UNIVERSIDAD NACIONAL DE LA PLATA

Facultad de Ciencias Agrarias y Forestales

CÁLCULO ESTADÍSTICO Y BIOMETRÍA

Curso 2024

GUIA DE APUNTES TEÓRICOS.

DOCENTES

Profesor Adjunto: Ing. Agr. Martín E. Delucis
Lic. Rodrigo Altamirano

Ayudantes Diplomados: Dra. Noelia Ferrando
Dr. Adrián Jauregui
Mgr. Laura Maly
Dra. Marina Pifano

Ayudante alumna: Paula Gertsmayer

INDICE:

Estadística descriptiva 2
- Medidas de posición 5
- Medidas de dispersión 6
- Tablas de frecuencia 9
- Representaciones gráficas 12
- Variables bidimensionales 14

Probabilidades 17
- Operaciones básicas con eventos 23

Distribuciones de probabilidad 28
- Variable aleatoria 28
- Esperanza y varianza 31
- Distribución binomial 38
- Distribución de Poisson 40
- Distribución Normal 44
- Estandarización 48

Estimación de parámetros 51
- Propiedades de un estimador 52
- Métodos de estimación 53
- Distribución de t 56
- Distribución de chi-cuadrado 60

Pruebas de hipótesis 63
- Pruebas no-paramétricas 72

Análisis de regresión y correlación 79
- Análisis de regresión lineal simple 80
- Relaciones no-lineales 89
- Regresión lineal múltiple 93
- Análisis de correlación lineal 93

Análisis de la variancia y diseño de experimentos 96
- Experimentación 96
- Principios básicos de la experimentación 101
- Supuestos análisis de la variancia 103
- Diseño completamente aleatorizado 106
- Diseño en bloques completos al azar 112
- Experimentos factoriales 118

ESTADÍSTICA DESCRIPTIVA

Podemos definir a la *Estadística* como la ciencia que estudia cuantitativamente los fenómenos aleatorios (cuyo resultado es azaroso). Su importancia no solo radica en el campo científico en áreas como la biología, economía, sociología, etc.; sino también en el ámbito profesional, habiendo llegado a convertirse en una herramienta importante para el tratamiento de la información.

En general podemos establecer que la *Estadística* se ocupa de los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar datos, siempre y cuando la variabilidad e incertidumbre sea una causa intrínseca de los mismos; así como de realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y en su caso a formular predicciones.

De aquí surgen dos tipos de estadística: *la Descriptiva y la Inferencial*.

La *Estadística Descriptiva* o *Análisis Exploratorio de Datos* tiene como objetivo el describir y resumir las características más sobresalientes de la información registrada proveniente de estudios observacionales o ensayos experimentales; para tal fin se vale de una serie de medidas (de posición o tendencia central, de dispersión, de forma, etc.), tablas y gráficos, que le permiten alcanzar dicho objetivo.

Por otro lado la *Estadística Inferencial* tiene como objetivo la búsqueda y la aplicación de procedimientos inferenciales necesarios para la toma de decisiones sobre una o más características de una población, en base a la información que está contenida en una muestra.

Antes de abordar los procedimientos de los cuales se vale la estadística descriptiva para analizar un conjunto de observaciones, se verán en forma sintética, algunos conceptos básicos que nos facilitaran la introducción a dichos procedimientos.

Conceptos Básicos

- Población

  Una población la podemos definir como un conjunto de elementos acotados en un tiempo y en un espacio determinados, ligados entre sí por una característica común observable y/o medible. Por ejemplo: las plantas de durazno de un monte frutal que en el mes de setiembre están en flor, etc.

- Tamaño poblacional

  Si la población es finita diremos que el tamaño poblacional es el número de elementos de la misma, y la denotaremos como **N**. En el ejemplo anterior de población sería el número de plantas de duraznos en flor.

- Muestra

  Generalmente es imposible y/o impracticable examinar toda la población, por lo que se examina una parte de ella (muestra). Por lo tanto podemos definir a una muestra como un subconjunto de elementos de la población, a partir de la cual describiremos las características más importantes de la misma.

- Unidad muestral

  La unidad muestral es el elemento o entidad sobre la cual relevaremos la información de interés. (la planta de durazno)

- Tamaño muestral

  Se entiende por tamaño muestral el número de elementos que conforman la muestra, y lo denotaremos como **n**.

- Variable

  Una variable es una característica, propiedad o atributo, con respecto a la cual los elementos de una población difieren. En general para denotar el conjunto de posibles valores que puede presentar una cierta variable se utiliza una letra mayúscula (**X**), y con la misma letra en minúscula (**x**) se hace referencia a un valor particular observable en un elemento de la población (dato). Por ejemplo: Si tenemos un grupo de cajas de petri cada una con 4 semillas y luego de un periodo de tiempo observamos las semillas germinadas, podremos denotar al número de semillas germinas en un caja de petri como **X** (variable observable), mientras **x** (dato) denotará el número de semillas germinadas en cada una de la cajas de petri.

- Variable aleatoria

  Vamos a definir a una variable aleatoria como un conjunto de mediciones u observaciones sobre unidades muestrales, cuyos posibles valores son azarosos, es decir no se puede prever con exactitud, existiendo sobre él un cierto grado de incertidumbre.

- Tipo de variables

  En la concepción moderna, los valores que puede asumir una variable pueden ser numéricos o atributos; si son números, la variable se denomina *cuantitativa*; mientras que si son atributos se denomina *cualitativa*. Una variable cuantitativa puede ser *continua* (si entre dos valores cualesquiera de su dominio existe un número infinito de posibles valores) o *discreta* (si entre dos valores cualesquiera de su dominio existe un número finito de posibles valores). Por otro lado una variable cualitativa también es conocida como variable *categórica* ya que dichas variables van a estar definidas por las clases o categorías que la componen. Son algunos ejemplos de variables:

Cuantitativas – Continuas	***Cuantitativas – Discretas***	**Cualitativas**
- Altura de árboles	- Número de árboles	- Orientación de los vientos
- Registros de temperatura	- Número de semillas	- Variedades de un cultivo
- Rendimiento de un cultivo	- Cantidad de frutos	- Grado de ataque de una plaga (severo-moderado-leve)
- Volumen de producción	- Número de mazorcas en plantas de maíz	- Color de corteza
- Peso de animales		

- Método de muestreo

  El método o diseño de muestreo puede definirse como el conjunto de procedimientos que se realizan para obtener una muestra lo más representativa posible de la población. Existen varios procedimientos de muestreo que presentan diferentes características y son apropiados en distintas situaciones, con dependencia entre otros aspectos de los datos disponibles, el costo de muestreo y la precisión requerida en las estimaciones. Los métodos básicos de muestreo son: el Muestreo Aleatorio Simple, el Muestreo Estratificado y el Muestreo por conglomerados.

- Parámetro y Estimador

  El *parámetro* es una constante que caracteriza a una población. Si por ejemplo consideramos que nuestra población esta formada por los datos (observaciones) del diámetro a la altura del pecho (DAP) de todos los árboles que hay en un monte forestal, el diámetro medio de todos los árboles (μ) es un parámetro; como también lo será la varianza de esos diámetros (σ^2). Pero si el estudio de la población lo hacemos a través de una muestra extraída de la misma población, el diámetro medio ($\overline{X}$) como la varianza (S^2) obtenidos a partir de dicha muestra serán variables y se denominarán *estimadores.* Si consideramos que de una población se pueden tomar más de una muestra, dispondremos entonces de tantos estimadores como muestras, razón por la cual debemos tener en cuenta que *los estimadores son variables aleatorias* con todas las características propias de las mismas.

Descripción y Resumen de la Información Registrada

Al registrar los datos provenientes de un estudio observacional o experimental, se puede simplemente realizar un listado de cada valor (datos sin agrupar) o se puede contabilizar las veces que se repiten los valores dentro de clases de extensión predeterminada (datos agrupados), básicamente en este caso para el análisis (numérico y gráfico) se parte de tablas de frecuencias.

Muchas veces el número de datos que se recoge es lo suficientemente grande como para que sea necesario agruparlos en clases, ya que por sí solos no nos dicen demasiado. La ventaja de agrupar las observaciones pasa por facilitar la representación gráfica de los datos, lo que da una idea más adecuada de la forma de la distribución de los mismos. Por otro lado es posible, a partir de datos agrupados, obtener estimaciones como por ejemplo

de la media, varianza, etc., pero se introduce una disminución en la precisión de dichas estimaciones.

A continuación se presentarán las medidas, tablas y gráficos comúnmente utilizados para la descripción de un conjunto de datos.

Medidas de Posición

Estas medidas permiten conocer el valor al que tiende el conjunto de las observaciones bajo análisis. Las más aplicadas son la media aritmética, la mediana, la moda y los cuantiles.

- Media: es el cociente entre la suma de todos los valores y el número de observaciones.

$$\overline{X} = \frac{\sum x_i}{n}$$

Con las siguientes propiedades más destacadas:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0$$

$$\overline{a + b \,.\, X} = a + b \,.\, \overline{X} \qquad \text{o} \qquad M(a + b \,.\, X) = a + b \,.\, M(X)$$

A pesar de ser un muy buen estadístico de posición, posee algunos inconvenientes:

- Uno de ellos es que es muy sensible a los valores extremos de la variable: ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema (atípica), hará que la media se desplace en esa dirección.

- En consecuencia, no es recomendable usar la media como medida central en las distribuciones muy asimétricas.

- Mediana: es el valor que divide a las observaciones, previa ordenación de las mismas, en dos subconjuntos de igual probabilidad, es decir que deja el mismo número de observaciones por debajo que por encima de ella. La mediana la calculamos ordenando los valores en orden creciente y tomando el valor central. Si hay un número impar de datos, la mediana viene dada por el valor medio, si el número de datos es par, la mediana viene dada por la semisuma de los valores medios. La posición de la mediana viene dada por (n+1)/2. Por ejemplo:

  Caso A: Observaciones: 1 – 3 – **4** – 9 – 12 , entonces la mediana será 4 (posición 3).

Caso B: Observaciones: 1 –1 – **3 – 5** – 9 – 12 , como el número de observaciones es par la mediana se calculara como la semisuma de 3 y 5, por lo tanto su valor será también 4. En este caso la posición de la Mediana es 3,5.

Si en el Caso B se reemplaza el número 12 por 120 la mediana no se vería afectada pero sí cambiaría la media. Se dice que *la mediana es una medida resistente (robusta) a la existencia de valores muy alejados (atípicos) mientras que la media no lo es.*

- Moda: es el valor más frecuente, el que se repite mayor número de veces. Como se verá luego si la distribución de los valores es unimodal (una sola moda) y tiende a ser simétrica entonces la mediana, la moda y la media tenderán a coincidir. Podemos tener distribuciones con 2 o más modas, en este caso las distribuciones se denominarán bimodales o multimodales respectivamente.

- Cuantiles: se obtienen de las series ordenadas de menor a mayor, los cuantiles pueden dividir a la serie en cuatro (cuartiles), diez (deciles) ó cien (percentiles) grupos iguales. Los cuantiles representan valores de la variables asociados con una determinada probabilidad

  Percentiles: valores que dividen el conjunto de datos ordenados en cien partes iguales: P_1, P_2, ..., P_{99}. Por ejemplo: P_8 = 14 mm/día, indica que hay un 8% de probabilidad que la precipitación sea menor o igual a 14 mm/día y un 92 % de probabilidad de superar dicho registro. El percentil 50 coincide con la mediana ya que sería el nivel que divide a la serie en dos grupos iguales, ambos con un 50% probabilidad.

  Deciles: valores que dividen el conjunto de datos ordenados en diez partes iguales: D_1, D_2, ..., D_9. Por ejemplo: D_1 = 2000 kg/ha, indica que hay un 10 % de probabilidad que el rendimiento sea menor o igual a 2000 kg/ha.

  Cuartiles: valores que dividen el conjunto de datos ordenados en cuatro partes iguales. En este caso el Q_2 coincide con la mediana. Por otro lado el Q_1 (o cuartil inferior) y el Q_3 (o cuartil superior) coinciden con el P_{25} y P_{75} respectivamente.

Medidas de dispersión

Miden cuán esparcidos, dispersos, se encuentran los datos; es decir permiten saber que tan grandes son los alejamientos respecto de las medidas de posición. Las más usadas son la varianza, el desvió estándar, el error estándar de la media, el coeficiente de variación, etc.

- Varianza muestral o cuadrado medio: es la media aritmética de los cuadrados de las diferencias de cada valor de la serie con respecto a la media, expresada en unidades cuadráticas.

$$S^2 = \frac{\sum \left(x_i - \overline{X}\right)^2}{n-1}$$

Propiedades: $V(a + b \cdot X) = b^2 \cdot V(X)$

- Desvió estándar: es la raíz cuadrada de la varianza muestral. Representa el valor promedio del alejamiento entre las observaciones y su media expresado en las mismas unidades de medición que las observaciones. También es conocido como la medida de dispersión absoluta

$$S = \sqrt{S^2}$$

- Error estándar de la media: El conjunto de números con el que se trabaja es normalmente un subconjunto tomado de una población de valores. Ese subconjunto es una muestra. Si se toman varias muestras de una población podrán obtenerse diferentes valores de la media sólo por la variación aleatoria que existe entre muestra y muestra. Esas diferencias serán menores cuanto menor variación tengan los datos en la población y cuanto mayor sea el número de observaciones de las muestras. La variabilidad entre esas medias muéstrales esta dada por el error estándar:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

  Si bien esta es una medida de la variabilidad entre medias de diferentes muestras, puede estimarse a partir de una sola muestra

- Rango: El rango es la diferencia entre el valor mas alto (máximo) y el mas bajo (mínimo) del conjunto de datos bajo análisis. Cuanto mayor es la variabilidad de los datos mayor es el rango.

- Rango intercuartílico: es la diferencia entre el cuartil superior (Q_3) y el inferior (Q_1). Abarca el 50 % central de los datos. Es una medida resistente a la presencia de valores atípicos.

- Coeficiente de variación: es el cociente entre el desvió estándar y la media, expresado en %. Es una medida de dispersión relativa – adimensional. Expresa cuanto representa el desvío sobre la media y por ser adimensional permite comparar la variabilidad entre variables expresadas en distintas unidades. Por ejemplo: nos permite comparar la variabilidad entre peso en kg. y altura en cm. Su valor nos da idea de la precisión de los datos, cuanto más bajo sea más precisos serán los datos.

$$CV = \frac{S}{\bar{X}} \bullet 100$$

  Generalmente podemos considerar valores de CV < al 5% como muy bajos, entre 5-15% bajos, entre 15-30 altos y > 30% muy altos.

Medidas de forma

Miden la forma de la distribución de los datos. De estas medidas las más empleadas son la asimetría y la curtosis.

- Asimetría: si la distribución de frecuencias de los datos tiene una "cola" más larga hacia la izquierda que hacia la derecha con respecto al máximo central, se dice que la distribución

está sesgada a la izquierda o que tiene asimetría negativa. Si por el contrario la distribución de frecuencias tiene una "cola" más larga hacia la derecha que hacia la izquierda con respecto al máximo central, se dice que la distribución está sesgada a la derecha o que tiene asimetría positiva. A las distribuciones de sesgo nulo se las denomina simétricas o con asimetría cero y en este caso la media, la mediana y la moda serán iguales.(Ver figura 1). Cuando realizamos un estudio descriptivo es altamente improbable que la distribución de frecuencias sea totalmente simétrica. En la práctica diremos que la distribución de frecuencias es simétrica si lo es de un modo aproximado. Por otro lado, aún observando cuidadosamente la gráfica, podemos no ver claro de qué lado están las frecuencias más altas. Conviene definir entonces unos estadísticos que ayuden a interpretar la asimetría, a los que llamaremos coeficientes de asimetría. Una de las formas (hay otras que no presentaremos) de calcular la Asimetría es a través del coeficiente del momento de asimetría:

$$As = \frac{m_3}{S^3} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^3}{n\, S^3}$$

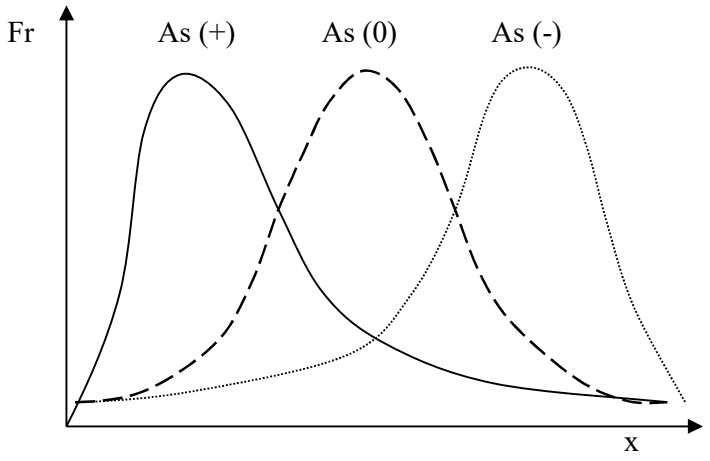


Figura 1

Otra medida más simple es:

$$As = \frac{\overline{x} - Mo}{s}$$

- Curtosis: mide cuán puntiaguda (aguda) es una distribución con un solo máximo (unimodal), con respecto a una distribución Mesocúrtica. El alejamiento de este "modelo" determina dos tipos más de distribuciones. Aquellas que presentan frecuencias mayores, que la mesocúrtica, en el centro de la distribución y menores en los extremos se llaman Leptocúrticas (distribución más aguda que la mesocúrtica), mientras que si las frecuencias mayores se dan en los extremos y las menores en el centro se llaman Platocúrticas (distribución más aplanada que la mesocúrtica) (Ver figura 2). Para calcular la curtosis se

pueden emplear distinto coeficientes, aquí presentaremos el coeficiente del momento de curtosis:

$$Kur = \frac{m_4}{S^4} - 3 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X} \right)^4}{n\, S^4} - 3$$

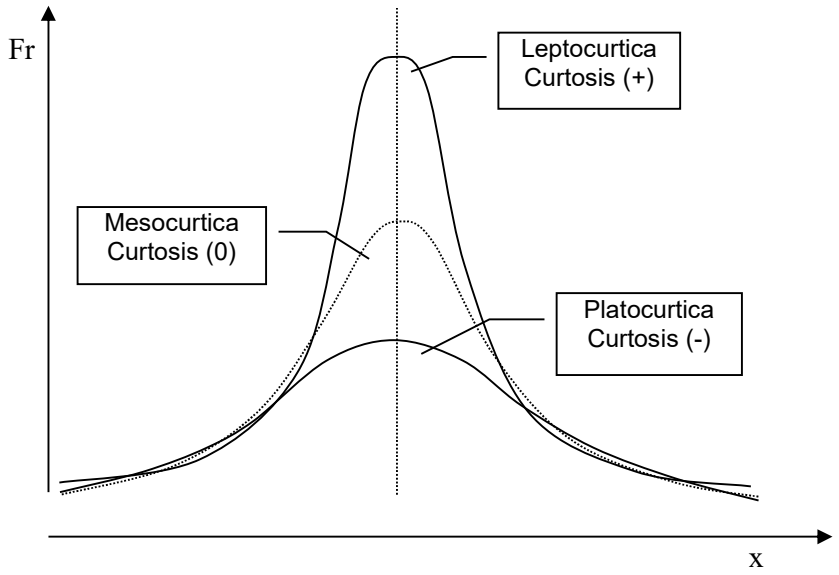


Figura 2

Tablas de frecuencia

Las *tablas de frecuencia* nos permiten agrupar los datos en clases de extensión predeterminada y observar rápidamente la *frecuencia absoluta*, es decir el número de observaciones de cada clase (para variables continuas) o bien el número de veces que se repite cada valor de la variable (para variables discretas) . Generalmente en una tabla de frecuencias no sólo se muestran las frecuencias absolutas, sino también se incluyen las frecuencias relativas y las frecuencias acumuladas (absolutas y relativas). Por otro lado la confección de una tabla de frecuencias puede ser necesario como paso previo para la realización de una serie de gráficos como son el histograma de frecuencias, el polígono de frecuencias, entre otros. (ver tabla 1).

A partir de aquí llamaremos *distribución de frecuencias* al conjunto de clases junto a las frecuencias correspondientes a cada una de ellas.

Una tabla de frecuencias se construye de la siguiente forma:

a) Primero se deben *seleccionar las clases*, para ello pueden seguirse los siguientes criterios en función del tipo de variable que estudiemos:

- Cuando se trate de variables cualitativas, las clases serán de tipo nominal;
- En el caso de variables cuantitativas, existen dos posibilidades:

  - Si la variable es discreta, las clases serán valores numéricos.

  - Si la variable es continua las clases vendrán definidas mediante lo que denominaremos intervalos. En este caso, cada intervalo o clase contendrá todos los valores numéricos comprendidos en cada uno de los mismos.

b) Posteriormente se debe *determinar la cantidad y amplitud de las clases o intervalos*.

Para el caso de variables cualitativas y discretas la cantidad de clases corresponderá a los valores observados de la variable, mientras la amplitud va a estar dada por las características de la variable.

Para variables continuas ambas determinaciones suelen ser arbitrarias, en cuanto a la cantidad de clases el óptimo esta entre 5 y 15 clases puesto que, si no hay suficientes intervalos habrá demasiada concentración de datos y si hay demasiados puede suceder que muchos no contengan observaciones lo que impedirá ver con claridad la estructura de los datos. Como referencia podemos tomar como un número aproximado de intervalos el que se obtiene de aplicar la siguiente expresión: Nº de Clases = 1 + 3,22 log n, siendo n el tamaño de la muestra .

Para determinar la amplitud del intervalo primero se debe calcular el rango (diferencia entre el valor máximo y mínimo del conjunto de datos) y luego dividir este por el número de clases que se quieren tomar. Es decir:

Amplitud del intervalo = rango / número de clases

Definidas la cantidad y amplitud de las clases se deben determinar los límites inferiores y superiores de cada clase. Para ello a partir del valor mínimo (limite inferior de la primer clase) sumamos la amplitud del intervalo y obtenemos el limite superior de la primer clase, el cual será también el limite inferior de la segunda clase y nuevamente volvemos a sumar la amplitud del intervalo a dicho valor y obtenemos el limite superior de la segunda clase, y así continuamos hasta formar las clases fijadas anteriormente. El valor promedio entre los limites del intervalo se llama *punto medio* del intervalo. Por ejemplo: si tenemos una amplitud del intervalo de 8 cm, un valor mínimo de 10 cm, las dos primeras clases tendrán los siguientes limites:

Clases	Limite inferior	Limite superior
1	10	18
2	18	26

Planteados los limites de esta forma si tenemos una observación con un valor de 18 no sabremos a que clase pertenece. Para salvar está situación podemos trabajar con un decimal más que los datos o bien trabajar con limites semiabiertos (incluye uno de los limites el inferior o el superior).

c) Finalmente se deben *calcular las frecuencia relativas y las frecuencias acumuladas absolutas y relativas* como se indica a continuación:

Frecuencia relativa de la clase i es el cociente h_i entre las frecuencias absolutas de dicha clase (f_i) y el número total de observaciones (n), es decir:

$$h_i = \frac{f_i}{n}$$

Obsérvese que h_i es el *tanto por uno* de observaciones que están en la clase i. Multiplicado por 100 representa el porcentaje de la población que comprende esa clase.

Frecuencia absoluta acumulada de la clase i (F_i), es el número de elementos de la población que surge de la acumulación de las frecuencias absolutas hasta la correspondiente clase i.

$$F_i = \sum_{j=1}^{i} f_{ji}$$

Frecuencia relativa acumulada de la clase i (H_i), es el tanto por uno de los elementos de la población que están en alguna de las clases inferiores o igual a la clase i .

$$H_i = \frac{F_i}{n} = \sum_{j=1}^{i} h_j$$

Clase	***L. inferior***	***L. Superior***	***Punto Medio*** (X_i)	***Frecuencia Absoluta*** (f_i)	***Frecuencia Relativa*** (h_i)	***Frec. Abs. Acumulada*** (F_i)	***Frec. Rel. Acumulada*** (H_i)
1	[10	18]	14	5	0.05	5	0.05
2	(18	26]	22	12	0.12	17	0.17
....	....	....]	....	....	....	....	....
10	(82	90]	86	2	0.02	100	1
Totales				**100**	**1**		

Tabla 1 (Tabla de frecuencia para una variable continua)

Como lo mencionamos anteriormente a partir de datos agrupados, es posible también obtener estimaciones de las medidas de posición y dispersión. Las expresiones para el calculo de algunas de ellas son las siguientes:

Media (para datos agrupados)

$$\overline{X} = \frac{\sum f_i \times X_i}{n}$$

Varianza (para datos agrupados)

$$S^2 = \frac{\sum \left(X_i - \overline{X}\right)^2 \times f_i}{n-1}$$

siendo X_i el punto medio de cada clase.

Representaciones gráficas

Existen muchos tipos de gráficos, de acuerdo a los datos, que queremos representar y con que fin. Hoy en día, planillas electrónicas como Excel o programas de estadística traen múltiples opciones para realizar gráficos de gran calidad, por lo que aquí mostramos los tipos básicos de gráficos, que tienen utilidad estadística

- Histograma / Polígono de frecuencias: Para el caso de *variables continuas* un histograma se construye a partir de la tabla de frecuencia, representando sobre cada intervalo, un rectángulo que tiene a este segmento como base y como altura el número de observaciones del conjunto de datos comprendidas dentro de cada intervalo (frecuencias de observaciones). Este tipo de gráfico permite visualizar la forma de la distribución de frecuencias. La frecuencia se representa en el **eje y**, mientras en el **eje x** se indica el límite inferior y superior de cada clase. (Figura 3). Si observamos la figura 3 vemos que existen 7 observaciones mayores a 620 y menores o iguales a 720, 9 observaciones mayor a 720 y menores o igual a 820. Las mayores a 820 se cuentan en la clase siguiente, y así sucesivamente.

En el **eje y** se pueden graficar tanto las frecuencia absolutas, las relativas, como las acumuladas absolutas o relativas. Si unimos los puntos medios de cada barra del histograma obtenemos el *Polígono de Frecuencias*.

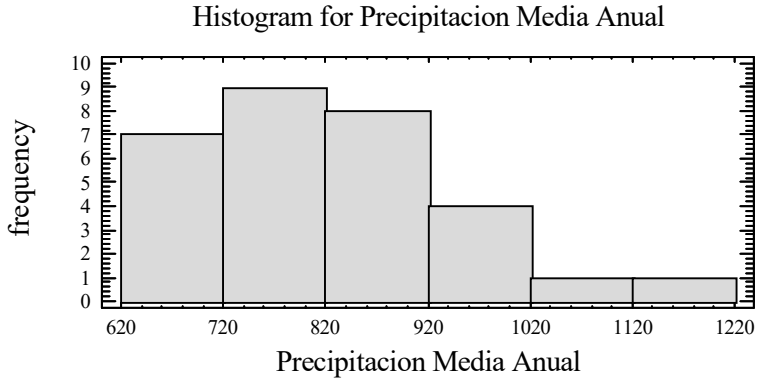


Figura 3: Histograma de frecuencias absolutas para variables continuas

Para variables discretas y cualitativas las frecuencias indican el número de veces que se repite cada valor de la variable o bien el número de observaciones que

corresponden a cada categoría de la variable respectivamente, tomando el nombre para estos casos de ***diagrama de barras*** (figura 4)

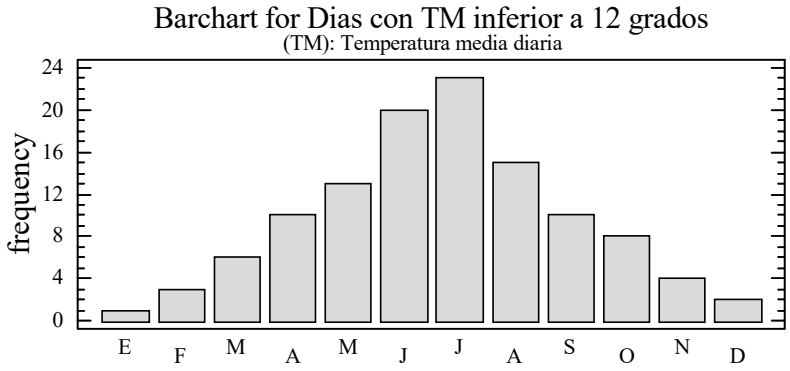


Figura 4: Diagrama de barras para variables cualitativas

Para fines estadísticos, y en el caso de variables continuas es muy importante representar la frecuencia relativa . La importancia del histograma de frecuencias relativas radica en que permite calcular, cuál es la probabilidad de que el elemento en cuestión adquiera un cierto valor, y si representamos las frecuencias relativas acumuladas podemos calcular cuál es la probabilidad de que un cierto valor adquiera un valor menor o igual a uno dado.

- Diagrama de caja (Box Plot o Box and Whisker Plot): es un tipo de gráfico utilizado para representar datos cuantitativos. Permiten visualizar la Asimetría de la distribución e incorpora medidas de posición y dispersión con el objeto de estudiar, también, la variabilidad de los datos y la concentración de los mismos. Las medidas de variación están dadas por el largo de la caja (rango intercuartilico) y por el largo de las barras externas a las cajas (rango), mientras las de posición por el signo suma (media) y por la línea vertical dentro de la caja (mediana). (figura 5).

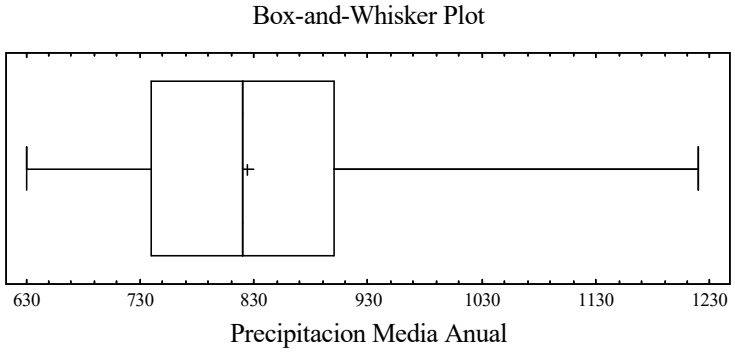


Figura 5: Diagrama de Caja

VARIABLES BIDIMENSIONALES

Lo tratado hasta este punto se refiere a variables unidimensionales. En muchos problemas se hace necesario estudiar simultáneamente dos variables, o lo que es lo mismo, estudiar una variable bidimensional. La información consiste entonces en pares ordenados (X;Y): $(x_1;y_1)$; $(x_2;y_2)$; ..... ; $(x_n;y_n)$.

Un cuadro en dos dimensiones que represente las frecuencias asociadas a este caso es el siguiente:

	y_1	y_2	...	y_j	...	y_k	X
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1k}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2k}	$f_{2.}$
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
x_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{ik}	$f_{i.}$
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
.	.	.	...	.	...	.	.
x_m	f_{m1}	f_{m2}	...	f_{mj}	...	f_{mk}	$f_{m.}$
Y	$f_{.1}$	$f_{.2}$	...	$f_{.j}$	...	$f_{.k}$	$f_{..} = n$

La columna y la fila exterior al cuadro representan la distribución univariada de X e Y respectivamente y se llaman distribuciones marginales de la distribución bivariada. Un cuadro similar puede representar frecuencias relativas.

Se llama *diagrama de dispersión* o diagrama de puntos al gráfico bidimensional sobre un sistema de coordenadas rectangulares, que se forma con todos los puntos que representan a los pares de valores $(x_i ; y_i)$ correspondientes a ambas variables.

Covariancia y correlación:

Existen numerosos coeficiente que cuantifican la relación entre dos variables, un primer coeficiente es el de ***covariacia***. Se lo defiene como la esperanza de los productos de los desvíos de las variables respecto de sus valores esperados.

$$COV(X,Y) = \sigma_{xy} = \frac{\Sigma (x_i - \mu_x) . (y_i - \mu_y)}{n}$$

que es estimada en una muestra por:

$$S_{xy} = \frac{\Sigma (x_i - \overline{x}) . (y_i - \overline{y})}{n - 1}$$

Para interpretar este estadístico, consideremos el diagrama de la Figura 6:

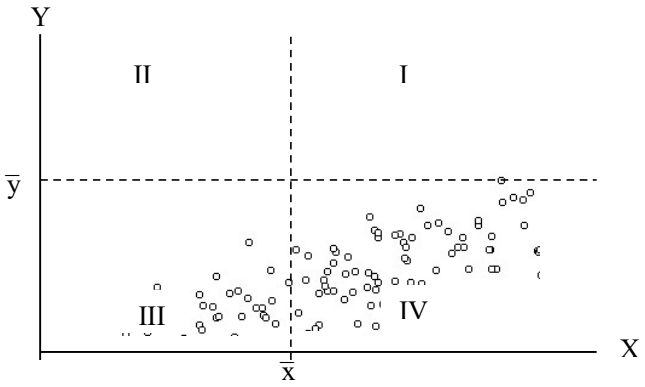


Figura 6

en él observamos cuatro regiones. Los puntos que pertenecen a la región I, verifican las siguientes relaciones:

$$(x_i - \overline{x}) > 0 \;;\; (y_i - \overline{y}) > 0 \qquad \text{luego} \qquad (x_i - \overline{x}) \,.\, (y_i - \overline{y}) > 0$$

y con igual razonamiento tendremos que en la región II y región IV este producto es menor que cero, y en la región III al igual que en la I mayor que cero.

En nuestro ejemplo por la configuración de la nube de puntos, prevalecerán los productos positivos, tanto por la cantidad de puntos situados en las regiones I y III, como por la magnitud de dichos productos. Intuitivamente aceptamos que S_{xy} es mayor que cero, y decimos que la correlación lineal entre X e Y es positiva o proporcional. En las Figuras 7 y 8 se observan casos en que la covariancia es negativa y aproximadamente igual a cero, es decir, con correlación lineal negativa o inversamente proporcional y con correlación lineal nula (lo que no implica independencia entre variables).

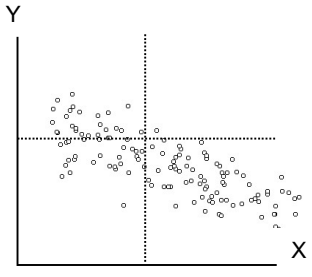


Fig. 7

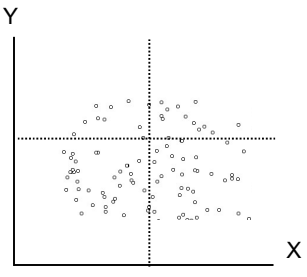


Fig. 8

El inconveniente de este coeficiente es que depende del sistema de unidades elegido y que no es acotado. Para lograr un coeficiente que subsane estas desventajas se define el ***coeficiente de correlación lineal*** que no es otra cosa que la covariancia dividida por los desvíos estándar de las variables.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \text{que se estima por} \qquad r_{xy} = \frac{S_{xy}}{S_x S_y}$$

la fórmula de cálculo usada es:

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

Es inmediato ver que r tomará valores entre -1 y 1. La correlación lineal es perfecta y positiva si r = 1 y, perfecta y negativa si r = -1, con toda la gama de valores entre estas situaciones extremas. Si r = 0 decimos que no hay correlación lineal entre las variables X e Y, pero puede haber correlación no lineal, tal como lo indica el caso de la Figura 9.

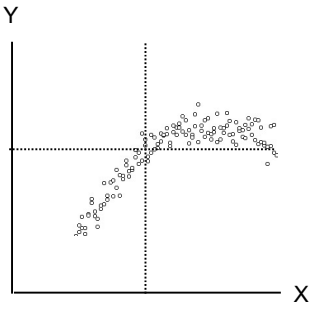


Fig. 9

PROBABILIDADES

Si el único propósito del investigador es describir los resultados de un experimento concreto, la estadística descriptiva puede considerarse suficiente. No obstante, si lo que se pretende es utilizar la información obtenida de una muestra para extraer conclusiones generales sobre la población, entonces la estadística descriptiva constituye sólo el principio del análisis, y debe recurrirse a métodos de inferencia estadística, los cuales implican el uso de la teoría de la probabilidad.

Hay muchos fenómenos reales que se comportan de una manera tan regular que son absolutamente predecibles (muchas leyes físicas), sin embargo existen fenómenos (**fenómenos aleatorios**) en los cuales los resultados no pueden predecirse con certeza, lo que lleva por lo tanto a un **estado de incertidumbre.**

Las ciencias experimentales presentan en sus resultados este tipo de incertidumbre, por ejemplo, saber si un descendiente será macho o hembra, si el crecimiento de un cultivo será afectado por las condiciones climáticas, el color del pelaje de un potrillo, etc; son todos sucesos con grados de incertidumbre variable.

La teoría matemática de la probabilidad pretende establecer reglas de comportamiento general en la ponderación de situaciones o en la ocurrencias de sucesos aleatorios, más que en establecer una definición de la probabilidad.

El cálculo de probabilidades nos suministrará las reglas para el estudio de los experimentos aleatorios o al azar, constituyendo la base para la estadística inductiva o inferencial. De alguna manera, el concepto de probabilidad, nos recordará las propiedades de la frecuencia relativa.

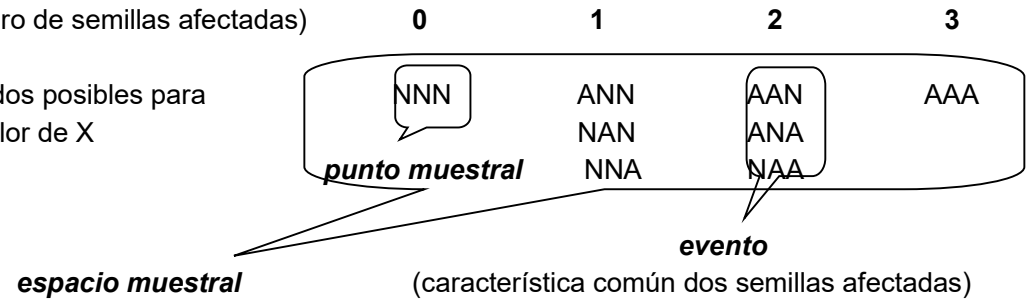
Tal como se citó anteriormente, en las aplicaciones prácticas es importante poder describir los rasgos principales de una distribución, es decir, caracterizar los resultados del experimento aleatorio mediante unos parámetros. Llegamos así al estudio de las características asociadas a una variable aleatoria introduciendo los conceptos de esperanza y varianza matemática, relacionándolos con los conceptos de media y varianza.

Para trabajar con el cálculo de probabilidades y posteriormente con esperanza matemática y sus aplicaciones es necesario fijar previamente ciertos conceptos básicos como punto muestral (suceso), espacio muestral, evento , etc.

Conceptos básicos

Para explicar ciertos conceptos, considerados indispensables, para el cálculo de probabilidades nos basaremos en el siguiente ejemplo:

Supongamos que de un grupo de semillas tomamos 3 semillas una a una, y observamos si las mismas están afectadas (A) o no (N) por una plaga, por lo tanto los resultados posibles serán:



A cada resultado del experimento lo denominaremos **suceso aleatorio o punto muestral**, el conjunto de resultados (sucesos) posibles lo llamaremos **espacio muestral** y a un subconjunto de sucesos del espacio muestral con una característica común que los una se lo denominará **evento.** Por otro lado el conjunto de todos los elementos del espacio muestral que no forman parte de un evento se denominan **complemento del evento.**

Otros ejemplos:

a) Arrojar una moneda

b) Arrojar un dado

c) Extraer elementos de un lote que contiene elementos defectuosos(D) y no defectuosos($\overline{D}$)

d) Extraer dentro de los números naturales los menores de 1000

e) Los sucesos al observar tres pariciones en ovejas, donde M significa macho y H hembra.

Los espacios muestrales asociados a cada ejemplo son :

a) $\Phi_1 = \{C\ ,\ S\}$

b) $\Phi_2 = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$

c) $\Phi_3 = \{D\ ,\ \overline{D}\}$

d) $\Phi_4 = \{0,\ 1,\ 2,\ ...\}$

e) $\Phi_5 = MMM\ ,MMH\ ,MHM\ ,HMM\ ,MHH\ ,HMH\ ,HHM\ ,HHH\ .$

Los espacios muestrales pueden ser ***discretos o continuos***. En los primeros se cuenta con un número finito o infinito numerable de sucesos. En los segundos, los elementos son todos los puntos existentes dentro de un intervalo determinado, siendo no numerables.

Todos los ejemplos vistos hasta ahora tienen espacios muestrales discretos. Un ejemplo de un espacio muestral continuo puede estar dado por la producción de una determinada droga **(x),** la cual varía entre un valor mínimo "a" y otro máximo "b". El espacio

muestral **(Φ)**que se genera al observar la producción en diversos instantes, es continuo, en símbolos:

$$\Phi = \{\ x \in \Phi\ /\ a \leq x \leq b\ \}$$

Hemos llamado **evento** a un subconjunto del espacio muestral. Si Consideramos, en el ejemplo **e)** el evento al menos dos machos (evento D), lo podemos representar como:

D = MMH, MHM , HMM , MMM

$$D \subset \Phi_5 \quad si\ \ x \in D\ /\ x \in \Phi_5$$

Entonces dado el evento D se dice que D está contenido en el espacio Φ_5 si $D \subset \Phi_5$ o D es un subconjunto de Φ_5 ; si **x** es un elemento de D pertenece también al conjunto Φ_5 .

Concepto de probabilidad

- Concepto a priori o Clásico

La probabilidad de un evento determinado es el cociente entre el número de casos favorables a la ocurrencia del evento y el total de casos posibles en el experimento, suponiendo que cada suceso tiene la misma probabilidad de ocurrencia y es posible conocer todos los casos posibles del experimento. Por lo tanto si consideramos a **N** como el número total de casos posibles y $\mathbf{n_A}$ el número de casos favorables a la ocurrencia del evento **A**, tendremos que la probabilidad de ocurrencia del suceso A [**P(A)**] será:

$$P\left(\ A\ \right) = \frac{n_A}{N}$$

Por lo tanto si expresamos la ocurrencia de un evento por valores numéricos tendremos tres situaciones :

si $n_A = N$	P(A) = 1	evento de ocurrencia cierta
si0 < n_A < N	P(A) = entre 0 y 1	evento de ocurrencia incierta
si $n_A = 0$	P(A) = 0	evento de imposible ocurrencia

Ejemplos:

1) Al arrojar un dado la variable X asume los siguientes valores

X = 1 , 2 , 3 , 4 , 5, 6 ; por lo tanto la probabilidad tendremos:

- P(X > 6) = 0 ; ya que en este caso el evento > a 6 es un evento de imposible ocurrencia.
- P(X ≤ 6) = 1 ; ya que en este caso el evento es de ocurrencia cierta.
- P (X = 3) = 1/6 ; ya que en este caso el evento es de ocurrencia incierta.

- P (X = un múltiplo de dos) = 3/6 , al presentar este evento tres casos favorables (2-4-6).

2) Realizo una llamada telefónica y me olvido una cifra y la marco al azar. ¿ Cuál es la probabilidad de que marque el número correcto?

$$P(C) = 1/10 = 0,1$$

- Análisis Combinatorio

En ciertos problemas para conocer el número de sucesos posibles de un experimento debe emplearse algún método del análisis combinatorio que nos facilite su cálculo, dentro de estos métodos veremos a continuación: Permutaciones, Permutaciones repetidas y Combinaciones.

1) Permutaciones con reposición

Se llaman permutaciones con reposición al conjunto de arreglos de r elementos de un total de n elementos tomados de uno en uno permitiendo la reposición de cada elemento luego de su extracción . El número de arreglos que cumplen con esta condición se pueden obtener a partir de la siguiente expresión:

$$P_r^n = n^r$$

2) Permutaciones sin reposición

Se llama permutaciones de n elementos a los diferentes arreglos de esos n elementos de forma que:

- en cada arreglo intervienen todos los n elementos sin repetirse ninguno.
- dos arreglos son diferentes si el orden de los elementos es distinto (influye el orden)

En el caso que se deseen seleccionar r elementos de un total de n elementos, el total de arreglos se determina de la siguiente forma:

$$n\,P\,r = \frac{n\ !}{(n - r)!}$$

Los arreglos de n objetos tomados de n en n son llamados permutaciones de los n objetos. Se tiene que el número de permutaciones de n objetos es n! (es un caso particular de la situación anterior donde r=n).

$$P\,n = n!$$

3) Permutaciones sin reposición pero con repetición de los elementos

Si tenemos n elementos indicados como A1, A2, ..., An donde cada uno de ellos aparecen r1, r2, ..., rk veces, los distintos grupos que pueden formarse con esos n elementos se llaman permutaciones repetidas de forma que:

- en cada arreglo intervienen todos los elementos
- dos arreglos se diferencian en el orden de alguno de sus elementos.

El número de permutaciones de n elementos con repetición de r elementos con multiplicidades r1, r2,..., rk es :

$$n\ P r_{1}\ r_{2}\ r_{k} = \frac{n!}{r_{1}! \times r_{2}! \times r_{k}!}$$

4) Combinaciones

Llamaremos combinaciones de n elementos tomados de x en x veces a todos los arreglos posibles que pueden hacerse con los n elementos de forma que:

- cada arreglo esta formado por x elementos distintos entre sí.
- dos arreglos distintos se diferencian en al menos un elemento, no influyendo el orden .

El número de combinaciones de n elementos tomados de x en x se obtiene por la siguiente expresión:

$$C_{r}^{n} = \binom{n}{r} = \frac{n!}{r! \times (n-r)!}$$

Ejemplo:

1) Si tenemos un grupo de 4 novillos un Hereford (E), un Charolais (C), un Holando (H) y un Abeerden Angus (A).

a) Si se extraen 2 novillos. ¿ Calcular el número de arreglos posibles, si luego de extraer el primer novillo este se repone al grupo?.
Tenemos: n = 4 y r = 2, entonces:

$$P_{2}^{4} = 4^{2} = 4 \times 4 = 16$$ (16 son los arreglos posibles)

(EC, CE, EH, HE, EA, AE, CH, HC, CA, AC, HA, AH, EE, CC, HH, AA)

b) Idem al inciso a pero sin reposición del novillo extraído en primer instancia y considerando que dos arreglos con los mismos elementos en orden diferentes son distintos.

Tenemos: n = 4 y r = 2, entonces:

$${}_4P_2 = \frac{4\ !}{(4-2)!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 12$$ (en este caso 12 son los arreglos posibles)

(EC, CE, EH, HE, EA, AE, CH, HC, CA, AC, HA, AH)

c) ¿ Calcular el número de arreglos posibles que se pueden formar si tomamos grupos de 4 novillos, sin reposición?

$$P\,4 = 4! = 4 \times 3 \times 2 \times 1 = 24$$

(24 son los arreglos posibles)

d) ¿ Calcular el número de arreglos posibles que se pueden formar si tomamos grupos de 2 novillos sin reposición y considerando que dos arreglos con los mismos elementos en orden diferentes son considerados como un único arreglo?

$$C_2^4 = \frac{4\ !}{2\ ! \times (4-2)!} = \frac{4!}{2! \times 2!} = 6$$ (6 son los arreglos posibles)

(EC, EH, EA, CH, CA, HA)

2) Si se cuenta con 5 semillas 3 de girasol y 2 de soja ¿ Cual es el número de arreglos posibles al extraer las 5 semillas?.

$${}_5P_{3,2} = \frac{5!}{3! \times 2!} = 10$$ (10 son los arreglos posibles)

(GGGSS, GGSGS, GSGGS, SGGGS, SGGSG, SGSGG, SSGGG, GSSGG, GGSSG, GSGSG)

Igualmente en este caso podríamos haber aplicado combinaciones:

$$C_3^5 = \frac{5\ !}{3\ ! \times (5-3)!} = \frac{5!}{3! \times 2!} = 10$$

- <u>Concepto a posteriori o Frecuerncial</u>

Existen experimentos cuyos resultados no presentan la misma probabilidad de ocurrencia y el número de sucesos posibles no se conoce, en este caso el concepto anterior es extendido a : **Probabilidad a posteriori o frecuencial.**

La existencia de cierta regularidad que presentan las **frecuencias relativas de un determinado evento** originado por un experimento reproducible en condiciones similares (repetido muchas veces), nos sugiere postular un número **"P",** llamado **probabilidad del suceso** y al cual la frecuencia relativa con que aparece dicho suceso en los experimentos tenderá a aproximarse a medida que aumente el número de repeticiones de los mismos. Es decir:

$$P(C) = \lim_{n\to\infty} \frac{f_C(n)}{n} = \lim_{n\to\infty} h_{(C)}$$

- Desarrollo axiomático de la probabilidad

Sea **S** un espacio muestral diremos que **P(.)** es la función que asigna a cada evento (**R**) del espacio muestral (**S**) un número real entre (**0 , 1**) que llamaremos probabilidad, si satisface los tres **axiomas** siguientes:

1) $0 \leq P(R) \leq 1$; para todo R de S. Siendo R un evento de S

2) **P(S) = 1**.

3) Si **R_1, R_2,** ... es una sucesión de subconjuntos (eventos) mutuamente excluyentes de S, es decir si $R_i \cap R_j = \varnothing$ (conjunto vacio) para i…j , tendremos entonces que:

P ($R_1 \cup R_2 \cup$ ...) = P(R_1) + P(R_2) + ... = P (S) = 1

Operaciones Básicas con eventos

- Adición de eventos

Dados 2 eventos A y B, se llama **unión de A con B** ($A \cup B$), al evento formado por los sucesos que pertenecen a A, a B o a ambos.

Se presentan dos variantes de adición de eventos:

1) *Adición de eventos mutuamente excluyentes*: dos eventos A y B serán **mutuamente excluyentes** cuando no se presentan simultáneamente, es decir cuando la ocurrencia de uno (A) excluye la ocurrencia del otro (B) . En ese caso no tendrán elementos en común:

$A \cap B = \varnothing$ $\varnothing$ = conjunto vacío.

Por lo tanto

P(A $\cup$ B) = P(A) + P(B)

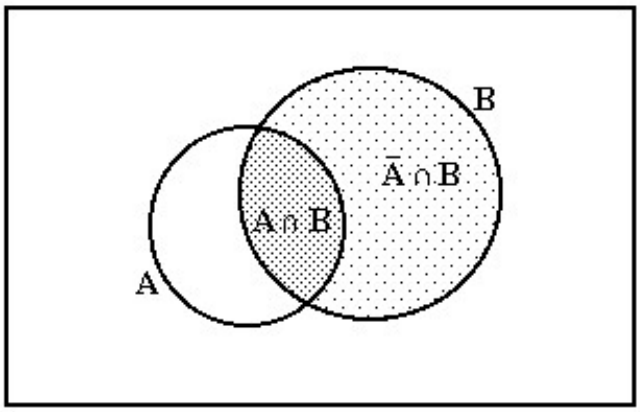
2) *Adición de eventos no mutuamente excluyentes*: en ese caso la ocurrencia de uno no excluye la presencia del otro, dos eventos A y B serán **no mutuamente excluyentes** cuando la ocurrencia de A no excluye la ocurrencia de B, o viceversa. En este caso:

$A \cap B \neq \varnothing$

y

P(A $\cup$ B) = P(A) + P(B) – P(A $\cap$ B)

En este caso P(A ∩ B) se resta para evitar el doble conteo. Esto se puede ver en el siguiente diagrama



Problema de aplicación: La probabilidad de lluvia (LL) en La Plata el día 24 de Diciembre es de 0,25; la de tormenta eléctrica (TE) es 0,08 y la de lluvia y tormenta eléctrica es de 0,04. ¿Cuál es la probabilidad de que llueva o haya tormenta eléctrica ese día?

P(LL) **= 0.25** ; P(TE) **= 0.08** ; P(LL ∩ TE) **= 0.04**

Entonces:

P(LL ∪ TE) = 0.25 + 0.08 - 0.04 **= 0.29**

Ejemplo:

Supongamos una empresa agrícola que cuenta con personal de sexo masculino y femenino, los que se encuentran realizando tareas en el área administrativa, técnica y de mantenimiento, distribuidos de la siguiente manera:

	Administrativos (A)	Técnicos (T)	Mantenimiento (M)	Total
Varones (V)	3	10	4	17 n_V
Mujeres (F)	7	4	2	13 n_F
	10 n_A	14 n_T	6 n_M	30 (N)

- Dos eventos excluyentes son personal Administrativos (A) y de Mantenimiento (M), por lo tanto se debe cumplir:

$$A \cap M = \varnothing$$

Administrativos A = {10 elementos} ; Mantenimiento M = {6 elementos} ; N = 30 elementos

$$P(A \cup M) = P(A) + P(M) = (10/30) + (6/30) = \mathbf{16/30} = \mathbf{8/15}$$

16 es el total de sucesos favorables a la ocurrencia de ambos eventos excluyentes sobre un total de 30 sucesos.

- Dos eventos no excluyentes son personal Administrativos (A) y Varones (V), por lo tanto se debe cumplir:

$$A \cap M \neq \varnothing$$

Administrativos A = { 10 elementos} ; Varones V = { 17 elementos}

En este caso $A \cap V =$ **3** .

$$P(A \cup V) = P(A) + P(V) - P(A \cap V) = (10/30) + (17/30) - (3/30) = \mathbf{24/30} = \mathbf{4/5}$$

24 es el total de sucesos favorables a la ocurrencia de ambos eventos no excluyentes

- Multiplicación de eventos

Dados 2 eventos C y D, se llama **multiplicación C con D** (A ∩ B), al evento formado por los sucesos que pertenecen a C y a D.

Se presentan dos variantes de multiplicación de eventos:

1) *Multiplicación de eventos independientes*: Dos eventos C y D, son **independientes** si la ocurrencia de C no está ligada en forma alguna con la ocurrencia de D. Por lo tanto la probabilidad conjunta de C y D se calcula como:

$$\mathbf{P(C \cap D) = P(C) \times P(D)}$$

2) *Multiplicación de eventos dependientes*: Dos eventos **C** y **D,** son **dependientes** cuando la ocurrencia de C esta ligada a la ocurrencia previa de D, o viceversa. Por lo tanto la probabilidad conjunta de C y D se calcula como:

$$\mathbf{P(C \cap D) = P(C/D) \times P(D)} \qquad \textbf{o bien} \qquad \mathbf{P(C \cap D) = P(D/C) \times P(C)}$$

Donde **P(C/D) y P(D/C)** se definen como **probabilidad condicional** . Para el caso **P(C/D)** como probabilidad condicional del evento **C** habiendo ocurrido **D** y puede expresarse como:

$$P(C/D) = \frac{P(C \cap D)}{P(D)}$$

Siendo: **P (C ∩ D) =** la probabilidad de los sucesos que presentan los atributos C y D
P (D) = la probabilidad marginal del evento D**.**

Ejemplo:

En el ejemplo anterior nos podemos preguntar:

- Caso 1: ¿ Cuál es la probabilidad de seleccionar dos personas y estás resulten Administrativo (A) la primera y Técnico (T) la segunda, habiendo restituido a la primer persona al grupo antes de seleccionar la segunda ?

- Caso 2: ¿ Cuál es la probabilidad que una persona seleccionada al azar sea Administrativo (A), sabiendo que es varón (V)?.

En el caso 1 los eventos **A** y **T** son **independientes**, es decir la probabilidad de A no depende de la ocurrencia o no de T, o viceversa, por lo tanto:

$$\mathbf{P (A \cap T) =} P(A) \times P(T) = (10/30) \times (14/30) = 7/45$$

pudiéndose corroborar que para sucesos independientes la **P (A / T) = P (A) y P (T / A) = P (T)**

En el caso 2 los eventos **A** y **V** son **dependientes**, es decir la ocurrencia de A depende de la ocurrencia previa de V, podemos decir en este caso que A está condicionado a la ocurrencia de V, por lo tanto:

$$\mathbf{P (A / V) =} P (A \cap V) / P(V) = (3/30) / (17/ 30) = \mathbf{3 / 17}$$

Problemas de Aplicación:

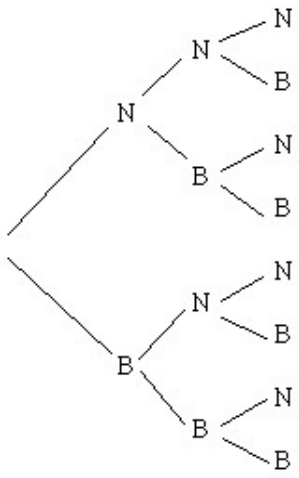
Un examen consta de 7 preguntas, para cada una de ellas se dan 5 respuestas posibles siendo solo una de ellas la correcta. Cuál es la probabilidad de contestar correctamente las 7 preguntas? (suponemos que las preguntas son independientes una de otra).

$$\mathbf{P (C,C,C,C,C,C,C)} = 1/5 \times 1/5 \times 1/5 \times 1/5 \times 1/5 \times 1/5 \times 1/5 = \mathbf{(1/5)^7}$$

Arboles de probabilidades

Cuando un experimento es la repetición de sucesivos experimentos simples, como por ejemplo tirar n veces una moneda y observar la figura que aparece en cada tirada, es útil representar secuencialmente los resultados posibles en un árbol de probabilidades, en que cada rama indica el resultado obtenido en cada uno de los pasos, así como la probabilidad del suceso.

Por ejemplo se extraen 3 bolillas de una urna que contiene 6 bolillas negras y 4 blancas, el árbol correspondiente resulta:



El espacio muestral representado es: Ω = { NNN, NNB, NBN, NBB, BNN, BNB, BBN, BBB}. Para asignar las probabilidades a cada rama, debemos completar la información en el sentido de que si las bolillas extraídas y observadas, son devueltas o no a la urna para la extracción siguiente (modelo con o sin reposición). En el primer caso la probabilidad de una extracción no depende de las anteriores, en cambio en el otro si. Esto se relaciona con la independencia o no de los sucesos, deducida de la independencia o no de los experimentos simples.

La probabilidad de un suceso conjunto (una rama) se obtiene multiplicando las probabilidades de cada tramo de la rama según la ley multiplicativa.

Variable aleatoria

Se llama **variable aleatoria** a todo valor real asignado a los sucesos de un experimento. Por ejemplo en el caso del problema de aplicación anterior del examen que consta de 7 preguntas, la variable correspondiente a respuestas correctas solo podrá tomar los siguientes valores:

X (Nº de respuestas correctas**)**: 0 1 2 3 4 5 6 7

Las probabilidades que se asignan a cada valor que asume la variable aleatoria **X** se obtienen a través de la denominada **función de cuantía [f (x)]** la cual puede ser una expresión matemática o reducirse solamente a una tabla de valores. Por ejemplo: Si tenemos una urna que contiene 3 bolillas negras, 2 rojas y 5 amarillas, y codificamos los colores con los números 1, 2 y 3 respectivamente, las probabilidades para cada valor que sume la variable en estudio puede ser obtenida sin necesidad de construir una expresión matemática, solamente limitándonos a tabular la función obtendremos dichos valores.

X (color que puede tomar la bolilla**)**:	**1**	**2**	**3**
f(x) (función de cuantía**):**	**0,3**	**0,2**	**0,5**

Estas **distribuciones de probabilidades** pueden ser **discretas o continuas**, dependiendo del tipo de variable en estudio.

En el ejemplo de respuestas correctas, podríamos preguntarnos:

$P(x > 4)$; $P(2 < x < 5)$; $P(1 \leq x \leq 4)$

para contestar deberemos definir una nueva función llamada **función acumulativa de probabilidades F(x)**

$$\mathbf{F(x) = \Sigma f(x_i)}$$

- Tipos de Variables Aleatorias

Discreta: Se dice que una variable aleatoria es discreta cuando toma un número finito o infinito numerable de valores en el eje **x** . Por ejemplo : Número de insectos, recuento de maleza en un ensayo , semillas por fruto, etc.

Ejemplo: Si tenemos 4 semillas que pueden estar sanas o enfermas, la variable aleatoria semillas sanas asumirá los siguientes valores:

X (Nº de semillas sanas): 0 1 2 3 4

Si tenemos la función f (x) que asigna las probabilidades a cada uno de estos valores, llamaremos **distribución de probabilidades** la conjunto de valores que puede asumir dicha variable asociados a su respectiva probabilidad..

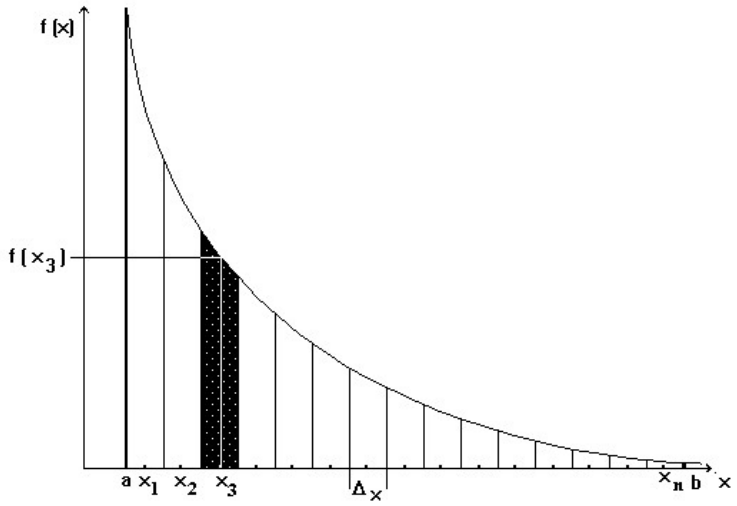
Para cualquier distribución de probabilidad de una variable aleatoria discreta, se debe cumplir:

1) $0 \leq$ **P (x_i)** ≤ 1, para todos los valores x_i de X.

2) Σ **P (x_i)** = 1 (la suma de las probabilidades de todos los valores posibles de X es uno)

Las distribuciones probabilísticas discretas más comunes son la Binomial, la de Poisson, la geométrica, la hipergeométrica, la multinomial , etc. Cada una de ellas se distingue por su función de probabilidad .

Continua: se dice que X es una variable aleatoria continua si asume infinito valores dentro de un intervalo, en el caso de variables discretas podemos asignar a cada valor de la variable su correspondiente probabilidad, en cambio en variables continuas esto no es posible. Abordando esta dificultad podemos considerar intervalos en lugar de puntos (ver figura)



Definimos un intervalo en X y determinamos la probabilidad que un valor x tomado al azar pertenezca a ese intervalo.

En variables continuas los posibles valores que conforman el espacio muestral pertenecen a una función y = f(x) denominada **función densidad**. Como los x pertenecen a una variable continua, su valor exacto no puede ser conocido, pero puede definirse un intervalo en X y determinar la probabilidad de que, tomado un x al azar, éste pertenezca a ese intervalo. Para hallarla sólo debemos realizar el producto de f (x) por el intervalo Δ_x .

En símbolos:

$$P(X_3 < X < X_3 + \Delta_X) = f(X_3)\Delta_X$$

Si reducimos la longitud de cada intervalo a valores infinitamente pequeños ($\Delta x \rightarrow 0$) abordamos el cálculo de probabilidades definiendo un diferencial de probabilidad:

dp = f(x) dx para **x ≤ X ≤ x + dx**

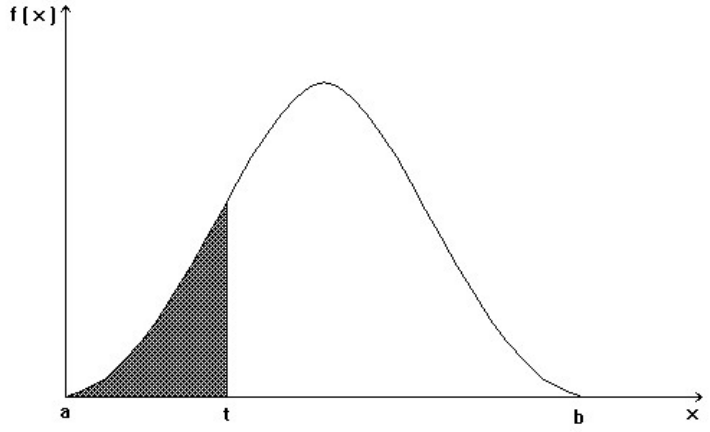
Ahora elegimos un valor "t ", para hallar la probabilidad de que la variable asuma un valor **menor** que t, para a ≤ x ≤ b y f(x) = 0 para todo x fuera del intervalo [a , b]. Para ello, introducimos el concepto de **función acumulativa de distribución de probabilidades F(t)**:

$$F(t) = \int_a^t f(x)dx = P(X \leq t)$$

y :

F(a) = P(X ≤ a) = 0 **F(b) = P(X ≤ b) = 1**

La gráfica de la función densidad se llama "**CURVA DE DISTRIBUCION**" de la variable.



La **función densidad f(x)** la definimos como la derivada primera de la función acumulativa de la distribución de probabilidades:
f (x) = F '(x)

por lo que **F(x)** es la primitiva de **f(x).** Para que **f(x)** sea una función densidad deberá cumplir los siguientes requisitos:

1) **f(x)** ≥ 0 para todo valor de x entre (- ∞ < x < ∞)

2) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

Ejemplo: Siendo la función densidad f(x) = 1/4 para todo -1 ≤ x ≤ 3. ¿Qué valor asume la función densidad para x ≤ -1 y para x ≥ 3? . ¿ Cuánto vale la función de distribución de probabilidades? ¿Cuál será la probabilidad de que la variable asuma valores entre 1 y 1,5?

Entonces:

$$f(x)\begin{cases} = 0 \quad para \quad toda \quad x \le 1 \\ = \frac{1}{4} \quad para \quad toda \quad 1 \le x \le 3 \\ = 0 \quad para \quad toda \quad x \ge 3 \end{cases}$$

la primitiva de f(x) = F(x), entonces F(x) = 1/4 x ,

y, por último:

$$P\left(1 \le x \le 1{,}5\right) = \frac{1}{4}\int_{1}^{1.5} x \, dx = \frac{1}{4}\left[\frac{x^2}{2}\right]_{1}^{1.5} = \frac{1.25}{8} = 15.6\,\%$$

Las distribuciones continuas más comunes son la Normal, la "t", la "F" y la χ^2 (ji cuadrado).

Esperanza y Varianza Matemática

Toda variable aleatoria está completamente determinada por la distribución de probabilidades [x_i ; p (x_i)] si es discreta y por su función de densidad f (x) si es continua. Comúnmente la función de distribución de una variable discreta o continua no se conoce y nos limitamos a conocer ciertos valores que describen la magnitud aleatoria en su totalidad.

Entre estas características numéricas merece destacarse la **esperanza matemática**, que se considera igual al valor medio de la variable aleatoria y la **varianza matemática**.

- Esperanza Matemática de una variable aleatoria discreta

Se la calcula como la suma de los productos de todos los valores que asume la variable por sus respectivas probabilidades. De esta definición , se desprende que la esperanza matemática de una variable aleatoria discreta es una **magnitud no aleatoria (parámetro)** que se puede obtener a partir de la siguiente expresión:

$$E(X) = \sum_{i=1}^{N} x_i \; . \; p_i = \mu$$

Ejemplo:

1) Determinar el valor medio o esperanza matemática del nacimiento de hembras al observar 3 pariciones, tomando que la relación macho:hembra es 1:1.

X (Nº de nacimientos hembras)	:	0	1	2	3
p(x)	:	1/8	3/8	3/8	1/8

Entonces:

E(x) = 0 . (1/8) + 1 . (3/8) + 2 . (3/8) + 3 . (1/8) **= (12/8) = 1,5**

2) Hallar la esperanza matemática del número de veces que ocurre el suceso A en una prueba, si su probabilidad de ocurrencia vale **"p"** y la de no ocurrencia vale **"q"** .

la variable aleatoria X sólo puede tomar dos valores:

$x_1 = 1$ (si el suceso A ocurrió); $P(x_1) = p$

$x_2 = 0$ (si el suceso A no ocurrió); $P(x_2) = 1 - p = q$.

La esperanza matemática buscada se calcula entonces como:

$$E(x) = 1 * p + 0 * q = p$$

Cuando la prueba tiene sólo dos maneras de presentarse como en este ejemplo, la esperanza matemática del número de apariciones de un suceso es igual a la probabilidad de ocurrencia de ese suceso.

- Propiedades de la esperanza

1) **E(c) = c** "La esperanza de una constante es el valor de la constante".

2) **E(cx) = c . E(x)** "La esperanza de una constante multiplicada por una variable es la constante por la esperanza de la variable".

3) **E(x . y) = E(x) . E(y)** "La esperanza del producto de dos variables aleatorias independientes es el producto de sus respectivas esperanzas".

4) **E(x + y) = E(x) + E(y)** "La esperanza de la suma de dos variables aleatorias independientes es la suma de sus respectivas esperanzas".

5) **E[x-E(x)] = 0** "La esperanza matemática de los desvíos de las observaciones respecto del valor medio es 0".

- Esperanza matemática del número de apariciones de un suceso en experimentos repetidos independientes

Si se realizan n pruebas independientes y, en cada una de ellas, la probabilidad de que aparezca el suceso A es constante e igual a "p". ¿Cuánto valdrá el valor medio de apariciones del suceso A en estas n pruebas?.

La respuesta la da el siguiente teorema:

"La esperanza matemática E(x) del número de apariciones del suceso A en n pruebas independientes es igual al producto del número de pruebas por la probabilidad que aparezca el suceso A":

$$\mathbf{E(x) = n \cdot p}$$

Sea X el número total de apariciones del suceso A en las n pruebas, entonces, será igual a la suma de las apariciones de él en las pruebas individuales.

$$\mathbf{X} = x_1 + x_2 + \ldots + x_n$$

$$\mathbf{E(X)} = E\,(x_1 + x_2 + \ldots + x_n)$$

pero, por la propiedad 4:

$$\mathbf{E(X)} = E\,(x_1) + E\,(x_2) + \ldots + E\,(x_n).$$

Puesto que:

$$\mathbf{E(X)} = E\,(x_1) = E\,(x_2) = \ldots = E\,(x_n) = \mathbf{p}$$

Luego: $\mathbf{E(x) = n \cdot p}$

Ejemplo**:**

Consideramos un experimento en el cual se han observado 5000 plantas de ciruelo, de las cuales cada una puede estar sana(S) o enferma (E). El número de sanas es de 3500 y el de enfermas 1500. Sin tomamos al azar 5 plantas, determinar el espacio muestral de plantas sanas y la esperanza matemática del número de plantas sanas.

Entonces: **p** (sana) = 3500/5000 = **0.7** ; **q** (enferma) = 1500/5000 = 0.3 = 1 – p = **0.3**

X (Nº número de plantas sanas)	=	0	1	2	3	4	5
p(x)	=	0,00243	0,02835	0,1323	0,3087	0,36015	0,168

Para X =2 la probabilidad de calcula como: $p^2 \cdot q^3 \cdot C_5^2 = (0.7)^2 \cdot (0.3)^3 \cdot 10 = 0{,}1323$

Luego calculamos la esperanza : **E(X) = 3,5**

y finalmente por lo expresado anteriormente **E(X) = n . p** = 5 . (0.7) = **3,5**

- Varianza de una variable aleatoria discreta

Sabemos que la varianza es una medida de la dispersión de las variables alrededor de la media; si la variable x es de carácter discreto, podemos expresarla de la siguiente manera:

$$V(X) = E\left[x - E(X) \right]^2$$

Ejemplo: Hallar la varianza de la variable aleatoria x que presenta la siguiente distribución de probabilidades:

X :	1	3	5
p (x):	0,3	0,5	0,2

Determinamos E(x):

E(x) = 1 . 0,3 + 3 . 0,5 + 5 . 0,2 = 2,80.

Calculamos los desvíos cuadráticos:

$[x_1 - E(x)]^2 = (1 - 2,8)^2 = 3,24.$

$[x_2 - E(x)]^2 = (3 - 2,8)^2 = 0,04.$

$[x_3 - E(x)]^2 = (5 - 2,8)^2 = 4,84.$

Luego:

V(x)= 3,24*0,3 + 0,04*0,5 + 4,84*0,2= 1,96.

También podríamos haber expresado la varianza como:

por lo tanto: $$V(X) = E\left(X^2 \right) - \left[E(X) \right]^2$$

$E (x^2) = 1 . 0,3 + 9 . 0,5 + 25 . 0,2 = 9,80$

$[E(x)]^2 = 2,80^2 = 7,84.$

V(x) = 9,80 - 7,84= 1,96

- Propiedades de la varianza

1) **V(c) = 0** "La varianza de una constante es 0".

2) **$V(c . x) = c^2 V(x)$** "La varianza de una constante multiplicada por una variable es el cuadrado de la primera por la varianza de la segunda".

3) **V(x + y)= V(x) + V(y)** "La varianza de la suma de dos variables aleatorias independientes es igual a la suma de sus respectivas varianzas".

4) **V(x - y)= V(x) + V(y)** "La varianza de la diferencia de dos variables aleatorias independientes es igual a la suma de sus respectivas varianzas".

- Dispersión del número de apariciones de un suceso en experimentos repetidos independientes

Supongamos que se realizan **n** experimentos independientes y, en cada uno de ellos, la probabilidad de que aparezca el suceso **A** es constante e igual a **"p".** ¿Cómo se calcula la dispersión del número de apariciones del suceso en estos experimentos?

Aquí también recurrimos a un teorema:

"La dispersión del número de apariciones de un suceso A en n pruebas independientes, en cada una de las cuales el suceso tiene una probabilidad de ocurrencia constante e igual a p, es el producto entre número de pruebas y la probabilidad de ocurrencia y la de no ocurrencia de ese suceso". En símbolos:

$$\mathbf{V(x) = n \,.\, p \,.\, q}$$

El total del número de apariciones del suceso A en n pruebas independientes es igual a la suma de las apariciones de él en las pruebas individuales.

$$\mathbf{X} = x_1 + x_2 + \ldots + x_n$$

Por la propiedad 3:

$$\mathbf{V(X)} = V(x_1) + V(x_2) + \ldots + V(x_n)$$

y **:**

$$\mathbf{V(x_1)} = E(x_1^2) - [E(x_1)]^2 =$$

Donde: $\mathbf{[E(x_1)]^2 = p^2}$ y $\mathbf{E(x_1^2)} = 1^2 . p + 0^2 . q = \mathbf{p}$

Entonces:

$$\mathbf{V(x_1)} = p - p^2 = p.(1 - p) = \mathbf{p \,.\, q}$$

Por las características del experimento:

$$\mathbf{V(X) = V(x_1) = V(x_2) = \ldots = V(x_n)}$$

y así llegamos a :

$$\mathbf{V(X) = n \,.\, p \,.\, q}$$

Problema de aplicación:

Se realizan 10 pruebas independientes y, en cada una de ellas, la probabilidad de que ocurra el suceso A es 0,6. Hallar la dispersión del número de apariciones del suceso en estas pruebas.

n = 10 ; p = 0,6 ; q = 1 - 0,6 = 0,4

Entonces: **V(x)** = 10 . 0,6 . 0,4 = **2,4**

- Esperanza Matemática de una variable aleatoria continua

Los conceptos de Esperanza Matemática tienen su aplicación también en el campo de las variables continuas. Sea una variable continua X, definida en el intervalo [a , b] a través de una función f(x) (función densidad).

La esperanza matemática de la variable X es la suma del producto de los posibles valores x_i por las respectivas probabilidades de ubicarse en un intervalo Δx (recuerde lo visto respecto al cálculo de probabilidades en el campo continuo)

En símbolos:

$$\sum x_i \,.\, f(x_i) \,.\, \Delta x$$

Si ($\Delta x \to 0$) resolveremos el problema a través de la integral definida:

$$\boxed{\int_a^b x_i \;.\; f(x_i) \;.\; dx}$$

que da el valor de la Esperanza Matemática de la Variable Continua X, cuyos posibles valores pertenecen al [a , b].

Si la variable estuviera definida en [$-\infty$; ∞] entonces:

$$\boxed{E(x) = \int_{-\infty}^{\infty} x_i \;.\; f(x_i) \;.\; dx}$$

De manera análoga a lo hecho para variables aleatorias discretas obtenemos la dispersión:

$$V(x) = \int_a^b [x - E(x)]^2 .\; f(x) \,.\, dx$$

Desarrollando el cuadrado del binomio nos queda:

$$\boxed{V(x) = \int_a^b x^2 .\; f(x) \;.\; dx - \left[\int_a^b x \;.\; f(x) \;.\; dx\right]^2}$$

Si la variable está definida en el intervalo [-∞ ; ∞] la expresión de la varianza estará definida dentro de ese intervalo.

De esto se desprende, por último, que:

$$\sigma(x) = \sqrt{V(x)}$$

- Covarianza **:**

  Se define como:

$$Cov(x, y) = E\ \{[x - E(x)]\ [y - E(y)]\} = E(xy) - E(x)\,E(y)$$

Para sucesos independientes, la covarianza es nula pues

$$E(x, y) = E(x) \cdot E(y)$$

Distribuciones de Probabilidades Teóricas

Hemos visto cómo los datos observados continuos se pueden agrupan en clases, obteniéndose las ***distribuciones de frecuencias***, las cuales no son otra cosa que un listado de las frecuencias observadas de todos los resultados de un experimento que realmente se efectuó.

Si los intervalos de clase se hacen suficientemente pequeños (es decir, aumentamos gradualmente el número de clases), el histograma llega a transformarse en una superficie limitada por una curva continua que llamamos ***curva de distribución de frecuencias***.

Si, para construir nuestra curva, en vez de frecuencias absolutas (f_i), utilizamos frecuencias relativas (h_i) la curva se transforma en una ***curva de probabilidad***, y la frecuencia relativa (h_i) nos indica la probabilidad de que un individuo de la distribución pertenezca a la clase *i* .

La construcción de una curva de probabilidad en base a los datos observados presenta muchas dificultades, sobre todo porque, para tener una curva que se aproxime a la realidad (que sea confiable), necesitamos un gran número de observaciones o, lo que es lo mismo, disponer de toda la población, cosa a menudo imposible. Una curva basada en una parte de la población (la muestreada), estará siempre expuesta a la llamada "fluctuación de la muestra" (debemos **suponer** que dicha muestra es **representativa** de la población).

Este problema, en apariencia complicado, tiene solución. En efecto, sobre la base de nuestro conocimiento de las causas que motivan la variación de los datos, o

simplemente, formulando una hipótesis acerca de estas causas, podemos deducir cómo "deberían distribuirse" los datos reales si nuestra hipótesis se cumpliera rigurosamente, dando lugar a las llamadas ***distribuciones de probabilidades teóricas***, que no se obtienen de la realidad, sino de hipótesis y deducciones, es decir que a diferencia de las distribuciones de frecuencias una distribución de probabilidad es un listado de las probabilidades de todos los posibles resultados que podrían obtenerse si el experimento se lleva a cabo.

Estas distribuciones han demostrado explicar con bastante certeza aquellos problemas biológicos que son aptos de ser descritos cuantitativamente.

Se han estudiado varias distribuciones teóricas, pero las más importantes y que intervienen en aspectos biológicos son: la Distribución Binomial, estudiada por Santiago Bernoulli, la Distribución Normal, analizada por De Moivre y asociada a Laplace y Gauss, conocida también como "Curva de los Errores" y la Distribución de Poisson, que lleva el nombre de quien la estudió.

En función de las variables analizadas estas distribuciones pueden ser ***Discretas o Continuas***, dentro de las primeras estudiaremos las distribuciones de **BINOMIAL** y de **POISSON**; mientras la distribución **NORMAL**, de variable continua, se analizará más adelante, así como también las distribuciones **t-STUDENT** y **CHI-CUADRADO**.

Distribuciones de Probabilidad Discretas

Distribución Binomial

A este tipo de distribución responden algunas variables discretas o discontinuas. Esta distribución puede emplearse para el calculo de probabilidades cuando contamos con **n** ensayos del mismo experimento donde (1) los resultados de cada ensayo sólo pueden pertenecer a la categoría "**éxito**" o "**fracaso**", (2) los n ensayos deben ser independientes, y (3) la probabilidad de ocurrencia (éxito), deberá permanecer constante en cada ensayo.

Por ejemplo si de una población de insectos extraemos muestras de 5 insectos cada una y examinamos separadamente la presencia o no de un determinado virus en cada uno de ellos, y además sabemos por estudios previos que la proporción (probabilidad) de que dichos insectos estén infectados con el virus es de un 40% , estaremos en presencia de un proceso binomial, ya que:

1) se cumple que el resultado de cada ensayo (observar cada insecto) tiene dos resultados posibles y mutuamente excluyentes como son insecto infectado (éxito) o insecto no infectado (fracaso)
2) los ensayos son independientes: es decir que un insecto este infectado o no no influye en que el siguiente lo este o viceversa.

3) se supone que la población es tan grande que el muestreo sea con o sin reposición no es importante para los fines prácticos, por lo tanto podemos considerar a la probabilidad constante.

El fundamento teórico de esta distribución ha sido detallado en el trabajo práctico anterior (Probabilidad y Esperanza Matemática de un suceso en experiencias repetidas independientes).

La variable aleatoria usual asigna un valor igual a 1 para los éxitos y 0 para los fracasos:

Ejemplo:

1) Si arrojamos una moneda y asignamos X=0 si sale ceca y X=1 si sale cara, la ecuación será de la forma:

$$P(X = x) = 0{,}5 \qquad \text{con } x = 0, 1.$$

Puede leerse así: "La probabilidad de que la variable aleatoria X tome el valor particular x es 0,5 para x = 0 y para x =1". *Esto constituye la distribución de probabilidad.*

Consideremos el problema de obtener una ecuación que dé, en un solo enunciado, todas las probabilidades necesarias de una cierta distribución binomial. Supóngase que un experimento aleatorio consiste en n ensayos independientes. Sea la **probabilidad de éxito** (obtener una cara) [P (éxito)] = **p** y **la probabilidad de fracaso** (obtener ceca) [P (fracaso)] = (1 – p) = **q** , ya que éxito y fracaso cubren todo el espacio muestral y son mutuamente excluyentes, tenemos que:

$$p + q = 1$$

Un resultado del experimento se representará como una sucesión de unos y ceros. Así, 5 lanzamientos de una moneda pueden resultar en (0,0,1,1,0), esto es, dos cecas, dos caras y una ceca al final.

La probabilidad de este resultado puede encontrarse, debido a la independencia de las pruebas, multiplicando las probabilidades que intervienen en cada etapa. Por lo tanto, la probabilidad de que el experimento descripto ocurra es:

$$P(X = 0,0,1,1,0) = (1 - p)\ (1 - p)\ pp\ (1 - p) = p^2\ (1 - p)^3 = p^2\ q^3$$

De esta manera, todos los resultados posibles de obtener (2 caras y 3 cecas) al realizar el experimento que consiste en lanzar una moneda 5 veces a cara o ceca, pueden ser generados por el desarrollo de:

$$\boxed{P(X = x) = C_x^n \cdot p^x \cdot q^{n-x}}$$

Donde, el termino C_x^n representa el número de arreglos (combinaciones) posibles al lanzar 5 veces una moneda, y, en este caso en particular **p** = 0,5; **q** = (1 – p) = 0,5 y **n** = 5.

Como puede observarse, este término surge del desarrollo del binomio:

$$(p+q)^n$$

donde cada término particular determina la probabilidad de ocurrencia de los distintos sucesos posibles en un orden natural siendo x = 0,1,2,3,...,n.

Además, si bien una distribución queda caracterizada por sus parámetros, **que en este caso son n y p**, es importante conocer cuál es la media y la varianza de la distribución.

En la distribución Binomial, la media se calcula como: $\overline{X} = n \,.\, p$

y la varianza: $\sigma^2 = n \,.\, p \,.\, q$

Distribución de Poisson

Esta distribución discreta se relaciona con la distribución binomial, ya que es considerada como una forma limite de la binomial cuando "p" es pequeño y "n" grande. Puede demostrarse que:

$$\lim_{\substack{n\to\infty \\ p\to 0}} = C_x^n \,.\, p^x \,.\, q^{n-x} = \frac{e^{-\lambda} \,.\, \lambda^x}{x!}$$

siendo el segundo término, la expresión de la distribución de **Poisson**, donde λ (lambda) es el parámetro que caracteriza a esta distribución.

También se demuestra que representa una distribución de probabilidades pues:

$$\sum_{x=0}^{\infty} \frac{e^{-\lambda} \,.\, \lambda^x}{x!} = 1$$

Esta distribución tiene dos **propiedades** importantes, a saber:

$$\sum_{x=0}^{\infty} x \,.\, \frac{e^{-\lambda} \,.\, \lambda^x}{x!} = \lambda = E(x) = \mu$$

$$\sum_{x=0}^{\infty} x^2 \,.\, \frac{e^{-\lambda} \,.\, \lambda^x}{x!} - [E(x)]^2 = \lambda = V(x)$$

Aplicaciones

La distribución de Poisson y la Binomial son las distribuciones discretas más ampliamente utilizadas.

Poisson es un modelo probabilistico apropiado para un gran número de fenómenos aleatorios: accidentes de aviación, número de particulas emitidas por una sustancia radioactiva, ocurrencia de mutaciones, recuento de organismos en un medio, de insectos en una parcela de cultivo, número de semillas de malezas en muestras de semillas, etc.

Por otro lado la distribución binomial permite calcular probabilidades cuando se trabaja con variables discretas como numero de semillas, de insectos, de plantas. Se aplica a situaciones en las que se dispone de dos estados mutuamente excluyentes como semillas predadas y no predadas, insectos juveniles y adultos, plantas con flores y sin flores.

Cuando se dispone de muestras de igual tamaño (igual numero de observaciones de semillas, insectos, plantas) es posible calcular el numero medio de semillas predadas, insectos adultos, plantas con flor por muestra. Por ejemplo: Si se toman 15 muestras de 20 semillas cada una se puede contar cuantas tuvieron 0, 1, 2 ... semillas atacadas:

Nro de semillas atacada **(X)**	Nro de muestras **(frecuencia)**	**(x . frecuencia)**
0	2	0
1	2	2
2	4	8
3	5	15
4	1	4
5	1	0
Total	**15**	**29**

El *numero medio de semillas atacadas por muestra* es la sumatoria de los productos del valor de la variable x (numero de semillas atacadas) por el numero de muestras (frecuencia) que presentaron ese valor, dividido el número total de muestras. Es decir:

$$\overline{X} = \frac{\sum x \,.\, f}{\sum f}$$

Para el ejemplo tenemos:

$$\overline{X} = \frac{29}{15} = 1{,}93$$

Por lo tanto el número medio de semillas atacadas por muestra es **1,93**.

La *probabilidad media de semillas atacadas* (***p***) es el numero medio de semillas atacadas por muestra dividido por el numero total de semillas por muestra que se denomina "**n**". Es decir:

$$p = \frac{\overline{X}}{n} = \frac{\sum x \,.\, f}{(\sum f) \,.\, n}$$

Para este ejemplo n = 20, entonces:

$$p = \frac{1{,}93}{20} = 0{,}096$$

Como **p** vale 0,096 la probabilidad promedio de semillas no atacadas que se denomina "***q*** ", que se calcula como 1 – p, tomará un valor de 0,904.

A partir de **p, q y n** se puede calcular la probabilidad de encontrar una muestra con 0, 1, 2... semillas atacadas, utilizando la expresión:

$$P(X = x) = C_{x}^{n} \,.\, p^{x} \,.\, q^{n-x}$$

Si se multiplican esas probabilidades por el numero de muestras (15) se obtienen las *frecuencias teóricas o esperadas*. Es decir:

$$\text{f. teórica}_{(x)} = P_{(x)} \,.\, N$$

Estas son las frecuencias que deberían observarse si la variable en estudio tiene una distribución binomial. Esto puede cumplirse o no dependiendo de la naturaleza del fenómeno estudiado.

Todos los casos en los que se trabaja con la distribución binomial debe estar bien definido el numero total de casos n sobre el cual se observa el estado de la variable. En el ejemplo el numero de casos por muestra fue n =20 semillas y el estado de la variable fue x= 0, 1, 2.... semillas atacadas.

En algunas situaciones n puede no estar definido. Por ejemplo si se observan 15 plantas y se estudia el numero de insectos por planta podría pensarse que 15 plantas es el numero de muestras como antes fueron 15 muestras de 20 semillas. En el ejemplo anterior había 20 semillas que podían estar atacadas o no. En el nuevo ejemplo la variable tiene dos estados que son insectos que están en la planta e insectos que no lo están. Se conoce el numero de insectos que aparecen en la planta pero no el numero total de insectos, ya que esto seria la suma de los que están en la planta mas los que no están en la planta.

En estos casos se utiliza la distribución de Poisson que no utiliza el numero total de casos y supone que el estado no considerado es mucho mayor que el considerado. Es decir que se supone que hay muchos mas insectos fuera de la planta que en la planta.

Numero de insectos en la planta ***(x)***	Numero de muestras ***(frecuencia)***	***(x . frecuencia)***
10	2	20
11	3	33
12	5	60
13	4	52
14	1	14
15	0	0
Total	**15**	**179**

El número medio de insectos por planta se calcula igual que antes como la sumatoria de los productos del valor de la variable x (numero de insectos por planta) multiplicado por la frecuencia (numero de plantas con ese valor).

El numero medio de insectos por planta se denomina lambda (λ). En este caso su valor es de 11,93 ; lo que indica que hay en promedio aproximadamente 12 insectos por planta.

A partir de λ se puede calcular la probabilidad de encontrar una planta con 11, 12... insectos, a partir de la siguiente expresión:

$$P\ (X=x) = \frac{e^{-\lambda} \cdot \lambda^{x}}{x!}$$

Posteriormente si es necesario se pueden calcular las frecuencias teóricas, aplicando la misma metodología que la vista para la distribución binomial.

Un caso típico de la distribución de Poisson es cuando se realizan recuentos en parcelas de superficie fija. Se analiza el numero medio de plantas por parcela. Los estados de la variable son plantas dentro de la parcela y plantas fuera de la parcela. El numero de plantas fuera de la parcela es desconocido y se supone mucho mayor que el numero de plantas dentro. También es posible utilizar la distribución de Poisson cuando el numero total es conocido pero n es muy grande en relación a los valores de x. Por ejemplo si se estudia la ocurrencia de semillas predadas en muestras de 1000 semillas y solo se encuentran muestras 1, 2 o 3 semillas predadas puede utilizarse la distribución de Poisson.

Distribuciones de Probabilidades Continuas

En la sección anterior hemos presentado y desarrollado las principales distribuciones de probabilidades de variables discretas de aplicación en fenómenos aleatorios biológicos como son la distribución Binomial y la de Poisson, en está sección trataremos la ***distribución de probabilidad Normal*** la más importante de las distribuciones de variable aleatoria continua.

Distribución Normal

La distribución Normal, desempeña un papel central en la teoría y la práctica de la estadística. Muchos fenómenos agronómicos, biológicos, químicos, físicos, etc., son estudiados a partir de datos distribuidos de manera normal.

El uso extendido de la distribución normal en las aplicaciones estadísticas puede explicarse, además, por otras razones. Muchos de los procedimientos estadísticos habitualmente utilizados asumen la normalidad de los datos observados. Aunque muchas de estas técnicas no son demasiado sensibles a desviaciones de la normal y, en general, esta hipótesis puede obviarse cuando se dispone de un número suficiente de datos, resulta recomendable contrastar siempre si se puede asumir o no una distribución normal de los datos.

Se dice que una variable aleatoria continua tiene distribución normal con media (μ) y varianza (σ^2), si su función de densidad está dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

La gráfica de la función densidad normal es una curva simétrica, mesocurtica, asintótica con respecto al eje x y centrada en la media (μ), con forma de campana. (ver fig 1)

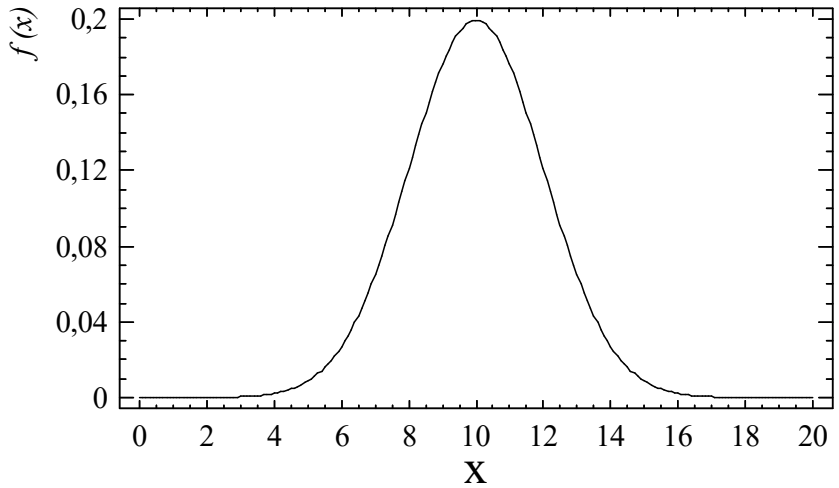


Figura 1: Curva con $\mu = 10$ y $\sigma = 2$

En el eje x del gráfico 1 van los posibles valores que puede tomar la variable aleatoria X y en el eje y los valores de la función densidad normal.

Propiedades de la distribución Normal

1.- Si determinamos el valor de la derivada primera en el punto 0 vemos que su valor es:

$$f'(\mu) = 0$$

lo que indica la existencia de un extremo relativo en ese punto. La derivada segunda vale:

$$f''(\mu) = \frac{f(\mu)}{\sigma^2}$$

lo que nos indica, finalmente que nuestra curva de distribución normal tiene un **máximo relativo** en μ.. Es decir que el valor máximo de ***f (x)*** se presenta en $\mathbf{x = \mu}$

2.- Se comprueba que:

$$f''(x) = 0 \qquad \begin{array}{c} si\ (x - \mu) = \pm\sigma \\ (o\ \ sea\ \ en\ \ x = \mu \pm \sigma) \end{array}$$

lo que nos permite afirmar que la curva de distribución normal tiene dos puntos de inflexión en:

$$x_1 = \mu + \sigma$$
$$x_2 = \mu - \sigma$$

3.- La función es simétrica, ya que:

$$f(\mu + k) = f(\mu - k)$$

siendo k cualquier valor. Esto significa que <u>*la media, la mediana y la moda son coincidentes*</u>.

4.- El área abarcada entre la curva y el eje **x** es igual a 1:

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

5.- Por la propiedad 3 (función simétrica con respecto a la media) existe una probabilidad de un 50% de observar un dato mayor que la media, y un 50% de observar un dato menor.

6.- El área bajo la curva comprendido entre los valores situados aproximadamente a una desviación estándar y a dos desviaciones estándar de la media es aproximadamente de 0.68 y 0.95 respectivamente. Es decir:

$$\mathbf{P}\ (\ \mu - \sigma \leq \mathbf{X} \leq \mu + \sigma) = \mathbf{0.68}$$

$$\mathbf{P}\ (\ \mu - 2\,\sigma \leq \mathbf{X} \leq \mu + 2\,\sigma) = \mathbf{0.95}$$

7.- Se trata de una función mesocúrtica, definida a partir del coeficiente de curtosis y, a partir de él hemos definido el grado de achatamiento o aguzamiento de las distribuciones respecto de la curva normal. (ver Estadística Descriptiva coeficiente de curtosis)

Entonces, la gráfica de la función de distribución de probabilidades de la Distribución Norma es:

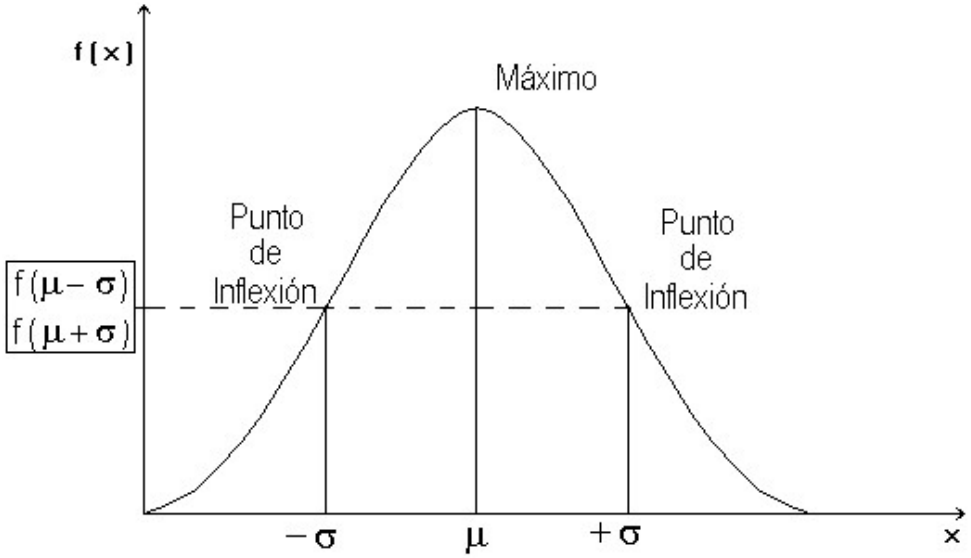


Figura 2: Curva de la Distribución Normal

En la figura 3 se muestra el efecto de cada uno de los parámetros, media (μ) y desvió estándar (σ), sobre la forma de la curva.

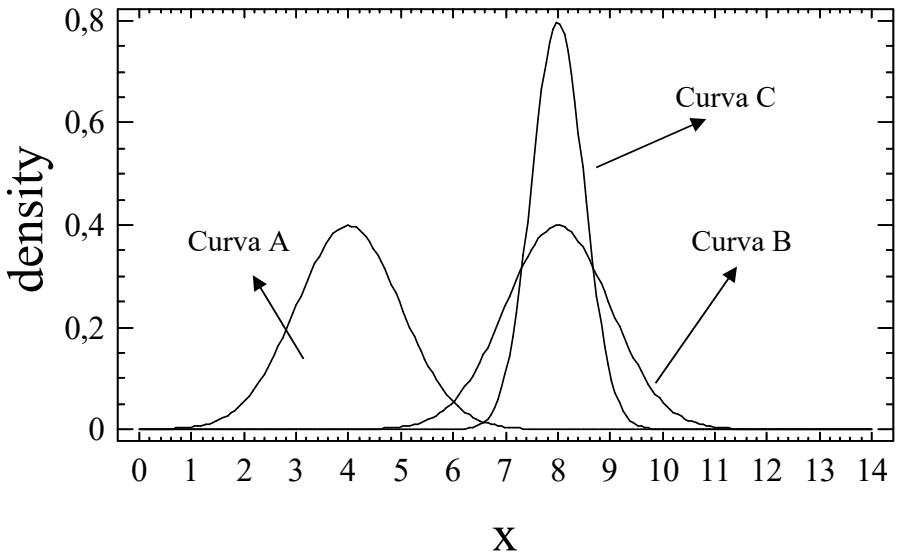


Figura 3: Curva **A** ($\mu = 4$; $\sigma = 1$); Curva **B** ($\mu = 8$; $\sigma = 1$); Curva **C** ($\mu = 8$; $\sigma = 0,5$)

Como se puede apreciar en la figura 3, la ***variación de la media (μ) manteniendo el desvió estándar (σ) fijo*** (caso de las curvas A y B) produce el desplazamiento de la función en el sentido del eje x.

En cambio, la ***variación del desvió estándar (σ) manteniendo la media (μ) fija*** (situación de las curvas B y C), produce el "aplastamiento" (para un σ grande) o "aguzamiento" (si el σ es chico) de la función.

Por lo visto hasta aquí la distribución de una variable aleatoria continua que presenta distribución normal quedará completamente determinada por los ***parámetros***, ***media*** *(μ)* *y* ***desvió estándar*** *(σ)*.

Cuando una variable tiene distribución normal se la suele indicar con una letra N, mayúscula de imprenta, y, entre paréntesis, **la media y la varianza**. Esto es:

$$\mathbf{X \sim N (\mu ; \sigma^2)}$$

Llamamos función de probabilidad a la función F(x) definida como:

$$F(x) = P(X \geq x) = \int_{x}^{\infty} f(x) \, . \, dx$$

siendo su valor integrable siempre positivo o nulo

Si se quiere conocer la probabilidad de que una variable distribuida normalmente (con media y desvió estándar conocido) tome un valor entre x_1 y x_2 , se deberá integrar la función densidad normal (*f (x)*) de la siguiente manera:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) \; . \; dx$$

Por lo tanto, a fin de evitar la necesidad de recurrir a la integración para obtener la probabilidad respectiva en cada problema que aparezca, ya que existe toda una familia de curvas normales, que difieren en la media y/o en el desvió estándar, se hace uso de una transformación que hace que variables distribuidas normalmente con funciones de densidad normal diferentes, se distribuyan de la misma manera, facilitando así los cálculos de probabilidades bajo cualquier combinación de los parámetros media y desvió estándar.

Estandarización

Se llama estandarización a la transformación de la variable X en una nueva variable Z. Está variable se obtiene a partir de la siguiente expresión:

$$\boxed{Z = \frac{(x - \mu)}{\sigma}}$$

Con esta transformación, nuestros datos, independientemente de sus valores numéricos, adquieren varianza 1 y media 0. Por lo tanto la función de densidad normal de está nueva variable es:

$$f(z) = \frac{1}{\sqrt{2\pi}} \, e^{-\frac{z^2}{2}} \qquad\qquad -\infty < Z < \infty$$

A partir de esta transformación se ha reducido el problema de tener muchas distribuciones, a tener una sola. Pero para hallar las probabilidades de que X tome un valor entre dos valores determinados se deberá aún integrar la función de densidad de la distribución normal estandarizada ($f_{(z)}$).

Afortunadamente, las integrales para el cálculo de probabilidades de está ultima función han sido calculadas y tabuladas (ver tabla al final de la sección).

Le valor de Z con el cual ingresamos a la tabla para el cálculo de las áreas (probabilidades), es el que se obtiene de aplicar la transformación precedente.

La probabilidad (área) que entrega la tabla es la comprendida entre el valor de Z (con el cual ingresamos a la tabla) y el infinito, y solo para los valores positivos de Z, es decir:

$$\boxed{P(z > 1{,}5) = \int_{1{,}5}^{\infty} f(z) \; dz = 0{,}0663}$$

Consideremos, por ejemplo, el siguiente problema: supongamos que se sabe que el peso de los cerdos de una determinada raza sigue una distribución aproximadamente

normal, con una media de 80 Kg y una desviación estándar de 10 Kg. ¿Podremos saber cuál es la probabilidad de que un cerdo, elegido al azar, tenga un peso superior a 100 Kg?

Como primer paso deberemos transformar la variable X en Z según la transformación:

$$Z = \frac{(x - \mu)}{\sigma} = \frac{(100 - 80)}{10} = 2$$

La probabilidad que se desea calcular se obtendrá directamente de la tabla y será:

$$P(X > 100) = P(Z > 2) = 0.0228$$

Por lo tanto, la probabilidad buscada de que un cerdo elegido al azar de esa población tenga un peso mayor de 100 Kg , es de 0.0228, es decir, aproximadamente de un 2.3%.

Si en lugar de buscar la probabilidad de que un cerdo, elegido al azar, tenga un peso superior a 100 Kg., deseamos calcular la probabilidad de que ese cerdo pese menos de 50 kg., debemos proceder de la siguiente forma:

$$Z = \frac{(x - \mu)}{\sigma} = \frac{(50 - 80)}{10} = -3$$

La probabilidad que se desea calcular será:

$$P(X < 50) = P(Z < -3)$$

Como se trata de una distribución simétrica (propiedad 3), tenemos:

$$P(Z < -3) = P(Z > 3)$$

Por lo tanto

$$P(X < 50) = P(Z < -3) = P(Z > 3) = 0.0013$$

Es decir que la probabilidad de que un cerdo elegido al azar de esa población tenga un peso inferior a 50 Kg , es de 0.0013, es decir, aproximadamente de un 0.13%.

Si quisiéramos conocer la probabilidad de seleccionar un cerdo al azar y que este tenga un peso comprendido entre 50 y 100 Kg. los pasos a seguir serian los siguientes:

Por la propiedad 4 el área total bajo la curva es igual a 1, por lo tanto se puede deducir que:

$$P(50 < X < 100) = 1 - [P(Z < -3) + P(Z > 2)]$$
$$P(50 < X < 100) = 1 - [0.0013 + 0.0228] = 0.9759$$

Finalmente, la probabilidad buscada de que un cerdo elegido al azar tenga un peso entre 50 y 100 Kg., es de 0.9759, es decir, aproximadamente de un 97.6%.

Importancia de la Distribución Normal

En los módulos posteriores veremos que la distribución normal desempeña un papel predominante. Numerosas poblaciones que son estudiadas en los más diversos campos parecen tener una distribución aproximada a la normal, por ejemplo medidas físicas del cuerpo, medidas de calidad de muchos procesos industriales o los errores no sistemáticos que se provocan en muchos procesos de medición. Una justificación de la frecuente aparición de la distribución normal es el *teorema central del límite*, el cual establece que cuando los resultados de un experimento se deban a un conjunto muy grande de causas independientes, que actúan sumando sus efectos, siendo su efecto individual de poca importancia respecto al conjunto, se espera que los resultados sigan una distribución normal.

Otra consideración es que las distribuciones en el muestreo basadas en una distribución madre que sea normal, resultan más cómodas desde el punto de vista analítico.

Al aplicar los métodos estadísticos basados en la distribución normal el investigador deberá conocer aproximadamente la forma general de la distribución que siguen sus datos. Si esta es normal usará directamente los métodos, en caso contrario transformará los datos de modo que las observaciones se aproximen a una distribución normal. Caso contrario se podrán aplicar métodos de distribución libre conocidos como métodos no paramétricos (los cuales no presentaremos en este curso).

Inferencia Estadística

Uno de los objetivos de la estadística es hacer inferencia con respecto a la población basándose en la información contenida en una muestra. Como las poblaciones se describen mediante medidas numéricas denominadas parámetros, el objetivo de la mayoría de las investigaciones estadísticas es hacer una inferencia con respecto a uno o más parámetros de la población.

En muchas aplicaciones prácticas a veces solo nos interesa ***estimar los valores de los parámetros*** de una distribución (media y varianza). Por ejemplo: saber cual es el rendimiento de una nueva variedad de trigo obtenida por mejoramiento genético.

En otras actividades muchas veces se deberá decidir si la declaración respecto a uno o más parámetros de una o más poblaciones es verdadera o falsa. Por ejemplo: probar que la producción en forraje de una nueva variedad de sorgo supera a las variedades que se siembran en la actualidad, que la utilización de una sierra sinfin adicional en un aserradero aumentará significativamente la producción de tablas, que la siembra directa es mejor que la siembra convencional, etc. Es decir que para tomar una decisión, que tal cosa es mejor, superior o bien que le conviene al productor o industrial poseerla o utilizarla debemos contar con procedimientos que nos permitan tomar dichas decisiones. Estos procedimientos se denominan ***pruebas de hipótesis.***

En base a lo expresado precedentemente podemos definir dos grandes áreas dentro de la inferencia estadística, ellas son:

1. *Estimación de Parámetros*
2. *Pruebas de Hipótesis*

- ESTIMACION DE PARAMETROS

La estimación es el proceso mediante el cual obtenemos estimadores que se aproximen al valor del parámetro poblacional partiendo de la información proveniente de una muestra.

La estimación de un parámetro desconocido puede ser ***Puntual*** (un único valor o punto) o ***por Intervalo*** (un rango de valores dentro del cual se espera el verdadero valor poblacional).

a) Estimación Puntual

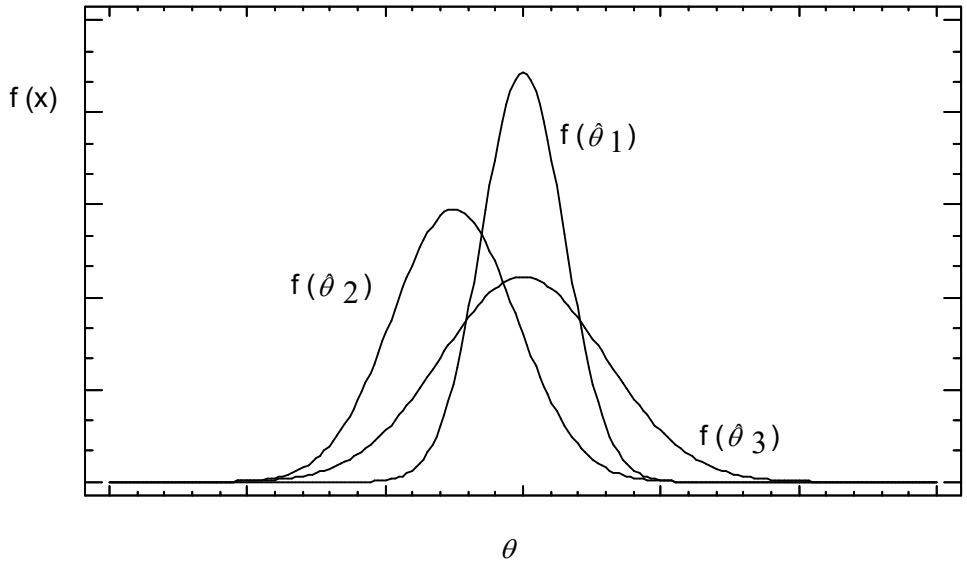
Si designamos a un parámetro poblacional desconocido como "θ " (tita); la estimación puntual consiste en elegir un estimador ($\hat{\theta}$), cuyo valor numérico particular se toma como aproximación del parámetro θ.

Por ejemplo, el valor de la media muestral $(\overline{X})$ calculado de una muestra de tamaño n, es un estimador puntual de la media poblacional (μ).

Debe quedar claro que todo estimador es una variable aleatoria, por lo cual pueden haber varios estimadores posibles para un parámetro. Si deseamos estimar la media de una variable aleatoria, podemos tomar, por ejemplo, tres muestras o considerar una función de la muestra que brinde al mejor estimador del parámetro.

Intuitivamente, perece obvio que la distribución de un buen estimador debería concentrarse lo más cerca posible del verdadero valor del parámetro poblacional.

Supongamos que $\hat{\theta}_1$, $\hat{\theta}_2$ y $\hat{\theta}_3$ son los diferentes estimadores del parámetro θ, con funciones de densidad de probabilidades que se muestran a continuación:



El estimador $\hat{\theta}_1$ se considerará mejor estimador que $\hat{\theta}_2$ y $\hat{\theta}_3$, porque su distribución muestral se concentra más cerca del verdadero valor de θ.

Para decidir qué estimador es el mejor, deberán estudiarse sus propiedades estadísticas y desarrollar algún criterio para comparar estimadores.

- Propiedades de un estimador

1. Insesgado

Un estimador de θ es insesgado si la esperanza matemática de dicho estimador es igual al parámetro.

$$E(\hat{\theta}) = \theta$$

Caso contrario el estimador es sesgado positivamente o negativamente.

Ejemplo: " la $\overline{X}$ es un estimador insesgado de μ ".

Para demostrarlo, partimos de n medias muestrales obtenidas de una variable con media μ y varianza σ^2 .

Entonces:

$$E(\overline{X}) = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n}\ E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i)\right] = \frac{n\mu}{n} = \mu$$

2. Consistente

Si $\hat{\theta}_n$ es una secuencia de estimadores de θ dado en una muestra aleatoria de tamaño n, se dice que $\hat{\theta}_n$ es consistente para el parámetro θ $\forall$que $\varepsilon > 0$:

$$Lím_{\ n \to \infty}\ P\left[\left|\hat{\theta}_n - \theta\right| < \varepsilon\right] = 1$$

En otras palabras, un estimador consistente es el que está, desde un punto de vista probabilistico, más cerca del parámetro a medida que crece el tamaño de la muestra.

3. Eficiente

Considerando todos los posibles estimadores insesgados de un parámetro, solo aquel con mínima varianza se lo considerará eficiente, "estimador eficiente de θ " o "estimador de mínima varianza de θ ".

Es decir dados dos estimadores ($\hat{\theta}_1$ y $\hat{\theta}_3$) insesgados de un parámetro θ , seleccionaríamos el estimador con la menor varianza ($\hat{\theta}_1$), permaneciendo constante todo lo demás.

Resumiendo diremos que el "mejor" estimador para un determinado parámetro será aquel que sea ***insesgado, consistente y eficiente***.

- Métodos de Estimación Puntual

Existen varios métodos de estimación puntual de los que sólo citaremos algunos:

- *Método de los momentos*
- *Método de los Cuadrados Mínimos Ordinarios*
- *Método de Máxima Verosimilitud*

b) Estimación a Intervalo

A diferencia de la estimación puntual, que nos ofrece solo un valor numérico a partir de alguna formula conocida, la estimación a intervalos nos ofrece un ***rango de valores y una magnitud de la medida de la seguridad*** de que ése rango de valores contenga al verdadero parámetro desconocido de la población. Está última es la magnitud del error debido al muestreo aplicado para obtener la serie de datos y proporciona información acerca de la precisión de la estimación obtenida.

Siendo $X_1, X_2, ..., X_n$ una muestra aleatoria de una población X cuya distribución de probabilidades es f(x; θ), donde θ_1 y θ_2 son dos valores del parámetro θ, tal que $\theta_1 < \theta_2$, entonces, puede determinarse la probabilidad de que:

$$P [\theta_1 < \theta < \theta_2] = \gamma = 1 - \alpha$$

en la cual:

- θ_1 es el límite inferior del intervalo de confianza para θ.
- θ_2 es el límite superior del intervalo de confianza para θ.
- γ es el coeficiente de confianza del intervalo.
- α es el nivel de significancia.

La probabilidad de que un intervalo de confianza contenga θ se conoce como coeficiente de confianza siendo su elección arbitraria pero estando relacionada con las caracteristicas de la experiencia. Sin embargo suelen utilizarse valores de 0,90 (90%) o 0,95 (95%) en general.

Si escogemos un γ de 0,95; establecemos que $[\theta_1 ; \theta_2]$ es el intervalo de confianza que tiene un 95% de probabilidades de contener al verdadero valor poblacional θ . Esto se conoce también como un ***Intervalo de Confianza Bilateral***, pues se especifican los límites inferior y superior. En cambio si solo especificamos la probabilidad de:

$$P [\theta_1 < \theta] = \gamma = 1 - \alpha$$

o bien

$$P [\theta < \theta_2] = \gamma = 1 - \alpha$$

entonces, el intervalo se extiende entre $[\theta_1 ; \infty]$ en el primer caso, denominado ***Intervalo de Confianza Unilateral Inferior***, y entre $[- \infty ; \theta_2]$ en el segundo, llamado ***Intervalo de Confianza Unilateral Superior***.

- Ejemplo:

Sea X una variable de una población cuya curva de distribución presenta μ desconocida y varianza σ^2 conocida. Hallar un intervalo de Confianza para μ. Esto es encontrar θ_1 y θ_2 tal que :

$$P [\theta_1 < \mu < \theta_2] = \gamma$$

Para hallarlos debemos:

1. Elegir el coeficiente de confianza (γ)
2. Tomar una muestra aleatoria y hallar la media muestral ($\overline{X}$)
3. Si el tamaño de la muestra (n) es suficientemente grande (n>30) entonces:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

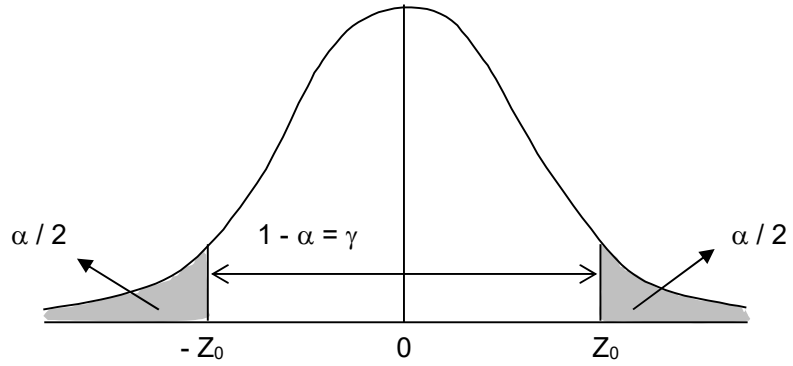
4. Transformar la variable (estandarizarla) para llevarla a una distribución N(0;1).

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

5. Fijado γ podemos determinar los valores Z_1 y Z_2 tal que:

$$P\,[Z_1 < Z < Z_2] = \gamma$$

Como la curva es simétrica, entonces $Z_1 = -Z_0$ y $Z_2 = Z_0$, como se ve en la figura:



Si γ vale 0,95 entonces α valdrá 0,05 ; por lo tanto $\alpha/2$ será 0,025, luego:

$$P\left(-Z_0 < Z < Z_0\right) = \gamma$$

$$P\ \left(-Z_0 < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_0\right) = \gamma$$

$$P\left(-Z_0 \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < Z_0 \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

$$P\left(-\overline{X} - Z_0 \frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + Z_0 \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

$$P\left(\overline{X} - Z_0 \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_0 \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

De donde se obtienen:

$$\ell_{1\mu} = \overline{X} - Z_0 \frac{\sigma}{\sqrt{n}} \qquad \ell_{2\mu} = \overline{X} + Z_0 \frac{\sigma}{\sqrt{n}}$$

que son los límites inferior (ℓ_1) y superior (ℓ_2) de un intervalo que tiene un 95% de probabilidad de contener al verdadero valor de μ.

Nótese que aquí para hacer una estimación del parámetro poblacional μ , necesitamos del valor de otro parámetro poblacional, el desvío estándar σ . En la práctica, generalmente σ es desconocido, por lo que se deberá recurrir a otra distribución que considere esta situación.

Distribución de t

Hasta el momento hemos definido a una variable z como:

$$z = \frac{\overline{X} - \mu}{\sigma_{\mu}}$$

Si la variable estaba referida a una muestra aleatoria de la población.

donde: μ = media poblacional.

$\overline{x}$ = media muestral.

$\sigma_{\mu} = \frac{\sigma}{\sqrt{n}}$ = desvío estándar de la media poblacional.

σ = desvió estándar

Como se ve, **z** depende de los valores de los parámetros μ y σ . Ahora bien, si σ es desconocido y se lo sustituye en la ecuación precedente , por su estimador **S** (error estándar o desvió estándar muestral) se genera una nueva variable aleatoria denominada **"t".**

Esta nueva variable genera una distribución que se origina en la relación entre una diferencia de medias y el error estándar. Esta diferencia puede ser entre una media muestral y su correspondiente media poblacional o entre dos medias muestrales. Para el primer caso, diferencia entre una media muestral y su correspondiente media poblacional, tenemos:

$$t = \frac{\overline{X} - \mu}{s_{\overline{x}}} \qquad (1)$$

donde: $S_{\overline{x}}$ es el error estándar de la media muestral, que es una medida de la dispersión de dicha media.

La distribución de t depende de los grados de libertad **(gl)** de **S** , calculados en este caso como (n – 1), de manera que si se quiere calcular probabilidades (bajo el supuesto de normalidad), la tabla de Z no sirve. Es por eso que se calculó una nueva tabla (al final del texto), en la que figuran las probabilidades de hallar un valor igual o menor de t (distribución acumulada). El uso de está distribución es recomendado en la comparación de medias cuando el tamaño de la muestra es menor que 30. Para tamaños de muestras > a 30 ambas curvas (normal y t) son prácticamente iguales.

La expresión (1) de la distribución de t se aplica para:

- Determinar la significancia de la diferencia entre una media muestral y un valor de μ hipotético.
- Establecer los límites de confianza dentro de los cuales se espera que se encuentre la media poblacional.

En términos más generales, la variable t también se puede definir como la razón entre la diferencia entre medias muestrales de dos poblaciones (por ejemplo: de dos raciones, dos fertilizantes, dos variedades, etc.) y el error estándar de dicha diferencia. Es decir:

$$t = \frac{\overline{X}_A - \overline{X}_B}{S_{\overline{x}_A - \overline{x}_B}} \qquad (2)$$

donde: $S_{\overline{X}_A - \overline{X}_B}$ es el error estándar de la diferencia de medias.

Cálculo del Error estándar de la diferencia de medias

Su cálculo dependerá de sí:

1. Las dos poblaciones tienen una varianza común (varianzas iguales) o no.
2. Las dos muestras son del mismo tamaño o no.

A continuación presentaremos el cálculo del error estándar de la diferencia de medias para la situación en que las poblaciones tienen una **varianza común**, es decir que $\sigma^2_A = \sigma^2_B$

- **Caso 1** ($n_A \neq n_B$)

$$S_{\overline{x}_A - \overline{x}_B} = \sqrt{S_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}$$

donde: S_p^2 es un promedio ponderado de las varianzas muestrales, el cual es calculado como:

$$S_P^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{(n_A - 1) + (n_B - 1)}$$

Los **grados de libertad** (gl.) para este caso se calculan como:

$$\textbf{gl.} = (n_A + n_B) - 2$$

- **Caso 2** ($n_A = n_B = n$)

$$S_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{2 \times S^2}{n}}$$

donde: S^2 es el promedio aritmético de las varianzas muestrales, el cual es calculado como:

$$S^2 = \frac{S_A^2 + S_B^2}{2}$$

Los **grados de libertad** (gl.) para este caso se calculan como:

$$\textbf{gl.} = 2\,(n - 1)$$

Para la situación en que las poblaciones tienen **varianzas distintas**, es decir que $\sigma^2_A \neq \sigma^2_B$ la expresión apropiada para el calculo del error estándar de la diferencia de medias es:

$$S_{\bar{x}_A - \bar{x}_B} = \sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}$$

En este caso, de varianzas distintas, los grados de libertad (gl) no se pueden calcular directamente, como en la situación de varianzas iguales, en su lugar se utiliza una ponderación de los mismos, los que se denominan **grados de libertad efectivos (υ)**.

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}\right)^2}{\frac{(s_1^2 / n_1)^2}{n_1 + 1} + \frac{(s_2^2 / n_2)^2}{n_2 + 1}} - 2$$

Aplicaciones:

1) Intervalo de confianza para la media poblacional con variancia desconocida

Ejemplo:

En un monte de Álamos ubicado en el Delta del Río Paraná, se midieron los diámetros a la altura del pecho (DAP) de 20 árboles, obteniéndose los siguientes estadísticos: media muestral 23,8 cm. y desvió estándar muestral 4,5 cm. Asumiendo que la variable DAP presenta distribución aproximadamente normal, se desea obtener un Intervalo de confianza para la media poblacional (μ) con un coeficiente de confianza (γ) del 95 %.

Resolución:

La expresión de t a utilizarse es la 1 de donde se deduce que:

$$\left(\overline{X} - t\, S_{\bar{x}} \leq \mu \leq \overline{X} + t\, S_{\bar{x}}\right)$$

Constituyendo $\left(\overline{X} - t\, S_{\bar{x}}\right)$el limite inferior y $\left(\overline{X} + t\, S_{\bar{x}}\right)$el limite superior del intervalo de confianza para la media poblacional.

El valor de "t" se obtiene de tabla con n-1 grados de libertad y con el correspondiente nivel de significancia (γ) para una prueba bilateral.

Para este caso:

- gl =19
- α = 0,05 que sale de hacer 1 – γ

Por lo tanto el valor de t "tabulado" (**t_t**) será de 2.093

El error estándar de la media se obtiene aplicando $\frac{S}{\sqrt{n}}$, para este caso su valor es de 1.006.

Finalmente tenemos:

$$(23{,}8 - 2.093 \text{ x } 1.006 \leq \mu \leq 23{,}8 + 2.093 \text{ x } 1.006)$$

$$\mathbf{(21.69 \leq \mu \leq 25.9)}$$

que constituye el *intervalo de confianza para la media poblacional* por lo que se concluye, que el intervalo comprendido entre 21.69 y 25.9 contiene con un nivel de confianza del 95% a la media poblacional

Distribución de Chi-cuadrado (χ^2)

La distribución de Chi-cuadrado (o Ji cuadrada) se define como la suma de los cuadrados de variables aleatorias independientes, normalmente distribuidas con media (μ) igual a 0 y varianza (σ^2) igual a 1. Entonces si tenemos k variables (Z_1, Z_2,...,Z_k) con dichas características, la distribución de Chi-cuadrado tendrá k grados de libertad y quedará definida por:

$$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_k^2 = \sum_k Z_k^2$$

La ***media*** *y la* ***varianza*** de una distribución de Chi-cuadrado son los grados de libertad y dos veces los grados de libertad respectivaqmente, es decir:

$$\mu = k \qquad y \qquad \sigma^2 = 2k$$

Como vemos esta distribución esta caracterizada por un único ***parámetro***, los ***grados de libertad***, por lo tanto habrá una distribución de Chi-cuadrado para cada número de grados de libertad.

Por otra parte sea X_1, X_2, ... , X_k una muestra aleatoria de tamaño k de una población normal con media μ y varianza σ^2 y sean $\overline{X}$ y S^2 la media muestral y la varianza muestral respectivamente, se puede demostrar por un teorema que:

$$\chi^2 = \frac{\sum_{i=1}^{k}\left(x_i - \overline{x}\right)^2}{\sigma^2} = \frac{(n-1)\,S^2}{\sigma^2}$$ Con k =n-1 grados de libertad

Aplicación: *Estimación de intervalos de confianza para la varianza*

La esencia del método es la misma que la que se utiliza en las distribución normal y en la de t. Consiste en fijar un valor de probabilidad (coeficiente de confianza) para el cálculo de un intervalo que contendrá al parámetro que estamos estimando.

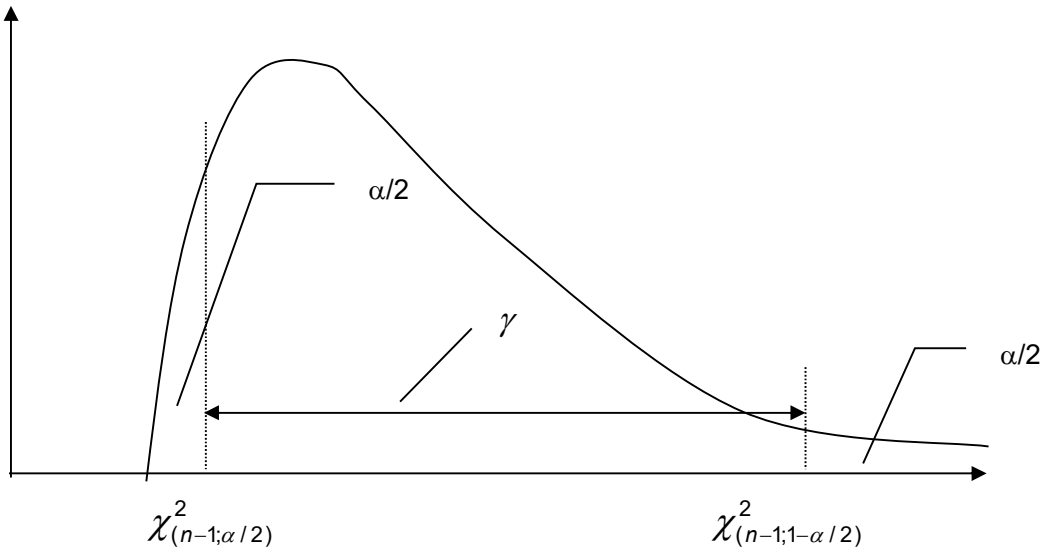
Para esto se utiliza el estadístico Chi-cuadrado definido como:

$$\chi^2 = \frac{(n-1)\,.\,S^2}{\sigma^2}$$

Por lo tanto si obtenemos una muestra de n observaciones que nos permita calcular S^2, podremos hallar un intervalo para σ^2, con una probabilidad γ de que este la contenga. Es decir:

$$P\left(Lim.\,Inferior \leq \sigma^2 \leq Lim.\,Superior \right) = \gamma$$

$$P\left[\frac{(n-1)\cdot S^2}{\chi^2_{(\alpha/2)}} \leq \sigma^2 \leq \frac{(n-1)\cdot S^2}{\chi^2_{(1-\alpha/2)}}\right] = \gamma$$



donde $\chi^2_{(\alpha/2)}$ y $\chi^2_{(1-\alpha/2)}$ son los valores de Chi-cuadrado tabulados con n –1 grados de libertad, dejando (o acumulando) áreas de $\alpha/2$ y 1- $\alpha/2$ respectivamente a la izquierda.

Se debe tener en cuenta que al no ser una distribución simétrica, los valores de Chi-cuadrado que limitan el intervalo deben ser calculados independientemente.

Para el cálculo de los limites de confianza del desvío estándar (σ) sólo hay que aplicarles las raíces cuadradas a los limites de confianza calculados para la varianza.

Ejemplo:

Dada una muestra de 10 observaciones provenientes de una población con distribución normal. ¿Calcular el intervalo para su varianza y su desvío estándar con un coeficiente de confianza (γ) del 90%?.

Las 10 Observaciones son: 5, 7, 11, 10, 8, 5, 6, 4, 2, y 9.

Resolución:

El primer paso consiste en hallar S^2 (varianza muestral) en este caso 8.011

Si queremos un intervalo del 0.9, implica que debemos dejar $\alpha/2$ a cada lado de la curva, en este caso un 5% del área. Como la tabla de Chi-cuadrado que presentamos en este escrito siempre indica el valor de probabilidad acumulada, debemos buscar el valor para el

5%, el cual utilizaremos para el límite inferior y el valor para el 95% que utilizaremos para el límite superior. En este caso y para 9 grados de libertad tenemos:

$$\chi^2_{(9)\,0.05} = 3.33 \qquad y \qquad \chi^2_{(9)\,0.95} = 16.9$$

por lo tanto

$$\left[\frac{(10-1)\cdot 8.01}{16.9} \leq \sigma^2 \leq \frac{(10-1)\cdot 8.01}{3.33} \right]$$

$$\left(4.26 \leq \sigma^2 \leq 21.68 \right)$$

Que constituye el *intervalo de confianza para la varianza*.

Si aplicamos las raíces cuadradas a esos valores, tendremos:

$$\left(2.06 \leq \sigma \leq 4.65 \right)$$

Que constituye el *intervalo de confianza para el desvío estándar*

PRUEBAS DE HIPOTESIS

Las ***pruebas de hipótesis*** nos permiten afirmar, con un cierto grado de probabilidad supuestos o declaraciones acerca de:

- características de los datos con los que se va a trabajar : independencia de las observaciones, homogeneidad de los datos, etc.
- la forma de la distribución de partida: binomial, poisson, normal, etc.
- de los parámetros de la distribución conocida su forma, etc.

Por ejemplo: que la ganancia promedio de peso en novillos con una determinada ración es significativamente superior a las raciones que se estaban utilizando, o que la producción promedio de tablas en un aserradero aumenta significativamente con la incorporación de una sierra sinfin adicional y que, por ejemplo, el costo promedio mensual de su funcionamiento no es significativamente diferente respecto a los costos antes de adquirir la sierra.

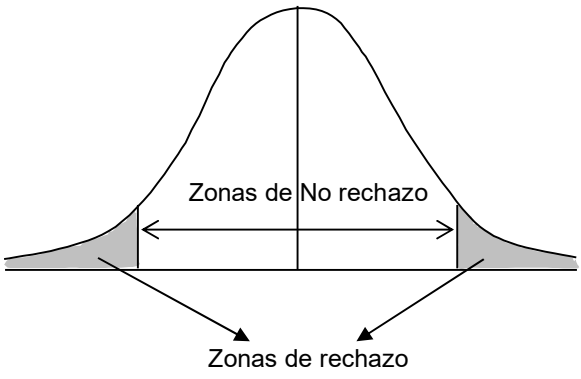
El propósito principal de la pruebas de hipótesis es hacer posible una elección adecuada entre dos hipótesis, denominadas ***Hipótesis nula*** (H_0) e ***Hipótesis alternativa*** (H_1).

La Hipótesis nula se plantea con intención de rechazarla, ya que fue planteada en forma opuesta a lo que se supone cierto. En caso de ser rechazada se asume la Hipótesis alternativa.

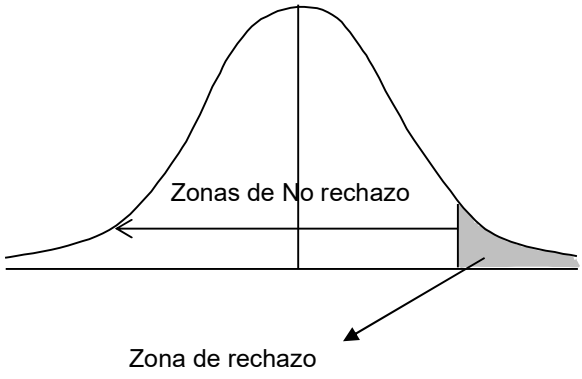
El procedimiento de contrastar dos hipótesis conlleva al uso de distribuciones ("t", "χ^2" (chi-cuadrado), " F ", etc.) para poder comprobar las suposiciones asumidas.

En las distribuciones muestrales se deben establecer dos zonas o regiones complementarias, llamadas ***Región de Rechazo*** y ***Región de No Rechazo*** de la hipótesis nula, cuyos valores quedarán definidos por el nivel de significancia (α) fijado y el tipo de prueba, bilateral o unilateral (*Ver figura*).

- Prueba bilateral



- Prueba unilateral a la derecha



Las ***pruebas bilaterales*** son llamadas de dos colas, se emplean cuando se

sospecha que el parámetro no es igual al valor postulado, siendo todos los demás valores posibles. Por otro lado las pruebas unilaterales, llamadas de una cola, se emplean, en problemas, en que se tiene algún indicio que el valor del parámetro es menor o mayor al postulado, en el primer caso hablariamos de ***pruebas unilaterales a la izquierda*** y en el segundo de ***pruebas unilaterales a la derecha***.

Los planteos de hipótesis según la prueba toman las siguientes formas:

- *Prueba Bilateral para una población*

$H_0 : \mu = k$
$H_1 : \mu \neq k$

- *Pruebas Unilaterales para una población*

$H_0 : \mu \leq k$ $H_0 : \mu \geq k$
$H_1 : \mu > k$ $H_1 : \mu < k$
Unilateral a la derecha Unilateral a la izquierda

- *Prueba Bilateral para dos poblaciones*

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

- *Pruebas Unilaterales para dos poblaciones*

$H_0 : \mu_1 \leq \mu_2$ $H_0 : \mu_1 \geq \mu_2$
$H_1 : \mu_1 > \mu_2$ $H_1 : \mu_1 < \mu_2$
Unilateral a la derecha Unilateral a la izquierda

En toda prueba de hipótesis es conveniente seguir los siguientes pasos:

1. Planteo de la hipótesis
2. Planificación del experimento o del esquema muestral conducente a obtener datos que permitan la validación o no de la hipótesis sometida a prueba.
3. Seleccionar el Estadístico de prueba adecuado
4. Establecer el nivel de significancia
5. Fijar la o las regiones de rechazo y no rechazo.
6. Realizar el ensayo o muestreo definido en el paso 2.
7. Calcular el valor del estadístico postulado y determinar si cae dentro o fuera de la región de rechazo.

Se mencionó más arriba que las pruebas de hipótesis nos permiten afirmar, con cierto grado de probabilidad que una cosa es igual, mayor, menor, diferente, superior, etc. que otra. Esta decisión que uno toma al poner a prueba una hipótesis está sujeta a dos tipos de errores. Por un lado podemos rechazar la H_0 cuando, en realidad, es verdadera, o bien aceptar H_0 cuando es falsa y alguna hipótesis alternativa es verdadera. Estos errores se llaman ***errores de TIPO I*** y ***de TIPO II***, respectivamente. Las dos opciones para la hipótesis nula (verdadera o falsa) junto con las dos decisiones que puede tener el experimentador, forman la tabla de decisiones.

- Tabla de decisiones

	Hipótesis Nula	
Decisión	**Verdadera**	**Falsa**
Rechazar H_0	*Error tipo I*	Decisión correcta
No Rechazar H_0	Decisión correcta	*Error tipo II*

Resumiendo:

α = P (Error Tipo I) = P (Rechazar H_0 / H_0 es verdadera)

β = P (Error Tipo II) = P (No Rechazar H_0 / H_0 es falsa)

Al valor complementario de β se le denomina ***potencia de la prueba*** y representa la capacidad o poder que tiene la prueba de reconocer correctamente que la H_0 es falsa.

Potencia de la prueba = 1 - β = P (Rechazar H_0 / H_0 es falsa)

Es obvio que, quien toma las decisiones, quiera reducir al máximo las probabilidades de cometer cualquiera de estos dos errores. La ***probabilidad de cometer el error tipo I*** se la simboliza con la letra griega α y su valor se fija de antemano y está de acuerdo con la importancia del problema, tomando usualmente valores entre un 5% y un 1%. En cambio la ***probabilidad de cometer el error tipo II*** se la simboliza con la letra griega β y su cálculo es bastante complejo dependiendo del valor del parámetro.

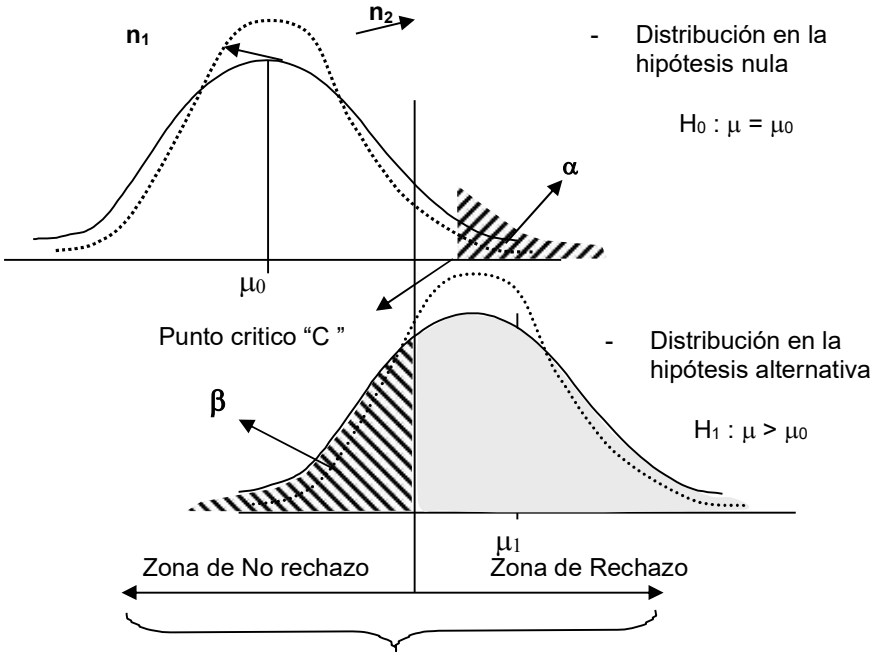
Una relación básica entre α y β para un tamaño de muestra fijo es el siguiente:

Si α aumenta $\rightarrow$ β disminuye ; Si α disminuye $\rightarrow$ β aumenta

Para lograr que ambos errores disminuyan debemos aumentar el tamaño de la muestra. Las gráficas siguientes tratarán de clarificar los conceptos vertidos sobre ambos errores.

- *Figura a*

- *Figura b*



- Distribución en la hipótesis nula

$H_0 : \mu = \mu_0$

- Distribución en la hipótesis alternativa

$H_1 : \mu > \mu_0$

Supongamos que tomamos dos muestras aleatorias de tamaño n_1 y n_2 , siendo $n_2 > n_1$, de una población con el propósito de probar la hipótesis:

H_0 $\mu = \mu_0$ H_1 $\mu > \mu_0$ (por ejemplo μ_1)

La distribución de la media muestral bajo el supuesto de la H_0 se ilustra en la ***figura a*** en la cual se debe localizar la zona de rechazo y no rechazo de la H_0 en base al nivel de significancia (α) fijado, lo que nos permite determinar el valor critico " C " tomando H_0 como verdadera. En caso de que la H_0 es falsa, la distribución de la media muestral tendrá como media poblacional a μ_1 en lugar de μ_0 , siendo $\mu_1 > \mu_0$, véase ***figura b***.

Si luego de realizada la prueba de hipótesis concluimos que no se rechaza la H_0 (para todo valor < a C) siendo μ_0 falso estaremos cometiendo el error de tipo II con una probabilidad β. En cambio si rechazamos la H_0 (para todo valor > a C) siendo μ_0 verdadero estaremos cometiendo el error tipo I con una probabilidad α. El poder de la prueba 1 - β queda determinado por el área sombreada bajo la curva de la ***figura b***.

De las ***figuras a*** y ***b*** observamos que:

- a medida que aumentamos el tamaño de la muestra la variabilidad disminuye por ende ambos errores disminuyen.
- para un tamaño de muestra fijo el valor de β dependerá de la distancia entre μ_1 y μ_0 .
- a medida que el valor de μ_1 se acerca a μ_0 la probabilidad de cometer el error tipo II aumenta.
- a medida que μ_1 se aleja de μ_0, se observa que la probabilidad de cometer el error tipo II disminuye, aumentando por otro lado el poder de la prueba es decir la probabilidad de rechazar la H_0 siendo esta falsa.

Contraste de hipótesis para la media muestral de una determinada población
(sin conocer la variancia poblacional)

Ejemplo:

Se analiza el contenido de vitamina C en 8 frutos de tomate y se hallan los siguientes contenidos en mg.:

250 270 265 230 275 245 268 272

De donde se calcularon la media muestral 259.4 mg y el error estándar de la media 5.637 mg.

Se desea probar la hipótesis de que estos frutos de tomates provienen en realidad de una población de frutos con una media de 300 mg de vitamina C.

Resolución:

Planteo de hipótesis (test bilateral)

$H_0 : \mu = 300$ mg.
$H_1 : \mu \neq 300$ mg.

Cálculo del valor de "t"

Para este caso aplicamos:

$$t = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \frac{259.4 - 300}{5.637} = -7.20$$

Obtención del valor de "t" tabulado

Debemos definir el nivel de significancia y los grados de libertad de la prueba, para entrar en la tabla de "t" y extraer su valor (tener en cuenta que se trata de un test bilateral).

- gl = n –1 = 8 –1 = 7
- nivel de significancia (α) = 0.05

- **"t" tabulado = 2.365**

Conclusiones

Como el valor del t calculado supera al valor del t tabulado para un nivel de significancia del 0,05; la hipótesis nula es rechazada, es decir que existe evidencia estadísticamente significativa para suponer que estos frutos no provienen de una población con media de 300 mg. de vitamina C.

Contraste de hipótesis para la diferencia entre medias de dos tratamientos aplicados a observaciones independientes (o muestras no pareadas)

Ejemplo:

Se desea comparar la ganancia de peso de novillos Aberdeen Angus sometidos a dos dietas diferentes. Un grupo de novillos fue alimentado con la ración A y otro grupo con la ración B, ambos grupos fueron seleccionados al azar. Los aumentos de peso en gramos fueron de:

Ración A	175	218	200	234	187	248	179	132	151	219	149	123	206	206
Ración B	142	337	302	253	236	211	249	311	262	195	199	216	176	214

Siendo las ganancias medias y la varianza de cada ración las siguientes:

	Media muestral	Varianza muestral
Ración A	187,64 grs.	1451,48 grs.^{2}
Ración B	235,93 grs.	2946,99 grs.^{2}

Se desea probar que el aumento de peso generado por las raciones es diferente. Considerar las varianzas poblacionales de donde provienen las muestras como iguales.

Resolución:

Planteo de hipótesis (test bilateral)

$H_0 : \mu_A = \mu_B$

$H_1 : \mu_A \neq \mu_B$

Cálculo del valor de "t"

Para este caso aplicamos:

$$t = \frac{\bar{x}_A - \bar{x}_B}{S_{\bar{x}_A - \bar{x}_B}}$$

$$t = \frac{187,64 - 235,93}{17,72} = -2,725$$

$$S_P^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{(n_A - 1) + (n_B - 1)} = 2.199,2 \qquad \text{y} \qquad S_{\bar{x}_A - \bar{x}_B} = \sqrt{S_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} = 17.72$$

Obtención del valor de "t" tabulado

Debemos definir el nivel de significancia y los grados de libertad de la prueba, para entrar en la tabla de "t" y extraer su valor (tener en cuenta que se trata de un test bilateral).

- **gl**= $(n_A + n_B) - 2 = (14 + 14) - 2 = 26$
- nivel de significancia (α) = 0.05

- **"t" tabulado = 2.056**

Conclusiones

Como el valor del t calculado supera al valor del t tabulado para un nivel de significancia del 0,05; la hipótesis nula es rechazada, es decir que existe evidencia estadísticamente significativa para suponer que la ganancia de peso difiere de una ración a otra.

Contraste de hipótesis para la diferencia entre medias de dos tratamientos aplicados a observaciones pareadas

La distribución de t también se la puede utilizar cuando trabajamos con unidades experimentales pareadas, ya sean plantas, vacas, maquinarias. Cada par estará constituido por dos miembros similares en la característica en que se va a medir los efectos de los tratamientos; hecho esto, a uno de los miembros de cada par, al azar, se le asigna uno de los tratamientos y al otro miembro el otro tratamiento. Por ejemplo: consideremos un experimento de racionamiento de cerdos, donde queremos probar dos raciones alimenticias. Antes de asignarle las raciones clasificamos a los cerdos por edad, peso, etc.,de manera de poder parear individuos similares. Es decir que el pareamiento se ha hecho antes de comenzar el experimento con base en respuestas similares cuando no hay efectos de tratamiento.

Si los pares de individuos covarían, puede aumentarse la capacidad del experimento para detectar una pequeña diferencia. La información sobre el pareamiento se usa para eliminar una fuente de variación extraña, que es la que hay de un par a otro.

En este caso la expresión de t es la siguiente:

$$t = \frac{\overline{D} - 0}{S_{\overline{D}}}$$

Como vemos en la expresión en el caso de observaciones pareadas, la diferencia de medías $(\overline{x}_A - \overline{x}_B)$ se sustituye por $\overline{D}$, que es la media aritmética de las diferencias entre pares, y, por ello, el estimador del error estándar de esta nueva diferencia se calcula con la siguiente expresión:

$$S_{\overline{D}} = \frac{S}{\sqrt{n}}$$ donde **S** es: $$S = \sqrt{\frac{\sum_{i=1}^{n} D_i^2 - \frac{\left(\sum_{i=1}^{n} D_i\right)^2}{n}}{n-1}}$$

En este caso los **grados de libertad** se calculan como:

gl = n – 1

siendo **n** el número de pares involucrados.

Ejemplo:

Se trata de un experimento para comparar el efecto de dos raciones en cerdos. Previamente a la asignación de las raciones los cerdos fueron clasificados por peso, edad, etc.; lo que permitió parear cerdos similares. Luego a cada integrante de un par, se le asigno una ración. Los datos fueron los siguientes:

Pares	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ración A	26	25	12	25	20	16	18	21	11	8
Ración B	23	22	16	29	24	15	24	25	16	14
D_i	3	3	-4	-4	-4	1	-6	-4	-5	-6

Siendo $\overline{D}$ de –2,6 kg. y el error estándar de la diferencia entre pares ($S_{\overline{D}}$) es de 1,117 kg.

Se desea probar que no hay diferencias en el aumento de peso entre ambas raciones contra la alternativa que la ración B produce un aumento mayor.

Planteo de hipótesis (test unilateral)

$$H_0 : \overline{D}_i = 0$$
$$H_1 : \overline{D}_i < 0$$

Cálculo del valor de “t”

Para este caso aplicamos:

$$t = \frac{\overline{D} - 0}{S_{\overline{D}}} \qquad t = \frac{-2.6}{1{,}117} = -2{,}326$$

Obtención del valor de “t” tabulado

Debemos definir el nivel de significancia y los grados de libertad de la prueba, para entrar en la tabla de “t” y extraer su valor (tener en cuenta que se trata de un test unilateral).

- **gl**= n – 1 = 10 – 1 = 9
- nivel de significancia(α) = 0.05

- **“t” tabulado = 1.833**

Conclusiones

Como el valor del t calculado supera al valor del t tabulado para un nivel de significancia del 0,05; la hipótesis nula es rechazada, es decir que existe evidencia estadísticamente significativa para suponer que la ganancia de peso de la ración B es mayor que la de la A.

Contraste de hipótesis acerca de una variancia muestral población

Ho: $\sigma^2 = \sigma_0^2$	Ho: $\sigma^2 = \sigma_0^2$	Ho: $\sigma^2 = \sigma_0^2$
H1: $\sigma^2 \neq \sigma_0^2$	H1: $\sigma^2 > \sigma_0^2$	H1: $\sigma^2 < \sigma_0^2$

$$\chi^2_{obs} = \frac{(n-1)\cdot S^2_{obs}}{\sigma_o{}^2}$$

Contraste de hipótesis para la igualdad de dos variancias

Esta prueba se realiza mediante el test F que compara dos variancias a los efectos de comprobar si las muestras en estudio pertenecen a una misma población de variancia σ^2 .

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \Rightarrow \qquad H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Esta comparación la realiza por la expresión:

$$F_0 = \frac{s_1^2}{s_2^2}$$

Este valor calculado se compara con un valor teórico de F para diferentes niveles de probabilidad 5 o 1%.

La hipótesis nula será rechazada o no en la medida de que el valor de F calculado supere o no los valores de F teóricos de la tabla para posniveles de significancia elegidos.

PRUEBAS NO PARAMETRICAS O CHI-CUADRADO

La distribución de Chi-cuadrado puede utilizarse también cuando tratamos con variables expresadas en frecuencias (Nominales, Discretas o Continuas) y queremos probar discrepancias entre frecuencias observadas y frecuencias esperadas o calculadas.

$$\chi^2 = \sum \frac{(observados - calculados)^2}{calculados}$$

A. PRUEBAS DE AJUSTE DE MODELOS GENÉTICOS (distribución de proporciones)

Muy utilizadas en estudios genéticos, *donde se desea verificar si los resultados obtenidos en una segregación corresponden a la hipótesis planteada*. En este caso se utiliza la siguiente expresión para hallar un χ^2 calculado, que luego se contrasta contra el obtenido de tabla (tabulado) según los grados de libertad y el coeficiente de confianza deseado, para luego concluir si se acepta o rechaza la hipótesis planteada.

Es decir se define a χ^2 como la sumatoria de los cocientes resultantes de dividir por las respectivas frecuencias calculadas o esperadas (Fe) a cada cuadrado de discrepancia entre las frecuencias observadas (Fo) y esperadas (Fe).

Esto implica la definición de una hipótesis previa que nos permitirá calcular valores teóricos y compararlos con los observados en la muestra.

Para el caso de un modelo genético (distribución de proporciones) el planteo de las hipótesis nulas que se ensayan podrían ser:

Ho: 'la segregación de cierto poroto responde a la teoría de Mendel 9:3:3:1'
(Lisos Amarillos – Rugosos Amarilos – Lisos Verdes – Rugosos Verdes)

Ho: 'la relación de moscas ojos blancos respecto a moscas ojos rojos es 1 a 3'

Ho: 'la proporción de hembras y machos es la misma en una determinada población de animales silvestres'

Si las Fo tienden a ser similares a las Fe entonces χ^2 tomará un valor bajo, es decir tenderá a cero y la Ho no será rechazada. A medida que aumentan las discrepancias entre las Fo y las Fe, aumenta χ^2 y cuando este supera el valor del Chi-cuadrado tabulado correspondiente al nivel de significancia (α) establecido la Ho será rechazada.

Ejemplo:

En ciertos casos de herencia del color de la aleurona del maíz interviene sólo un par de genes:

P = Aleurona de color púrpura.
p = Aleurona de color blanco.

Cuando se cruza una planta de maíz de aleurona púrpura con otra de aleurona blanca, todas las plantas de la primera generación tienen granos con aleurona púrpura. Si se autofecunda la descendencia, deben aparecer en la segunda generación, de acuerdo a las leyes de la herencia, plantas con granos púrpuras y blancos en la proporción 3 a 1 (3:1).

En una experiencia se contaron en la segunda generación, 1654 granos de aleurona púrpura y 466 con aleurona blanca, totalizando 2120 granos.

Compruebe si existe concordancia de los valores observados con las leyes de herencia .

Resolución:

1.- Planteo de hipótesis:

H_0: *La proporción de maíces con aleurona color púrpura respecto a los que tienen aleurona color blanco es 3 a 1*
H_1: *La proporción de maíces con aleurona color púrpura respecto a los que tienen aleurona color blanco es distinta de 3 a 1*

2.- Determinación de χ^2 calculado

Para ello debemos construir una tabla que nos ayudará a su calculo:

Clases	Observados	Calculados	(Obs-Cal)	$(Obs-Cal)^2 / Cal$
P	1654	1590	64	2.576
p	466	530	- 64	7.728
Totales	2120	2120	0	$\chi^2 = 10.3$

Los valores de las frecuencias calculadas se determinan aplicando una regla de tres simple, teniendo presente que, de acuerdo con la hipótesis, por cada 4 granos, 3 son de aleurona púrpura, por lo que en un total de 2120 granos tenemos:

P (granos aleurona púrpura)= 2120 x 3/4= **1590**

El otro valor sale por diferencia entre el total y el calculado para P:

p (granos aleurona blanca)= 2120 - P= 2120 – 1590 = **530**

3.- Obtención del valor de χ^2 tabulado

Para la obtención del χ^2 tabulado necesitamos saber los grados de libertad apropiados para la prueba y el nivel de significancia (α)

- Grados de libertad (gl) = número de clases – 1= 2 – 1 = **1**
- Nivel de significancia (α) = **0.05**

- De la tabla obtenemos un χ^2 **tabulado = 3,84**

4.- Conclusión:

Como el valor de χ^2 calculado resulta mayor que el tabulado para el 0.05, la hipótesis nula se rechaza, es decir que la probabilidad (p) de que Ho sea cierta es menor que α o sea (p<0,05). Por lo tanto el resultado de este ensayo nos permite admitir que las desviaciones respecto a la proporción 3:1 son altamente significativas.

B. PRUEBAS DE INDEPENDENCIA DE FACTORES

Con frecuencia los individuos se clasifican de acuerdo a varias variables. Por ejemplo una persona puede clasificarse como fumadora o no fumadora y, al mismo tiempo, como un individuo con o sin enfermedad coronaria; una vaca puede ser clasificada como vacunada o no vacunada y también como con o sin brucelosis.

Este tipo de datos se suelen registrar en una tabla de doble entrada, las que se denominan **tablas de contingencia**.

		Vía Columna		**Total**
		A	**B**	
Vía Fila	**1**	A1	B1	**1**
	2	A2	B2	**2**
Total		**A**	**B**	**N**

Descripción de la tabla de contingencia:

Una tabla de contingencia de dos vías (dos criterios de clasificación) consta de:

Total marginal de fila: Es la suma según la fila. (1 y 2)
Total marginal de columna: Idem, según la columna. (A y B)
Dato: La observación correspondiente a cada una de las Celdas. Ej.: (A1, B1, etc.)
Total: La suma de los totales marginales de fila y de columna. (N)

La *hipótesis nula* que se ensaya en estas pruebas siempre plantea *la independencia de los criterios de clasificación*, lo que implicaría que las frecuencias observadas son iguales a las frecuencias esperadas.

Por lo tanto, la probabilidad de que un individuo seleccionado al azar pertenezca a la celda A1 bajo el supuesto de que la hipótesis nula es cierta (independencia de los criterios de clasificación) será:

$$P(A1) = P(A) \times P(1)$$

Pero como la prueba de Chi-cuadrado compara frecuencias y no probabilidades, podemos indicar, como regla general para calcular la frecuencia esperada o calculada correspondiente a una celda de una tabla de contingencia, la siguiente expresión:

$\frac{\textit{Total de fila x Total de columna}}{\textit{Total general}}$, para la celda A1 tendríamos: $\frac{A \times 1}{N}$

Ejemplo:

"La tabla de contingencia de 2 x 2 que se muestra a continuación, clasifica 708 vacas según sean de raza Aberdeen Angus (1) o Hereford (2) con pariciones normales (A) o con abortos (B). Se desea saber si hay asociación entre los criterios de clasificación, es decir si los abortos son más frecuentes en una de las razas (hay asociación o dependencia) o sí la frecuencia de aborto no tiene relación con la raza (son independientes).

	Pariciones		
Razas	Normales (A)	Abortos (B)	Total marginal fila
Aberdeen Angus (1)	123	35	**158**
Hereford (2)	470	80	**550**
Total marginal columna	**593**	**115**	**708**

1.- Planteo de hipótesis

Como las frecuencias esperadas para la situación de independencia pueden calcularse a partir de los valores marginales, el planteo es:

H_0: Que la parición sea normal o con aborto es independiente de la raza del animal
H_1: Que la parición sea normal o con aborto depende de la raza del animal

2.- Cálculo del valor de χ^2 calculado

Las frecuencias esperadas o calculadas se obtienen para cada clase. Por ejemplo para la clase A1 (Aberdeen Angus y Pariciones Normales) se tiene la siguiente frecuencia calculada:

$$(A1)_{cal} = \frac{(A).(1)}{N} = \frac{593 \times 158}{708} = 132.34$$

Donde (A1) es la frecuencia calculada suponiendo cierta la hipótesis.
(A) es el total marginal de la columna A.
(1) es el total marginal de la fila 1.
N es el total de animales observados.

Clases	Observados	Calculados	(Obs-Cal)	(Obs-Cal) ^{2} / C
A1	123	132.34	-9.34	0.659
A2	470	460.66	9.34	0.189
B1	35	25.66	9.34	3.4
B2	80	89.34	-9.34	0.976
Totales	**708**	**708**	**0**	**χ^2 = 5.22**

3.- Obtención del valor de χ^2 tabulado

Los grados de libertad: Se puede ver que basta calcular la frecuencia esperada de una celda y las demás frecuencias se obtienen por diferencia. En consecuencia, para la prueba de independencia de 2 x 2 como la estudiada, existe sólo un grado de libertad. En general para tablas de contingencia, los grados de libertad se calculan como:

$$gl = (f-1).(c-1) = \mathbf{1}$$

donde: f es el número de filas y c el de columnas en la tabla de contingencia.

- Nivel de significancia = 0.05

- De la tabla (pagina 72) obtenemos un χ^2 **tabulado = 3,84**

4.- Conclusión:

Como el valor de χ^2 calculado resulta mayor que el tabulado para el 0.05, en consecuencia la hipótesis nula se rechaza. Es decir que el resultado de este ensayo nos permite admitir que las razas y los tipos de parición no son criterios de clasificación independiente. Al analizar las frecuencias observadas de la tabla se aprecia que los abortos en Aberdeen Angus (A.A.) (35) son más frecuentes que lo esperado (25,66) y lo contrario en Hereford (H). Los abortos son más frecuentes en A.A. que en H.

C. PRUEBAS DE AJUSTE DE MODELOS DE DISTRIBUCION DE PROBABILIDAD

(Pruebas de Bondad de Ajuste)

Existen ocasiones en las cuales el investigador debe estudiar y encontrar si la variable que esta estudiando, sigue algún modelo de distribución conocido. Es decir debe ajustar su variable a una distribución y comprobar si esta sigue el patrón de la misma.

El investigador cuenta con los datos que resultaron de sus mediciones o conteos, que corresponden a las frecuencias observadas de cada clase. Si a su vez calcula cuales deberían haber sido las frecuencias teóricas (esperadas o calculadas) si dicha variable siguiera al patrón de distribución del modelo al que se intenta asociarla, podría utilizar el método de Chi-cuadrado para estimar si las desviaciones de lo observado respecto a lo teórico resultan o no significativas según un valor de probabilidad preestablecido.

Un punto que debe tomarse en cuenta en la aplicación de este procedimiento de prueba se refiere a la magnitud de las frecuencias esperadas. Si estas son demasiado pequeñas, entonces χ^2 no reflejará la diferencia entre lo esperado y lo observado, sin solamente la pequeñez de las frecuencias esperadas. Por lo general se considera que las frecuencias esperadas no deberían ser inferiores a 4 o 5. Si este es el caso se pueden agrupar clases adyacentes dado que no es necesario que las clases tengan el mismo tamaño.

Los *grados de libertad* se calculan como el número de clases menos uno (1) menos el número de parámetros estimados para el cálculo de las frecuencias esperadas, número que depende de la distribución a la que estamos ajustando. Por lo tanto:

Distribución	Parámetros	Grados de libertad
Binomial	p	gl. : nº de clases –1 – 1
Poisson	λ	
Normal	μ y σ^2	gl. : nº de clases – 1 – 2

Ejemplo:

Se realizaron muestreos de la sangre de animales domésticos y se contó el número de virus por cm^3. Suponemos que el número de virus por cm^3 se distribuye según el modelo de Poisson. Las frecuencias esperadas (o calculadas) fueron obtenidas de la forma conocida (Práctica Nº 3)

Clases	F Obs.	F Cal.
0	45	40.23
1	23	27.66
2	6	9.51
3	4	2.2
4	2	0.40
Totales	**80**	**80**

1.- Planteo de hipótesis

H_0: la variable número de virus por cm^3 sigue una distribución de Poisson.

2.- Cálculo del valor de χ^2 calculado

Debido a que las últimas clases tienen una frecuencia menor a 3, se las agrupa en una sola clase de la siguiente forma

Clases	**F Obs.**	**F cal.**	**O - C**	**$(O - C)^2$**	**$(O - C)^2/C$**
0	45	40.23	4.77	22.75	0.565
1	23	27.66	-4.66	21.72	0.785
2	6	9.51	-3.51	12.32	1.295
≥3	6	2.60	3.4	11.56	4.446
Totales	**80**	**80**	**0**		**$\chi^2 = 7.09$**

3.- Cálculo del valor de χ^2 tabulado

En este caso, para la distribución de Poisson se necesita estimar un parámetro (lambda) para el cálculo de las frecuencias esperadas, por lo que los grados de libertad se calculan como:

- g.l = 4 clases - 1 parámetro -1 = 2
- Nivel de significancia = 0.05

- De la tabla (pagina 72) obtenemos un **χ^2 tabulado = 5,99**

4.- Conclusión:

Como el valor de χ^2 calculado resulta mayor que el tabulado para el 0.05, en consecuencia la hipótesis nula se rechaza. Es decir que el resultado de este ensayo nos permite admitir que el número de virus por cm^3 en sangre de animales no se distribuye según el modelo de Poisson.

ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

En la aplicación de los métodos estadísticos estudiados hasta el momento, se ha tratado una única variable de interés. A estas variables se le determinaron y examinaron varias medidas que describen su comportamiento, además se aplicaron diversas técnicas de inferencia estadística, como intervalos de confianza y pruebas de hipótesis, para hacer estimaciones y sacar conclusiones acerca de ellas. En esta sección se tratarán los métodos estadísticos que han de aplicarse en situaciones donde se observan dos o más variables cuantitativas sobre cada unidad experimental, con el fin de establecer y medir las relaciones existentes entre ellas.

En numerosos trabajos encontramos que existe relación entre las variables en estudio y, generalmente es importe conocer la relación funcional como así también la intensidad de asociación entre ellas.

Por ejemplo, en un cultivo sería de relevancia el estudio de la relación entre el rendimiento y las dosis de un fertilizante nitrogenado o entre el rendimiento y la densidad de siembra, la relación entre la dosis de un insecticida y la mortalidad de los insectos tratados, el efecto de la irradiación sobre la tasa de fotosíntesis, etc.

La estadística para dar respuesta a estas problemáticas cuenta con dos técnicas: el *Análisis de Regresión* y el *Análisis de Correlación*.

El Análisis de Regresión estudia la relación funcional que existe entre dos o más variables. Identifica el modelo o función que liga a las variables, estima sus parámetros y, generalmente prueba hipótesis acerca de ellos. Una vez estimado el modelo es posible predecir valores de la variable denominada dependiente en función de la o las variables independientes que conforman el modelo. Para que este análisis sea aplicable es esencial que la variable dependiente sea aleatoria mientras que la o las independientes sean fijadas por el investigado.

El Análisis de Correlación estudia la intensidad y el sentido de la asociación que hay entre un conjunto de variables, tomadas dos a dos y, a diferencia del análisis de regresión, no se identifica ni se estima un modelo funcional para las variables, ya que este siempre se supone lineal, ni hay necesidad de distinguir entre variables dependientes e independientes, ya que ningunas de la variables puede ser fijada por el investigador. Es decir ambas deben ser aleatorias.

Análisis de Regresión

En un modelo de regresión se supone que una variable independiente fija X tienen una influencia esencial en el comportamiento de otra variable Y, aleatoria, y además existe un conjunto de factores, aleatorios no observables, individualmente de poca influencia, que englobamos dentro del término denominado residuo aleatorio o error y que denotaremos por ε. De tal manera la hipótesis estructural básica del modelo de regresión simple es:

$$Y = f(x) + \varepsilon$$

Y: Variable dependiente (Aleatoria).
X: Variable independiente (Fija).
f (x) : Función de regresión. Modelo matemático a determinar
ε : Residuo o error aleatorio.

La función f(x) es la que hipotéticamente define cómo se comporta la esperanza matemática o valor medio teórico de Y en función de X. Esta función suele ser desconocida, y su estimación se realiza a partir de los datos disponibles y basándonos en la suposición que conocemos la familia a la que pertenece (esta suposición proviene de la forma de la nube de puntos).

En un análisis de regresión, que incluye una variable dependiente Y y una variable independiente X, los pares de valores individuales se representan por un punto en un sistema de coordenadas cartesianas ortogonales.

Se llama diagrama de dispersión o diagrama de puntos al gráfico bidimensional que se forma con todos los puntos que representan a los pares de valores $(x_i;y_i)$ correspondientes a las variables independientes y dependientes respectivamente.

Por ejemplo, si la forma de la nube lo sugiere podemos suponer que la función es lineal. Pero puede ocurrir que sea una función exponencial o de otro tipo la que sugiere la nube de puntos. Pero aún cuando la suposición de linealidad no es aparentemente muy acertada, pero se pronostica que la función es monótona, es posible efectuando las transformaciones oportunas con los datos, trasladar el problema a la búsqueda de una función lineal. Por ejemplo, si se aprecia una nube de puntos (x,y) de apariencia exponencial, se apreciará una nube de puntos (x, ln y) de apariencia lineal. De estos comentarios se desprende la importancia del modelo lineal en estos problemas y de ahí que la suposición básica que se establece es la hipótesis de linealidad, bien a partir de los datos originales o bien a partir de datos transformados convenientemente.

Cuando en un análisis de regresión participan más de una variable independiente (fija), el mismo se denomina **"múltiple"**, si participa sólo una variable independiente se denomina **"simple"** y si suponemos que la relación es una línea recta la regresión es **lineal simple**.

1) Análisis de Regresión Lineal Simple

El Análisis de Regresión Lineal Simple se utiliza para la estimación del valor de la variable dependiente Y a partir del valor de una variable independiente X, cuando el comportamiento de Y está explicada por una función lineal o sea, por una recta.

Suponiendo que la verdadera relación X e Y es una línea recta y que a cada valor de X le corresponde un valor aleatorio de Y, entonces el modelo de regresión lineal simple toma la siguiente forma:

$$E\ (Y / X\ ) = Y\ = \beta_0 + \beta_1 . X$$

Siendo:

- **E (Y/X)** el valor teórico (esperado) de Y para un valor dado de X.
- β_0 y β_1 Parámetros desconocidos los cuales se deben estimar.
- β_0 el **intercepto** u **ordenada al origen** (valor que asume Y cuando la recta corta a la ordenada)
- β_1 el **coeficiente de regresión** o **pendiente de la recta**.

Si en la expresión anterior reemplazamos los parámetros β_0 y β_1 por sus estimadores **a** y **b** obtenemos el modelo estimado de regresión lineal simple:

$$\hat{Y}_i = a + b\ X_i + e_i$$

Donde:

- $\hat{Y}_i$ valor estimado de Y_i.
- **a** el estimador de β_0
- **b** el estimador de β_1
- $\mathbf{e_i}$ residuo o error aleatorio con distribución N (0, σ^2) y Cov (e_i, e_j) = 0.

Para poder estimar β_0 y β_1 se necesitan tener n pares de valores (x_1, y_1) ; (x_2, y_2) ; ..... (x_n, y_n) siendo el procedimiento usado para estimar puntualmente β_0 **y** β_1 la *teoría de los cuadrados mínimos*, que nos permite determinar las expresiones para **a** y **b**, estimadores puntuales de β_0 **y** β_1 respectivamente. Este método consiste en calcular los valores **a** y **b** de modo tal que se minimice la suma del cuadrado de los residuos, es decir que los valores observados Y_i presenten mínima desviación respecto de los valores estimados $\hat{Y}_i$ a partir de la recta. Esa desviación mínima significa que :

$$\mathbf{S.C._{RES.}} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} e_i^2 : [\text{mín}]$$

o la expresión equivalente

$$\mathbf{S.C._{RES.}} = \sum_{i=1}^{n}(Y_i - a - b\,X_i)^2 : [\text{mín}]$$

Haciendo las derivadas parciales de **S.C.$_{RES.}$** con respecto a los estimadores ***a*** y ***b*** respectivamente, e igualándolas a cero, se obtiene el siguiente sistema de ecuaciones llamado **SISTEMA DE ECUACIONES NORMALES**.

$$\begin{cases} a \cdot n + b \cdot \sum X_i = \sum Y_i \\ a \cdot \sum X_i + b \cdot \sum X_i^2 = \sum X_i Y_i \end{cases}$$

Resolviendo este sistema se obtienen los estimadores **a** y **b** respectivamente:

$$a = \frac{\sum_{i=1}^{n} Y_i - b\sum_{i=1}^{n} X_i}{n} = \overline{Y} - b\ \overline{X}$$

y :

$$b_{y/x} = \frac{\sum_{i=1}^{n} \delta_x \ \delta_y}{\sum_{i=1}^{n} \delta_x^2} = \frac{\sum_{i=1}^{n} X\,Y - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}} = \frac{\sum_{i=1}^{n} \left[\left(X - \overline{X}\right).\left(Y - \overline{Y}\right)\right]}{\sum_{i=1}^{n} \left(X - \overline{X}\right)^2} = \frac{Cov\left(X,Y\right)}{S_x^2}$$

A fin de poder determinar qué tan bien explica el comportamiento de la variable dependiente Y, el comportamiento de la variable independiente X, se necesita desarrollar medidas de variación que descompongan la variabilidad total de la variable Y en la variabilidad explicada por el modelo de la regresión y la no explicada o variabilidad residual, entre ellas tenemos: (Ver gráfico).

Suma de cuadrado total (SCT)

Se define como la sumatoria de la distancia al cuadrado que hay entre cada valor Y_i de una observación y la media aritmética $\overline{\mathbf{Y}}$.

$$SCT = \sum \left(Y_i - \overline{Y}\right)^2 = SCExp. + SC\,\mathrm{Re}\,s.$$

Suma de cuadrado explicada por el modelo (SCExp.)

Se interpreta como la sumatoria de la distancia al cuadrado que hay entre cada valor estimado de Y_i a partir de la recta de regresión ($\hat{Y}_i$) y la media aritmética $\overline{\mathbf{Y}}$.

$$SCExp. = \sum \left(\hat{Y}_i - \overline{Y}\right)^2$$

Suma de cuadrados residual o no explicada por el modelo (SCRes.)

Se define como la sumatoria de la distancia al cuadrado que hay entre cada valor Y_i de una observación y el valor estimado de Y_i a partir de la recta de regresión ($\hat{Y}_i$)

$$SC\,\mathrm{Re}\,s. = \sum \left(Y_i - \hat{Y}_i\right)^2$$

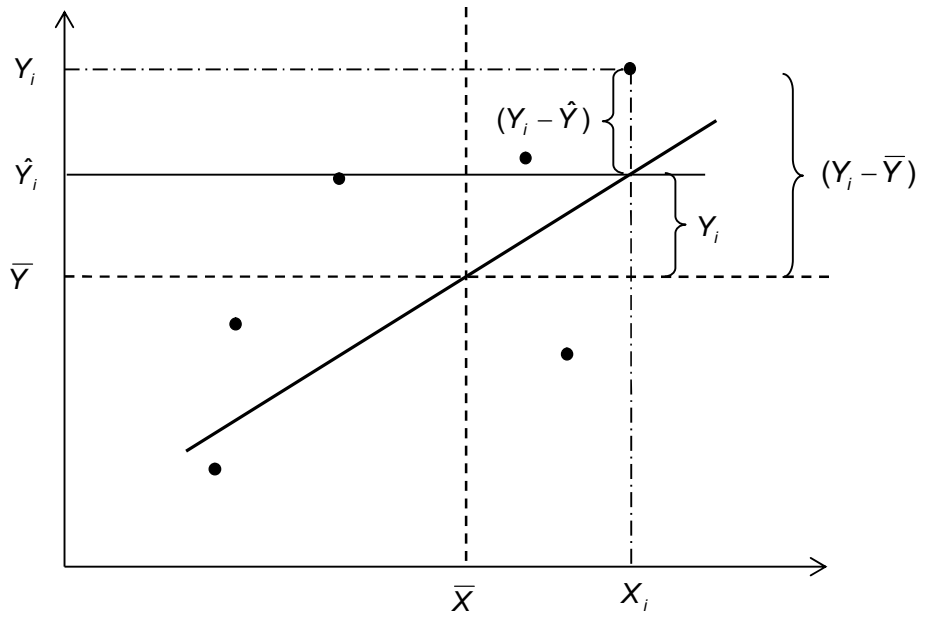
A partir de la relación entre la Suma de cuadrado explicada por el modelo estimado (SCExp.) y la Suma de cuadrado total (SCT) se obtiene un coeficiente que se denomina *Coeficiente de Determinación* ($\mathbf{R^2}$), el cual mide la proporción de la variación total explicada por la regresión. Es decir indica la proporción de la varianza de Y que es explicada por la variable X. También se lo suele interpretar como el grado de ajuste de los datos al modelo, a mayor R^2 mayor ajuste de los datos al modelo estimado.

$$\mathbf{R^2} = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2} = \frac{SCExp.}{SCT}$$

Este coeficiente puede tomar valores entre cero y uno, su valor se acerca a 0 cuando no existe relación entre las variables o bien la misma es muy debil y tiende a 1 a medida que aumenta la intensidad de la asociación.

La ***significancia de la regresión*** se obtiene mediante un análisis de varianza que pone a prueba la hipótesis nula de que no existe una relación lineal entre las variables, es decir que la varianza explicada de la variable dependiente por el modelo de regresión lineal no es significativamente mayor a la no explicada (varianza residual) es decir:

$$H_0 : R^2 = 0$$
$$H_1 : R^2 \neq 0$$



Donde:

- ● Pares de observaciones (Y_i ,X_i)
- —— Recta de regresión ajustada a los pares de valores (Y_i ,X_i)

La ***Tabla de Análisis de la Varianza de la regresión*** (ANOVA) toma la siguiente forma:

Fuentes de Variación	***Suma de Cuadrados***	***Grados de Libertad***	***Cuadrado Medio***	***F Calculado***
Explicada por la regresión	*SCExp.*	$gl_{Exp.} = k$	$CM_{Exp.} = \frac{SCExp}{gl_{Exp}}$	$F = \frac{CM_{Exp}}{CM_{Res}}$
Error	*SCRes.*	$gl_{Res.} = n - k - 1$	$CM_{Res.} = \frac{SCRes.}{gl_{Res.}}$	
Total	*SCT*	$glt = n - 1$		

Siendo: k número de variables indpendientes (para la regresión simple k =1)

Como se ve el valor de **F calculado** se define como:

F = Varianza explicada por el modelo (CM_{Exp}) / Varianza residual o no explicada (CM_{Res})

Los grados de libertad para obtener el **F tabulado** se corresponden con el número de variables independientes (k) y con el número de observaciones menos el número de variables independientes menos uno (n – k – 1) para el numerador y denominador respectivamente. Para el caso exclusivo de la **regresión simple** siempre tendremos **1 grado de libertad para el numerador** y **n - 2 para el denominador**, ya que solo hay una varible independiente.

Por otro lado en el análisis de regresión se puede examinar la *significancia de los parámetros del modelo, es decir de la pendiente y de la ordenada al origen* mediante una ***prueba de t*** con n – k – 1 grados de libertad, que para el caso de la **regresión simple** siempre tendremos **n - 2 grados de libertad**, siendo las hipótesis a probar y las expresiones a utilizar las siguientes:

Prueba de hipótesis para la ordenada al origen

- Planteo de hipótesis

$$H_0: \quad \beta_0 = 0$$
$$H_1: \quad \beta_0 \neq 0$$

- Expresión de cálculo de t para la ordenada (t_a)

$$t_a = \frac{a}{S_a}$$

Siendo: $\mathbf{S_a}$ el error estandár de la ordenada al origen, el cual se puede calcular a partir de la siguiente expresión:

$$S_a = \sqrt{S_a^2} = \sqrt{\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}\right) CM_{Res}}$$

Prueba de hipótesis para la pendiente

- Planteo de hipótesis

$$H_0: \quad \beta_1 = 0$$
$$H_1: \quad \beta_1 \neq 0$$

- Expresión de cálculo de t para la pendiente (t_b)

$$t_b = \frac{b}{S_b}$$

Siendo: $\mathbf{S_b}$ el error estandár de la pendiente, el cual se puede calcular a partir de la siguiente expresión:

$$S_b = \sqrt{S_b^2} = \sqrt{\frac{CM_{Res.}}{\sum (X_i - \overline{X})^2}}$$

Características del coeficiente de regresión

1.- Es una estimación del de la población (β).

2.- Mide la pendiente de la línea de regresión, a mayor valor absoluto, mayor pendiente.

3.- Si b_1 es positivo, la línea de regresión es ascendente de izquierda a derecha, y, en caso contrario, desciende de manera inversa.

4.- "+ b" es el promedio de los incrementos de "y" debido a aumentos unitarios en "x" y "- b"
es el promedio de disminuciones de "y" debido a aumentos unitarios en "x".

5.- Sus unidades son las de la variable "y".

Características de la recta de regresión

1.- Los valores estimados por la recta son más confiables **dentro de los valores observados de x**.

2.- Deberá pasar por el punto $(\overline{\mathbf{x}}, \overline{\mathbf{y}})$.

3.- Corta al eje de las ordenadas en "a".

4.- El sentido de la recta de regresión está determinado por el signo del coeficiente de regresión.

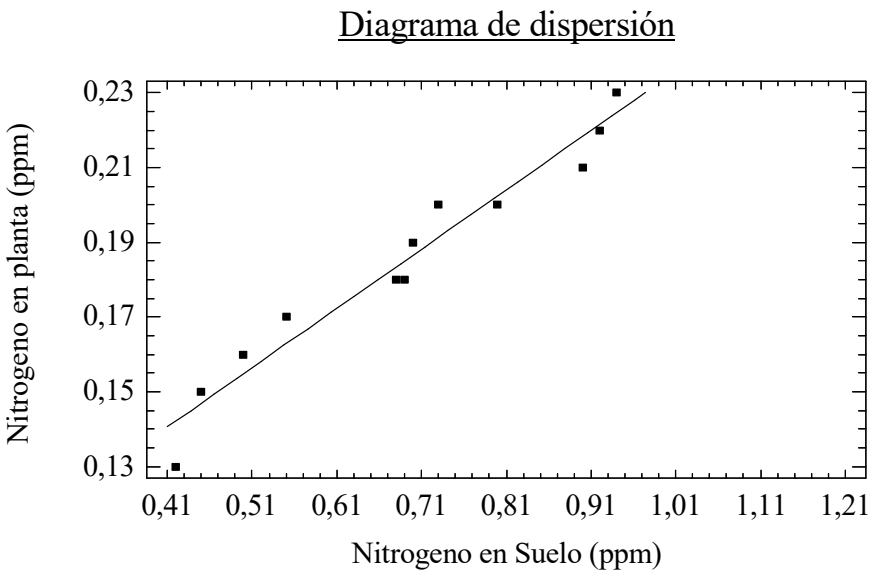
Ejemplo de aplicación

En un ensayo sobre Trigo se desea cuantificar la relación que hay entre la disponibilidad de Nitrogeno en el suelo y la cantidad de Nitrogeno en la planta. Se obtuvieron datos para 12 parcelas, con diferente contenido de nitrogeno en el suelo (X) y valores promedios de nitrogeno por planta (Y). Los resultados se presentan a continuación:

Parcela	X : Nitrogeno en Suelo (ppm)	Y : Nitrogeno en planta (ppm)
1	0,42	0,13
2	0,45	0,15
3	0,50	0,16
4	0,55	0,17
5	0,68	0,18
6	0,69	0,18

Parcela	X : Nitrogeno en Suelo (ppm)	Y : Nitrogeno en planta (ppm)
7	0,70	0,19
8	0,73	0,20
9	0,80	0,20
10	0,90	0,21
11	0,92	0,22
12	0,94	0,23

Una primera aproximación de la relación entre las variables la observamos por medio del diagrama de dispersión, el cual en este caso indica que hay una tendencia a una relación lineal y positiva entre ambas variables.



El paso siguiente es la estimación de la recta de regresión lineal. Para ello debemos calcular los estimadores **a** y **b** de los parámetros poblacionales β_0 **y** β_1 .

- Cálculos previos a la determinación de los estimadores:

$$\sum X\,Y = 1{,}5588$$
$$\sum X = 8{,}28 \qquad \overline{X} = 0{,}69$$
$$\sum Y = 2{,}22 \qquad \overline{Y} = 0{,}185$$
$$\sum X^2 = 6{,}0728$$

Hechos los calculos previos determinamos **a** y **b**:

$$b\ y/x = \frac{\sum_{i=1}^{n} \delta_x\ \delta_y}{\sum_{i=1}^{n} \delta_x^2} = \frac{\sum_{i=1}^{n} X\,Y - \frac{\sum_{i=1}^{n} X \sum_{i=1}^{n} Y}{n}}{\sum_{i=1}^{n} X^2 - \frac{\left(\sum_{i=1}^{n} X\right)^2}{n}} = \frac{1{,}5588 - \frac{8{,}28 \times 2{,}22}{12}}{6{,}0728 - \frac{8{,}28^2}{12}} = 0{,}159$$

$$a = \overline{Y} - b\ \overline{X} = 0{,}185 - 0{,}159 \times 0{,}69 = 0{,}076$$

Por lo tanto la recta de regresión estimada es:

Y = 0,076 + 0,159 X

La ordenada al origen **a** representa el valor de Y cuando X es igual a cero. En este problema, la ordenada al origen es el contenido de nitrogeno en la planta que no varía con la cantidad de nitrogeno en el suelo.

La pendiente **b** representa la variación de Y, cuando X varía en una unidad. En este problema la pendiente representa el incremento del contenido de nitrogeno en la planta (0,159 ppm) cuando la cantidad de nitrogeno en el suelo aumenta una unidad (1 ppm).

Posteriormente calculamos el Coeficiente de determinación $\mathbf{R^2}$ a partir de la suma de cuadrados explicada por el modelo y la suma de cuadrados del total.

$$\mathbf{R^2} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{SCExp.}{SCT} = \frac{0.009}{0.0095} = 0{,}9473$$

R^2 expresado en porcentaje, indica que el 94,73% de la variación del contenido de nitrogeno en la planta está explicada por el contenido de nitrogeno en el suelo.

Para analizar la significancia de la regresión, es decir para analizar si la varianza explicada por el modelo estimado de la variable dependiente es significativamente mayor que

la no explicada (varianza residual), realizamos el ANOVA cuyo planteo de hipótesis es el siguiente:

$$H_0 : R^2 = 0$$
$$H_1 : R^2 \neq 0$$

Fuentes de Variación	*Suma de Cuadrados*	*Grados de Libertad*	*Cuadrado Medio*	*F Calculado*
Explicada por la regresión	0.009	1	0.009	180
Error	0.0005	10	0.00005	
Total	0.0095	11		

El **F tabulado** lo obtenemos de tablas en base al nivel de significancia (5%) y los grdaos de libertad del numerador y denominador, para este caso 1 y 10 respectivamente. Su valor es de **4,96.**

Como el valor del F calculado (180) es mayor al valor de F tabulado (4,96) la hipótesis nula es rechazada, es decir que la varianza explicada por el modelo es significativamente mayor que la no explicada, es decir que hay un 94,73 % de la varianza de la cantidad de nitrogeno en la planta que es explicada por la cantidad de nitrogeno en suelo y que solo hay un 5,27 % de variabilidad que se debe a otros factores (variables) que se encuentran contenidos en el error.

Finalmente realizamos las pruebas de hipótesis para analizar si los estimadores (a y b) de los parámetros (β_0 y β_1) son significativamente distintas de cero.

Prueba de hipótesis para la ordenada al origen

- Planteo de hipótesis

$$H_0 : \quad \beta_0 = 0$$
$$H_1 : \quad \beta_0 \neq 0$$

- Cálculo del t_a

$$S_a = \sqrt{S_a^2} = \sqrt{\left(\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2}\right) CM_{Res}} = \sqrt{\left(\frac{1}{12} + \frac{0{,}69^2}{0{,}3596}\right) \cdot 0{,}00005} = 0{,}00839$$

$$t_a = \frac{a}{S_a} = \frac{0{,}076}{0.00839} = 9{,}06$$

Prueba de hipótesis para la pendiente

- Planteo de hipótesis

$$H_0: \quad \beta_1 = 0$$
$$H_1: \quad \beta_1 \neq 0$$

- Cálculo del t_b

$$S_b = \sqrt{S_b^2} = \sqrt{\frac{CM_{Res.}}{\sum (X_i - \overline{X})^2}} = \sqrt{\frac{0{,}00005}{0{,}3596}} = 0{,}0118$$

$$t_b = \frac{b}{S_b} = \frac{0{,}159}{0{,}0118} = 13{,}48$$

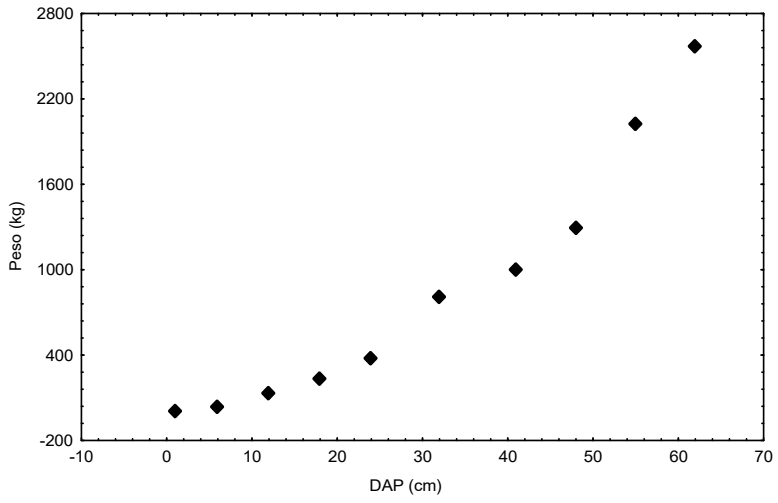
El **t tabulado** lo obtenemos de tablas en base al nivel de significancia (5%) y los grdaos de libertad calculados como n – k – 1 , para este caso 10 gl. Su valor es de **2,228.**

Las hipótesis nula, tanto para la ordenada al origen como para la pendiente son rechazadas ya que los valores de $\mathbf{t_a}$ y $\mathbf{t_b}$ son **mayores** que el valor de **t tabulado**, por lo tanto podemos concluir que la recta no pasa por el origen y que las variables X e Y (nitrogeno en suelo y nitrogeno en planta) presentan una relación lineal positiva y significativamente diferente de cero.

2) Relaciones no lineales

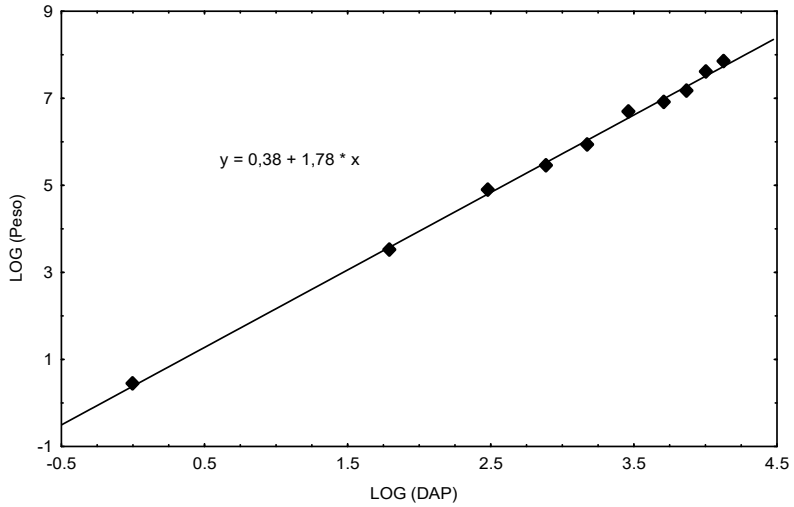
La relación entre dos variables puede ser no lineal mostrando un comportamiento curvilíneo. Existen muchas funciones matemáticas que representan líneas curvas algunas de las cuales son modelos comúnmente usados en problemas prácticos. Algunos modelos son preferidos a otros porque para poder aplicar regresión lineal, la función debe poder transformarse en una recta. Por ejemplo, el peso de una árbol puede predecirse a partir del diámetro del tronco medido a 1,3 m de altura (diámetro a la altura del pecho). Un modelo muy usado para esa relación es el ***modelo potencial***:

$$y = a \cdot x^b$$



que se transforma en una recta mediante una transformación logarítmica:

$$\log (y) = \log (a) + b \,.\, \log (x)$$



log (y) y log (x) son la nueva variable dependiente e independiente respectivamente, log (a) es la ordenada al origen y b es la pendiente de la recta.

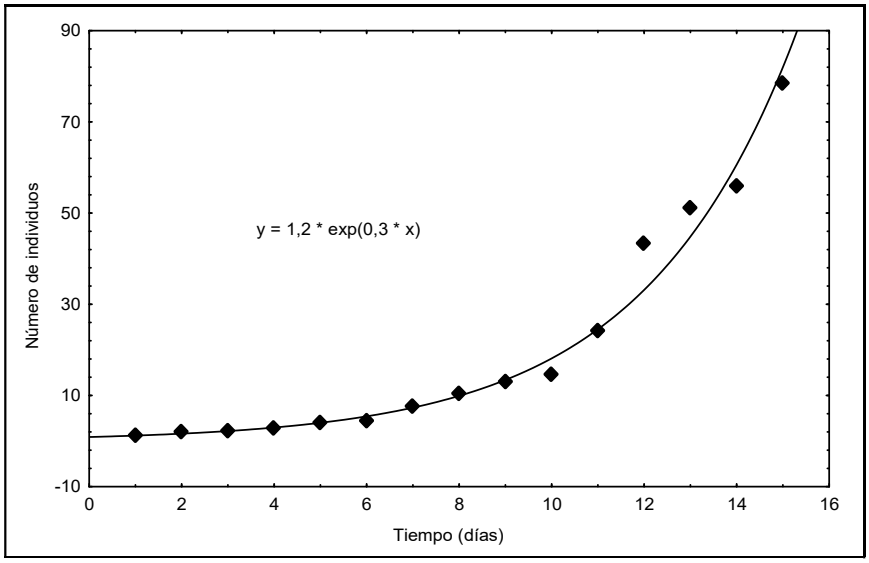
Otro ejemplo de función no lineal es el ***modelo exponencial***:

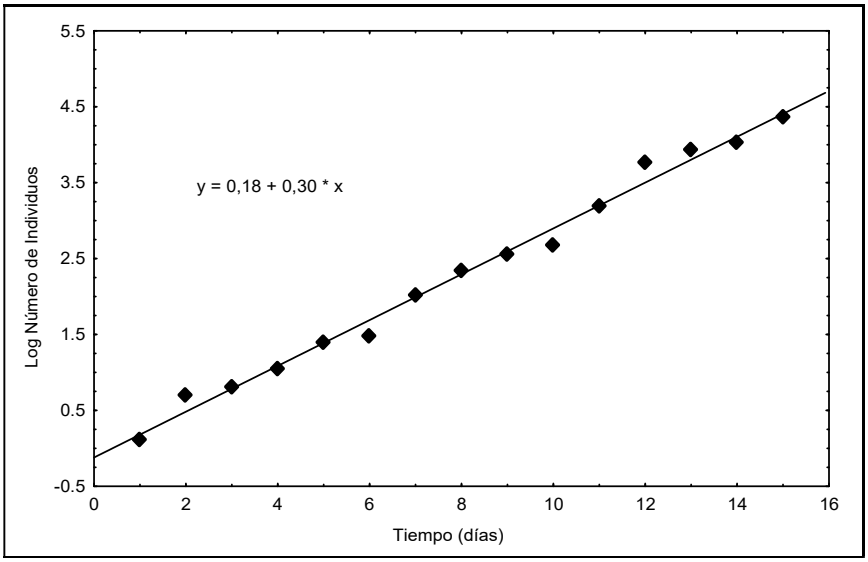
$$y = a \,.\, e^{b.\,x}$$

que se lineariza aplicando logaritmos naturales:

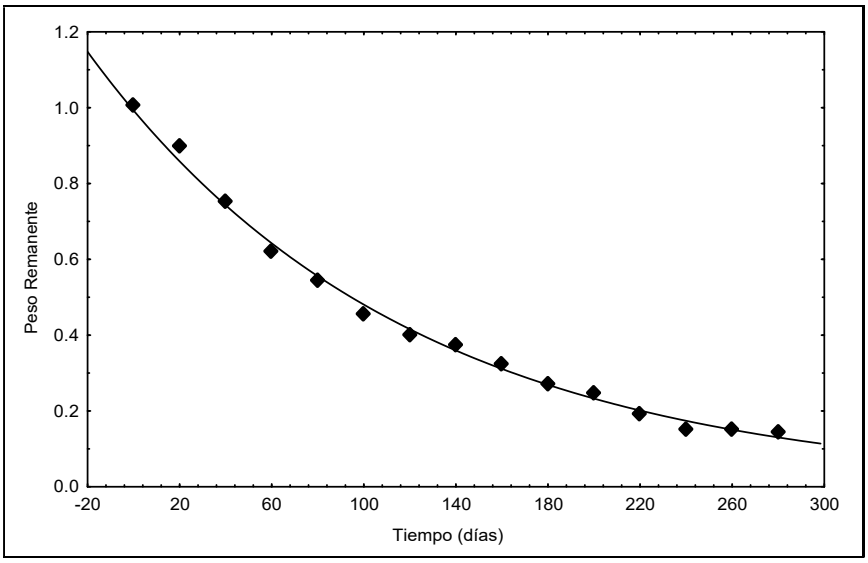
$$\log_e (y) = \log_e (a) + \log_e (e^{b.x}) = \log_e (a) + b.x$$

Esta relación puede describir el aumento del número de individuos de una población (y) en función del tiempo (x) y es útil para plagas en sus fases iniciales de desarrollo cuando el aumento es muy acelerado.



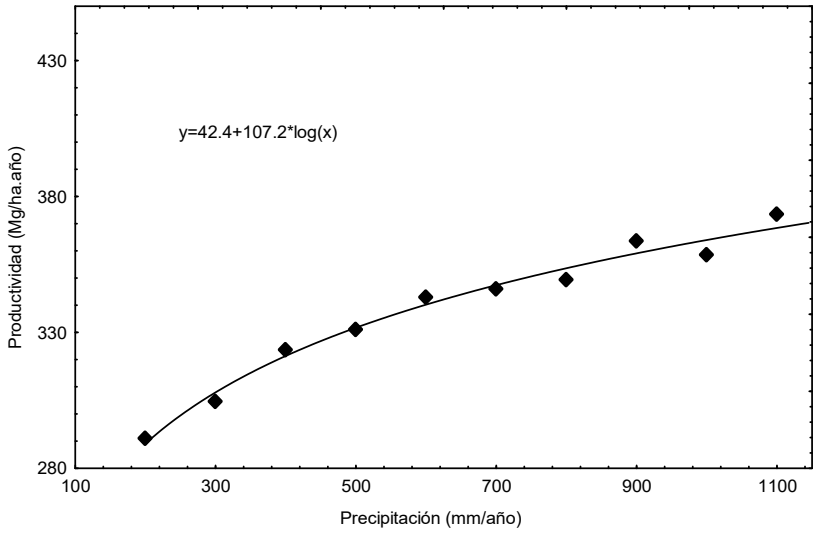


La forma negativa de estas funciones se utiliza para describir la pérdida de peso de materia vegetal que se descompone. Este es un proceso importante de retorno de nutrientes al suelo en ecosistemas naturales y agroecosistemas.



Las funciones logarítmicas positivas se ajustan a situaciones en las que "y" aumenta con "x" pero la magnitud de los incrementos va disminuyendo con el aumento de "x".

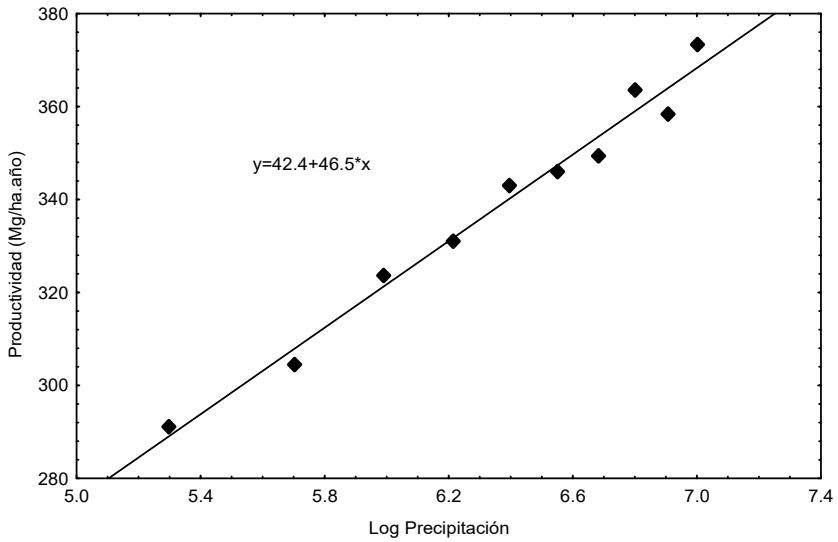
$$y = a + b \log(x)$$



Una situación que se ajusta a estos modelos es la respuesta del crecimiento de las plantas al aumento de la disponibilidad de un recurso limitante. Si se analiza la productividad de un pastizal para distintos valores de precipitación anual podría hallarse una relación logarítmica.

Podría considerarse que “y”depende linealmente de log(x) ya que si se establece que esa es una nueva variables: log(x) = x_1 , se ve que “y” depende linealmente de x_1: y = a + b.x_1

La forma de linealizar la relación es representar “y” en función de log(x) y no en función de x. La regresión se calcula entre “y” y la variable “x” tranformada logarítmicamente.



3) Regresión Lineal Múltiple

En ciertos casos es necesaria más de un variable independiente para realizar predicciones ajustadas de una variable dependiente. Existen modelos que predicen la productividad de bosques o pastizales en función de características climáticas. Algunos de ellos utilizan precitpitación y temperatura o temperatura y evapotranspiración real. Aunque esos modelos en general no son lineales podría suponerse que se comportan linealmente en un rango acotado de valores.

$$Y = a + b_1 . X_1 + b_2 . X_2$$

En esta ecuación Y depende linealmente de dos variables independientes que podrían representar temperatura y precipitación en el modelo de productividad. Si la estimación mejora con el uso de las dos variables independientes entonces $\mathbf{R^2}$ será mayor en la regresión múltiple que en cualquiera de las dos regresiones simples de Y sobre X_1 o X_2. La significancia de R^2 se prueba mediante **la prueba de F** cuyos grados de libertad son:

número de variables independientes para el numerador = (k)

número de observaciones - nro variables independientes – 1 para el denominador = (n–k–1)

En la regresión múltiple la ordenada al origen representa el valor de Y cuando todas las variables independientes valen cero. Existe un valor de pendiente asociado a cada variable independiente y su significancia se prueba separadamente mediante pruebas de “t” independientes para cada pendiente.

Análisis de Correlación Lineal

El Análisis de Correlación Lineal como lo expresamos al comienzo de esta sección es un método estadístico que permite medir el grado de asociación y el sentido del mismo entre variables aleatorias que se supone se relacionan linealmente.

La medida del grado de asociación entre las variables X e Y se realiza por el Coeficiente de Correlación muestral "**r**", el cual constituye el estimador de rho (ρ) que es el Coeficiente de Correlación Poblacional.

Este coeficiente mide la "**intensidad de asociación entre variables**".

Se puede demostrar que el coeficiente de correlación lineal, “r”, es un número que necesariamente está entre menos uno y uno, es decir:

$$-1 \leq r \leq 1$$

A continuación se ilustran tres tipos diferentes de valores extremos de asociación entre variables:

r = - 1 : Perfecta relación lineal inversa entre las variables. Todos los puntos pertenecen a una recta de pendiente negativa.

r = 1 : Perfecta relación lineal directa entre las variables. Todos los puntos pertenecen a una recta de pendiente positiva

r = 0 : No hay relación lineal entre las variables. Ya sea porque, las variables no están asociadas, o porque la relación entre ellas no es lineal

El estudio clásico de la correlación se basa en la suposición de que la distribución de valores (X_i, Y_i) es una distribución normal bidimensional y se representa en gráficos tridimensionales.

Cálculo del Coeficiente de Correlación:

$$\rho = \frac{\sigma_{x\,y}}{\sqrt{\sigma_x \,.\, \sigma_y}} = \frac{Cov(x\,,\,y)}{\sqrt{V_{(x)}\,.\,V_{(y)}}} = \frac{Cov(X\,,\,Y)}{S_x \,.\, S_y}$$

El coeficiente de correlación muestral se calcula como:

$$r = \frac{\sum_{i=1}^{n} \delta_x\, \delta_y}{\sqrt{\delta_x{}^2\, \delta_y{}^2}} = \frac{\sum_{i=1}^{n} x_i\, y_i \;-\; \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i{}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] \cdot \left[\sum_{i=1}^{n} y_i{}^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

Recordando los conceptos vistos en el análisis de regresión, se comprueba que:

$$r = \sqrt{b_{y/x} \,.\, b_{x/y}}$$

Siendo:

$$b\ y/x = \frac{\sum_{i=1}^{n} \delta_x\, \delta_y}{\sum_{i=1}^{n} \delta_x{}^2}$$

$$b\ x/y = \frac{\sum_{i=1}^{n} \delta_x\, \delta_y}{\sum_{i=1}^{n} \delta_y{}^2}$$

Prueba de hipótesis para el coeficiente de correlación poblacional

La prueba de significancia consiste en probar la hipótesis nula que el coeficiente de correlación poblacional es cero, contra la alternativa que es distinto de cero mediante una ***prueba de t*** con **n - 2 grados de libertad**, es decir:

- Planteo de hipótesis

$$H_0: \quad \rho = 0$$
$$H_1: \quad \rho \neq 0$$

- Expresión de cálculo de t_r

$$t_r = \frac{r}{S_r}$$

Siendo: $\mathbf{S_r}$ el error estandár del coeficiente de correlación, el cual se puede calcular a partir de la siguiente expresión:

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

EXPERIMENTACION

Si se acepta la premisa de que el conocimiento nuevo se obtiene muy frecuentemente a través de análisis y la interpretación cuidadosa de los datos, entonces es muy importante que se deba dedicar tiempo y esfuerzo considerables al planeamiento y recolección de los mismos, con el objeto de obtener la máxima información con el menor costo de recursos.

En muchas oportunidades, la obtención de los datos ha sido realizada con anterioridad, y nuestra tarea solo será la de efectuar los procedimientos estadísticos apropiados al problema que se quiere resolver y luego interpretar los resultados obtenidos. Sin embargo, hay situaciones en las cuales no disponemos de los datos, y es nuestra responsabilidad generarlos de la mejor y más correcta forma posible.

Hay situaciones donde los datos fueron obtenidos descuidando ciertos principios básicos, que, de haberse atendido, con el mismo costo se hubiera logrado una mayor eficiencia en el análisis de la información.

Por lo expresado precedentemente podemos concluir que *"al momento de decidir realizar un experimento o bien al asesorar sobre como idearlo para generar la información deseada debemos tener presente los objetivos del mismo, de forma tal de diseñarlo adecuadamente para poder aplicar los procedimientos estadísticos convenientes y lograr la mayor eficiencia con el menor costo".*

Para comprender y poder utilizar adecuadamente la **experimentación**, se deberán conocer previamente los siguientes conceptos:

Investigación: Labor de observación y razonamiento para conocer naturaleza, significado y consecuencias de cualquier conjunto particular de circunstancias, con el objeto de generalizar sus resultados y poder aplicarlos a situaciones más complejas.

Experimentación: Herramienta del método científico mediante la cual sometemos a prueba hipótesis o estimamos diferencias entre los efectos de dos o más tratamientos.

Experimento: Se puede considerar a un experimento como una búsqueda planeada para obtener nuevos conocimientos o para confirmar o no resultados de experimentos previos. Tal investigación ayudará en la toma de decisiones administrativas, tales como la recomendación de una variedad, un procedimiento o un pesticida, etc.

Tratamiento: Es un procedimiento cuyo efecto se mide u observa. Puede ser un factor o combinación de factores a ser estudiados, por ejemplo: Dosis de fertilizantes, insecticidas, funguicidas, razas de ganado, variedades de una especie, distancias de plantación, sistemas de labranzas, etc. Cuando se utilizan combinaciones de factores como: especies de Eucalyptus y distancias de plantación, variedades de soja y épocas de siembra, los ensayos tendrán tantos tratamientos como especies x distancias o variedades x épocas se prueben respectivamente.

Experimento Simple: Experimento en el cual se analiza un solo factor (Tratamiento), donde el conjunto de tratamientos se corresponde con los niveles del factor bajo análisis.

Experimento Factorial: Experimento en el cual se analizan dos o más factores, en este caso el conjunto de tratamientos a ensayar surge de la combinación de los niveles de los factores a estudiar.

Hipótesis Estadística: Suposición que se verifica sólo por métodos estadísticos. Ya sea sobre un tipo de distribución o sobre los parámetros de distribuciones desconocidas.

Ejemplo:"Tal conjunto de datos está distribuido según la Distribución de Poisson" (Se supone sobre el tipo de la distribución desconocida); "Las varianzas de dos distribuciones son iguales" (sobre los parámetros de dos distribuciones desconocidas).

Prueba de Significancia: Técnica estadística que permite resolver si se rechaza o no una hipótesis.

Diseño experimental: Plan al cual se ajusta un experimento con el objeto de reunir la información apropiada para dar respuesta al problema planteado.

Análisis de la Varianza: Es una técnica estadística que sirve para analizar la variación total de los resultados de un experimento montado en un diseño particular, descomponiéndolo en fuentes de variación independientes atribuibles a cada uno de los efectos en que se constituye el experimento.

Tipos de modelos estadísticos: De acuerdo a la selección de los tratamientos se tiene la siguiente clasificación:

- *Modelo I* (Efectos Fijos):

Se presenta cuando los tratamientos son fijados por el investigador; es decir, no se efectúa una elección aleatoria. En estos casos las conclusiones del análisis de varianza solamente son válidas para los tratamientos usados en el experimento. En este apunte solo consideraremos el caso de modelo de efectos fijos, por ser el que se presenta con mayor frecuencia en la experimentación agropecuaria.

- *Modelo II* (Efectos aleatorios):

Se presenta cuando los tratamientos que intervienen en un experimento son elegidos al azar de una población. En estos casos las conclusiones del análisis de varianza son válidos, tanto para los tratamientos usados, así como para toda la población de tratamientos.

- *Modelo III* (Modelo Mixto):

Este modelo es la combinación de los dos anteriores y se presenta cuando algunos factores son fijados y otros son elegidos al azar. En estos casos las conclusiones del análisis de

varianza serán válidas para toda la población de factores cuando estos son elegidos al azar, y solamente para los factores usados cuando estos son fijados.

Pruebas de comparaciones múltiples de medias: Es propósito de todo investigador que realiza un análisis de varianza de un experimento en particular, realizar la prueba sobre el efecto de los tratamientos en estudio, para ello hace uso de la prueba F la cual indicará (para el caso de un experimento simple) si los efectos de todos los tratamientos son iguales o hay al menos uno que difiere de otro; en caso de aceptar la hipótesis de que todos los tratamientos no tienen el mismo efecto (es decir que hay uno que difiere de otro), entonces es necesario realizar pruebas de comparación de medias a fin de saber entre que tratamientos hay diferencias significativas y entre cuales so las hay por el azar. Las pruebas de comparación de medias más utilizadas son: el test de Tukey, la prueba de Duncan, prueba de Scheffe, prueba de comparación de Dunnet, etc.

Parcela: Es la unidad experimental a la cual se le aplica un tratamiento. Ejemplo: un número de vacas, un cuadro de alfalfa, nueve plantas de Sauces plantadas en cuadro o en fila, una parcela de suelo, etc.

Efecto de Borde: En los experimentos pecuarios la unidad experimental por lo general esta conformada por un animal (novillo, cerdo, etc.), en los experimentos forestales la unidad experimental en la mayoría de los casos esta conformado por un árbol, mientras en los experimentos agrícolas, la unidad experimental es una parcela de tierra; es en este último caso con frecuencia se presenta lo que se llama *efecto de borde*.

Se define como *efecto de borde* a las diferencias, que generalmente se presentan, en el crecimiento y la producción de las plantas que están situadas en los perímetros de la parcela en relación con aquellas plantas situadas en la parte central. Estas diferencias pueden causar sobre-estimación o sub-estimación de las respuestas de los tratamientos, llegando con esto a comparaciones sesgadas entre ellos.

El efecto de bordes puede ser causado por:

- Vecindad de las parcelas ó áreas no cultivadas, que hace que las plantas en los perímetros tengan menor competencia de luz y nutrientes.

- Competencia entre tratamientos, que depende de la naturaleza de los tratamientos vecinos.

Para controlar el efecto de borde se acostumbra a evaluar solamente las plantas centrales de las parcelas. Estas plantas centrales constituyen lo que se suele llamar parcela neta experimental.

Error Experimental: Es una medida de la variación existente entre parcelas tratadas en forma similar. Dicha variación puede provenir de dos fuentes principales: (1) de la variabilidad inherente al material experimental al cual se aplican los tratamientos; (2) de una variación resultante de cualquier falta de uniformidad en la realización física del experimento.

.

Ejemplo: Si a 50 gallinas se las enjaula juntas y se las alimenta con la misma ración, la unidad experimental es una parcela de 50 gallinas. Se necesitan otras jaulas de 50 gallinas antes de poder medir la variación entre unidades tratadas en forma semejante. Por otra parte cada gallina será genéticamente diferente de las otras (variabilidad inherente al material) y también las jaulas podrán estarán expuestas a diferentes condiciones de luz, humedad, temperatura, etc. (falta de uniformidad).

El *control del error experimental* puede lograrse mediante:

1.- Una selección adecuada del diseño experimental.
2.- Un análisis de covarianza (que no desarrollaremos en este curso).
3.- La elección del tamaño y la forma de las parcelas o unidades experimentales

El **tamaño de parcela** en ensayos a campo ha sido un tema de bastante discusión entre los investigadores, debido generalmente a que es una característica particular de los experimentos que puede variar según una serie de factores. Lo cierto es que cuando un investigador va a planificar sus ensayos a campo, lo ideal sería que contara con un tamaño de parcela experimental adecuado que le permitiera disminuir al máximo posible el error experimental y así poder detectar como significativas las diferencias que pudieran existir entre tratamiento, si es que las hay.

A muchos de los investigadores, cuando les llega el momento de seleccionar el tamaño que va a tener la unidad experimental de su ensayo, generalmente lo que hacen es: a) seguir criterios de tipo personal sin ninguna consideración ni estadística ni económica, b) revisión de literatura extranjera, lo cual no es totalmente deseable ya que el tamaño de parcela es una característica muy local influenciada mucho por las características de la zona donde se desarrolla el experimento, c) seguir criterios estadísticos, económicos y prácticos; claro que para ello se necesita una investigación anterior que realmente casi no existe.

Entre los factores que más influyen en el tamaño y la forma de la parcela tenemos:

- La extensión superficial del terreno disponible.

Cuando se dispone de un terreno suficientemente grande se puede utilizar el tamaño de parcela necesario para que la variabilidad del error sea mínima. Cuando contamos con terrenos muy pequeños, debemos reducir el tamaño de la parcela en proporción al número de repeticiones para que los resultados tengan la confiabilidad suficiente.

- El tipo de suelo.

Cualquier persona pensaría, y de hecho ha ocurrido así que si el suelo es heterogéneo deberían usarse parcelas pequeñas, para no caer en zonas de diferente fertilidad, pero esto, hasta cierto punto, no es cierto, ya que si revisamos cualquier mapa de heterogeneidad de suelo, las líneas casi siempre se presentan siguiendo una irregularidad muy marcada. De hecho entonces, si utilizamos parcelas pequeñas, habrá algunas que caerán en zonas de fertilidad muy diferente y contribuirán, por lo tanto, a aumentar el error. En tanto que

si utilizamos parcelas grandes las diferencias no debidas a tratamientos sino a heterogeneidad de suelo son menos notables, ya que en cada parcela como se incluyen más de una línea de heterogeneidad, las variaciones con relación a las otras parcelas son menores.

- La clase de cultivo.

Sin duda el cultivo a ensayar influirá en el tamaño de parcela a utilizar. Cuando cada planta ocupa un espacio grande, como sucede con los árboles frutales, es difícil que cada unidad experimental tenga un elevado número de plantas, como sucede con cultivos de menor tamaño, en cuya unidad experimental puede haber un gran número de individuos, pero también debe tomarse en cuenta que este número no puede reducirse demasiado, pues las variaciones que pueden existir en cultivos perennes de planta a planta, podrían aumentar mucho el error experimental.

- Los métodos de cultivo.

Se refiere a los medios de que se dispone para la preparación de las parcelas o para su plantación. Si tanto la preparación como la siembra y cosecha se hacen a mano, es muy posible que el tamaño y la forma sean diferentes a cuando se utiliza la maquinaria agrícola. Algunas veces el empleo de la maquinaria forma parte misma del experimento, por consiguiente, el tamaño y la forma de la parcela tiene que ser la más adecuada para la aplicación de la maquinaria prevista.

- El grado de precisión deseado.

Cuando el grado de precisión que se desea es grande, debido a que la experiencia del investigador indica que es necesario disminuir al máximo el error porque las diferencias entre sus tratamientos se consideran pequeñas, es adecuado utilizar el tamaño más grande dentro del rango recomendado de tamaño de parcela.

Es sabido que la **forma de la parcela** tiene menos influencia que el tamaño en la disminución del error experimental. Sin embargo, algunas veces la forma de parcela puede ser muy importante, ella depende mucho del manejo de las diferentes prácticas culturales, forma general del campo y exigencia del cultivo que se trate.

Existe una gran variación en la forma de la parcela, puede haber parcelas rectangulares de diferentes dimensiones y en diferentes sentidos, al igual que parcelas de forma cuadrada. La forma rectangular tiene las ventajas de que muchas veces facilita las prácticas culturales del cultivo, uso de maquinaria, riego, fertilización, control de plagas, etc.

Los estudios de ensayos de uniformidad, es decir estudios de datos obtenidos de experimentos donde no se aplican tratamientos, han indicado que las parcelas individuales rectangulares son generalmente menos variables que las parcelas cuadradas; por lo tanto es más conveniente utilizar éstas formas de parcelas.

Por último, se hace necesario mencionar que se han desarrollado varios métodos con el objeto de determinar el tamaño y la forma más conveniente de parcela.

Principios básicos de la experimentación: Repetición – Aleatorización – Bloqueo.

- **Repetición**:

Es el número de veces que se repite un mismo tratamiento. Las funciones de la repetición son:

1.- Permitir una estimación del error experimental.

2.- Mejorar la precisión de un experimento mediante la reducción del error estándar de la media de un tratamiento.

3.- Ejercer control sobre la varianza del error.

La determinación del número de repeticiones es uno de los problemas más interesantes de la experimentación. Numerosa soluciones han sido propuestas, pero ninguna es enteramente satisfactoria para todas las situaciones.

A través de los años la experimentación agrícola y zootécnica indica que difícilmente se arribe a resultados satisfactorios desde el punto de vista estadístico con ensayos que tengan menos de 20 unidades experimentales.

Así en un experimento con dos tratamientos, debemos tener por lo menos 10 repeticiones por tratamiento. Otra indicación útil, en general, es que debemos tener como mínimo 12 grados de libertad para el error experimental.

Estas dos indicaciones, pueden ser dejadas de lado en algunos casos. Esto puede ocurrir en experimentos de gran precisión (ensayos físicos, químicos, etc.) donde se requiere disminuir el error experimental y por ende un aumento en el número de repeticiones, o bien cuando tenemos un necesidad de realizar un grupo numeroso de ensayos que serán estudiados en conjunto, teniendo en vista únicamente resultados generales, se puede disminuir el numero de repeticiones en cada ensayo individual, afín de, con los recursos disponibles, aumentar el número de ensayos.

Entre las variadas técnicas que existen para determinar el número de repeticiones se encuentran las curvas características de operación, el test de Tukey, el procedimiento de Stein, etc.; cuyos desarrollos y aplicaciones exceden a los alcances de este curso.

- **Aleatorización**:

También llamada casualización, consiste en la distribución al azar de los tratamientos, a los efectos de asegurar que un tratamiento particular no resulte favorecido en forma consistente en repeticiones sucesivas por alguna fuente externa de variación conocida o desconocida y evitar una subestimación o sobreestimación del error experimental como podría acontecer en un diseño sistemático, en el cual los tratamientos se aplican a las unidades experimentales de una manera no aleatoria y seleccionada.

El cumplimiento de este principio básico de la experimentación asegura la validez de las pruebas de significancia al eliminar la correlación entre los errores.

Bloqueo o " control local":

Es un principio básico de la experimentación de uso muy frecuente, pero no obligatorio como los dos anteriores. En la práctica, el experimentador puede algunas veces advertir a priori que su material experimental es heterogeneo, entonces, para mejorar la precisión o sensibilidad del experimento, debe tratar de agrupar las unidades experimentales en bloques homogeneos y dentro de cada bloque realizar una asignación aleatoria de cada tratamiento.

Supongamos que queremos probar el rendimiento de dos variedades de soja (1 y 2) y que nuestro campo experimental se encuentra dividido en 10 parcelas o unidades experimentales, y que luego de realizar la distribución al azar de las variedades a las unidades experimentales resulta la siguiente asignación de parcelas para las variedades 1 y 2:

1	1	1	1	2
2	1	2	2	2

Si hubiera un aumento de fertilidad en el sentido de la flecha, la variedad 1 podría estar favorecida y, si su rendimiento fuera igual al de la variedad 2, y no existiera otra causa de variabilidad, fuera de la fertilidad, se obtendría un resultado erróneo $\overline{X}_1 > \overline{X}_2$.

La situación planteada se podría corregir y evitar un resultado sesgado del análisis si previo al ensayo se agrupan las parcelas en bloques homogeneos, para ello se deben disponer los bloques en forma perpendicular a la heterogeneidad dada por la fertilidad y asignar al azar dentro de cada bloque una repetición de cada variedad, es decir:

	1	2	2	1	2
	2	1	1	2	1
BLOQUE	**I**	**II**	**III**	**IV**	**V**

Con dicha asignación se compensan los efectos que tenia la variación de fertilidad sobre el rendimiento de las variedades, afectando a ambas variedades por igual.

Por otro lado, los bloques no necesariamente deben estar uno al lado del otro, pueden estar distribuidos por todo el campo experimental.

En caso de ensayos con animales, cada bloque debe encerrar animales de constitución genética, peso, edad, etc, bien semejantes.

En los ensayos de laboratorio las muestras de cada bloque deben, en lo posible, ser analizadas simultáneamente, es decir el mismo día y por el mismo analista.

Análisis de la Varianza

El **Análisis de la Varianza (ANOVA)** fue ideado por Sir Ronald Fisher y es esencialmente un procedimiento aritmético que descompone una suma total de cuadrados en componentes asociados con fuentes de variación reconocida. Se ha usado con provecho en todos los campos de la investigación en los que los datos se miden cuantitativamente.

Los resultados numéricos del análisis de la varianza se presentan en una **Tabla conocida como Tabla de Análisis de la Varianza** la cual toma la forma según el tipo de experimento (simple o factorial) y el diseño experimental usado en el montaje del ensayo.

La **hipótesis nula** que se pone a prueba en el análisis de la varianza en un experimento simple plantea la igualdad de la medias de los tratamientos, es decir que el efecto de los tratamientos es nulo, contra la **hipótesis alternativa** que hay al menos una diferencia significativa entre dos tratamientos. Por lo tanto tendremos:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t \Leftrightarrow H_0 : \tau_1 = \tau_2 = \cdots = \tau_t = 0$$
$$H_1 : \textit{Hay al menos una diferencia}$$

Supuestos del Análisis de la Varianza

Todo análisis de la varianza se basa en un modelo lineal cuyas componentes varían en función de las fuentes de variación que se desea controlar.

La **linealidad** se debe a que las variables que participan están elevadas a la potencia 1. Todo análisis de la varianza presupone la existencia de un modelo lineal que cumple ciertos supuestos básicos:

1) **ADITIVIDAD:** Los factores que participan en el modelo son aditivos. Cualquier observación es un buen ejemplo de aditividad. En estadística, un modelo corriente que describa la naturaleza de una observación consta de una **media más un error**. Este es un **modelo lineal aditivo**.

El modelo más simple es de la forma:

$$Y_i = \mu + \varepsilon_i$$

es decir, "la observación i-ésima es una observación de la media (μ), pero está sujeta a un error de muestreo (ε_i) que actúa en forma aditiva sobre ella.

2) INDEPENDENCIA: Los errores (ε_i) son independientes (no presentan correlación).

3) HOMOCEDASTICIDAD: Los errores (ε_i) tienen la misma varianza σ^2.

4) NORMALIDAD: Los errores (ε_i) tienen distribución Normal.

En general podemos decir que si estos supuestos se verifican sólo aproximadamente, el modelo tendrá validez, aunque existen pruebas y métodos numéricos para asegurarnos su cumplimiento.

Los factores que participan de un modelo varían en función del diseño experimental que se implementa, pero todos ellos deberán cumplir los supuestos enunciados. Respecto a ello, las pruebas de significancia aplicadas (t o F) no experimentan cambios en su validez si el supuesto de normalidad se verifica parcialmente.

Favorecemos que los errores de una distribución respondan a una Distribución Normal, haciendo uso de las **Transformaciones,** recordar que utilizamos transformaciones para llevar una variable **X** a una **Z**, con distribución **N** (0,1).

Cuando la distribución de los errores se aparta de la Normalidad, podemos superar este inconveniente realizando una transformación de la variable para aproximar la variable a la normalidad. Por ejemplo, si las variables son en porcentaje (Distribución Binomial), podemos usar la transformación:

$$Z_{ij} = arc\ sen\sqrt{(\%)}$$

Esta es una transformación "arco seno", pero existen muchas más (logarítmica, raíz cuadrada, etc.) que normalizarán en mayor o menor grado las variables, debiendo escoger la más conveniente.

Pruebas de Comparaciones Múltiples de Medias

Test de Tukey:

Es una prueba de uso muy simple y exacta, en el caso de que el número de repeticiones sea igual para todos los tratamientos y permite comparar las medias de a pares. La expresión por la cual calculamos las denominadas "diferencias mínimas significativas" (d.m.s.) es:

$$\Delta_{\%} = q_{\%} \cdot \frac{S}{\sqrt{r}}$$

Donde "q" es la "amplitud total estudentizada", obtenida de tabla (pagina 93) en función del nivel de significancia, los grados de libertad del error y el número de tratamientos; "S" es el error estándar o raíz cuadrada del cuadrado medio del error; "r" el número de repeticiones en un **DCA** o el número de bloques en un **DBCA.**

Calculadas las diferencias entre las medias de tratamientos se observa si estos valores superan o no a las d.m.s para el nivel de significancia fijado. Si la diferencia entre dos medias no supera a la d.m.s ambas medias se pueden identificar con una misma letra minúscula, caso contrario se deben identificar con letras diferentes. Es decir que dos medias con una misma letra indicará que no presentan diferencias significativas. Otra forma de

identificar la significancia es indicar ambas medias con una cruz en una misma columna cuando no difieren significativamente y en distinta columna si sí difieren.

Test de Duncan:

Es de aplicación más dificultosa que Tukey, no obstante nos permite comparar más de dos medias y discrimina con mayor facilidad entre los tratamientos, indicando significancia en casos en que Tukey no la detecta.

Para ser exacta exige igual número de repeticiones en los tratamientos. La expresión que calcula la d.m.s. para comparar la diferencia entre la **mayor y la menor media del grupo** es:

$$D_{\%} = Z_{\%} \cdot \frac{S}{\sqrt{r}}$$

Donde "z" es el valor de la amplitud total estudentizada obtenido de tabla (pagina 94) en función del nivel de significancia, del número de medias que abarca el contraste (comparación) y los grados de libertad del error.

Si el valor del contraste supera al valor de D calculado, es porque la media mayor difiere significativamente de la menor, procediendo luego a calcular un nuevo D para poder comparar una nueva diferencia, existiendo tantos valores de D como medias menos uno.

El resultado final es dado por la distribución de las medias en orden creciente o decreciente y aquellas que no difieren son unidas por una barra o por una misma letra minúscula.

Prueba de Scheffé:

Se aplica exclusivamente cuando la prueba de F en el análisis de la varianza ha sido significativa ya que por lo menos existirá una comparación significativa. La prueba es más general que la de Tukey y Duncan, permitiendo comparar cualquier contraste.

Su d.m.s. se calcula como: $S = \sqrt{(n-1).F.V_{(\bar{Y})}}$

Donde: F = es el valor obtenido por tablas (paginas 90 a 92) al nivel del 5 o del 1% de probabilidad, es función de los grados de libertad de tratamientos y los grados de libertad del error.

$V_{(\bar{Y})}$ = es la varianza estimada del contraste.
n = es el número de tratamientos.

El procedimiento de comparación del contraste con el valor de S es similar a las anteriores pruebas.

Prueba de Dunnet:

Esta prueba es útil cuando el experimentador está interesado en determinar si un tratamiento difiere o no de un testigo, control o tratamiento estándar, y no en hacer todas las comparaciones posibles (que pasarían a una segunda prioridad); es decir, cuando se quiere comparar el testigo con cada uno de los tratamientos en estudio. Este procedimiento requiere una sola d.m.s para juzgar la significancia de las diferencias entre cada tratamiento y el control.

Coeficiente de variación (CV)

El coeficiente de variación es una medida de variabilidad que generalmente acompaña a todo análisis de la varianza, ya sea de un experimento simple o factorial, y que da una idea de la precisión del experimento. Se calcula como:

$$CV = \frac{S}{\overline{Xg}} \bullet 100$$

Donde "S" es el error estándar o raíz cuadrada del cuadrado medio del error y $\overline{Xg}$ la media general del ensayo.

En ensayos agrícolas de campo podemos considerar valores de CV < al 10% como bajos, entre 10-20% medios, entre 20-30 altos y > 30% muy altos.

A continuación se analizan los **Experimentos Simples** montados en diseños experimentales "**completamente aleatorizados**", "**bloque completos al azar**" y en "**cuadrado latino**"; sus características, sus ventajas y desventajas, cómo se aplica el análisis de la varianza y cómo se ajustan al modelo lineal aditivo.

Experimentos simples

Diseño Completamente Aleatorizado (DCA)

Es el más simple de los diseños. Su característica distintiva es la distribución totalmente al azar de los tratamientos en las unidades experimentales, sin presentar esta casualización ninguna restricción.

Es considerado un diseño básico, ya que los restantes se originan por la aplicación, en mayor o menor medida, de restricciones en la distribución de los tratamientos en las unidades experimentales. El modelo lineal para este diseño es el siguiente

$$Y_{ij} = \mu + t_i + \varepsilon_{ij}$$

Y_{ij} = Observación del tratamiento i repetición j.
μ = Media general del ensayo.
t_i = Efecto del tratamiento i.

ε_{ij} = Contribución del azar (factores no controlados).

Su aplicación es recomendable cuando las unidades experimentales son relativamente homogéneas, o sea cuando la variación entre ellas es pequeña. Por ello su uso generalizado en experimentos de laboratorio, donde se ejerce gran control sobre las condiciones experimentales. En el caso de trabajar con macetas (invernáculos, vidrieras), si cambiamos frecuentemente su posición no es necesario recurrir a otros diseños (bloques).

Ventajas:

- El diseño completamente aleatorio es flexible en cuanto a que el número de tratamientos y de repeticiones sólo está limitado por el número de unidades experimentales disponibles.

- El número de repeticiones puede variar de un tratamiento a otro, aunque generalmente lo ideal sería tener un número igual por tratamiento. El análisis estadístico es simple aún en el caso en que el número de repeticiones difiera con el tratamiento y si algunas unidades experimentales o tratamientos enteros faltan o se descartan.

- La presencia de parcelas perdidas no es obstáculo para practicar el análisis de la varianza con los debidos recaudos. Esto no ocurre en los otros diseños, en los que se deberá recalcular la unidad experimental faltante para poder realizar el ANOVA.

- El número de grados de libertad para estimar el error experimental es máximo; esto mejora la precisión y es importante con experimentos pequeños, por ello se recomienda que todo experimento no deba tener menos de 20 parcelas y que los grados de libertad del error no sean menores que 12.

Desventajas:

- El error experimental incluye toda la variación entre las unidades experimentales excepto la debida a los tratamientos, por ello frecuentemente es muy grande. Como alternativa podemos agrupar las unidades experimentales de manera de disminuir esa variación y aumentar la eficiencia del ensayo.

En la **Tabla de Análisis de la Varianza** (Tabla 1) se resumen los estadísticos y cálculos básicos para obtener el Cuadrado Medio Entre o Cuadrado Medio de Tratamientos **(CME)** y el Cuadrado Medio Dentro o Cuadrado Medio del Error **(CMD)**, estadísticos estos claves para la prueba de **hipótesis nula de igualdad de medias** (Prueba de F).

Fuentes de Variación	***Suma de Cuadrados***	***Grados de Libertad***	***Cuadrado Medio***	***F Calculado***
Entre Tratamientos	*SCE*	*gle = t – 1*	$CME = \frac{SCE}{gle}$	$F = \frac{CME}{CMD}$
DentroError Experimental	*SCD = SCT - SCE*	*gld = t x (r -1)*	$CMD = \frac{SCD}{gld}$	
Total	*SCT*	*glt = N –1*		

Tabla 1

(Donde: t = número de tratamientos; r = número de repeticiones;

N = número de observaciones totales = t x r)

Para el cálculo de las sumas de cuadrado se pueden utilizar las siguientes expresiones:

Suma de cuadrado total (SCT)

$$SCT = \sum_{i,j} Y_{ij}^2 - C = \sum_{i,j} Y_{ij}^2 - \frac{\left(\sum_{i,j} Y_{ij}\right)^2}{N}$$

Siendo: Y_{ij} la observación j-ésima bajo el tratamiento i-ésimo, i = 1, 2, ..., t. y j = 1, 2, ..., r.

C = termino corrector o factor de corrección.

Suma de cuadrado entre o Suma de cuadrado de tratamientos (SCE)

$$SCE = \frac{Y_1^2 + \cdots + Y_t^2}{r} - C$$

Siendo: Y_1 hasta Y_t los totales de los tratamiento, desde el tratamiento 1 hasta el tratamiento t.

Suma de cuadrado dentro o Suma de cuadrado del error (SCD)

$$SCD = SCT - SCE$$

Una vez completada la tabla de análisis de la varianza y obtenido el valor de **F calculado** debemos buscar el valor de **F tabulado** para contrastar la hipótesis de nulidad.

El valor de **F tabulado** se extrae de las tablas correspondientes, en función del nivel de significancia y de los grados de libertad del numerador y del denominador (grados de libertad de tratamientos y grados de libertad del error respectivamente).

Probada la **hipótesis de nula** si esta **no es rechazada**, se asume que no existen diferencias significativas entre medias; en cambio **si es rechazada**, se asume que si existen diferencias significativas entre medias. En esta última situación para detectar cuales son verdaderamente diferentes y cuales solo lo son por el azar debemos realizar las **"Comparaciones múltiples de Medias"**, como el **Test de Tukey**, el **Test de Duncan**, entre otros.

Ejemplo de Aplicación:

Se ha realizado un ensayo para observar el efecto que provocan 4 raciones (A, B, C y D) en el aumento de peso en cerdos. Dado que contamos con animales relativamente homogéneos, adjudicamos al azar los tratamientos (raciones). El ensayo consta de 5 repeticiones y los resultados fueron:

RACIONES (tratamientos)

	A	***B***	***C***	***D***	*Totales*
	35	40	39	27	**141**
	19	35	23	12	**89**
	31	46	20	13	**110**
	15	41	29	28	**113**
	30	33	45	30	**138**
Totales	**130**	**195**	**156**	**110**	**591**

1) *Planteo de Hipótesis a probar*: "Las medias por tratamiento son iguales". (Consideramos así, que el efecto de las raciones es nulo).

2) *Cálculo de las Sumas de cuadrado* y *completado de la Tabla de Análisis de la Varianza*.

$$C = \frac{\left(\sum_{i,j} Y_{ij}\right)^2}{N} = \frac{(591)^2}{4 \times 5} = 17464{,}05$$

$$SCT = \sum_{i,j} Y_{ij}^2 - C = (35^2 + ... + 30^2 + ... + 27^2 + ... + 30^2) - 17464{,}05 = 1960{,}95$$

$$SCE = \frac{Y_1^2 + \cdots + Y_t^2}{r} - C = \frac{130^2 + ... + 110^2}{5} - 17464{,}05 = 808{,}15$$

$$SCD = SCT - SCE = 1960{,}95 - 808{,}15 = 1152{,}8$$

Fuentes de Variación	***Suma de cuadrados***	***gl***	***Cuadrados Medios***	***F Calculado***
Entre Tratamiento	808,15	3	269,38	3,74
Dentro Error Experimental	1152,8	16	72,05	
Total	1960,95	19		

3) *Obtención del F tabulado*

Para un α = 0,05; gle = 3 y gld = 16 a partir de la tabla pagina 91 extraemos el valor de F tabulado, en este caso es de **3,24**

4) *Conclusión parcial:* Como el valor que corresponde al **F calculado** (3,74) es **mayor** que el correspondiente al **F tabulado** (3,24) , la **hipótesis nula es rechazada**, por lo tanto podemos concluir que existen diferencias significativas entre medias. Para detectar cuales difieren significativamente y cuales solo por el azar aplicaremos el Test de Tukey y el de Duncan para detectar tales diferencias.

5) *Comparaciones múltiples de medias*

Test de Tukey (para un nivel de significancia del 5%)

Recordar que este test nos permite comparar las medias tomadas dos a dos, es decir de a pares. Como punto de partida debemos calcular la diferencia mínima significativa (d.m.s) como:

$$d.m.s_{5\%} = \Delta_{5\%} = q_{5\%} \cdot \frac{S}{\sqrt{r}} = 4{,}05 \frac{\sqrt{72{,}05}}{\sqrt{5}} = 15{,}374$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (5%), del número de tratamientos (4 en este caso) y de los grados de libertad del error (para este caso gld =16).

La diferencia mínima significativa (d.m.s) calculada por este test es comparada con la diferencia entre las medias de los tratamientos tomados de a pares. Cuando la diferencia entre las medias no supera la d.m.s se concluye que ambas medias no difieren significativamente y que solo difieren por el azar, mientras que si la diferencia entre las medias supera la d.m.s podemos concluir que ambas medias difieren significativamente, es decir que la diferencia entre ambas medias es debido a un efecto del tratamiento.

Previo a las conclusiones por este test, para facilitar las comparaciones se ordenan los tratamientos según sus medias, de menor a mayor.

Conclusiones de este test:

Tratamiento	*Media*	*Significancia*	*Significancia*
D	22	X	a
A	26	X X	a b
C	31,2	X X	a b
B	39	X	b

Test de Duncan (nivel de significancia del 5%)

A diferencia del test de Tukey este test compara medias de a grupos. Por lo tanto para este caso que contamos con 4 medias necesitamos calcular 3 d.m.s. (una para comparar grupos de dos medias, otra para comparar grupos de tres medias y la última para comparar grupos de 4 medias). La expresión general para el cálculo de las d.m.s es:

$$d.m.s._{5\%} = D_{5\%} = Z_{5\%} \cdot \frac{S}{\sqrt{r}}$$

Donde: Z se obtiene de la tabla de la pagina 94 en función del nivel de significancia (5%), del número de tratamientos que forman el grupo (2, 3 o 4 según corresponda) y de los grados de libertad del error (para este caso gld =16).

Los valores de las d.m.s. para esta prueba se resumen en la siguiente tabla:

Número de medias que forman el grupo	$Z_{5\%}$	*d.m.s*
2	3,00	$D_2 = 11,39$
3	3,15	$D_3 = 11,96$
4	3,23	$D_4 = 12,26$

El paso siguiente es ordenar los tratamientos según sus medias, de menor a mayor, para luego comparar la d.m.s correspondientes al grupo que contiene todas las medias (D_4) contra la diferencia que se obtiene entre la media mayor y la menor del grupo, si esta comparación es no significativa el análisis concluye, por el contrario, si resulta significativa se continua con las comparaciones de los grupos formados por tres medias y así sucesivamente. La significancia o no de las comparaciones se pueden detallar en forma semejante al test de Tukey.

Conclusiones de este test:

Tratamiento	*Media*	*Significancia*	*Significancia*	*Significancia*
D	22	X	a	
A	26	X	a	
C	31,2	X X	a b	
B	39	X	b	

Nota: Como se puede ver al comparar el test de Tukey con el de Duncan, este último es menos riguroso, esto es, da diferencias significativas con más facilidad que el primero.

6) *Conclusión final:*

Según el test de Tukey la ración D es la única que presenta diferencias significativas con la B, por lo que se puede concluir que la ración D es la que genera menor aumento de Peso; en cambio como las diferencias entre las otras 3 raciones (A, B y C) no son significativas, es decir son debidas al azar, se puede considerar que cualquiera de ellas generan un mayor e idéntico aumento de peso.

Para el test de Duncan las raciones B y C son las que producen el mayor aumento de peso, siendo la diferencia entre ellas debido al azar.

Diseño en Bloques Completos Al Azar (DBCA)

Este diseño consiste en dividir el material experimental (unidades experimentales) en grupos homogéneos denominados **bloques**, cada uno de los cuales constituye una repetición. La eficiencia del diseño descansa en la homogeneidad de las unidades que constituyen un bloque, **no importando la heterogeneidad entre ellos**.

El número de parcelas serán tantas, dentro de un bloque, como tratamientos se prueben y así como lo permitan las condiciones de homogeneidad del medio. (Suelo, animales, árboles, etc.). Su tamaño será función del tipo de material a experimentar. Sin embargo no está de más tener en mente lo siguiente:

a) Evitar el empleo de parcelas grandes, siendo preferible aumentar el número de repeticiones.

b) Evitar trabajar con numerosas parcelas por bloque.

La disposición de los bloques dependerá de las características topográficas del terreno y de la existencia de gradientes de alguna índole (de fertilidad, de drenaje, de exposición, etc.). Siempre la ubicación de los bloques deberá ser perpendicular al gradiente considerado.

En experimentos zootécnicos los bloques serán constituidos con animales de características semejantes, por ejemplo, si estudiamos en vacas lecheras su respuesta a determinadas raciones, los bloques serán formados con animales de edades, épocas de parición y producción de leche similares.

El modelo matemático del diseño es:

$$Y_{ij} = \mu + t_i + b_j + \varepsilon_{ij}$$

Y_{ij} = Observación del tratamiento i repetición j.

μ = Media general del ensayo.
t_i = Efecto del tratamiento i.
b_j = Efecto del bloque j.
ε_{ij} = Contribución del azar (factores no controlados).

Es decir que aparece una **nueva fuente de variación** respecto al diseño completamente aleatorizado que es la debida a **bloques**.

Un diseño en bloques completos al azar con 6 tratamientos y 4 repeticiones se muestra a continuación:

C	**F**	**D**
A	**B**	**E**

D	**E**	**B**
F	**C**	**A**

E	**D**	**F**
B	**C**	**A**

D	**F**	**A**
C	**E**	**B**

El número de parcelas de cada bloque es igual al número de tratamientos (Por ello es COMPLETO) y la distribución de los tratamientos dentro de los bloques se realiza al azar.

La **Tabla de Análisis de la Varianza** (Tabla 2) es similar al realizado para un Diseño Completamente aleatorizado **(DCA)**, con una nueva fuente de variación debida al efecto de los Bloques.

Los cálculos de las sumas de cuadrado son similares que para un DCA. Para la S*uma de cuadrado de bloque (SCB)* podemos utilizar la siguiente expresión:

$$SCB = \frac{B_1^2 + \cdots + B_r^2}{t} - C$$

Siendo: B_1 hasta B_r los totales de los bloques, desde el bloque 1 hasta el bloque *r*.

Suma de cuadrado dentro o Suma de cuadrado del error (SCD)

$$SCD = SCT - SCE - SCB$$

Fuentes de Variación	***Suma de Cuadrados***	***Grados de Libertad***	***Cuadrado Medio***	***F Calculado***
Entre Tratamientos	*SCE*	*gle = t – 1*	$CME = \frac{SCE}{gle}$	$F = \frac{CME}{CMD}$
Bloques	*SCB*	*glb = r – 1*	$CMB = \frac{SCB}{\text{glb}}$	
Dentro Error Experimental	*SCD=SCT–SCE–SCB*	*gld =(t -1)* x *(r -1)*	$CMD = \frac{SCD}{gld}$	
Total	*SCT*	*glt = N –1*		

Tabla 2 **(Donde: *t* = número de tratamientos; *r* = número de bloques; N = *t* x *r*)**

El procedimiento del **test de hipótesis de nulidad** es similar al realizado para un diseño completamente aleatorizado. En las **"Comparaciones múltiples de Medias"** no hay que olvidar que el número de bloques sustituye el número de repeticiones en el cálculo de las diferencias mínimas significativas (d.m.s), después el procedimiento es similar.

Ejemplo de Aplicación:

En un ensayo comparativo de rendimientos de 8 variedades de papa conducido en Balcarce, en bloques al azar con 4 repeticiones, las producciones obtenidas, en tn/ha, fueron las siguientes:

	Bloques				Totales *Variedad*
Variedades	**1**	**2**	**3**	**4**	
Kennebec (1)	9,2	13,4	11,0	9,2	**42,8**
Huinkul (2)	22,1	27,0	26,4	25,7	**101,2**
Sta. Rafaela (3)	22,6	29,9	24,2	25,1	**101,8**
Buena Vista (4)	15,4	11,9	10,1	12,3	**49,7**
B 25-50 E (5)	12,7	18,0	18,2	17,1	**66**
B 1-52 (6)	20,0	21,1	20,0	28,0	**89,1**
B 116-51 (7)	23,1	24,2	26,4	16,3	**90**
B 72-53 A (8)	18,0	24,6	24,0	24,6	**91,2**
Totales Bloque	**143,1**	**170,1**	**160,3**	**158,3**	**631,8**

(*) Abreviación de las variedades para facilitar la interpretación.

1) *Planteo de Hipótesis a probar*: "Los rendimientos medios de las variedades son iguales". (Consideramos así, que el efecto de las variedades es nulo).

2) *Cálculo de las Sumas de cuadrado* y *completado de la Tabla de Análisis de la Varianza*.

$$C = \frac{\left(\sum_{i,j} Y_{ij}\right)^2}{N} = \frac{(631{,}8)^2}{8 \times 4} = 12474{,}10$$

$$SCT = \sum_{i,j} Y_{ij}^2 - C = (9{,}2^2 + ... + 13{,}4^2 + ... + 11^2 + ... + 9{,}2^2 + ... + 24{,}6^2) - 12474{,}10 = 1153{,}46$$

$$SCE = \frac{Y_1^2 + \cdots + Y_t^2}{r} - C = \frac{42{,}8^2 + ... + 91{,}2^2}{4} - 12474{,}10 = 930{,}61$$

$$SCB = \frac{B_1^2 + \cdots + B_r^2}{t} - C = \frac{143{,}1^2 + ... + 158{,}3^2}{8} - 12474{,}10 = 46{,}72$$

$$SCD = SCT - SCE - SCB = 1153{,}46 - 930{,}61 - 46{,}72 = 176{,}13$$

Fuentes de Variación	***Suma de cuadrados***	***gl***	***Cuadrados Medios***	***F Calculado***
Entre Tratamientos	930,61	7	132,95	15,85
Bloques	46,72	3	15,57	
Dentro Error Experimental	176,13	21	8,39	
Total	1153,46	31		

3) *Obtención del F tabulado*

Para un α = 0,05; gle = 7 y gld = 21 a partir de la tabla pagina 91 extraemos el valor de F tabulado, en este caso es de **2,49**

4) *Conclusión parcial:* Como el valor que corresponde al **F calculado** (15,85) es **mayor** que el correspondiente al **F tabulado** (2,49) , la **hipótesis nula es rechazada**, por lo tanto podemos concluir que existen diferencias significativas entre medias. Para detectar cuales difieren significativamente y cuales solo por el azar aplicaremos el Test de Tukey para detectar tales diferencias.

5) *Comparaciones múltiples de medias*

Test de Tukey (para un nivel de significancia del 5%)

$$d.m.s_{5\%} = \Delta_{5\%} = q_{5\%} \cdot \frac{S}{\sqrt{r}} = 4{,}745 \frac{\sqrt{8{,}39}}{\sqrt{4}} = 6{,}87$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (5%), del número de tratamientos (8 en este caso) y de los grados de libertad del error (para este caso gld =21).

Para facilitar las comparaciones se ordenan los tratamientos (Variedades) según sus rendimientos medios, de menor a mayor.

Conclusiones de este test:

Tratamiento	*Media*	*Significancia*	*Significancia*
1	10,7	X	a
4	12,425	X	a
5	16,5	X X	a b
6	22,275	X X	b c
7	22,5	X X	b c
8	22,8	X X	b c
2	25,3	X	c
3	25,45	X	c

6) *Conclusión final:*

Según el test de Tukey se puede observar que existe un grupo de variedades de alta producción (Sta. Rafaela (3), Huinkul (2), B 72-53 (8), B 116-51 (7), B 1-52 (6)) que se destacan nítidamente de las demás, desde el punto de vista estadístico.

Diseño en cuadrado latino (CL)

Este diseño es aconsejado en suelos heterogéneos pero de topografía plana, su nombre es accidental ya que originalmente la representación de los tratamientos se realizaba en letras latinas.

Es un diseño en el cual el número de parcelas es un cuadrado perfecto, obteniéndose al elevar al cuadrado el número de tratamientos; éstos se distribuyen en hileras y columnas, existiendo tantas hileras y columnas como tratamientos hay en un estudio. Cada hilera y columna contiene a todos los tratamientos repetidos una sola vez.

El modelo matemático del diseño es:

$$Y_{ijk} = \mu + t_i + c_j + h_k + \varepsilon_{ij}$$

$Y_{i\,j\,k}$ = Observación del tratamiento i, columna j y hilera k.
μ = Media general del ensayo.
t_i = Efecto del tratamiento i.
c_j = Efecto de la columna j.
h_k = Efecto de la hilera k.
$\varepsilon_{i\,j\,k}$ = Contribución del azar (factores no controlados).

Es un diseño menos flexible que el de bloques al azar, el menor cuadrado latino es el de 5 x 5, ya que los grados de libertad del error experimental deberán ser como mínimo 12

La presencia de hileras y columnas permite eliminar la heterogeneidad del suelo en dos direcciones.

Originado para solucionar problemas en la experiencia a campo, es bastante utilizado también en la experimentación industrial, zootecnia, etc.

Una limitación seria de este tipo de diseño es que el número de columnas y de hileras es igual al de tratamientos por lo que, cuando el número de tratamientos es grande, el cuadrado latino se torna impracticable.

Experimentos Factoriales

Se denominan *Experimentos Factoriales*, a aquellos ensayos en los que se *estudian simultáneamente dos o más factores*, diferenciándose de los Experimentos Simples, en los cuales se contempla el estudio de un solo factor, llámense plaguicidas, híbridos, raciones, dosis de nitrógeno, dosis de fósforo, etc.

Estos experimentos se usan en diversos campos y son de gran valor en trabajos exploratorios cuando se sabe muy poco acerca del comportamiento de numerosos factores que inciden por ejemplo en un cultivo, por ejemplo que variedad es la más rendidora para una región particular, las respuestas de las mismas a distintas densidades de siembra y/o a las épocas de siembra, etc.; con una sola densidad y una sola época de siembra podría suceder que la densidad de siembra escogida o la época de siembra no sea la apropiada para una o para todas las variedades con lo cual los resultados no serían los correctos.

En los experimentos factoriales todo factor (elemento) a estudiar proporcionara diferentes estados o cantidades. Estas cantidades o estados diferentes de un factor se denominan *niveles*.

Así, un experimento factorial es aquel en el que el conjunto de tratamientos consiste en todas las combinaciones posibles de los niveles de varios factores. Entendiéndose en estos experimentos que cada combinación resultante se considera un tratamiento.

Por ejemplo, si deseamos establecer un ensayo para estudiar el comportamiento de 5 variedades de soja a tres densidades de siembra, el experimento se llama *Experimento Factorial 5 x 3*, con 5 niveles del factor variedad y 3 niveles del factor densidad, el cual constará de 15 tratamientos, los cuales surgen de la combinación de las 5 variedades con las 3 densidades de siembra. En símbolos (V = variedades , d = densidades)

V1d1	V2d1	V3d1	V4d1	V5d1
V1d2	V2d2	V3d2	V4d2	V5d2
V1d3	V2d3	V3d3	V4d3	V5d3

Si en un experimento factorial participan, como tratamientos, todas las combinaciones posibles de los factores, se los denomina *Factoriales Completos* caso contrario se los denomina *Factoriales Incompletos.*

La información que se obtiene del análisis de los experimentos factoriales es más amplia que la producida por ensayos simples debido a que permite analizar el efecto de cada uno de los factores y detectar posibles interacciones entre ellos.

El efecto de un factor lo podemos definir como el cambio en la respuesta de la variable analizada producido por un cambio en el nivel del factor. A esto se lo denomina ***efecto***

principal del factor. Por ejemplo, considérense los datos de la Tabla 1 que corresponden al rendimiento de dos variedades de un cultivo sometido a dos épocas de siembra.

	Factor B (época)	
Factor A (variedad)	B_1	B_2
A_1	20	30
A_2	50	10

Tabla 1: Considerar A_1 y A_2 niveles del factor variedad y, B_1 y B_2 niveles del factor época

El efecto principal del factor A es la diferencia entre la respuesta promedio en el primer nivel de de A y la respuesta promedio en el segundo nivel de A, es decir:

$$A = \frac{50+10}{2} - \frac{20+30}{2} = 5$$

Esto es, cambiar el factor A del nivel 1 al nivel 2 (cambiar la variedad A_1 por la A_2) provoca un incremento de 5 unidades en el rendimiento promedio del cultivo. De manera semejante, el efecto principal de B es:

$$B = \frac{30+10}{2} - \frac{20+50}{2} = -15$$

En este caso, cambiar el factor B del nivel 1 al nivel 2 (cambiar la época B_1 por la B_2) provoca una disminución de 15 unidades en el rendimiento promedio del cultivo
Si la diferencia en respuesta entre los niveles de un factor no es la misma en todos los niveles del otro factor, se dice que existe ***interacción*** entre los factores. Por ejemplo, analizando los datos de la Tabla 1, vemos que en el primer nivel del factor A (A_1) el efecto de B es

$$B = 30 - 20 = 10$$

y en el segundo nivel del factor A (A_2) , el efecto de B es

$$B = 10 - 50 = -40$$

Puesto que el efecto de B depende del nivel seleccionado para el factor A, se dice que existe interacción entre A y B. Es decir que el efecto de la época de siembra sobre el rendimiento del cultivo no es el mismo para las variedades A_1 y A_2, mientras que con la variedad A_1 se produce un aumento promedio de 10 unidades en el rendimiento, con la variedad A_2 se produce una disminución promedio de 40 unidades en el mismo.

Veamos ahora una situación hipotética en donde no hay interacción entre los factores. Para ello analicemos los datos de la Tabla 2

	Factor B (época)	
Factor A (variedad)	B_1	B_2
A_1	20	30
A_2	50	60

Tabla 2

Para esta situación en el primer nivel del factor A (A_1) el efecto de B es

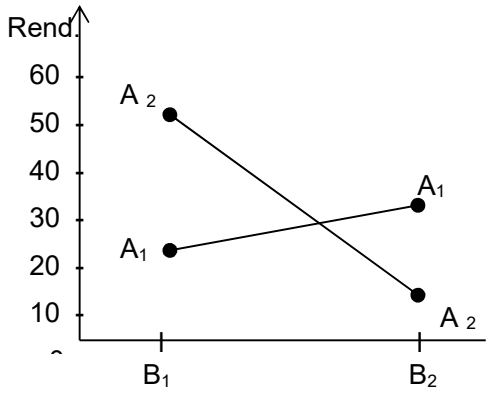
$$B = 30 - 20 = 10$$

y en el segundo nivel del factor A (A_2) , el efecto de B es

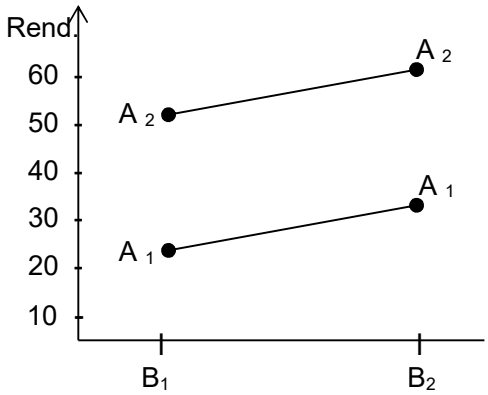
$$B = 60 - 50 = 10$$

Puesto que el efecto de B no depende del nivel seleccionado para el factor A, se dice que no existe interacción entre A y B. Es decir que el efecto de la época de siembra sobre el rendimiento del cultivo es el mismo para ambas variedades.

El concepto de interacción puede ilustrarse gráficamente, lo cual puede ser una rápida y útil aproximación a los resultados de un experimento. Sin embargo no debe ser la única técnica para analizar los datos, porque su interpretación es subjetiva y su apariencia, a menudo, puede resultar engañosa. A continuación se muestran las gráficas de la respuesta de los datos de las Tabla 1 y 2 para los niveles del factor B para ambos niveles del factor A.



Gráfica 1: Corresponde a Tabla 1



Gráfica 2: Corresponde a Tabla 2

En la gráfica 1 se observa que las rectas A_1 y A_2 no son paralelas. Esto muestra una interacción entre los factores A y B. En cambio en la gráfica 2 se observa que las rectas A_1 y A_2 son paralelas. Esto indica que no hay interacción entre los factores. No siempre estas gráficas son como las presentadas hay otras situaciones intermedias como pueden ser rectas no paralelas pero que no llegan a cruzarse.

Los diseños generalmente utilizados en el montaje de un experimento factorial van desde los diseños más simples como lo son el diseño completamente aleatorizado (DCA)

y el diseño en bloques completos al azar (DBCA), hasta diseños más complejos como son en parcela dividida (DPD), diseños en franja (DF) o en bloque divididos (DBD), etc.

Elección de factores

Hemos expresado anteriormente que en un experimento factorial podemos considerar a un factor como un elemento conformado por diferentes niveles relacionados entre si o que pertenecen a una misma clasificación. Por ejemplo: un factor antibiótico se puede presentar en tres dosis (0 mg, 3 mg, 6 mg.), un factor raza puede estar constituido por razas Landrace, Duroc yersey, Poland china, etc.

No es recomendable trabajar con más de tres factores en los ensayos especialmente agronómicos ya que las interacciones de orden triple o mayores en general no presentan significancia estadística especialmente en ensayos de fertilidad.

Los factores que se estudian en los experimentos se pueden clasificar en:

- *Factores cuantitativos:* presentan la característica que sus niveles son cantidades numéricas de una variable cuantitativa. Estos factores son frecuentes en química, agricultura, medicina, física, etc. Ejemplos: dosis crecientes de nitrógeno, temperaturas crecientes, presiones, cantidades crecientes en la concentración de un reactivo, etc.

- *Factores cualitativos:* no presentan un orden natural establecido y cada nivel tiene un interés intrínseco. Por ejemplo: variedades de trigo, técnicas en la poda de árboles, etc. También pertenecen a este grupo los factores constituidos por categorías, por ejemplo: diferentes clases de lana, diferentes laboratorios que participan en ensayos cada uno con sus equipos, material, personal, etc.

También podemos considerar en esta clasificación aquellos factores cuyos niveles aunque no se expresen cuantitativamente presentan un cierto orden; por ejemplo: un rodeo de vacas que presenta mastitis pueden haber sido clasificadas en: ligeramente atacadas, moderadamente atacadas, muy atacadas, severamente atacadas.

Ventajas de los Experimentos Factoriales

- Permite el estudio de los efectos principales, los de interacción, y los efectos simples.
- Permite que todas las unidades intervengan en la determinación de los efectos principales y en la interacción de los factores.
- El número de grados de libertad del error es elevado contribuyendo a disminuir la varianza del error y aumentando la precisión del ensayo.

Desventajas de los Experimentos Factoriales

- Requiere un mayor número de unidades experimentales lo que torna compleja la conducción del ensayo. Por ej. un factorial de 4^3 implica la presencia de 3 factores cada

uno con cuatro niveles, las combinaciones entre los niveles de un factor con cada nivel de los restantes factores generan 64 tratamientos; si el ensayo tuviese 4 repeticiones tendremos 256 parcelas.

- En los ensayos factoriales completos se deben combinar todos los niveles de cada factor entre si a los efectos que el análisis sea balanceado, el resultado es que a veces algunas combinaciones no tienen ningún interés práctico, como se observa en el siguiente ejemplo: tenemos el factor labranza, sin labranza (labranza 0), con una labranza (labranza 1), con dos labranzas (labranza 2) y el factor encalado, testigo (0 tn), 4 tn y 8 tn. Las combinaciones correspondientes a encalado sin labranza no tienen ningún interés práctico ya que la cal quedaría sin enterrar.

A continuación presentaremos sintéticamente la metodología de análisis de los experimentos factoriales de dos factores en los diseños más simples.

Experimentos Factoriales de dos factores en un Diseño Completamente Aleatorizado (DCA)

El **modelo lineal** al cual responde un experimento factorial de 2 factores (A y B) montado en un Diseño Completamente Aleatorizado (DCA) es el siguiente:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

Donde:

Y_{ijk} = Observación bajo el i-ésimo nivel del factor A, j-ésimo nivel del factor B, repetición k.
μ = Media general del ensayo.
τ_i = Efecto del tratamiento A.
β_j = Efecto del tratamiento B .
$(\tau\beta)_{ij}$ = Efecto de la interacción (A x B).
ε_{ijk} = Error experimental (factores no controlados).

La **Tabla de Análisis de la Varianza** para un experimento factorial montado en un DCA con r repeticiones por tratamiento (Tabla 3) es similar a la de un experimento simple (unifactorial), con la disgregación de los tratamientos en efectos principales (A y B) e interacción entre los factores (A x B). Es decir con la partición de los grados de libertad y de las suma de cuadrados de tratamientos en las componentes atribuibles a cada efecto principal y a las interacciones.

Fuentes de Variación	***Suma de Cuadrados***	***Grados de Libertad***	***Cuadrado Medio***	***F Calculado***
Factor A	SC_A	$gl_A = a - 1$	$CM_A = \frac{SC_A}{gl_A}$	$F = \frac{CM_A}{CMD}$
Factor B	SC_B	$gl_B = b - 1$	$CM_B = \frac{SC_B}{gl_B}$	$F = \frac{CM_B}{CMD}$
Interacción *(A x B)*	SC_{AxB}	$gl_{AB} = (a - 1)(b - 1)$	$CM_{AB} = \frac{SC_{AB}}{gl_{AB}}$	$F = \frac{CM_{AB}}{CMD}$
Error Experimental	SCD	$gl_E = ab(r - 1)$	$CMD = \frac{SCD}{gl_E}$	
Total	SCT	$glt = abr - 1$		

Tabla 3**: (Donde: a = niveles del factor A; b = niveles del factor B; r = número de repeticiones)**

Para el cálculo de las sumas de cuadrado se pueden utilizar las siguientes expresiones:

Suma de cuadrado total (SCT)

$$SCT = \sum_{i,j,k} Y_{ijk}^2 - C = \sum_{i,j,k} Y_{ijk}^2 - \frac{\left(\sum_{i,j,k} Y_{ijk}\right)^2}{N}$$

Siendo: Y_{ijk} Observación bajo el i-ésimo nivel del factor A, i = 1,2, …, a; j-ésimo nivel del factor B, j = 1,2, …, b; repetición k-ésima, k = 1,2, …, r.

C = termino corrector o factor de corrección.

Suma de cuadrado para el factor A (SC_A)

$$SC_A = \frac{Y_1^2 + \cdots + Y_a^2}{rb} - C$$

Siendo: Y_1 hasta Y_a los totales de los niveles del factor A desde el nivel 1 hasta el nivel a.

Suma de cuadrado para el factor B (SC_B)

$$SC_B = \frac{Y_1^2 + \cdots + Y_b^2}{ra} - C$$

Siendo: Y_1 hasta Y_b los totales de los niveles del factor B desde el nivel 1 hasta el nivel *b*.

Suma de cuadrado para la interacción AxB (SC_{AxB})

$$SC_{AxB} = SCE - SC_A - SC_B$$

Siendo: SCE la suma de cuadrado entre o suma de cuadrado de tratamientos

Suma de cuadrado entre o Suma de cuadrado de tratamientos (SCE)

$$SCE = \frac{Y_{11}^2 + \cdots + Y_{ab}^2}{r} - C$$

Siendo: Y_{11} hasta Y_{ab} los totales de las combinaciones (tratamientos), desde la combinación 11 hasta la combinación *ab*.

Suma de cuadrado dentro o Suma de cuadrado del error (SCD)

$$SCD = SCT - SC_A - SC_B - SC_{AxB}$$

Las **Hipótesis que se someten a prueba** son:

Hipótesis 1: *"No hay diferencias entre los niveles del factor A"*

Hipótesis 2: *"No hay diferencias entre los niveles del factor B"*

Hipótesis 3: *"No hay interacción entre los factores (independencia)"*

La primer hipótesis que debemos contrastar es la 3 por que si se rechaza (F calculado > al F tabulado de la Interacción) quiere decir que *los efectos principales* (efectos de A y B) *no son independientes* es decir que hay interacción de los factores, por lo que no deben tenerse en cuenta los efectos principales (ya que no hay independencia de los mismos). La interacción significa que el efecto de un factor no es el mismo para los diferentes niveles del otro.

Por otro lado si la hipótesis 3 no se rechaza (F calculado < al F de la interacción) nos indica que los *efectos principales son independientes* (no hay interacción), por lo tanto en este caso el paso siguiente es contrastar las hipótesis 1 y 2 para ver que pasa con los efectos principales de A y de B.

Experimentos Factoriales de dos factores en un Diseño en Bloques Completos al Azar (DBCA)

El **modelo lineal** es similar a la de un DCA con la inclusión de una nueva fuente de variación debida al Bloque. Por lo tanto el modelo al cual responde un experimento factorial de 2 factores (A y B) montado en un DBCA es el siguiente:

$$Y_{ijk} = \mu + \lambda_k + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

Donde:

γ_k = Efecto debido a los bloques

La **Tabla de Análisis de la Varianza** también es similar a la de un DCA con el agregado de una nueva fuente de variación provocada por los bloques (Tabla 4).

Fuentes de Variación	***Suma de Cuadrados***	***Grados de Libertad***	***Cuadrado Medio***	***F Calculado***
Bloque	*SCB*	*glb = r – 1*	$CMB = \frac{SCB}{\text{glb}}$	
Factor A	SC_A	$gl_A = a - 1$	$CM_A = \frac{SC_A}{gl_A}$	$F = \frac{CM_A}{CMD}$
Factor B	SC_B	$gl_B = b - 1$	$CM_B = \frac{SC_B}{gl_B}$	$F = \frac{CM_B}{CMD}$
Interacción *(A x B)*	SC_{AxB}	$gl_{AB} = (a - 1)(b - 1)$	$CM_{AB} = \frac{SC_{AB}}{gl_{AB}}$	$F = \frac{CM_{AB}}{CMD}$
Error Experimental	*SCD*	$gl_E = (ab - 1)(r - 1)$	$CMD = \frac{SCD}{gl_E}$	
Total	*SCT*	*Glt = abr –1*		

Tabla 4**: (Donde: a = niveles del factor A; b = niveles del factor B; r = número de bloques)**

Los cálculos de las sumas de cuadrado para el total, para los factores (A y B) y para la interacción (AxB) son similares que para un DCA. Para la S*uma de cuadrado de bloque (SCB)* podemos utilizar la siguiente expresión:

$$SCB = \frac{B_1^2 + \cdots + B_r^2}{ab} - C$$

Siendo: B_1 hasta B_r los totales de los bloques, desde el bloque 1 hasta el bloque *r*.

Suma de cuadrado dentro o Suma de cuadrado del error (SCD)

$$SCD = SCT - SCB - SC_A - SC_B - SC_{AxB}$$

Comparaciones Múltiples de Medias

Test de Tukey

Cuando *los efectos principales son independientes* (es decir, no hay interacción significativa entre los factores) y la prueba de F fuera significativa para algunos de los factores o para ambos, debemos analizar el efecto que producen los distintos niveles del factor o de los factores, respectivamente, sobre la variable analizada. En estos casos la forma que toma la expresión del cálculo de la *d.m.s* para la prueba de Tukey es la siguiente:

1) Para comparar los efectos principales del factor A

$$\Delta = q_{(\alpha;gl_E;a)} \cdot \frac{S}{\sqrt{rb}}$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (generalmente 5%), del número de niveles del factor A y de los grados de libertad del error.

2) Para comparar los efectos principales del factor B

$$\Delta = q_{(\alpha;gl_E;b)} \cdot \frac{S}{\sqrt{ra}}$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (generalmente 5%), del número de niveles del factor B y de los grados de libertad del error.

Por otro lado cuando *los efectos principales no son independientes* (es decir, hay interacción significativa entre los factores) el análisis de los efectos principales no corresponde. En su lugar una forma posible de analizar los datos es comparar todos los tratamientos de a pares por el Test de Tukey mediante la siguiente expresión:

$$\Delta = q_{(\alpha;gl_E;ab)} \cdot \frac{S}{\sqrt{r}}$$

Donde: q se obtiene de la tabla de la correspondiente en función del nivel de significancia (generalmente 5%), del número de tratamientos y de los grados de libertad del error.

Ejemplo de aplicación sin interacción significativa entre los factores

Deseamos probar dos dosis de nitrógeno y dos de fósforo en un cultivo de maíz. Este ensayo presenta dos factores (N y P), N con dos niveles (0,1) y P con dos niveles (0,1) se lo denomina factorial de 2 x 2 o 2^2 , con 4 (cuatro) tratamientos.

Siempre el exponente es el número de factores y la base el número de niveles de cada factor.

Supongamos que el diseño del ensayo es un D.B.C.A. con 5 repeticiones

Bloques	***Tratamientos***				***Totales de Bloque***
	N_0P_0	**N_1P_0**	**N_0P_1**	**N_1P_1**	
I	1,00	1,50	3,20	3,80	9,50
II	1,60	2,30	4,50	5,00	13,40
III	1,20	1,10	5,60	6,00	13,90
IV	1,30	1,40	5,50	6,20	14,40
V	1,30	1,60	4,40	4,80	12,10
Totales Tratamientos	6,40	7,90	23,20	25,80	63,30
Media Tratamientos	1,28	1,58	4,64	5,16	

1) *Planteo de Hipótesis que se someterán a prueba*:

Hipótesis 1: No hay diferencias de rendimiento entre los dos niveles de nitrógeno (N0 = N1)

Hipótesis 2: No hay diferencias de rendimiento entre los dos niveles de fósforo (P0 = P1)

Hipótesis 3: No hay interacción entre los factores (independencia).

2) *Cálculo de las Sumas de cuadrado* y *completado de la Tabla de Análisis de la Varianza*.

Suma de cuadrado del Total

$$C = \frac{\left(\sum_{i,j,k} Y_{ijk}\right)^2}{N} = \frac{(63,3)^2}{2x2x5} = 200,3445$$

$$SCT = \sum_{i,j,k} Y_{ijk}^2 - C = \left(1,00^2 + \cdots + 1,50^2 + \cdots + 3,20^2 + \cdots + 3,80^2\right) - 200,3445 = 69,6855$$

Para facilitar los cálculos posteriores creamos una tabla de doble entrada como la siguiente:

	N_0	**N_1**	*Totales*
P_0	6,40	7,90	14,30
P_1	23,20	25,80	49,00
Totales	29,60	33,70	63,30

Suma de cuadrado para el factor A (SC_A)

$$SC_A = \frac{Y_1^2 + \cdots + Y_a^2}{rb} - C = \frac{\left(29{,}60^2 + 33{,}70^2\right)}{5x2} - 200{,}3445 = 0{,}8405$$

Suma de cuadrado para el factor B (SC_B)

$$SC_B = \frac{Y_1^2 + \cdots + Y_b^2}{ra} - C = \frac{\left(14{,}30^2 + 49{,}00^2\right)}{5x2} - 200{,}3445 = 60{,}2045$$

Suma de cuadrado entre o Suma de cuadrado de tratamientos (SCE)

$$SCE = \frac{Y_{11}^2 + \cdots + Y_{ab}^2}{r} - C = \frac{\left(6{,}4^2 + 7{,}90^2 + 23{,}20^2 + 25{,}80^2\right)}{5} - 200{,}3445 = 61{,}1055$$

Suma de cuadrado para la interacción AxB (SC_{AxB})

$$SC_{AxB} = SCE - SC_A - SC_B = 61{,}1055 - 0{,}8405 - 60{,}2045 = 0{,}0605$$

Suma de cuadrado para Bloques (SCB)

$$SCB = \frac{B_1^2 + \cdots + B_r^2}{ab} - C = \frac{\left(9{,}5^2 + 13{,}40^2 + 13{,}90^2 + 14{,}40^2 + 12{,}10^2\right)}{2x2} - 200{,}3445 = 3{,}853$$

Suma de cuadrado dentro o Suma de cuadrado del error (SCD)

$$SCD = SCT - SCB - SC_A - SC_B - SC_{AxB} = 69{,}6855 - 3{,}853 - 0{,}8405 - 60{,}2045 - 0{,}0605 = 4{,}727$$

Fuentes de Variación	***Suma de cuadrados***	***gl***	***Cuadrados Medios***	***F Calculado***
Bloque	3,853	4	0,963	
Factor A (N)	0,8405	1	0,8405	2,134
Factor B (P)	60,2045	1	60,2045	152,836 *
Interacción (N x P)	0,0605	1	0,0605	0,154
Error Experimental	4,727	12	0,3939	
Total	69,6855	19		

3) *Obtención de los F tabulados*

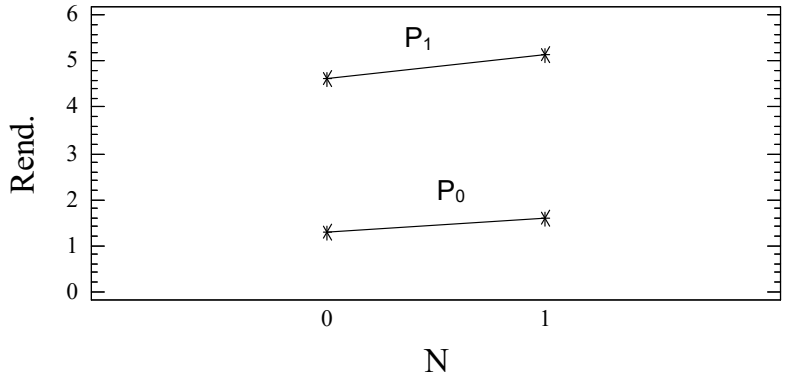
Fuentes de Variación	*gl del numerador*	*gl del denominador*	*Nivel de significancia*	*F Tabulado*
Nitrógeno	1	12	0,05	4,75
Fósforo	1	12	0,05	4,75
Interacción	1	12	0,05	4,75

4) *Conclusión parcial:* Como el valor que corresponde al **F calculado para la interacción** (0,154) es **menor** que el correspondiente al **F tabulado** (4,75), la **hipótesis nula que plantea la no interacción de los factores no es rechazada** (hipótesis 3), por lo tanto podemos concluir que **los efectos principales son independientes**, lo cual nos permite analizar los efectos principales.

Para el **factor Nitrógeno** la **hipótesis nula** que plantea la igualdad de rendimiento para ambos niveles de dicho factor (hipótesis 1) **no es rechazada**, es decir que el **efecto que provoca el nitrógeno sobre el rendimiento del maíz no es significativo**.

En cambio para el **factor Fósforo** la **hipótesis nula** que plantea la igualdad de rendimiento para ambos niveles de dicho factor **es rechazada**, es decir que en este ensayo **se comprueba que hay un efecto significativo del fósforo sobre el rendimiento del maíz**. Para detectar cuales niveles del factor fósforo difieren significativamente y cuales solo por el azar aplicaremos el Test de Tukey. Antes observaremos y analizaremos el gráfico de interacción.

5) *Gráfico de Interacción*



El paralelismo entre las rectas P_1 y P_0 nos indica la no interacción de los factores nitrógeno y fósforo, lo cual queda reflejado al observar que el efecto que provoca el nitrógeno es el mismo tanto en presencia como en ausencia del fósforo. Por otro lado este gráfico nos manifiesta la posible superioridad de rendimiento del cultivo de maíz cuando fertilizamos con Fósforo mas allá de la fertilización o no con nitrógeno.

6) *Comparaciones múltiples de medias*

La comparación de las medias sólo se realiza sobre los diferentes niveles de Fósforo sin importar el tratamiento de Nitrógeno que hayan recibido ya que este factor no presentó un efecto significativo sobre el rendimiento.

Para este ensayo no sería necesario aplicar un Test de comparación de medias, esto se debe a que el factor fosforo solo presenta dos niveles 0 y 1 y al haber sido significativa la prueba de F, esta directamente nos indica que el nivel de fósforo con mayor media es el que provocará el mayor rendimiento. Por lo tanto como P_1 presenta un rendimiento medio de 4,90 tn/ha contra 1,43 tn/ha de P_0 el fósforo en la dosis 1 es el que maximizará el rendimiento de maíz. Igualmente a modo de ejemplo aplicaremos el Test de Tukey.

Test de Tukey (para un nivel de significancia del 5%)

Como punto de partida debemos calcular la diferencia mínima significativa (d.m.s) como:

$$d.m.s_{5\%} = \Delta_{5\%} = q_{5\%} \cdot \frac{S}{\sqrt{ra}} = 3{,}08 \frac{\sqrt{0{,}3939}}{\sqrt{5x2}} = 0{,}6113$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (5%), del número de niveles del factor Fósforo (2 en este caso) y de los grados de libertad del error (para este caso gld =12).

Para facilitar las comparaciones se ordenan los tratamientos según sus medias, de menor a mayor.

Conclusiones de este test:

Tratamiento	*Media*	*Significancia*	*Significancia*
P_0	1,43	X	a
P_1	4,90	X	b

Como se puede ver de la aplicación del Test de Tukey surgen las mismas conclusiones que las obtenidas a partir de la prueba de F, es decir que el fósforo en la dosis 1 es el que maximiza el rendimiento.

Ejemplo con interacción significativa entre los factores

Deseamos probar dos dosis de nitrógeno y dos de fósforo en un cultivo de maíz. Este ensayo presenta dos factores (N y P), N con dos niveles (0,1) y P con dos niveles (0,1) se lo denomina factorial de 2 x 2 ó 2^2 , con 4 (cuatro) tratamientos.

Supongamos que el diseño del ensayo es un D.B.C.A. con 5 repeticiones

1) *Planteo de Hipótesis que se someterán a prueba*:

Hipótesis 1: No hay diferencias de rendimiento entre los dos niveles de nitrógeno (N0 = N1)

Hipótesis 2: No hay diferencias de rendimiento entre los dos niveles de fósforo (P0 = P1)

Hipótesis 3: No hay interacción entre los factores (independencia).

2) *Completado de la Tabla de Análisis de la Varianza*

Del Análisis de los datos (que no se presentan) surge el siguiente cuadro de análisis de la varianza:

Fuentes de Variación	***Suma de cuadrados***	***Gl***	***Cuadrados Medios***	***F Calculado***
Bloque	1.317	4	0.32925	
Factor A (N)	57.8	1	57.8	43.59
Factor B (P)	87.362	1	87.362	65.89
Interacción (N x P)	45.602	1	45.602	34.39 *
Dentro Error Experimental	15.911	12	1.32592	
Total	207.992	19		

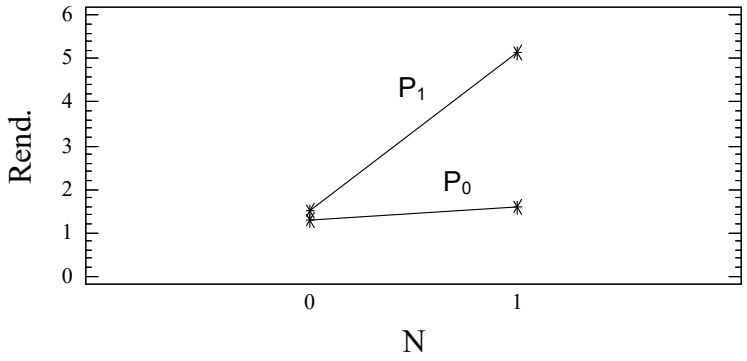
3) *Obtención de los F tabulados*

Fuentes de Variación	*gl del numerador*	*gl del denominador*	*Nivel de significancia*	*F Tabulado*
Nitrógeno	1	12	0,05	4,75
Fósforo	1	12	0,05	4,75
Interacción	1	12	0,05	4,75

4) *Conclusión parcial:* Como el valor que corresponde al **F calculado para la interacción** (34,39) es **mayor** que el correspondiente al **F tabulado** (4,75), la **hipótesis nula que plantea la no interacción de los factores es rechazada** (hipótesis 3), por lo tanto podemos concluir que **los efectos principales no son independientes**, lo cual no nos permite analizar los efectos principales.

La interacción significa que el efecto del nitrógeno no es el mismo para los diferentes niveles de fósforo o viceversa.

5) *Gráfico de Interacción*



El no paralelismo entre las rectas P_1 y P_0 nos indica la posible interacción de los factores nitrógeno y fósforo. Por otro lado permite apreciar que el agregado de nitrógeno no ejerce ningún efecto sobre el rendimiento cuando el nivel fósforo es cero, en cambio cuando el nivel de fósforo es uno el agregado de nitrógeno incrementa marcadamente el rendimiento, lo cual nos manifiesta la potencial interacción.

6) *Comparaciones múltiples de medias*

La presencia de interacción significativa entre los factores nos conduce a realizar las comparaciones de medias sobre el total de tratamiento o combinaciones.

Test de Tukey (para un nivel de significancia del 5%)

Como punto de partida debemos calcular la diferencia mínima significativa (d.m.s) como:

$$d.m.s_{5\%} = \Delta_{5\%} = q_{5\%} \cdot \frac{S}{\sqrt{r}} = 4{,}20\frac{\sqrt{1{,}3259}}{\sqrt{5}} = 2{,}163$$

Donde: q se obtiene de la tabla de la pagina 93 en función del nivel de significancia (5%), del número de tratamientos (4 en este caso) y de los grados de libertad del error (para este caso gld =12).

Previo a las conclusiones por este test, para facilitar las comparaciones se ordenan los tratamientos según sus medias, de menor a mayor.

Conclusiones de este test:

Tratamiento	*Media*	*Significancia*	*Significancia*
N_0P_0	1,28	X	a
N_0P_1	1,52	X	a
N_1P_0	1,58	X	a
N_1P_1	5,16	X	b

De la aplicación del Test de Tukey se verifica que el tratamiento $\mathbf{N_1P_1}$ supera a todos los demás, maximizando el rendimiento del maíz.