

# StadHelp 2.0k – Medal Project

## Tutorial BASICO DE ESTADISTICA

Editor: Dr. Jorge Raúl Rodríguez Yañez / Medal. Agosto, 2000

### Indice Temático

ACLARACIONES PRELIMINARES  
DEFINICIONES BASICAS  
VARIABLES EN BIOESTADISTICA  
DISTRIBUCION  
SIGNIFICACION ESTADISTICA  
MEDIDAS ESTADISTICAS  
METODO CIENTIFICO  
TIPOS DE ESTUDIOS CIENTIFICOS  
ANALISIS ESTADISTICO  
ANALISIS BIVARIADO  
TABLAS DE CONTINGENCIA  
PRUEBA DE T STUDENT  
ANALISIS DE VARIANZA  
CORRELACION  
REGRESION  
ANALISIS ACTUARIAL  
ANALISIS MULTIVARIADO  
MODELOS PREDICTIVOS  
SENSIBILIDAD Y ESPECIFICIDAD  
USO DE BASES DE DATOS  
BASES Y TABLAS  
FORMULARIOS  
CONSULTAS  
INFORMES Y REPORTEES  
BIBLIOGRAFIA

**Copyright:** SPSS es una marca registrada de SPSS Inc. Chicago, IL (<http://www.spss.com> ).  
Las demás marcas registradas pertenecen a sus respectivos autores. Medal pertenece al  
*Medical Algorithms Project*, USA. <http://www.medal.org> y <http://www.medalorg.ar> .

## ACLARACIONES PRELIMINARES



Esta pequeño Tutorial tiene simplemente la misión de permitir al medico general o especialista y demás miembros del equipo de salud, una mejor comprensión de la Bioestadística. De ninguna manera se pretende dar información valida para que cualquier profesional aplique la estadística, ya que para eso se necesita estudiar en forma directa la misma, capacitarse y ganar experiencia.

La estadística es un arma de doble filo. Se deben cumplir ciertos requisitos para poder desarrollar trabajos científicos o interpretarlos correctamente. El primer punto es establecer los objetivos de estudio dentro del marco correcto, ético y legal.

Segundo elegir que datos se analizaran a posterior. Tercero recolectar esos datos sin viciar la muestra. Cuarto volcar los datos a una base para su posterior análisis. Y por último, analizar estadísticamente los datos según la hipótesis planteadas. Para ello se debe conocer que pruebas y en que tiempo se deben utilizar.

Si no se siguen estos pasos y el análisis lo efectúa alguien sin la experiencia suficiente los resultados podrán ser muy “ bonitos “, pero seguramente no serán válidos frente a la realidad, y algo más, probablemente determinen conductas que perjudiquen de alguna forma al enfermo.

Se debe ser totalmente honesto en la practica científica. Se pueden “inventar“ datos que arrojaran resultados excepcionales, todo depende de cada profesional y su ética. La estadística no es la panacea de nada, solamente intenta dar una explicación a un fenómeno observado en la realidad, pero ojo, la realidad es la realidad y nada la puede reemplazar.

En este humilde manual, no se colocan fórmulas, ya que el objetivo no es inundar con guarismos matemáticos al lector, sino dejar claros algunos conceptos básicos.

Muchas Gracias

Dr. Jorge Raúl Rodríguez Yañez  
Medal Project. Bioestadística e Informática en Medicina  
jrrodri@intramed.net.ar

## DEFINICIONES BASICAS



**Estadística:** es una disciplina de estudio relacionada con la recopilación, organización y resumen de **datos** y la obtención de inferencias a partir de esos datos. La estadística descriptiva, describe los resultados globales de los datos recogidos. La estadística **inferencial**, analiza los datos y aporta conclusiones a una **hipótesis** previa.

**Individuo:** es la unidad mínima que se estudia. En medicina habitualmente es el paciente y en el caso de personas sanas se denomina sujeto o persona. También pueden estudiarse otros como: animales de experimentación, datos de laboratorio, exámenes, etc. (en estos casos se denomina observación).

**Población:** conjunto de individuos, sujetos u observaciones con alguna característica en común. Conjunto de elementos de la misma especie que se pretende estudiar en una investigación científica y de la cual se obtiene una muestra.

Las poblaciones pueden ser clasificadas básicamente como sigue:

- **Población General o Madre:** población real que se pretende estudiar y a la cual se extenderán las conclusiones de la muestra perteneciente a la misma.
- **Población Hipotética:** conjunto formado por todas las poblaciones, en las que se podría efectuar la investigación llevada a cabo en la población actual de estudio.
- **Población Estándar:** población patrón, que sirve de base para comparaciones con otras poblaciones.

**Muestra:** es el grupo de pacientes u observaciones que se estudiará, la cual debe haberse elegido al azar (**Aleatorio**) y ser representativa de la población a la cual pertenece, esto quiere decir: sin **sesgos**. En general la muestra es toda parte representativa de un conjunto, población o universo, cuyas características debe reproducir en pequeño lo más exacto posible. A partir del análisis de la muestra, obtenida correctamente y al azar, se pueden hallar conclusiones que sean extrapolables a la población de origen. Para elegir la muestra debe apelarse a un determinado método de muestreo. Existen varios métodos de muestreo de acuerdo al objetivo que se quiera llegar con la muestra. El método habitual es el de muestreo al azar o aleatorio, pero también puede hacerse en determinados casos un muestreo controlado para evitar la incorporación de factores no deseables en la muestra

# VARIABLES EN BIOESTADISTICA



**Variable:** es una característica o propiedad determinada del **individuo**, sea medible o no. Esta propiedad hace que las personas de un grupo puedan diferir de las de otro grupo en la **muestra** o **población** de estudio.

**Las variables se clasifican en:**

**Variable Cuantitativa:** es la que se puede medir. Habitualmente es llamada variable **Numérica** o **Continua**, o sea que posee una continuidad. Por ejemplo la edad, hematócrito, transporte de oxígeno, altura, peso, frecuencia cardiaca o respiratoria, dosis de un medicamento.

**Variable Cualitativa:** son variables que representan cualidades de la muestra, como por ejemplo la evolución del paciente hacia la mejoría o la muerte, color de ojos de un grupo de personas, sexo, etc. Estas variables también son llamadas categóricas o discretas, por dividirse en **categorías**.

**Las variables cualitativas se clasifican en:**

**Variables Categóricas Dicotómicas:** son las que tienen **dos valores fijos y excluyentes** entre si como la evolución, presencia o ausencia de una enfermedad o característica en la muestra.

**Variables Categóricas Nominales:** son variables cualitativas que no permiten establecer un orden, por ejemplo la raza, que puede ser blanca, negra, caucásica, etc., o los grupos sanguíneos A, B, AB o 0. También son excluyentes entre si, o sea que cada paciente pertenece a una u otra categoría pero no a dos al mismo tiempo.

**Variables Categóricas Ordinales:** estas si permiten establecer un orden determinado, por ejemplo los grupos de Apache son I a IV, un paciente del grupo II tiene menor probabilidad de mortalidad en UTI que el del grupo IV. La clasificación de la Disnea en grados ( I a IV) es otro ejemplo. También son excluyentes entre sí.

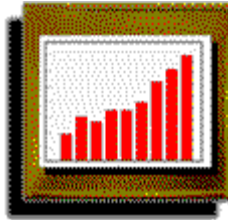
Además de lo expuesto anteriormente, existe otra forma de clasificar a las variables (v) que es también de suma importancia en estadística: en dependientes, independientes y asociadas.

**Variable Dependiente:** es la v. motivo de nuestro interés, cuyos valores dependen de otras variables que pueden influir en ella. También se la llama v. de respuesta. Por ejemplo la sobrevivida, respuesta al tratamiento, evolución, etc.

**Variable Independiente:** es la que modifica de una u otra manera a la v. dependiente, llamándose también según el caso factor de riesgo, factor predictivo, etc.

**Variable Asociada:** se denomina así a aquella v. independiente que no modifica por su sola presencia a la v. dependiente, pero que al combinarse con otra variable, si influye notoriamente a la anterior.

## DISTRIBUCION



En Bioestadística, la distribución se refiere en general a toda lista o tabla de **datos** estadísticos, ordenada según un criterio determinado. Una distribución se define por ciertas propiedades de su **variable** componente: **medidas de tendencia central** y **medidas de dispersión** de la variable.

Existen varios conceptos dentro de Distribución que se deben aclarar, a saber:

- **Distribución de Frecuencias:** **tabla** de datos, referentes a una variable en cuestión, en la que se exponen varias **categorías** de la misma, junto con sus **frecuencias** o número de veces que se repite en la muestra (puede expresarse también en porcentaje). La tabla puede tener diferentes formatos y es llamada tabla de frecuencias. Cuando se comparan la frecuencia de dos variables, se compone una **tabla de contingencia**, en la cual una variable ocupa las filas y la otra las columnas.

Ejemplo de una tabla de frecuencias simple de la variable COMA, en sus categorías: Ausente y Presente. (SPSS 9.0).

**Tabla de frecuencia COMA**

		Frecuencia	Porcentaje
Categorías	Ausencia	602	82,7
	Presencia	126	17,3
	Total	728	100,0

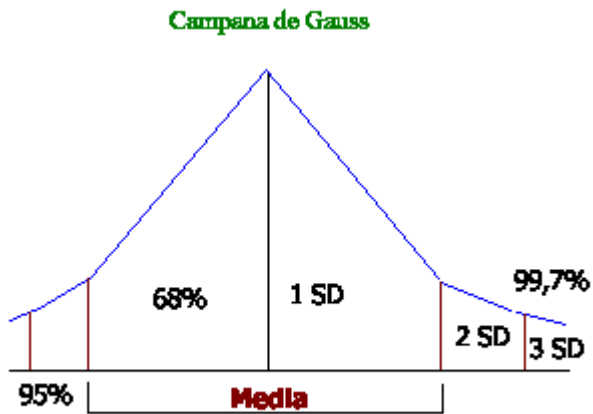
Ejemplo de Tabla de Contingencia de dos variables categóricas (DMO: Disfunción Orgánica y Shock: Estado de Shock) generada por el programa SPSS 9.0. Obsérvese que cada variable presenta dos categorías: Ausente y Presente.

**Tabla de contingencia DMO y Shock**

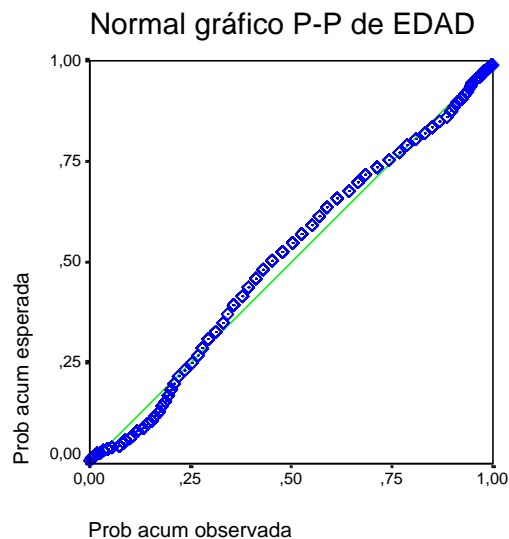
			Shock		Total
			Ausente	Presente	
DMO	Ausente	Recuento	428	82	510
	Presente	Recuento	62	151	213
Total		Recuento	490	233	723

- Distribución Normal:** la distribución normal es en forma de campana, habitualmente llamada distribución de Gauss. Es simétrica en torno a su **media** ( $\mu$ ); la media, mediana y modo son iguales; el área total de la curva por encima del eje basal  $x$  es la unidad del área = 1, por lo tanto cada sector de derecha e izquierda tiene un valor de 0,5. Si se trazan líneas perpendiculares a un **desvío estándar** ( $\sigma$ ) de distancia de la media, se obtiene un 68% del área de la curva. Dos desvíos estándar encierran un 95% y tres un 99,7% de la curva.

En el gráfico se observa la campana de Gauss, representante de la distribución normal y sus desvíos estándares.

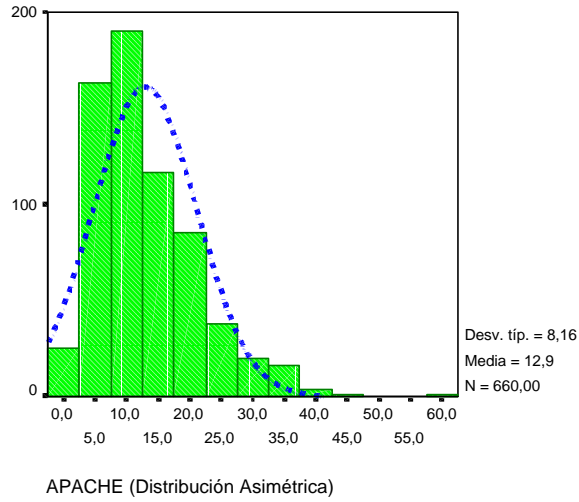


En el gráfico siguiente se expone un ejemplo de distribución normal generado por un gráfico PP de SPSS 9.0. Obsérvese que la distribución normal teórica está representada por la línea verde media y la distribución de la variable edad por los puntos azules. Se evidencia claramente que la distribución de la edad se acerca a la normal.

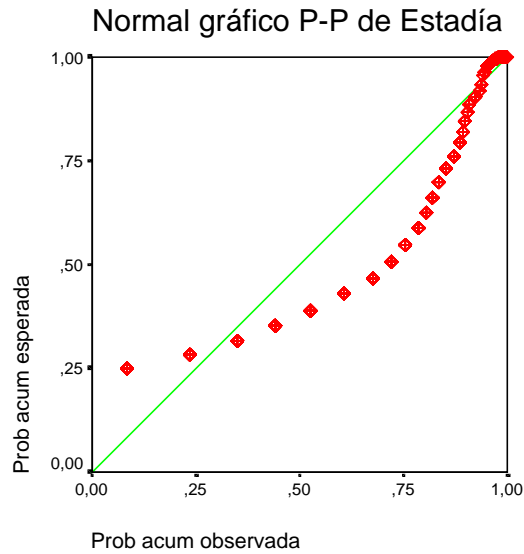


**Curiosis:** grado de afilamiento o achatamiento de una curva de distribución con relación a la curva normal.

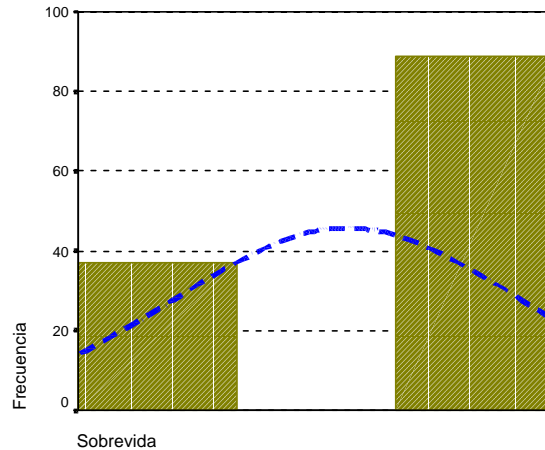
- **Distribución Asimétrica:** aquella que no guarda la relación de gauss en su distribución de frecuencia. En el gráfico siguiente se observa una distribución asimétrica en torno a la edad de una muestra. (SPSS 9.0).



En el gráfico siguiente se observa la distribución de la variable Estadía en relación a la normal. Se evidencia que la Estadía está distribuido en forma asimétrica.

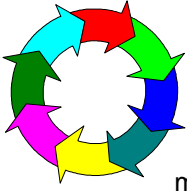


- Distribución Binomial:** la distribución Binomial afecta a las variables discretas solamente. Se deduce a partir del ensayo de Bernoulli, en donde se expone que cuando un experimento, solo puede conducir a dos resultados probables y mutuamente excluyentes, ambos son rotulados como éxito y fracaso, la probabilidad de éxito es  $p$  y la probabilidad de fracaso es  $q = 1 - p$ . Un ejemplo claro de distribución Binomial es la frecuencia de distribución de variables dicotómicas excluyentes como la sobrevida.



- Distribución de Poisson:** la distribución de Poisson también se observa en variables discretas. La ley de Poisson dice: la frecuencia de un evento es independiente de otros. La frecuencia de un evento en un intervalo de espacio o tiempo, no tiene efecto sobre la probabilidad de una segunda frecuencia del evento en el mismo intervalo o en cualquier otro. La distribución de Poisson se utiliza cuando se hacen registros de eventos que se distribuyen al azar en un espacio o tiempo determinado. Puede esperarse que cierto proceso obedezca la ley de Poisson y ante esta suposición se puede calcular la probabilidad de que ese evento se presente en una unidad de tiempo.

## SIGNIFICACION ESTADISTICA



Este concepto es una forma de expresar matemáticamente si dos grupos son o no diferentes dentro de una **muestra** o si dos **variables** tienen diferencias dentro de un mismo grupo y esas diferencias no son debidas a factores **aleatorios**. El método utilizado para hallar la significación estadística (ss), es un tipo especial de método matemático que se llama **análisis** estadístico. Es necesario crear una unidad de medida de ss para lo cual se usa el valor de p, al estudiar distribución de frecuencias, o el estudio de las colas de las distribuciones, o el área bajo una determinada curva, etc.

Por lo tanto **p** es la probabilidad de error al comparar dos o más muestras o grupos cuando aseguramos que ambos son diferentes. O sea que p es la probabilidad en el sentido de la significación estadística. Obtener una **p < 0.05** significa que tenemos un 5% de probabilidades de error en las conclusiones, por lo cual la probabilidad de equivocarnos es baja.

En otras palabras, en la estadística, se dice que un evento, suceso o valor, es significativo, cuando es poco probable y por lo tanto, seguramente no se debe al azar, sino a factores específicos.

De forma más estricta, significación estadística, hace referencia a la cuestión de determinar estadísticamente, si un valor o resultado obtenido de una muestra, es poco probable, de modo que no puede explicarse por las fluctuaciones propias de esa muestra en cuestión. En este caso, las conclusiones pueden ser extensibles a la población de la cual derivó la muestra, dando el basamento de rechazo de la hipótesis nula.

### Error tipo Alfa y tipo Beta

En el campo de la investigación en ciencias biológicas, un margen de error de hasta un 5% es aceptable desde el punto de vista estadístico. Este margen de error significa que las observaciones o resultados derivados de la investigación en curso, pueden deberse al azar en hasta un 5% de los casos. (valor de p 0.05). Esto significa también que los resultados se encuentran presentes en el 95% de los casos estudiados (intervalo de confianza del 95% de la media) y que se pueden generalizar a la población a la cual pertenecen.

La decisión de un investigador de rechazar o no una HP Nula, se basa en la consideración de la probabilidad de que las diferencias halladas se deban o no al azar. Como el investigador no cuenta con los datos de toda la población, siempre se puede incurrir en errores. Existen dos tipos de errores en los cuales se puede caer en la inferencia estadística:

**Error tipo Alfa:** o tipo I, es rechazar una HP Nula verdadera.

**Error tipo Beta:** o tipo II, es aceptar una HP Nula falsa.

# MEDIDAS ESTADISTICAS



Existen algunos conceptos que se aplican al análisis de las variables que deben quedar en claro. Algunas medidas son generales y otras especiales de acuerdo al tipo de variable mensurada.

## Medidas generales

- **n**: es el número de casos de la muestra
- **N**: es la suma del número de casos de varias muestras.
- **x**: cada uno de los datos de la muestra
- **Σ**: sumatoria de los datos de una serie

## 2 Medidas para variables numéricas:

- **Media** : es la suma de todos los valores dividido por el numero de casos **n**.
- **Mediana**: corresponde al valor central de la serie de datos observada.
- **Modo**: valor más frecuente de una serie de datos.
- **Desvío** : es la diferencia entre la  $\bar{x}$  y c/u de los valores de la muestra (**x**).
- **Varianza**: es el promedio de los cuadrados de los desvíos; la Varianza mide la dispersión de los valores y marca el punto de inflexión de las curvas en los histogramas.
- **Desvío Estándar (SD)** : es la raíz cuadrada de la Varianza y corresponde al 68% del área de la curva alrededor de  $\bar{x}$ .
- **Error Estándar**: es la división entre el **SD** y la raíz cuadrada de **n**.
- **Intervalo de Confianza para la Media (IC)**: se define como el espacio o intervalo comprendido por los valores extremos de la muestra, en el que teóricamente se va a encontrar la **media de la población o universo**. Habitualmente se utiliza el IC del 95% de la media, representado por un valor mínimo y máximo.

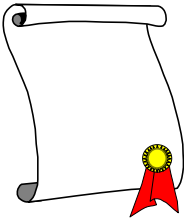
## 2 Medidas para variables discretas:

- **Frecuencia (f)**: numero de veces en que se repite una característica o las categorías de una variable discreta. Generalmente se acompaña del porcentaje relacionado a esa frecuencia.
- **Riesgo Relativo**: el riesgo relativo (**RR**) mide o cuantifica el grado de asociación existente entre la presencia de un factor y la aparición de un suceso. Se calcula como la razón entre dos tasas de incidencia. (Frecuencia de enfermos / frecuencia de casos expuestos). A menudo se utiliza en estudios cohorte, en los cuales se quiere identificar variables que estén

relacionadas con la aparición de un hecho concreto, una enfermedad, o la mortalidad, por ejemplo.

- **Odds:** consiste en el cociente entre dos frecuencias de una categoría de una misma variable. Representa la proporción de poder pertenecer a una u otra posibilidad de la misma categoría.
- **Odds Ratio:** razón entre dos Odds diferentes. El Odds Ratio (**OR**) se utiliza en estudios retrospectivos de caso – control, en donde un grupo padeció la enfermedad o fue tratado y el otro grupo fue control. (Evidentemente no se puede calcular una tasa de incidencia como el RR en este tipo de estudios retrospectivos). El cálculo del OR se efectúa de la siguiente manera:  $OR = (n \text{ casos positivos} / n \text{ casos negativos}) / (n \text{ casos control positivos} / n \text{ casos control negativos})$ .

## METODO CIENTIFICO



El método de investigación para el conocimiento de la realidad observable, que consiste en formularse interrogantes sobre esa realidad, con base en la teoría ya existente, tratando de hallar soluciones a los problemas planteados. El método científico (mtc) se basa en la recopilación de datos, su ordenamiento y su posterior análisis.

### Pasos del Método Científico:

- **Observación:** el primer paso es la observación de una parte limitada del universo o **población** que constituye la **muestra**. Anotación de lo observable, posterior ordenamiento, **tabulación** y selección de los **datos** obtenidos, para quedarse con los más representativos.
- **Hipótesis:** se desarrolla en esta etapa, el planteamiento de las hipótesis que expliquen los hechos ocurridos (observados). Este paso intenta explicar la relación causa – efecto entre los hechos. Para buscar la relación causa – efecto se utiliza la analogía y el método inductivo. La HP debe estar de acuerdo con lo que se pretende explicar (atingencia) y no se debe contraponer a otras HP generales ya aceptadas. La HP debe tener matices predictivos, si es posible. Cuanto más simple sea, mas fácilmente demostrable (las HP complejas, generalmente son reformulables a dos o más HP simples). La HP debe poder ser comprobable experimentalmente por otros investigadores, o sea ser reproducible.
- **Experimentación:** la hipótesis debe ser comprobada en estudios controlados, con autentica veracidad.

### Hipótesis en Investigación:

Hipótesis significa literalmente “lo que se supone”. Está compuesta por enunciados teóricos probables, referentes a variables o relaciones entre ellas. En el campo de la investigación, la hipótesis, supone soluciones probables al problema de estudio.

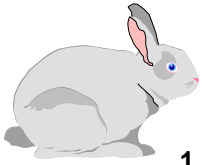
El proceso estadístico se basa en la comprobación de hipótesis (HP).

Existen dos tipos de HP, a saber:

- **HP. Alterna o Científica:** es la HP que pretende comprobar el investigador en su muestra de pacientes. Básicamente significa que la media de una característica o propiedad de un grupo es diferente a la media del otro grupo o grupos, o que la distribución y frecuencia de un evento en un grupo es diferente del otro. **H1 : grupo 1 grupo 2**
- **HP. Nula:** es lo contrario de la anterior, o sea que no existen diferencias entre dos o más grupos o muestras. **H0 : grupo 1 = grupo 2**

El **valor de p** es entonces la medida de la evidencia contra la H0. Cuanto menor sea el valor de p, menor será la posibilidad de que la HP. Nula sea cierta, por lo cual se rechazará, aceptando a la HP. Científica como verdadera.

# TIPOS DE ESTUDIOS CIENTIFICOS



Los estudios en Investigación Científica o Clínica se dividen en **Estudios de Observación y Experimentales**.

## 1. Estudios de Observación:

**1.1 Estudio de Casos en Serie** : generalmente se trata de la descripción de observaciones interesantes en un grupo de ptes. que corresponden a un periodo relativamente corto. No incluyen a un grupo control.No plantea ninguna **hipótesis** a investigar. Por su importante función descriptiva anteceden muchas veces a otros tipos de estudios que se encargaran de comprobar HP.

**1.2 Estudio de Caso-Control** : En los estudios de caso-control se comienza con la presencia o ausencia de una determinada característica o propiedad y luego se investiga hacia atrás en el tiempo tratando de detectar causas o factores de riesgo posibles,que generalmente han sido marcados por un estudio de casos en serie previo. Existen Criterios de Inclusión y Exclusión para el seguimiento de ptes. y los casos-control son sujetos sin esa enfermedad o característica. De esta manera se estudia las historias o evoluciones de casos-control y enfermos para determinar sus diferencias y el grado de significación de las mismas. O sea que el **Estudio de Caso-Control** es un estudio **Longitudinal Retrospectivo**.

**1.3 Estudio Transversal** : en el ETV se analiza datos de un grupo de ptes. o sujetos sanos en un momento dado, en lugar de un periodo determinado. La pregunta básica es: ¿que es lo que está pasando en este momento ?.Son también llamados estudios de **Prevalencia**. Un ejemplo clásico son las encuestas. Se utilizan para describir una enfermedad o proporcionar información respecto al dg. o etapa de una enfermedad.

**1.4 Estudios Cohorte** : los ECH son **Longitudinales y Prospectivos**. Una **Cohorte** es un grupo de Individuos que tienen algo en común y que forman parte de un grupo por un largo periodo. En medicina los sujetos de un ECH se relacionan por alguna característica definida o mas de una, o por uno o mas factores de riesgo para una determinada situación, patología o evolución. La pregunta básica es : ¿Que pasará ? Un ECH puede servir para varios fines como para la investigación de Factores de Riesgo, de la evolución de una enfermedad, de un tratamiento, etc.

## 2. Estudios Experimentales o Pruebas Clínicas :

**2.1 Pruebas Clínicas Controladas** : Las PCC pueden diseñarse de varias formas, por ejemplo con un grupo control, con autocontroles o con control histórico. La PCC con un grupo de estudio y otro control es la mas usada en medicina. Básicamente a un grupo con enfermedad se lo somete a un tratamiento y se lo compara con un grupo control, o sea sin el tratamiento que se quiere probar o se administra placebo. Al grupo que recibe tto. se lo denomina grupo experimental que puede ser uno o mas. Cuando el medico conoce el tto. y el pte. no, la prueba se llama **Ciega** y a su vez cuando ninguno de los dos conoce si se esta administrando tto. o placebo, la prueba se denomina a **Doble Ciego**. Los ptes. deben asignarse a cada grupo en forma aleatoria o randomizada.

**2.2 Pruebas Clínicas No Controladas** : Las PCNC son aquellas en donde no se establece ningún grupo control. Para la investigación de una tratamiento o intervención terapéutica son

totalmente inapropiadas. A veces son apropiadas si para investigar sobre determinadas características de una muestra de ptes. solamente.

# ANÁLISIS ESTADÍSTICO



El análisis estadístico se divide en tres grandes tipos : **univariado, bivariado y multivariado**.

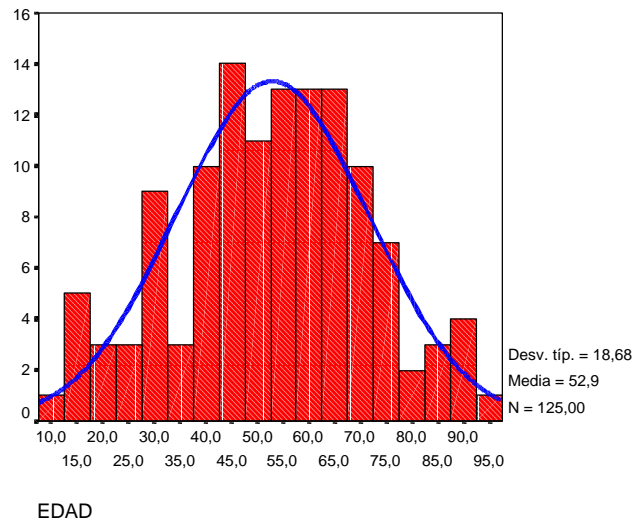
- En el análisis univariado se describen las características de una variable por vez. **También se lo llama estadística descriptiva**.
- En el análisis bivariado se investiga la influencia de una **variable** que es **independiente**, por vez, con respecto a la **variable dependiente**.
- En el análisis multivariado se investiga la influencia de dos o mas variables independientes, junto o no a una o mas variables asociadas (**covariables o cofactores**) sobre una o más variables dependientes.

## Análisis Univariado de Variables Numéricas

El análisis de los datos tiene como objetivo el responder a las preguntas que se hicieron los investigadores, pero para llegar a ese punto primero se debe describir las variables o datos que se recogieron durante el estudio.

Para describir una **variable numérica** se la puede ordenar de mayor a menor y observar cuantos pacientes corresponden a cada cifra (**histograma**), encontrar su media, SD, valores mínimos y máximos, etc., dependiendo de cada estudio en particular.

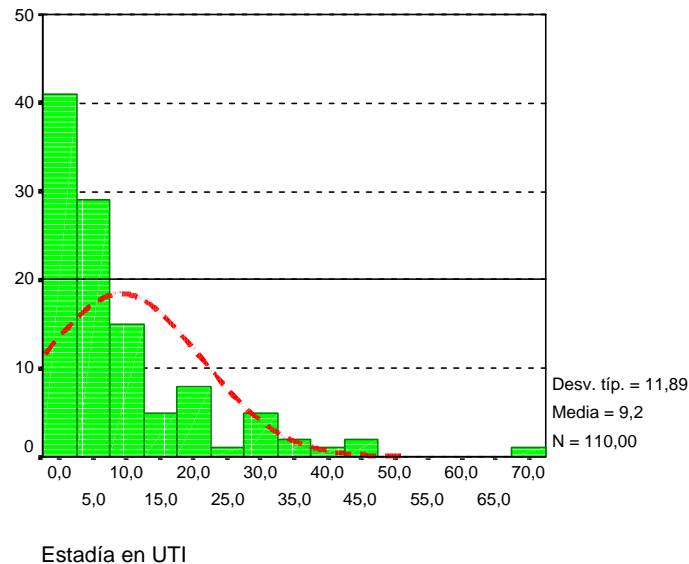
La idea de la tendencia central de esa v. numérica, es el promedio aritmético de la v. en cuestión. El histograma representa la frecuencia de ptes. dentro de determinados rangos de la v numérica. Esto se denomina **distribución de frecuencias**.



Histograma de Frecuencias de la Variable Edad en una muestra de 125 pacientes de UTI. **Distribución normal**. (SPSS 9.0). Como se observa en la figura, la distribución de la Edad se acerca a la distribución normal o Gaussiana lo cual se denomina distribución de t.

La figura representa el histograma de la edad en UTI de la misma muestra, tal como se observa la distribución no es igual a la anterior, denominándose **distribución asimétrica** o

simplemente anormal.

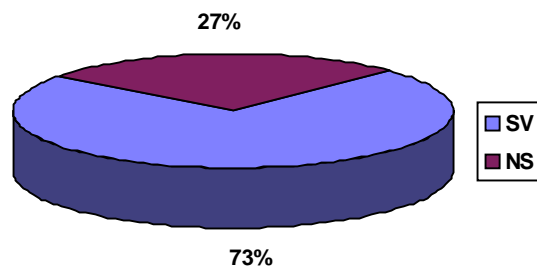


## Análisis Univariado de Variables Categóricas

### Variables Categóricas Dicotómicas:

el ejemplo mas común en la investigación medica es la evolución (sobrevida o no sobrevida (SV-NS)) como tipo de **variable categórica dicotómica**.

Por ejemplo en un periodo determinado se siguen 100 ptes. y fallecen 27. Simplemente se divide  $27/100$  y se obtiene 27% de mortalidad. La representación gráfica de este porcentaje puede hacerse en barras o sectores (tortas). Siempre debe escribirse al lado del porcentaje el n de la muestra.



### Variables categóricas nominales y ordinales:

como en el caso anterior se calcula los porcentajes de cada grupo en particular acompañándolo de los n correspondientes.

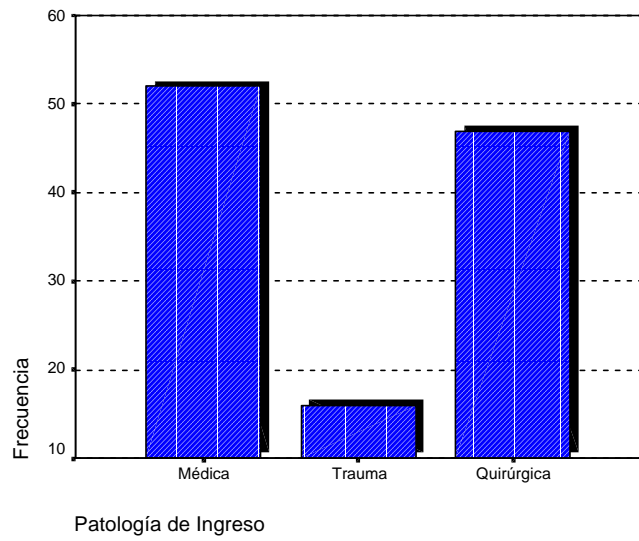


Gráfico que muestra la frecuencia de patología de ingreso a sala en pacientes críticos (SPSS 9.0).

## ANALISIS BIVARIADO



El análisis bivariado enfrenta a cada una de las **variables independientes con la dependiente**, por separado. Por ejemplo, tenemos una muestra en donde queremos analizar el tipo y dosis de ATB administrado a 150 ptes con faringitis para observar su curación. Primero haremos un análisis univariado o descriptivo, de tipo de ATB, dosis y curación, en una **tabla** :

Item (variables)	n total = 150	Porcentaje o Media $\pm$ SD
<b>ATB</b>		
<b>Amoxicilina</b>	50	33%
<b>Ampicilina</b>	50	33%
<b>Penicilina</b>	50	33%
<b>Curación global</b>	100	66%
<b>Dosis Amoxicilina</b>	50	2,2gr $\pm$ 0,3
<b>Dosis Ampicilina</b>	50	1,5gr $\pm$ 0,5
<b>Dosis Penicilina</b>	50	1 ml. Unid $\pm$ 0,25

Luego debemos saber si el índice de curación fue mejor con alguno de estos ATB por lo cual le aplicamos el análisis bivariado correspondiente. Tomamos primero a numero de ptes con y sin Amoxicilina y la enfrentamos a la curación y así sucesivamente con la Ampicilina y penicilina por separado. De esta manera cada ATB se convierte en v. independiente frente a la curación que es la v. dependiente. Lo mismo haremos con la dosis de cada ATB. Esto es un análisis bivariado.

De acuerdo a la distribución y tipo de variable en juego ( sea **numérica** o **discreta**), se debe elegir entre las pruebas **paramétricas** y **no paramétricas** para aplicar. Si la distribución es **normal**, como ocurre habitualmente en muchas variables numéricas, se aplican pruebas paramétricas. De lo contrario (**distribución anormal**), como ocurre en las variables categóricas, se utilizan pruebas no parametricas. Veamos la siguiente tabla para una mejor comprensión:

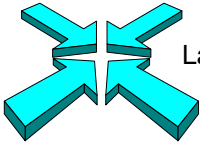
Tipo de Variable	Distribución	Prueba a usar
Numérica	Simétrica	t Student / ANOVA
Numérica	Asimétrica	U Mann – Withney Kruskal – Wallys
Categórica	Binomial	Prueba Chi Cuadrado Riesgo Relativo / Odds

En caso de v. numéricas, lo que se hace es comparar la media de un grupo con respecto a la media del otro grupo. En nuestro ejemplo compararíamos la dosis media de Amoxicilina en los pacientes con curación si y no. Luego procederíamos de la misma forma para la peni y ampi. Si

la dosis de amoxi posee una distribución normal se puede usar la prueba de t student y si por ejemplo la dosis de peni y ampi describen distribuciones asimétricas o anormales, deberíamos utilizar la prueba de U de Mann-Withney. Las pruebas nos arrojarán un valor de resultado (valor de t o valor de U) que si corresponde a una  $p < 0,05$ , nos habla de que solo en el 5% de los casos las variaciones de las dosis en los grupos citados se deben al azar, lo cual es **significativo** para la medicina. En este caso concluiríamos que los antibióticos influenciaron benéficamente en la curación.

En caso de que las v. a analizar sean categóricas se debe usar **tablas de contingencia** en las cuales se colocan las categorías de una de las variables en las columnas y las categorías de la otra en las filas.

## TABLAS DE CONTINGENCIA



Las tablas de contingencia están compuestas por filas (horizontales) y columnas (verticales) que delimitan celdas donde se vuelcan la **frecuencia** de cada **categoría** analizada. En el ejemplo siguiente, efectuado con el programa SPSS 9.0 se observa la tabla de contingencia de dos variables en una población de pacientes críticos hipotéticos: la evolución (SV/NS) y la presencia o ausencia de coma al ingreso. Las celdas delimitadas por estas dos variables comprenden cuatro tipos:

- Frecuencia de pacientes sin coma sobrevivientes
- Frecuencia de pacientes sin coma no sobrevivientes
- Frecuencia de pacientes con coma sobrevivientes
- Frecuencia de pacientes con coma No Sobrevivientes

**Tabla de contingencia EVOL \* COMA**

Recuento		COMA		Total
		NO	SI	
EVOL	SV <sup>a</sup>	484	37	521
	NS <sup>b</sup>	118	89	207
Total		602	126	728

a. SV = Sobreviviente

b. NS = No Sobreviviente

La pregunta es ¿es el coma al ingreso un factor de riesgo para la mortalidad?

Para contestarnos esa pregunta debemos apelar al uso de la prueba de Chi Cuadrado de Independencia. Esta prueba contrasta la hipótesis: ¿las categorías de las dos variables son independientes entre sí o no?. El análisis del chi cuadrado arroja un valor de p determinado, que si es inferior a 0.05, indica que existe una relación entre las categorías estudiadas, o sea que las variables no son independientes entre sí.

En general la prueba de chi cuadrado presenta ciertos puntos a tener en cuenta:

- Si el N casos es pequeño, se utiliza la prueba exacta de Fisher para obtener el valor de chi cuadrado (X<sup>2</sup>).
- Si el N > 40 casos se puede utilizar la corrección de continuidad de Yates para obtener el X<sup>2</sup>.
- Para hallar correctamente el valor de X<sup>2</sup>, la tabla de 2x2 debe estar integrada por valores de una muestra aleatoria, con distribución multinomial y los valores esperados no deben ser < 5.
- Los métodos estadísticos más usados para hallar el valor del X<sup>2</sup> son el método de Pearson y el de razón de verosimilitud, funcionan muy bien para muestras grandes.

Pruebas de chi-cuadrado

	Valor	gl	Valor p
Chi-cuadrado de Pearson	133,353	1	,000
Corrección de continuidad	130,857	1	
Razón de verosimilitud	120,913	1	
Estadístico exacto de Fisher			
Asociación lineal por lineal	133,170	1	
N de casos válidos	728		

Según el análisis el valor del estadístico chi cuadrado es de 133, 353 correspondiendo a un valor de  $p = 0.000$  (según el test de Fisher), es decir una  $p < 0.001$ , sumamente **significativa**, lo cual indica que existe una relación entre coma al ingreso y sobrevida en pacientes críticos. El valor de gl representa los grados de libertad de la muestra estudiada.

Para determinar si la presencia de coma al ingreso empeora la sobrevida o no, debemos determinar el **Riesgo Relativo** del cruce de las dos variables. El valor sería **3,604**, por lo cual la presencia de coma al ingreso determina un mayor riesgo de mortalidad.

Estimación de riesgo de mortalidad para coma al ingreso

	Valor	Intervalo de confianza al 95%	
		Inferior	Superior
Coma Si	3,604	2,959	4,389
Coma No	,365	,278	,480

## PRUEBA DE t STUDENT



La prueba de t Student, es un método de análisis estadístico, que compara las medias de dos categorías dentro de una variable dependiente, o las medias de dos grupos diferentes. Es una prueba paramétrica, o sea que solo sirve para comparar variables numéricas de **distribución normal**. En caso de tener que analizar variables numéricas de **distribución anormal**, se debe utilizar otro tipo de pruebas no paramétricas, como la prueba U de Mann – Withney.

La prueba t Student, arroja el valor del estadístico t. Según sea el valor de t, corresponderá un valor de significación estadística determinado.

En definitiva la prueba de t Student contrasta la HP Nula de que la media de la variable numérica “y”, no tiene diferencias para cada grupo de la variable categórica “x”.

La **prueba t para muestras independientes** se utiliza para comparar la media de dos grupos o dos categorías dentro de una misma variable dependiente.

Por ejemplo, supongamos la comparación de la edad en 566 pacientes con Hipertensión esencial y 214 con Hipertensión secundaria. Los resultados arrojan que los pacientes del grupo de hipertensión esencial presentan una edad media de 55 12 años, mientras que los hipertensos secundarios 26 8 años. El valor de la prueba t se establece mediante el estadístico t que en este caso es de 38,9 correspondiendo a una  $p < 0.0001$ . Esto implica que la diferencia de edad entre ambos grupos de hipertensos no es **aleatoria**, o sea que la hipertensión secundaria se observa en grupos etarios más jóvenes. (se rechaza la HP Nula **HP alterna**)

La **prueba t para muestras dependientes** se utiliza para comparar las medias de un mismo grupo en diferentes etapas, como por ejemplo pre y post tratamiento. Supongamos el grupo de 566 Hipertensos sometidos a tratamiento durante un mes. Los valores de tensión arterial media (TAM) pretratamiento fueron de 125 15 mmHg, que descendieron a 88 10 mmHg postratamiento. Comparando ambas medias observamos un valor de t de 78,9 correspondiendo a una  $p < 0.0001$ . Esto implica que el descenso de la TAM con el tratamiento no se produjo al azar.

### Prueba U de Mann – Withney

La U de Mann – Withney es una prueba no paramétrica para grupos independientes, que mide las diferencias entre medias, asignando rangos a cada grupo. La suma de rangos para los 2 grupos puede compararse por la obtención de la cifra estadística U)

La prueba de **Suma de Rangos de Wilcoxon** es semejante a la prueba U, pero se utiliza para muestras de grupos dependientes o apareados.

## ANALISIS DE VARIANZA



El **ANOVA** es una prueba semejante a la **prueba t Student**, en cuanto a la práctica, pero la comparación entre grupos no es a través de la **media** y su **SD**, sino a través de la **varianza** de la variable numérica "y", en cada grupo de la variable categórica "x".

Básicamente el **análisis de Varianza**, se utiliza para corroborar si la significación de diferencias entre medias de dos o mas grupos, son o no debidas al azar. La cifra estadística obtenida con el Anova es la razón F.

Suponiendo que se analizan 2 grupos, el Anova, analiza las variaciones entre los dos grupos (inter-grupal) y la compara con la variación dentro de cada grupo (intra-grupal), para obtener mediante una suma de cuadrados el valor de F.

Si las diferencias de **varianza** entre cada grupo son mayores que las intra-grupales, seguramente existen diferencias significativas entre los grupos que no son debidas al azar.

Los grupos se definen como en la prueba t eligiendo una variable categórica. La variable a analizar debe ser numérica y de distribución simétrica.

Utilizando la misma muestra de pacientes con 566 hipertensos esenciales y 214 secundarios, el valor de F es 109,43, lo cual corresponde a un valor de  $p < 0.001$ . Esto implica que las diferencias de medias de edad entre ambos grupos no es debida al azar.

También existe un modelo de Anova multivariado, llamado **MANOVA**, en el cual se comparan mas de una variable numérica en dos o más grupos.

En caso de tener que analizar las varianzas de variables numéricas de distribución asimétrica, se debe apelar a otro tipo de métodos semejantes a la prueba U pero modificados. La **prueba de Kruskal – Wallis** es uno de los más utilizados.

# CORRELACION



Otra forma de **análisis bivariado** es la correlación y regresión de variables **numéricas** y **discretas**. El concepto de correlación y regresión se basa en el grado de relación que poseen dos variables numéricas entre sí.

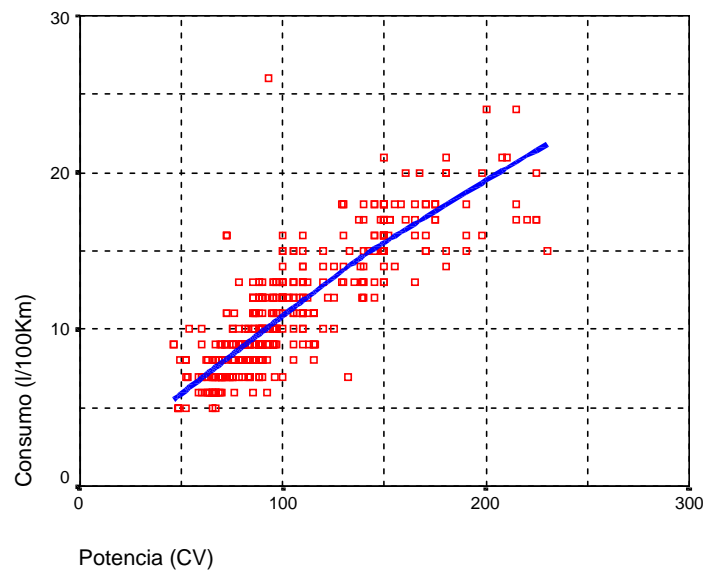
El coeficiente de correlación permite predecir si entre dos variables existe o no una relación o dependencia matemática.

Supongamos que queremos estudiar la correlación existente entre peso y altura de un grupo de personas tomadas al azar. Sometemos los datos recogidos de peso y altura al análisis de correlación y encontramos el coeficiente de correlación entre ambas, que se representa con la letra  $r$ . El  $r = 0.78$ . Esto significa que a mayor altura correspondería mayor peso.

Los coeficientes de correlación  $r$  siempre oscilan entre valores de 1 y  $-1$ . El valor cero 0 significa que no existe correlación entre ambas variables. Un valor positivo indica que a incrementos en la variable A se producen incrementos proporcionales en B y un valor negativo indica lo contrario.

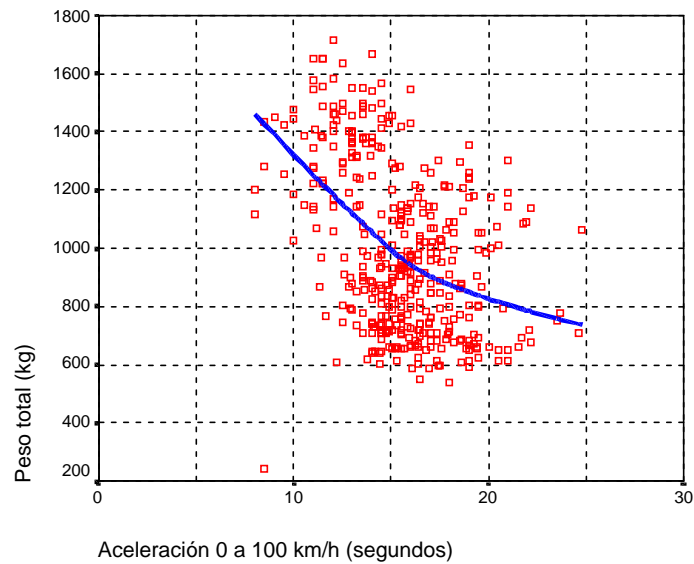
Podemos graficar la correlación entre las dos variables a través de una gráfica de dos ejes (abscisas y ordenadas) cartesianos.

En el siguiente gráfico observamos la correlación entre potencia de motor de un automóvil y consumo en Litros por cada 100 Km. El  $r = 0.87$  (correlación positiva). (SPSS 7.5). Evidentemente a mayor potencia se observa mayor consumo de combustible. El valor de significación para ese  $r$  es de una  $p < 0.01$ . Esto quiere decir que la correlación entre potencia y consumo no es **aleatoria**.



En el siguiente gráfico encontramos la relación existente entre peso del automóvil en kg. y aceleración 0 a 100 Km. / hora en segundos. El  $r = -0.56$  con una  $p < 0.05$ . Esto significa que

existe una correlación negativa significativa, entre peso del auto y respuesta de la aceleración. Automóviles más pesados presentan una respuesta más tardía y viceversa. (SPSS 7.5)



Para interpretar el coeficiente de correlación, **Colton** a dado los siguientes lineamientos generales:

- Valor de  $r$  de 0 a 0.25 implica que no existe correlación entre ambas variables.
- Valor de  $r$  de 0.25 a 0.50 implica una correlación baja a moderada.
- Valor de  $r$  de 0.50 a 0.75 implica correlación moderada a buena.
- Valor de  $r$  de 0.75 o mayor, implica una muy buena a excelente correlación.
- Estos rangos de valores se pueden extrapolar a correlaciones negativas también.

Se debe tener cuidado al analizar la correlación entre dos variables, de que ambas varíen juntas permanentemente. Esto parece redundante, pero es importante. Por ejemplo, si correlacionamos edad y altura. La altura irá aumentando con la edad hasta un determinado punto en donde ya no aumentará más.

## REGRESION



Se puede definir a la Regresión, como una **correlación** matemática basada en la ecuación de la recta modificada. Existen varios tipos de regresión y todos se basan en modificaciones de la formula de regresión lineal :

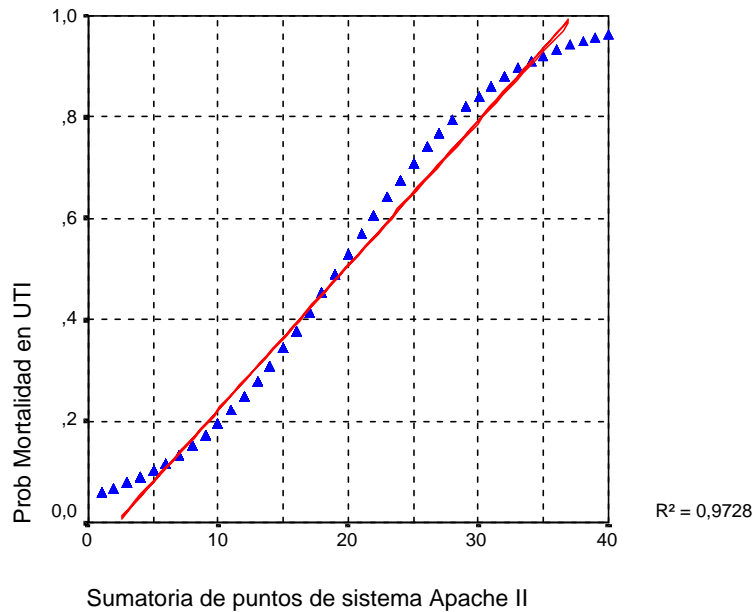
$$Y = a + b \cdot X \quad (\text{ecuación matemática de la recta})$$

**Y** es la variable dependiente (de estudio) y **X** la variable independiente.

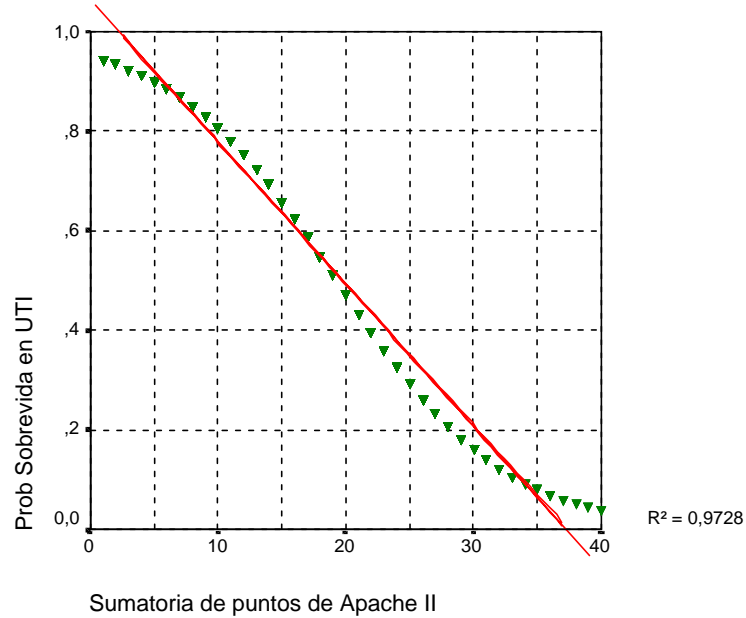
**a** y **b** son factores derivados de la ecuación matemática de la recta.

Básicamente, por medio de la regresión se pretende predecir el valor de una variable llamada genéricamente "**Y**", a través de otra variable llamada "**X**".

La regresión se representa mediante un coeficiente R que oscila entre - 1 y + 1. Cuando la variable dependiente Y aumenta ante incrementos de la variable independiente X , el R es positivo y oscila entre 0 y 1. A su vez cuando Y disminuye ante incrementos de X el R es negativo, entre 0 y -1. Veamos algunos ejemplos para una mejor comprensión :



El gráfico muestra la relación existente entre sumatoria de puntos del Sistema Apache II a las 24 horas del ingreso y la probabilidad de mortalidad en terapia intensiva. El valor de R es 0,98 para un nivel de  $p < 0.001$ . El valor de R<sup>2</sup> es 0.97. El R<sup>2</sup> es un coeficiente importante en regresión. Se deduce de la elevación al cuadrado de R y es representativo del grado de relación entre variables. Un R<sup>2</sup> de 0.97, significa que el valor de la probabilidad de mortalidad podrá ser predecido en un 97% de las veces por el valor del Apache II.



Si analizamos la Probabilidad de sobrevivida en UTI con respecto al valor del Apache II, obtenemos una curva de regresión similar pero negativa, con un  $R = -0,98$  y nuevamente un  $R^2$  de 0.97.

Por lo tanto el valor de **R<sup>2</sup>** indica el porcentaje de variabilidad de los valores de Y que pueden ser explicadas en función de la variabilidad de los valores de X.

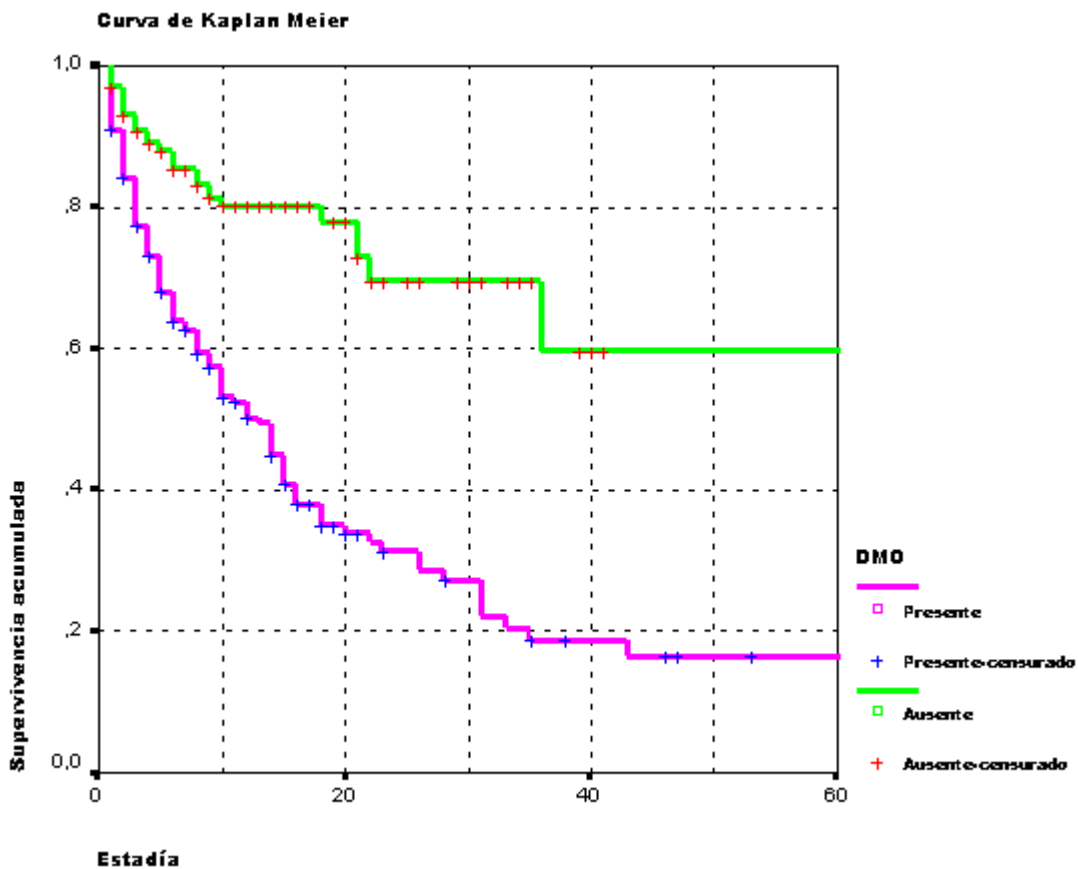
Los modelos de regresión no siempre son lineales y se basan en la ecuación pura de la recta. Existen también modificaciones de esta ecuación de tal manera que se pueden practicar análisis de regresión cuadrática, cúbica, logarítmica, logística, etc. Además la regresión puede ser simple o múltiple, constituyendo un tipo de **análisis multivariado**.

## ANALISIS ACTUARIAL



Se denomina **análisis actuarial** a un grupo especial de estudios estadísticos que tienen en cuenta el tiempo como factor primordial. Habitualmente se analiza el efecto durante el tiempo de una variable independiente, que puede denominarse **factor**, sobre una variable dependiente, generalmente **dicotómica**.

Existen varias pruebas actuariales, pero las más difundidas son la de **Kaplan Meier** y la **Regresión de Riesgos Proporcionales de Cox**. El Análisis de Curvas de Kaplan Meier, permite enfrentar a una variable independiente numérica, dicotómica, nominal u ordinal a otra variable dependiente dicotómica, veamos un ejemplo:



Curva de Kaplan Meier efectuada por SPSS 9.0, en donde se observa la supervivencia en pacientes con y sin DMO hasta los 60 días de evolución en UTI ( $p < 0.001$ ). En este gráfico podemos observar claramente que la supervivencia del grupo con DMO es significativamente menor en relación a la estadía en UTI. En este caso estamos ante la presencia de un análisis actuarial bivariado. El método actuarial permite computar el valor de un estadístico denominado Log Rank y observar su grado de significación. Un valor de Log Rank elevado presenta una correspondencia con un  $p < 0,05$ , lo cual indica que la diferencia de supervivencia de los dos grupos estudiados no se

debe al azar.

También existen métodos como la Regresión de Cox que permiten efectuar un análisis de tipo actuarial multivariado. Este método mide la interacción entre dos o más variables o factores independientes sobre por ejemplo la evolución en el tiempo.

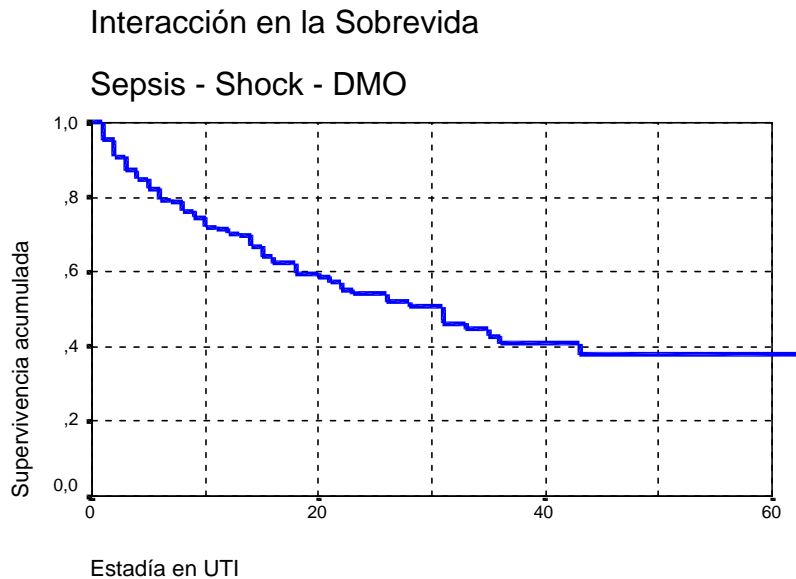


Gráfico generado con SPSS 9.0, aplicando regresión de Cox en donde se analiza la sobrevida a 60 días de UTI para pacientes con interacción Sepsis – Shock y DMO. En este caso estamos efectuando una análisis actuarial multivariado, tomamos a cada variable independiente, observamos su interacción y calculamos la sobrevida para esa interacción. Estamos enfrentando tres v. independientes categóricas a otra v. categórica dependiente.

El análisis actuarial se basa en la observación en el tiempo de estudio, que se determina de antemano, de individuos que presentaron el fenómeno problema, por ejemplo mortalidad y los que no lo presentaron, los cuales se denominan casos censurados. A partir de ahí y según la prueba utilizada se aplican formulas estadísticas que permiten determinar si las diferencias en los grupos estudiados se deben o no al azar.

Los resultados se exponen en una gráfica de dos ejes representados por los factores (eje y) y la variable de medición del tiempo (eje x).

## ANALISIS MULTIVARIADO



El análisis multivariado posee la propiedad de poder enfrentar a diferentes variables o **factores independientes**, juntas, (asociadas o no a **covariables**) con una o más v. dependientes.

Existen varios tipos de métodos multivariados, vamos a describir sintéticamente el fundamento de los más usados.

- **Regresión Lineal Múltiple:** el fundamento es el de la regresión lineal simple, pero la diferencia es que se estudia la relación entre dos o más v. independientes con una o más v. dependientes, hallándose el R<sup>2</sup> que las asocia, estableciéndose así un modelo predictivo lineal.
- **Regresión Logística:** la R logística posee una fórmula logarítmica que calcula la relación entre una o más v. independientes con una v. dependiente categórica dicotómica, como por ejemplo la evolución. Se utiliza mucho para investigar a variables predictivas de un evento determinado en la población y para la confección de modelos de scores de probabilidad.
- **Análisis Discriminante :** discrimina la pertenencia a diferentes grupos dentro de una muestra, asignándole diferentes pesos a cada variable independiente analizada y estableciendo su relación con una v. dependiente categórica nominal u ordinal generalmente.
- **Análisis Multivariado de Varianza (Manova):** utiliza la metodología del Anova con modificaciones que le permiten analizar el efecto producido por varios factores independientes sobre una o más variables dependientes categóricas.

Las pruebas multivariadas poseen fórmulas mucho más complejas y difíciles de comprender para el médico, deben ser manejadas por individuos con experiencia ya que pueden reproducir resultados ficticios al igual que cualquier prueba estadística mal utilizada.

## MODELOS PREDICTIVOS



La metodología usada en la confección de modelos predictivos se basa en la aplicación de la regresión logística para la construcción del modelo, la calibración del mismo y la discriminación.

Para comprender los resultados de la regresión logística en cada variable, vamos a revisar sus fundamentos. La regresión logística es un método de análisis multivariado, en donde se enfrentan uno o mas variables que se sospecha pueden jugar un papel como **factores de riesgo o pronósticos** (Sepsis, Edad, Apache II, etc.) a una variable dependiente que es dicotómica, es decir con dos valores posibles (Evolución en la internación: sobrevida u óbito). Por el método logístico, se averigua si esa posible variable pronostica, implica o marca realmente un pronóstico sobre la otra variable.

El fundamento se basa en que el valor de la constante y del coeficiente beta de **regresión** (modificada logísticamente) de cada variable sea significativo, o sea presente un valor de  $p < 0,05$ . Esto quiere decir que la variable en cuestión marca un pronóstico.

El valor pronosticado por un modelo, es un valor de probabilidad de que el evento ocurra, en este caso de probabilidad de sobrevida u óbito en la internación. Esta probabilidad se expresa por un valor que oscila de 0 a 1. Un valor de 0 significa probabilidad alta de óbito y el valor de 1 alta probabilidad de sobrevida. La regresión logística computa el valor de probabilidad de cada paciente, si este es inferior o igual a 0,5 lo interpreta como negativo (óbito). Si el valor es superior a 0,5 lo interpreta como positivo (sobrevida). Por lo tanto 0,5 es el punto de corte de probabilidad del evento para la regresión logística

La calibración del modelo con la realidad se estudió por medio del estadístico Goodness of Fit (Gof) de Hosmer y Lemeshow (ver bibliografía). Este método permite observar si los resultados pronosticados por el modelo (éxito o fracaso) para cada paciente individual se corresponden con el resultado real observado. Una buena calibración significa que el pronóstico dado por modelo, coincide con la realidad observada y está representado por un Gof inferior a 15,5.

El tercer paso es averiguar la **Sensibilidad y Especificidad** para un punto de corte de probabilidad determinado en el modelo.

## SENSIBILIDAD Y ESPECIFICIDAD



El análisis de la sensibilidad y especificidad se denomina discriminación y se efectúa generalmente por medio de la metodología ROC.

La gráfica de la curva ROC se construye con cada punto de S y E de cada valor de la **variable o factor (independiente)** que se está estudiando con respecto a una variable dependiente **categorica dicotómica**.

Un ejemplo típico de la construcción de una curva ROC es en la determinación de la S y E de un Sistema de Score Predictivo.

Para interpretar la curva ROC se analiza el área bajo la curva (AUC) que para una buena discriminación debe ser superior a 0.80. Para comprender la discriminación vamos a repasar los conceptos fundamentales que se comprenden claramente observando la tabla de 2x2 de outcome:

<b>Outcome Pronosticado</b>	<b>Outcome</b>	<b>Observado</b>	
	<b>NS</b>	<b>SV</b>	<b>Totales</b>
<b>NS</b>	<b>A</b>	<b>C</b>	<b>A+C</b>
<b>SV</b>	<b>B</b>	<b>D</b>	<b>B+D</b>
<b>Totales</b>	<b>A+B</b>	<b>C+D</b>	

**Sensibilidad (S):** se define como el número de NS pronosticados y verdaderos (correctos) sobre el número de NS totales.  $[A / A+B]$

**Especificidad (E):** se define como el número de SV pronosticados y verdaderos (correctos) sobre el número de SV totales.  $[D / C+D]$

**Tasa de Falsos Negativos (FN):** aquellos pronosticados como SV que fueron NS sobre el número de NS totales.  $[B / A+B]$

**Tasa de Falsos Positivos (FP):** aquellos pronosticados como NS que fueron SV sobre el número de SV totales.  $[C / C+D]$

Para valorar una prueba diagnóstica también se utilizan los conceptos de S y E, vamos a ver una tabla de ejemplo:

<b>Resultado de la prueba</b>	<b>Infarto</b>	<b>Miocardio</b>	
	<b>Presente</b>	<b>Ausente</b>	<b>Totales</b>
<b>Positiva</b>	<b>A</b>	<b>B</b>	<b>A+B</b>
<b>Negativa</b>	<b>C</b>	<b>D</b>	<b>C+D</b>
<b>Totales</b>	<b>A+C</b>	<b>B+D</b>	

**Sensibilidad:**  $a / a + c$

(n prueba positiva en enfermedad positiva / casos totales con enfermedad positiva)

**Especificidad:**  $d / b + d$

(n prueba negativa en enfermedad negativa / casos totales con enfermedad negativa)

**Valor predictivo de una prueba positiva:**  $a / a + b$

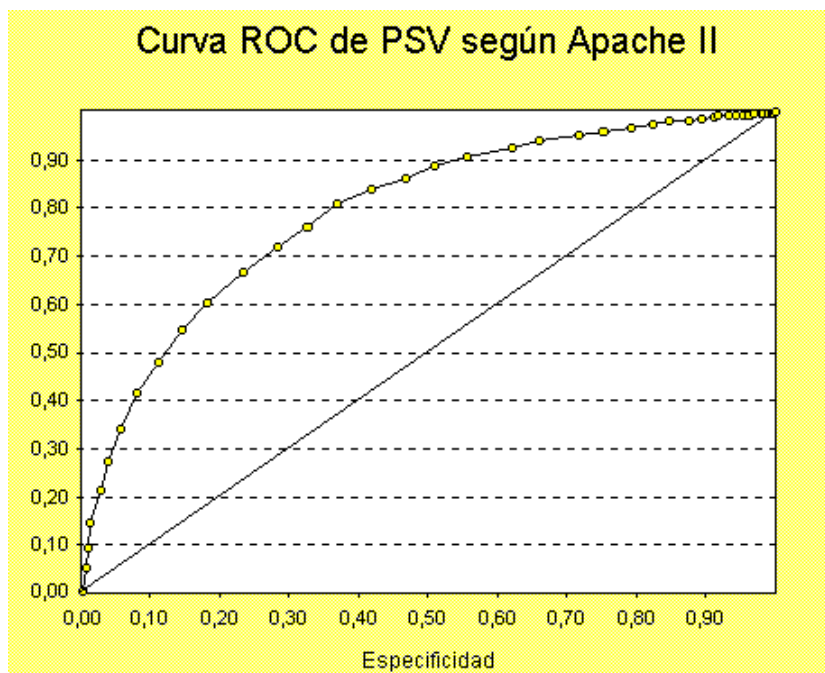
(n pruebas positivas enfermedad positiva / total de pruebas positivas)

**Valor predictivo de una prueba negativa:**  $d / c + d$

(n pruebas negativas enfermedad negativa / total de pruebas negativas)

**Prevalencia:**  $a + c / a + b + c + d$

En el gráfico siguiente generado con el programa Simstat 1.21, se observa la curva ROC de la Probabilidad de Sobrevida según puntaje Apache II en 3000 pacientes críticos. El AUC fue de 0.81, altamente satisfactoria para una buena discriminación.

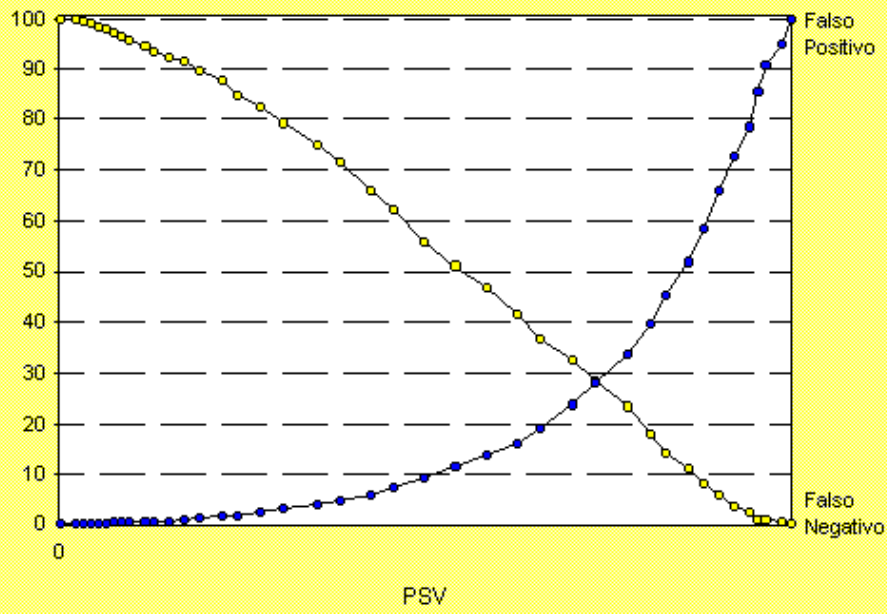


Para poder establecer la S y E se toma un punto de corte de probabilidad, por ejemplo 0.5. Aquellos pacientes con probabilidad calculada por el programa 0.5 serán tomados como sobrevivientes y viceversa.

La sensibilidad de esta muestra de pacientes para un punto de corte de 0.5 de probabilidad de sobrevida según puntaje Apache II, fue de 51% y la especificidad del 88%.

La tasa de falsos negativos y positivos también puede expresarse en un gráfico de error como se observa a continuación:

## Curvas de razón de error de PSV



# GLOSARIO ESTADISTICO



**Análisis de Varianza:** se utiliza para corroborar si la significación de diferencias entre varianzas de dos o mas grupos, son o no debidas al azar. La cifra estadística obtenida con el Anova es la razón F.

## **Análisis Estadísticos**

- En el análisis bivariado se investiga la influencia de una variable independiente por vez con respecto a la variable dependiente.
- En el análisis multivariado se investiga la influencia de dos o mas variables independientes, junto o no a una o mas variables asociadas (covariables o cofactores) sobre una o más variables dependientes.
- En el análisis univariado se describen las características de una variable por vez. También se lo llama estadística descriptiva.

**Correlación:** El concepto de correlación y regresión se basa en el grado de relación que poseen dos variables numéricas entre si.

El coeficiente de correlación permite observar, si entre dos variables existe o no una relación o dependencia. El coeficiente de regresión permite predecir matematicamente el grado de relación matemática entre dos variables.

**Desvío :** es la diferencia entre la  $m$  y c/u de los valores de la muestra ( $x$ ).

**Desvío Estándar (SD)  $s$  :** es la raíz cuadrada de la Varianza y corresponde al 68% del área de la curva alrededor de  $m$ .

**Distribución Asimétrica:** aquella que no guarda la relación de gaus en su distribución de frecuencia.

**Distribución Normal:** la distribución normal es en forma de campana, habitualmente llamada distribución de Gaus. Es simétrica en torno a su media ( $m$ ); la media, mediana y modo son iguales; el área total de la curva por encima del eje basal  $x$  es la unidad del área = 1, por lo tanto cada sector de derecha e izquierda tiene un valor de 0,5. Si se trazan líneas perpendiculares a un desvío estándar ( $s$ ) de distancia de la media, se obtiene un 68% del área de la curva. Dos desvíos estándar encierran un 95% y

**Error Estándar:** es la división entre el SD y la raíz cuadrada de  $n$ .

**Factor asociado o Cofactor:** que ayuda a otros factores a que se produzca el evento.

**Factor de Riesgo:** que implica una tendencia hacia la aparición del evento.

**Factor Pronóstico:** que desencadena la evolución hacia un evento.

**Frecuencia (f):** numero de veces en que se repite una característica o las categorías de una variable discreta. Generalmente se acompaña del porcentaje relacionado a esa frecuencia.

**HP. Alterna o Científica:** es la HP que pretende comprobar el investigador en su muestra de pacientes. Plantea la posibilidad de que exista una relación causa efecto en el evento de estudio.

**HP. Nula:** es lo contrario de la alterna o científicar, o sea que no existen relación causa efecto y el evento de estudio, se debe puramente al azar

**Individuo:** es la unidad mínima que se estudia. En medicina habitualmente es el paciente y en el caso de personas sanas se denomina sujeto o persona. También pueden estudiarse otros como: animales de experimentación, datos de laboratorio, exámenes, etc. (en estos casos se denomina observación).

Las poblaciones pueden ser clasificadas básicamente como sigue:

**Media :** es la suma de todos los valores dividido por el numero de casos n.

**Mediana:** corresponde al valor central de la serie de datos observada.

**Modo:** valor más frecuente de una serie de datos.

**Muestra:** es el grupo de pacientes u observaciones que se estudiará, la cual debe haberse elegido al azar y ser representativa de la población a la cual pertenece. En general la muestra es toda parte representativa de un conjunto, población o universo, cuyas características debe reproducir en pequeño lo más exacto posible. A partir del análisis de la muestra, obtenida correctamente y al azar, se pueden hallar conclusiones que sean extrapolables a la población de origen. Para elegir la muestra debe apelarse

**p :** es la probabilidad de error al comparar dos o más muestras o grupos cuando aseguramos que ambos son diferentes. O sea que p es la probabilidad en el sentido de la significación estadística. Obtener una  $p < 0.05$  significa que tenemos un 5% de probabilidades de error en las conclusiones, por lo cual la probabilidad de equivocarnos es baja.

**Población:** conjunto de individuos, sujetos u observaciones con alguna característica en común. Conjunto de elementos de la misma especie que se pretende estudiar en una investigación científica y de la cual se obtiene una muestra.

· Población General o Madre: población real que se pretende estudiar y a la cual se extenderán las conclusiones de la muestra perteneciente a la misma.

· Población Hipotética: conjunto formado por todas las poblaciones

**Prueba de t Student:** es un método de análisis estadístico, que compara las medias de dos categorías dentro de una variable independiente, o las medias de dos grupos diferentes. Es una prueba paramétrica.

**Variable Asociada:** se denomina así a aquella v. independiente que no modifica por su sola presencia a la v. dep

**Variable Cualitativa:** son variables que representan cualidades de la muestra, como por ejemplo la evolución del paciente hacia la mejoría o la muerte, color de ojos de un grupo de personas, sexo, etc. Estas variables también son llamadas Categóricas o Discretas.

**Variable Cuantitativa:** es la que se puede medir. Habitualmente es llamada variable Numérica o Continua, o sea que posee una continuidad. Por ejemplo la edad, hematócrito, transporte de oxígeno, altura, peso, frecuencia cardiaca o respiratoria, dosis de un medicamento.

**Variable Dependiente:** es la v. motivo de nuestro interés, cuyos valores dependen de otras variables que pueden influir en ella. También se la llama v. de respuesta. Por ejemplo la sobrevida, respuesta al tratamiento, evolución, etc.

**Variable Independiente:** es la que modifica de una u otra manera a la v. dependiente, llamándose también según el caso factor de riesgo, factor predictivo, etc.

**Variable:** es una característica o propiedad determinada del individuo, sea medible o no. Esta propiedad hace que las personas de un grupo puedan diferir de las de otro grupo en la muestra o población.

**Variables Categóricas Dicotómicas:** son las que tienen dos valores fijos y excluyentes entre si como la evolución, presencia o ausencia de una enfermedad o característica en la muestra.

**Variables Categóricas Dicotómicas:** son las que tienen dos valores fijos y excluyentes entre si

como la evolución, presencia o ausencia de una enfermedad o característica en la muestra.

**Variables Categóricas Nominales:** son variables cualitativas que no permiten establecer un orden, por ejemplo la raza, que puede ser blanca, negra, caucásica, etc., o los grupos sanguíneos A, B, AB o 0. También son excluyentes entre sí, o sea que cada paciente pertenece a una u otra categoría pero no a dos al mismo )

**Variables Categóricas Ordinales:** estas si permiten establecer un orden determinado, por ejemplo los grupos de Apache son I a IV, un paciente del grupo II tiene menor probabilidad de mortalidad en UTI que el del grupo IV. La clasificación de la Disnea en grados ( I a IV) es otro ejemplo. También son excluyentes entre sí.

**Varianza:** es el promedio de los cuadrados de los desvíos; la Varianza  $n$  mide la dispersión de los valores y marca el punto de inflexión de las curvas en los histogramas.

## BIBLIOGRAFIA



1. MH. Kolef; DP. Schuster – Predicting intensive care unit outcome with scoring systems: Underlying concepts and principles. – Crit. Care Clin.; 10:1; 1-18; 1994
2. UE. Ruttimann – Statistical approaches to development and validation of predictive instruments. - Crit. Care Clin.; 10:1; 19 -36; 1994
3. P. Armitage; G. Berry – Estadística para la Investigación Biomédica – Tercera Edición; Editorial Harcourt Brace; Barcelona 1997
4. Marija J. Norusis; Advance Statistics 6.0 – SPSS Inc. Chicago, IL. – 1993
5. Marija J. Norusis; Base Statistics 6.0 – SPSS Inc. Chicago, IL. – 1993
6. R. Alvarez Cáceres; Estadística multivariante y no paramétrica con SPSS – Editorial Diaz de Santos. – Madrid, 1995
7. Gonzalez López Varcacel R; Analisis multivariante, aplicación al ámbito sanitario. – SG Editores, Barcelona 1991.
8. Bisquerra, R; Analisis multivariante. - Editorial PPU, Barcelona 1991.
9. Etxeberria, J.; Joaristi, L; Lizasoain, L; Programación y análisis estadísticos básicos con SPSS / PC+. – 3ra Edición - Editorial Paraninfo SA. – Madrid, 1995
10. Bakke OM; Carné X; García Alonso F; Ensayos clínicos con medicamentos. – Fundamentos básicos y metodología. – Doyma – Barcelona, 1994
11. Elston R; Johnson WD; Principios de Bioestadística. – Manual Moderno. – México, 1990
12. Daniel W; Bioestadística. – 3ra edición – Editorial Limusa, Grupo Noriega Editores. – México, 1993
13. Norman G; Streinar D; Bioestadística. – Mosby Doyma Libros SA. – Madrid, 1996
14. Sackett D; Haynes R; Guyatt G; Tugwell P; Epidemiología Clínica. – 2da edición - Editorial Médica Panamericana – Bs. As., 1994
15. Dawson Sanders B; Trapp R; Bioestadística Médica. – Manual Moderno – México, 1993
16. Daniels S; Flanders W; Eley J; Boring J; Epidemiología Médica. – Manual Moderno – México, 1995
17. Day R; Como escribir y publicar trabajos científicos. – 3ra edición. – Organización Panamericana de la Salud. – The Oryx Press, 1990
18. Bertranou E; Manual de Metodología de la Investigación Clínica. – Librería Akadia Editorial. – Bs. As., 1995
19. Coronado JL; Corral A; López P; Estadística aplicada con Statsgraphics – Editorial Ra-Ma – MADRID, 1994
20. Canales FH; Alavarado EL; Pineda EB; - Metodología de la Investigación – Organización Panamericana de la Salud - 1986
21. Marija J. Norusis - SPSS Professional Statistics 7.5 – SPSS Inc., Chicago, IL, 1997.

22. Marija J. Norusis - SPSS Advanced Statistics 7.5 – SPSS Inc., Chicago, IL, 1997.
23. S. Lemeshow; D. Hosmer - A Review of goodness of fit statistics for use in the development of logistic regression models. *Am. Journal of Epidemiology*; Vol.115, N°1:92-106; January 1982.
24. R. Sierra Bravo: - *Diccionario Practico de Estadística*. – Ed. Paraninfo, Madrid, 1991.