

# Análisis de la Investigación Cuantitativa

## Métodos clásicos



Eduardo Jiménez Marqués

## 6. ANÁLISIS DE LOS DATOS

Una vez efectuadas correctamente todas las fases expuestas hasta el momento del proceso de Investigación Comercial pasamos a la siguiente: el análisis de los datos.

El análisis de los datos es la interpretación de los resultados obtenidos en las anteriores etapas de la investigación. Hemos de recordar que por muy excelente que sea este análisis, no se eliminan los posibles errores cometidos en las fases anteriores (sesgos, no respuestas y demás errores no muestrales) e, incluso, puede que éstos se acentúen más en los resultados que obtengamos en esta fase.

La cantidad de procedimientos estadísticos es muy amplia. El investigador deberá conocer el número suficiente de técnicas o herramientas que le permitan sacar utilidad de los datos así como conocer cuando es interesante su aplicación, tanto por el tipo de datos de los que dispone como por el tipo de resultados que le proporcionan. Podemos afirmar que cuantas más herramientas conozca el investigador, más posibilidades tendrá de realizar un buen análisis.

Antes de utilizar una determinada técnica hemos de tener en cuenta los siguientes aspectos:

El objetivo que perseguimos. El análisis de los datos persigue expresar de una forma sencilla y reducida las relaciones entre los datos con referencia al problema objeto de la Investigación Comercial. Estamos buscando respuestas al problema de la Investigación Comercial, no tratando de realizar una tesis doctoral o una investigación exhaustiva.

La escala básica de medida en que hemos obtenido los datos, ya que según se trate de una escala nominal, ordinal, de intervalo o de razón, las posibilidades de tratamiento varían.

El número de variables a analizar simultáneamente. Podemos considerar análisis de una sola variable (univariable), de dos variables (bivariable), o de más de dos variables o análisis multivariable.

En este capítulo vamos a estudiar una serie de técnicas para el análisis de una y dos variables, es decir, del análisis univariable y del bivariable.

## 6.1 MÉTODOS ESTADÍSTICOS CLÁSICOS

### 6.1.1 CONCEPTOS BÁSICOS

Vamos a reseñar brevemente algunos conceptos de la estadística aplicables a la Investigación Comercial a través de muestras.

#### 6.1.1.1 TIPOS DE ESTADÍSTICA

Normalmente las técnicas estadísticas se clasifican en tres grupos:

- La estadística **descriptiva** utiliza métodos numéricos y gráficos con el propósito de analizar el comportamiento, resumir y presentar la información contenida en un conjunto de datos, sin pretender generalizar los resultados obtenidos.
- La **estadística inferencial** (inferencia estadística) utiliza los datos de una muestra para hacer estimaciones, predicciones u otras generalizaciones sobre un conjunto de datos más amplio (la población), siendo una ayuda en el proceso de toma de decisiones. Al ser una inferencia (estimación o predicción), está sujeta a cierto error, con lo que se deberá calcular el nivel de confianza, medida de la seguridad con que se efectúa la inferencia.
- La **estadística relacional** utiliza los datos de una muestra para medir el grado de relación existente entre dos o más variables. Al ser también una inferencia estará sujeta a cierto error, con lo que tendrá un determinado nivel de confianza.

#### 6.1.1.2 MEDIR

Medir es asignar números o categorías a objetos, sucesos o casos, siguiendo reglas determinadas. De esta manera, en cada una de las unidades estudiadas se recoge información sobre una serie de características que varían de una unidad a otra. A la serie de características se les denomina variables o atributos. Cada variable toma un determinado valor en cada caso (por ejemplo: edad, estado civil, sexo, etc.).

#### 6.1.1.3 ATRIBUTO

Es el carácter de una población no susceptible de ser medida numéricamente. Las diferentes formas que puede presentar el atributo se denominan *modalidades*, también

# Análisis de la Investigación Cuantitativa

---

llamadas *variables cualitativas* por otros autores. El nombre de variable cualitativa es el más utilizado en la Investigación Comercial.

## 6.1.1.4 VARIABLE

En un sentido estadísticamente académico, una variable es cualquier carácter de una población o universo susceptible de tomar valores numéricos.

Las variables pueden ser continuas y discretas.

**Variable continua** es aquella en la que, entre dos valores determinados, siempre puede encontrarse otro valor. Por ejemplo, la edad; entre 46 y 47 años siempre podemos encontrar casos intermedios, por ejemplo 46 años y 5 meses, etc.

**Variable discreta** es aquella en la que, entre dos valores determinados, no se encuentra ningún otro. Por ejemplo, dos familias de 5 y 6 miembros (es imposible encontrar familias intermedias); número de personas en un hogar, (nunca se encontrarán valores intermedios como 5´4 miembros).

En la Investigación de Marketing muchas variables que son continuas (dinámicas), se discretizan; por ejemplo la edad.

*“En la práctica de la investigación comercial se acostumbra a hablar de variables cuantitativas y variables cualitativas o categóricas que corresponden, estas últimas, a lo definido como atributo”.*

## 6.1.2 ESCALAS DE MEDIDA

Para realizar mediciones se establecen escalas. Las escalas básicas más usuales son: nominal, ordinal, de intervalo y de razón.

### 6.1.2.1 ESCALA NOMINAL

En este tipo de escala se asigna un código o nombre a cada caso. La operación lógica que puede realizarse es el establecimiento de igualdad o desigualdad entre los diferentes casos.

Por ejemplo: la variable sexo, con sus categorías de: masculino y femenino; podemos decir que Pepe=Juan; y Tere  $\neq$  José. ( = igual.  $\neq$  no igual, diferente ).

## Análisis de la Investigación Cuantitativa

---

Al medir en escala nominal podemos obtener frecuencias (nº de repeticiones o de casos), porcentajes o frecuencias relativas, y si los resultados se expresan en tanto por uno se denominan proporciones. El porcentaje también se denomina frecuencia relativa.

Ejemplo:

Sexo	Código	Frecuencia	Porcentaje	Proporción
Masculino	1	70	60'87	0'6087
Femenino	0	45	39'13	0'3913
	Total	115	100	1

### 6.1.2.2 ESCALA ORDINAL

En esta escala, además de las relaciones de igualdad y desigualdad, se establece un orden lógico de categorías, por ejemplo, clase social: alta, media-alta, media-media, media-baja, y baja.

Entendemos de esta manera que un individuo de la clase alta está mejor situado que otro de la clase media-alta, y así sucesivamente.

En la escala ordinal, además de asignar un nombre, código o categoría a cada caso o individuo (igual que se hace en la escala nominal), se establece un orden lógico entre las categorías.

A pesar de establecerse un orden, no puede decirse que la distancia o diferencia entre diferentes categorías sea la misma. Es decir no existe igualdad entre los intervalos.

Ejemplo: clase social (alta, media alta, media, media baja y baja)

No podemos decir que la diferencia entre la clase social alta y la media alta sea la misma que la existente entre la clase media baja y la baja.

Clase social	Código
Alta	1
Media alta	2
Media	3
Media baja	4
Baja	5

### 6.1.2.3 ESCALA DE INTERVALO

Se establecen las relaciones de igualdad / desigualdad, así como el orden lógico de categorías, y la igualdad de intervalos. Para ello es preciso fijar una unidad de medida constante y uniforme a lo largo de todos los posibles valores de la variable.

El punto cero es un valor consensuado, pero que no tiene un significado de ausencia total del aspecto medido. Por ejemplo: la diferencia entre 21°C y 28°C es la misma que entre 31°C y 38°C.

Es decir, podemos establecer: igualdad y desigualdad  $20^{\circ}\text{C} = 20^{\circ}\text{C}$ , y  $25^{\circ}\text{C} \neq 18^{\circ}\text{C}$ .

Existe orden  $20^{\circ} > 19^{\circ} > 18^{\circ}$ , etc.

Igualdad de intervalo, es decir el aumento de 1°C, representa lo mismo a 20°C que a 4°C.

Y el origen o cero de la escala, ( 0°C ), es un valor consensuado que no significa la ausencia de valor. Recordar el conocido chiste ¡Qué bien estamos a 0<sup>0</sup> C! No hace “ ni frío ni calor”.

### 6.1.2.4 ESCALA DE RAZÓN O PROPORCIÓN

Aparte de todas las propiedades de las escalas de intervalo, se establece un cero absoluto, que representa la ausencia total del aspecto que se está valorando. La existencia de este cero absoluto es un requisito necesario para poder realizar comparaciones mediante cocientes. Las operaciones matemáticas que utiliza el hombre con mayor frecuencia para realizar comparaciones son la resta y la división.

Es decir, en las escalas de razón o proporción existen las relaciones de igualdad, orden, igualdad de intervalos y además, el cero absoluto.

### 6.1.2.4.1 RESUMEN ESCALAS BÁSICAS DE MEDIDA

En la tabla siguiente se resumen las características y relaciones de las diferentes escalas

Escala	Relaciones que la definen
Nominal	Equivalencia
Ordinal	Equivalencia Mayor que
De intervalo	Equivalencia Mayor que Razón conocida entre cualesquiera intervalos. Cero consensuado
De razón	Equivalencia Mayor que Razón conocida entre cualesquiera intervalos Razón conocida entre cualesquiera de dos valores de la escala Cero como ausencia de la característica que se mide

### 6.1.3 ELEMENTOS DE LA ESTADÍSTICA INFERENCIAL

La estadística inferencial tiene los siguientes elementos característicos:

**Población:** un determinado conjunto de elementos (personas, objetos, transacciones, sucesos...) de los que queremos estudiar una característica o propiedad (por ejemplo, ¿cuál es el consumo medio de...?, o ¿cuál es el porcentaje de consumidores de...?).

**Muestra:** un subconjunto de elementos de la población. Son los que realmente estudiamos, de los que obtenemos información.

**Estimador o inferencia estadística:** una estimación, predicción u otra generalización sobre la característica estudiada de la población, basada en la información contenida en la muestra.

**Nivel de confianza:** una medición de la seguridad con que se efectúa la inferencia (estimación o predicción).

## 6.2 ANÁLISIS UNIVARIABLE

Vamos a reseñar esquemáticamente las técnicas estadísticas utilizadas en el estudio de una sola variable.

### 6.2.1 DESCRIPCIÓN DE LOS DATOS

Para obtener una descripción de los datos se realizan básicamente las siguientes mediciones:

1. Frecuencias, absolutas y relativas
2. Medidas de tendencia central
3. Medidas de dispersión
4. Medidas en cuanto a la forma de distribución, coeficientes de asimetría y apuntamiento o curtosis
5. Otras mediciones como ratios o relaciones por cociente, normalización de la variable y medidas de concentración tales como curva de Lorenz, índice de Gini o medial o mediana.

A continuación, resumimos las cuatro primeras de estas medidas<sup>3</sup>:

### 6.2.2 FRECUENCIA

Es el número de veces que se presenta un determinado valor de una variable. Se distinguen dos tipos de frecuencias.

A la que nos viene indicada por el número de veces o casos que presenta una variable objeto de estudio, se le denomina frecuencia absoluta o simplemente frecuencia.

Si calculamos el porcentaje (porcentaje es la proporción en que contestaron o eligieron una modalidad, por cien), se denomina frecuencia relativa.

En la tabla siguiente resumimos las principales características

---

<sup>3</sup>Para las muestras utilizaremos letras latinas, mientras que cuando nos refiramos al universo se utilizarán griegas.



## Análisis de la Investigación Cuantitativa

---

### TABLA: CONCEPTOS RELACIONADOS CON LA DISTRIBUCIÓN DE FRECUENCIAS

**Frecuencia absoluta (*frequency*):** número de veces que se repite cada valor en el conjunto de los datos.

**Porcentaje o frecuencia relativa (*percent*):** frecuencia absoluta dividida por el número total de observaciones. Se suele expresar en tantos por ciento. Por tanto, es el porcentaje de repeticiones de un determinado valor en el total de la muestra, pero incluye el porcentaje correspondiente a los valores perdidos (*missing cases*).

**Porcentaje válido (*valid percent*):** frecuencia absoluta dividida por el número de observaciones con información válida. Se excluyen del cálculo los valores perdidos, por lo que indica el porcentaje de respuestas sobre el total de la información conocida.

**Porcentaje acumulado (*cum percent*):** porcentaje de observaciones que están por encima o por debajo de cierto valor.

**Intervalos de clase:** grupo de observaciones con valores semejantes. El intervalo queda definido por un valor mínimo, un valor máximo y una marca de clase (el valor central). Por regla general, todas las clases que se definen tienen la misma amplitud (distancia entre el mínimo y el máximo de cada clase).

**Frecuencias absolutas y relativas de un intervalo de clase:** número o porcentaje de observaciones con un valor que pertenece al intervalo de clase.

Si el valor máximo y el valor mínimo de dos clases consecutivas coinciden, se considera que dicho valor concreto es de la clase superior (donde es mínimo).

### 6.2.3 MEDIDAS DE TENDENCIA CENTRAL

También se denominan medidas de posición. Son las que resumen la información aportada por todos los elementos objeto de estudio en un valor central. La tendencia central nos da una descripción concisa del promedio o funcionamiento característico del grupo como un todo.

Las medidas de la tendencia central nos permiten comparar dos o más grupos en función de su funcionamiento característico.

Existen varias medidas de tendencia central. Su elección se basa en el tipo de escala que utilicemos. Recordemos que estas escalas son las siguientes: métrica, ordinal y nominal.

Investigación Comercial

Toda la información medida en una escala superior, se puede convertir en otra inferior.

Las medidas de posición o de tendencia central son las siguientes:

La media, la moda y la mediana.

*La media* es el promedio aritmético de los valores obtenidos por la variable.

$$m = \sum_{i=1}^n \frac{x_i}{n}$$

Donde  $m$  es la media de la muestra,  $n$  es el tamaño de la muestra y  $x_i$  cada valor que toma la variable “ $i$ ”

*La moda* es el valor o intervalo de la variable que más se repite, el valor más frecuente.

*La mediana* es el valor de la variable que deja, por encima y por debajo, el 50% de los casos. En la práctica es difícil encontrar un valor de la variable que parta la distribución de frecuencias en dos. Por ello, formalmente, la mediana es el mínimo de los valores tal que *la frecuencia acumulada es igual o mayor al 50%*. Los datos deben estar ordenados en orden ascendente o descendente.

## **Consideraciones generales acerca de las medidas de posición o de tendencia central:**

La moda es la única medida de posición que podemos aplicar a los datos medidos en escala nominal.

Para poder calcular la mediana se necesita, al menos, disponer de datos medidos en una escala ordinal, no siendo posible su cálculo en el caso de datos medidos en escala de tipo nominal.

En el cálculo de la media se precisa disponer, al menos, de una unidad de medida. Esto sólo lo cumplen las métricas, es decir, las escalas de intervalo y las de proporción.

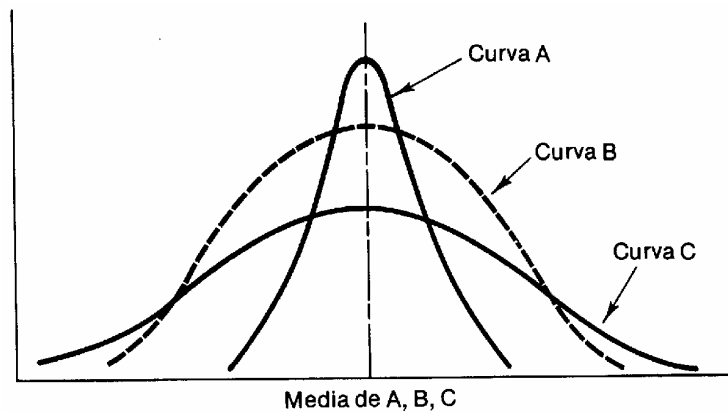
## **6.2.4 MEDIDAS DE DISPERSIÓN**

Las medidas de tendencia central o de posición, aunque son muy útiles para resumir la información contenida en una variable, no aportan toda la realidad de la misma. Con el fin de dar solución a este problema se diseñaron las medidas de dispersión.

Las medidas de dispersión están diseñadas para dar información sobre el error que se comete al considerar todos los casos iguales a la tendencia central.

Existen muchas posibilidades de calcular la dispersión de una variable. Las medidas de dispersión más usuales son: amplitud o rango, la desviación intercuartil, la varianza, la desviación típica y el coeficiente de variación.

Las medidas de tendencia central, la media, la mediana y la moda, sólo facilitan una parte de la información contenida en los datos. Para conocer mejor el comportamiento de las observaciones es conveniente conocer también su **dispersión o variabilidad** (cómo varían los valores de la variable o hasta qué punto son todos muy parecidos o muy diferentes).



### 6.2.4.1 AMPLITUD O RANGO

Es la diferencia entre el valor más pequeño y el más alto de la variable en estudio. Por ejemplo, si tenemos los siguientes datos para una variable

5, 7, 8, 9, 15, 17, la amplitud sería  $17 - 5 = 12$

La amplitud también se llama recorrido; es una buena medida de dispersión por su claro significado, dependiendo sólo de los dos valores extremos. En el caso de una variable muy concentrada, pero con dos extremos muy diferenciados, obtenemos una medida de dispersión poco representativa al calcular la amplitud.

## 6.2.4.2 RECORRIDO INTERCUARTÍLICO

Se define como la diferencia entre el tercer y el primer cuartil.

Viene dado por la siguiente fórmula:

$$R_I = Q_3 - Q_1$$

## 6.2.4.3 LA DESVIACIÓN INTERCUARTIL

Es la diferencia entre los dos valores que ocupan los percentiles 25% y 75% dividida por dos.

**Tabla: RESUMEN MEDIDAS ELEMENTALES DE DISPERSIÓN**

**Mínimo (*mínimum*):** es el valor más bajo de las observaciones.

**Máximo (*máximum*):** es el valor más alto de las observaciones.

**Rango o amplitud (*range*):** es la diferencia ente el máximo y el mínimo.

### FRACTILES

**Fractil:** valor por debajo del cual se encuentra una fracción o proporción de los datos.

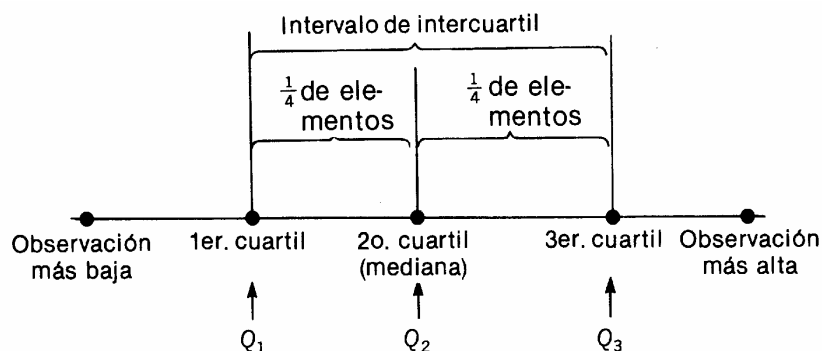
**Percentil:** la variable se ordena de menor a mayor, entonces se divide en 100 partes iguales (100 grupos con el mismo número de observaciones) y se estudia el valor que alcanza en cada parte. Ejemplo: el percentil al 50% es la mediana.

**Cuartil:** los datos se dividen en 4 partes iguales, obteniendo:

$$Q_1, Q_2, Q_3 \text{ y } Q_4$$

(Q2 es la mediana y Q4 coincide con el máximo).

Rango o amplitud intercuartil: diferencia entre Q3 y Q1.



Entre el primer y tercer cuartil, en la amplitud intercuartil, se encuentran el 50% de las observaciones centrales.

### 6.2.4.4 DESVIACIÓN MEDIA

Se define como la media aritmética de las desviaciones entre los valores de la variable y la media aritmética, en valor absoluto. Su fórmula es:

$$DM = \frac{\sum_{i=1}^n |x_i - m|}{n}$$

Si todas las observaciones fueran iguales, la DM sería cero. Cuanto mayor sea el valor de la DM mayor será el grado de dispersión.

Esta medida es poco utilizada debido a que no se puede manipular algebraicamente.

### 6.2.4.5 VARIANZA Y DESVIACIÓN TÍPICA O ESTÁNDAR

La varianza es una medida de dispersión de variables métricas.

Si pretendemos evaluar exactamente cual es el error que cometemos al asignar el valor de la media a todos los casos de la distribución de la variable, hay que calcular el promedio de las diferencias con la media de todos los valores.

Para poder hacer un promedio, necesitamos una unidad de medida; por consiguiente, la variable deberá estar, al menos, en escala de intervalo.

La varianza es el promedio de las diferentes desviaciones de variable respecto de la media al cuadrado. Se representa mediante  $S^2$  ó  $\sigma^2$ , según se refiera a la muestra o al universo. Su fórmula es:

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

La desviación típica es la raíz cuadrada de la varianza, el error promedio que cometemos al asignar la media a cada caso. La desviación típica es la distancia de los distintos valores de la variable a la media.

Cuanto menor es la desviación típica, más representativa y más próxima a la realidad es la representación de la variable mediante la media.

La desviación típica presenta la siguiente propiedad:

La desviación típica nos permite realizar rápidamente cálculos sobre entre qué valores se encuentran el 95,5% de los casos, en las variables que siguen una distribución normal (la mayor parte de las variables sigue este tipo de distribución). Para realizar el cálculo,

Investigación Comercial

## Análisis de la Investigación Cuantitativa

---

hay que sumar y restar dos veces la desviación típica al valor de la media. Se representa por:

$$m \pm 2s \quad \text{o bien} \quad \mu \pm 2s$$

Según nos refiramos a la muestra o al universo o población.

### TABLA: CONCEPTOS VARIANZA Y DESVIACIÓN TÍPICA

**Varianza (*variance*) de la población:** suma de cuadrados de las distancias entre la media y cada elemento, dividido por el número total de observaciones de la población. Es una distancia promedio a la media.

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

**Varianza de la muestra:** es equivalente a la de la población, pero se divide por el número de observaciones de la muestra menos 1 (ya que la media de la muestra es un dato conocido):

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

**Desviación estándar o típica (*Std deviation*):** es la raíz cuadrada positiva de la varianza, ya sea de la población o de la muestra.

Está medida en las mismas unidades que la variable estudiada.

### 6.2.4.6 COEFICIENTE DE VARIACIÓN

Es una medida de dispersión relativa, que se utiliza para comparar variables con diferentes unidades de medida.

Consiste en dividir la desviación típica por la media y después expresarla en porcentaje.

Este coeficiente está exento de unidades.

Investigación Comercial

Su fórmula es:

$$CV = \frac{s}{m} \times 100$$

Tiene que ser  $m \neq 0$ .

El coeficiente de variación tiene la ventaja de que pierde las unidades de medida, por lo que puede servir para comparar las dispersiones relativas de dos variables que tienen diferente unidad de medida.

### 6.2.5 MEDIDAS RELATIVAS A LA FORMA DE LA DISTRIBUCIÓN

#### 6.2.5.1 DISTRIBUCIÓN

Una distribución es una serie de valores separados, colocados u ordenados según una magnitud. Un conjunto de códigos ordenados y sus frecuencias correspondientes se llaman “distribución de frecuencias”.

#### 6.2.5.2 COEFICIENTE DE SESGO O ASIMETRÍA (SKEWNESS)

Con las medidas de asimetría se intenta medir si las observaciones están dispuestas simétrica o asimétricamente respecto a un valor central, generalmente la media aritmética.

El coeficiente más utilizado es el de R. A. Fisher, que viene expresado por la siguiente fórmula:

$$CA = \frac{\sum_{i=1}^n (x_i - m)^3}{ns^3}$$

Se dan las siguientes alternativas:

$CA = 0$  se trata de una distribución simétrica.

$CA > 0$  es asimétrica a la derecha. Positiva.

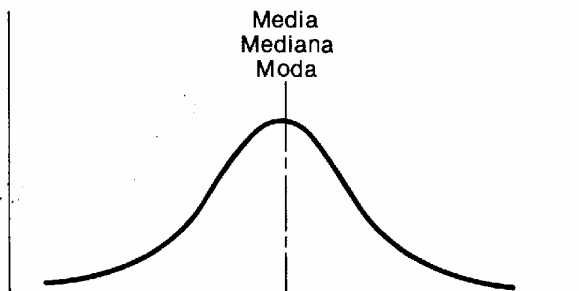
$CA < 0$  es asimétrica a la izquierda. Negativa.

Las distribuciones **simétricas o insesgadas** son aquellas que dejan el mismo número de observaciones a la izquierda que a la derecha de la media.

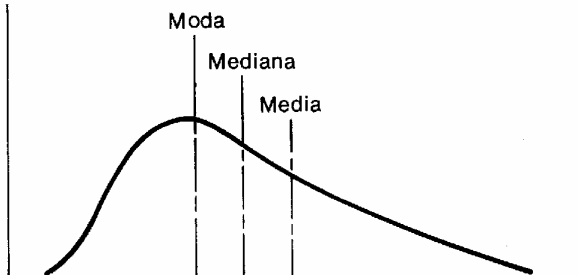
La simetría se mide a través del **sesgo** (*skewness*).

En la gráfica siguiente exponemos las correspondientes situaciones<sup>4</sup>

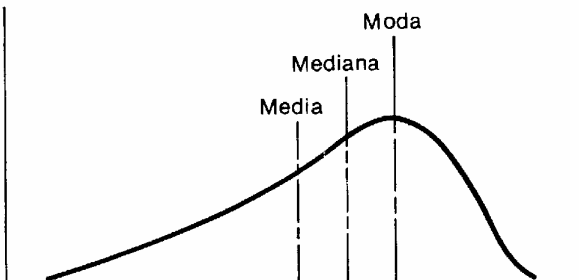
Por tanto, cuando la distribución es simétrica, sesgo = 0, la media es igual a la mediana (e iguales a la moda, si sólo existe una):



Si la mediana es más pequeña que la media la distribución está sesgada a la derecha (hacia valores grandes), sesgo > 0:



Si la mediana es más grande que la media la distribución está sesgada a la izquierda (hacia valores pequeños), sesgo < 0:



### 6.2.5.3 COEFICIENTE DE CURTOSIS O APUNTAMIENTO

Las medidas de apuntamiento o curtosis son aplicables a distribuciones campaniformes, unimodales simétricas (no en U) o con una ligera asimetría. Se mide la mayor o menor concentración de las frecuencias alrededor de la media, es decir, su nivel de apuntamiento. Se toma como distribución tipo la distribución normal. La distribución normal cumple que:

$$m_4 = 3 s^4 \quad s = 1 \text{ y } m = 0$$

Donde  **$m_4$**  es el momento de orden 4 respecto de la media o momento central que viene dado por la expresión siguiente:

---

<sup>4</sup> Universidad Autónoma Barcelona. Apuntes de Cátedra Investigación Comercial I Teresa Obis  
Investigación Comercial



$$m_4 = \frac{\sum_{i=1}^n (x_i - m)^4}{n}$$

El coeficiente de curtosis viene dado por la siguiente expresión:

$$CK = \frac{m_4}{s^4} = \frac{\sum_{i=1}^n (x_i - m)^4}{ns^4}$$

Como CK para la distribución normal es igual a 3 se suele usar el coeficiente corregido, también llamado coeficiente de exceso y que es:

$$CK_2 = \frac{\sum_{i=1}^n (x_i - m)^4}{ns^4} - 3$$

Los valores que toma son:

$CK_2 = 0$  es decir  $CK = 3$  Es una distribución mesocúrtica, sin exceso.

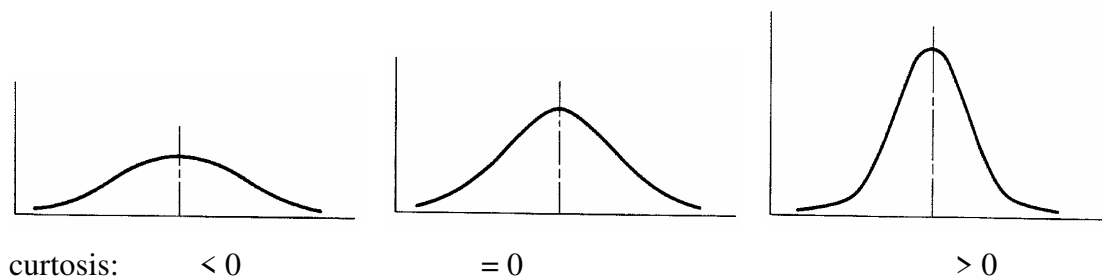
$CK_2 > 0$  es decir  $CK > 3$  Distribución leptocúrtica. Con exceso. Puntigrada.

$CK_2 < 0$  es decir  $CK < 3$  Distribución platicúrtica. Es decir achatada

En el gráfico siguiente se pueden observar las diferentes formas de la “curtosis” o apuntamiento

**Curtosis:** grado de apuntamiento (pico) de la distribución. Una distribución en "forma de campana" tiene una curtosis igual a 0.

Si está más concentrada (puntigrada) la curtosis es  $> 0$ ; si es más plana (con mayor dispersión) la curtosis es  $< 0$ .



## Análisis de la Investigación Cuantitativa

---

### 6.2.6 ¿CÓMO REALIZAR INFERENCIAS?

Cuando lo que pretendemos es comparar los valores obtenidos en la investigación con otros predeterminados o bien realizar inferencias, lo que se hace es aplicar determinadas pruebas o test en función de la escala de medida de la variable objeto de estudio. Lo podemos resumir en la siguiente tabla:

Escala	Frecuencia	Tendencia central	Dispersión	Prueba
Nominal	Absoluta, relativa	Moda		Chi cuadrado, binomial
Ordinal	Absoluta, relativa, acumulación, percentiles	Moda, Mediana	Desviación intercuartil	
Intervalo	Absoluta, relativa, acumulación, percentiles	Moda, Mediana Media <sup>5</sup>	Desviación intercuartil Varianza, desviación típica	
Proporción razón	Absoluta, relativa, acumulación, percentiles	Moda, Mediana Media	Desviación intercuartil Varianza, desviación típica Coeficiente de variación	

---

<sup>5</sup> En ocasiones no tiene ningún significado, debido al desconocimiento de los intervalos de clase con exactitud.

## 7. CONTRASTE DE HIPÓTESIS

### 7.1 CONCEPTOS GENERALES

#### 7.1.1 HIPÓTESIS

La palabra “hipótesis” tiene su origen en los términos griegos *thesis*, que significa “lo que se pone”, e *hipo*, que equivale a “debajo”. Hipótesis es, por tanto, lo que se pone debajo, es decir lo que se supone.

Desde el punto de vista de la investigación comercial, las hipótesis se pueden definir como soluciones probables previamente seleccionadas al problema planteado, que se tienen que confirmar a través del proceso de investigación con los hechos.

La primera prueba a la que debemos someter una idea nueva es a la de su coherencia. Una vez superado este primer paso, la idea o conjetura pasa a denominarse hipótesis y deberemos contrastarla a través de la denominada prueba de hipótesis.

#### 7.1.2 PRUEBAS DE HIPÓTESIS

Para K. R. Popper en su obra *El desarrollo del conocimiento científico* (Ediciones Paidós, 1967), una hipótesis es “una conjetura expresada en términos de alto contenido informativo”.

Para que la hipótesis se convierta en teoría deberá ser validada mediante pruebas de hipótesis. Estas pruebas actúan como un filtro, eliminando las hipótesis falsas y manteniendo las verdaderas. Podemos afirmar que cuantas más pruebas haya superado una hipótesis, más convencimiento tendremos acerca de su veracidad.

En la praxis se ha escogido que el nivel de filtro sea del 95%, es decir, se eliminan el 95% de las hipótesis falsas.

En el campo científico es de gran importancia la repetición de la prueba de hipótesis por diferentes equipos de investigadores y en diferentes partes del mundo. Esto significa que cuantas más oportunidades se han dado a una hipótesis de ser abandonada y más veces ha superado las pruebas, más seguros estaremos de su veracidad.

En el campo de la Investigación Comercial hemos de tener presente que trabajamos con muestras, que los estadísticos son aleatorios y que, para el estadístico, muestras distintas pueden generar diferentes valores.

Una hipótesis estadística es una afirmación con respecto a alguna característica de interés. En las pruebas de hipótesis hay que depurar los resultados mediante estudios más profundos.

Contrastar una hipótesis es decidir si se rechazan o no unos determinados supuestos que planteamos acerca de un universo o población, partiendo de los datos obtenidos con una muestra, midiendo a la vez el riesgo de error correspondiente a cada una de las posibles decisiones.

Para realizar la prueba de hipótesis tenemos que definir previamente la hipótesis nula ( $H_0$ ) y la hipótesis alternativa ( $H_1$ )

## 7.1.2.1 HIPÓTESIS NULA

Es la hipótesis que se realiza sobre un determinado fenómeno con el propósito de comprobarla o rechazarla. Se representa por  $H_0$ .

## 7.1.2.2 HIPÓTESIS ALTERNATIVA

Cualquier otra hipótesis que se formule diferente a la hipótesis nula. La hipótesis alternativa se representa por  $H_1$ .

Ejemplo:

Partimos de dos supermercados, A y B; queremos demostrar que la diferencia de sus ventas es debida a la diferencia de permanencia de los clientes en el establecimiento.

Para ello, medimos el tiempo medio de estancia en la tienda de los clientes de A y B.

Definiremos las siguientes hipótesis:

$H_0$  : las medias de estancia, de los clientes, en la tienda son iguales  $m_A = m_B$

$H_1$  : las medias de estancia en la tienda son diferentes  $m_A \neq m_B$

## 7.1.3 CARACTERÍSTICAS Y METODOLOGÍA

Las principales características de las hipótesis nula y alternativa son:

- Una y sólo una de las dos puede ser cierta

# Análisis de la Investigación Cuantitativa

---

- Mientras  $H_0$  indica una única posibilidad, la hipótesis alternativa comprende infinitas
- Lo que creemos y tratamos de demostrar es la hipótesis alternativa
- Si podemos sospechar que la hipótesis nula no es cierta, entonces obligatoriamente deberá ser cierta la hipótesis alternativa

La metodología para realizar un contraste de hipótesis consiste en la realización de los siguientes pasos:

- Definir  $H_0$  y  $H_1$
- Definir un parámetro de medida que relacione los parámetros muestrales y poblacionales y la función de probabilidad
- Establecer un criterio para juzgar si el valor calculado del parámetro es compatible con  $H_0$
- Realizar el muestreo
- Obtener conclusiones de conformidad con lo expuesto

Con el fin de comprender mejor los diferentes conceptos, vamos a realizar un ejemplo sencillo y clásico.

Para nuestro estudio, utilizaremos una moneda cargada.

Las correspondientes hipótesis serían:

Hipótesis nula ( $H_0$ ). La probabilidad de que salga cara es 0'5, es decir, la moneda no está cargada  $H_0 : p = 0'5$ .

Hipótesis alternativa ( $H_1$ ). La moneda esta cargada. Esto es,  $H_1 : p \neq 0'5$ .

Para elegir una de las dos hipótesis lanzamos la moneda al aire 100 veces.

Si  $H_0$  fuera cierta el resultado que obtendríamos sería 50 veces cara y 50 veces cruz.

Supongamos que obtenemos cara en 51 ocasiones y cruz en 49. Esta pequeña diferencia sería totalmente explicable por el azar o la suerte. Por consiguiente, nuestra conclusión sería que no hemos encontrado nada que nos haga sospechar que la moneda esté cargada, luego no podríamos decir que la hipótesis nula fuese falsa (no podríamos rechazar  $H_0$ ).

Supongamos que al lanzar 100 veces la moneda, obtenemos 97 veces cara y 3 cruz. La diferencia es lo suficientemente significativa como para hacernos dudar de la validez de

la hipótesis nula y afirmaríamos que la moneda está cargada o bien que abandonamos la hipótesis nula  $H_0$  (se puede rechazar la  $H_0$ ).

Pero supongamos que después de 100 lanzamientos de la moneda, hemos obtenido 68 caras y 32 cruces. En esta situación unas personas no rechazarían la  $H_0$ , considerando que la desviación es fruto del azar, mientras que otras sí que rechazarían la  $H_0$ .

Con el fin de dar solución a este tipo de situaciones se establece un criterio homogeneizador. La estadística propone como criterio rechazar la hipótesis nula  $H_0$  cuando la probabilidad de un resultado tanto más extremo, más pequeña que un cierto valor escogido *a priori*. Basándose en la experiencia se acepta que ese valor sea 0'05.

Una vez aceptado ese valor deberemos calcular las probabilidades de los diferentes resultados muestrales suponiendo cierta la  $H_0$ .

Teniendo en cuenta que el producto de  $p * n$  es mayor que 5, la proporción  $p$  obtenible en las infinitas muestras sigue una distribución normal de media  $p$  y desviación típica

$$S_p = \sqrt{\frac{p(1-p)}{n}}$$

siendo  $n$  el número de casos de la muestra.

En el ejemplo tenemos que  $H_0$  nos dice que  $p = 0'5$ . Si lanzamos la moneda 100 veces tendríamos que

$$S_p = \sqrt{\frac{0'5(1-0'5)}{100}} = 0'05$$

Se simboliza como  $p \rightarrow n(0'5, 0'05)$

Esto significa que la proporción de caras obtenidas en una muestra de 100 casos sigue una distribución normal de media 0'5 y desviación típica 0'05.

Siguiendo las propiedades de la distribución normal vamos a calcular entre qué dos valores se situarán el 95% de los resultados obtenidos con las infinitas muestras de 100 lanzamientos de la moneda.

Si recordamos que para el valor del 95% la razón crítica  $Z = 1'96$ . Aplicando la propiedad de la distribución normal  $\mu \pm Z\sigma$  obtenemos

$0'5 \pm 1'96 * 0'05$ , lo cual nos conduce a los siguientes valores  $p_1 = 0'402$  y  $p_2 = 0'598$

## Análisis de la Investigación Cuantitativa

---

Por consiguiente, si es cierta la  $H_0$ , la probabilidad de encontrar una proporción muestral de caras entre 0'402 y 0'598 será igual al 95% (0'95).

La probabilidad de encontrar valores fuera de ese intervalo será de 0'05.

De conformidad con lo expuesto, estableceríamos un sistema de decisión del tipo siguiente:

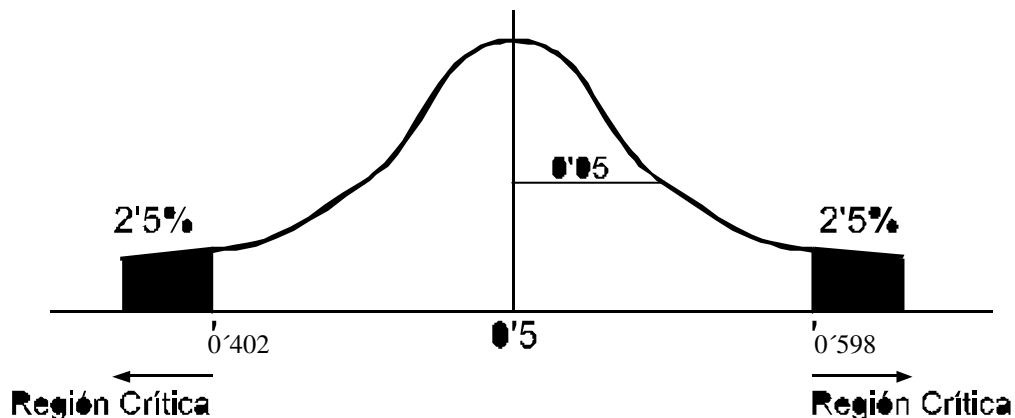
Si la proporción de caras, al lanzar 100 veces la moneda, se sitúa entre 40'2 y 59'8%, la  $H_0$  será cierta.

Si la proporción se sitúa por debajo del 40'2 y por encima del 59'8, se dirá que la probabilidad de observar estos resultados es tan pequeña (0'05) como para sospechar que la  $H_0$  es falsa.

A la zona extrema y alejada, que tiene una pequeña probabilidad de ocurrir si es cierta la  $H_0$  y que si se observa un resultado en ella nos permite rechazar la hipótesis nula, se la denomina *región crítica*.

La representación gráfica para nuestro ejemplo es la siguiente:

### Gráfica



### Esquema de decisión

Si el resultado se sitúa en la región crítica se rechaza la hipótesis nula  $H_0$

Si el resultado no se sitúa en la región crítica, no se rechaza la hipótesis nula  $H_0$ , rechazando entonces la hipótesis alternativa  $H_1$ .

## 7.2 TEST DE HIPÓTESIS

También se denomina test de significación o contraste. Se utiliza para designar el procedimiento que se utiliza para contrastar la validez de una hipótesis. Son pruebas estadísticas que se utilizan para determinar si los resultados obtenidos con una muestra o dos elegidas al azar difieren marcadamente de aquellos que habría que esperar con la hipótesis planteada y la variación debida al muestreo.

### 7.2.1 OBJETIVO

El objetivo de este tipo de test es:

Observar si la variable estudiada se comporta aleatoriamente

Verificar si la media de la muestra estudiada pertenece a la media del universo estudiado

Comprobar si la proporción obtenida en el estudio muestral es la misma que la de la población

Ver si los datos obtenidos en la muestra corresponden a algún tipo de distribución conocida

Observar si los valores obtenidos siguen unos patrones esperados

## 7.3 METODOLOGÍA DEL TEST DE HIPÓTESIS

Tal y como indicamos en el punto anterior, el test de hipótesis es una forma de decidir, de manera objetiva, si los resultados obtenidos a través de la Investigación Comercial muestran una realidad o son simplemente una consecuencia de la aleatoriedad de la muestra.

En este proceso se diferencian las siguientes etapas:

1. Formulación de las hipótesis
2. Elección del nivel de significación
3. Elección de la prueba o test
4. Interpretación



## 7.3.1 FORMULACIÓN DE LAS HIPÓTESIS

Desde el punto de vista de la investigación comercial, las hipótesis se pueden definir como soluciones probables previamente seleccionadas al problema planteado, que se tienen que confirmar con los hechos en el proceso de investigación.

Una hipótesis es, por tanto, una proposición relativa a un problema, a la forma o a los parámetros de una distribución. Cuando se refiere a un valor concreto para un parámetro, se denomina hipótesis simple. Si se refiere a un intervalo para un parámetro, hipótesis compuesta.

En el campo de la Investigación Comercial hemos de tener presente que trabajamos con muestras, y que los estadísticos son aleatorios y que muestras distintas pueden generar diferentes valores para el estadístico.

Una hipótesis estadística es una afirmación con respecto a alguna característica de interés. En las pruebas de hipótesis se trata de depurar los resultados mediante estudios más profundos.

Contrastar una hipótesis es decidir si se rechazan o no como ciertos los supuestos que planteamos acerca de un universo o población, partiendo de los datos obtenidos con una muestra, midiendo a la vez el riesgo de error correspondiente a cada una de las posibles decisiones.

El proceso de la prueba o test de hipótesis comienza definiendo la hipótesis, que se confirmará o rechazará de conforme al resultado que obtengamos en el test.

Para realizar la prueba de hipótesis tenemos que definir previamente la hipótesis nula  $H_0$  y la hipótesis alternativa  $H_1$  que, recordemos de nuevo, son:

Hipótesis nula. Hipótesis que se realiza sobre un determinado fenómeno con el propósito de comprobarla o rechazarla<sup>6</sup>. Se representa por  $H_0$ . Se refiere a determinadas características de la población, denota ausencia de diferencias y es la que se mantendrá en el caso de que los resultados de la prueba o test no muestren falsedad.

---

<sup>6</sup>Siendo estrictos en vez de decir se acepta la hipótesis nula debemos decir “no se rechaza la hipótesis nula”

Hipótesis alternativa. Cualquier otra hipótesis que se formule diferente a la hipótesis nula. Se representa por  $H_1$ . Denota la existencia de diferencia; es una forma de negación de la  $H_0$ .

La formulación de ambas hipótesis debe hacerse de forma muy rigurosa y precisa.

Las principales características de las hipótesis nulas y alternativas son:

1. Una y solamente una de las dos puede ser cierta
2. Mientras  $H_0$  indica una única posibilidad, la hipótesis alternativa comprende infinitas
3. Lo que creemos y tratamos de demostrar es la hipótesis alternativa
4. Si podemos sospechar que la hipótesis nula no es cierta, entonces obligatoriamente deberá ser cierta la hipótesis alternativa

En función de la  $H_0$  se van a definir los valores esperados (teóricos). La comparación entre estos valores y los observados nos proporcionaran la decisión de acuerdo con las normas preestablecidas.

La hipótesis se puede enunciar de forma unidireccional (de una cola), si el sentido o dirección de la hipótesis alternativa es conocido. Ejemplo. la cuota de audiencia de una cadena local de TV es superior al 25%. Las hipótesis las podemos formular

$$H_0 : P \leq 25\% \text{ y la alternativa será } H_1 : P > 25\%$$

También se pueden formular de forma bidireccional, bilateral o test de dos colas (Two-tailed test), si la hipótesis alternativa puede tomar cualquier sentido. Siguiendo con el ejemplo anterior, supongamos que conocemos que la cuota es diferente del 25%, entonces la formulación será como sigue:

$$H_0 : P = 25\% \text{ y la alternativa será } H_1 : P \neq 25\%$$

En Investigación Comercial se suele utilizar más el test de una cola, ya que se suele conocer el sentido de la afirmación. El bilateral se usa cuando no hay preferencia sobre la dirección del resultado.

## 7.3.2 ELECCIÓN DEL NIVEL DE SIGNIFICACIÓN

En el mundo científico, para que la hipótesis se convierta en teoría deberá ser validada mediante pruebas de hipótesis. Estas pruebas actúan como un filtro, eliminando las hipótesis falsas y manteniendo las verdaderas. Podemos afirmar que cuantas más pruebas haya superado una hipótesis, más convencimiento tendremos acerca de su veracidad.

## Análisis de la Investigación Cuantitativa

---

En la praxis se ha escogido que el nivel de filtro sea del 95%, es decir, se eliminan el 95% de las hipótesis falsas.

En las pruebas de hipótesis se presentan cuatro situaciones que resumimos en la siguiente tabla:

	<b>H<sub>0</sub> es verdadera</b>	<b>H<sub>0</sub> es falsa</b>
<b>No rechazar H<sub>0</sub></b>	1. Decisión correcta Nivel de confianza Probabilidad $p = 1 - \alpha$	3. Error tipo II Probabilidad $p = \beta$
<b>Rechazar H<sub>0</sub></b>	2. Error tipo I Nivel de significación Probabilidad $p = \alpha$	4. Decisión correcta Poder de prueba Probabilidad $p = 1 - \beta$

1. La decisión correcta de no rechazar  $H_0$  cuando es verdadera. El nivel de confianza proporciona el porcentaje de situaciones en los que la hipótesis nula se aceptaría siendo verdadera. No se refiere a la probabilidad de que los resultados sean observados en términos de muestreo, sino a la probabilidad de que la hipótesis sea verdadera con los datos obtenidos. La elección del nivel de significación y, por consiguiente, del nivel de confianza, es convencional y se fija a priori. En investigación comercial se suele trabajar con valores de nivel de significación  $\alpha = 0.01$  (1%) y  $\alpha = 0.05$  (5%). Los niveles de confianza vendrán dados por  $1 - \alpha$ . Es decir, fijamos a priori la probabilidad  $\alpha$  de rechazar la  $H_0$  siendo cierta. También en ocasiones se emplean niveles de significación menos exigentes, pudiendo llegar a veces hasta valores de  $\alpha$  del 10%, que se denominan cuasi significativos. Rara vez se pasa de este valor del 10% para el nivel de significación.
2. Error de tipo I. Es el riesgo que se asume. El nivel de significación  $\alpha$  representa el porcentaje de veces que se rechaza la  $H_0$  cuando es verdadera.
3. Error de tipo II. No se rechaza  $H_0$  cuando en realidad es falsa. A la probabilidad de que esto ocurra se le denomina  $\beta$ .

4. Se rechaza la hipótesis nula y es falsa. Se trata de una decisión correcta. La probabilidad de que esto ocurra se denomina poder o potencia de la prueba y viene dado por  $1 - \beta$

La determinación de  $\beta$  es compleja. La relación entre  $\alpha$  y  $\beta$  es de tipo inverso, de forma que al disminuir el error de tipo I aumenta  $\beta$  o la probabilidad de que ocurra un error tipo II, para una determinada muestra. El poder de la prueba  $1 - \beta$  está ligada al valor del parámetro probado en la población objeto de estudio.

En la práctica,  $\beta$  se ignora o bien se determina después de estar seleccionada la muestra.

**El interés del investigador se centra en no cometer un error de tipo I.**

Como conclusión, podemos decir que en la práctica lo que se hace es fijar un nivel de significación ( $\alpha$ ) del 1% ó 5%. Y se rechaza la  $H_0$  si hay menos del 1% ó 5% de posibilidades de que las diferencias obtenidas sean debidas al azar y se dice que la diferencia es significativa.

### 7.3.3 ELECCIÓN DEL TEST

La Estadística ha tenido un enorme desarrollo. En la actualidad, para cualquier diseño de investigación, nos encontramos con diversas pruebas estadísticas válidas para decidir acerca de  $H_0$ .

Por ello, se hace necesario elegir criterios racionales para determinar qué prueba estadística es la más adecuada para analizar los datos de una Investigación Comercial. Se pueden utilizar diversos criterios. Los más usuales son:

la potencia, el procedimiento de obtención de los datos, el universo (N) de donde se obtuvo la muestra (n), las hipótesis que deseamos probar o la escala de medición utilizada.

### VALIDEZ Y POTENCIA

La potencia de un análisis estadístico es, en parte, una función de la prueba estadística que se usa en un análisis. Una prueba estadística es válida si la probabilidad de rechazar  $H_0$ , cuando ésta es verdadera, es igual al valor elegido  $\alpha$ . Se dice que una prueba es potente si tiene gran probabilidad de rechazar  $H_0$  cuando ésta es falsa.

## MODELO ESTADÍSTICO

En el momento en el que identificamos la naturaleza del universo o población (N) y la técnica de muestreo a aplicar, hemos establecido un modelo estadístico. Cada prueba estadística se asocia a un modelo y a un requisito de medida.

Todas las decisiones tomadas para el uso de cualquier prueba estadística deben llevar consigo la siguiente condición.

*“Si el modelo utilizado es correcto y los requisitos de medida fueron cumplidos, entonces....”.*

Cuanto más pobres o débiles sean las suposiciones que definen el modelo, necesitaremos simplificar más la decisión alcanzada por la prueba estadística asociada al modelo y, por consiguiente, más generales serán las conclusiones.

Las pruebas más potentes son aquéllas que tienen las suposiciones más fuertes o extensas.

Las pruebas paramétricas (como la t, la Z o la F), tienen fuertes suposiciones subyacentes a su uso. Si estas suposiciones son válidas, las pruebas basadas en las mismas son las que tienen mayor probabilidad de rechazar  $H_0$  cuando es falsa.

Hemos de resaltar que los requerimientos de los datos de investigación deben ser adecuados para la prueba.

## EJEMPLO:

Las condiciones que debe satisfacer la prueba “t” para ser la más potente y aceptar así las conclusiones obtenidas con ella con el adecuado nivel de confianza, son:

1. Las observaciones deben ser independientes
2. Las observaciones deben derivar de poblaciones normalmente distribuidas
3. En el análisis de dos grupos, las poblaciones deben tener la misma varianza
4. Las variables serán medidas al menos en una escala de intervalo

Todas estas condiciones son elementos del modelo estadístico asociado con la distribución normal. De ordinario, estas suposiciones no son probadas en el curso del análisis estadístico, sino que son presunciones aceptadas y su certeza o falsedad determina la exactitud y significación de la probabilidad establecida mediante la prueba paramétrica.

# Análisis de la Investigación Cuantitativa

---

Las pruebas paramétricas prueban hipótesis acerca de parámetros específicos. Se presupone que las hipótesis acerca de tales parámetros son idénticos a las hipótesis de investigación.

Cuando existen razones para creer que las condiciones se encuentran en los datos objeto de análisis, es posible elegir una prueba estadística paramétrica.

## ¿Qué ocurre si estas condiciones no se encuentran?

Cuando no se encuentran las suposiciones que constituyen el modelo estadístico de una prueba entonces esta no puede ser válida. Esto significa que el estadístico de prueba puede caer en la región de rechazo con una probabilidad mayor que alfa ( $\alpha$ ).

## EFICACIA

Cuando comparamos dos muestras de tamaños diferentes, la potencia de la prueba aumenta conforme aumenta el tamaño de la muestra estadística. Por consiguiente, podremos utilizar una prueba menos potente con un tamaño de muestra más grande.

El concepto de potencia-eficacia se relaciona con el tamaño de la muestra que es necesario para lograr que la prueba B sea tan potente como la A cuando el nivel de significación y el tamaño de la muestra de la prueba A se mantienen constantes.

Si la prueba A es la prueba conocida más potente de su tipo y en prueba es una prueba para el mismo diseño de investigación, entonces

Potencia eficacia de prueba B

$$P_B = \frac{100n_a}{n_b}$$

*Eficacia relativa asintótica de un estadístico*

Es un modo de determinar el tamaño de muestra necesario para que la prueba B tenga la misma potencia que la prueba A.

Eficacia relativa

$$P_B = 100 \lim_{n_a \rightarrow \infty} \frac{n_a}{n_b}$$

## ESCALA DE MEDIDA

Las escalas de medidas utilizadas en la Investigación Comercial son: nominal, ordinal, intervalo y de razón. Resmimos en la siguiente tabla sus características y relaciones:

Investigación Comercial

## Análisis de la Investigación Cuantitativa

---

Escala	Relaciones que la definen
Nominal	Equivalencia
Ordinal	Equivalencia Mayor que
De intervalo	Equivalencia Mayor que Razón conocida entre cualesquiera intervalos
De razón	Equivalencia Mayor que Razón conocida entre cualesquiera intervalos Razón conocida entre cualesquiera de dos valores de la escala

Como conclusión, podemos decir que hemos de elegir el test más potente. Como regla general recordemos que:

- Los test unidireccionales son más potentes que los bidireccionales
- Los test paramétricos son más potentes que los no paramétricos
- Las pruebas que usan datos métricos son más potentes que las que no los utilizan

### 7.3.4 INTERPRETACIÓN DE LA PRUEBA

La prueba o test elegido tiene una variable asociada a un estadístico cuya distribución es conocida y cuyo valor se obtiene mediante la correspondiente fórmula con los datos de la muestra.

Para el nivel de significación  $\alpha$  conocemos el valor teórico (valor de tablas) del test ,que será el que compararemos con el obtenido de aplicar la fórmula a los datos obtenidos de la muestra.

Esta comparación es la que nos sirve como norma de decisión. En general

**“Si el valor calculado es mayor que el teórico o de tablas para un determinado nivel de significación  $\alpha$ , se rechaza la  $H_0$ . En caso contrario no se rechaza la  $H_0$ ”.**

No rechazar la  $H_0$  supone que los datos obtenidos en una investigación concreta son compatibles con la hipótesis propuesta. La interpretación ha de realizarse con prudencia

y criterio, recordando la frase de Henry Clay que dice: “las estadísticas no son sustitutos del criterio”.

En algunas pruebas, es preciso conocer los grados de libertad o el número de observaciones diferentes que obtenemos de una variable si descontamos el o los estadísticos calculados. En el caso de una tabla de  $r$  filas y  $c$  columnas, los grados de libertad son  $(r - 1)$  para las filas,  $(c - 1)$  para las columnas y  $(r - 1)(c - 1)$  para la tabla.

### RESUMEN

La metodología para realizar un contraste de hipótesis consiste en la realización de los siguientes pasos:

1. Definir  $H_0$  y  $H_1$
2. Definir un parámetro de medida que relacione los parámetros muestrales y poblacionales y la función de probabilidad
3. Establecer un criterio para juzgar si el valor calculado del parámetro es compatible con  $H_0$
4. Realizar el muestreo
5. Obtener conclusiones de conformidad con lo expuesto

### 7.4 TIPOS DE TEST DE HIPÓTESIS

Como hemos visto ya, los test de hipótesis se utilizan para designar el procedimiento que contrastará la validez de una hipótesis. Son pruebas estadísticas que se utilizan para determinar si los resultados obtenidos con una muestra o dos elegidas al azar difieren marcadamente de aquellos que habría que esperar con la hipótesis planteada y la variación debida al muestreo.

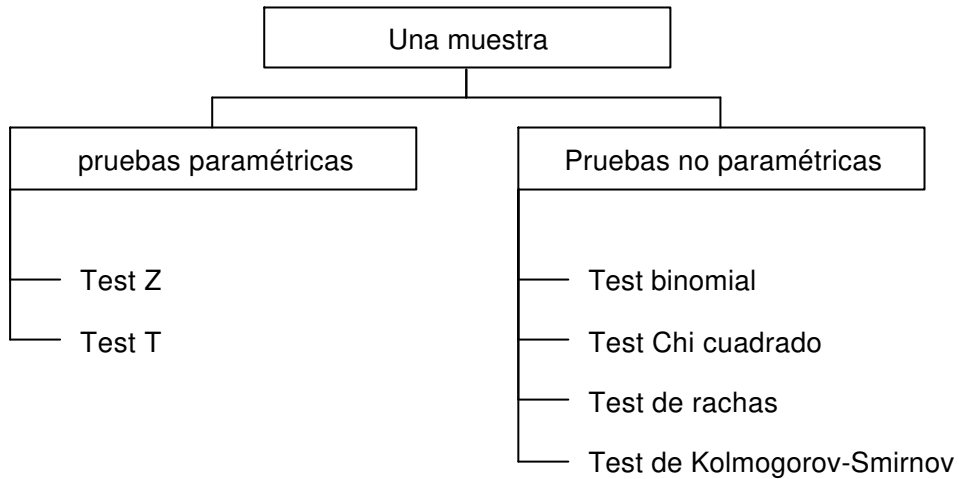
Existen diferentes tipos de test. La forma más habitual de clasificarlos es la siguiente:

- Dependiendo del tipo de variables a contrastar, se distingue entre pruebas paramétricas y pruebas no paramétricas. Las primeras se aplican a variables medidas con escalas de ratio y de intervalo. Las segundas, a variables ordinales y nominales.
- Según realicemos el contraste sobre una muestra o sobre dos muestras, que a su vez se pueden diferenciar en muestras independientes o relacionadas (pareadas), aplicaremos diferentes tipos de test de hipótesis.

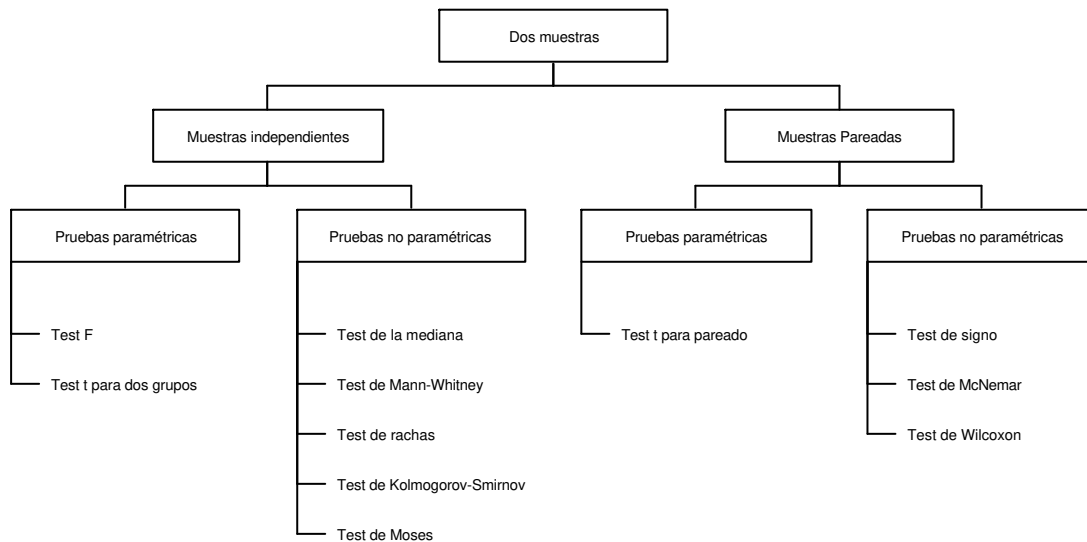
Los más usuales los resumimos en el siguiente esquema,  
Investigación Comercial



## Contraste de hipotesis



## CONTRASTE DE HIPÓTESIS PARA DOS MUESTRAS



## 7.5 BREVE DESCRIPCIÓN DE LOS TEST

Vamos a reseñar brevemente en qué consiste cada uno de los procedimientos resumidos en el esquema anterior.

### 7.5.1 CONTRASTES PARA UNA MUESTRA

#### 1. Pruebas paramétricas

Test Z. Está basado en la distribución normal. Se utiliza para contrastar estadísticos de una distribución con respecto a valores de parámetros del universo. El paquete estadístico calcula el valor de Z y luego calcula la probabilidad asociada al mismo. Se compara la probabilidad con el nivel de significación seleccionado y no se rechaza la  $H_0$  en el caso de que la probabilidad sea mayor que dicho nivel. Se acepta la hipótesis alternativa cuando la probabilidad es menor que el nivel de confianza.

Test t. Se basa en la t de Student. Parte del conocimiento de la media de la muestra y de la desviación típica de la media y de la media del universo. La interpretación de los resultados es idéntica al test Z.

#### 2. Pruebas no paramétricas

Test binomial. Se utiliza para variables dicotómicas, con la finalidad de contrastar si una variable procede de una población binomial con una probabilidad determinada de que se produzca un suceso.

Test Chi cuadrado. Se suele utilizar en variables medidas en escala nominal. Se basa en la distribución Chi cuadrado.

Test de rachas (test runs). Las rachas se definen como el número de veces que se produce el cambio de un valor a otro dentro de la distribución de la variable dicotómica. Su finalidad es comprobar si la variable se distribuye aleatoriamente o no. Se basa en el orden de ocurrencia de los dos valores de una variable de tipo dicotómico.

Test de Kolmogorov-Smirnov. Se utiliza para contrastar la hipótesis de que la muestra estudiada se distribuye como alguna de las principales distribuciones, normal, uniforme o de Poisson.

### 7.5.2 CONTRASTES PARA DOS MUESTRAS INDEPENDIENTES

#### 1. Pruebas paramétricas

Test t para dos grupos independientes. Se aplica a variables medidas en escalas de intervalo o de razón. Se utiliza para contrastar si dos muestras independientes proceden de poblaciones con la misma media.

Test F (F de Barlett Box). Se utiliza para contrastar si las muestras proceden de poblaciones con idénticas varianzas. Realiza el cociente entre las medias cuadráticas de las dos muestras, teniendo en cuenta el número de grados de libertad del numerador y del denominador, asignando así la probabilidad.

#### 2. Pruebas no paramétricas

Test de la mediana. Se usa para contrastar si dos o más muestras independientes pertenecen a poblaciones con la misma mediana.

Test de rachas de Wald-Wolfowitz. Contrasta la hipótesis de que dos muestras proceden de la misma población. Se aplica normalmente a variables medidas en escala ordinal. Cuando existe un número elevado de rachas, las muestras pertenecen a la misma población.

Test de Kolmorov-Smirnov (para dos muestras). Sirve para contrastar si las dos muestras provienen de la misma distribución. El test utiliza las diferencias existentes entre mediana, dispersión, asimetría e incluso más estadísticos de ambas muestras, comparando las distribuciones acumuladas y las diferencias con patrones conocidos.

Test de Mann-Whitney. Se aplica a variables medidas en escala ordinal. Es una prueba parecida a la t.

Test de Moses. Se aplica a variables medidas en escala ordinal. Consiste en contrastar la hipótesis de que la variable estudiada afecta a unos elementos en una dirección y a otros en la dirección opuesta.

### 7.5.3 CONTRASTES PARA DOS MUESTRAS RELACIONADAS

#### 1. Pruebas paramétricas

Test t para pareado. Se utiliza para comprobar si dos muestras provienen de poblaciones con igual media.

#### 2. Pruebas no paramétricas

Test de McNemar. Se aplica sobre dos variables dicotómicas relacionadas. El contraste lo realiza a través de la distribución Chi cuadrado, analizando la probabilidad de ocurrencia de las situaciones posibles (0,1) y (1, 0).

Test del signo. Se utiliza para contrastar la hipótesis de que las dos variables tienen idéntica distribución. Si las dos muestras tienen la misma distribución, la mitad de las diferencias deberían ser positivas y la otra mitad negativas; por consiguiente el método que sigue se basa en la dirección (signo) de las diferencias entre las dos variables.

Test de Wilcoxon. Se utiliza para contrastar la hipótesis de que las dos variables tienen idéntica distribución. Tiene en cuenta la magnitud de las diferencias dentro de los pares y pondera con un valor mayor a los pares que presentan mayores diferencias.

## 8. ANÁLISIS BIVARIABLE

El análisis bivariable trata de analizar la relación entre dos variables. Este análisis permite comprobar si existe asociación entre esas variables así como medir la fuerza de esa asociación. Esta asociación no implica necesariamente causalidad (no es preciso que se cumpla que la variable causa preceda a la variable efecto).

Como las variables a estudiar pueden ser de tipo cuantitativo y cualitativo, las situaciones y principales técnicas de estudio se recogen en la tabla siguiente

TIPO DE VARIABLE	CUANTITATIVA	CUALITATIVA
CUANTITATIVA	Coeficiente de correlación (Pearson) Regresión Coeficiente alfa ( $\alpha$ ) de Cronbach	Análisis de varianza (Test F) Test “t” de medias
CUALITATIVA		Tabla de contingencia (Chi cuadrado) Correlación de rangos (Spearman) Coeficiente de Cramer

### 8.1 RELACIÓN ENTRE DOS VARIABLES CUALITATIVAS

#### 8.1.1 TABLAS DE CONTINGENCIA

##### CONCEPTO

La forma más frecuente de presentación del análisis es la tabulación cruzada, en ella se presentan las dos variables cualitativas con un número no excesivamente grande de categorías, en la praxis 4 ó 5 como mucho.

Una tabla de contingencia (Crosstab) es el resultado de clasificar los elementos de la muestra con arreglo a dos variables cualitativas (nominales), cada una de ellas diversificadas en modalidades o categorías mutuamente excluyentes.

Con la tabla de contingencia se puede verificar si existe relación de dependencia entre las dos variables. En el punto siguiente indicamos un ejemplo de una tabla de contingencia obtenida con el paquete estadístico SPSS para Windows

## Análisis de la Investigación Cuantitativa

### 8.1.1.1 TABLA DE CONTINGENCIA : SEXO Y CONOCIMIENTO DE INFORMÁTICA

Las variables que se cruzan son sexo de los encuestados y conocimiento de informática de los mismos. La tabla resultante del tratamiento estadístico es la siguiente:

**Tabla:Sexo by conoci.**

Count	Ninguno	Poco	Bueno	Elevado	Experto	
Exp Val						
Row Pct						Row
Col Pct						Total
Tot Pct	1	2	3	4	5	
Mujer	9	9	12	3	3	36
	7'1	6'2	6'7	12'0	4'0	44'4%
1	25'0%	25'0%	33'3%	8'3%	8'3%	
	56'3%	64'3%	80'0%	11'1%	33'3%	
	11'1%	11'1%	14'8%	3'7%	3'7%	
Hombre	7	5	3	24	6	45
	8'9	7'8	8'3	15'0	5'0	55'6%
2	15'6%	11'1%	6'7%	53'3%	13'3%	
	43'8%	35'7%	20'0%	88'9%	66'7%	
	8'6%	6'2%	3'7%	29'6%	7'4%	
Column	16	14	15	27	9	81
Total	19'8%	17'3%	18'5%	33'3%	11'1%	100'0%

Chi-square	Value	DF	Significance
Pearson	23'41527	4	'00010
Likelihood Ratio	25'80228	4	'00003
Mantel-Haenszel test for linear association	8'74396	1	'00311

Minimum Expected Frequency 4'000

Cells with Expected Frequency < 5 1 OF 10 (10'0%)

Number of Missing Observations: 21

### 8.1.1.2 SIGNIFICADO DE LOS ELEMENTOS QUE COMPONEN LA TABLA

Las dos variables cruzadas en esta tabla, sexo (A) y conocimiento de la informática (B), se dividen en las siguientes categorías:

Variable A, 2 categorías: 1 (Mujer) y 2 (Hombre).

Variable B, 5 categorías: 1 (Ninguno), 2 (Poco), 3 (Bueno), 4 (Elevado), y 5 (Experto).

Las cifras enmarcadas indican el número de observaciones totales válidas. En este caso, hay 81 casos, lo que representa el 100%.

En la columna titulada *Row Total* (Total fila) aparecen el número de observaciones de cada una de las categorías de la variable A, tanto en valores absolutos (frecuencias absolutas) como en porcentajes sobre el total casos (frecuencias relativas o probabilidad). Así, en el ejemplo anterior, el número de mujeres que han contestado la encuesta es de 36, que representa el 44,4% del total de casos válidos. El número de hombres es de 45 y su porcentaje el 55,6%.

Evidentemente, la suma de las frecuencias absolutas de cada categoría de esta variable coincide con el total de casos ( $36 + 45 = 81$ ), y las frecuencias relativas suman 100.

En la fila titulada *Column Total* (Totales de la columna) tenemos el total de casos que pertenecen a cada una de las categorías de la variable B. Así, en el ejemplo, el total de la primera columna nos indica que 16 personas (hombres+mujeres) no tienen ningún conocimiento de la informática (frecuencia absoluta). Esto representa el 19,8% de las observaciones válidas ( $16 : 81$ ) % (frecuencia relativa o probabilidad).

Como en el total de la fila, la suma de las frecuencias absolutas de cada categoría de esta variable coincide con el total de casos ( $16 + 14 + 15 + 27 + 9 = 81$ ), y las frecuencias relativas suman 100.

Dentro de cada casilla o celda obtenemos datos correspondientes a la intersección (cruce) entre las categorías de las variables ( $A_i$  y  $B_j$ ). La información contenida en cada casilla queda reflejada en el margen superior izquierdo de la tabla (no tiene por qué ser siempre la misma). En este ejemplo, las cifras de cada celda corresponden, correlativamente, a:

*Count (Frecuencia)* es el número de casos que han contestado a la vez, a la categoría  $i$  de la variable A y la categoría  $j$  de la variable B. En el ejemplo 9 es el número de mujeres que han contestado “Ningún conocimiento de informática”.

La suma de las frecuencias de A en una modalidad y que cumplen las diferentes categorías de B, nos da la frecuencia de A en esa modalidad. En el ejemplo, para la categoría mujer, sería:  $9 + 9 + 12 + 3 + 3 = 36$

*Exp Val (Valor esperado.)* Corresponde al número de casos que deberían aparecer en la casilla, si las dos variables fueran independientes entre sí. Luego se verá cómo se calcula y su utilidad.

*Row Pct (Porcentaje fila).* Porcentaje de casos de un cruce sobre el total de casos de la fila (observaciones de la categoría  $i$  de la variable A). Las 9 observaciones del cruce Mujer y Ningún conocimiento, son un 25% del total de Mujeres. O dicho de otra forma, del total de mujeres, un 25% tiene “Ningún conocimiento de informática”. Por tanto, esta frecuencia relativa es una probabilidad condicionada: probabilidad de  $B_j$  condicionada a  $A_i$ ,  $P(B_j / A_i)$ .

*Col Pct (Porcentaje columna).* Porcentaje de casos de un cruce sobre el total de casos de la columna (observaciones de la categoría  $j$  de la variable B). En el ejemplo, las 9 observaciones del cruce Mujer y Ningún conocimiento, son un 56,3% del total de personas que tienen “Ningún conocimiento de informática”. O bien, un 56,3% de los que tienen “Ningún conocimiento de informática” son mujeres. Esta frecuencia relativa también es una probabilidad condicionada. En este caso, la probabilidad de  $A_i$  condicionada a  $B_j$ ,  $P(A_i / B_j)$ .

*Tot Pct (Porcentaje total).* Porcentaje de casos de un cruce sobre el total de casos. Por tanto, equivale a la frecuencia relativa o probabilidad de la intersección de  $A_i$  con  $B_j$ ,  $P(A_i \cap B_j)$ . En el ejemplo, las 9 observaciones del cruce Mujer y Ningún conocimiento, son un 11,1% del total de casos válidos ( $9 : 81$ )%.

### 8.1.1.3 ANÁLISIS DE INDEPENDENCIA ENTRE LAS DOS VARIABLES. ESTADÍSTICO $\chi^2$ (CHI CUADRADO)

El análisis de independencia entre las dos variables se hace con el estadístico  $\chi^2$  (Chi cuadrado) que aparece después de la tabla. De las tres líneas que proporciona el ordenador hay que fijarse únicamente en la primera, la  $\chi^2$  de Pearson, que es la más genérica.



Para calcular la  $\chi^2$  primero hay que determinar el Valor esperado de una celda. Tal como hemos indicado anteriormente, el Valor esperado es el número de casos que deberían aparecer en una casilla si las dos variables fueran independientes.

El Valor esperado es igual al número de observaciones del total de la fila a la que pertenece la casilla, multiplicado por el número de observaciones del total de la columna y dividido por el número de observaciones válidas. En nuestro ejemplo, para la primera celda, el Valor esperado es igual a 7,1 obtenido con la siguiente operación matemática:  $(36 \times 16) / 81 = 7,1$ .

También podemos obtener este valor si aplicamos el porcentaje del total de la fila al número de observaciones total de la columna ( $44,4 \% \times 16 = 7,1$ ) o el porcentaje del total de la columna al número de observaciones total de la fila ( $19,8 \% \times 36 = 7,1$ ). Con estas nuevas fórmulas se puede interpretar el Valor esperado de una manera más simple. Si las dos variables son independientes el número de mujeres con “Ningún conocimiento de informática” debería ser igual al porcentaje de personas que tienen “Ningún conocimiento de informática” (19,8%) multiplicado por el número de mujeres de la muestra (36) o al porcentaje de mujeres (44,4%) aplicado al total de personas que tienen “Ningún conocimiento de informática” (16).

De forma genérica, si las dos variables son independientes, los porcentajes del total de la fila (*Row Total*) deberían ser iguales en cada columna en los porcentajes sobre el total de la columna (*Col Pct*), ya que la variable representada en las filas no influye en la otra variable. De igual forma, los porcentajes del total columna (*Column Total*) deberían ser los mismo en cada fila que los porcentajes sobre el total de la fila (*Row Pct*). En nuestro ejemplo, el 44,4% de mujeres y el 55,6% de hombres debería repetirse para todas las categorías de la variable “Conocimiento de informática”. O bien, los porcentajes de “Conocimiento de informática” correspondientes a las personas (Total columna), deben aparecer tanto para hombres como para mujeres.

La columna del ejemplo en la situación teórica de independencia entre las dos variables quedaría como sigue:

Mujeres 1	Count	7'1
	Exp Val	7'1
	Row Pct	19'8
	Col Pct	44'4
	Tot Pct	8'8
Hombres 2	Count	8'9
	Exp Val	8'9
	Row Pct	19'8
	Col Pct	55'6
	Tot Pct	11'0

La  $\chi^2$  se calcula de la siguiente manera:

$$\sum_{i=1}^r \sum_{j=1}^c \frac{[n_{ij} - E(n_{ij})]^2}{E(n_{ij})}$$

donde los sumatorios recogen los valores de todas las celdas, siendo  $n_{ij}$  la frecuencia observada (*Count*) y  $E(n_{ij})$  el valor esperado (*Exp Val*) de cada celda.

La  $\chi^2$  no puede utilizarse si cualquiera de las celdas tiene un valor esperado menor que 1 o si más de un 20% de las celdas tienen un valor esperado menor que 5. En nuestro ejemplo, podemos aplicarla, ya que el valor mínimo esperado es 4 y sólo en un 10% de las celdas (1 de 10) hay un valor esperado menor que 5.

Minimum Expected Frequency - 4.000

Cells with Expected Frequency < 5 - 1 OF 10 ( 10.0%)

La última línea del resultado de una tabla de contingencia nos indica el número de observaciones “*missing*”, es decir, aquéllas para las que el encuestado no ha contestado a alguna de las variables cruzadas o a ambas. En el ejemplo, en 21 casos no se dispone de información referida a las variables cruzadas.

### 8.1.2 COEFICIENTE V DE CRAMER

Se trata de un coeficiente que toma el valor 1 cuando hay una asociación perfecta entre los atributos considerados, con independencia del número de filas y columnas de la tabla de contingencia.

Su fórmula matemática es:

$$V = \sqrt{\frac{\chi^2}{m n}}$$

Donde m es el mínimo de los grados libertad de las filas y columnas. Por tanto

$$m = \min. (h - 1, k - 1)$$

n es el tamaño de la observación (muestra) y  $\chi^2$  es el estadístico Chi cuadrado.

### 8.1.3 CORRELACIÓN DE RANGOS DE SPEARMAN

Este coeficiente determina la correlación entre dos variables cualitativas medidas en escala ordinal. Este coeficiente varía entre  $-1$  y  $+1$ ; un valor positivo nos indica una correlación en el mismo sentido (directa), mientras que los valores negativos ponen de manifiesto una relación inversa.

Para su determinación se aplica la fórmula de Spearman:

$$r = 1 - \frac{6 \sum_{i=1}^n D_i^2}{N(N^2 - 1)}$$

Donde N es el número de pares de observaciones y D la diferencia entre los grados de los valores correspondientes (x e y).

## 8.2 MÉTODOS DE MEDICIÓN EN EL ANÁLISIS ENTRE DOS VARIABLES CUANTITATIVAS

La asociación entre dos variables no implica necesariamente una relación causal, sino simplemente conociendo la asociación entre dos variables se podrá anticipar la variación de una variable conociendo el comportamiento de la otra.

A continuación vamos a reseñar las medidas de asociación más habituales en la Investigación Comercial.

### 8.2.1 CORRELACIÓN

Es la asociación entre las variaciones de los valores de dos variables. La asociación puede ser directa (mismo sentido) o inversa (sentidos opuestos).

La asociación es la relación entre el comportamiento de las dos variables. Esta asociación puede deberse a la “casualidad” o bien ser causal.

Se dice que la relación es causal cuando un cambio en una de las variables (independiente o explicativa) produce un cambio en la otra (dependiente).

Para que exista causalidad se deben cumplir al menos las siguientes condiciones<sup>7</sup>:

- Variación Concomitante. Supone que ambas variables varían a la vez.
- Temporalidad en la variación. La causa debe ir por delante del efecto.
- Control sobre otros factores distorsionadores. Se deben encontrar las posibles correlaciones espurias (correlaciones debidas a la asociación entre dos variables, bien de forma accidental o casual, bien por efecto de una tercera variable, sin que se produzca ninguna relación causal entre ellas).

La correlación entre dos variables cuantitativas se estudia habitualmente mediante el análisis de correlación lineal, que se mide mediante el coeficiente de Pearson.

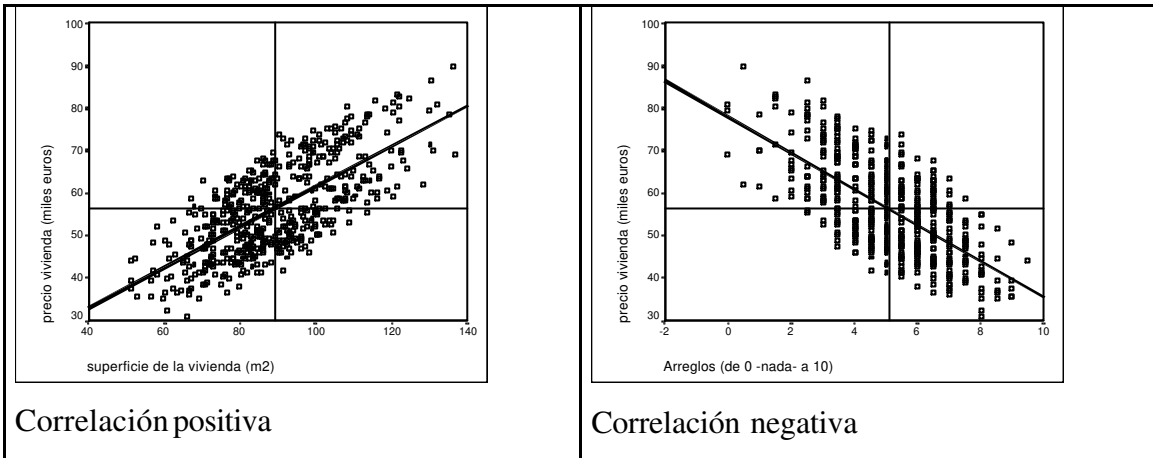
Este coeficiente puede tomar valores entre +1 y -1.

El valor 0 indica la ausencia de correlación. Si el coeficiente es positivo indica que la correlación es directa (en el mismo sentido), en caso contrario se dice que es indirecta.

---

<sup>7</sup> Dillon y otros Marketing Research in a Marketing environment Irwin 1994  
Investigación Comercial

## Gráfica



El coeficiente entre las dos variables x e y se calcula dividiendo la covarianza de ambas variables entre el producto de las desviaciones típicas de ambas variables.

$$r = \frac{s(x, y)}{s(x)s(y)}$$

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$$

## Análisis de la Investigación Cuantitativa

---

### 8.2.1.1 EJEMPLO

Supongamos que queremos conocer la correlación existente entre la inversión en promoción de ventas (x) y la cifra de ventas (y), en miles de euros

Los datos los resumimos en la siguiente tabla

X	y	xy	x <sup>2</sup>	y <sup>2</sup>
2	50	100	4	2500
2'5	70	175	6'25	4900
3	75	225	9	5625
3	80	240	9	6400
4	85	340	16	7225
3'5	90	315	12'25	8100
$\Sigma x = 18$	$\Sigma y = 450$	$\Sigma xy = 1395$	$\Sigma x^2 = 56'5$	$\Sigma y^2 = 34750$

El coeficiente de correlación lineal (s e u o) será

$$r = \frac{6 \bullet 1395 - 18 \bullet 450}{\sqrt{(6 \bullet 56'6 - 18^2)(6 \bullet 34750 - 450^2)}} = 0'88$$

Por tanto, podemos suponer que existe muy buena relación entre ambas variables.

Para concluir si la r de Pearson es significativa al 5% ó al 1%, deberemos consultar con la correspondiente tabla.

# Análisis de la Investigación Cuantitativa

## 8.2.1.2 TABLA COEFICIENTE DE CORRELACIÓN R DE PEARSON

Prueba unilateral: nivel de significación				
	0,05	0,025	0,01	0,005
Prueba bilateral: nivel de significación .				
Grados de libertad	0,10	0,05	0,02	0,01
1	0,988	0,997	0,999	0,999
2	0,900	0,195	0,980	0,990
3	0,805	0,878	0,934	0,959
4	0,729	0,811	0,882	0,917
5	0,669	0,784	0,833	0,874
6	0,622	0,707	0,789	0,834
7	0,582	0,666	0,750	0,798
8	0,549	0,632	0,716	0,765
9	0,521	0,602	0,685	0,735
10	0,497	0,576	0,658	0,708
11	0,576	0,553	0,634	0,684
12	0,458	0,532	0,612	0,661
13	0,441	0,514	0,592	0,641
14	0,426	0,497	0,574	0,623
15	0,412	0,482	0,558	0,606
16	0,400	0,468	0,542	0,590
17	0,389	0,456	0,528	0,575
18	0,378	0,444	0,516	0,561
19	0,369	0,433	0,503	0,549
20	0,360	0,423	0,492	0,537
21	0,352	0,413	0,482	0,526
22	0,344	0,404	0,472	0,515
23	0,337	0,396	0,462	0,505
24	0,330	0,388	0,453	0,496
25	0,323	0,381	0,445	0,487
26	0,317	0,374	0,437	0,479
27	0,311	0,367	0,430	0,471
28	0,306	0,361	0,423	0,463
29	0,301	0,355	0,416	0,486
30	0,296	0,349	0,409	0,449
35	0,275	0,325	0,381	0,418
40	0,257	0,304	0,358	0,393
45	0,243	0,288	0,338	0,372
50	0,231	0,273	0,322	0,354
60	0,211	0,250	0,295	0,325
70	0,195	0,232	0,274	Q303
80	0,183	0,217	0,256	0,283
90	0,173	0,205	0,242	0,267
100	0,164	0,195	0,230	0,254

### 8.2.2 REGRESIÓN SIMPLE

En términos generales, podemos definir la regresión lineal simple como el estudio entre una variable a explicar con respecto a otra, que denominamos explicativa. Ambas variables deben ser cuantitativas.

El modelo de regresión lineal de primer orden viene expresado por la siguiente ecuación:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

donde:

$y$  = variable **dependiente** o variable a explicar

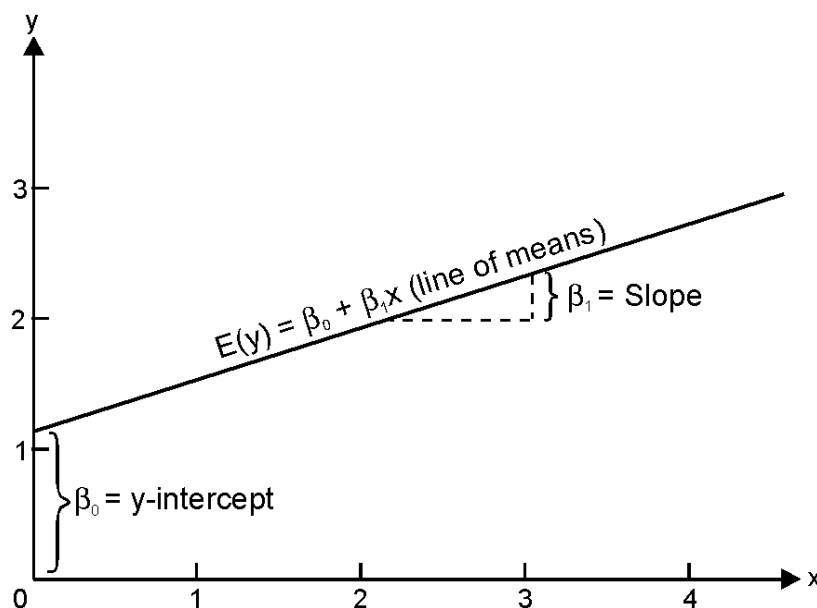
$x$  = variable **independiente** o variable explicativa

$\varepsilon$  (épsilon) = **error** o perturbación aleatoria

$\beta_0$  = **origen** de la recta: punto donde la recta corta el eje de ordenadas o eje de la  $y$

$\beta_1$  = **pendiente** de la recta o coeficiente de regresión: nos indica cuánto aumenta o disminuye la variable dependiente por cada incremento en 1 unidad de la variable independiente

La correspondiente representación gráfica es:





### 8.2.2.1 OBJETIVOS

Los objetivos que se pretenden con este tipo de análisis son varios:

Determinar la forma de la relación entre las dos variables, dependiente e independiente

Comprobar la hipótesis

Predecir los valores que tomará la variable dependiente

**Para dar satisfacción a estos objetivos nos interesa saber:**

Cómo se calculan los coeficientes de regresión,  $\beta_0$  y  $\beta_1$

Cómo se interpretan

Cómo se determina si son o no estadísticamente significativos

Cómo se comprueban las hipótesis del modelo

Estimación del modelo de regresión por mínimos cuadrados ordinarios

Con los datos de la muestra se pueden estimar los parámetros desconocidos del modelo del siguiente modo:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\text{SSE (Suma de Errores al Cuadrado)} = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

La recta de mínimos cuadrados ordinarios es, precisamente, aquella que minimiza la suma de los errores cuadrados.

Fórmulas para obtener los estimadores mínimos cuadrados

$$\text{Pendiente: } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \text{Origen: } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde

$$SS_{xy} = \sum_{i=1}^n x_i y_i - \frac{\left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n}$$

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

$n$  = tamaño de la muestra

Los estimadores mínimos cuadrados son los mejores que se pueden conseguir (insesgados, eficientes y consistentes) si se cumplen ciertas hipótesis sobre  $\varepsilon$  (las perturbaciones).

### 8.2.2.2 EJEMPLO REGRESIÓN LINEAL

En los últimos diez años, las ventas de un fabricante respecto de un determinado producto han sido las siguientes:

Año	1	2	3	4	5	6	7	8	9	10
Ventas	40	42	50	49	47	50	52	51	55	60

Determinar la recta de regresión (tendencia).

La recta vendrá dada por:

$y = \alpha + \beta x$  donde  $y$  son las ventas  $x$  los años

Las ecuaciones a resolver son :

$$\sum y_i = n\alpha + \beta \sum x_i$$

$$\sum x_i y_i = \alpha \sum x_i + \beta \sum x_i^2$$

## Análisis de la Investigación Cuantitativa

---

Para resolver el correspondiente sistema de ecuaciones realizamos la siguiente tabla

X	y	xy	x <sup>2</sup>
1	40	40	1
2	42	84	4
3	50	150	9
4	49	196	16
5	47	235	25
6	50	300	36
7	52	364	49
8	51	408	64
9	55	495	81
10	60	600	100
<b>55</b>	<b>496</b>	<b>2872</b>	<b>385</b>

El sistema de ecuaciones, a resolver queda como sigue:

$$496 = 10 \alpha + \beta 55$$

$$2872 = 55 \alpha + \beta 385$$

Luego

$$2872 = 55 \left( \frac{496 - 55\beta}{10} \right) + 385\beta$$

Una vez realizados los correspondientes cálculos obtenemos que  $\beta = 1,745$  y  $\alpha = 40$

La recta de regresión será:  $y = 40 + 1,745 x$

### 8.2.3 COEFICIENTE ALFA DE CRONBACH

El coeficiente alfa de Cronbach es un estimador de la consistencia interna de una escala de medida. La expresión matemática de este coeficiente es:

$$\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_s^2} \right)$$

Donde  $k$  es el número de ítem de la escala,  $s_i^2$  es la varianza del ítem “i” y  $s_s^2$  es la varianza de toda la escala.

El valor de alfa tiende a aumentar a medida que se incrementa el número de ítem de la escala. Este coeficiente varía entre 0 y 1, y puede tomar valores negativos cuando existen ítem correlacionados negativamente, en esta situación el coeficiente alfa de Cronbach no es el adecuado para medir la confiabilidad (grado en el que la medida de una variable está libre de error aleatorio y por consiguiente proporciona resultados consistentes).

La principal aplicación de este índice es medir la confiabilidad de una escala.

Un valor del coeficiente alfa de Cronbach inferior a 0.7 indica una baja consistencia interna.

Cuando  $\alpha < 0.7$  nos indica que la escala mide varios fenómenos y, por consiguiente, puede no ser apropiada para los objetivos de la investigación; en estas circunstancias de escalas multidimensionales suele ser mejor aplicar análisis factorial.

### 8.3 RELACIÓN ENTRE UNA VARIABLE CUANTITATIVA Y OTRA CUALITATIVA

#### 8.3.1 ANÁLISIS DE LA VARIANZA

El análisis de la varianza, en el análisis bivariable, es una técnica estadística que relaciona una variable cualitativa con otra cuantitativa.

Determina la existencia de diferencias significativas entre las medias de una variable dependiente.

Es de gran utilidad en la investigación experimental.

La variable dependiente debe estar medida en escala métrica, mientras que la variable independiente es cualitativa. El proceso es como sigue:

El análisis de la varianza (ANOVA) se utiliza cuando no conocemos previamente el comportamiento de la variable dependiente (cuantitativa) sin la influencia del factor principal controlado (variable cualitativa).

### 8.3.1.1 MÉTODO ANOVA TRADICIONAL UNIDIRECCIONAL

En este método, el test de significación que se utiliza es el F, que compara los efectos de los diferentes tratamientos recogidos por la varianza de la dispersión factorial con los efectos de los factores controlados representados por la varianza de la dispersión residual a través de una relación.

$$F = \frac{\text{Varianza.de.la.dispersión.factorial}}{\text{Varianza.de.la.dispersión.residual}}$$

El examen de las diferencias entre las medias implica la descomposición de la variación total observada en la variable dependiente. Luego será:

$$\text{Dispersión total} = \text{Dispersión factorial} + \text{Dispersión residual}$$

Y las varianzas de las dispersiones son:

#### 1- Dispersión total

Viene dada por la siguiente fórmula:

$$S_t^2 = \frac{\sum_{ij} (x_{ij} - m)^2}{n - 1}$$

donde n es el tamaño de la muestra.

$x_{ij}$  es el valor de cada uno de los datos individuales considerados

m es el valor de la media general

#### 2- Dispersión factorial

Su expresión matemática es:

$$S_f^2 = \frac{\sum_i (m_i - m)^2}{K - 1}$$

donde  $m_i$  es la media de los diferentes grupos estudiados,

m es la media general y

k es el número de grupos considerados

### 3- Dispersión Residual

Su formulación es:

$$S_r^2 = \frac{\sum_{ij} (x_{ij} - m_i)^2}{n - K}$$

donde k es el número de grupos considerados

$m_i$  es la media de los diferentes grupos estudiados

$x_{ij}$  es el valor de cada uno de los datos individuales considerados

n es el tamaño de la muestra

Cuando el valor de F es igual a la unidad o muy próximo a este valor, significa que las dos dispersiones apenas difieren y, como consecuencia, se puede afirmar que los diferentes tratamientos no han tenido eficacia.

Si el valor de F es muy superior a la unidad, se admite que el efecto de los tratamientos ha sido eficaz.

Como el número de observaciones que se realizan es limitado, no podemos conocer el valor exacto de la varianza. Por ello, el valor de F oscilará en torno a la unidad como consecuencia de las variaciones del muestreo.

Este efecto se mitiga con la utilización de las tablas de Snedecor. De esta manera, cuando el valor de F calculado es superior al valor crítico de F indicado en las tablas, significa que los tratamientos aplicados son eficaces.

Si el valor calculado de F es inferior al valor obtenido en las tablas, esta diferencia se debe a las variaciones de muestreo.

## Análisis de la Investigación Cuantitativa

---

**TABLA ESTADÍSTICA: DISTRIBUCIÓN DE LA F (nivel de confianza 95%)**

“n”	“m”				
	1	2	3	4	5
1	161´4	199´5	215´7	224´6	230´2
2	18´51	19	19´16	19´25	19´30
3	10´13	9´55	9´28	9´12	9´01
4	7´71	6´94	6´59	6´39	6´26
5	6´61	5´79	5´41	5´19	5´05
6	5´99	5´14	4´76	4,53	4´39
7	5´59	4´74	4´35	4´12	3´97
8	5´32	4´46	4´07	3´84	3´69
9	5´12	4´26	3´86	3´63	3´48
10	4´96	4´10	3´71	3´48	3´33
11	4´84	3´98	3´59	3´36	3´20
12	4´75	3´89	3´49	3´26	3´11
13	4´67	3´81	3´41	3´18	3´03
14	4´6	3´74	3´34	3´11	2´96
15	4´54	3´68	3´29	3´06	2´90

Siendo **m** los grados de libertad del numerador y **n** los grados de libertad del denominador.

### 8.3.1.1.1 EL PROCESO DEL MÉTODO ANOVA UNIDIRECCIONAL

Se determinan las siguientes dispersiones:

1.- Dispersión total (DT)

Mide la suma de las dispersiones.

2.- Dispersión factorial (DF)

Mide la dispersión entre los grupos creados por las diferentes alternativas del factor o factores estudiados.

Dependiendo del tipo de experimento, pueden existir varias dispersiones factoriales, correspondientes al factor principal y a los factores de bloque.

3.- Dispersión residual (DR)

Mide la dispersión dentro de los grupos creados por las diferentes alternativas del factor o factores estudiados.

$$DT = DF + DR \quad DR = DT - DF$$

4.- Se calcula el cuadrado medio total (CMT)

Se trata de la dispersión total dividida por el número de grados de libertad

$$CMT = DT / gl \quad \text{donde } gl \text{ son los grados de libertad.}$$

5.- Se calcula el cuadrado medio factorial (CMF)

Se trata de la dispersión factorial dividida por el número de grados de libertad.

$$CMF = DF / gl$$

Dependiendo del tipo de experimento pueden existir varias varianzas factoriales, que corresponden al factor principal y a los factores bloque.

6.- Se calcula el cuadrado medio residual (CMR)

Se trata de la dispersión residual dividida por el número de grados de libertad.

$$CMR = DR / gl$$

7.- Se realiza el test de la F.

Para cada factor estudiado se calcula:

7-1.- Se calcula el estadístico F.

$$F = CMF / CMR$$

Si el valor de F es menor que uno, es decir  $CMF < CMR$ , no existe un efecto significativo del factor estudiado sobre la variable dependiente, y por tanto no es necesario realizar la comparación de F con el correspondiente valor de las tablas.



## Análisis de la Investigación Cuantitativa

---

7-2.- Se determina el valor de F en las tablas estadísticas de la distribución de la F, en base a los grados de libertad del numerador y del denominador.

7-3.- Se comparan ambos valores.

La hipótesis nula  $H_0$  es: NO EXISTE EFECTO SIGNIFICATIVO DEL FACTOR ESTUDIADO.

Entonces:

Si  $F > F_t$  (tabla), se rechaza  $H_0$  y por tanto el factor estudiado tiene una influencia significativa sobre la variable dependiente.

Si  $F = F_t$  (tabla), no se rechaza  $H_0$

### EJEMPLO DE ANOVA UNIDIRECCIONAL

Vamos a desarrollar lo expuesto anteriormente mediante un caso práctico:

Un banco lleva a cabo un experimento comercial realizando tres promociones diferentes para el lanzamiento de un nuevo producto. Estas promociones consisten en:

P1: regalo de una bicicleta, P2: regalo de un ordenador, P3: regalo de los electrodomésticos de la cocina.

Cada promoción se prueba en cinco sucursales diferentes durante un mes. Los resultados obtenidos, en cuanto a unidades de producto colocadas entre la clientela, se recogen en el cuadro siguiente:

	S1	S2	S3	S4	S5
P1	65	50	30	40	65
P2	30	25	15	20	35
P3	15	10	10	25	50

### SOLUCIÓN

Definiremos las siguientes características:

Factor principal: los diferentes tipos de promoción P1, P2, P3, luego  $K = 3$

## Análisis de la Investigación Cuantitativa

---

Unidades experimentales: 15 (5 sucursales x 3 tipos de promoción)

Variable dependiente: unidades vendidas

Hipótesis nula ( $H_0$ ) Los resultados obtenidos son independientes del tipo de promoción, es decir las medias de ventas de las tres promociones serán iguales.

$$H_0 : \mu_{p1} = \mu_{p2} = \mu_{p3}$$

Número total de mediciones:  $n=15$

Número de mediciones por cada tratamiento (promoción)  $n_j=5$

$x_{ij}$ = unidades físicas vendidas en cada sucursal

$m_j$ = media de unidades vendidas por tratamiento

$m$ = media total

Partiendo del cuadro de resultados, obtenemos los valores de  $m_j$  y  $m$ , los cuales son:

	S1	S2	S3	S4	S5	s	$m_j$
P1	65	50	30	40	65	250	50
P2	30	25	15	20	35	125	25
P3	15	10	10	25	50	110	22

por tanto,  $m = 32'333$

Una vez obtenidos estos datos, pasamos a realizar los cálculos de la técnica ANOVA

Dispersión total:

$$D T = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - m)^2$$

Sustituyendo por los correspondientes valores obtenemos:

$$\begin{aligned} DT &= (65 - 32'3)^2 + (50 - 32'3)^2 + (30 - 32'3)^2 + (40 - 32'3)^2 + (65 - 32'3)^2 + \\ &- 32'3)^2 + (25 - 32'3)^2 + (15 - 32'3)^2 + (20 - 32'3)^2 + (35 - 32'3)^2 + (15 - 32'3)^2 + (10 - 32'3)^2 + \\ &- 32'3)^2 + (10 - 32'3)^2 + (25 - 32'3)^2 + (50 - 32'3)^2 = 4.693'333 \end{aligned}$$

Dispersión factorial:

$$DF = \sum_{j=1}^k n_j (m_j - m)^2$$

Sustituyendo obtenemos

$$DF = 5(50 - 32'3)^2 + 5(25 - 32'3)^2 + 5(22 - 32'3)^2 = 2.363'333$$

Dispersión residual:

$$DR = DT - DF \quad \text{Luego } DR = 4.693'33 - 2.363'33 = 2.330$$

Cuadrado medio factorial (CMF):

$$CMF = \frac{DF}{gl} = \frac{DF}{k - 1}$$

Sustituyendo obtenemos  $CMF = 1.181'6667$

Cuadrado medio residual (CMR):

$$CMR = \frac{DR}{gl} = \frac{DR}{n - k}$$

Sustituyendo obtenemos  $CMR = 194'1667$

Test de la F :

$$F = \frac{CMF}{CMR}$$

Sustituyendo obtenemos  $F = 6'0858$

Si buscamos el valor de F en tablas para un nivel del 95% y  $gl = 2$  y 12, obtenemos que  $F = 3'89$

Como  $6'0858 > 3'89$ , existe un efecto significativo de los diferentes tratamientos estudiados para un nivel de confianza del 95%.

**La conclusión es que los diferentes tipos de promoción afectan significativamente a la demanda.**

**Es decir rechazamos la hipótesis nula o de independencia**

# Análisis de la Investigación Cuantitativa

## 8.3.2 TEST t DE MEDIAS

El test t de medias es una prueba estadística de contraste de hipótesis que nos permite contrastar si existen diferencias significativas entre dos valores medios.

El proceso es

- Se fija la hipótesis nula
- Se fija el nivel de significación  $\alpha$ , normalmente 5% (Error tipo I)
- Determinación del valor observado o medido
- Determinación del valor crítico
- Contraste de hipótesis

Este tipo de test es aplicable en el tratamiento de muestras pequeñas. ( $n < 30$  unidades).

### EJEMPLO: DIFERENCIA DE MEDIAS, DOS POBLACIONES INDEPENDIENTES

¿Existen sueldos medios diferentes según el sexo del empleado? o ¿Existe diferencia de medias en la experiencia previa según el sexo de los empleados?

Estadísticos del grupo

Sexo del empleado		N	Media	Desviación típ.	Error típ. de la media
Sueldo actual en miles de ptas.	Hombre	217	5426,88	2729,35	185,28
	Mujer	191	3361,17	1041,14	75,33
Experiencia previa (en años)	Hombre	216	1,7870	,6231	4,240E-02
	Mujer	191	1,7359	,5943	4,300E-02

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Intervalo de confianza para la diferencia	
									Inferior	Superior
Sueldo actual en miles de ptas.	Se han asumido varianzas iguales	80,26	,000	9,847	406	,000	2065,71	209,78	1653,33	2478,10
	No se han asumido varianzas iguales			10,328	284,482	,000	2065,71	200,01	1672,02	2459,40
Experiencia previa (en años)	Se han asumido varianzas iguales	1,763	,185	,844	405	,399	5,1E-02	6,057E-02	-7,E-02	,1702
	No se han asumido varianzas iguales			,846	402,675	,398	5,1E-02	6,039E-02	-7,E-02	,1698

## Análisis de la Investigación Cuantitativa

**TABLA ESTADÍSTICA: DISTRIBUCIÓN DE  $\chi^2$**

Valores de la función de distribución

g.l. = grados de libertad

$\chi^2_c$  tal que  $p(\chi^2 \leq \chi^2_c) = p$

Probabilidad p											
g.l.	0,995	0,990	0,975	0,950	0,900	0,500	0,100	0,050	0,025	0,010	0,005
1	7,88	6,63	5,02	3,84	2,71	0,45	0,01	0,00	0,00	0,00	0,00
2	10,60	9,21	7,38	5,99	4,61	1,39	0,21	0,10	0,05	0,02	0,01
3	12,84	11,34	9,35	7,81	6,25	2,37	0,58	0,35	0,22	0,12	0,07
4	14,86	13,28	11,14	9,49	7,78	3,36	1,06	0,71	0,48	0,30	0,21
5	16,75	15,09	12,83	11,17	9,24	4,25	1,61	1,15	0,83	0,55	0,41
6	18,55	16,81	14,45	12,69	10,64	5,35	2,20	1,64	1,24	0,87	0,68
7	20,28	18,48	16,01	14,07	12,02	6,35	2,83	2,17	1,69	1,24	0,99
8	21,96	20,09	17,53	15,51	13,36	7,34	3,49	2,73	2,18	1,65	1,34
9	23,59	21,67	19,02	16,92	14,68	8,34	4,17	3,33	2,70	2,09	1,73
10	25,19	23,21	20,48	18,31	15,99	9,34	4,87	3,94	3,25	2,56	2,16
11	26,76	24,73	21,92	19,68	17,28	10,34	5,58	4,57	3,82	3,05	2,60
12	28,30	26,22	23,34	21,03	18,55	11,34	6,30	5,23	4,40	3,57	3,07
13	29,82	27,69	24,74	22,36	19,81	12,34	7,04	5,89	5,01	4,11	3,57
14	31,32	29,14	26,12	23,68	21,06	13,34	7,79	6,57	5,63	4,66	4,07
15	32,80	30,58	27,49	25,00	22,31	14,34	8,55	7,26	6,26	5,23	4,60
16	34,27	32,00	28,85	26,30	23,54	15,34	9,31	7,96	6,91	5,81	5,14
17	35,72	33,41	30,19	27,59	24,77	16,34	10,09	8,67	7,56	6,41	5,70
18	37,16	34,81	31,53	28,87	25,99	17,34	10,86	9,39	8,23	7,01	6,26
19	38,58	36,29	32,85	30,14	27,20	18,34	11,65	10,12	8,91	7,63	6,84
20	40,00	37,67	34,27	31,41	28,41	19,34	12,44	10,85	9,59	8,26	7,43
21	41,40	38,93	35,48	32,67	29,62	20,34	13,24	11,59	10,28	8,90	8,03
22	42,80	40,29	36,78	33,92	30,81	21,34	14,04	12,34	10,98	9,54	8,64
23	44,18	41,64	38,08	35,17	32,01	22,34	14,85	13,09	11,69	10,20	9,26
24	45,56	42,98	39,36	36,42	33,20	23,34	15,66	13,85	12,40	10,86	9,89
25	46,93	44,31	40,65	37,65	34,38	24,34	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	35,56	25,34	17,29	15,38	13,84	12,20	11,16
27	49,64	46,96	43,29	40,11	36,74	26,34	18,11	16,15	14,57	12,83	11,81
28	50,99	48,28	44,46	41,34	37,92	27,34	18,94	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	39,09	28,34	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43,77	40,26	29,34	20,60	18,49	16,89	14,96	13,78
40	66,77	63,69	59,34	55,76	51,81	39,34	29,05	26,51	24,43	22,16	20,71
60	91,95	88,38	83,30	79,08	74,40	59,34	46,56	43,19	40,48	37,43	35,58

## 9.5 DISTRIBUCIÓN F DE FISHER

Cuando estudiamos la significación de la diferencia de las desviaciones típicas de dos muestras pequeñas independientes obtenidas de una misma población, podemos extraer consecuencias averiguando si las diferencias entre las desviaciones típicas pueden ser 0 por azar o no, en cuyo caso concluiremos que una muestra posee características propias que la hacen ser diferente de la otra.

La distribución F de Fisher es la razón entre dos estimaciones de la varianza de la población y sirve para poner a prueba el hecho de que dos estimaciones lo sean realmente de una varianza igual.

El proceso consiste en saber si una razón F, calculada entre dos muestras determinadas, es compatible con la hipótesis de que ambas muestras procedan de una única población y sólo difieran entre sí por azar.

Una vez calculada la F es necesario acudir a las correspondientes tablas, teniendo en cuenta los grados de libertad correspondientes a cada muestra ( $n - 1$ ) y el nivel de significación, normalmente el 5%.

Si la F calculada resulta mayor que la F de tablas a un nivel de significación del 5%, podemos afirmar que F no puede darse ni en el 5% de los casos; es decir, las dos muestras son significativamente diferentes, no pudiendo resultar iguales por azar ni en el 5% de los casos.

Para el cálculo se utiliza la siguiente fórmula:

$$F = \frac{s_1^2}{s_2^2} = \frac{\frac{\sum (x_i - m_1)^2}{n_1 - 1}}{\frac{\sum (x_i - m_2)^2}{n_2 - 1}}$$

En el numerador se pone siempre la varianza mayor.

### 10.3.3 PRUEBA DE DIFERENCIA DE MEDIAS INDEPENDIENTES.

A su vez se divide en función de si la varianza del universo es conocida o desconocida

#### 10.3.3.1 CON VARIANZA CONOCIDA

Si se conoce la varianza por algún estudio previo y la distribución es normal, se aplicará la prueba Z si la muestra es grande y el test t si la muestra es pequeña.

El estadístico correspondiente es

$$Z = \frac{m_a - m_b}{\sigma_{m_a - m_b}}$$
$$\text{Siendo } \sigma_{m_a - m_b} = \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$$

Donde  $m_a$  y  $m_b$  representa las medias de las muestras a y b,  $n_a$  y  $n_b$  representan el tamaño para las muestras a y b.  $\sigma_{m_a - m_b}$  es el error estándar y  $\sigma_a$  y  $\sigma_b$  las correspondientes desviaciones típicas de la población a y b.

#### EJEMPLO:

Realizado un estudio sobre 1.000 hogares, se comprueba que la compra media mensual de detergente líquido ha sido de 7.000 cajas, con una desviación típica de 49 cajas.

Se realiza una campaña de comunicación. Se hace un nuevo estudio de mercado, con una muestra de 2.000 hogares, obteniéndose los siguientes resultados: venta media mensual 7.500 cajas con una desviación típica de 15 cajas.

Podemos afirmar que la campaña ha sido eficaz a un nivel del 1%.

Solución:

$$m_1 = 7.500 \quad s_1 = 15 \quad n_1 = 2.000 \quad m_2 = 7.000 \quad s_2 = 49 \quad n_2 = 1.000$$

$$s_{m_1} - s_{m_2} = \sqrt{\frac{15^2}{2000} + \frac{49^2}{1000}} = 1'59$$

$$\frac{\text{Diferencia Observada}}{\text{Error Estándar}} = \frac{7500 - 7000}{1'59} = 314'5$$

Como  $314'5 > 2'58$ , podemos decir que la campaña ha sido eficaz.

### 10.3.3.2 CON VARIANZAS DESCONOCIDAS

En este caso, se estima la desviación típica poblacional a partir de las muestrales.

Cuando se trata de una distribución normal pero el tamaño de muestra es pequeño, el estadístico adecuado es el test t con  $(n_a + n_b - 1)$  grados de libertad. En el caso de que la muestra sea grande, se utiliza el test Z.

Si la distribución no es normal y la muestra es pequeña se deben aplicar test no paramétricos. Si la muestra es grande por el teorema central del límite se admite la normalidad.

El error estándar en este caso es:

$$S_{m_a - m_b} = \sqrt{s_{m_a}^2 + s_{m_b}^2} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

Si las varianzas, aún siendo desconocidas, podemos considerarlas iguales, utilizaremos la conjunta. La correspondiente desviación típica será

$$S = \sqrt{\frac{\sum_{i=1}^{n_a} (x_{ia} - m_a)^2 + \sum_{i=1}^{n_b} (x_{ib} - m_b)^2}{n_a + n_b - 2}}$$

Y el error estándar por tanto será

$$S_{m_a - m_b} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}} = \sqrt{S^2 \left( \frac{1}{n_a} + \frac{1}{n_b} \right)}$$

Para decidir si las varianzas son iguales o diferentes se aplica el test de la F con las siguientes hipótesis:

$H_0$ , las varianzas son iguales y  $H_1$ , las varianzas son diferentes.

El estadístico F se calcula con la siguiente fórmula:

$$F = \frac{s_a^2}{s_b^2}$$

El teórico se obtiene para los siguientes grados de libertad  $gl = (n_a - 1) (n_b - 1)$  mirando en las correspondientes tablas. Para calcular el test de las diferencias de medias, si se rechaza la  $H_0$  se utilizan las varianzas separadas; en caso contrario, se utiliza la varianza conjunta. Ver tabla adjunta Distribución F para un nivel de significación del 5%