desproporcionado, aún más que la media, por la existencia de algunos valores extremos en el conjunto.

Siguiendo con nuestro ejemplo (y tomando $\bar{x} = 4.372$):

Código	*Frecuencia*	$(x_i - \bar{x})^2$	$fi(x_i - \bar{x})^2$
1	12	11.37	136.44
2	18	5.63	101.27
3	29	1.88	54.59
4	154	0.14	21.31
5	287	0.39	113.19
Total	**500**		**426.81**

Por tanto, la varianza es $426.81 / 499 = 0.85$.

La raíz cuadrada positiva de la varianza recibe el nombre de desviación estándar y se denota por:

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

Considerando el ejemplo, tendríamos: $s = 0.9239$

La varianza y la desviación estándar no son medidas de variabilidad esencialmente distintas, las diferencias son las unidades en las que están expresadas, la varianza está en unidades al cuadrado.

2.4 Estimación.

El estimador es la estadística que proporciona el pronóstico de un parámetro de cierta población. Hay dos tipos de estimación:

- Puntual. Cuando se expresa el parámetro de la población mediante un número o dato.

- Intervalo. Si se presenta la estimación expresada por dos medidas, que delimitan un intervalo que contiene el valor del parámetro poblacional estudiado, llamado intervalo de confianza.

Símbolos para población y muestra.

Variable	Población	Muestra
Media	μ	$\bar{x}$
Proporción	π	p
Varianza	σ^2	s^2
Desviación estándar	σ	s
Tamaño	N	N
Varianza estandarizada	$\frac{x-\mu}{\sigma}$	$\frac{x-\bar{x}}{s}$
Coeficiente de variación	$\frac{\sigma}{\mu}$	$\frac{s}{\bar{x}}$

2.4.1 Intervalo de confianza.

Como las estimaciones de punto rara vez serán iguales a los parámetros que se suponen estiman, por lo general es deseable darnos alguna libertad de acción mediante el uso de estimaciones de intervalo. Una estimación de intervalo de un parámetro θ es un intervalo de la forma $\theta_1 < \theta < \theta_2$ en donde θ_1 y θ_2 dependen del valor que tome el estimador θ en una muestra dada y también en la distribución muestral de θ. Por ejemplo, si se nos pidiera determinar el coeficiente intelectual en promedio de un grupo de estudiantes muy grande sobre la base de la muestra aleatoria, podríamos llegar a la estimación de un intervalo $109 < \mu < 117$ sobre la base de la media de la muestra n = 113 así como información acerca de la distribución de la muestra de *n*.

Este intervalo $\theta_1 < \theta < \theta_2$, determinado en relación con una muestra en particular, recibe el nombre de intervalo de confianza del $(1-\alpha)100\%$. La fracción $(1-\alpha)$ se conoce como coeficiente de confianza o grado de confianza y los extremos θ_1 y θ_2 reciben el nombre de límites de confianza inferior y superior. Por ejemplo, cuando $\alpha = 0.02$ el grado de confianza es 0.98 y obtenemos un intervalo de confianza del 98%.

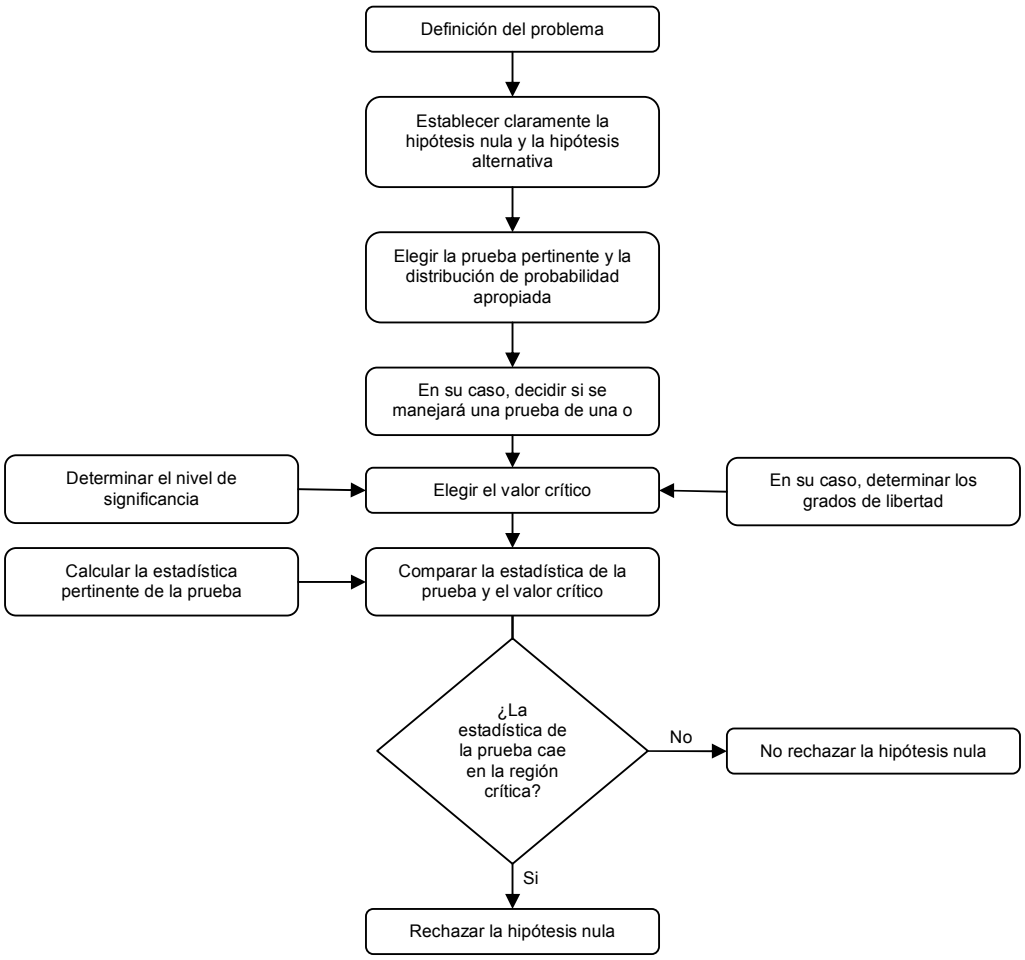
2.5 Prueba de hipótesis.

La prueba de hipótesis es un proceso para determinar si existe una diferencia significativa entre el resultado obtenido y el resultado esperado. El procedimiento que más se utiliza comprende los siguientes pasos:

2.5.1 Formulación de la hipótesis nula (H_0).

Esta es llamada la hipótesis nula, simbolizada por H_0, el subíndice cero implica "cero diferencia". La hipótesis nula es una afirmación que será aceptada si los datos de la muestra no nos proveen de evidencia convincente de que es falsa, es decir, si se

acepta la hipótesis nula decimos que la evidencia no es suficiente para rechazarla pero no podemos afirmar que es verdadera. Si se rechaza, puede aceptarse entonces la hipótesis alterna (H_1). La hipótesis alterna es la aseveración operacional de la hipótesis de investigación del experimentador. La hipótesis de investigación es la predicción que se deriva de la teoría que se está probando.



2.5.2 Especificación del nivel de significancia (α).

Nuestro procedimiento es rechazar H_0 para aceptar H_1, la prueba de significancia estadística busca probar que existe una diferencia real entre dos grupos y además, que esta diferencia no es al azar (para ello usamos la probabilidad, que no es más que el grado de significación estadística, el cual suele representarse con la letra α**)**. Valores comunes de α son 0.05 y 0.01. En otras palabras, si la probabilidad asociada con lo que ocurre en H_0, es decir, cuando la hipótesis de nulidad es verdadera, del valor particular producido por una prueba estadística es igual o menor que α**,** rechazamos H_0 y aceptamos H_1.

Hay dos tipos de errores que pueden cometerse al decidir acerca de H_0, el primero, el *error tipo I* es rechazar H_0 siendo verdadera. El segundo, *el error tipo II*, es aceptar H_0 siendo falsa.

p (error tipo I) = α

p (error tipo II) = β

La *potencia de la prueba* se define como la probabilidad de rechazar H_0 cuando es realmente falsa. Esto es: potencia = 1 – probabilidad del error tipo II = 1 - β.

	Aceptar H_0	**Rechazar H_0**
H_0 es verdadera	Decisión correcta	Error tipo I
H_0 es falsa	Error tipo II	Decisión correcta

No hay un nivel de significancia para todos los estudios, se puede utilizar cualquier valor de probabilidad entre 0 y 1. Tradicionalmente, el nivel de .05 es aplicado a proyectos de investigación, el nivel .01 a control de calidad, y .10 a sondeos políticos.

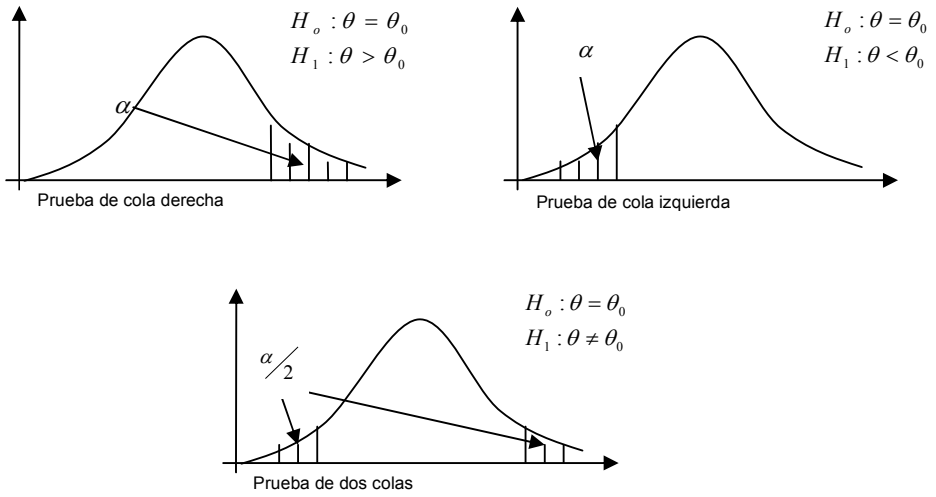
2.5.3 La distribución muestral.

El tipo de distribución se determinará dependiendo de la naturaleza de la hipótesis y del tamaño de la muestra. Cuando la hipótesis es relativa a medias poblacionales y las muestras son grandes (n > 30) se utiliza la distribución Normal. Cuando la hipótesis es relativa a la media y la muestra es chica (n ≤ 30) se utiliza la distribución *t* de Student.

2.5.4 La región de rechazo o región crítica.

La región de rechazo consiste en un conjunto de valores posibles tan extremos que, cuando H_0 es verdadera, es muy pequeña la probabilidad (α) de que la muestra observada produzca un valor que esté entre ellos, La probabilidad *asociada* con cualquier valor de la región de rechazo es igual o menor que α.

La localización de la región de rechazo es afectada por la naturaleza de H_1. Si H_1 indica la dirección predicha de la diferencia, entonces se requiere una prueba de dos colas. Las pruebas de una y de dos colas se distinguen en la localización (pero no en el tamaño) de la región de rechazo. Esto es, en una prueba de una cola, la región de rechazo está totalmente en un extremo (o cola) de la distribución muestral. En una prueba de dos colas, la región de rechazo está en ambos extremos de la distribución muestral.



2.5.5 La decisión.

Si la prueba estadística da un valor (probabilidad asociada) que está en la región de rechazo, se rechaza H_0. El razonamiento en que se apoya este proceso es muy simple, si es muy pequeña la probabilidad asociada con la ocurrencia conforme a la hipótesis de nulidad de un valor particular en la distribución muestral, podemos explicar la ocurrencia efectiva de ese valor de dos maneras: suponiendo que la hipótesis de nulidad es falsa o que el evento raro e improbable ha sucedido, la probabilidad de que la segunda explicación sea correcta está dada por α, pues el rechazo de H_0 cuando es verdadera es el error tipo I.

Cuando la probabilidad asociada con un valor observado de una prueba estadística es igual o menor que el valor previamente determinado de α, concluimos que H_0 es falsa. El valor observado es llamado "significativo". La hipótesis en prueba, H_0 se rechaza siempre que ocurra un resultado significativo. Se llama valor "significativo" a aquel cuya probabilidad asociada de ocurrencia de acuerdo con H_0 es igual o menor que α.

Ejemplo. Dos muestras aleatorias de entrevistados fueron utilizadas para medir el efecto de una campaña de publicidad para el uso del cinturón de seguridad en los automóviles.

Los resultados fueron los siguientes:

	Antes de la campaña	**Después de la campaña**
Tamaño de la muestra	200	220
Incidencia del uso del cinturón de seguridad	24%	30%

Se podría obtener la probabilidad de alcanzar un incremento observado estrictamente al azar (cuando la incidencia de uso en realidad no cambió), entonces podemos estar en la posición de responder la pregunta. Si la probabilidad fue muy pequeña, podríamos decir que el incremento observado no es razonablemente atribuido al azar y tener una explicación alternativa: el efecto de la campaña publicitaria es más aceptable. Por otro

lado, si la probabilidad fuera grande, podríamos razonablemente atribuir el incremento observado al azar y concluir que no es estadísticamente significativa.

En esencia, las pruebas de significancia estadística consisten solamente en determinar la probabilidad de obtener los resultados observados (diferencias, etc.) por azar. Si esta probabilidad es muy pequeña, menor al valor especificado llamado nivel de significancia, podemos decir que los resultados son significantes en el nivel predeterminado de probabilidad.

El procedimiento es formalizado para especificar dos hipótesis entre si:

$H_0 : \theta = \theta_1$

$H_1 = \theta > \theta_1$

El procedimiento conceptual de pruebas significativas consiste en determinar la probabilidad de obtener resultados de muestras (o resultados más extremos que los observados) cuando de hecho H_0 es verdadera. En otras palabras, deseamos determinar la probabilidad de obtener los incrementos observados (o más) al azar. Si esta probabilidad es tan pequeña como la seleccionada previamente, estaremos dispuestos a aceptar como "límite de fluctuación al azar" y podemos rechazar H_0 y aceptar la hipótesis H_1.

La probabilidad es determinada a través del uso de distribuciones muestrales apropiadas (pruebas estadísticas).

En nuestro ejemplo, la prueba estadística apropiada (prueba de hipótesis para dos proporciones de dos grupos independientes) podría producir una probabilidad de .102 de obtener un incremento (o más) observando en la incidencia del uso de cinturón de seguridad al azar.

Si nosotros establecemos un 5% como el límite aceptable de fluctuación al azar, podríamos rechazar H_0 si la probabilidad calculada es menor a 5% y en consecuencia aceptar la alternativa H_1. Si la probabilidad calculada es más grande de 5% podríamos decidir no rechazar H_0 (aceptarla) o, en otras palabras, concluir que no hay evidencia para rechazarla.

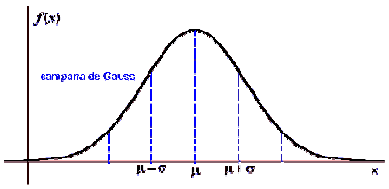
En nuestro ejemplo .102>.05, por lo que no rechazo H_0 y se concluye que el incremento observado en el conocimiento no es estadísticamente significativo.

2.6 Distribuciones comúnmente utilizadas.

Si bien existe un gran número de distribuciones, sólo alguna son las más utilizadas en la Investigación de mercados como: La Normal, Binomial, Poisson, la distribución *t* de Student, la Ji cuadrada y la distribución *F*, que describiremos a continuación.

2.6.1 La Normal.

La distribución de uso más frecuente es la distribución Normal. Fue reconocida por primera vez por el francés Abraham de Moivre (1667-1754). Posteriormente, Carl Friedrich Gauss (1777-1855) elaboró desarrollos más profundos y formuló la ecuación de la curva; de ahí que también se la conozca más comúnmente como la "campana de Gauss". La distribución de una variable Normal está completamente determinada por dos parámetros, su media y su desviación estándar, denotadas generalmente por μ y σ respectivamente.



Es una distribución continua y su función de densidad está dada por:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad x \in R$$

Propiedades de la distribución Normal:

La distribución Normal posee ciertas propiedades importantes que conviene destacar:

i. Tiene una única moda, que coincide con su media y su mediana.
ii. La curva Normal es asintótica al eje de abscisas. Por ello, cualquier valor entre $-\infty$ y $+\infty$ es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.
iii. Es simétrica con respecto a su media μ. Según esto, para este tipo de variables existe una probabilidad de un 50% de observar un dato mayor que la media, y un 50% de observar un dato menor.
iv. La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación típica (σ). Cuanto mayor sea σ, más aplanada será la curva de la densidad.
v. El área bajo la curva comprendida entre los valores situados aproximadamente a dos desviaciones estándar de la media es igual a 0.95. En concreto, existe un 95% de posibilidades de observar un valor comprendido en el intervalo $(\mu - 1.96\sigma, \mu + 1.96\sigma)$.
vi. La forma de la campana de Gauss depende de los parámetros μ y σ. La media indica la posición de la campana, de modo que para diferentes valores de μ la gráfica es desplazada a lo largo del eje horizontal. Por otra parte, la desviación estándar determina el grado de apuntamiento de la curva. Cuanto mayor sea el valor de σ, más se dispersarán los datos en torno a la media y la curva será más plana. Un valor pequeño de este parámetro indica, por tanto, una gran probabilidad de obtener datos cercanos al valor medio de la distribución.

No existe una única distribución Normal, sino una familia de distribuciones con una forma común diferenciadas por los valores de su media y su varianza. De entre todas

ellas, la más utilizada es la distribución Normal estándar, que corresponde a una distribución de media 0 y varianza 1. Así, la expresión que define su densidad se puede obtener:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

Es importante conocer que, a partir de cualquier variable X que siga una distribución $N(\mu,\sigma)$, se puede obtener otra característica Z con una distribución Normal estándar, sin más que efectuar la transformación:

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

Esta propiedad resulta especialmente interesante en la práctica, ya que para una distribución $N(0,1)$ existen tablas publicadas a partir de las que se puede obtener de modo sencillo la probabilidad de observar un dato menor o igual a un cierto valor Z, y que permitirán resolver preguntas de probabilidad acerca del comportamiento de variables de las que se sabe o se asume que siguen una distribución aproximadamente Normal.

2.6.2 Binomial.

La distribución Binomial representa una distribución discreta que indica la probabilidad de x sucesos en n intentos. La distribución Binomial representa una situación en la que se tiene:

- n sucesos independientes.
- dos posibles resultados.

La utilizamos cuando tenemos poblaciones conformadas por sólo dos clases, por ejemplo, masculino-femenino, letrado e iletrado, etc. Todas las observaciones posibles caerán en una u otra de las dos clasificaciones.

En cualquier población de dos clases, al saber que la proporción de casos en una clase es p, entonces la proporción en la otra clase será $1 - p$.
Para una n de gran tamaño, la Binomial se aproxima a una Normal, aunque si la p es cercana a ½ se aproxima más rápidamente y se acerca más lentamente conforme tiende a 0 o a 1.

Su distribución de probabilidad está dada por:

$$P(X) = \binom{n}{X} p^x (1-p)^{n-x} \qquad x = 0,1,2,...,n$$

Donde:

p = probabilidad de éxito en el suceso dado. También
$\mu = np$

$$\sigma = \sqrt{np(1-p)}$$

2.6.3 Poisson.

Es una distribución discreta, con frecuencia se le contempla como una distribución de la cantidad de éxitos en un periodo determinado de tiempo. Su función de densidad está dada por:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0,1,2,...,n$$

Donde:
λ : n*p (número de veces (n) que se realiza el experimento por la probabilidad p)
x: el número de éxitos cuya probabilidad se está calculando
e: 2.71828…

Su media y desviación estándar están dadas por:

$$\mu = \lambda$$
$$\sigma = \sqrt{\lambda}$$

Para una λ grande, la Poisson se aproxima bastante a una Normal.

2.6.4 t de Student.

La distribución *t* de Student es otra distribución en forma aproximadamente de campana, pero menos "crecida" que la Normal. La media y desviación estándar de la distribución *t* están dadas por:

$$\mu = 0$$
$$\sigma = \sqrt{\frac{\upsilon}{\upsilon - 2}}$$

Respecto a υ grande, la distribución *t* es aproximada a la Normal.

2.6.5 Ji cuadrada.

Es la distribución de la suma de las variables Normales estándares independientes elevadas al cuadrado. Se encuentra caracterizada por un solo parámetro *k* que recibe el nombre de grados de libertad, ésta distribución interviene en la inferencia estadística y de manera especial al hacer inferencia con respecto a las varianzas.

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} \qquad x > 0$$

Espacio paramétrico: grados de libertad $k \in \{1,2,3,...\}$
Valor esperado: k
Varianza: $2k$

Si una variable aleatoria X tiene distribución Ji cuadrada con k grados de libertad, entonces si k es grande, la variable aleatoria $Z = \frac{X-k}{\sqrt{(2k)}}$ tiene distribución aproximada Normal estándar.
En la práctica, si k es grande, si es requiere la probabilidad acumulada $F(x)$ con F distribución Ji cuadrada, se puede obtener su valor aproximado buscando en la tabla Normal

$$F_N\left(\frac{x-k}{\sqrt{(2k)}}\right)$$

en que F_N es la distribución Normal estándar. Se puede utilizar como criterio la condición $k > 200$.

2.6.6 Distribución *F.*

Recibió este nombre en honor a Sir Ronald Fisher, uno de los fundadores de la estadística moderna. Esta distribución de probabilidad se usa como estadística prueba en varias situaciones. Se emplea para probar si dos muestras provienen de poblaciones que poseen varianzas iguales. Esta prueba es útil para determinar si una población Normal tiene una mayor variación que la otra y también se aplica cuando se trata de comparar simultáneamente varias medias poblacionales. La comparación simultánea de varias medias poblacionales se conoce como análisis de varianza (ANOVA). En ambas situaciones, las poblaciones deben ser normales y los datos tener al menos la escala de intervalos.

Características de la distribución F:

1. Existe una "familia" de distribuciones F. Un miembro específico de la familia se determina por dos parámetros: los grados de libertad en el numerador y en el denominador . Existe una distribución F para la combinación de 29 grados de libertad en el numerador y 28 grados en el denominador. Existe otra distribución F para 19 grados en el numerador y 6 en el denominador.
2. La distribución F es una distribución continua.
3. F no puede ser negativa.
4. La distribución F tiene un sesgo positivo.
5. A medida que aumentan los valores, la curva se aproxima al eje x, pero nunca lo toca.

Su distribución está dada por:

$$f(X) = \left\{ \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} X^{\frac{v_1}{2}-1} \left(1+\frac{v_1}{v_2}X\right)^{-\frac{1}{2}(v_1+v_2)} \right. \qquad x > 0$$

Donde:

v_1 y v_2 : grados de libertad

3. Calculando el tamaño de muestra.

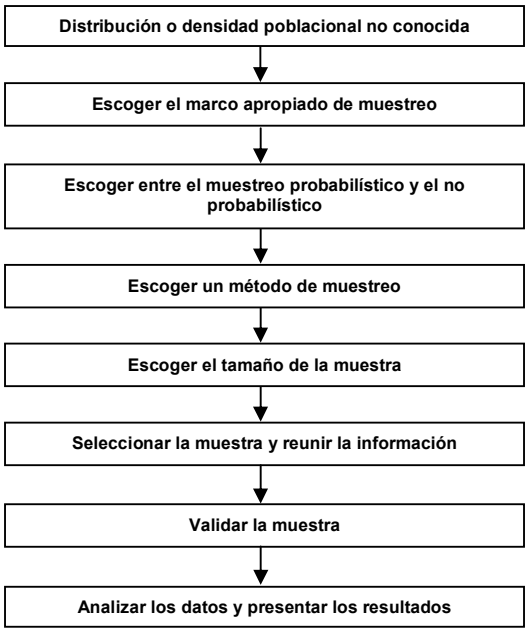
En un estudio de tipo cuantitativo debemos cuidar los detalles con el fin de no afectar negativamente la calidad de la información, considerando que uno de los aspectos más importantes es el diseño de la muestra. La mayoría de los estudios de mercado cuantitativos basan la recolección de la información en diseños muestrales en los cuales se establece a cuantos individuos debemos o queremos entrevistar del total del universo, en caso de que se conozca el total del universo. La idea es hacer un número tal de entrevistas que, con cierto margen de error, podamos decir que los resultados son "representativos" del resto de la población.

3.1 Necesidades básicas del muestreo.

En el momento en que se decide levantar una encuesta y no investigar a todo el universo, de entrada estamos aceptando que el estudio tendrá un margen de error estadístico, pero en general no se tiene otra alternativa ya que sería prácticamente incosteable en términos de tiempo y dinero llevar a cabo un censo. En tal caso es necesario mantener los tamaños de muestra en números tales que la toma de decisiones que en los resultados se base sea perfectamente confiable porque los resultados corresponden a la mayoría de la población, probabilísticamente hablando.

Uno de los aspectos más importantes a destacar en un estudio de mercado es el tamaño de la muestra. A continuación mostramos cuales son los pasos a seguir en un proceso común de muestreo.

Pasos del proceso de muestreo:



Como primer paso debemos establecer los objetivos de manera clara y concisa, manteniendo los objetivos suficientemente simples. Definimos la población que va a ser muestreada, por ejemplo, si la muestra consiste en adultos, definimos qué se entiende

por adulto (aquellos con más de 18 años de edad por ejemplo). Seleccionamos el marco de tal manera que la lista de unidades muestrales y la población objetivo concuerden lo más posible. Seleccionamos el diseño de muestreo, incluyendo el número de elementos de la muestra, así la muestra nos proporcionará suficiente información para los objetivos de la encuesta, una vez que hemos levantado nuestra encuesta se analizan los datos, desde un análisis descriptivo hasta un análisis de conglomerados por ejemplo, según se requiera para presentar los resultados.

Cuando diseñamos una muestra tenemos que asumir o suponer algunas cosas que en realidad no sabemos, pero cuyas suposiciones pueden ser bastante predecibles o poco riesgosas. Por ejemplo, tradicionalmente asumimos que la distribución de probabilidad del comportamiento de la población se perfila como una Normal.

En la mayoría de los casos esta suposición es bastante aceptable, por ejemplo, si son suficientes datos agregados o promediados. Se pueden utilizar distintos niveles o intervalos de confianza para el diseño y análisis de un estudio de mercado pero tradicionalmente es el 95% el de mayor uso por ser lo suficientemente "estricto" para muchas aplicaciones reales.

3.2 Cálculo del tamaño de muestra y margen de error estadístico.

Cálculo de la muestra que estadísticamente nos da representatividad del universo y el margen del error estadístico que la misma tendrá simplemente por ser una muestra y no un censo se determinará de la siguiente manera:

$n = \left[\dfrac{4pqN}{s^2(N-1) + 4pq} \right]$, donde n = máximo entero $\leq$ que la cantidad entre paréntesis.

$s = \sqrt{\dfrac{\dfrac{4pqN}{n} - 4pq}{N-1}}$, donde máx$\{s\}$ es cuando $p = q = \frac{1}{2}$.

Donde:

N = tamaño del universo que puede ser entrevistado.
n = tamaño de la muestra.
s = margen del error estadístico.
p = probabilidad de ocurrir el evento y
q = probabilidad de no ocurrir el evento (1-p).

Supuestos:

- Una curva de distribución de tipo Normal.
- Un intervalo de confianza del 95%.
- El máximo para la desviación estándar s se alcanza con $p = q = \frac{1}{2}$ de tal manera que al sustituir estos valores en s, nos da una cota superior para n.

Pensemos que queremos hacer un estudio de refrescos en las cuidades de México, Guadalajara y Monterrey, que el 50% de la población consume la categoría y que el

margen de error estadístico que queremos en el proyecto es del 5% (es decir, s = .5). También sabemos que los universos son los siguientes:

Cuidad	Universo	Población consumidora
México	25,000,000	12,500,000
Guadalajara	5,000,000	2,500,000
Monterrey	4,000,000	2,000,000

Si calculamos el tamaño de muestra (en números enteros) que necesitamos en cada cuidad para mantener el margen de error en 5% es:

$$\text{México} = \left[\frac{4(.5)(.5)(12{,}500{,}000)}{(.5)^2(12{,}500{,}000-1)+4(.5)(.5)}\right] = \left[\frac{12{,}500{,}000}{31{,}251}\right] = 400,$$

$$\text{Gdl.} \quad = \left[\frac{4(.5)(.5)(2{,}500{,}000)}{(.5)^2(2{,}500{,}000-1)+4(.5)(.5)}\right] = \left[\frac{2{,}500{,}000}{6{,}251}\right] = 400 \quad \text{y}$$

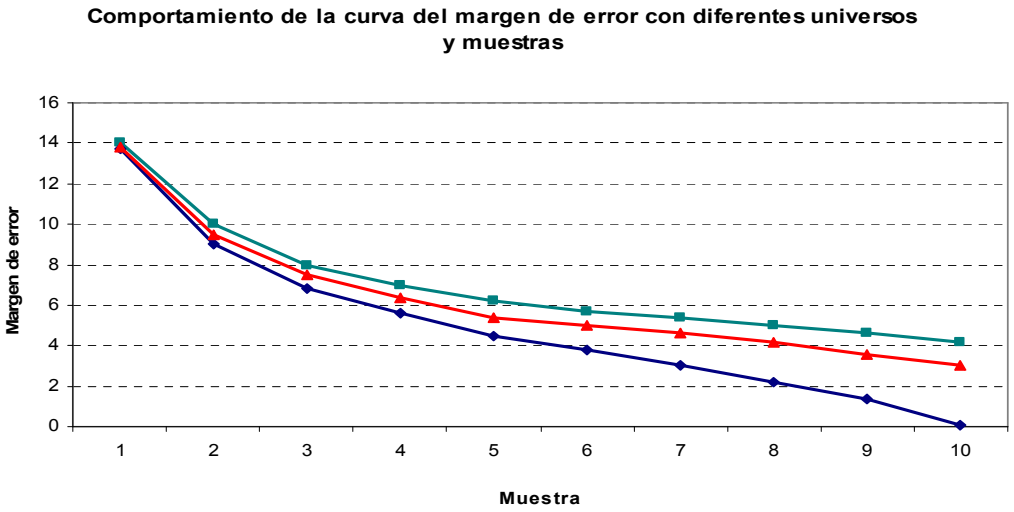
$$\text{Mty.} \quad = \left[\frac{4(.5)(.5)(2{,}000{,}000)}{(.5)^2(2{,}000{,}000-1)+4(.5)(.5)}\right] = \left[\frac{2{,}000{,}000}{5001}\right] = 400$$

Con esto podemos observar que, aquella afirmación que se podría pensar, de que se tienen que hacer más entrevistas en México que en Guadalajara y Monterrey, desde el punto de vista estadístico es una equivocación, ya que los universos son de tal tamaño, que el margen de error es el mismo si hacemos una muestra del mismo tamaño en cada plaza.

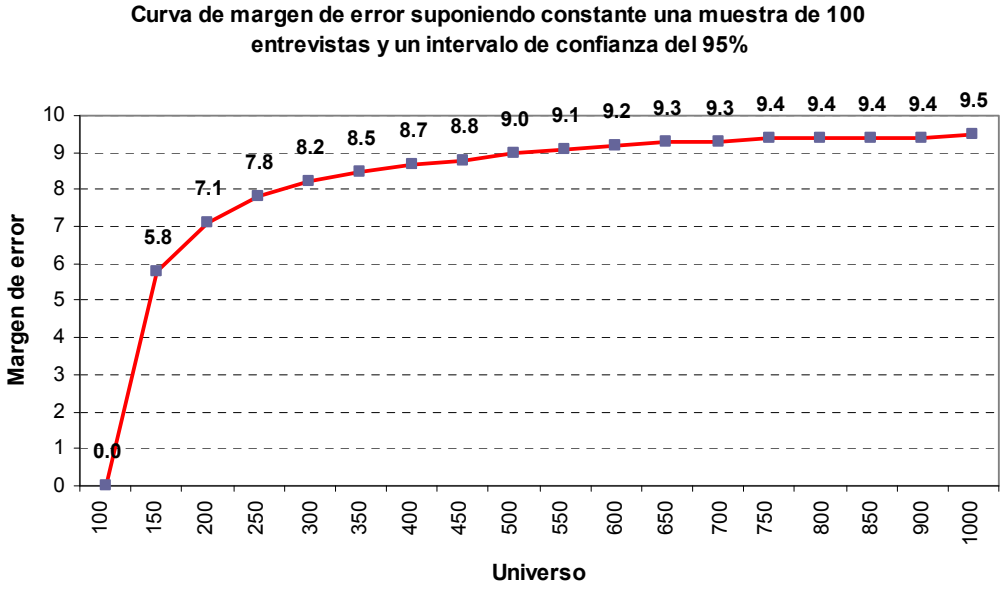
3.2.1 Clasificación de los universos.

En el ejemplo anterior vimos, que no obstante que los universos de Guadalajara y Monterrey eran mucho más pequeños que el de la cuidad de México, el comportamiento del margen de error era exactamente igual, sin embargo, llega un momento en que a universos pequeños el margen de error comienza comportarse de forma distinta; ahí es donde encontramos la diferencia en definición de un universo prácticamente infinito y uno finito.

En la gráfica observamos que cuando el universo es pequeño, el margen de error tiene más variación.



En esta otra gráfica observamos que cuando el universo es más grande, el margen de error no tiene tanta variación. Esto nos indica que las curvas comienzan a cambiar cuando el universo es mayor o menor que 4,000, por lo que, a partir de la curva diremos que: universo prácticamente infinito = cuando el universo es ≥ 4,000 y universo finito = cuando el universo es < 4,000.



3.3 Selección del tipo de muestra.

Para seleccionar una muestra con frecuencia el primer paso es ver si existe alguna lista u otro desglose organizado (por ejemplo, geográfico) del universo meta, si no, las únicas alternativas a nuestra disposición son el muestreo de localización o el marcar números aleatorios en el teléfono.

Una distinción comúnmente establecida en muestreo la constituye las muestras probabilísticas y las basadas en su finalidad (no necesariamente probabilísticas). En las muestras probabilísticas, cada miembro del universo objetivo tiene una probabilidad fija (con frecuencia igual) de ser un miembro de la muestra objeto. Las muestras no probabilísticas por otra parte le prestan mayor importancia a algunos segmentos del universo objetivo y se derivan para "representar excesivamente" a estos segmentos importantes.

Por ejemplo, para una compañía que vende autos la muestra probabilística podría ser el total de la población, mientras que la basada en su finalidad podría obtenerse de la población masculina mayor de edad con ingresos promedio mensual superiores a 30,000.

3.4 Muestreo probabilístico.

El muestreo probabilístico implica cuatro consideraciones 1. Se debe especificar la población meta, el grupo acerca del cual se está buscando información. 2 Se necesita desarrollar el método para seleccionar la muestra. 3. Debemos determinar el tamaño de la muestra, el cual dependerá de las necesidades de exactitud, la atracción dentro de la población y el costo y 4. Se debe considerar el problema de la ausencia de respuesta.

Algunos de los enfoques de muestra más importantes incluyen el muestreo simple aleatorio, sistemático, estratificado, universal (censo), grupal y secuencial.

3.4.1 Muestro simple aleatorio.

La forma más conocida y "democrática" para seleccionar la muestra y sus puntos, es la aleatoria. Este método básicamente asegura que cada persona en el objetivo tiene igual probabilidad de quedar comprendido dentro de la muestra. Esto se logra utilizando una tabla de números aleatorios para generar *n* (el número deseado a muestrear) números al azar entre 1 y el número de nombres de la lista. Este método es equivalente a la lotería clásica en donde los nombres se incluyen o colocan en un sombrero para luego seleccionar al azar, sin embargo, la muestra aleatoria no es un procedimiento 100% seguro, por tanto, las muestras al azar representan al universo de manera adecuada con cierto margen de error, podemos decir que ciertas muestras seleccionadas al azar y en especial las pequeñas, podrán no representar a la población general adecuadamente.

Ejemplo. Consideramos los datos la siguiente tabla que representa los 32 estado de la República Mexicana:

No	Estado	Total	Hombre	Mujer
1	Aguascalientes	1,065,416	515,364	550,052
2	Baja California	2,844,469	1,431,789	1,412,680
3	Baja California Sur	512,170	261,288	250,882
4	Campeche	754,730	373,457	381,273
5	Coahuila de Zaragoza	2,495,200	1,236,880	1,258,320
6	Colima	567,996	280,005	287,991
7	Chiapas	4,293,459	2,108,830	2,184,629
8	Chihuahua	3,241,444	1,610,275	1,631,169
9	Distrito Federal	8,720,916	4,171,683	4,549,233
10	Durango	1,509,117	738,095	771,022
11	Guanajuato	4,893,812	2,329,136	2,564,676
12	Guerrero	3,115,202	1,499,453	1,615,749
13	Hidalgo	2,345,514	1,125,188	1,220,326

14	Jalisco	6,752,113	3,278,822	3,473,291
15	México	14,007,495	6,832,822	7,174,673
16	Michoacán de Ocampo	3,966,073	1,892,377	2,073,696
17	Morelos	1,612,899	775,311	837,588
18	Nayarit	949,684	469,204	480,480
19	Nuevo León	4,199,292	2,090,673	2,108,619
20	Oaxaca	3,506,821	1,674,855	1,831,966
21	Puebla	5,383,133	2,578,664	2,804,469
22	Querétaro Arteaga	1,598,139	772,759	825,380
23	Quintana Roo	1,135,309	574,837	560,472
24	San Luis Potosí	2,410,414	1,167,308	1,243,106
25	Sinaloa	2,608,442	1,294,617	1,313,825
26	Sonora	2,394,861	1,198,154	1,196,707
27	Tabasco	1,989,969	977,785	1,012,184
28	Tamaulipas	3,024,238	1,493,573	1,530,665
29	Tlaxcala	1,068,207	517,477	550,730
30	Veracruz de Ignacio de la Llave	7,110,214	3,423,379	3,686,835
31	Yucatán	1,818,948	896,562	922,386
32	Zacatecas	1,367,692	659,333	708,359
	Total	**103,263,388**	**50,249,955**	**53,013,433**

Fuente: II Conteo de Población y Vivienda 2005, INEGI

Para obtener una muestra al azar de 10 estados se introdujeron números aleatorios, ordenándolos en orden creciente y seleccionando los primeros diez, quedando la lista de la siguiente forma:

No	Estado	Total	Hombre	Mujer	Aleatorio
10	Durango	1,509,117	738,095	771,022	2
28	Tamaulipas	3,024,238	1,493,573	1,530,665	5
17	Morelos	1,612,899	775,311	837,588	6
20	Oaxaca	3,506,821	1,674,855	1,831,966	12
4	Campeche	754,730	373,457	381,273	14
22	Querétaro Arteaga	1,598,139	772,759	825,380	18
25	Sinaloa	2,608,442	1,294,617	1,313,825	19
12	Guerrero	3,115,202	1,499,453	1,615,749	19
21	Puebla	5,383,133	2,578,664	2,804,469	19
11	Guanajuato	4,893,812	2,329,136	2,564,676	21
Total		**28,006,533**	**13,529,920**	**14,476,613**	

Observamos que en cierto sentido, esta muestra no representa bien a la República Mexicana, ya que simplemente deja fuera el estado más "importante" (Distrito Federal). Por tanto, es importante mencionar que, en *promedio*, las muestras aleatorias representan al universo de manera adecuada, ciertas muestras al azar y en especial, las pequeñas muestras al azar, podrán no representar a la población general adecuadamente.

3.4.2 Muestreo sistemático.

Este procedimiento genera puntos de muestra objeto, seleccionando un punto de arranque arbitrario y luego, seleccionando una enésima persona en una sucesión de una lista. El principal problema con este procedimiento es que existe un ciclo en los datos, el cual se relaciona con el intervalo existente entre entrevistados. La principal ventaja del muestreo del enésimo nombre es la facilidad del muestreo aleatorio; en primer lugar, no es necesario que se genere un conjunto de números aleatorios, en segundo lugar, estos números no tienen que compararse con entrevistados individuales.

Ejemplo. Considerando la tabla 2.4, una muestra del enésimo nombre la constituirían los estados 1, 11, 21, 31 por ejemplo:

No	Estado	Total	Hombre	Mujer
1	Aguascalientes	1,065,416	515,364	550,052
11	Guanajuato	4,893,812	2,329,136	2,564,676
21	Puebla	5,383,133	2,578,664	2,804,469
31	Yucatán	1,818,948	896,562	922,386
	Total	**103,263,388**	**50,249,955**	**53,013,433**

Si bien, no se trata de una muestra perfecta, suele considerarse tan útil como una muestra aleatoria.

3.4.3 Muestreo estratificado.

Para muchos estudios el universo objetivo puede dividirse en segmentos con diferentes características. Para este caso, la información respecto a los segmentos (estratos) puede ser utilizado para diseñar el plan de muestreo. Específicamente, diferentes planes de muestreo pueden establecerse para cada estrato, esto garantiza que cada estrato habrá de encontrarse adecuadamente representado.

Utilizando la tabla 2.4 dividiremos la República mexicana en 6 estratos de acuerdo a su localización geográfica. Entonces tenemos $N = 32$ (Total), $N_1 = 11$(Pacífico), $N_2 = 4$ (Golfo), $N_3 = 2$ (Caribe), $N_4 = 3$ (Norte), $N_5 = 7$ (Meseta central), $N_6 = 5$ (Altiplano).

Pacífico	Golfo de México	Caribe	Norte	Meseta central	Altiplano
Baja California	Tamaulipas	Yucatán	Chihuahua	Durango	Hidalgo
Baja California Sur	Veracruz	Quintana Roo	Coahuila	Zacatecas	México
Sonora	Tabasco		Nuevo León	San Luís Potosí	Distrito Federal
Sinaloa	Campeche			Aguascalientes	Morelos
Nayarit				Guanajuato	Puebla
Jalisco				Querétaro	
Michoacán				Tlaxcala	
Guerrero					
Oaxaca					
Colima					
Chiapas					

3.4.4 Muestreo universal.

En la mayor parte de las investigaciones por encuesta entre consumidores, resulta sumamente costoso entrevistar a todos los posibles clientes. Es importante señalar que un verdadero censo casi nunca se podrá obtener.

3.4.5 Muestreo grupal.

Las muestras agrupadas son lo que su nombre indica, muestras recopiladas por grupos. La motivación básica para el muestreo en grupo, lo constituye la reducción de costos.

Podemos dividirlo en etapas como:
1. Seleccionar datos.
2. Seleccionar municipios dentro de los estados.
3. Escoger grupos censales por cuadras o áreas postales dentro de los municipios.
4. Escoger entrevistas al azar dentro de los territorios del censo, etc.

Obsérvese que las muestras de agrupación difieren de las muestras estratificadas en que las muestras por agrupación mucho o la mayor parte del estrato (por ejemplo, estados) se dejan fuera del plan de muestreo. Sin embargo, es muy común seleccionar grupos y luego proceder a usar muestreo estratificado (por ejemplo, en base a ingresos) en cada grupo.

3.4.6 Muestreo secuencial.

En este método se establece una pequeña muestra y los resultados se analizan. Si los resultados son suficientemente claros, se toma una decisión y el resto de la muestra no se hace. De lo contrario, otra muestra se selecciona subsecuentemente. Este método ofrece economía potencial, al reducir el tamaño de la muestra pero por otra parte, este método tarda más en días calendario que un estudio de una sola ocasión. Por consideración de tiempo y por competencia, el muestreo secuencial, rara vez se emplea en el área de Investigación de mercados.

3.5 Muestreo no probabilístico.

En el muestreo no probabilístico se eliminan los costos y el desarrollar un marco de muestreo, pero también la precisión con la que se puede presentar la información resultante, de hecho, los resultados pueden contener sesgos ocultos o incertidumbres que los hacen peores que si no se tuviera información en absoluto. Debe observarse que los problemas no disminuyen al incrementar el tamaño de muestra. Por esta razón, se evitan los diseños no probabilísticos, sin embargo, con frecuencia se usan de manera legítima y eficaz.

El muestreo no probabilístico se usa comúnmente en situaciones como 1) las etapas exploratorias de un proyecto de investigación; 2) la prueba preliminar de un cuestionario; 3) el manejo de una población homogénea; 4) cuando un investigador carece de conocimiento estadístico; 5) cuando se requiere facilidad operativa. Distinguiremos cuatro tipos de procedimientos de muestreo no probabilística: de criterio, de bola de nieve, por cuotas y de conveniencia/localización.

3.5.1 Muestreo de criterio.

En un muestreo de criterio un "experto" emplea su propio criterio para identificar muestras representativas.

Ejemplo. Varios estados podrían ser seleccionados para representar a la República Mexicana.

El muestreo de criterio generalmente está asociado con una diversidad de sesgos obvios y no tan obvios, no hay una forma de cuantificar realmente el sesgo resultante y la incertidumbre, debido a que el marco de muestreo no se conoce y el procedimiento de muestreo no está bien especificado.

Existen situaciones en las que el muestreo de criterio es útil e inclusive aconsejable. Por ejemplo, tal vez sea imposible conseguir una lista de vendedores ambulantes pero un muestreo de criterio podría ser apropiado en este caso.

3.5.2 Muestreo de bola de nieve.

El muestreo de bola de nieve es una forma de muestreo de criterio que resulta muy conveniente cuando es necesario llegar a poblaciones pequeñas y específicas. Bajo un diseño de bola de nieve, a cada encuestado, después de realizar la entrevista, se le pide que identifique a uno o más miembros del campo en particular. El resultado puede ser

una muestra muy útil. Este diseño puede usarse para llegar a cualquier población pequeña, como los buzos en aguas profundas, personas confinadas a sillas de ruedas, propietarios de vehículos ligeros para playa y arena, familias con trillizos, etc. Un problema es que las personas que son visibles socialmente tienen más probabilidad de ser seleccionadas.

3.5.3 Muestreo cuota.

Un muestreo cuota se basa en la idea preconcebida, que ciertas características individuales deben estar representadas adecuadamente, si la muestra ha de ser proyectable. Esencialmente representa un punto intermedio entre una muestra estratificada y una muestra de conveniencia/localización.

Ejemplo. Una empresa desea conocer las opiniones de cuando menos 30 amas de casa entre los 40 y 55 años, por tanto, una muestra cuota podría generarse pidiendo que el entrevistador obtenga datos de las primeras 30 mujeres que estén dentro de la categoría y que estén de acuerdo en participar.

Tal procedimiento no hace que la muestra cuota sea tan buena como una muestra al azar pero si garantiza que, en términos de ciertas características, la muestra habrá de representar el universo objetivo. Se puede hacer escogiendo al azar y eliminando a los que no se encuentran dentro del perfil hasta alcanzar el número *n* de entrevistados.

3.5.4 Muestreo de conveniencia/localización.

Se considera la forma más barata de diseño de muestra ya que es a cualquier persona. Si bien es cuestionable su grado de proyectabilidad, las muestras de conveniencia son muy útiles para generar hipótesis, así como para pruebas-piloto para entrevistas. Una forma relativamente útil de muestreo de conveniencia es el localizado en un punto céntrico, como sería una encuesta de consulta en centros públicos. Estas se aprovechan de lugares céntricos en donde se encuentran grandes cantidades de personas del universo objetivo.

La elección del método dependerá del tipo de información y del objetivo deseado y, lo más importante, del presupuesto y tiempo disponible.

4. Cuestionario.

Primero, deberíamos definir qué entendemos por cuestionario. El cuestionario es un instrumento utilizado para la recolección de información, diseñado para poder cuantificar, universalizar la información y estandarizar el procedimiento de la entrevista. Su finalidad es conseguir la comparación de la información. En términos genéricos, cuando hablamos de cuestionarios estamos hablando muchas veces de escalas de evaluación. Las escalas de evaluación son aquellos instrumentos/cuestionarios que permiten un escalamiento acumulativo de sus preguntas, dando puntuaciones globales al final de la evaluación.

Uno de los elementos fundamentales para poder hacer un correcto estudio de mercado, no importando si es segmentación, recordación publicitaria, actividades y actitudes, etcétera, lo constituye el diseño de un buen cuestionario. Un aspecto importante es lo que llaman los americanos "depurar", junto con la planeación de la muestra y el levantamiento de campo. En la elaboración de un cuestionario apropiado se imponen varias restricciones, por ejemplo, el número, forma y orden de las preguntas específicas están determinados parcialmente por el método de recolección de datos, la disposición y capacidad del entrevistado para responder así como el formato final del cuestionario.

Aún cuando cada cuestionario debe diseñarse teniendo en mente los objetivos específicos de la investigación, hay una secuencia de pasos lógicos que todo investigador debe seguir para elaborar un buen cuestionario:

Fundamentos para el éxito. Definición de propósitos y parámetros de la encuesta. Antes de proceder a medir algo debemos tener una idea muy clara de lo que queremos medir. Ello puede requerir la realización de una revisión de la bibliografía y la consulta con expertos en la materia. Sean actitudes, conductas o conocimientos, se debe definir en forma clara y precisa el objeto de la medida y, a ser posible, determinar y conocer las teorías que sustentan la definición que se acuerde. Un problema puede definirse desde distintas perspectivas teóricas y, por tanto, pueden proponerse definiciones diferentes de un mismo tema.

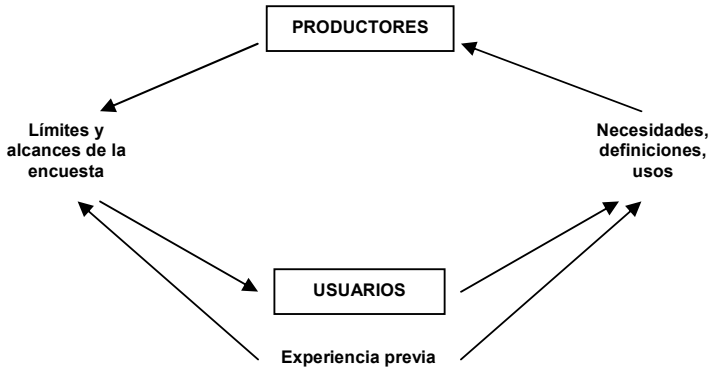
Formulación de cuestionarios. Se trata de establecer el contenido del cuestionario, definir la población a la que va dirigida, la forma de administración y el formato del cuestionario. El propósito de la escala va a determinar en gran medida el contenido de sus preguntas y algunos aspectos relacionados con su estructura y la logística de la recogida de los datos. Por ejemplo, si nuestro tema o aspecto a medir fuera la calidad de vida, deberíamos revisar exhaustivamente todas las posibles características que conforman la calidad de vida: independencia física, vitalidad, equilibrio emocional, sueño, capacidad para relacionarse con otros, etc.

Preparación del resumen estadístico. Explorar la información con análisis descriptivo por ejemplo.

Análisis más especializado. Elegir alguna herramienta estadística que nos apoye en la toma de decisión.

Evaluación del cuestionario. Aquí nos referiremos a que si el cuestionario y por tanto, las preguntas elegidas son indicadores de lo que queremos medir.

Vínculos con los usuarios:



Tanto las entrevistas como los cuestionarios basan su información en la validez de la información verbal de percepciones, sentimientos, actitudes o conductas que transmite el encuestado, información que, en muchos casos, es difícil de contrastar y traducir a un sistema de medida, a una puntuación. Es esta característica lo que hace tan complejo establecer los criterios de calidad de este tipo de instrumentos. A continuación mencionaremos algunas secciones que se incluyen en un cuestionario de manera "general".

4.1 Sección factores demográficos.

Debemos incluir una serie de preguntas que nos indiquen el perfil demográfico, tales como:

- Sexo.
- Edad.
- Nivel socioeconómico.
- Estado civil.
- Profesión.
- Nivel de escolaridad.

4.2 Sección factores perfil general.

Este apartado tiene que ver con las actividades de la vida diaria tales como:

- Actividades en el tiempo libre.
- Incidencia de participar en algún deporte o pasatiempo.
- Posesión de artículos.
- Hábitos de exposición a medios publicitarios.

4.3 Sección factores categóricos.

Esta parte se le conoce como "el cuerpo" del mismo y normalmente incluye la información principal que se puede encontrar en un estudio base:

- Conocimiento de marcas.
- Recordación publicitaria.
- Consumo de marcas.
- Hábitos de compras.
- Elasticidad de precio demanda.

Es necesario aclarar que es tan ineficiente hacer un cuestionario demasiado largo como hacer uno excesivamente escueto, hay que recordar el fin del estudio ya que de nada nos sirve un "x" número de segmentos si no podemos analizar lo que buscamos.

4.4 Sección factores actividades/actitudes en general.

Para esta categoría debemos siempre referirnos a conductas, actitudes, percepciones o estados anímicos y no a hechos o juicios de valor. Actividades: medidas de ocupación real en diversas actividades tales como deportes, oficios, mirar televisión, leer revistas, etc. Actitudes: opiniones de la vida en general, instituciones (educación, iglesia), aspectos particulares (programas de bienestar social) e intereses particulares (lectura).

- Soy feliz.
- Con frecuencia me encuentro desesperado.
- Si pudiera, me iría a vivir a otro país.
- Me siento orgulloso de ser mexicano.
- Prefiero ganar poco pero seguro, que mucho pero con riesgo.
- Prefiero los productos de marca aunque sean más costosos.

Un conjunto de variables de estilos de vida:

Actividades	Intereses	Opiniones	Demográficos
Trabajo	Familia	Ellos mismos	Edad
Pasatiempos	Hogar	Problemas sociales	Educación
Eventos sociales	Trabajo	Política	Ingreso
Vacaciones	Comunidad	Negocios	Ocupación
Entretenimiento	Recreación	Economía	Tamaño de familia
Membresía de un club	Moda	Educación	Vivienda
Comunidad	Alimentos	Productos	Geografía
Compras	Medios	Futuro	Tamaño de cuidad
Deportes	Logros	Cultura	Fase en ciclo vital

Fuente: Joseph T. Plummer, "The Concept and Application of life Segmentation"

4.5 Consejos sobre la presentación.

La apariencia física de un cuestionario es la imagen del investigador con el encuestado. Su misma forma motiva o impide su lectura, en cuestionarios largos hay que identificar cada página con alguna marca por si se separan las hojas, lo mejor es no hacer cuestionarios largos. Una reducción pobre influye en el resultado y también en la calidad de las respuestas obtenidas. Hay que utilizar lenguaje común y corriente, no deben utilizarse palabras vagas ni ambiguas o que tengan varios significados. Las preguntas no deben estar en negativo.

Una vez diseñado el borrador definitivo, corresponde llevar a cabo la realización de una prueba piloto y la evaluación de las propiedades métricas de la escala.

4.6 Escalas.

El tipo de escala (nominal, ordinal, intervalo o de proporciones) dependerá del análisis deseado y de las respuestas que obtengamos. La medición puede definirse como

un proceso estandarizado de asignación de números u otros símbolos a ciertas características de los objetos de interés, de acuerdo con algunas reglas especificadas previamente. La medición a menudo tiene que ver con números, debido a que los análisis matemáticos y estadísticos sólo pueden realizarse sobre números y pueden ser comunicados de la misma manera. Para que la medición sea un proceso de asignación estandarizado, son necesarias dos características 1) debe haber una correspondencia de uno a uno entre el símbolo y la característica del objeto que se está midiendo 2) las reglas de asignación no deben variar con el tiempo y los objetos que se están midiendo.

Existen cuatro escalas de medición fundamentales: nominal, ordinal, de intervalo y de proporciones (o razón), sus propiedades se resumen en el cuadro posterior a la definición:

4.6.1 Escala nominal.

En una escala nominal a los objetos se les asignan categorías etiquetadas, que si bien son mutuamente excluyentes, no existen relaciones necesarias entre ellas, es decir, no se implica ningún orden o espacio. Si a una entidad se le asigna el mismo número que a otra, son idénticas respecto a una variable nominal. El género, la ubicación geográfica y el estado civil son variables de una escala nominal. La única operación aritmética que puede realizarse en una escala de este tipo es un conteo de cada categoría. Las escalas nominales son útiles sólo para procesar frecuencias.

Ejemplo. Para saber la preferencia respecto a un color, una de las maneras más evidentes para obtener una medición de escala, es marcar una sola respuesta en un conjunto de alternativas, sería del siguiente tipo:

¿Qué tipo de color prefiere?
_____ 1. Azul
_____ 2. Rojo
_____ 3. Verde
_____ 4. Amarillo

4.6.2 Escala ordinal.

Se obtiene asignándole un lugar a un objeto u ordenándolos con respecto a una variable común, la escala más evidente es la jerarquización forzada. La escala ordinal proporciona información acerca de qué tanta diferencia hay entre dos objetos. Debido a que no conocemos la cantidad de diferencia entre los objetos, las operaciones aritméticas permitidas se limitan a valores estadísticos como la mediana o la moda.

Ejemplo 1. Para una aplicación de jerarquización:

Favor de jerarquizar los siguientes colores en términos de su preferencia, marcando con un 1 el color preferido, con un 2 el siguiente, etc.
Azul _____
Rojo _____
Verde _____
Amarillo _____

Ejemplo 2. Comparación por pareja, en este método se escoge la alternativa de mayor preferencia (o la más llamativa, o lo que se quiera medir) de dos alternativas al mismo tiempo:

Azul, rojo	_____
Azul, verde	_____
Azul, amarillo	_____
Rojo, verde	_____
Rojo, amarillo	_____
Verde, amarillo	_____

4.6.3 Escala de intervalos.

En una escala de intervalos, los números usados para clasificar a los objetos también representan incrementos iguales del atributo que se está midiendo, esto significa que pueden compararse las diferencias. La ubicación del punto cero no está fija, ya que el cero denota la ausencia del atributo. Las temperaturas Fahrenheit y Celsius se miden con diferentes escalas de intervalo y tienen diferentes puntos cero. Las escalas de intervalos tienen propiedades muy deseables, debido a que virtualmente pueden emplearse toda la gama de operaciones estadísticas para analizar el número resultante, incluyendo la suma y la resta. En consecuencia, es posible calcular una medida aritmética a partir de las mediciones en escalas de intervalos.

Ejemplo 1. Más que atribuir una descripción a cada una de las categorías de respuesta, sólo las dos categorías extremas se titulan, llamamos escala **Adjetivo bipolar**:

Le desagrada mucho				**Le agrada mucho**
1	2	3	4	5

Ejemplo 2. Escala de **Acuerdo-desacuerdo**, es una variante de la escala adjetivo bipolar:

Totalmente en desacuerdo				**Totalmente de acuerdo**
1	2	3	4	5

4.6.4 Escala de proporciones.

Una escala de proporciones es una clase especial de escala de intervalos que tiene un punto cero significativo. Con una escala de este tipo (de peso, participación de mercado o dólares en cuenta de ahorros por ejemplo) es posible decir cuantas veces es mayor o menor un objeto que otro. Este es el único tipo de escala que nos permite hacer comparaciones de magnitud absoluta.

Ejemplo. La mejor manera de obtener información en una escala de proporción es preguntar directamente el valor de la construcción que aparece en la escala de proporción, como:

¿Cuántos pantalones de mezclilla tiene usted? _____

¿Qué edad tiene usted? _____

Finalmente, para la redacción de un buen cuestionario, se dan las siguientes sugerencias:

- Hacer preguntas sencillas, no complejas.
- No utilizar tecnicismos.
- Agrupar preguntas semejantes.
- Utilizar escalas equilibradas o balanceadas.
- Secuencia de preguntas.
- Prueba previa y revisión.
- Codificado de cuestionario.

Tabla. Tipo de escala y sus propiedades.

Tipo de escala	**Tipos de escalas de actitudes**	**Reglas para asignar un número**	**Aplicación típica**	**Estadísticas/Pruebas estadísticas**
Nominal	Escalas dicotómicas "si" o "no"	Los objetos son o idénticos o diferentes	Clasificación (por género, área geográfica, clase social	Porcentajes, Moda/Ji cuadrada
Ordinal	Comparativa, por posición, categoría de elementos, comparación por pares	Los objetos son más grandes o más pequeños	Clasificación (preferencia, posición en la clase)	Porcentajes, Moda/Ji cuadrada, Percentil, Mediana, Correlación/ANOVA Friedman
De intervalos	Asociativa	Los intervalos entre las posiciones adyacentes son iguales	Números índices, escalas de temperatura, medidas de actitudes	Porcentajes, Moda/Ji cuadrada, Percentil, Mediana, Correlación/ANOVA Friedman, Media, Desviación estándar, Correlaciones de momentos/Pruebas *t*, ANOVA, Regresión, Análisis de Factores
De proporción	Ciertas escalas con instrucciones especiales	Existe un cero significativo, por lo que es posible una comparación de magnitudes absolutas	Ventas, ingresos, unidades producidas, costos, edad	Porcentajes, Moda/Ji cuadrada, Percentil, Mediana, Correlación/ANOVA Friedman, Media, Desviación estándar, Correlaciones de momentos/Pruebas *t*, ANOVA, Regresión, Análisis de Factores, Coeficiente de variación

4.7 El diseño de la investigación.

Un diseño de investigación es un marco general o plan para realizar el proyecto de la investigación de mercado. Ahí se detallan los procedimientos para la obtención de la información así como el propósito del estudio, es importante considerar las propiedades de las técnicas estadísticas, en particular su objetivo y sus premisas, algunas técnicas son apropiadas para explorar diferencias entre variables; otras para evaluar magnitudes de las relaciones entre variables y otras más para hacer pronósticos. Las técnicas parten también de premisas distintas y algunas toleran mejor que otras las infracciones a sus supuestos. En general, diversas técnicas pueden ser apropiadas para analizar los datos de determinado proyecto. El diseño lo podemos ordenar de la siguiente manera:

1. Definición de la información necesaria.
2. Análisis de los datos secundarios.
3. Investigación cualitativa.
4. Métodos para el acopio de datos cuantitativos.
5. Procedimiento de medición.

5. Análisis de los datos.

El primer paso del análisis de los datos, consiste en analizar cada pregunta o medirla en sí misma y esto lo podemos hacer mediante la tabulación de los datos. La tabulación consiste en contar el número de casos que caen en las diversas categorías, su uso principal consiste en:

1. Determinar la distribución empírica (distribución de frecuencias) de la variable en cuestión.
2. Calcular la estadística descriptiva (resumen), particularmente la media o los porcentajes.

Posteriormente, los datos se someten a las tabulaciones cruzadas para evaluar si está presente alguna asociación entre dos variables (típicamente nominales). Si las variables se miden como intervalos o proporciones, se transforman en variables de escala nominal para el propósito de una tabulación cruzada.

5.1 Tabulación cruzada.

También se conoce como Tablas cruzadas, Clasificación cruzada y análisis de Tablas de contingencia. En una Tabulación cruzada la muestra se divide en subgrupos, a fin de ver cómo varía la variable dependiente de subgrupo a subgrupo. Las tablas de la Tabulación cruzada requieren menos supuestos para su construcción y sirven como base de varias técnicas estadísticas como el análisis de la Ji cuadrada. Se calculan porcentajes con base en cada celda o por filas o columnas. Cuando los cálculos son por filas o columnas, las tablas de Tabulación cruzada generalmente se conocen como Tablas de contingencia, ya que los porcentajes están condicionados básicamente a los totales de las filas o columnas.

Ejemplo.

Rangos de edad	Hombre	Mujer	Total
De 0 a 4 años	5,169,799	5,004,490	10,174,289
De 5 a 9 años	5,330,945	5,164,996	10,495,941
De 10 a 14 años	5,532,966	5,394,448	10,927,414
De 15 a 19 años	4,967,162	5,100,065	10,067,227
De 20 a 24 años	4,198,535	4,699,604	8,898,139
De 25 a 29 años	3,754,202	4,288,734	8,042,936
De 30 a 34 años	3,702,464	4,180,315	7,882,779
De 35 a 39 años	3,338,421	3,734,343	7,072,764
De 40 a 44 años	2,849,145	3,140,509	5,989,654
De 45 a 49 años	2,373,138	2,623,196	4,996,334
De 50 a 54 años	1,950,556	2,127,659	4,078,215
De 55 a 59 años	1,492,146	1,616,365	3,108,511
De 60 a 64 años	1,239,578	1,375,615	2,615,193
De 65 a 69 años	919,756	1,032,533	1,952,289
De 70 a 74 años	700,873	790,278	1,491,151
De 75 a 99 años	1,010,720	1,215,861	2,226,581
De 100 y más años	6,606	10,766	17,372
No especificado	92,749	91,565	184,314
Total	**48,629,761**	**51,591,342**	**100,221,103**

Fuente: II Conteo de Población y Vivienda 2005

Ésta es una técnica para analizar los datos como ya lo habíamos mencionado pero antes de emprender un estudio formal de técnicas estadísticas avanzadas, veremos los diversos factores que influyen en la elección de una técnica de análisis. Esto nos ayudará a identificar la(s) técnica(s) adecuada(s) con base en nuestras necesidades.

5.2 La meta del análisis.

Nunca debemos olvidar cual es la meta del análisis, de los datos en términos de nuestro objetivo de un estudio de mercado. El análisis de los datos no es un fin en sí mismo, su finalidad es brindar información que sirva para abordar algún problema. La selección de una estrategia de análisis de datos debe comenzar por una consideración de las primeras etapas del proceso.

5.3 El enfoque del análisis.

El enfoque del análisis depende de varios factores, la definición del problema, la determinación de la hipótesis, la selección del tipo de estudio, la precisión de la muestra y el plan del análisis. Además de considerar estos pasos en un sentido práctico, el enfoque obedece a otros factores que son tiempo y dinero así como a procedimientos establecidos por la compañía que solicita el estudio. Es adecuado considerar los siguientes puntos:

5.3.1 Tipo de datos.

El tipo de datos desempeña un papel importante en la elección de la técnica estadística que se emplea para el análisis de los datos. Como ya se señaló, una clasificación útil de datos comprende las escalas de medición nominal, ordinal, de intervalo y de proporciones.

Los datos de una investigación pueden ser resumidos por:

- Describiendo el total de la población, la forma más sencilla es mediante tabulación.
- Proporcionando una medida de tendencia central (media, moda o mediana).
- Proporcionando una medida de variación (rango, rango intercuartiles, varianza o desviación estándar y coeficiente de variación) y/o
- Especificando otras medidas como simetría, máximos de la distribución, sesgo.

5.3.2 Calculando diferencias.

El cálculo de las diferencias puede tomarse de la siguiente manera:

- Comparando un conjunto individual de respuestas, por ejemplo, comparando la distribución ocupacional de los entrevistados.
- Comparando dos conjuntos de respuestas para determinar si hay diferencias de un conjunto a otro. Por ejemplo, comparando el consumo de cigarrillos entre hombres y mujeres.

Estas diferencias pueden ser en términos de toda la distribución, medidas de posición (media, varianza), medidas de dispersión (varianza), otras medidas.

5.3.3 Determinando las asociaciones.

Las asociaciones en una investigación pueden ser expresadas en términos de:

- Un índice de asociación.
- Una forma funcional de relación.
- Agrupamiento de variables u objetos.

La determinación de asociación puede conjuntarse con sólo dos variables y más de dos variables. Las variables pueden ser:

- Divididas en dos casos llamados predictor (o independiente) y criterio (o dependiente). En este caso, el objetivo es analizar la dependencia entre las dos variables.
- No divididas, en este caso, el objetivo es analizar la interdependencia entre las variables.

Algunas de las decisiones que debemos enfrentar implican la dependencia de las observaciones, el número de observaciones por objeto, el número de grupos que están siendo analizados y el control que ejerce sobre las variables de interés.

5.4 Asociación de índices bivariados.

Vamos a medir algo parecido a un rango definitivo, es decir, queremos conocer el valor con asociación perfecta y que tiene la necesidad completa de asociarse. Algún valor que tiene un rango fijo puede ser obligado a otro rango como 0.0 a -1.0 o de -1.0 a +1.0. Para definir límites como éstos podemos reorganizar cuando las asociaciones existen en un relativo nivel de asociación.

Ejemplo. Se asume que el consumo de cigarros tiene un índice de asociación de +.85 para hombres y un índice de asociación de +.40 para mujeres. Si sabemos que el índice de asociación es de -1.0 a +1.0 en donde 0 indica no asociación, podemos decir que la asociación es alta para los hombres y es positiva para ambos.

5.5 Significancia estadística de un índice de asociación.

Un índice de asociación basado en datos de una muestra es exactamente un estadístico, no es diferente de una medida muestral o de una proporción de muestra. Como tal, si la meta del estudio es inferir estadísticamente, entonces la prueba apropiada de significancia debe ser ejecutada.

La prueba de significancia más común consiste en determinar si el índice difiere significativamente de 0. El resultado de la significancia debe ser separado de la significancia substantiva. Así por ejemplo, un índice de .10 basado en una muestra grande puede ser estadísticamente significativo, pero no importante para la práctica.

5.6 Índices comúnmente usados en asociación para dos variables dicotómicas.

Los más conocidos y usados son el índice Phi (ϕ), Q y Y de Yule.

5.6.1 El coeficiente Phi (ϕ).

El coeficiente Phi puede variar de -1 a 1, sin embargo, el límite superior depende de los cuatro totales marginales en la tabla 2 x 2 de los coeficientes calculados.

Cálculo

	+	-	
+	A	B	A+B
-	C	D	C+D
	A+C	B+D	n=A+B+C+D

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

El valor absoluto del coeficiente Phi puede también ser obtenido como:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Ejemplo. Un estudio de cigarros fumados entre 200 matrimonios proporcionan los siguientes datos en la incidencia de fumadores entre esposos y esposas.

	Fumadores	No fumadores	
Fumadoras	80	20	100
No fumadoras	40	60	100
	120	80	200

Por lo tanto, el índice de asociación es

$$\phi = \frac{80 \times 60 - 20 \times 40}{\sqrt{100 \times 100 \times 120 \times 80}} = .41$$

Para esta prueba de significancia, $n\phi^2 = 200 \times .41^2 = 33.6$

De 33.6 > 3.84 (de las tabla 5, Ji cuadrada, para $\alpha = .05$ y un grado de diferencia), concluimos que el coeficiente ϕ es estadísticamente significativo.

Por lo tanto, de los datos del estudio, dado que la asociación entre la incidencia de fumadores de esposos y esposas es positiva (.41) es estadísticamente significativo a un nivel de .05.

5.6.2 Q y Y de Yule.

Es un cálculo fácil para medir la asociación de dos variables dicotómicas. Usando la terminología empleada para el coeficiente ϕ, Q es definido como sigue:

$$Q = \frac{AD - BC}{AD + BC}$$

Así para el ejemplo de incidencia de fumadores que se expuso anteriormente tenemos:

$$Q = \frac{80 \times 60 - 20 \times 40}{80 \times 60 + 20 \times 40} = .714$$

5.7 Índices de asociación comúnmente usados para dos variables multicotómicas.

Varios índices de asociación basados en χ^2 intentan estandarizar el valor de la misma para eliminar la dependencia del índice en el tamaño de muestra como el número de categorías en las dos variables. Cramer´s V tiene al menos propiedades no reprochables.

5.7.1 V de Cramer.

El coeficiente *V* tiene rango de 0 a 1 y es un buen índice de asociación entre dos variables multicotómicas, especialmente cuando dos variables no tienen el mismo número de categorías.

Procedimiento del cálculo

$$V = \sqrt{\frac{\chi^2}{n[\min(r-1, c-1)]}}$$

Donde *min(r-1, c-1)* es el mínimo de *(r-1)* y *(c-1), r* es el número de filas y *c* es el número de columnas en la tabla de contingencia y *n* es el tamaño de muestra.

Prueba de significancia

El cálculo del valor de *V* es considerado estadísticamente significativo si el valor correspondiente a χ^2 es significativo. Por lo tanto, comparar χ^2 con el valor en el cálculo *V* a χ^2* de tablas para *(r-1)(c-1)* grados de libertad, donde *r* es el número de filas y *c* es el número de columnas. Si $\chi^2 > \chi^2$* entonces se concluye que *V* es estadísticamente significativo. O comparar el valor de la probabilidad correspondiente al nivel de confianza (α) preestablecido a χ^2.

Ejemplo. Los datos de un estudio de preferencia por bebidas alcohólicas de 200 entrevistados seleccionados aleatoriamente son los siguientes:

	Negro	Blanco	Hispano	Total
Licor	25 (10)	15 (20)	0 (10)	**40**
Vino	10 (25)	60 (50)	30 (25)	**100**
Cerveza	15 (15)	25 (30)	10 (15)	**60**
Total	**50**	**100**	**50**	**200**

Las frecuencias que se esperan correspondientes a cada celda se presentan en la esquina. Ha sido calculado de la manera usual, es decir, multiplicando los dos totales marginales correspondientes a la celda y dividiendo por el total global.

Para la tabla tenemos:

$$\chi^2 = \frac{(10-25)^2}{10} + \frac{(15-20)^2}{20} + \frac{(0-10)^2}{10} + \ldots + \frac{(20-15)^2}{15} = 48.25$$

Por lo tanto,

$$V = \sqrt{\frac{48.25}{200 \times 2}} = .35$$

Del valor χ^2 de la tabla 5 para (3-1)(3-1) = 4 grados de libertad y $\alpha = .05$, encontramos $\chi^2 = 9.49$.
Dado que 48.25 > 9.49, concluimos que el valor calculado de la V de Cramer es estadísticamente significativo.

5.8 Otros Ji cuadrados χ^2 basados en índices de asociación.

5.8.1 T de Tschuprow.

Es definido como sigue:

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}$$

Donde la notación es la misma que la utilizada para Cramer´s V. Cuando el número de filas en la tabla es igual al de las columnas, *T* toma el rango de 0 a 1.
Usando los datos del ejercicio anterior tenemos:

$$T = \sqrt{\frac{48.25}{200\sqrt{2 \times 2}}} = .35$$

Después el valor χ^2 fue presentado antes de ser estadísticamente significativo, el valor *T* de .35 es declarado significativo. Cramer´s V y Tschuprow´s T siempre tienen el mismo valor para tablas donde el número de filas es igual al número de columnas.

5.8.2 Coeficiente de contingencia de Pearson C.

Es aplicable cuando se desea evaluar la asociación entre dos variables ordinales o entre una variable ordinal y otra continua.

Es definido como:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

El límite inferior de *C* es cero pero el límite superior depende del número de filas y columnas. El límite superior se incrementa como se incrementa el número de filas y columnas, aunque siempre es menor a 1.
Usando los datos del ejemplo anterior, tenemos:

$$C = \sqrt{\frac{48.25}{48.25 + 200}} = .44$$

Sin embargo, el valor máximo del coeficiente de contingencia puede ser alcanzado en base al número de filas y columnas en la tabla. Para una tabla de 2x2 el valor máximo es de .707 y para una tabla de 3x3 el valor máximo es de .756. Podemos estandarizar el coeficiente de contingencia si dividimos por el valor máximo. En este caso .44/.756=.58. En otras palabras, el coeficiente de contingencia alcanza un valor máximo de 58% y es para una tabla de 3x3.
Dado que el valor χ^2 de 48.25 fue estadísticamente significativo, el coeficiente de contingencia también es significativo.

5.9 Otros índices de asociación.

5.9.1 Lambda de Goodman-Kruskal (λ).

Es una medida de reducción en el error cuando el valor de una variable es usado para predecir valores de otras, en rango de 0 a 1. Un valor de 0 es equivalente a concluir que el valor A no necesita ayudar para la predicción del valor B, mientras que el valor de 1 implica que la variable A perfectamente especifica la categoría de B. Tres versiones de Lamdba son típicamente calculados, dos Lambdas asimétricas (una usa la variable de las filas como predictor y la otra usa la variable de las columnas como predictor) y una Lambda simétrica que utiliza como predictor las variables de las filas y las columnas con igual frecuencia.

5.9.2 Tau de Goodman-Kruskal.

Es otro índice basado en el concepto de reducción proporcional del error, es decir, el radio de la medida del error en la predicción de los valores de una variable basados en el conocimiento de la variable única y de su misma medida de error aplicada en la predicción basada en el conocimiento de una variable adicional.

5.10 Índices de asociación comúnmente usados para dos variables ordinales.

5.10.1 Coeficiente de Correlación de rango de Spearman.

El coeficiente de Correlación de rango de Spearman (r_s) es la mejor conocida de todas las medidas de asociación ordinal.

El coeficiente (r_s) tiene tangos de -1.0 para Correlación negativa a +1.0 para Correlación positiva y es particularmente útil cuando tratamos con variables ordinales y pocos o sin resultados vinculados.

Cálculo

- Convierte el resultado original en rangos para cada variable por separado, en principio con el rango 1. Los resultados vinculados son asignados al promedio del rango correspondiente a estos resultados, si el resultado tiene la forma de rango, el paso anterior es innecesario.

- Calcula la diferencia del resultado (*D*) correspondiente a cada par de rangos calculando el coeficiente:

$$r_s = 1 - \frac{6\sum D^2}{n\left(n^2 - 1\right)}$$

Donde *n* es el número de respuestas apareadas.

Prueba de significancia

Cuando el tamaño de la muestra es >10, podemos usar la prueba de significancia del coeficiente de Correlación de rango de Spearman, la cual se calcula de la siguiente manera:

$$t = \frac{r_s}{\sqrt{\frac{1 - r_s^2}{n - 2}}} = r_s\sqrt{\frac{n-2}{1-r_s^2}}$$

Ejemplo. Una empresa de equipo industrial desea conocer si existe alguna Correlación entre el entrenamiento técnico y el desempeño laboral de sus representantes de ventas. Se ha seleccionado una muestra de 10 personas, los datos se presentan en la siguiente tabla:

Representante	Entrenamiento (A)	Desempeño (B)	Diferencia (B-A) =D	D^2
1	1	4	3	9
2	8	12	4	16
3	9	8	-1	1
4	7	13	6	36
5	4	3	-1	1
6	3	2	-1	1
7	4	6	2	4
8	8	6	-2	4
9	7	7	0	0
10	4	5	1	1
Total				**73**

Por tanto

$$r_s = 1 - \frac{6 \times 73}{10(10^2 - 1)} = .56$$

La prueba de significancia estadística de r_s se puede calcular:

$$t = .56\sqrt{\frac{10-2}{1-.56^2}} = 1.91$$

De las tablas de distribución de *t* para $\alpha = .05$ (una cola) y 8 grados de libertad, *t** es igual a 1.86. De 1.91>1.86 concluimos que la Correlación observada entre el entrenamiento técnico y el desempeño es estadísticamente significativo, es decir, significativamente mayor a cero.

5.11 Número de variables.

La mayoría de los estudios de mercado obtienen datos de diversas variables basados en el número de variables empleadas simultáneamente, en el análisis se clasifican de la siguiente manera:

- Una variable.
- Dos o más variables.

5.11.1 Técnicas de una variable.

Son apropiadas cuando hay una sola medición de cada uno de los *n* objetos de la muestra o cuando hay varias mediciones de cada una de las *n* observaciones pero cada variable se analiza de manera aislada.

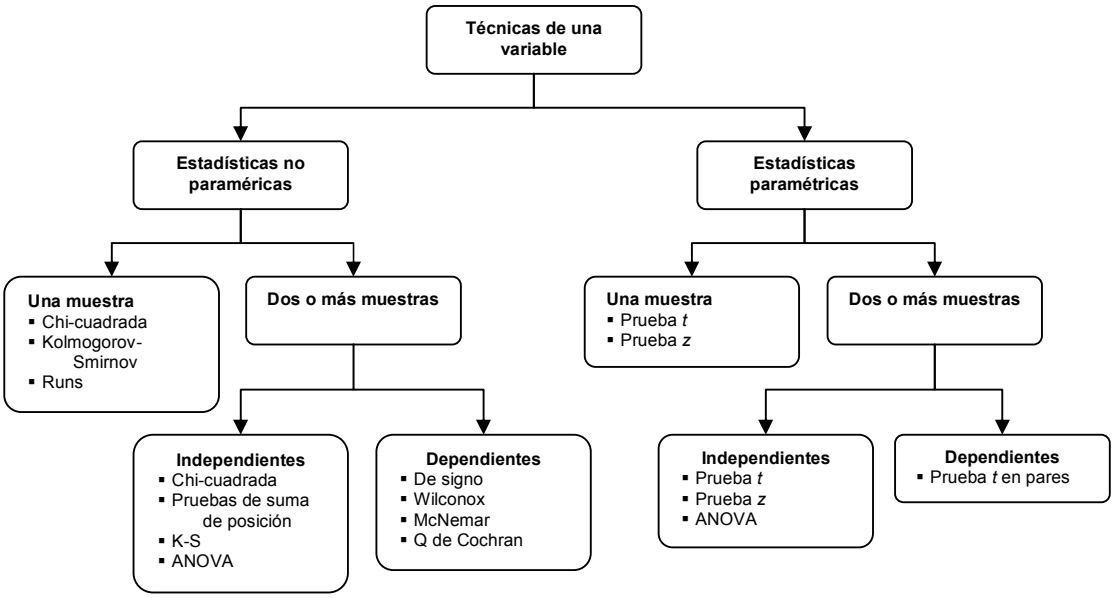
Las técnicas de una variable se pueden clasificar a su vez con base en el tipo de datos, ya sean métricos o no métricos. Los datos no métricos se miden en una escala nominal u ordinal, en tanto que los datos métricos se miden en una escala de intervalos o razones.

Una limitación que tienen es que no son aplicables a casos en los que se desean manejar muchas variables al mismo tiempo (para estos casos, si se requeriría una prueba paramétrica), lo que si se requiere y en general es el supuesto que se debe cumplir en la mayoría de las pruebas no paramétricas para confiar en ellas, es que la muestra haya sido seleccionada en forma probabilística.

Además del problema de los supuestos, algunos experimentos o estudios que se deseen realizar producen respuestas que no es posible evaluar con la escala que tiene más ventajas, por ejemplo, cuando los datos solamente se encuentran en una escala ordinal como cuando se evalúan las habilidades de los vendedores, o el atractivo de cinco modelos de casas, o la preferencia por sopas de cinco marcas diferentes. En general, aspectos como la habilidad o preferencias de un alimento o producto, solamente los podemos ordenar; resultados de este tipo se presentan frecuentemente en estudios de mercado y en otros del campo de las ciencias sociales.

Las pruebas que se mencionarán son las que se podrían necesitar con mayor frecuencia en estudios de mercado, se mencionarán sus principales características y aplicaciones, además de la prueba paramétrica a la que podrían sustituir en caso necesario, así como los supuestos en los que se basa la prueba, que como se podrá ver, son menos rigurosos que para las pruebas paramétricas.

Técnicas estadísticas de una variable

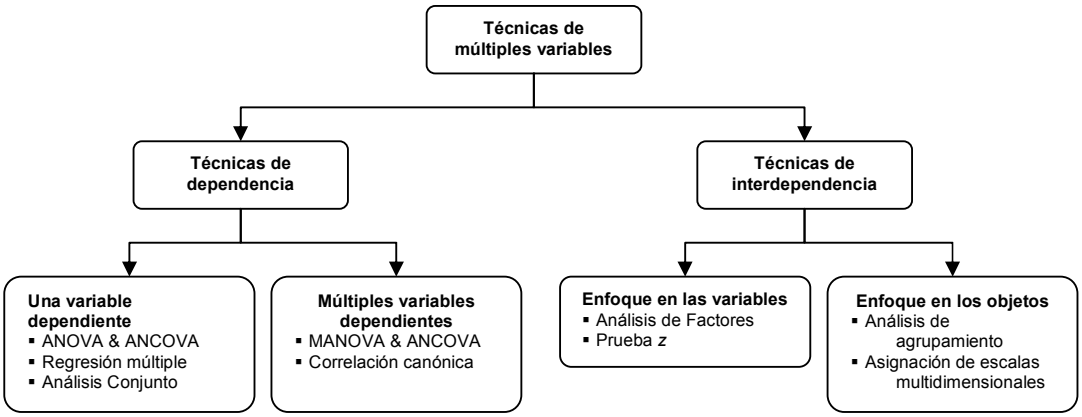


5.11.2 Técnicas de múltiples variables.

Son apropiadas para analizar datos cuando hay dos o más mediciones de cada observación y las variables se van a analizar de manera simultánea.

Las técnicas estadísticas de múltiples variables pueden definirse de manera general como "un conjunto de procedimientos para analizar la asociación entre dos o más series de mediciones que se hicieron sobre cada objeto en una o más muestras de objetos".

Técnicas estadísticas de múltiples variables



Con base en el primer factor, las técnicas de múltiples variables pueden clasificarse de manera general como técnicas de dependencia y técnicas de interdependencia. Las técnicas de dependencia son apropiadas cuando una o más variables se pueden identificar como variables dependientes y las restantes como variables independientes.

La elección apropiada de las técnicas de dependencia adicionalmente está en función de la existencia de una o más variables dependientes en el análisis.

En las técnicas de interdependencia, las variables no se clasifican como dependientes o independientes, más bien, se examina todo el conjunto de relaciones de interdependencia. Las técnicas de interdependencia se pueden clasificar a su vez, según la orientación en variables o en objetos.

5.12 Algunas consideraciones.

La suposición de distribución está implícita en las pruebas estadísticas. Algunas suposiciones generalmente pueden satisfacerse empleando varias pruebas estadísticas comunes.

El término "robustez" es utilizado normalmente para describir las insensibilidades de una prueba particular a las violaciones de los supuestos implícitos. Los procedimientos estadísticos ofrecen la clasificación dentro de las categorías basadas en estructuras supuestas:

- Paramétricas.
- No paramétricas.

Como se mencionó anteriormente, los procedimientos requieren de varios supuestos acerca de la distribución y los parámetros de la población. Los no-paramétricos o distribuciones libres no requieren de tales supuestos, es decir, no se está suponiendo una forma funcional "a priori" para los parámetros, aunque por supuesto siempre hay hipótesis, pudiendo ser independencia, aunque para estos hay otros métodos.

La elección entre pruebas estadísticas paramétricas y no paramétricas puede ser mejor comprendida con una lista de ventajas y desventajas de las pruebas estadísticas no paramétricas.

Prueba de hipótesis	No. grupos/muestras	Propósito	Prueba estadística	Supuestos/comentarios
Distribuciones de frecuencias	Uno	Bondad de ajuste	χ^2	
	Dos	Pruebas de independencia		
Proporciones	Uno	Comparar proporciones de muestras y proporciones	Z	Si σ es conocida y para muestras grandes
			t	Si σ es desconocida y para muestras pequeñas
	Dos	Comparar las proporciones de dos muestras	Z	Si σ es conocida
			t	Si σ es desconocida
Medias	Uno	Comparar la media de la muestra y la población	Z	Si σ es conocida
			t	Si σ es desconocida
	Dos	Comparar las medias de dos muestras	Z	Si σ es conocida
		Comparar las medias de dos muestras (independientes)	t	Si σ es desconocida
		Comparar las medias de dos muestras (relacionadas)	t	Si σ es desconocida
	Dos o más	Compara medias de muestras múltiples	F	Utilizando el marco de trabajo del análisis de varianza
Varianza	Uno	Comparar varianzas de muestras y poblaciones	χ^2	
	Dos	Comparar varianzas de muestras	F	

5.13 Presentación de los resultados.

En un momento dado, el investigador debe desarrollar algunas conclusiones a partir del análisis de los datos y presentar los resultados. La presentación ya sea oral, escrita o

en ambas formas, puede jugar un papel fundamental en la capacidad final de la investigación para influir en las decisiones.

6. Herramientas estadísticas.

Herramientas no paramétricas

Cuando el tamaño de muestra es pequeño, tenemos como alternativa la estadística no paramétrica, estas pruebas son apropiadas para escalas nominales y ordinales, lo que las hace menos restrictivas.

6.1 Prueba del signo.

La prueba del signo basa su nombre en la forma de registrar la información. Si el juicio emitido sobre un resultado del tratamiento es positivo, se registra con el signo (+) y si es negativo con el signo (-), por tanto, el resultado de una muestra tiene la apariencia que presenta la serie: $++-+--+---++++++-+...$

Funciones:

- ✓ Cuado se trata de muestras pequeñas pero la población no es Normal, utilizaremos la prueba del signo.

Para que se utiliza

- ✓ Cuando las observaciones cualitativas tienen relación con el cumplimiento o no de cierto atributo.

Cálculo

Se define como estadístico a X que indica el número de signos positivos y se distribuye según la Binomial con media y varianza:

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

Ejemplo. Se quiere evaluar el impacto publicitario después de una campaña de publicidad a 10 personas. Se desea demostrar que la publicidad es convincente.

Si la opinión es favorable se anota el signo +, si es desfavorable se anota el signo -.

H_0: la publicidad es indiferente para el consumidor.

H_1: la publicidad es relevante para el consumidor.

La probabilidad de obtener un signo + es de ½ y como la muestra consta de 10 personas, el número de apariciones del signo + se distribuye como una Binomial, con media y desviación:

$$\mu = np = 10(.5) = 5$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{10(.5)(.5)} = 1.58$$

Si el resultado de la prueba es de $++-++--+++$, el valor de X es de 7.

Para probar si el resultado (7) es significativo, determinamos P(X = 7) en el esquema Binomial con parámetros $n = 10$ y $p = .5$.

Con $\alpha = 0.025$ se busca en la tabla 1 la probabilidad de lograr un resultado como el anterior o mayor y obtenemos:

$Prob\,(X \geq 7) = 1 - P(X \leq 6) = 1 - F(6{,}10{,}.05) = 0$

Conclusión

Se rechaza la hipótesis nula y concluimos que, la campaña de publicidad fue convincente.

6.2 Prueba de rangos con signos de Wilcoxon.

Esta herramienta utiliza la dirección de los datos pero agrega la magnitud.

Funciones

- ✓ Registra la diferencia entre los resultados de dos muestras o de una muestra contra un valor de referencia.

Para que se utiliza

- ✓ Comparación de dos productos.

Cálculo

La estadística de esta prueba se conforma a partir de las diferencias apareadas que se ordenan en términos de su magnitud absoluta, asignándoles un número de rango, así el primero de estos para la menor diferencia, el segundo para la diferencia siguiente y así sucesivamente. En caso de empates, se asigna como número rango el promedio calculado a partir de la suma de los rangos que corresponderían si no hubiese empate. Solo se consideran las diferencias no nulas.

Esta estadística designa la más pequeña entre dos sumas (W+ y W-), que son la suma de los rangos de orden con diferencias positivas y las de diferencias negativas respectivamente.

La hipótesis nula establece que la población es simétrica, y por lo tanto, la distribución de rangos es igual para los casos positivos y negativos.

Tenemos entonces que:

$H_0 : \mu_1 = \mu_2$: la cual se puede rechazar a favor de la hipótesis alternativa.

$H_1 : \mu_1 < \mu_2$: cuando W+ es pequeña y W- es grande, o

$H_1 : \mu_1 > \mu_2$: en caso de que W+ es grande y W- es pequeña o

$H_1 : \mu_1 \neq \mu_2$: si ambos, W+ y W- son pequeños.

Ejemplo. Se hizo una prueba de producto a 10 personas con los siguientes resultados:

Persona	1	2	3	4	5	6	7	8	9	10
A	9	9	7	7	8	6	9	7	8	8

B	8	7	7	8	6	7	7	8	8	7

La diferencia cero se descarta.

Persona	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	9	9	7	7	8	6	9	7	8	8
B	8	7	7	8	6	7	7	8	8	7
A – B	1	2	0	-1	2	-1	2	-1	0	1
#	**3**	**7**		**3**	**7**	**3**	**7**	**3**		**3**

Para la diferencia mínima de 1 o -1 existen cinco valores que les corresponden los rangos 1, 2, 3, 4, 5, cuyo promedio es (1+2+3+4+5) / 5 = 3.

Para la segunda diferencia de 2 o -2 existen tres valores y el rango promedio es (6+7+8) / 3 = 7.

Entonces:
$W+$ = (3+7+7+7+3) = 27
$W-$ = (3+3+3) = 9
Por tanto W= 9

Conclusión

El valor crítico para $\alpha = .05$ y $n = 8$ proporciona el valor 4 (tabla 2), entonces como W= 9, queda por fuera de la región crítica y se rechaza H_0.

Como la hipótesis nula establece que $\mu_a = \mu_b$, es decir, que la calificación es la misma para ambos productos, se aplica la tabla bilateral. La región crítica es w >4 que contiene a W, por lo cual se acepta que la calificación dada por las personas a favor del producto es significativa.

6.3 U de Mann-Whitney.

También conocida como prueba de la suma de rangos, la prueba U de Mann-Whitney es la más conocida de las pruebas para dos muestras independientes.

Funciones

- ✓ Compara medias de dos poblaciones no Normales, cuando provienen de dos muestras independientes.

Para que se utiliza

- ✓ Probar si dos grupos independientes han sido tomados de la misma población.

Cálculo

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = S_2 - \frac{n_2(n_2+1)}{2}, \qquad U = \min[U_1, U_2]$$

Donde

n_1, n_2**:** son la suma de los rangos.
S_1, S_2**:** muestras 1 y 2 respectivamente.

La hipótesis nula $H_0 : \mu_1 = \mu_2$ establece que las poblaciones que se comparten tienen distribuciones idénticas. La hipótesis alternativa es una de las siguientes:

$H_1 : \mu_1 < \mu_2$, para proceder con U_1.
$H_1 : \mu_1 > \mu_2$, para proceder con U_2.
$H_1 : \mu_1 \neq \mu_2$, para proceder con U.

Si se rechaza H_0 porque U_1 es más pequeño que el valor crítico, significa que $\mu_1 < \mu_2$ y, por tanto, la población favorece a la muestra 2.

Ejemplo. Se llevo a cabo un estudio con dos grupos de personas de diferentes estados para determinar el tiempo que utilizan para sus actividades libres. Los resultados son los siguientes:

Puntajes	N	Resultados										S
Guadalajara	N_1=10	10	11	16	18	9	17	14	20	19		S_1=134
Monterrey	N_2=11	8	2	15	13	7	16	12	4	3	19	S_2=99

$$U_1 = S_1 - \frac{n_1(n_1+1)}{2} = 134 - \frac{10(11)}{2} = 79$$

$$U_2 = S_2 - \frac{n_2(n_2+1)}{2} = 99 - \frac{11(12)}{2} = 33$$

Conclusión

Para $\alpha = .05$, $n_1 = 10$ y $n_2 = 11$ en la tabla de Mann-Whitney (tabla 3) obtenemos el valor crítico de 26, por tanto, se rechaza H_0 puesto que 33 > 26. Entonces, las personas de ambos estados tienen diferente distribución acerca de la forma en que utilizan su tiempo libre.

6.4 Kruskal-Wallis.

También llamada prueba *H*, es una generalización de la prueba *U* (o de Mann-Whitney).

Funciones

- ✓ Permite decidir si puede aceptarse la hipótesis de que *k* muestras independientes proceden de la misma población o de poblaciones idénticas con la misma mediana.

Para que se utiliza

- ✓ Cuando se quiere probar la hipótesis nula de más de dos muestras independientes que pertenecen a poblaciones idénticas.

Cálculo

Se ordenan las *n* observaciones de menor a mayor y se les asignan rangos desde 1 hasta *n*. A continuación se obtiene la suma de los rangos correspondientes a los elementos de cada muestra R_j y se halla el rango promedio. Si la hipótesis nula es cierta, es de esperar que el rango promedio sea aproximadamente igual para las *k* muestras; cuando dichos promedios sean muy diferentes es un indicio de que H_0 es falsa.

El estadístico de prueba es:

$$H = \frac{12}{n(n-1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n-1)$$

Donde

k: número de muestras.
n_j: número de casos en la muestra de orden *j*.
n: el número de casos de todas la muestras combinadas.
R_j: suma de rangos de la muestra de orden *j*.
$\sum_{j=1}^{k}$: indica la suma de las *k* muestras.

Está distribuida aproximadamente como una Ji cuadrada con $gl = k - 1$, cuando se cumple que $n_j > 5$.
La hipótesis nula establece que todas las poblaciones se distribuyen de forma idéntica y se plantea:

$$H_0 : n_1 = n_2 = n_k$$

La hipótesis nula se rechaza si *H* supera el valor crítico, esto es $H \geq \chi^2$ para un nivel de confiabilidad de α.

Ejemplo. Se realizó el lanzamiento de un nuevo producto, alimento para niños, considerando para la prueba 3 sabores, los cuales fueron probados por los niños pertenecientes a una muestra. Los sabores se ordenaron después en orden de preferencia con los siguientes resultados:

Puntajes			Rangos		
A	**B**	**C**	**A**	**B**	**C**
96	82	115	4	2	7
128	124	149	9	8	13
83	132	166	3	10	14
61	135	147	1	11	12
101	109		5	6	
n_1=5	n_2=5	n_3=4	R_1=22	R_2=37	R_3=46

Tenemos:
$\alpha = .05$
$N = 14$

$$H = \frac{12}{n(n-1)}\sum_{j=1}^{k}\frac{R_j^2}{n_j} - 3(n-1) = \frac{12}{14(14+1)}\left[\frac{(22)^2}{5} + \frac{(37)^2}{5} + \frac{(46)^2}{4}\right] - 3(14+1) = 6.4$$

Conclusión

De acuerdo a la tabla 4, cuando n_j son 5, 5, y 4, $H \geq 6.4$ tiene una probabilidad de ocurrencia bajo la hipótesis de nulidad de $p < 0.049$, como esta probabilidad es menor que $\alpha = 0.05$, la decisión es rechazar H_0 y aceptar H_1. Concluimos que hay una preferencia significativa por alguno de los sabores.

6.5 Ji cuadrada χ^2 para dos muestras independientes.

Cuando los datos de una investigación consisten en frecuencias de categorías discretas puede usarse la χ^2 para determinar la significación de las diferencias entre los grupos independientes.

Funciones

La hipótesis que usualmente se pone a prueba supone que los dos grupos difieren con respecto a alguna característica.

Para que se utiliza

- ✓ Probar si dos grupos difieren por ejemplo en su acuerdo o desacuerdo con alguna opinión.

Cálculo

La hipótesis de nulidad la podemos probar mediante:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})}{E_{ij}}$$

Donde:

O_{ij}: es el número observado de casos clasificados en la fila *i* de la columna *j*.

E_{ij}: es el número esperado de casos esperados conforme a H_0 que clasificarán en la fila *i* de la columna *j*.

$\sum_{i=1}^{r}\sum_{j=1}^{k}$: indica sumar en todas las filas (*r*) y en todas las (*k*) columnas, es decir, sumar en todas las celdillas.

Ejemplo. Una compañía de cervezas tiene dos tipos de cerveza: clara y oscura. Antes de lanzar su nueva propaganda desea saber si existen diferencias en la preferencia del

tipo de cerveza entre los hombres y las mujeres. Se hizo una encuesta a 150 personas y obtuvo la siguiente información:

Género / tipo cerveza	Clara	Negra	Total
Mujeres	55	15	**70**
Hombres	45	35	**80**
Total	**100**	**50**	**150**

Hipótesis:

H_0: las preferencias por el tipo de cerveza y el género son independientes.

H_1: las preferencias por el tipo de cerveza y el género no son independientes.

Grados de libertad: (2-1)(2-1) = 1

Valor esperado:

Género / tipo cerveza	Clara	Negra
Mujeres	(70 x 100)/150 = 46.67	(70 x 50)/150 =23.33
Hombres	(80 x 100)/150 =53.33	(80 x 50)/150 =26.67

$$\chi^2 = \frac{(55-46.67)^2}{150} + ... + \frac{(35-26.67)^2}{150} = 8.371$$

Conclusión

Para determinar la significación de $\chi^2 = 8.371$ cuando $gl = 1$, vamos a la tabla 5. La tabla muestra que este valor χ^2 es significativo más allá del nivel .01. Por lo tanto, se rechaza H_0. Podemos entonces decir que la preferencia por el tipo de cerveza depende del género.

6.6 Kolmogorov-Smirnov.

Esta prueba para una muestra se considera un procedimiento de "bondad de ajuste", es decir, permite medir el grado de concordancia existente entre la distribución de un conjunto de datos y una distribución teórica específica. Su objetivo es señalar si los datos provienen de una población que tiene la distribución teórica especificada.

Funciones

- ✓ Se interesa en el grado de acuerdo entre la distribución de un conjunto de valores de la muestra (puntajes observados) y alguna distribución teórica específica.
- ✓ Determina si razonablemente puede pensarse que los puntajes en la muestra provengan de una población que tenga una distribución teórica.

Para que se utiliza

- ✓ Compara la función de distribución acumulativa observada de una variable con una distribución teórica especificada, que puede ser Normal, Uniforme o Poisson, etc.
- ✓ Prueba de dos muestras independientes. Compara dos grupos de casos de una variable.

Cálculo

La distribución de los datos F_n para *n* observaciones y_i se define como:

$$D_i = f_t - f_i$$

Donde

f_t : frecuencias teóricas.
f_i : frecuencias observadas.
D : máxima discrepancia entre ambas.

Ejemplo. En una investigación, consistente en medir la talla de 100 niños de 5 años de edad, se desea saber si las observaciones provienen de una población Normal.

H_0: Las diferencias entre los valores observados y los teóricos de la distribución Normal de deben al azar.
H_1: Los valores observados de las frecuencias para cada clase son diferentes de las frecuencias teóricas de una distribución Normal.

Nivel de significancia: para todo valor de probabilidad menor o igual a .05 se acepta H_1 y se rechaza H_0.

Zona de rechazo: para todo valor de probabilidad mayor que 0.5 se acepta H_0 y se rechaza H_1.

Los datos son los siguientes:

Rangos estatura	f_i	f_a
90-93	7	7
94-97	20	27
98-101	45	72
102-105	22	94
106-109	6	100
Total	**100**	

Los valores de $\bar{x} + \sigma$ son 99.2 ± 2.85.

Los cálculos para *Z* son de la siguiente forma:

$$Z_{90} = \frac{x - \bar{x}}{\sigma} = \frac{90 - 99.2}{2.85} = -3.23$$

$$Z_{90} = \frac{x - \bar{x}}{\sigma} = \frac{93 - 99.2}{2.85} = -2.18$$

Y así sucesivamente.

Para cada valor de *Z,* se localiza el área de la curva tipificada de la tabla de números aleatorios. A partir de estos valores, se obtiene la diferencia entre los límites de clases entre el superior y el inferior, por ejemplo: 0.4997 - 0.4793 = 0.020, 0.4793 - 0.2357 = 0.2436, 0.2357 - (-0.2794) = 0.5151, -0.2794 - (-0.4854) = 0.206 y -0.4854 - (-0.4994) = 0.014. Estos resultados de diferencias se multiplican por el tamaño de la muestra, luego se obtienen las frecuencias y finalmente las frecuencias acumuladas.

Límites de clase	Valor *Z* de los límites	Área bajo la curva	Diferencias entre clases	Diferencias N(100)=F	Fa
90	-3.23	-0.4994			
93	-2.18	-0.4854	0.014	1.4	1.4
97	-0.77	-0.2794	0.206	20.6	22
101	0.63	0.2357	0.5151	51.5	73.5
105	2.04	0.4793	0.2436	24.4	77.9
109	3.44	0.4997	0.0200	2	99.9
Total				**100**	

Las frecuencias acumuladas teóricas y las observadas se arreglan en los rangos correspondientes y posteriormente, se aplica la fórmula de Kolmogorov-Smirnov.

Rangos	1	2	3	4	5
f_t	1.4	22	73.5	97.9	99.9
Acumulada	100	100	100	100	100
f_i	7	27	72	94	100
Acumulada	100	100	100	100	100
$f_t - f_i$	-0.056	-0.05	0.015	0.039	-0.001

La diferencia máxima D es igual a -0.049, valor que se compara con los valores críticos de D en la prueba muestral de Kolmogorov-Smirnov y se obtiene la probabilidad de la existencia de esa magnitud de acuerdo con la prueba de Kolmogorov-Smirnov. El valor ***N*** 100 es mayor y el mayor número de ***N*** en la tabla es 35, por lo que se aplica la fórmula al final de la tabla:

$$\frac{1.36}{\sqrt{100}} = 0.136$$

Lo anterior quiere decir que para cada valor menor que el crítico para una probabilidad de .05, la probabilidad correspondiente es mayor que .05.

Conclusión

En virtud de lo anterior, el estadístico Kolmogorov-Smirnov obtenido es menor que el crítico y su probabilidad mayor que 0.05, por lo tanto, no se rechaza H_0 ya que no hay evidencia para que aceptemos H_1.
Las frecuencias observadas y las teóricas acumuladas no difieren significativamente, por lo que, las observaciones tienen una distribución Normal.

6.7 Prueba de McNemar.

La prueba McNemar para la significación de los cambios es particularmente apropiada para los diseños de "antes y después" en los que cada persona es usada como su propio control y en la medida tiene la fuerza de una escala nominal y ordinal, así podrá usarse para probar la efectividad de un tratamiento particular dirigido a las preferencias de los votantes por los diferentes candidatos por ejemplo. O podrá usarse

para probar, por ejemplo, los desplazamientos del campo a la cuidad en la afiliación política de la gente.

La siguiente tabla muestra las diferentes respuestas que podría tener una persona:

		Después	
		-	+
Antes	+	A	B
	-	C	D

Una persona es clasificada en la celda A si cambió de + a -, es clasificado en la celda D si cambió de – a +. Si no hay cambio va a la celda B o C.

Cálculo

$$\chi^2 = \frac{\left(|A - D| - 1\right)^2}{A + D}$$

con $gl = 1$

Ejemplo. Se realizó una prueba de ejercicios físicos a un grupo de estudiantes de una primaria como parte de un proceso para una mejora escolar. Los resultados fueron los siguientes:

		Después	
		-	+
Antes	+	4	3
	-	5	13

Entonces:

H_0 : para cualquier estudiante antes de la prueba (P_A) o después de la misma (P_D) es igual a ½, es decir, $P_A = P_D = 1/2$.

$H_1 : P_A > P_D$

Nivel de significancia: $\alpha = .05$, $N = 25$.

Distribución muestral: se aproxima fuertemente a una χ^2 con $gl = 1$.

Región de rechazo: Ya que H_1 predice la dirección de la diferencia, la región de rechazo es de una cola. La región de rechazo está compuesta por todos los valores de χ^2 (calculados con datos en los que A>D) tan grandes que tienen una probabilidad de una cola asociada con su recurrencia conforme H_0 de .05 o menos.

De acuerdo a los datos:

$$\chi^2 = \frac{\left(|4 - 13| - 1\right)^2}{4 + 13} = 3.76$$

$p = .05$

Conclusión

Rechazamos H_0 con $\alpha = .05$. Con esto concluimos que los estudiantes muestran un mejor rendimiento escolar incluyendo una rutina de ejercicios.

6.8 ANOVA (Análisis de varianza) de un factor.

El análisis de la varianza (Anova) se debe al estadístico-genético Sir Ronald Aylmer Isher (1890-1962), autor del libro "Statistics Methods for Research Workers" publicado en 1925 y pionero de la aplicación de métodos estadísticos en el diseño de experimentos, introduciendo el concepto de aleatorización.

Para que se utiliza

- ✓ Para encontrar diferencias significativas entre medias, es decir, cuando se desea comparar el desempeño de varios aspectos bajo observación.
- ✓ Identifica si el comportamiento de dichos aspectos es o no es estadísticamente significativo.
- ✓ Para realizar regresiones.

Cálculo

$$Y_{ij} = \overline{Y} + \overline{Y}_j + \varepsilon_{ij}$$

Donde

Y_{ij}**:** es la *i*-ésima observación del *j*-ésimo tratamiento.

$\overline{Y}$**:** es la media total.

$\overline{Y}_j$**:** es el efecto sobre la respuesta debido al *j*-ésimo tratamiento.

ε_{ij}**:** es el error experimental para la *i*-ésima observación del *j*-ésimo tratamiento.

El ANOVA tradicional parte de:

Variación total = Variación entre + Variación intra

Lo anterior podemos expresarlo:

$$\underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y})^2}_{\text{Var. Total SCT}} = \underbrace{\sum_{j=1}^{g} n_j(\overline{Y}_j-\overline{Y})^2}_{\text{Var. Entre SCE}} + \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{n_j}(Y_{ij}-\overline{Y}_j)^2}_{\text{Var. Intra SCD}}$$

Siendo consecuentes con la expresión $Y_{ij} = \overline{Y} + \overline{Y}_j + \varepsilon_{ij}$ tenemos:

$Y_j = \overline{Y}_j - \overline{Y}$ y $\varepsilon_{ij} = Y_{ij} - \overline{Y}_j$

Podemos expresar:
SCT = suma cuadrática total.

SCE = Suma cuadrática del tratamiento.
SCD = Suma cuadrática del error.

Los grados de libertad entre grupos (GLE), dentro de los grupos (GLD) y total (GLT) se calculan de la siguiente manera:

GLE = r – 1
GLD = n – r
GLT = n – 1

El cuadrado medio entre grupos (CME) y el cuadrado medio dentro de grupos (CMD) se calculan de la siguiente manera:

CME = SCE / GLE
CMD = SCD / GLD

El estadístico de contraste para realizar la prueba ANOVA se construye de la siguiente forma:

F = CME / CMD

Que se distribuye según una F-Snedecor con GLE grados de libertad del numerador y GLD grados de libertad en el denominador.

Tabla ANOVA

	Suma de cuadrados	G.L.	Cuadrado medio	F-valor	p-valor
Entre grupos	SCE	GLE	CME	F	P
Dentro de grupos	SCD	GLD	CMD		
Total	SCT	GLT			

Ejemplo. Se tienen los siguientes datos, donde se ha recogido información de dos variables, la variable explicativa "estatus" es nominal y la variable "respuesta" es cuantitativa:

Respuesta	Estatus	Respuesta	Estatus
125	1	115	2
145	1	169	2
156	1	184	2
158	1	163	3
198	1	128	3
153	1	144	3
147	1	111	3
123	1	155	3
129	2	198	3
179	2	137	3

$$SCE = 8(150.62 - 150.85)^2 + 5(155.2 - 150.85)^2 + 7(148 - 10.85)^2 = 151.875$$
$$SCD = 467633 - 455266 = 12367$$
SCT=151.875 + 12367 = 12518.55

$GLE = r - 1 = 3 - 1 = 2$
$GLD = n - r = 20 - 3 = 17$
$GLT = n - 1 = 20 - 1 = 19$

$CME = SCE / GLE = 75.94$
$CMD = SCD / GLD = 727.45$
$F = CME / CMD = .1044$

$R^2 = SCE / SCT = 1.21\%$

Conclusión

Con lo que se tendría, que el modelo ANOVA o más específicamente, la variable que forma los grupos, explica el 1.21% de la variabilidad de la variable respuesta.

Técnicas multivariables.

Los procedimientos modernos se basan en técnicas multivariadas, que involucran muchas variables y permiten analizar la información proveniente de grandes volúmenes de datos. Estas técnicas nos permiten obtener conclusiones más precisas, objetivas e interesantes.

6.9 Análisis de Regresión lineal simple.

El caso más sencillo de análisis de Regresión es la situación donde una variable depende sólo de otra variable única.

La Regresión en la Investigación de mercados se puede ver afectada si se limita el uso a variables independientes que se representen en escalas de intervalos. Es aquí cuando se hace uso de la Regresión logística donde se pueden utilizar las variables independientes nominales en términos de Regresión. Este tipo de Regresión convierte las variables nominales en variables binarias que se codifican cero-uno (dummy). Por ejemplo, en el caso de promociones bancarias del tipo de pagos sin intereses se podría medir la respuesta de los clientes con un cero para aquellos que no usaron la oferta y un uno para los que si la usaron.

Funciones

- ✓ Identifica basándose en los datos, la existencia de una ecuación matemática que represente el comportamiento de la variable de respuesta sobre la base de los cambios en las variables de predicción.
- ✓ Identifica una asociación entre la variable de respuesta y las variables de predicción (más no detecta una relación causa-efecto entre las variables).
- ✓ Permite realizar predicciones mediante el uso de la ecuación que se obtiene, modificando variables para cada variable de predicción.

Para que se utiliza

- ✓ Para ajustar alguna función a un conjunto de datos.
- ✓ Para predecir el valor o comportamiento de la variable de respuesta dados los valores de las variables de predicción.
- ✓ Para establecer asociaciones entre las variables de interés en las cuales la relación usual no es casual.

Cálculo

Para encontrar la B_0 y B_1 que generen la "mejor" recta se aplicarán las siguientes fórmulas:

$$B_1 = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

y

$$B_0 = \bar{Y} - B\bar{X}$$

Ecuación de Regresión: $Y' = B_0 + B_1 X$

Donde

B_0: El valor que Y asumiría si X fuera cero, esto es, la constante o intersección.

B_1: Se interpreta como la cantidad Y que se aumentaría si X se incrementara en una unidad, llamaremos coeficiente de Regresión o pendiente. Un coeficiente de Regresión negativo significa que a medida que X se incrementa, Y disminuye su valor.

Ejemplo. Una empresa está entrevistando y seleccionando nuevos vendedores. La empresa ha diseñado una prueba que ayudará a realizar la mejor elección posible de su Fuerza de ventas. Con el fin de probar la validez para predecir las ventas semanales, se eligieron vendedores experimentados y se aplicó la prueba a cada uno, la calificación de cada vendedor de acuerdo a sus ventas es la siguiente:

Vendedor	Calificación X	Ventas semanales Y	X^2	$X*Y$	Y^2
1	5	6	25	30	36
2	9	12	81	108	144
3	6	4	36	24	16
4	8	9	64	72	81
5	10	11	100	110	121
Media	**7.6**	**8.4**	**61.2**		**79.6**
Suma	**38**	**42**	**306**	**344**	**398**

$$B_1 = \frac{344 - 5(7.6)(8.4)}{306 - 5(57.76)} = 1.44$$

$$B_0 = 8.4 - 1.44(7.6) = -2.50$$

Para predecir las ventas semanales de un aspirante a vendedor que obtuvo una calificación de 9 en la prueba por ejemplo, se aplica la ecuación de Regresión:

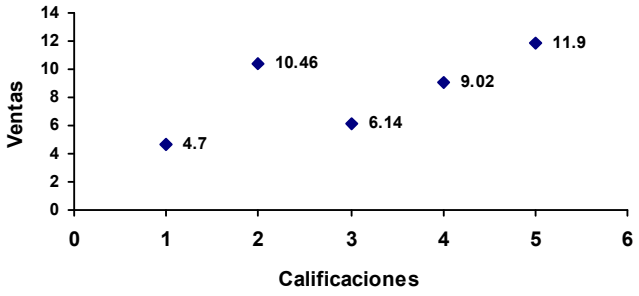
$$Y' = -2.50 + 1.44(9) = 10.46$$

Para determinar los puntos de la línea de Regresión se sustituyen los valores de la variable independiente de la ecuación de Regresión:

Vendedor	Calificación	Solución	Predicción de Ventas

1	5	Y'= -2.50+1.44(5)	4.7
2	9	Y'= -2.50+1.44(9)	10.46
3	6	Y'= -2.50+1.44(6)	6.14
4	8	Y'= -2.50+1.44(8)	9.02
5	10	Y'= -2.50+1.44(10)	11.9

Gráfico de dispersión



La predicción perfecta es prácticamente imposible, por lo tanto es necesaria una medida que indique que tan precisa es un predicción de *Y* basada en *X*. Esta medida es el error estándar. El error estándar es el mismo concepto de desviación estándar, que mide la dispersión alrededor de la media, el error estándar mide entonces la dispersión alrededor de la media de dispersión. Su fórmula es la siguiente:

$$S_{X,Y} = \sqrt{\frac{\sum Y^2 - B_0 \sum Y - B_1 \sum XY}{n-2}}$$

Siguiendo nuestro ejemplo:

$$S_{X,Y} = \sqrt{\frac{398 + 2.50(42) - 1.44(344)}{5-2}} = 1.78$$

Intervalos de predicción.

Existen dos razones para construir una Regresión lineal. Una es predecir los valores de respuesta de la variable dependiente Y' a un valor de la variable independiente *X*.

Hay dos tipos de intervalo de predicción:

1. El intervalo de predicción del valor medio de *Y* para un valor dado de *X*.
2. El intervalo de predicción del valor individual *Y* para un valor dado de *X*.

Para determinar el intervalo de confianza del valor medio de *Y* que se simboliza μ_y para un valor dedo de *X,* la fórmula es:

$$\mu_y = Y' \pm tS_{Y,X} \sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum X^2 - \frac{\sum (X)^2}{n}}}$$

Considerando una variable dependiente y una variable independiente, la media de los errores igual a cero, la varianza de los errores constantes, la covarianza de los errores igual a cero y la esperanza matemática de la media de los errores igual a cero. Los errores independientes unos de los otros.

Donde:

Y' : es el valor de la predicción Y para un valor dado de X.

t : es el valor de t para $\alpha/2$ y $\phi = n-2$.

Estamos suponiendo un intervalo de confianza del 95% (para tamaños de muestra grandes).

Ejemplo. Siguiendo con el ejemplo, calcularemos el intervalo de predicción para el valor medio de Y del 95% para un valor dado de X=8.

$$\mu_y = 9.02 \pm 3.18(1.78)\sqrt{\frac{1}{5} + \frac{(8-7.6)^2}{306 - \frac{1444}{5}}}$$

$$\mu_y = 9.02 \pm 2.59$$

$$P = (6.43 \le \mu_y \le 11.61) = .95$$

Entonces, para un grupo de aspirantes que obtuvieron calificaciones exactamente de 8, hay una probabilidad del 95% que sus ventas semanales promedio se localicen entre 6.43 y 11.61.

6.10 Coeficiente de Correlación.

Probablemente el método más popular para resumir rápidamente el grado de relación entre dos variables lo constituye el coeficiente de Correlación.

Es una técnica estadística que mide la asociación entre variables, indica como varían en conjunto.

Funciones

- ✓ Mide la magnitud de la relación entre las variables.
- ✓ Especifica si la Correlación es **positiva o negativa**.
- ✓ Provee una medida del monto de varianza compartida, es decir, si una variable cambia (ya sea que aumente o disminuya su valor) el coeficiente de Correlación indica el cambio de la otra variable que se esté estudiando (ya sea que también aumente o disminuya su valor).
- ✓ Sugiere, (más **no implica** una relación casual) entre variables.

Para que se utiliza

- ✓ Seleccionar atributos claves para la satisfacción de los consumidores (y otras investigaciones).

- ✓ Identificar aquellas variables correlacionadas más fuertemente.
- ✓ Como dato de entrada para otro tipo de estudios como la Regresión múltiple y otras técnicas multivariadas.
- ✓ Comparar cambios en la magnitud de las relaciones entre las variables en el tiempo o en grupos diferentes.
- ✓ Estimar relaciones divariadas.

Supuestos teóricos de la Correlación

- ✓ Las variables son distribuidas como una Normal bivariada.
- ✓ Los casos representan una muestra aleatoria de la población y las calificaciones de cada una de las variables son consideradas independientes de las demás variables.

Cálculo

Se necesita calcular la media o promedio para cada variable, una vez calculada se aplica la siguiente fórmula:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

Donde

X y *Y*: Posibles variables correlacionadas o asociadas.

$\bar{X}$: Media o promedio de la variable *X*.
$\bar{Y}$: Media o promedio de la variable *Y*.
∑: Indica que se debe realizar la suma de todos los valores de las variables.

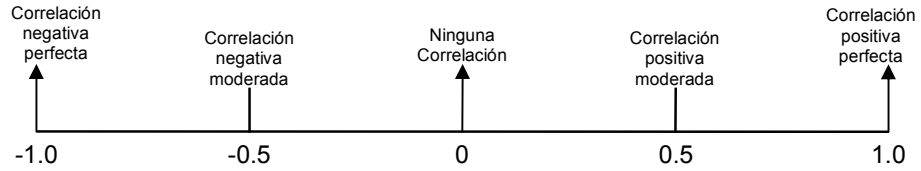
r = covarianza atributo 1 y 2 / (varianza atributo 1)(varianza atributo 2)

En otras palabras:
r = como varían los atributos 1 y 2 / como varía cada atributo de manera independiente

El valor de *r* se aproxima a +1 cuando la Correlación tiende a ser lineal directa (mayores valores de *X* significan mayores valores de *Y*), y se aproxima a –1 cuando la Correlación tiende a ser lineal inversa. Es importante notar que la existencia de Correlación entre variables no implica causalidad.

¡Atención!: si no hay Correlación de ningún tipo entre dos v.a., entonces tampoco habrá Correlación lineal, por lo que $r = 0$. Sin embargo, el que ocurra $r = 0$ sólo nos dice que no hay Correlación lineal, pero puede que la haya de otro tipo.

El siguiente diagrama resume el análisis del coeficiente de Correlación entre dos variables:



Ejemplo. Especialistas sugieren que los individuos que leen demasiado (casi todos los días) tienden a ser muy buenos lectores, con velocidad en la lectura y buena dicción, investigadores desean probar esta afirmación y al mismo tiempo obtener un índice de la magnitud de la relación que pueda existir. No hay ninguna razón para creer que existe una relación causa-efecto, únicamente existe un interés en obtener un grado en el cual las mediciones de la habilidad en la lectura y el tiempo dedicado a la lectura varían juntas en una población dada.

Dos variables aparecen en el estudio:

X : puntaje para la prueba de lectura.
Y : puntaje para el número de horas que la gente invierte en la lectura.

Individuo	X	Y	Dif. Con la media X	Dif. Con la media Y	$X*Y$	X^2	Y^2
1	20	5	0	0	0	0	0
2	5	1	-15	-4	60	225	16
3	5	2	-15	-3	45	225	9
4	40	7	20	2	40	400	4
5	30	8	10	3	30	100	9
6	35	9	15	4	60	225	16
7	5	3	-15	-2	30	225	4
8	5	2	-15	-3	45	225	9
9	15	5	-5	0	0	25	0
10	40	8	20	3	60	400	9
Promedio	**20**	**5**			**$\sum XY$=370**	**$\sum X^2$=2050**	**$\sum Y^2$=76**

El coeficiente de Correlación se define como:

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

$$r = \frac{370}{\sqrt{(2050)(79)}}$$

$$r = 0.937$$

Conclusión

Dado que el valor obtenido del coeficiente de Correlación es positivo muy alto, se concluye que: existe una gran asociación entre leer con rapidez y dicción y dedicar varias horas a la lectura, además el coeficiente positivo indica que si se dedican más horas a la lectura con el tiempo se incrementará la rapidez en la lectura y dicción al leer (el valor de 0.937 es un índice muy cercano a 1, la asociación es muy alta).

Prueba de hipótesis

El cálculo del coeficiente de Correlación *r* supone que las variables, cuya relación está siendo probada, son métricas. Si esta suposición no se cumple, parcial o completamente, afecta el valor de la *r*. Se puede realizar una prueba de hipótesis sencilla para verificar la significancia de la relación entre dos variables, medidas por *r*. Esto implica probar:

La hipótesis nula: $H_0 : \theta = 0$ vs

La hipótesis alterna: $H_1 : \theta \neq 0$

Consideramos el ejemplo anterior. Para probar la significancia de esta relación, la estadística de prueba *t* puede calcularse usando:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

En nuestro ejemplo, n=10 y r=0.937, por tanto:

$$t = 0.937\sqrt{\frac{10-2}{1-0.937^2}} = 7.59$$

Si la prueba se hace a un valor $\alpha = .05$ con $gl = 8$, entonces el valor crítico de *t* se puede obtener de la tabla 6 como 1.860. Debido a que 7.59>1.860, podemos rechazar la hipótesis nula.
¿Qué significa esto? La prueba estadística de significancia revela que el valor de la Correlación *r* de la muestra (que es .937) es significativamente diferente de cero. En otras palabras, existe una asociación sistemática entre las variables.

6.11 Análisis Discriminante.

El análisis Discriminante es un análisis estadístico multivariante desarrollado por Fisher en 1930. Es una técnica estadística de clasificación que permite identificar aquellas variables que discriminan entre dos o más grupos definidos con anterioridad. Por lo tanto, permite establecer diferencias entre dichos grupos.

Las técnicas del análisis Discriminante se utilizan para clasificar individuos en uno de dos o más grupos alternativos (o poblaciones) con base en un conjunto de mediciones.

Se sabe que las poblaciones son distintas y cada individuo pertenece a un grupo de ellas, esta técnica también puede usarse para identificar las variables que contribuyen a hacer la clasificación. Por tanto, la predicción y la descripción, como en el caso del análisis de Regresión, son los principales usos del análisis Discriminante.

Funciones

- ✓ Ayuda a comprender las diferencias entre grupos. Explica, en función de características métricas observadas por qué los objetos/sujetos se encuentran asociados a distintos niveles de un factor.

Para que se utiliza

- ✓ Distinción entre usuarios frecuentes de cierto producto y usuarios poco frecuentes: hombres de mujeres, por ejemplo.
- ✓ Compradores de marcas nacionales y compradores de marcas privadas.
- ✓ Predecir cuando una campaña va a ser exitosa y cuando no.

Cálculo

$$f = w_1 x_1 + w_2 x_2 + ... + w_k x_k$$

Donde

x_k : las características medidas.
f: el índice (la función discriminatoria).
w_k : la ponderación (coeficiente discriminatorio) de la característica *i* al discriminar entre los dos grupos.

Los coeficientes w_k se eligen de tal manera que se consiga la máxima separación entre los grupos existentes, es decir, tratando de que los valores que toman estas funciones discriminantes en los *k* grupos sean lo más diferentes posibles. Estadísticamente este criterio equivale a maximizar la varianza "entre grupos" frente a la varianza "dentro de grupos". Por tanto, los coeficientes w_k se elegirán de tal forma que se consiga maximizar el valor del coeficiente:

$$\lambda = \frac{\text{var}\,ianza \quad entre \quad grupos}{\text{var}\,ianza \quad dentro \quad de \quad grupos} = \frac{\sum n_j \left(\bar{f}_j - f\right)^2}{\sum\sum \left(f_j - \bar{f}_j\right)^2}$$

Los coeficientes w_k se determinarán siguiendo el procedimiento análogo de los coeficientes de Regresión:

$$w_1 \sum \left(x_1 - \bar{x}_1\right)^2 + w_2 \sum \left(x_1 - \bar{x}_1\right)\left(x_2 - \bar{x}_2\right) = \bar{x}_{1s} - \bar{x}_{1n}$$
$$w_1 \sum \left(x_1 - \bar{x}_1\right)\left(x_2 - \bar{x}_2\right) + w_2 \sum \left(x_2 - \bar{x}_2\right)^2 = \bar{x}_{2s} - \bar{x}_{2n}$$

Los x_j son los centroides por grupo y $\bar{x}_{is}$ y $\bar{x}_{in}$ son los centroides de los casos favorables y no favorables respectivamente.

Para determinar cuántas funciones discriminantes son significativas existen varios criterios, entre los más utilizados tenemos:

- **Porcentaje relativo**. Este criterio compara entre sí las funciones discriminantes cuantificando, en términos relativos, el poder discriminatorio de cada una de ellas con respecto al total, es decir, calculando el porcentaje del poder

discriminante correspondiente a cada una de ellas sobre el total acumulado por todas las funciones.

- **Coeficiente de Correlación canónica *eta* (η).** Este coeficiente mide, para cada función discriminante *Y*, el grado en que difieren las medias de dicha función en los distintos grupos. Un valor alto del coeficiente *eta* indica una fuerte relación entre el grupo de pertenencia y los valores de la función discriminante correspondiente.
- **Estadístico Lambda de Wilks (λ).** El estadístico λ de Wilks es también una medida de las diferencias entre los grupos debidas a las funciones discriminantes, que se utiliza, no tanto para contrastar la significación de una función concreta, sino para medir de forma secuencial el poder discriminatorio de cada una de las funciones que se van construyendo, empezando siempre por la primera, que es la de mayor capacidad discriminatoria.

Una vez obtenidas las funciones discriminantes, nuestro objetivo es establecer la contribución relativa de las distintas variables a la discriminación, o lo que es lo mismo, determinar cuáles son las variables que más contribuyen a discriminar entre un grupo y otro.

Ejemplo. Se desea realizar una clasificación entre usuarios y no usuarios de un nuevo producto, considerando precio y calidad en una de dos categorías: gustar vs disgustar. La información es la siguiente:

Persona	Evaluación	Precio (x_1)	Calidad (x_1)	$(x_1)^2$	$(x_2)^2$	$x_1 x_2$
1	Disgusto	2	4	4	16	8
2	Disgusto	3	2	9	4	6
3	Disgusto	4	5	16	25	20
4	Disgusto	5	4	25	16	20
5	Disgusto	6	7	36	49	42
Media		**4**	**4.4**			
6	Gusto	7	6	49	36	42
7	Gusto	8	4	64	16	32
8	Gusto	9	7	81	49	63
9	Gusto	10	6	100	36	60
10	Gusto	11	9	121	81	99
Media		**9**	**6.4**			
Media total		**6.5**	**5.4**			
Desv. Est.		**3.028**	**2.011**			

Calcular los pesos discriminantes:

Precio: $\bar{x}_1 - \bar{x}_1$ = gustan – disgustan= 9 – 4 = 5

Calidad $\bar{x}_2 - \bar{x}_2$ = disgustan – gustan = 6.4 – 4.4 = 2

Medias corregidas de sumas de cuadrados y productos cruzados:

	Disgustan	Gustan	Total
$\sum(x_1)^2$	$90 - 5(4)^2 = 10$	$415 - 5(9)^2 = 10$	20
$\sum(x_2)^2$	$110 - 5(4.4)^2 = 13.2$	$218 - 5(6.4)^2 = 13.2$	26.4
$\sum x_1 x_2$	96 – 5(4)(4.4) = 8	296 – 5(9)(6.4) = 8	16

Donde:

$$\sum x_1^2 = \sum x_1^2 - n\bar{x}_1^2$$

$$\sum x_2^2 = \sum x_2^2 - n\bar{x}_2^2$$

$$\sum x_1 x_2 = \sum x_1 x_2 - n\bar{x}_1\bar{x}_2$$

Para hallar w_1 y w_2 resolvemos las ecuaciones simultáneas:

$$\sum x_1^2 w_1 + \sum x_1 x_2 w_1 = \bar{x}_1 \text{(gustan)} - \bar{x}_1 \text{(disgustan)}$$
$$\sum x_1 x_2 w_1 + \sum x_2^2 w_2 = \bar{x}_2 \text{(gustan)} - \bar{x}_2 \text{(disgustan)}$$

Tenemos entonces:

$$20\, w_1 + 16\, w_2 = 5$$
$$16\, w_1 + 26.4\, w_2 = 2$$

$w_1 = 0.368$ y $w_2 = -.147$

Función discriminante: $f = 0.368\, x_1 - 0.147\, x_2$
Podemos entonces hallar los puntajes discriminatorios para la tabla de casos, así por ejemplo tenemos: $f = 0.368(2) - 0.147(4) = .0148$

También hallaremos los puntajes discriminantes para los centroides de los dos grupos y la media total:
$\bar{f}$ (disgustan) = 0.368(4) – 0.147(4.4) = 0.824
$\bar{f}$ (gustan) = 0.368 (9) – 0.147(6.4) = 2.368
$\bar{f}$ (media total) = 0.368(6.5) – 0.147(5.4) = 1.596

Gustan		Disgustan	
Persona	**Puntaje discriminante**	**Persona**	**Puntaje discriminante**
1	0.148	6	1.691
2	0.809	7	2.353
3	0.735	8	2.279
4	1.250	9	2.794
5	1.176	10	2.721
Media	**0.824**		**2.368**
Media total	1.596		

Variabilidad entre grupos:
$5(0.824 - 1.596)^2 + 5(2.368 - 1.596)^2 = 5.96$

Variabilidad en grupos:
Disgustan $(0.148 - 0.824)^2 + (0.809 - 0.824)^2 + ... + (1.176 - 0.824)^2 = 0.772$
Gustan $(1.691 - 2.368)^2 + (2.353 - 2.368)^2 + ... + (2.721 - 2.368)^2 = 0.772$
Total: 1.544

Criterio discriminante: $C = \dfrac{5.96}{1.544} = 3.86$

La variabilidad entre grupos se calcula:
$5(4 - 6.5)^2 + 5(9 - 6.5)^2 = 62.5$

La variabilidad conjunta dentro de grupos sobre x_1 se calcula:
$\sum x_1^2$ (disgustan) = 10

$\sum x_1^2$ (gustan) = 10

El criterio discriminante en este caso es: $C = \frac{62.5}{20} = 1.125$

Entonces la función óptima es: $f = 0.368x_1 - 0.147x_2$. Vemos que x_2 está altamente correlacionado con x_1 y sirve de supresora, dando a x_2 un peso negativo en la función discriminante en la cual x_1 tiene un peso positivo, la predictibilidad del criterio se aumenta aún más.

Puesto que las variables x_1 y x_2 fueron expresadas en unidades diferentes y exhiben desviaciones estándar diferentes, el análisis generalmente normaliza los pesos discriminantes antes de examinar su importancia relativa.

Consideraremos dos métodos de estandarización:

Método 1. Multiplica cada peso discriminante por la desviación estándar de la muestra total de esa variable.
Las desviaciones estándar de la muestra total son $s_{x_1} = 3.028$ y $s_{x_2} = 2.011$, los pesos discriminantes estandarizados son:

$$k_1^{s(t)} = 3.028(0.368) = 2.240$$
$$k_2^{s(t)} = 2.011(-0.147) = -0.296$$

Método 2. Multiplica el peso por la desviación estándar conjunta dentro de grupos de esa variable.

$$s_{x_1}^{(z)} = \sqrt{\frac{\sum x_1^2}{8}} = \sqrt{\frac{20}{8}} = 1.581$$
$$s_{x_2}^{(z)} = \sqrt{\frac{\sum x_2^2}{8}} = \sqrt{\frac{26.4}{8}} = 1.817$$

Donde 8 = n – 2 grados de libertad para los grupos conjuntos (donde n denota el tamaño de la muestra total). Los pesos estandarizados son:

$$k_1^{s(z)} = 1.581(0.368) = 0.582$$
$$k_2^{s(z)} = 1.817(-0.147) = -0.267$$

Aunque los dos métodos producen valores numéricos diferentes, el ordenamiento concuerda con el ordenamiento original.

Conclusión

Tenemos entonces que x_1 (precio) es "más importante" que x_2 (calidad).

6.12 Análisis Conjunto.

El análisis Conjunto es una técnica multivariante utilizada específicamente para entender cómo los encuestados desarrollan preferencias para productos o servicios. Se basa en la sencilla premisa de que los encuestados evalúan el valor o utilidad de un producto/servicio/idea (real o hipotética) procedente de la combinación de las cantidades separadas de utilidad suministradas por cada atributo.

El análisis Conjunto supone que tanto los atributos como las posiciones de las alternativas en los atributos se conocen. El procedimiento, luego, intenta atribuir valores a cada uno de los diversos atributos. Los datos de entrada constituyen normalmente evaluaciones de diversas combinaciones de niveles de atributos y el resultado representa la utilidad en cada uno de los diferentes niveles de diversos atributos.

Funciones

- ✓ Mide los intercambios que los consumidores realizan en los atributos de los productos. El análisis Conjunto principia con un orden de rangos de las preferencias de productos y luego calcula los valores de utilidad para las características centrales que describen al producto.
- ✓ Lo que pretende es encontrar un conjunto de utilidades que expliquen el orden en que se clasificaron los productos.

Para que se utiliza

- ✓ Lanzamiento de nuevos productos.
- ✓ Preferencias de un producto en cuanto a precio, utilidad, desempeño.

Cálculo

El cálculo parte de una asignación arbitraria de puntajes iniciales, que se suman para calcular la utilidad total y derivar un orden jerárquico de acuerdo a esas asignaciones. El principio consiste en correlacionar esta jerarquización con la obtenida directamente de los participantes en las evaluaciones. Si la Correlación es baja, se corrigen los puntajes iniciales de los atributos y se vuelven a calcular los coeficientes y así se procede sucesivamente, sobre una base de puntajes, derivando cada vez un nuevo orden y confrontándolo con le real. El procedimiento se repite iterativamente en forma de tanteo y error, hasta que se llegue a una Correlación aceptable.

- Mínimos cuadrados ordinarios. El modelo a estimar mediante la metodología de mínimos cuadrados ordinarios, en función de los atributos y de los niveles, es el siguiente:

$$y_t = \alpha + \sum_{i=1}^{I} \sum_{j=1}^{k_i} \beta_{ij} x_{ij} + e_t$$

Donde:

y_t : es el orden o la valoración de la preferencia sobre el estímulo *t.*

α : es el término constante.

β_{ij}: es la utilidad o partworth asociado al nivel *j*-ésimo $j = 1,2,...k_i$ del atributo *i*-ésimo $i = 1,2,...I$.

$x_{ij} = 1$: si el nivel j-ésimo del atributo i-ésimo está presente en el estímulo *t.*

$x_{ij} = 0$: si el nivel j-ésimo del atributo i-ésimo no está presente en el estímulo *t.*

La base de la interpretación de resultados es el vector de utilidades o partworths β. Un valor alto de partworth significa que el nivel asociado proporciona al entrevistado una utilidad alta, mientras que un partworth bajo significa que el nivel asociado proporciona una utilidad baja.

A partir de los partworths, el análisis Conjunto también calcula la importancia relativa que los individuos atribuyen a los diferentes atributos que componen el producto. Un atributo será más importante cuanto más grande sea la diferencia entre el partworth más elevado y el más bajo (en valores absolutos), por tanto, se puede obtener la importancia de un atributo mediante:

$$imp_i = \left| máx(\beta_{ij}) - mín(\beta_{ij}) \right| \quad \forall i = 1,...,I \qquad \forall j = 1,...,k_i$$

Para poder comparar la importancia de cada factor se utiliza la importancia relativa:

$$Rimp_i = \frac{imp_i}{\sum_{i=1}^{I} imp_i} 100$$

- Logit ordenado. El modelo Logit ordenado de respuesta múltiple se relaciona a una variable Y_i con las variables $X_{11},...X_{ki}$ a través de la siguiente ecuación:

$$Y_i^* = F(X_i\beta) + e_i$$

Donde:

Y_i^*: es una variable latente que cuantifica las distintas categorías.

$X_i\beta$: es una combinación lineal de las variables o características independientes.

e_i: es una variable aleatoria.

La relación existente entre la variable real u observada, Y_i, y la variable latente o no observada, Y_i^*, es la siguiente:

$$Y_i = \begin{cases} 0 & si \quad Y_i^* \le c_1 \\ 1 & si \quad c_1 \le Y_i^* \le c_2 \\ 2 & si \quad c_2 \le Y_i^* \end{cases}$$

La probabilidad de elegir cada una de las categorías de la variable Y_i viene definida por la siguiente relación:

$$P(Y_i = 0 / X_i, \beta, c) = F(c_1 - X_i\beta)$$
$$P(Y_i = 1 / X_i, \beta, c) = F(c_2 - X_i\beta) - F(c_1 - X_i\beta)$$
$$P(Y_i = 2 / X_i, \beta, c) = 1 - F(c_2 - X_i\beta)$$

donde la función *F* es la función de distribución logística.

- Probit ordenado. El modelo Probit ordenado sigue la misma estructura que el modelo Logit ordenado, con la salvedad de que la función *F* es la función de distribución Normal. En nuestro estudio, tanto en el modelo Logit ordenado como en el modelo Probit ordenado hemos procedido a clasificar la variable dependiente con los valores 0, 1 y 2. Esta clasificación la hemos hecho a partir de las siguientes consideraciones:

Ordenación jerárquica Ordenación divisiva Rangos	Ratio 1 a 4	Ratio 1 a 7	Ratio 1 a 11
0 si rango ≥ 13 1 si 7 ≤ rango ≤ 12 2 si rango ≤ 6	0 si puntuación = 1 1 si 2 ≤ puntuación ≤ 3 2 si puntuación = 4	0 si puntuación ≤ 2 1 si 3 ≤ puntuación ≤ 5 2 si puntuación ≥ 6	0 si puntuación ≤ 4 1 si 5≤puntuación ≤ 7 2 si puntuación ≥ 8

- Tobit doblemente censurado. El modelo Tobit se deriva de un clásico modelo de Regresión lineal. Una variable se dice que está censurada cuando todos sus valores en un cierto rango son sustituidos por un valor fijo. En su formulación se emplea una variable latente, Y^*, de la siguiente forma:

$$Y_i^* = X_i\beta + e_i$$

Esta variable latente o no observable está relacionada con una variable real u observable, *Y*, de la siguiente manera:

$$Y_i \begin{cases} a_1 & si \quad Y_i^* \le a_1 \\ Y_i^* & si \quad a_1 < Y_i^* < a_2 \\ a_2 & si \quad Y_i^* \ge a_2 \end{cases}$$

En nuestro estudio los límites de censura que hemos establecido son los siguientes:

Límites de censura en el modelo Tobit doblemente censurado

Ordenación jerárquica Ordenación divisiva Rangos	Ratio 1 a 4	Ratio 1 a 7	Ratio 1 a 11
2 y 15	2 y 4	2 y 6	2 y 10

La justificación de establecer estos límites de censura en nuestra investigación radica en que los estímulos ordenados o puntuados con valores altos o valores bajos tienen una baja influencia en las preferencias de los entrevistados ya que se consideran "estímulos obvios", es decir, estímulos que todos o la mayoría de los individuos los prefieren o estímulos que todos o la mayoría de los individuos no los prefieren.

Ejemplo. Se estudia la implementación de una tarjeta de crédito empresarial para satisfacer las necesidades crediticias de una industria. El objeto del estudio es conocer la relevancia de los atributos individuales sobre la base de la forma actual que presenta tres

configuraciones A, B y C. Hay tres atributos claves: tasa de descuento en las compras, el tiempo para el reintegro del dinero y el medio para checar la vigencia de la tarjeta.

La investigación presentó los siguientes resultados:

B = 180
A = 70
C = 50

Atributos	Tasa de descuento	Tiempo de reintegro	Medio de chequeo
Niveles	10%	15 días	Teléfono
	7%	7 días	Computador

	Conjunto *A*	Conjunto *B*	Conjunto *C*
Tasa	10%	7%	10%
Tiempo	7 días	15 días	15 días
Medio	Teléfono	Computador	Teléfono

Los resultados de acuerdo al orden jerárquico suministrado por los consultantes fueron los siguientes:

Atributos	Tasa de descuento		Tiempo de reintegro		Medio de chequeo	
	Nivel	Utilidad	Nivel	Utilidad	Nivel	Utilidad
Niveles	10%	0.3	15 días	0.4	Teléfono	0.1
	8%	0.5	12 días	0.6	Computador	0.2
	7%	0.9	7 días	0.8	Ninguno	0.3

Puntajes de configuraciones:

Atributos	Tasa de descuento	Tiempo de reintegro	Medio de chequeo	Total
Conjunto A	0.3	0.7	0.1	1.1
Conjunto B	0.9	0.4	0.2	1.5
Conjunto C	0.3	0.4	0.3	1.0
Conjunto D	0.5	0.6	0.3	1.4

Conclusión

Es posible diseñar un nuevo producto con base en una configuración basada en niveles para los atributos, diferentes a los que actualmente se ofrecen en el mercado.

6.13 Análisis de Factores (Factor Análisis).

El propósito del análisis de Factores consiste en agrupar variables que estén altamente relacionadas, el propósito fundamental es la simplificación, esto es, encontrar un modo de condensar la información contenida en un número de variables originales en un conjunto más pequeño de variables (factores) con una pérdida mínima de información.

Funciones

- ✓ Puede ser usada para analizar interrelaciones a través de un gran número de variables en escala cuantitativa en términos de sus dimensiones en común (factores).
- ✓ Trata de encontrar una manera de condensar la información contenida en un número de variables originales dentro de un pequeño conjunto de dimensiones (factores) con una mínima pérdida de información.

Para que se utiliza

- ✓ La reducción de datos.
- ✓ La identificación de estructuras.
- ✓ La transformación de datos.
- ✓ Desarrollo de escalas de personalidad.
- ✓ Identificación de atributos de productos clave.
- ✓ Segmentos de mercado basados en datos psicográficos.
- ✓ Similitud entre productos.

Cálculo

El modelo sobre el cual radica el Análisis factorial implica que los datos observados (X_s) son realmente "producidos" por algunos factores no observados y básicos (f_s).

$$X_{ik} = \lambda_{i1} f_{1k} + \lambda_{i2} f_{2k} + ... + \lambda_{im} f_{mk} + e_{ik}$$

Donde

X_{ik} : valor de la variable *i* para la observación *k.*

λ_{ij} : relación de la variable *i* con el factor común *j* habiendo *m* factores y *p* variables, $m \leq p$.

f_{jk} : valor del factor *j* para observación *k* (comúnmente llamado calificaciones de factor).

Como funciona

Consiste en derivar un "buen" conjunto de λ s (asignaciones). Un buen conjunto de λ s tiene dos propiedades básicas, en primer lugar, las λ s deben producir *X*s, que deben asemejarse a las *X*s observadas. En segundo lugar, las λ s deberán indicar claramente qué variables pertenecen a cuáles factores, el problema es que tanto los valores de factores como las cargas se desconocen, el método para llevar esto a cabo requiere de dos pasos básicos:

1. **Componentes principales**. Los componentes principales (*PCs*) se obtienen de las *X*s originales por medio del modelo:

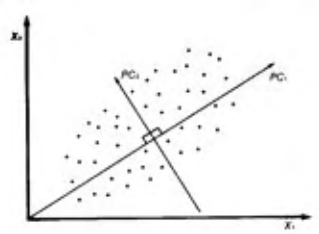
   $$PC_s = a_{j1} x_1 + a_{j2} x_2 + ... + a_{jp} x_p$$

   En donde:

   PC_j : el componente principal *j*.

   a_{ji} : el coeficiente que relaciona la variable *i* con el componente *j*.

Estos componentes se obtienen de manera matemática, de manera que el primero contiene información total para todas las *p* originales posibles, el segundo (que es independiente del primero) contiene la máxima información restante posible.



2. **Normalización de la matriz de datos.** El proceso se inicia a partir de los datos originales, la forma normalizada implica media cero y desviación uno. Para ello, se aplica a partir de los promedios y las varianzas, la siguiente estandarización:

$$y_{i,j} = \frac{x_{i,j} - \bar{x}}{s_{x_j}}$$

Donde $\bar{x}_j$ es la media y s_{x_j} la desviación estándar.

Ejemplo. Se realizó un estudio sobre aceites, una pregunta obvia que surge es si existen patrones de consumo entre estas personas. Aplicaremos un análisis factorial de 36 variables.

Los resultados son los siguientes:

	Componente										
	1	2	3	4	5	6	7	8	9	10	H^2
La tecnología	**.700**			-.389							.64
Diversidad en presentaciones	**.638**			-.345	-.427						.71
La calidad del aceite	**.617**	-.423									.56
Que sea aceite multigrado	**.545**		.360	-.314							.53
Que no se queme rápidamente		**-.733**	-.307								.63
Que aguante altas temperaturas	.495	**-.696**									.73
Que dure mucho tiempo		**-.688**	.364								.61
Que tenga garantía		**-.608**		-.365				-.339			.62
Que resista grandes velocidades	.345	**-.565**		-.402						-.363	.73
Que tenga calidad internacional		-.509		-.414		-.410				-.427	.78
Que lo pueda encontrar en las gasolineras		-.346	**.786**								.74
Que lo pueda comprar en las tiendas de autoservicio		-370	**.686**								.61
Que se pueda comprar en las esquinas de las calles			**.605**	-.310			.411	-.321			.73
Que lo vendan en las refaccionarias	.547		**.547**	-.323							.70
El prestigio de la marca				**-.794**							.63
La solidez de la marca	.354	-.331		**-.767**							.82
La experiencia que tiene la marca en aceites				**-.721**				-.336			.63
La antigüedad de la marca				**-.699**							.49
Que sea una marca reconocida	.427	-.305		**-.623**	-.325					-.330	.88
Que encuentre aceites en cualquier lugar			.448	-.507			.348	-.405			.74
La practicidad de los envases				-.333	**-.773**			-.372			.85
Que pueda guardar el envase con el aceite que no utilice		-.312	.398	-.344	**-.723**						.89

Que venga en un envase de plástico					**-.669**		.461				.66
El envase de los aceites				-.365	**-.611**			-.470		-.398	.89
Que sea una empresa mexicana						**.851**					.72
Que venga en lata					-.440	.531				-.350	.60
Que me la recomiende el mecánico							**.792**				.63
Las promociones que ofrecen								**-.722**			.52
Que tenga diversidad de viscosidad				-.360	-.347			**-.717**			.76
El precio del aceite									**.826**		.68
Que sea de una marca extranjera										**-.730**	.53
Que patrocinen las carreras de coches						.301				**-.723**	.61
La publicidad de los aceites					-.321			-.509		**-.640**	.77
Que tenga publicidad en talleres			.308		-.306		.370	-.426		**-.566**	.83
Que sea una marca líder en el mercado	.328	-.346		-.480				-.352		**-.531**	.86
La imagen en general	.318			-.367			.453			**-.495**	.69

Método de extracción: Análisis de componentes principales.

Método de rotación: Normalización Oblimin con Kaiser.

Comunalidad:

H^2: indica la proporción de varianza en cada variable representada por los diez factores. La comunalidad representa la suma de las cargas al cuadrado de las variables hacia cada factor, por ejemplo, $.64 = 700^2 + (-.389)^2$.

Ahora, para facilitar el análisis, se debe aplicar el criterio y dar un nombre o etiqueta a cada uno de los 10 factores que encontramos:

	Temas	**Etiqueta**
Factor 1	Calidad, tecnología, presentación, que sea multigrado.	Intrínsecos
Factor 2	Resistencia y durabilidad	Resistencia
Factor 3	Disponibilidad en el mercado, en diferentes puntos de venta	Disponibilidad en el mercado
Factor 4	Valores mercadológicos	Marca
Factor 5	Envases, materiales y practicidad	Envases
Factor 6	Que provenga de una empresa mexicana	Empresa mexicana
Factor 7	Recomendación por parte del mecánico	Recomendación
Factor 8	Diversidad de viscosidades, y las promociones que ofrece	Promociones
Factor 9	Precio del aceite	Precio
Factor 10	Esfuerzos comerciales de la marca, así como el apoyo internacional	Apoyos, internacional

Si bien las etiquetas nos ayudarán mucho para le análisis de los resultados, es siempre útil tener a la mano los componentes individuales de cada factor porque a veces en algunas ocasiones, agrupamos atributos similares pero no iguales.

6.14 Análisis de Correspondecia (AC) (Correspondence Analysis).

Es una técnica gráfica para la exploración y búsqueda de afinidad de datos en Tablas de contingencia y categorías de datos multivariados.

Funciones

- ✓ Representa gráficamente el patrón de asociación entre dos variables con escala nominal. Se inicia a partir de una tabla de contingencia donde cada renglón y cada columna definen un perfil.
- ✓ Compara los perfiles de los renglones y las columnas por separado para colocarlos en un punto de la gráfica, de tal manera que perfiles semejantes estén asociados a puntos cercanos.

Para que se utiliza

- ✓ Crea mapas perceptuales para la definición de mercado, posicionamiento y seguimiento de producto, así como estudios de imagen.
- ✓ Evalúa publicidad.
- ✓ Prueba conceptos y nombres de productos.
- ✓ Desarrolla un sistema de segmentación, (Análisis de Correspondencia Múltiple).

Cálculo

El análisis de Correspondencia se aplica en Tablas de contingencia, matriz representativa de 2 caracteres o propiedades generales de tipo cualitativo, expresado en forma de modalidades exhaustivas y exclusivas entre sí. Por ejemplo, sea la siguiente tabla:

		Caracter 2				
		1	**2**	**3**	**…**	**P**
Caracter 1	**1**	k_{11}	k_{12}	k_{13}	…	K_{1p}
	2	k_{21}	k_{22}	k_{23}	…	K_{2p}
	3	k_{31}	k_{32}	k_{33}	…	K_{3p}
	.					
	.	…	…	…	…	…
	.					
	n	k_{n1}	k_{n2}	k_{n3}	…	k_{np}

Donde el elemento k_{ij} representa el número de individuos (frecuencia absoluta) que cumplen tanto la condición de la modalidad *i* perteneciente al primer carácter como la condición de la modalidad correspondiente al segundo carácter estudiado sobre una muestra.

Teniendo como matriz de datos de partida una tabla de frecuencias:

		1	**2**	**3**	**…**	**j**	**…**	**p**	
	1	k_{11}	k_{12}	k_{13}	…	k_{1p}		k_{1p}	$\mathbf{k_1}$
	2	K_{21}	k_{22}	k_{23}	…	k_{2p}		k_{2p}	$\mathbf{k_2}$
	3	k_{31}	k_{32}	k_{33}	…	k_{3p}		k_{3p}	$\mathbf{k_3}$
$\mathbf{K_{(n\times p)}}$ =	.	…	…	…	…	…	…	…	…
	:								
	:								
	I	k_{i1}	k_{i2}	k_{i3}	…	k_{ip}	…	k_{ip}	$\mathbf{k_.}$

N	k_{n1}	k_{n2}	k_{n3}	…	k_{np}	…	k_{np}	k_n
	k_1	k_2	k_3	…	k_p	…	k_p	K

Para aplicar el análisis de Correspondencias, no se emplea directamente la tabla de frecuencias sino que se transforma en una matriz de probabilidades, de acuerdo a la siguiente expresión:

$f_{ij} = k_{ij} / K$

A partir de la información anterior se tiene la posibilidad de definir, sobre R^p, las características de cada punto fila, que vienen dadas por los siguientes elementos:

masa = $f_i = k_i / K$

perfil = (f_{ij} / f_i) $\quad j = 1,...,p$

coordenadas = $(f_{ij} / (f_i f_j)^{1/2})$ $\quad j = 1,...,p$

Observando que el perfil fila no es nada mas que la frecuencia condicionada [$(f(j/i)$], y por otra parte, que el perfil medio es equivalente al marginal de la tabla de frecuencias.

En R^n, cada punto columna j vendrá dado por los siguientes elementos:

masa = $f_j = k_j / K$

perfil = (f_{ij} / f_j) $\quad j = 1,...,p$

coordenadas = $(f_{ij} / (f_j f_i)^{1/2})$ $\quad j = 1,...,p$

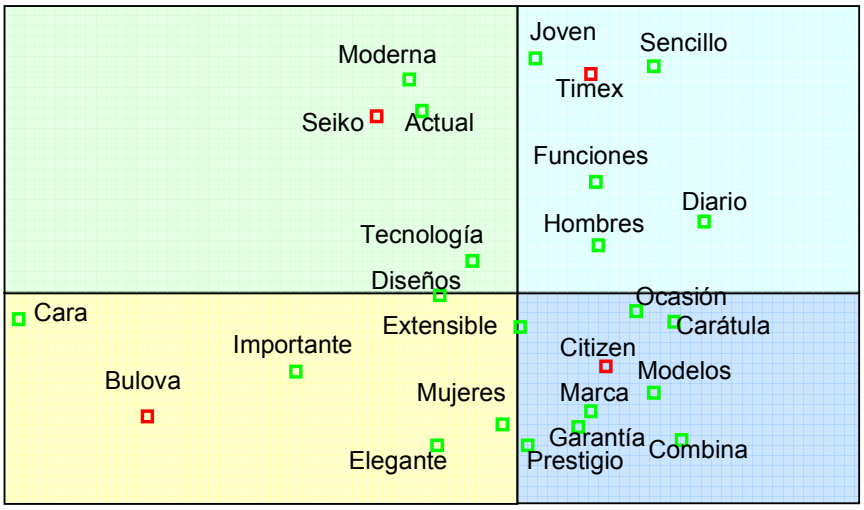
La masa de una fila (f_i) o columna (f_j) debe entenderse como la importancia relativa dentro de la tabla de datos. Sirven para atenuar la preponderancia que podría tener alguna fila o columna en el análisis. Por otra parte, el perfil fila o perfil columna identifica cada modalidad en cuanto a su importancia relativa.

Ejemplo. En un estudio de relojes se evaluaron una serie de atributos para poder conocer el perfil por marca es decir, se quiere saber con que marcas se asocian los atributos:

Atributos	**CITIZEN (%)**	**TIMEX (%)**	**BULOVA (%)**	**SEIKO (%)**
A. Es la mas moderna	46	16	11	23
B. Es la mas actual	44	16	11	19
C. Es la mas cara	32	10	26	18
D. Tiene los mejores diseños	48	12	12	14
E. Tiene la mejor tecnología	47	10	9	16
F. Es la marca que usa la gente importante	41	9	16	13
G. Tiene la mejor garantía	52	9	8	9
H. Su extensible/correa es la mejor	47	11	9	11
I. Tiene el mayor prestigio	60	11	12	10
J. Es la marca que debo usar	56	10	8	10
K. Tiene la mayor variedad de funciones	43	13	5	12
L. Es para gente joven	42	18	7	15
M. Por su diseño se puede usar en cualquier ocasión	56	13	6	11
N. Es para mujeres	50	10	11	9
O. Es para uso diario	53	17	4	9
P. Es para hombres	51	13	6	13
Q. Es sencillo	47	22	5	11

R. Su carátula es fácil de leer	54	13	5	9
S. Es elegante	57	10	15	12
T. Es la que tiene mayor variedad de modelos	54	11	6	8
U. Combina con la ropa	53	9	5	7

Mapa perceptual:



Conclusión

Citizen se percibe principalmente como la marca que debo usar, la que combina con la ropa y, la que tiene mayor variedad de modelos principalmente. Timex, es para gente joven, sencillo y tiene la mayor variedad de funciones. Seiko es moderna, actual y, tiene la mejor tecnología, mientras que Bulova se considera una marca que usa la gente importante y tiene los mejores diseños.

6.15 Escalas multidimensionales.

El Escalamiento multidimensional tiene como propósito posicionar los objetos y las variables provenientes de una encuesta en un espacio multidimensional o mapa perceptual.

Funciones

- ✓ Transformar los juicios de similitudes o preferencias de los consumidores (por ejemplo, preferencias sobre tiendas o marcas) en distancias representadas en el espacio multidimensional. Si los objetos A y B son juzgados por los que responden como los que son más similares comparados con todos los otros posibles pares de objetos A y B de tal manera que la distancia entre ellos en el espacio multidimensional será menor que la distancia entre otro par de objetos.

Para que se utiliza

- ✓ La identificación de los atributos notables de un producto, percibidos por el comprador.
- ✓ Los productos que se consideran como sustitutos y aquellos que se diferencian entre sí.
- ✓ Los segmentos viables que existen en el mercado.
- ✓ Investigación sobre el cambio de marcas.

Cálculo

El procedimiento, en términos muy generales, sigue algunas ideas básicas en la mayoría de las técnicas. El punto de partida es una matriz de disimilaridades entre *n* objetos, con el elemento $\partial_{i,j}$ en la fila *i* y en la columna *j*, que representa la disimilaridad del objeto *i* al objeto *j*. También se fija el número de dimensiones, *p*, para hacer el gráfico de los objetos en una solución particular. Generalmente el camino que se sigue es:

- Arreglar los *n* objetos en una configuración inicial en *p* dimensiones, esto es, suponer para cada objeto las coordenadas $x_1, x_2, ...x_p$ en el espacio de *p* dimensiones.
- Calcular las distancias euclidianas entre los objetos de esa configuración, esto es, calcular las $d_{i,j}$, que son las distancias entre el objeto *i* y el objeto *j*.
- Hacer una Regresión de $d_{i,j}$ sobre $\partial_{i,j}$. Esta Regresión puede ser lineal, polinomial o monótona. Por ejemplo, si se considera lineal se tiene el modelo $d_{i,j} = a + b\,\partial_{i,j} + \varepsilon$ y utilizando el método de los mínimos cuadrados se obtienen estimaciones de los coeficientes *a* y *b*, y de ahí puede obtenerse lo que genéricamente se conoce como una "disparidad" $\overline{d}_{i,j} = \overline{a} + \overline{b}\,\partial_{i.j}$. Si se supone una Regresión monótona, no se ajusta una relación exacta entre $d_{i,j}$ y $\partial_{i,j}$, sino se supone simplemente que si $\partial_{i,j}$ crece, entonces $d_{i,j}$ crece o se mantiene constante.
- A través de algún estadístico conveniente se mide la bondad de ajuste entre las distancias de la configuración y las disparidades. Existen diferentes definiciones de este estadístico, pero la mayoría surge de la definición del llamado índice de esfuerzo (en inglés: STRESS).

$$STRESS1 = \sqrt{\frac{\sum\sum\left(d_{i,j} - \overline{d}_{i,j}\right)^2}{\sum\sum d_{i,j}^2}}$$

$$SSTRESS1 = \sqrt{\frac{\sum\sum\left(d_{i,j}^2 - \overline{d}_{i,j}^2\right)^2}{\sum\sum d_{i,j}^4}}$$

Todas las sumas sobre *i* y *j* van de 1 a *p* y las disparidades dependen del tipo de Regresión utilizado en el tercer paso del procedimiento. El STRESS1 es la fórmula introducida por Kruskal quien ofreció la siguiente guía para su interpretación:

Tamaño del STRESS1	Interpretación
0.2	Pobre
0.1	Regular
0.05	Bueno
0.025	Excelente
0.00	Perfecto

- Las coordenadas $x_1, x_2, ...x_t$ de cada objeto se cambian ligeramente de tal manera que la medida de ajuste se reduzca.

Los pasos del 2 al 5 se repiten hasta que al parecer la medida de ajuste entre las disparidades y las istancias de configuración no puedan seguir reduciéndose. El resultado final del análisis es entonces las coordenadas de los *n* objetos en las *p* dimensiones. Estas coordenadas pueden usarse para elaborar un gráfico que muestre cómo están relacionados los objetos. Lo ideal sería encontrar una buena solución en menos de tres dimensiones, pero esto no es siempre posible.

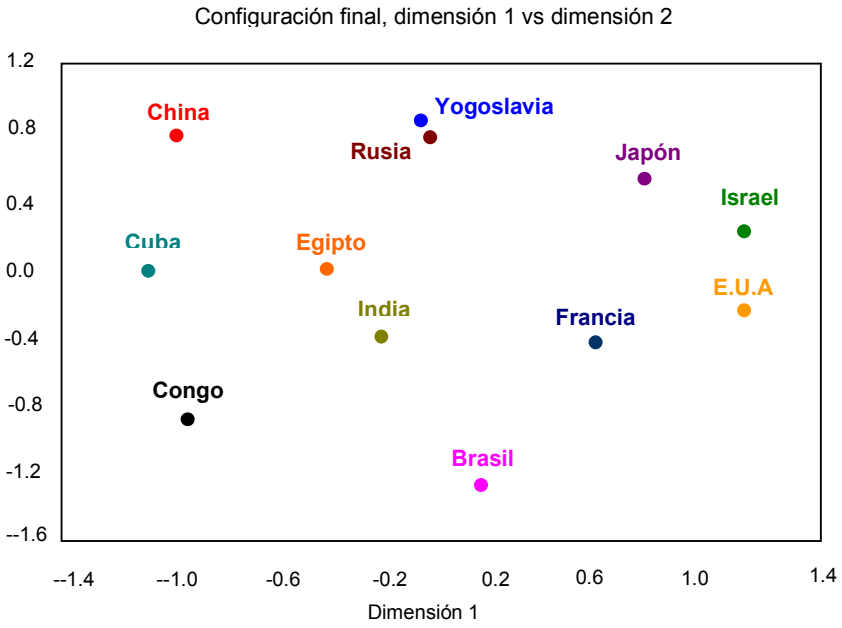
La matriz de similaridades puede consistir de medidas de similaridad $\partial_{i,j}$ entre los objetos *i* y *j* que son iguales a $\partial_{i,j}$ (matriz simétrica), pero también puede ser que $\partial_{i,j}$ no es igual a $\partial_{i,j}$ (matriz asimétrica). Las principales técnicas para matrices simétricas son:

- Escalamiento métrico. Entre estos métodos puede citarse el análisis de coordenadas principales que usa los valores $\partial_{i,j}$ directamente. El proceso computacional de las coordenadas principales es similar al del análisis de componentes principales. Un caso especial del análisis de coordenadas principales ocurre cuando los datos son distancias euclidianas obtenidas de la matriz de datos originales Y de *n* filas y *p* columnas; en este caso, existe una dualidad entre un análisis de coordenadas principales sobre una matriz de dispersión y un análisis de componentes principales sobre una matriz de distancias.
- "Unfolding". Cuando la matriz de similaridades tiene valores perdidos, el escalamiento métrico es impracticable, pero un caso especial es cuando de la matriz triangular inferior, por ejemplo, se tiene una tabla de dos entradas de dimensión *r* x *s* y con sólo estos datos pueden encontrarse las coordenadas de los *r* + *s* puntos.
- Escalamiento no métrico. Aunque en esencia estos métodos son similares a su contrapartida métrica en que encuentran localizaciones para los *n* objetos que mantienen aproximadamente las asociaciones entre los objetos, esto se hace utilizando una escala ordinal de tal manera que el criterio de bondad de ajuste depende sólo del orden de los valores $\partial_{i,j}$ y no de sus valores absolutos. Para las matrices asimétricas existen dos clases de métodos, a saber:
  - Métodos de primera clase. Se modela la tabla completa.
  - Métodos de segunda clase. Se modela la tabla en dos partes, primero la componente simétrica y después la asimétrica.

Ejemplo. En un estudio piloto sobre la percepción sobre diferentes naciones llevado a cabo a principio del año 70, cada uno de los 18 estudiantes que participaron en el estudio atribuyó una nota entre 1 (para los países muy diferentes) y 7 (para los muy

similares) a cada uno de los 66 pares formados en el conjunto de las 12 naciones consideradas.
Un ejemplo clásico, los datos son citados en Kruskall (1976). Los datos son los siguientes:

	BRA	CON	CUB	EGY	FRA	IND	ISR	JAP	CHI	RUS	USA	YUG
BRA	7.0	4.8	5.3	3.4	4.7	4.5	3.8	3.5	2.4	3.1	5.4	3.2
CON	4.8	7.0	4.6	5.0	4.0	4.8	3.3	3.4	4.0	3.4	2.4	3.5
CUB	5.2	4.6	7.0	5.2	4.1	4.0	3.6	2.9	5.5	5.4	3.2	5.1
EGY	3.4	5.0	5.2	7.0	4.8	5.8	4.7	3.8	4.4	4.4	3.3	4.3
FRA	4.7	4.0	4.1	4.8	7.0	3.4	4.0	4.2	3.7	5.1	5.9	4.7
IND	4.5	4.8	4.0	5.8	3.4	7.0	4.1	4.5	4.1	4.5	4.3	4.0
ISR	3.8	3.3	3.6	4.7	4.0	4.1	7.0	4.8	3.0	4.2	5.9	4.4
JAP	3.5	3.4	2.9	3.8	4.2	4.5	4.8	7.0	4.2	4.6	6.1	4.3
CHI	2.4	4.0	5.5	4.4	3.7	4.1	3.0	4.2	7.0	5.7	2.6	5.1
RUS	3.1	3.4	5.4	4.4	5.1	4.5	4.2	4.6	5.7	7.0	5.0	6.7
USA	5.4	2.4	3.2	3.3	5.9	4.3	5.9	6.1	2.6	5.0	7.0	3.6
YUG	3.2	3.5	5.1	4.3	4.7	4.0	4.4	4.3	5.1	6.7	3.6	7.0



Puede observarse que los valores más bajos en la tabla indican que, por ejemplo, China y Brasil y USA y el Congo son percibidos con las mayores diferencias, mientras que Yugoslavia y Rusia y USA y Japón se perciben como los más similares ya que tienen los valores más altos. El STRESS1 obtenido fue de 0.189 que según la guía ofrecida anteriormente puede calificarse de "regular".

Existe un posicionamiento muy interesante de los países sobre el plano:
- USA, Japón, Francia e Israel son países capitalistas desarrollados.
- Congo, Egipto e India son países subdesarrollados.
- Cuba, China, Rusia y Yugoslavia son países comunistas.

6.16 Segmentación y Cluster.

La Segmentación de mercado describe la división de mercado en grupos homogéneos en donde cada uno responderá en forma diferente a promociones, comunicaciones, publicidad y otra mezcla de variables mercadológicas.

Es el arte de dividir a objetos o individuos en grupos. Para la mayoría de los tipos de segmentación estos grupos son mutuamente exclusivos (cada individuo es asignado a un solo segmento). La segmentación es muy útil para identificar y evaluar patrones que no son fáciles de distinguir con un análisis simple. Un análisis de este tipo puede ayudar a identificar cuales son los clientes rentables de una institución, a entender como interactúan con la institución y como se sienten con respecto a los servicios que reciben.

Funciones

- ✓ Facilidad de mercadeo. Es más sencillo cubrir las necesidades de pequeños grupos de consumidores, particularmente si tienen muchas características en común (por ejemplo: buscan los mismos beneficios, tiene la misma edad y/o sexo, etc.).
- ✓ Encontrar nichos de mercado. Identificar sub-cubiertos o no-cubiertos. Utilizando "mercadotecnia de nichos", la segmentación puede permitir la introducción de una compañía o de un nuevo producto mediante la identificación de los futuros compradores y pude ayudar a un producto maduro a encontrar nuevos compradores.
- ✓ Eficiencia. Los recursos que se poseen para una estrategia de mercado pueden ser optimizados enfocándolos a los segmentos para su ofrecimiento–producto, precio, promoción y plaza (distribución).

Para que se utiliza

- ✓ Proporcionar servicios más eficientes.
- ✓ Direccionando estrategias a clientes existentes y prospectos más efectivas.
- ✓ Fortaleciendo la relación con el cliente.
- ✓ Desarrollando productos que satisfagan necesidades específicas de los clientes.
- ✓ Posicionando y diferenciando productos.
- ✓ Entendiendo, evaluando el tamaño y priorizando las oportunidades.

Método de aproximación

Los métodos de agrupación se basan en comparaciones por parejas, y para ello, se debe establecer uno de los dos criterios para medir la semejanza entre los objetos: el primero hace referencia a la *proximidad* y el otro, a la *similitud* entre los objetos.

Proximidad: se referencia en un espacio métrico, utilizando el sentido euclidiano de medida, en donde las dimensiones sin los atributos que caracterizan los objetos. Si x_{ik} es la medida normalizada de la proyección sobre la dimensión *k* del punto u objeto *i* en un espacio ortogonal de componentes principales, le medida de proximidad entre los dos puntos *i* y *j* se expresa:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Aplicando esta expresión en el caso de dos dimensiones, se tiene que la distancia entre los puntos 1 y 2 es:

$$d_{12} = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

Similitud: la similitud se refiere a la semejanza o grado de acoplamiento que presentan las variables que pertenecen a las escalas categóricas, ordinales y nominales. Para ello se utilizan coeficientes que miden la capacidad de los objetos para compartir atributos. El coeficiente de similitud entre dos objetos se define:

$$S = \frac{M}{N} = \frac{num \quad atributos \quad compartidos}{num \quad atributos \quad estudiados}$$

Cuando se utilizan medidas en escalas categóricas o en escalas mezcladas, se puede definir como coeficiente de similitud:

$$S = 1 - \frac{|x_{ik} - x_{ik_1}|}{R_k}$$

En donde la medida *S* varía entre 0 y 1 y R_k es el rango de la variable.

Métodos de agrupación:

Los algoritmos empiezan con *n* conglomerados (uno por cada observación individual). En pasos sucesivos los conglomerados se van uniendo por afinidad, de modo que en cada etapa los conglomerados son subconjuntos de los que se obtienen en las etapas subsiguientes. Distintos criterios para calcular las similitudes (o distancias) entre clusters son los siguientes:

- ✓ **Método de encadenamiento simple o vecino más próximo**: la similitud entre dos conglomerados se calcula mediante la máxima similitud entre sus respectivos elementos.
- ✓ **Método de encadenamiento completo o vecino más lejano**: La similitud entre dos conglomerados se calcula mediante la mínima similitud entre sus respectivos elementos.
- ✓ **Método de agrupación de centroides**: se calcula la distancia entre los vectores de medias.

Estructura general de un algoritmo aglomerativo.

a) Comenzamos con *n* clusters (cada uno contiene una observación) y con una matriz de similitudes (o con una matriz de distancias).
b) En la matriz de similitudes (o de distancias) buscamos la pareja de clusters más parecidos.
c) Unimos la pareja de clusters del paso anterior y actualizamos la matriz de similitudes del siguiente modo: eliminamos las filas y las columnas que corresponden a la pareja de clusters. Añadimos una fila y una columna que contendrá las similitudes del nuevo cluster con los restantes clusters.
d) Los pasos 2 y 3 se repiten n − 1 veces.

El proceso del algoritmo puede resumirse en el historial de aglomeración, que es una tabla en la que se indica en cada paso los conglomerados que se unen y las distancias o similitudes entre ellos. También puede representarse gráficamente en un dendograma.

El resultado final del algoritmo puede variar mucho dependiendo de la medida de similitud o distancia utilizada y de cual sea el criterio (vecino más próximo, vecino más lejano, agrupación de centroides) que se utilice para definir la distancia entre los conglomerados.

El número de conglomerados se puede decidir a partir del historial de aglomeración o del dendograma, deteniendo el proceso de aglomeración en el momento en que la aplicación del algoritmo lleve a unir conglomerados que están muy distantes.

Métodos de agrupación no jerárquicos o de partición: El método de K-medias.

a) Repartimos las *n* observaciones en *K* grupos. Esta primera asignación se puede hacer aleatoriamente. En cada uno de los grupos se obtiene el vector de medias (centro del grupo).
b) Asignamos secuencialmente cada observación al grupo cuyo centro esté más cercano (usualmente, utilizamos la distancia euclídea de las observaciones a los centros de los grupos). En cada etapa se re-calcula el centro del grupo al que se añade una observación y el centro del grupo del que se elimina esa observación.
c) Repetir el paso anterior hasta que no haya re-asignaciones.

Un criterio para medir la homogeneidad de los grupos.

Es la "suma de cuadrados dentro de grupos":

$$SCDG = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k)'(x_{ik} - \bar{x}_k) = \sum_{k=1}^{K}\left(n_k \sum_{i=1}^{p} s_{ik}^2\right)$$

Donde

n_k : es el número de observaciones en el grupo *k.*

x_{ik} : es la *i*-ésima observación del grupo *k.*

$\bar{x}_k$: es el vector de medias (y centro) del cluster *k.*

s_{ik}^2 : es la varianza de la variable *i* en el grupo *k.*

Una regla empírica para seleccionar el número de grupos es pasar de *K* a *K + 1* grupos si

$$F = \frac{SCDG(K) - SCDG(K+1)}{SCDG(K+1)/(n-K-1)} > 10$$

donde *SCDG(K)* y *SCDG(K + 1)* denotan la suma de cuadrados dentro de los grupos cuando utilizamos *K* y *K + 1* grupos, respectivamente.
El objetivo del algoritmo de *K*-medias es encontrar una partición en *K* conglomerados que de un mínimo aproximado de la *SCDG.*

Ejemplo. Se tiene una muestra de siete entrevistados que responden a una encuesta de diez preguntas, los resultados son los siguientes:

		Respuesta									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Encuestado	**1**	a	B	b	c	a	B	B	a	a	D
	2	a	C	b	c	d	E	E	a	b	C
	3	c	B	b	c	d	A	B	c	a	D
	4	a	B	e	c	a	D	B	a	a	C
	5	c	C	b	b	d	A	B	c	d	D
	6	a	C	e	c	d	C	E	a	e	D
	7	b	B	c	a	a	A	B	c	a	B

Se usará como distancia entre casos el número (o la fracción, dividiendo el número por 10) de respuestas diferentes, y la distancia entre conglomerados, la del vecino más próximo.

Iteración 1 (D_1)**.**

La matriz de distancias entre los encuestados es la siguiente, siendo cada caso un conglomerado:

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0	6	4	3	7	6	6
2	6	0	7	6	7	4	10
3	4	7	0	6	3	7	5
4	3	6	6	0	9	6	6
5	7	7	3	9	0	7	7
6	6	4	7	6	7	0	10
7	6	10	5	6	7	10	0

Se unen (1,4) y (3,5) a la distancia 3.

Iteración 2 (D_2)**.**

	(1,4)	**2**	**(3,5)**	**6**	**7**
(1,4)	0	6	4	6	6
2	6	0	7	4	10
(3,5)	4	7	0	7	5
6	6	4	7	0	10
7	6	10	5	10	0

Se unen (1,4) con (3,5) y (2,6) a la distancia 4.

Iteración 3 (D_3)**.**

	(1,4,3,5)	**(2,6)**	**7**
(1,4,3,5)	0	6	5
(2,6)	6	0	10
7	5	10	0

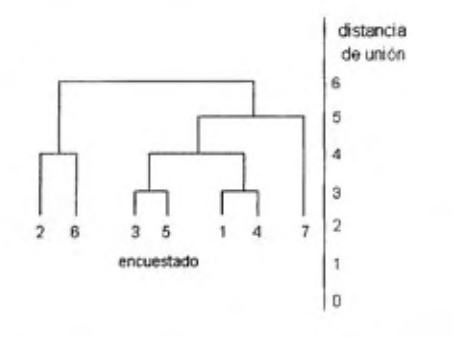
Se unen (1, 3, 4, 5) con 7 a la distancia 5.

Iteración 4 (D_4)**.**

	(1,4,3,5,7)	**(2,6)**
(1,4,3,5,7)	0	6
(2,6)	6	0

Última iteración, se unen todos en un solo conglomerado a la distancia 6.

El gráfico siguiente es el Dendograma, que ilustra la forma en como se fueron uniendo los conglomerados.



La escala Horizontal corresponde a la distancia en que se produjeron las uniones en cada caso. De este gráfico se despende si queremos tener dos conglomerados, estos serían (1, 3, 4, 5, 7) y (2, 6), si queremos tres serían (7), (1, 3, 4, 5) y (2, 6). Si queremos 5, estos serían (1, 4), (3, 5), (2), (6) y (7).

6.17 Resumen.

El **análisis de Correlación** expresa el grado de asociación entre variables, mientras que el **análisis de Regresión** es la forma, el cómo es esa asociación. El objetivo **análisis Discriminante** consiste en seleccionar atributos que permiten diferenciar objetos entre sí, en contraste, el **análisis de Factores** agrupa atributos que son similares. El **análisis de Correspondencia** genera un mapa perceptual en el que se posicionan tanto los elementos de los atributos como las marcas. El **Escalaminento multidimensional** emplea como datos de entrada las proximidades entre diferentes objetos. El **análisis Conjunto** es un modelo de descomposición, para posteriormente, estimar preferencias. El objetivo del **análisis de Conglomerados (Cluster)** es obtener grupos de objetos de forma que los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí, pero muy diferentes a los demás grupos.

Comentarios finales.

A través del tiempo y en la medida en que los mercados cambian, las estrategias de mercado se hacen más evidentes, las condiciones que alguna vez fueron determinantes para el éxito de un negocio, se pueden volver adversas para su supervivencia en otra época, los consumidores hoy en día ya no son considerados homogéneos, en lugar de ello, se desarrollan estrategias dirigidas a segmentos con características demográficas y psicográficas específicas.

La Investigación de mercados propone principios y procedimientos, como el diseño de un buen cuestionario, tipo de escala a emplearse, la selección del tipo y tamaño de muestra, esto sin dejar de lado el objetivo de la investigación como proceso inicial. La siguiente etapa es la recolección de la información y el análisis de la misma, donde el investigador puede emplear por ejemplo desde un análisis descriptivo hasta una técnica multivariada o ambos. Para el análisis estadístico debemos considerar los siguientes aspectos:

- el tipo de variables
- si tenemos una o más variables dependiente (s)
- nivel o escala de los datos (métrica o no métrica)

según estos datos, el análisis utilizado variará, por ejemplo, si hay una sola variable dependiente y varias variables independientes y si la variable dependiente es métrica (de intervalos o proporciones), se puede utilizar Regresión o Correlación, si la variable dependiente es no métrica, es decir, es clasificatoria (escala ordinal o nominal) podemos utilizar análisis Discriminante. Si hay varias variables dependientes y estas son métricas, se aplica el análisis de Varianza. En el caso de que los datos no estén clasificados como dependientes e independientes, si los datos son métricos se utiliza por ejemplo el análisis de Factores o el análisis de Conglomerados, cuando los datos no son métricos se puede aplicar Escalamiento multidimensional; el investigador tendrá que seleccionar la técnica más apropiada para así apoyar el desarrollo de una empresa y lograr sus objetivos.

En la actualidad existen paquetes estadísticos que llevan a cabo lo laborioso del cálculo, la finalidad de este trabajo es dar un panorama general tanto de su procedimiento, como de su aplicación.

Como una aplicación de la Actuaría, las herramientas estadísticas aplicadas a la Investigación de mercados nos pueden ser muy útiles hoy en día para conocer las necesidades del consumidor considerando entre otras cosas, las crisis económicas, el "bombardeo" publicitario", además de adentrarnos en valores, actitudes, conductas, para así cumplir con la definición de un profesional de la Actuaría.

Bibliografía.

Investigación de mercados
Aaker – Kumar – Day
Ed. Limusa Wiley
2001

Investigación de mercados
Ronald M. Weiers
Ed. Prentice Hall
1986

Investigación de mercados, un enfoque aplicado
Kinnear, Taylor, Rosas, Santana y Londono
Ed. MacGraw-Hill
1981

Investigación y análisis de mercado
Lehmann
Ed. CECSA
1993

Investigaciones de mercadeo
Paul E. Green, Donald S. Tull
Ed. Prentice Hall
1985

Estadística no paramétrica
Sidney Siegel
Ed. Trillas
1991

Probabilidad y Estadística
George C. Canavos
Ed. McGraw Hill
1999

Investigación de mercados, concepto y práctica
Arturo Orozco J.
Ed. Norma
1999

Marketing Research, an applied orientation
K. Malhotra
Ed. Prentice Hall
1999

Elementos de muestreo
Scheaffer, Mendenhall y Ott
Grupo Editorial Iberoamérica

1987

Artículos de Estadística
Pértega Díaz S, Pita Fernández S.
www.estadistico.com

Estadística
Manuel Salvador Figueras
http: //ciberconta.unizar.es

Estadística
Juan Domingo Gispert López
www.ciber-bbn.uned.es

Correlación Lineal y Análisis de Regresión
Alicia Vila, Máximo Sedano, Ana López, Ángel a. Juan
www.uoc.edu

Apuntes de Estadística
Rodrigo Basurto Barrera
dcb.fi-c.unam.mx

Análisis de conglomerados
Jorge Galbiati R.
www.jorgegalbiati.cl

Evaluación de los procedimientos de medición de la variable respuesta en el análisis conjunto bajo distintas alternativas de estimación.
Ramírez y Rondán.
www.asepelt.org/ficheros

Aplicaciones Ji cuadrada
Dra. Ada Ray
www.aray1.com

Tabla 1. Valores de la función de distribución acumulativa Binomial.

$$P(X \le x) = F(x;n,p) = \sum_{i=1}^{x} \binom{n}{i} p^x (1-p)^{n-x}$$

		P										
N	**x**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
2	**0**	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
	1	0.9999	0.9975	0.9900	0.9775	0.9600	0.9375	{).9100	0.8775	0.8400	0.7975	0.7500
	2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
3	**0**	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
	1	0.9997	0.9928	0.9720	0.9392	0.8960	0.8438	0.7840	0.7183	0.6480	0.5748	0.5000
	2	1.0000	0.9999	0.9990	0.9966	0.9920	0.9844	0.9730	0.957!	0.9360	0.9089	0.8750
	3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4	**0**	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.240!	0.1785	0.1296	0.0915	0.0625
	1	0.9994	0.9860	0.9477	0.8905	0.8192	0.7383	0.6517	0.5630	0.4752	0.3910	0.3125
	2	1.0000	0.9995	0.9963	0.9880	0.9728	0.9492	0.9163	0.8735	0.8208	0.7585	0.6875
	3	1.0000	1.0000	0.9999	0.9995	0.9984	0.9961	0.9919	0.9850	0.9744	0.9590	0.9375
	4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
5	**0**	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
	1	0.9990	0.9774	0.9185	0.8352	0.7373	0.6328	0.5282	0.4284	0.3370	0.2562	0.1875
	2	1.0000	0.9988	0.9914	0.9734	0.9421	0.8965	0.8369	0.7648	0.6826	0.5931	0.5000
	3	1.0000	1.0000	0.9995	0.9978	0.9933	0.9844	0.9692	0.9460	0.9130	0.8688	0.8125
	4	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990	0.9976	0.9947	0.9898	0.9815	0.9688
	5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	**0**	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
	1	0.9985	0.9672	0.8857	0.7765	0.6564	0.5339	0.4202	0.3191	0.2333	0.1636	0.1094
	2	1.0000	0.9978	0.9842	0.9527	0.9011	0.8306	0.7443	0.6471	0.5443	0.4415	0.3438
	3	1.0000	0.9999	0.9987	0.9941	0.9830	0.9624	0.9295	0.8826	0.8208	0.7447	0.6563
	4	1.0000	1.0000	0.9999	0.9996	0.9984	0.9954	0.9891	0.9777	0.9590	0.9308	0.8906
	5	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9982	0.9959	0.9917	0.9844
	6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	**0**	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
	1	0.9980	0.9556	0.8503	0.7166	0.5767	0.4449	0.3294	0.2338	0.1586	0.1024	0.0625
	2	1.0000	0.9962	0.9743	0.9262	0.8520	0.7564	0.6471	0.5323	0.4199	0.3164	0.2266
	3	1.0000	0.9998	0.9973	0.9879	0.9667	0.9294	0.8740	0.8002	0.7102	0.6083	0.5000
	4	1.0000	1.0000	0.9998	0.9988	0.9953	0.9871	0.9712	0.9444	0.9037	0.8471	0.7734
	5	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9962	0.9910	0.9812	0.9643	0.9375
	6	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9994	0.9984	0.9963	0.9922
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
8	**0**	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
	1	0.9973	0.9428	0.8131	0.6572	0.5033	0.3671	0.2553	0.1691	0.1064	0.0632	0.0352
	2	0.9999	0.9942	0.9619	0.8948	0.7969	0.6785	0.5518	0.4278	0.3154	0.2201	0.1445
	3	1.0000	0.9996	0.9950	0.9786	0.9437	0.8862	0.8059	0.7064	0.5941	0.4770	0.3633
	4	1.0000	1.0000	0.9996	0.9971	0.9896	0.9727	0.9420	0.8939	0.8263	0.7396	0.6367
	5	1.0000	1.0000	1.0000	0.9998	0.9988	0.9958	0.9887	0.9747	0.9502	0.9115	0.8555
	6	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9987	0.9964	0.9915	0.9819	0.9648
	7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9998	0.9993	0.9983	0.9961
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
9	**0**	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0207	0.0101	0.0046	0.0020
	1	0.9966	0.9288	0.7748	0.5995	0.4362	0.3003	0.1960	0.1211	0.0705	0.0385	0.0195
	2	0.9999	0.9916	0.9470	0.8591	0.7382	0.6007	0.4628	0.3373	0.2318	0.1495	0.0898
	3	1.0000	0.9994	0.9917	0.9661	0.9144	0.8343	0.7297	0.6089	0.4826	0.3614	0.2539
	4	1.0000	1.0000	0.9991	0.9944	0.9804	0.9511	0.9012	0.8283	0.7334	0.6214	0.5000
	5	1.0000	1.0000	0.9999	0.9994	0.9969	0.9900	0.9747	0.9464	0.9006	0.8342	0.7461
	6	1.0000	1.0000	1.0000	1.0000	0.9997	0.9987	0.9957	0.9888	0.9750	0.9502	0.9102
	7	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9986	0.9962	0.9909	0.9805
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9992	0.9980
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
10	**0**	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0135	0.0060	0.0025	0.0010
	1	0.9957	0.9139	0.7361	0.5443	0.3758	0.2440	0.1493	0.0860	0.0464	0.0233	0.0107
	2	0.9999	0.9885	0.9298	0.8202	0.6778	0.5256	0.3828	0.2616	0.1673	0.0996	0.0547
	3	1.0000	0.9990	0.9872	0.9500	0.8791	0.7759	0.6496	0.5138	0.3823	0.2660	0.1719
	4	1.0000	0.9999	0.9984	0.9901	0.9672	0.9219	0.8497	0.7515	0.6331	0.5044	0.3770
	5	1.0000	0.9999	0.9986	0.9936	0.9803	0.9527	0.9051	0.8338	0.7384	0.6230	9.0000
	6	1.0000	1.0000	1.0000	0.9999	0.9991	0.9965	0.9894	0.9740	0.9452	0.8980	0.8281
	7	1.0000	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9952	0.9877	0.9726	0.9453
	8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9995	0.9983	0.9955	0.9893
	9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9997	0.9990
	10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Tabla 2. Valores críticos de *T* en la prueba de los rangos señalados de pares igualados de Wilcoxon.

N	Niveles de significación para prueba de una cola		
	.025	.01	.005
	Niveles de significación para prueba de dos colas		
	.05	.02	.01
6	0		
7	2	0	
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	80	77	68

Tabla 3. Valores críticos de *U* en la prueba Mann-Whitney.
Valores críticos de *U* para una prueba de una cola en $\alpha = 0.025$ o para una prueba de dos colas en $\alpha = 0.05$.

n_1 \ n_2	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

Tabla 4. Tabla de valores asociadas con valores tan grandes como los valores observados de *H* en el análisis de varianza de una clasificación por rangos de Kruskal-Wallis.

Tamaño de muestras			*H*	*p*
n_1	n_2	n_3		
5	2	2	6.5333	0.008
			6.1333	0.013
			5.1600	0.034
			5.0400	0.056
			4.3733	0.090
			4.2933	0.122
5	3	1	6.4000	0.012
			4.9600	0.048
			4.8711	0.052
			4.0178	0.095
			3.8400	0.123
5	3	2	6.9091	0.009
			6.8218	0.010
			5.2509	0.049
			5.1055	0.052
			4.6509	0.091
			4.4945	0.101
5	3	3	7.0788	0.009
			6.9818	0.011
			5.6485	0.049
			5.5152	0.051
			4.5333	0.097
			4.4121	0.109
5	4	1	6.0545	0.008
			6.8400	0.011
			4.9855	0.044
			4.8600	0.056
			3.9873	0.098
			3.9600	0.102
5	4	2	7.2045	0.009
			7.1182	0.010
			5.2727	0.049
			5.2632	0.050
			4.5409	0.098
			4.5182	0.101
5	4	3	7.4449	0.010
			7.3949	0.011
			5.6564	0.049
			5.6308	0.050
			4.5487	0.099
			4.5231	0.103
5	4	4	7.7604	0.009
			7.7440	0.011
			5.6571	0.049
			5.6176	0.050
			4.6187	0.100
			4.5527	0.102
5	5	1	7.3091	0.009
			6.8364	0.011
			5.1273	0.046
			4.9091	0.053
			4.1091	0.086
			4.0364	0.105
5	5	2	7.3385	0.010
			7.2692	0.010
			5.3385	0.047
			5.2462	0.051
			4.6231	0.097
			4.5077	0.100
5	5	3	7.5780	0.010
			7.5429	0.010
			5.7055	0.046
			5.6204	0.051
			4.5451	0.100
			4.5363	0.102
5	5	4	7.8229	0.010
			7.7914	0.010
			5.6657	0.049
			5.6429	0.050
			4.5229	0.099
			4.5200	0.101
5	5	5	8.0000	0.009
			7.9800	0.010
			5.7800	0.049
			5.6600	0.051
			4.5600	0.100
			4.5000	0.102

Tabla 5. Tabla de valores críticos de Ji cuadrada.

gl	Probabilidad conforme a H_0 de que $x^2 \geq$ Ji cuadrada													
	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	0.00016	0.00063	0.0039	0.016	0.064	0.15	0.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	0.02	0.04	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	0.12	0.18	0.35	0.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	0.30	0.43	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	0.55	0.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	0.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.78	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.82	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.57	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	18.31	18.10	21.24	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

Tabla 6. Valores de cuantiles de la distribución *t* de Student.

gl	Nivel de significación para prueba de una cola					
	.10	.05	.025	.01	.005	.0005
	Nivel de significación para prueba de dos colas					
	.20	.10	.05	.02	.01	.001
1	3.078	6.314	12.706	31.820	63.656	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291