INTRODUCCIÓN
TÉCNICAS ESTADÍSTICAS
APLICADAS A
CIENCIAS DE LA SALUD

Pedro Cuesta Álvaro
Apoyo a Investigación
Servicios Informáticos UCM

INTRODUCIÓN

- **Necesidad de la estadística**. Las Ciencias de la Salud son esencialmente Experimentales

- Necesidad de **razonamientos inductivos** a partir de datos: Se hacen afirmaciones acerca de un colectivo de individuos u objetos, habiendo observado en realidad sólo una parte de ellos.

- Definición de Estadística:

  Conjunto de métodos necesario para recoger, clasificar, representar y resumir datos, así como para hacer inferencias científicas a partir de ellos.

- ***Estadística Descriptiva***. Con el estudio de ciertos estadísticos se conocen magnitudes que representan a la globalidad de los datos disponibles de forma resumida

- ***Inferencia Estadística***. La segunda fase es la formulación y confirmación de hipótesis, Se cuantifica el grado de certidumbre con el que se pueden establecer afirmaciones sobre los datos: Se obtienen conclusiones a partir de una información incompleta

- ***Población***: conjunto de los objetos que se desean estudiar (Todos los pacientes hipertensos del mismo medio)

- ***Muestra***: parte de esa población seleccionada para el experimento (Pacientes a los que se les suministra un tratamiento para la hipertensión)

- El análisis de la información será de gran ayuda para la **toma de decisiones** y la realización de investigaciones

- No hay que olvidar que los datos disponibles suministrarán una información parcial del proceso en estudio y aunque la estadística valide unas hipótesis, el investigador deberá dar un **significado real** a las conclusiones en el contexto correspondiente.

Algunos conceptos básicos:

- ***Observación***: una observación es un objeto individual que nos sirve como fuente de datos para la realización de nuestra investigación. Reciben diferentes denominaciones: Unidades muestrales, Individuos, Observaciones, Casos, Objetos, Unidades experimentales, …

- ***Variable***: Es una característica del individuo que puede tomar distintos valores. Cuando medimos algo representamos por un modelo numérico aquello que medimos. Por ejemplo, la altura de una persona: asignamos un número a cada persona. Las medidas físicas, como altura y peso, se miden con un instrumento físico. Otras propiedades abstractas tales como razonamiento, depresión, inteligencia se miden indirectamente.

- ***Valor***: son los distintos estados en los que se puede encontrar una característica de un individuo. Estos pueden ser cualitativos (masculino, femenino) o cuantitativos (163 cm).

PROBLEMAS TÍPICOS

1. Determinar las Unidades Experimentales o Unidades Muestrales

   - Paciente o localización en Odontología
   - Huesos: derecho, izquierdo, …
   - Familias o personas en encuestas

2. Homogeneidad respecto a otras características que puedan influir

3. Obtención de las medidas: Grandes errores

ORGANIZACIÓN RECTANGULAR

En general, los datos a analizar consistirán de un conjunto de *p* variables medidas en *n* unidades muestrales.

Grabados en Hojas de cálculo (EXCEL, LOTUS), Bases de Datos (DBASE, ACCESS) o Programas Estadísticos (STATGRAPHICS, SPSS)

NUMERO	SEXO	EDAD	EJERCICIO	ALCOHOL	TABACO	…
001	H	67	3	35	0	
003	M	76	1	56	10	
004	H	56	2	112	15	
005	M	63	4	67	25	
…						

Diferentes — Visitas, Localizaciones, Momentos de tiempo → No suponen nuevas muestras. Son nuevas medidas en la misma muestra

CLASIFICACIÓN DE VARIABLES

Se pueden considerar tres clasificaciones de variables:

1. Según la escala: Nominal, Ordinal, de Intervalo y de Razón.
2. Cualitativas y Cuantitativas
3. Discretas y Continuas

Según la escala:

Una clasificación comúnmente aceptada especifica cuatro tipos de variables: nominal, ordinal, intervalo, de razón.

Variables nominales

Una escala nominal es un sistema de clasificación que sitúa a personas, objetos u otras entidades dentro de categorías mutuamente excluyentes

Podemos usar símbolos (`H/M , SI/NO`) para representar las dos categorías.

Algunos programas de análisis de datos tratan sólo símbolos numéricos, por lo que es preferible esta representación. Puesto que las categorías pueden considerarse en cualquier orden cualquier conjunto de números será válido para su representación: 0/1, 1/2 (para no confundir ceros con blancos), 1/6 (para evitar errores de grabación).

Variables Ordinales

En este caso se usan categorías, pero existe un orden conocido entre ellas. Por ejemplo una escala de niveles de dureza de minerales, un estatus socioeconómico, etc. Puede usarse cualquier secuencia de números crecientes para su representación. Para definir una variable ordinal la operación básica es determinar si una observación es mayor que otra.

Variables de intervalo

Una variable intervalo es una variable ordinal especial, en la que las diferencias entre dos valores sucesivos es siempre la misma. Por ejemplo, la variable temperatura en grados Fahrenheit.

Variables de razón.

Son variables de intervalo en las que además hay un punto natural representando el origen: punto cero. Por ejemplo, la altura.

Cualitativas y Cuantitativas:

Las variables *cuantitativas* son aquellas en la que los valores son números. Cuantifican características que unos poseen en mayor **cantidad** que otros

En las *cualitativas*, también llamadas categóricas o de clasificación, los diferentes valores representan **grupos** distintos a los que el sujeto puede pertenecer.

Continuas y Discretas:

Una variables es continua si puede tomar cualquier valor en un rango específico. Ejemplo altura, peso, densidad, tiempo, resistencia,… .

Una variable que no es continua es discreta. Puede tomar sólo ciertos valores específicos. Ejemplo: número de hijos, sexo, partido votado. A veces a las variables de este tipo se les denomina también atributos.

Esta última clasificación lleva, posteriormente, a considerar las posibles **distribuciones de las variables** que se suponen en los análisis. De esta forma una variable discreta puede seguir una distribución Binomial, de Poisson, etc., mientras que la distribución Normal se usa para describir la distribución de las variables continuas.

ENTRADA DE DATOS

NUMERO	SEXO	EDAD	EJERCICIO	ALCOHOL	TABACO	ALIM_GRA	COLEST	ANT_FAM	PROBCOR
001	H	67	3	35	0	600	185	1	0
003	M	76	1	56	10	690	210	2	1
004	H	56	2	112	15	-1	195	1	1
005	M	63	4	67	25	650	200	2	0
006	H	55	1	-1	0	750	230	9	-1
Identif	**CUALI**	**CUANT**	**ORDINAL**					**CUALIT**	**CUALIT**

1. Nombres de las variables, nombres de códigos para descodificación

2. Codificar preferiblemente en números

3. Codificación detallada. Ejemplo: Edad, Tabaco con valor exacto. No intervalos que se pueden generar posteriormente.

4. Chequeo de rangos, máximos, mínimos,… en formularios de entrada (ACCESS, EXCEL)

5. Valores Missing: definición, codificación.
   - Los análisis multivariantes requieren casos completos
   - Valorar la supresión de un caso o una variable con alta proporción de valores missing

6) Copias de seguridad

7) Chequeo inicial: frecuencias, máximos y mínimos, gráficas, detectar valores no admisibles, inconsistentes, errores, etc.

CONTRASTES DE HIPÓTESIS

Se formulan hipótesis acerca de leyes o fenómenos físicos o naturales, que es necesario demostrar o rechazar por medio de "contrastes" (tests) o "pruebas". La prueba de la hipótesis es el ***Contraste*** de la hipótesis, lo que nos llevará a su aceptación o rechazo.

El procedimiento estándar consiste en recopilar información en forma de observaciones numéricas que serán la base de nuestra decisión. Por ejemplo si tiramos una moneda 100 veces y obtenemos siempre cara podemos percibir que la hipótesis de que la moneda no está trucada no es aceptable. Sin embargo es posible obtener este resultado con una moneda no trucada, por consiguiente no podremos estar completamente seguros de nuestra decisión.

Los procedimientos de ***Inferencia Estadística*** nos posibilitan, bajo ciertas hipótesis, establecer la probabilidad de aceptar hipótesis falsas o rechazar hipótesis verdaderas. Es decir permiten calcular **la probabilidad de cometer error con nuestra decisión.**

El objetivo de un contraste de hipótesis es comprobar si los datos muestrales apoyan la hipótesis nula, o por el contrario rechazan H_0 , lo cual nos llevaría a aceptar H_1. En un enfoque totalmente práctico hay que tener en cuenta dos cosas:

a) La hipótesis nula que se contrasta

b) el *p*-valor obtenido.

Se puede interpretar el *p*-valor de dos formas:

i. La probabilidad de error (o sea de equivocarse) si se rechaza la hipótesis nula cuando realmente es cierta. Es el error llamado de Tipo I
ii. La probabilidad de que las diferencias observadas sean debidas al azar.

Por ese motivo se rechaza la hipótesis nula cuando el *p*-valor es pequeño. El valor fijo a partir del cual el *p*-valor se considera pequeño es el nivel de significación α (0.10, 0.05, 0.01, 0.001).

MUESTREO

El requerimiento básico de una muestra es que sea representativa de la población

La forma de seleccionar los individuos que han de constituir la muestra tiene, como es lógico, una importancia capital para garantizar que ésta permita obtener conclusiones que puedan extrapolarse válidamente a la población de la que la muestra procede. No hay que olvidar nunca que el objeto final del estudio es siempre la población y que la muestra es sólo un medio para obtener información sobre ésta

Con el fin de permitir inferir conclusiones válidas sobre una población la muestra debe ser **'representativa'** de ésta En teoría la única forma de garantizar la representatividad de una muestra es seleccionando al **azar** los individuos. Aunque esta forma de proceder rara vez sea aplicable de forma estricta en la práctica, siempre hay que extremar las precauciones para que la forma real de obtener la muestra sea lo más parecida posible a la ideal.

En realidad en muchos casos un **conocimiento previo sobre la población** es indispensable para decidir si una muestra puede considerarse o no representativa de la misma.

Algunas técnicas de muestreo son:

- Muestreo aleatorio simple con reposición: Una vez seleccionado un individuo es reincorporado a la población.

- Muestreo aleatorio simple sin reposición: Una vez seleccionado un individuo no se reincorpora a la población para la siguiente extracción.

- Muestreo estratificado. Se particiona la población en estratos, buscando homogeneidad. La Afijación de los tamaños muestrales en cada estrato se lleva a cabo en función de unos pesos. Afijación proporcional cuando los pesos son los tamaños del estrato.

- Muestreo sistemático. Por ejemplo para datos ordenados. Se selecciona el primero o un caso aleatoriamente y el resto se selecciona cada un número fijo determinado por el tamaño de la población y la muestra.

Tamaño Muestral

Fijada una variable y conocida su varianza o una cota puede determinarse el tamaño muestral adecuado para estimar parámetros como una media o una proporción con una determinada precisión. Por ejemplo:

$$n \geq \frac{1.96^2 \cdot \sigma^2}{\varepsilon^2}$$

para estimar la media con una precisión de ε con una confianza del 95% de una población de varianza σ^2.

y

$$n \geq \frac{1.96^2 \cdot p(1-p)}{\varepsilon^2}$$

para estimar la proporción con una precisión de ε con una confianza del 95% de una población.

Ejemplo: Se sabe que el porcentaje de enfermos está entre un 0.05 y 0.10. ¿Qué número hay que examinar para estimar la proporción de enfermos con una precisión de ε=0.02? Solución

$$n \geq \frac{1.96^2 \cdot 0.10 \cdot 0.90}{(0.02)^2} = 864.36 \Rightarrow n = 865$$

ANÁLISIS DE DATOS

Un análisis de datos suele seguir tres fases:

- **Análisis exploratorio de los datos**: Estadística descriptiva de cada variable por separado. Se obtienen medidas de tendencia central, variabilidad, representaciones gráficas, etc. Se pretende conocer cada variable así como detectar errores, valores extremos, etc.

- **Estadística Bivariable**: Estudia las relaciones entre pares de variables, utilizando estadísticos como el coeficiente de correlación, chi-cuadrado, *t* de Student, etc. y representaciones gráficas diversas.

- **Análisis multivariante**: Analiza simultáneamente dos o más variables. Los métodos pueden ser ***predictivos*** cuando existe una variable criterio o independiente que se explica o identifica por un conjunto de variables independientes, predictoras o explicativas (Regresión Lineal, Regresión Logística, Análisis Discriminante, Árboles de Segmentación, Análisis de la Varianza) o ***reductivos*** cuando se estudian las relaciones entre un conjunto de variables o casos sin que exista una variable a identificar (Componentes Principales, Análisis Factorial, Correspondencias Binarias, Correspondencias Múltiples).

Cómo se usan las variables en el análisis

Las variables pueden ser definidas para medir una determinada salida o respuesta o bien para explicar por qué se obtiene una determinada salida. Por ejemplo en el estudio de una enfermedad, las variables edad, antecedentes, severidad del estado, tratamiento son variables *explicativas* o ***independientes***. La variable discreta sana/no-sana es la variable *a explicar* o ***dependiente***.

En ciertos análisis exploratorios todas las variables se usan como un **único conjunto**, sin distinción entre independientes y dependientes.

Selección del análisis apropiado

Hay dos motivos por los que resulta difícil la **elección** de la técnica estadística adecuada para un investigador con datos reales.

El primero es que los cursos y libros estadísticos se presentan en un orden lógico desde el punto de vista de la enseñanza de las materias, pero no desde el punto de vista del **proceso del análisis de datos**. La segunda es que los datos reales contienen mezclas de tipos de datos que hacen la elección del análisis arbitraria.

Un problema cada vez más acentuado es que se suele elegir la técnica que se conoce, o la que está incorporada al **programa informático estadístico** que se dispone. Es muy raro aplicar una técnica de un libro o artículo, aunque sea la adecuada, si no está implementada en un software asequible.

Una buena estrategia consiste en aplicar **diferentes análisis** al mismo conjunto de datos, lo que nos proporcionará información variada sobre el fenómeno en estudio.

Para decidir el análisis apropiado clasificamos las variables como sigue:

1. **Independientes** frente **Dependientes**
2. **Nominal** u **ordinal** frente **intervalo** o **razón**

En la Tabla se muestran los análisis estadísticos más usuales en función del tipo de variables.

Hay que tomar decisiones especiales cuando el investigador tiene, por ejemplo, una variable dependiente de intervalo junto con cinco variables independientes de las cuales tres son de intervalo, una ordinal y una nominal.

Tabla. Análisis estadístico sugerido.

	Variables Independientes			
	Nominal u Ordinal		**Intervalo o Razón**	
Variables dependientes	**1 variable**	**> 1 variable**	**1 variable**	**> 1 variable**
No variables dependient.	Ajuste χ^2	Medidas de asociación Modelo log-lineales Contraste χ^2 de independencia	Estadísticos univariantes (ej.: *t* para una muestra) Medidas descriptivas Contraste de normalidad	Matriz de correlación Componentes principales Análisis factorial Análisis Cluster
Nominal u Ordinal				
1 variable	Contraste χ^2 Contraste exacto de Fisher	Modelo log-lineales Regresión logística Segmentación	Función discriminante Regresión logística Estadísticos univariantes (*t*)	Función discriminante Regresión logística
> 1 variable	Modelo log-lineales	Modelo log-lineales	Función discriminante	Función discriminante
Intervalo o Razón				
1 variable	Contraste *t* Análisis Varianza Análisis Supervivencia	Análisis Varianza Segmentación Análisis Supervivencia	Regresión lineal Correlación Análisis Supervivencia	Regresión lineal múltiple Análisis Supervivencia
> 1 variable	Análisis multivariante Varianza Análisis varianza en Componentes Principales T^2 Hotelling Análisis Perfiles	Análisis multivariante Varianza Análisis varianza en Componentes Principales	Correlación canónica	Correlación canónica Análisis Path Modelos estructurales (LISREL, EQS)

ANALISIS BIVARIANTE.
RELACION ENTRE DOS VARIABLES

INDICE

Se estudian las relaciones entre dos variables. La técnica y contrastes estadísticos utilizado dependerá del tipo de las variables consideradas, según se resume en la tabla siguiente:

2 cuantitativas	Correlación. Regresión Lineal Simple
2 cualitativas	Tabulación Cruzada. Porcentajes. Estadístico de la χ^2 y otros

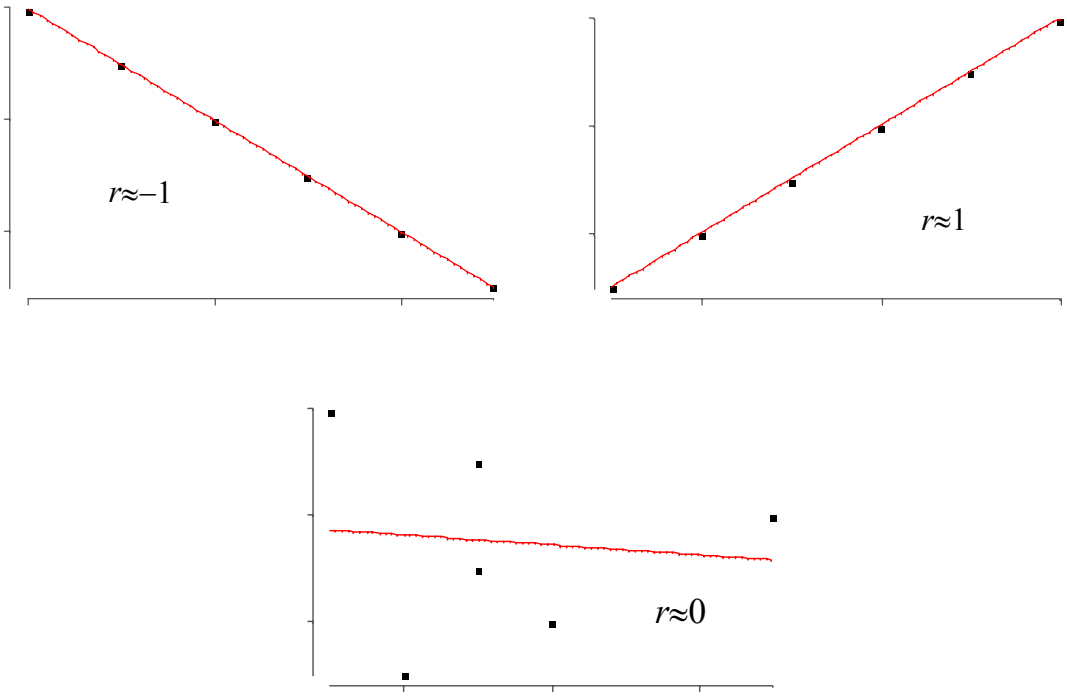
1 cuantitativa **1 cualitativa** Se constrastan diferencias entre los valores medios. Globalmente y pareados	**ANOVA (análisis de la varianza)**		
		Paramétrico	**No paramétrico**
	2 grupos	T de Student	Wilcoxon-Mann-Whitney
	Más de 2 grupos	Test F	Kruskall-Wallis

1. DOS VARIABLES CUANTITATIVAS

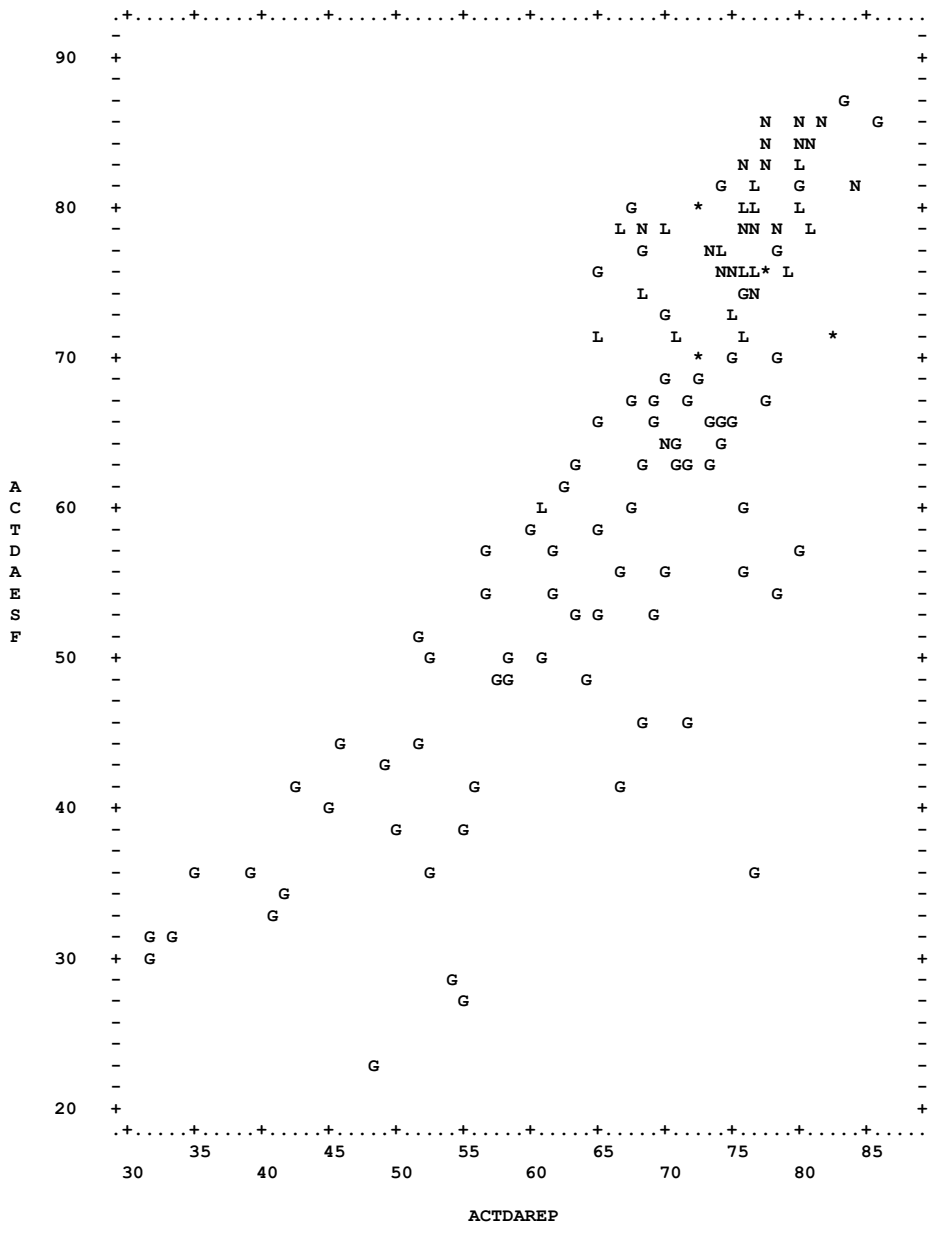
El **coeficiente de correlación de Pearson** indica si hay una relación lineal entre las dos variables $r \in [-1, 1]$.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y}$$

Una correlación positiva indica que cuando los valores de una variable se incrementan los de la otra variable tienden a incrementarse.



En este ejemplo se ve gráficamente la relación de dos variables ACTDAESF y ACTDAREP con una variable cualitativa *DA*:(*N*: Normal, *L*: Leve, *G*: Grave). Se observa que ACTDAESF < 70 está asociada con el grupo *G* independientemente del valor de ACTDAREP. Se deduce también la imposibilidad de discriminar entre *N* y *L* con las dos variables.



2. DOS VARIABLES CUALITATIVAS

Se analiza el grado de asociación entre dos variables categóricas. En los programas informáticos hay números estadísticos para tablas de frecuencias generales, tablas 2×2, tablas con categorías ordenadas, etc.

Chi-Cuadrado

Ejemplo TABLA DE ALCOHOL POR TABACO

ALCOHOL TABCANT

Frecuencia Celda Chi-cuadra Fila p.c.	T_0	T_1-10	T_11-20	_>20	Total
A<=33	22 6.7233 37.93	11 0.0748 18.97	21 2.1682 36.21	4 0.7856 6.90	58
A34-100	26 0.5015 25.24	17 0.0543 16.50	53 0.0527 51.46	7 1.4697 6.80	103
A101-200	18 0.4535 18.75	21 1.0691 21.88	49 0.0269 51.04	8 0.5041 8.33	96
A>200	12 4.22 12.24	13 0.9896 13.27	54 0.5403 55.10	19 6.9034 19.39	98
Total	78 21.97	62 17.46	177 49.86	38 10.70	355 100.00

ESTADÍSTICOS PARA TABLA DE ALCOHOL POR TABCANT

Estadístico	DF	Valor	Prob
Chi-cuadrado	9	26.537	0.002

Odds Ratio

	Fuma ≤ 10	Fuma > 10
Al ≤ 100	$n_{11} = 40$	$n_{12} = 44$
Al > 100	$n_{21} = 31$	$n_{22} = 71$

Odds ratio, también llamado CROSS-PRODUCT RATIO

$$\frac{n_{11} * n_{22}}{n_{12} * n_{21}} = \frac{n_{22}/n_{21}}{n_{12}/n_{11}} = \frac{n_{22}/n_{12}}{n_{21}/n_{11}} = \frac{\text{Tasa fuma en Ed} > 21}{\text{Tasa fuma en Ed} \leq 21}$$

Si no hay asociación entre las dos variable categóricas vale 1.

En el ejemplo posterior, dentro de la categoría Alcohol > 100 la tasa de fumar más de 10 es 71/31. Dentro de Alcohol < 100 la tasa es 44/40.

El odds-ratio para estas dos categorías es el cociente $\frac{71/31}{44/40} = 2.082$

Las tasas (odds) de fumar mucho es dos veces más en los que beben mucho.

Se calcula un intervalo de de confianza para contrastar si el odds-ratio es uno.

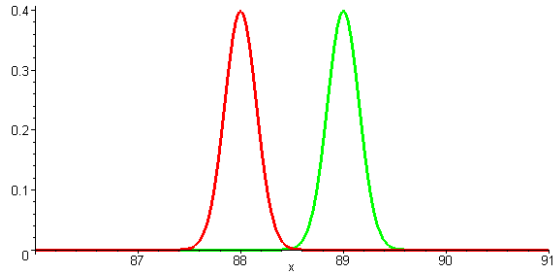
3. ANALISIS DE LA VARIANZA

3.1 Dos muestras. TEST T

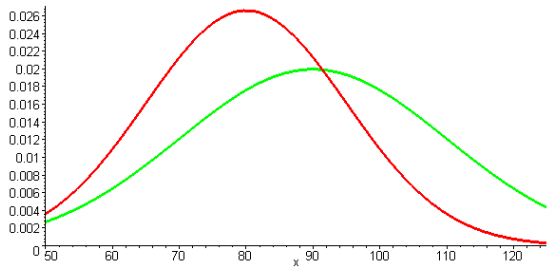
Medimos las diferencias entre dos conjuntos de datos: Resultados producidos por dos procesos de producción, dos tratamientos médicos, opiniones de dos grupos de personas, etc.

Se examinan las diferencias de las medias, teniendo en cuenta la variabilidad de cada conjunto de datos. Dos medias serán significativamente diferentes o no dependiendo de las variaciones.

Por ejemplo:



$\bar{x}_1 = 88$ y $\bar{x}_2 = 89$ pueden ser diferentes si s.e.m. son pequeñas.



$\bar{x}_1 = 80$ y $\bar{x}_2 = 90$ pueden no ser diferentes si s.e.m. son grandes.

Se realizan contrastes de localización de medias para una o dos muestras. En el análisis de dos muestras, éstas pueden ser independientes o dependientes (datos pareados).

Dos muestras independientes pueden aparecer, bien en un experimento donde hay una asignación aleatoria a dos tratamientos o bien en diseños clínicos donde se compararn dos grupos. Muestras dependientes o pareadas son frecuentemente dos medidas en el mismo sujeto realizadas en dos momentos de tiempo o bajo diferentes condiciones. Tambien pueden ser dos sujetos diferentes que se seleccionan en un par por ser homogéneos respecto al resto de características que pueden influir en la que se contrasta.

Muestras independientes

Se comparan los valores medios de los datos de dos grupos:

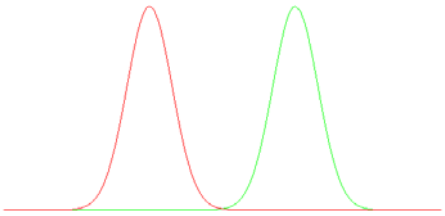
	CONTROL	PRUEBA
	90	100
	93	103
	87	104
	89	99
	90	102
Medias	88.6	101.6
Desviación	7.3	2.07
s.e.m.	3.26	0.927

H_0: Control y Prueba tienen la misma media y las diferencias muestrales observadas son debidas al azar.
Se contrasta con T de Student. Los estadísticos T para las medias contrastan la hipótesis nula

$$H_0:\mu_1-\mu_2=0 \quad \text{frente a} \quad H_1:\mu_1-\mu_2\neq 0$$

Existen dos formas:

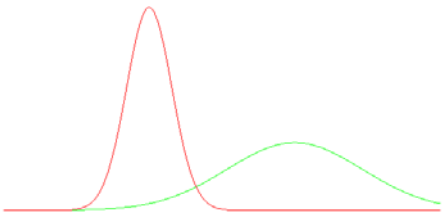
Varianzas iguales: En `EQUAL T` se estima una varianza conjunta como una media ponderada de las varianzas dentro de cada grupo



$$t_{ss}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}}\approx t_{n_1+n_2-2},$$

$$s_p^2=\frac{(n_1-1)^2 s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Varianzas no iguales. En `UNEQUAL T` no se asume igualdad de varianzas en las dos subpoblaciones. La estimación de la varianza de la diferencia de medias es la suma de las varianzas de cada media muestral, por lo que el estadístico t vale



$$t_{ss}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}=-3.83\approx t_{n_1+n_2-2}$$

Ejemplo muestras independientes: Diamétros de un hueso en hombres y mujeres.

Diametro	1	2	3	4	5	6	Media
HOMBRE	29.9	11.4	25.3	16.5	21.1		20.84
MUJER	26.6	23.7	28.5	14.2	17.9	24.3	22.53

La diferencia de medias es −1.693, que no es significativa.

```
 TTEST PROCEDURE

Variable: POSTEST

ALCOHOL       N           Mean        Std Dev      Std Error        Minimum        Maximum
-------------------------------------------------------------------------------------------
 A<=100     161   114.53416149    12.55738381     0.98966049    82.00000000   148.00000000
  A>100     194   117.74226804    14.56646610     1.04581132    84.00000000   182.00000000

Variances        T       DF    Prob>|T|
----------------------------------------
Unequal    -2.2281    352.5      0.0265
Equal      -2.1977    353.0      0.0286

For HO: Variances are equal, F' = 1.35    DF = (193,160)    Prob>F' = 0.0520


*******************************************************************************************

Variable: DINAEST

ALCOHOL       N           Mean        Std Dev      Std Error        Minimum        Maximum
-------------------------------------------------------------------------------------------
 A<=100     161   144.22360248    20.56330930     1.62061580    98.00000000   206.00000000
  A>100     193   147.76165803    21.32206363     1.53479572    60.00000000   216.00000000

Variances        T       DF    Prob>|T|
----------------------------------------
Unequal    -1.5851    344.7      0.1139
Equal      -1.5799    352.0      0.1150

For HO: Variances are equal, F' = 1.08    DF = (192,160)    Prob>F' = 0.6363
```

Datos pareados

Comparaciones pareadas::

x_a	x_d	Diferencia
		$x_a - x_d$
		$\bar{d}, s_d^2$

El estadístico t se calcula por

$$t = \frac{\bar{d}}{\sqrt{\frac{s_d^2}{n}}} \approx t_{n-1}$$

Pueden obtenerse dos test no parámetricos

1. El test de los signos (SIGN TEST) no usa rangos sino el número de diferencias positivas N_+ y negativas N_- , realizando un contraste binomial bajo la hipótesis nula de igual probabilidad de obtener los signos + y −.
2. El test de WILCOXON (Signed Rank) calcula, en este caso, la suma de los rangos de las N_+ diferencias positivas, R_+ , la suma de los rangos de las N_- diferencias positivas, R_- y realiza la inferencia con R_{min} = Minimo(R_+,R_-) suponiendo que la población es continua y simétrica.

Ejemplo datos pareados: Se estudia las medidas de un hueso, considerando Derecho e Izquierdo en 10 personas:

Hueso	1	2	3	4	5	6	7	8	9	10
Long_D	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
Long_I	14.0	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

La diferencia de medias es –0.410 con un *valor-t* significativo.

```
Variable=DIF

 N            10      Sum pond      10      Núm ^= 0      10   Núm > 0        2
 Media     -0.41      Suma         -4.1     M(Signo)      -3   Pr>=|M|    0.1094
 Std Dev    0.387     Varianza    0.149     Sgn Rang   -24.5   Pr>=|S|    0.0078
 CV       -94.428     Std Mean    0.122

 T:Med=0   -3.34888   Pr>|T| 0.0085
```

3.2 Análisis de la Varianza de un Factor

Extendemos la comparación de dos medias al caso en el que una variable categórica establece más de dos grupos o se examina el efecto de más de una variable categórica (factores) en las medias de una variable respuesta.

Ejemplo: Se considera el tiempo de recuperación en 60 personas sometidas a 4 terapias.

Variables: TIEMPOR, TERAPIA(A,B,C,D)

Llamamos "*tratamientos*" a los niveles de la variable categórica (TERAPIA). En el análisis de la varianza contrastamos la hipótesis de que los tratamientos tienen el mismo efecto.

Suponemos que el mecanismo de respuesta de la variable TIEMPOR es

$$\hat{Y}_j = (\textit{media global}) + (\textit{efecto tratamiento}_j)$$

$\hat{Y}_j$ es el valor ajustado al j-ésimo tratamiento.

En ausencia de tratamientos la mejor representación para una respuesta típica es la media global de la muestra. Si los tratamientos están presentes hacemos ajustes en esta media según el efecto de cada tratamiento.

Si todos los tratamientos tienen el mismo efecto no es necesario hacer ajustes y el efecto común se incorporaría a la media global y

variación total = variación debido al error (*respecto a la media*)

Si los efectos son diferentes

variación total = (*variación debido a tratamientos)* +
(*variación debido al error*)

Se deben inspeccionar gráficamente los tiempos en cada terapia para observar diferentes efectos de los tratamientos.

Suma de cuadrados Entre y Dentro

Ejemplo. Respuestas en tres tipos.

	A	B	C
1	6	8	0
2	4	11	2
3	3	5	1
4	3	4	1
Medias	4	7	1

Media total = 4.

¿Son las tres medias significativamente diferentes?
Se calculan tres sumas de cuadrados:

Suma de Cuadrados Total:

$$(6-4)^2+(4-4)^2+\ldots+(1-4)^2+(1-4)^2=110\,.$$

Suma de Cuadrado Dentro:

$$\left.\begin{array}{l}\mathrm{A}:(6-4)^2+(4-4)^2+(3-4)^2+(3-4)^2=6\\ \mathrm{B}:(8-7)^2+(11-7)^2+(5-7)^2+(4-7)^2=30\\ \mathrm{C}:\ (0-1)^2+(2-1)^2+(1-1)^2+(1-1)^2=\ \ 2\end{array}\right\}=38$$

Suma de Cuadrados Entre:

$$4(4-4)^2+4(7-4)^2+4(1-4)^2=72$$

$$\text{Se compara}\ \frac{\text{Entre}}{\text{Dentro}}=\frac{72/2}{38/9}=F$$

Interesa numerador grande y denominador pequeño para que las diferencias observadas dentro de cada grupo sean pequeñas respecto a las diferencias observadas entre los grupo. De esta forma admitimos que el tipo influye en la variable respuesta.

Se realiza un test estadístico comparando el valor *F-VALUE* con una distribución F con $(k-1,n-1)$ grados de libertad.

El nivel de significatividad calculado nos indica la probabilidad de cometer error si rechazamos la igualdad de efectos y decidimos que los tratamientos son diferentes.

Comparaciones múltiples

Cuando el valor global F es significativo nos indica que las medias en los grupos no son iguales. Los test de comparaciones pareadas y de rangos múltiples permiten analizar qué medias específicas difieren.

Se comparan los grupos de dos en dos.
Tukey, *Bonferroni*, *Scheffé*, *Dunnet*

Se obtienen intervalos de medias.
Duncan, *Student-Newman-Keuls*

```
PROC ANOVA ;
 CLASS SOBREPES ;
 MODEL DINAEST= SOBREPES ;
 MEANS SOBREPES /SNK SCHEFFE
```

```
                          Analysis of Variance Procedure
                             Class Level Information

                        Class    Levels    Values
                        SOBREPES      4    A> 15 S-5A+5 S5A15 S<=-5

                          Number of observations in data set = 355
NOTA: Due to missing values, only 351 observations can be used in this analysis.

Dependent Variable: DINAEST

Source                   DF           Sum of Squares             Mean Square    F Value     Pr > F

SOBREPES                  3           13599.69008461           4533.23002820     11.08     0.0001

Error                   347          141962.79424588            409.11468082

Corrected Total         350          155562.48433048

                         Student-Newman-Keuls test for variable: DINAEST

                              Alpha= 0.05  df= 347  MSE= 409.1147

                          Number of Means         2         3         4
                          Critical Range  6.6536253 7.9627851 8.7332115

                  Means with the same letter are not significantly different.

                         SNK Grouping              Mean      N  SOBREPES

                                   A           157.191     47  A> 15

                                   B           149.212     99  S5A15
                                   B
                                   B           144.658    149  S-5A+5

                                   C           135.286     56  S<=-5
```

TEST NO PARAMETRICOS

Los contrastes T de Student y F suponen, entre otras cosas, que las distribuciones son normales.

Cuando este no es el caso se construyen otro tipo de contrastes que trabajan con rangos.

Ejemplo: Para dos grupos

HOMBRES		MUJERES		
Valor	Rango	Valor	Rango	
1	1	4	3	
2	2	5	4	
18	6	6	5	
$\bar{x}_1 = 7$		$\bar{x}_2 = 5$		$\bar{x}_1 > \bar{x}_2$
Suma de rangos = $9 \Rightarrow \bar{R}_1 = 3$		Suma de rangos = $12 \Rightarrow \bar{R}_2 = 4$		$\bar{R}_1 < \bar{R}_2$

El test de *Wilcoxon-ManN-Whitney* contrastas si las diferencias observadas en las sumas de rangos son debidas al azar. Si p = 0.008 quiere decir que si decidimos que el grupo de HOMBRES es diferente a MUJERES la probabilidad de estar equivocados es p = 0.008, es decir hay un 0.8% de posibilidades de estar equivocado, que es menor que un 1%, luego la confianza en nuestra decisión es de un 99.2% > 99%.

```
PROC NPAR1WAY WILCOXON;
 CLASS SOBREPESO ;
 VAR DINAEST;
RUN ;
```

Para más de dos grupos el test de rangos es KRUSKAL-WALLIS:

Wilcoxon Scores (Rank Sums) for Variable DINAEST
Classified by Variable SOBREPES

SOBREPES	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
S<=-5	56	7190.0000	9856.0	695.656889	128.392857
S-5A+5	149	25137.5000	26224.0	938.985504	168.708054
A> 15	47	10413.0000	8272.0	646.957556	221.553191
S5A15	99	19035.5000	17424.0	854.885471	192.277778

Average Scores Were Used for Ties

Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ = 25.151 DF = 3 Prob > CHISQ = 0.0001

1. REGRESION LINEAL SIMPLE

Método estadístico para modelizar las relaciones entre variables continuas. Referimos la respuesta de una variable **dependiente** a los valores de las variables **independientes** o **explicativas**.

Los datos surgen de dos formas:

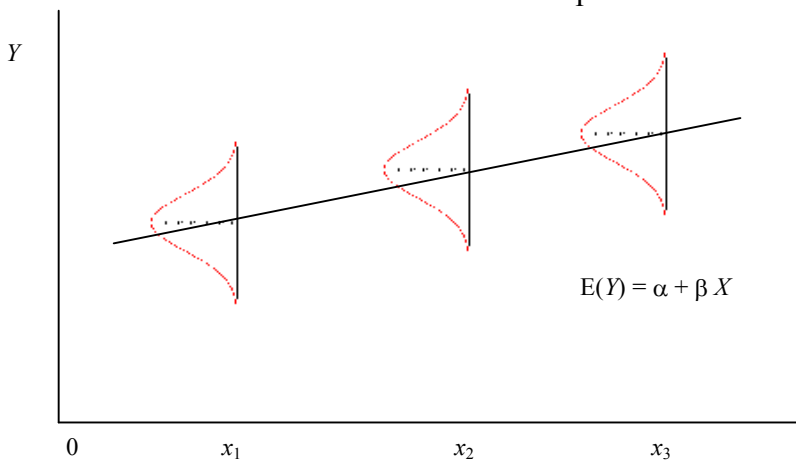
1. *X* fija : Años→ventas, temperatura→concentración.
2. *X* variable: 2 variables aleatorias medidas en la población

Regresión y correlación se usan para dos propósitos:

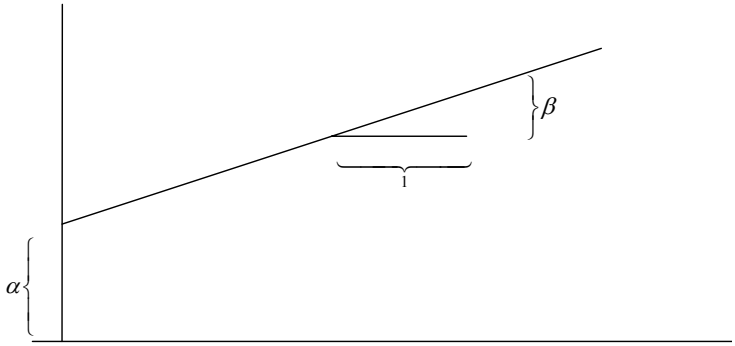
1. **Descriptiva**: Tipo de relación. Ecuaciones, representación gráfica, contrates de hipótesis, intervalos de confianza.
2. **Predicción**: Predecir *Y* dado una valor de *X*.

MÉTODOS DE REGRESION

Hipótesis: Fijado *X* la media de *Y* se sitúa sobre una misma recta y la varianza de Y es la misma para todos los valores de *X*: σ^2.



α término independiente, valor de *Y* cuando *X* = 0.
β pendiente, cantidad de cambio de *Y* por incremento unitario de *X*.

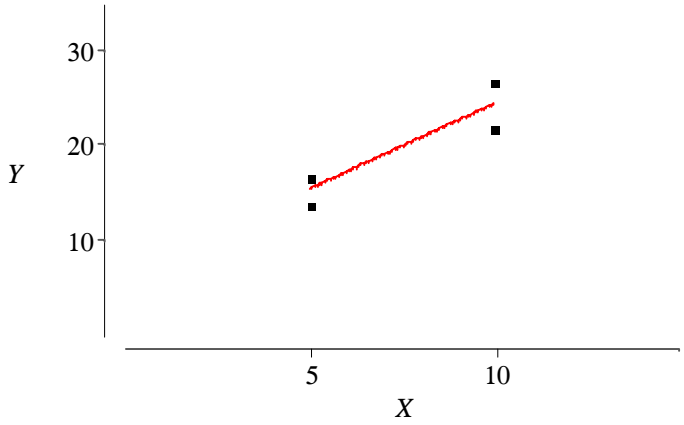


MINIMOS CUADRADOS

Estimamos α y β por el método de mínimos cuadrados, obteniendo A y B.

Ejemplo: $x_1 = 5,\ x_2 = 5,\ x_3 = 10,\ x_4 = 10$

$y_1 = 14,\ y_2 = 17,\ y_3 = 27,\ y_4 = 22$



	Media	Desv. St.	Regresión	Res. Med. Cuad.
X	7.5	2.88		
Y	20.0	5.71	Y=6.5 + 1.8 X	8.5

$$\underset{A,B}{\text{Min}} \sum_{i=1}^{4} e_i^2 = \sum_{i=1}^{4} [y_i - (A + Bx_i)]^2$$

$$\left.\begin{array}{l} B = \dfrac{\sum_i (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sum (x_i - \bar{x})^2} \\ \\ A = \bar{y} - B\bar{x} \end{array}\right\} \hat{Y} = A + BX$$

Residuales $e_1 = 14 - (6.5 + 1.8 \times 5) = -1.5$

Media de cuadrados residual

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} \quad \text{es una estimación de } \sigma^2.$$

En el ejemplo, $S^2 = \dfrac{17}{4-2} = 8.5$

Errores estándar de A y B.

Permiten calcular intervalos de confianza para los coeficientes:

$$IC_B = B \pm t_{n-2,\alpha/2} SE(B)$$
$$IC_A = A \pm t_{n-2,\alpha/2} SE(A)$$

Intervalos de Predicción.

Estimamos la media de Y en un valor fijo de $X = x_0$.

$$\hat{Y} = A + B\,x_0$$

Intervalo de confianza: Banda de confianza. Más estrecha cerca de la media $\bar{x}$.

Test de hipótesis.

$$H_0 \equiv \beta = \beta_0 \quad \text{se contrasta con} \quad t = \frac{B - \beta_0}{SE(B)} \approx t_{n-2}$$

Caso particular $\beta_0 = 0$ (Y no cambia, cuando la variable X cambia. La predicción de Y no mejora al conocer la variable X).

Significatividad de la regresión. Tabla ANOVA.

La tabla ANOVA contrasta la hipótesis $H_0 \equiv \beta = 0$ y proporciona una Medida de la calidad del ajuste.

Fuente	Suma de cuadrados	g.l	Media cuadrados	F
Regresión	$\sum(\hat{y}_i - \bar{y})^2$	1	SC/1	MC_{reg}/MC_{res}
Residual	$\sum(y_i - \hat{y}_i)^2$	$n-2$	SC/$n-2$	
Total	$\sum(y_i - \bar{y})^2$	$n-1$		

$R^2 = \frac{\text{SC reg}}{\text{SC tot}} \in [0,1]$ medida de calidad del ajuste. Proporción de las variaciones de la variable Y que pueden ser atribuidas a X.

Test F para medir significatividad $H_0 \equiv \beta = 0$.

Análisis de residuales

Detectar outliers, Influencia de observaciones individuales,
Normalidad, Independencia.

2. REGRESION LINEAL MÚLTIPLE

Relación entre una variable respuesta Y y un conjunto de variables X_i independientes.

El modelo general (con p+1 parámetros) se escribe

$$Y_i = a + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \qquad i = 1,\ldots,n$$

con ε_i variables aleatorias normales independientes $N(0,\sigma)$, σ desconocido.

β_1 es el ritmo de cambio de la variable Y por unidad de cambio de X_1 cuando el resto de variables permanece constante.

ESTIMACION POR MINIMOS CUADRADOS

Si $\hat{Y} = A + B_1 X_1 + B_2 X_2 + \ldots + B_p X_p$ se minimiza la suma de cuadrados de los errores

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

***Ejemplo*:** Sistema de distribución. Tres variables, TIEMPO empleado en el reparto, número de PUNTOS que tiene que recorrer y DISTANCIA máxima.

OBSERVAC.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PUNTOS	10	15	10	20	25	18	12	14	16	22	24	17	13	30	24
DISTANCIA	30	25	40	18	22	31	26	34	29	37	20	25	27	23	33
TIEMPO	24	27	29	31	35	33	26	28	31	39	33	30	25	42	40

```
Model:Dependent Variable: TIEMPO
                          Analysis of Variance

                                 Sum of          Mean
Source                 DF       Squares        Square      F Value      Prob>F

Model                   2     331.35860     165.67930       16.795      0.0003
Error                  12     118.37473       9.86456
C Total                14     449.73333

    Root MSE          3.14079    R-square        0.7368
    Dep Mean         30.86667    Adj R-sq        0.6929

                          Parameter Estimates

                   Parameter       Standard    T for H0:                 Standardized
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|       Estimate

INTERCEP   1        2.311202     5.85730272         0.395        0.7001     0.00000000
PUNTOS     1        0.877205     0.15303460         5.732        0.0001     0.92862129
DISTANC    1        0.455921     0.14676233         3.107        0.0091     0.50327125
```

La ecuación obtenida para el ejemplo es

$$TIEMPO = 2.311 + .877\ PUNTOS + .456\ DISTANCIA$$

Si queremos predecir la variable tiempo para la observación 1 obtenemos

$$TIEMPO = 2.311 + .877\ (10) + .456\ (30) = 24.76$$

Pero también podemos predecir para otras combinaciones de valores no presentes en la muestra.

Obs	Dep Var TIEMPO	Predict Value	Std Err Predict	Lower95% Mean	Upper95% Mean	Lower95% Predict	Upper95% Predict	Residual	Std Err Residual
1	24.0000	24.7609	1.397	21.7173	27.8045	17.2714	32.2504	-0.7609	2.813
2	27.0000	26.8673	1.107	24.4547	29.2799	19.6113	34.1233	0.1327	2.939
3	29.0000	29.3201	1.874	25.2379	33.4023	21.3518	37.2884	-0.3201	2.521
4	31.0000	28.0619	1.594	24.5886	31.5351	20.3877	35.7360	2.9381	2.706
..	..								
13	25.0000	26.0247	1.164	23.4878	28.5616	18.7264	33.3230	-1.0247	2.917
14	42.0000	39.1135	1.864	35.0512	43.1758	31.1554	47.0716	2.8865	2.528
15	40.0000	38.4095	1.608	34.9061	41.9128	30.7217	46.0973	1.5905	2.698
16	.	33.5329	0.954	31.4541	35.6117	26.3809	40.6849	.	.

Obs	Student Residual	-2-1-0 1 2	Cook's D
1	-0.270	| | |	0.006
2	0.045	| | |	0.000
3	-0.127	| | |	0.003
4	1.086	| |** |	0.136
5	-3.268	|****** |	0.803
6	0.255	| | |	0.002
7	0.461	| | |	0.016
8	-0.718	| *| |	0.028
9	0.474	| | |	0.006
10	0.206	| | |	0.008
11	0.185	| | |	0.003
12	0.461	| | |	0.007
13	-0.351	| | |	0.007
14	1.142	| |** |	0.237
15	0.590	| |* |	0.041
16	.		.

Sum of Squared Residuals 118.3747

Es importante resaltar que si obtenemos otra muestra de 15 observaciones es normal que la ecuación cambie y por lo tanto las predicciones. Entonces es importante conocer para nuestra muestra lo siguiente:

1. La cantidad de posible error presente en las predicciones.
2. La significatividad de las estimaciones de los parámetros.
3. Cuánto explica la ecuación de nuestros datos.

INTERPRETACION DE RESULTADOS

Intervalo de confianza para una predicción

Significatividad de los parámetros

Contribución individual de una variable: Podemos construir test de significatividad estadística de los parámetros del modelo:

$$t = \frac{(\text{estimacion}) - (\text{valor hipotético})}{(\text{Error std de la estimacion})} \approx t_{n-p-1}$$

El *valor-t* en la salida corresponde a la hipótesis "parámetro = 0". Esto equivale a suponer que la variable *X* correspondiente no influye en la *Y*.

La decisión de rechazar esta hipótesis tiene asociado un error o nivel de significatividad, determinado por la correspondiente variable *t* de Student.

Para X_1≡PUNTOS, $H_0 \equiv \beta_1 = 0$ $\qquad t = \frac{B_1 - 0}{SE(B_1)} = \frac{0.877}{0.153} = 5.732 \approx t_{15-3}$

Significatividad conjunta de la ecuación:

$$H_0 \equiv \beta_1 = \beta_2 = ... = \beta_p = 0$$

Fuente	Suma de cuadrados	g.l	Media cuadrados	F
Regresión	$\sum(\hat{y}_i - \bar{y})^2$	p	SS_{reg}/p	MC_{reg}/MC_{res}
Residual	$\sum(y_i - \hat{y}_i)^2$	$n-p-1$	$SS_{res}/n-p-1$	
Total	$\sum(y_i - \bar{y})^2$	$n-1$		

En el ejemplo F=16.795 y es significativo con p =0.0003 < 0.001

Variación explicada

El coeficiente de correlación entre y_i e $\hat{y}_i$ es el coeficiente de correlación múltiple: medida de asociación entre la variable *Y* y un conjunto $\{X_i\}_{i=1}^{p}$.

Coeficiente de Determinación $= \frac{SS_{reg}}{SS_{res}} =$ (coeficiente de correlación múltiple)^2 =

Porcentaje de información acerca de *Y* suministrado por la ecuación de regresión.

Es una medida de cómo el modelo explica la variable respuesta. Es el cociente

$$R^2 = \frac{\text{Suma de cuadrados debido a la regresión}}{\text{Suma de cuadrados total, ajustada por la media}} = \frac{331.359}{449.733} = .7368 \approx 73.7\%$$

VALIDACIÓN DEL MODELO

Para que la inferencia y predicción realizada en el apartado anterior sean válidas es necesario comprobar si se producen desviaciones de las hipótesis del modelo. Esta comprobación se basa en el estudio de residuales:

Gráficos de residuales:

Estadísticos de residuales:

SELECCIÓN DE VARIABLES EN REGRESION

Situaciones

a) Mucha variables independientes
b) Interés en determinar un orden de importancia
c) Obtención de una ecuación de predicción de bajo error lo más sencilla posible.

Interesa determinar la "mejor" ecuación de regresión.

Criterios de selección del mejor subconjunto:

- Minimizar Suma de Cuadrados Residual: $SC_R = \sum (y_i - \bar{y})(1 - R^2)$ equivalente a maximizar coeficiente de determinación.
- Como R^2 aumenta al añadir nuevas variables, una alternativa es el R^2 ajustado

$$R_a^2 = R^2 - \frac{p(1-R^2)}{n-p-1}.$$

- C_p criterio. Se recomienda seleccionar el conjunto de variables que minimizan C_p.

SELECCIÓN POR PASOS:

SELECCIÓN DE LOS MEJORES MODELOS:

1. ANALISIS DISCRIMINATE. CONCEPTOS BÁSICOS

OBJETIVO

Clasificar a un individuo respecto a dos o más poblaciones en función de los valores de una o más variables. Las poblaciones son grupos alternativos, conocidos y distintos.

Cada individuo pertenece a uno de ellos. Las técnicas identifican las variables que contribuyen a la clasificación, por lo que se tiene un doble objetivo: descripción y predicción.

EJEMPLOS

- Identificación de pacientes que responden a un tratamiento en función de características clínicas de la persona y carácterísticas de la enfermedad.

- Prediccion del sexo en función de las medidas de huesos.

Inicialmente suponemos que las variables observadas para realizar la discriminación son **cuantitativas**, por lo que tenemos un estudio similar al de la regresión lineal múltiple pero con la variable ***Y* cualitativa**.

Variables predictoras explicativas independientes		Variable a predecir a explicar dependiente
$X_1, X_2, ..., X_p$ cuantitativas	$\longrightarrow$ Y	Cuantitativa: REGRESION
		Cualitativa: AN. DISCRIMINANTE

El primer estudio que haríamos sería la relación de cada una de las variables X_i con la variable dependiente Y construyendo una tabla del tipo:

	I: HOMBRE		**II: MUJER**	
	Media	**Desviación**	**Media**	**Desviación**
Var 1	**30**	6.3	**50**	10.1
Var 2	**0.3**	0.15	**0.7**	0.12
.....	.....	.....	.....	.....

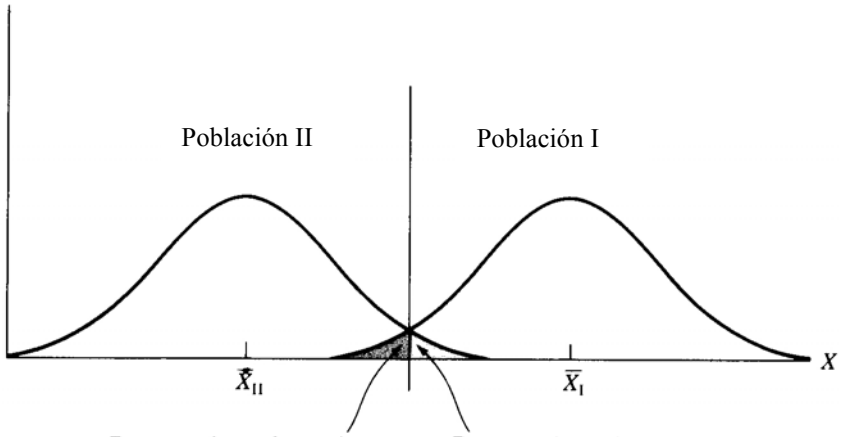
En esta tabla puede identificarse qué características tienen medias diferentes en un grupo y otro. (Test de comparación de dos medias)

Sin embargo, como en el análisis de regresión lineal, el uso de varias variables simultáneamente mejorará los resultados en cuanto a descripción y predicción respecto al uso de una única variable.

2. FUNCION DISCRIMINANTE PARA DOS GRUPOS

En el análisis discriminante clásico de Fisher se obtienen unas **funciones de clasificación** que permiten asignar los individuos a un grupo u otro y evaluar la calidad de los resultados.

Suponemos que un individuo puede pertenecer a una de dos poblaciones I y II y debemos clasificarle en una de estas dos poblaciones en función de una característica *X*. suponemos que tenemos una muestra representativa de cada población, lo que nos permite estimar las medias y distribuciones de *X* en cada población, como se representa en la figura



De la figura se deduce que individuos con valores bajos de *X* deben ser clasificados como II y con valores altos como I. es necesario dar un punto de corte, que usualmente será $C = \dfrac{\bar{X}_I + \bar{X}_{II}}{2}$ y asignar a $\begin{cases} \text{I} & \text{cuando } X \geq C \\ \text{II} & \text{cuando } X < C \end{cases}$

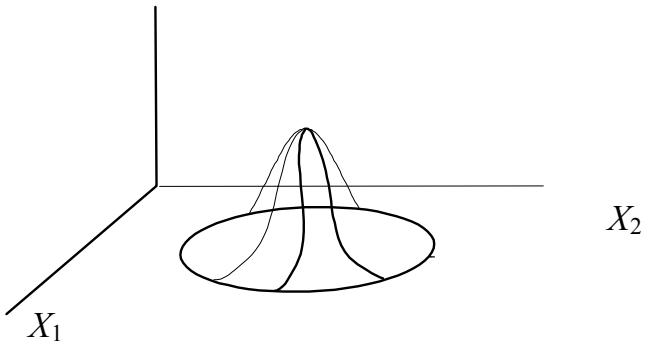
Si las dos distribuciones tienen diferentes desviaciones σ_I, σ_{II} la regla de asignación para el caso de una variable consiste en asignar al mínimo de

$$\left(\frac{X-\bar{X}_I}{\sigma_I}\right)^2, \left(\frac{X-\bar{X}_{II}}{\sigma_{II}}\right)^2$$

En la práctica, una sola variable no proporciona resultados satisfactorios. Se mejoran las clasificaciones combinando dos o más variables.

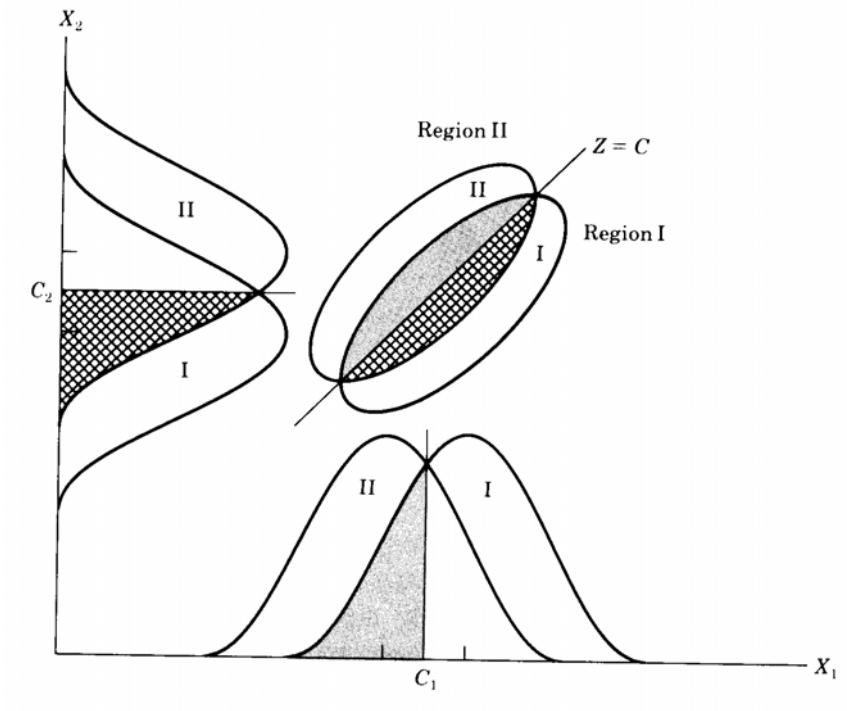
DOS VARIABLES

Es el caso de dos variables (X_1, X_2) con distribuciones normales, la función de densidad se representa por un elipsoide (en cada punto del plano (X_1, X_2) una altura, indicando la concentración de valores alrededor de ese punto). Al proyectar con un nivel del 95% obtenemos elipses que son con las que trabajamos.



Cuanto más separados estén los centroides mayor seguridad se obtiene de una buena asignación a I o II.

La figura representa la clasificación en dos grupos, con dos variables:



La línea recta que pasa por las intersecciones divide el plano en 2 regiones, cada región corresponde a una población. Un nuevo individuo se asigna a I ó II dependiendo de la región donde se posicione.

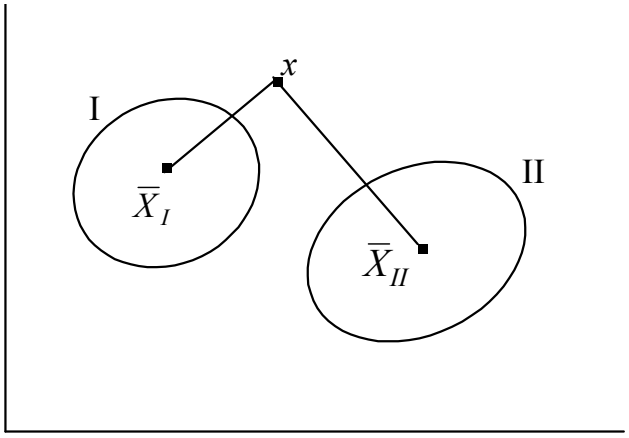
Las zonas sombreadas muestran porcentajes de clasificaciones incorrectas.

3. FUNCIONES DE CLASIFICACION

Los programas informáticos calculan "**Funciones de clasificación**".

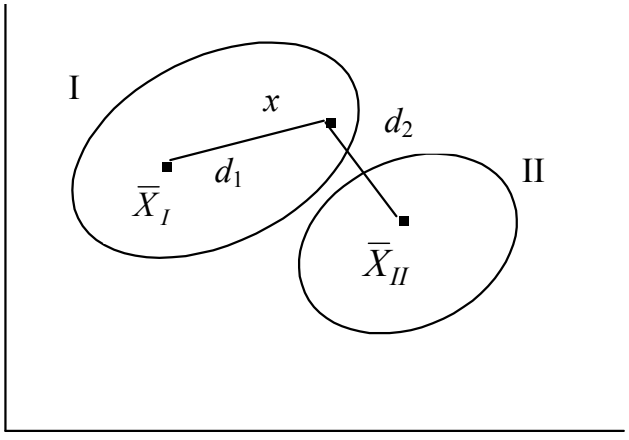
El grupo *i* tiene asociado una función de clasificación $g_i(X_1, X_2)$.

El cálculo de las funciones de clasificación se basa en la **distancia de Mahalanobis** que generaliza la distancia euclídea a raíz de la observación siguiente:



Se tiene el problema de clasificar el punto *x*. Entonces considerando una distancia asigno el punto *x* al grupo tal que la distancia de *x* al centroide sea mínima.

Puede surgir el problema siguiente:



La distancia d_2 es menor que d_1 pero *x* está en el elipsoide de concentración (al 95%) de $\overline{X}_I$.

Es decir, es más probable que sea del grupo I.

Para salvar esta anomalía hay que tener en cuenta el comportamiento de las **distribuciones**. Entonces el proceso de clasificación consiste en asignar un punto *x* al grupo al que la distancia de Mahalanobis sea menor.

Al transformar las distancias en similaridades se obtienen las funciones de clasificación $g_i(X_1, X_2) = a_0^i + a_1^i X_1 + a_2^i X_2$ para *i* = 1, 2, ..., *K* grupos.

"**La asignación consiste en evaluar las *K* funciones de clasificación y asignar al grupo que presente el máximo**".

Ejemplo:

```
       GROUP =   VARON        MUJER        ALL GPS.
VARIABLE
 34 longi       230.72084    203.30612    215.62733
 37 minmdd       11.40417      9.18707     10.18352

COUNTS               40.          49.          89.
```

CLASSIFICATION FUNCTIONS **una variable**

```
        GROUP =   VARON         MUJER
VARIABLE
 37 minmdd       17.48269     14.08387

CONSTANT       -100.38092    -65.38793
```

CLASSIFICATION MATRIX

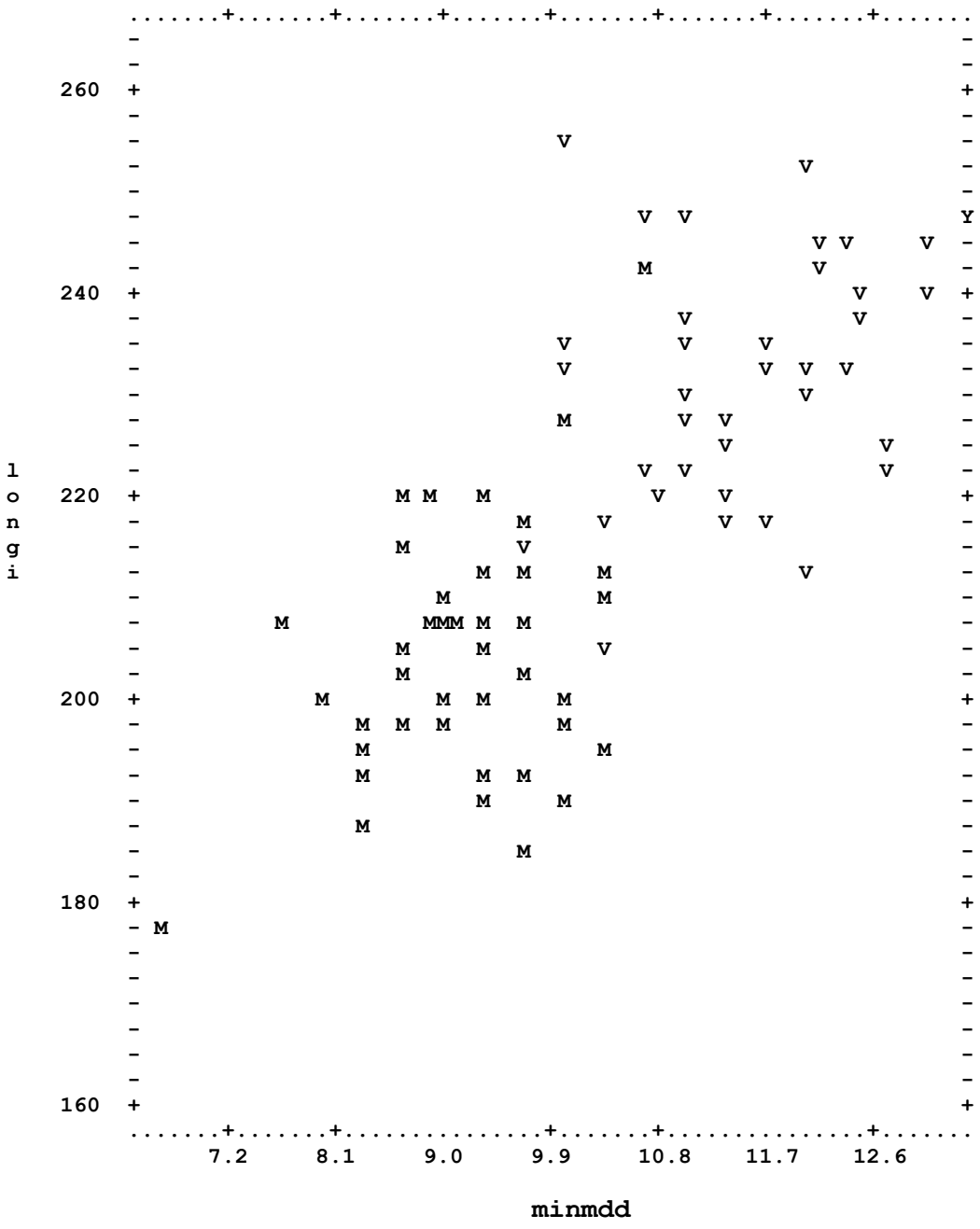
```
GROUP      PERCENT   NUMBER OF CASES CLASSIFIED INTO GROUP -
           CORRECT
                     VARON    MUJER
 VARON      90.0     36        4
 MUJER      91.8      4       45

TOTAL       91.0     40       49
```

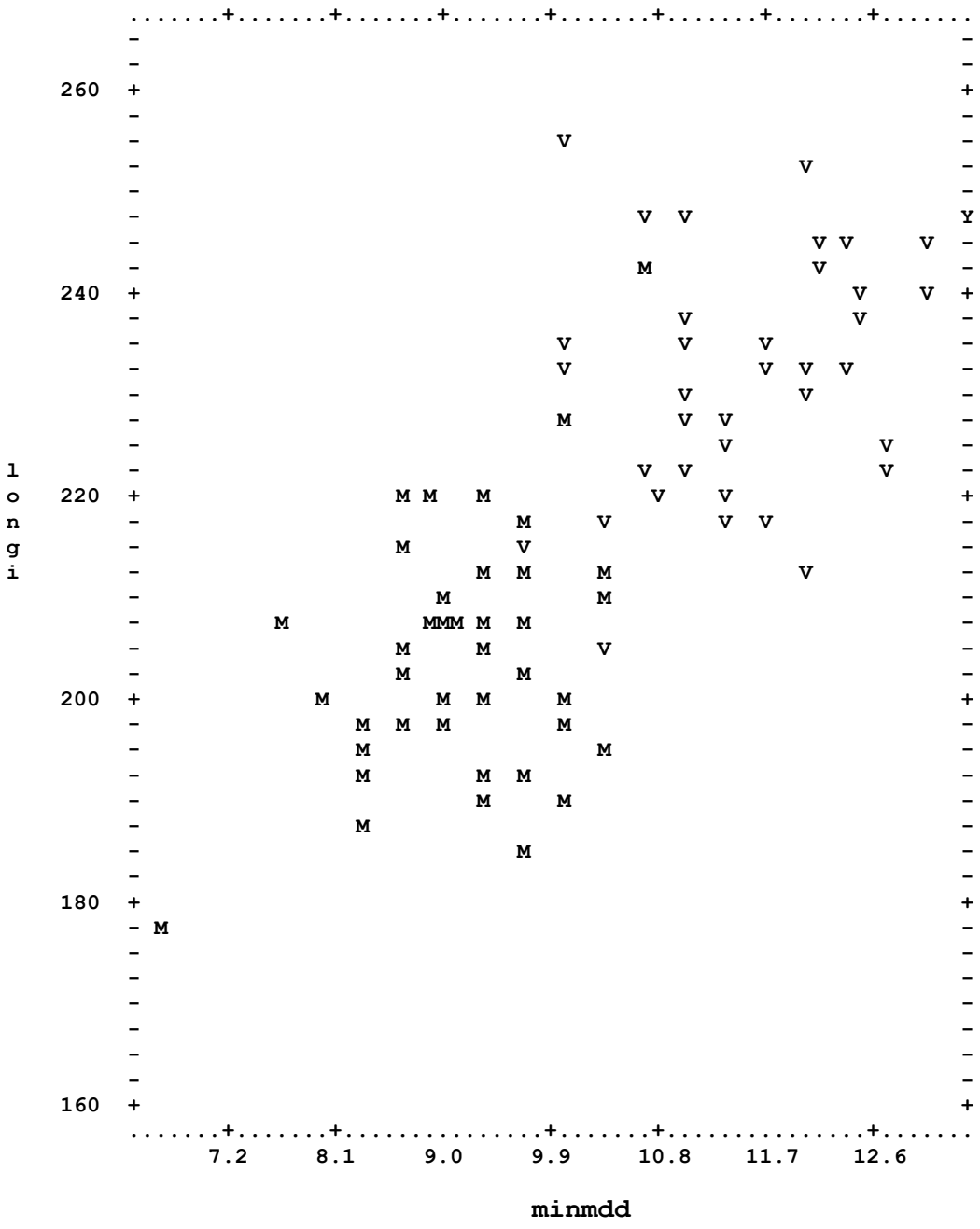
CLASSIFICATION FUNCTIONS **dos variables**

```
        GROUP =   VARON         MUJER
VARIABLE
 34 longi         1.47537      1.33043
 37 minmdd       11.09498      8.32370

CONSTANT       -234.15735   -174.17070
```

CLASSIFICATION MATRIX

```
GROUP      PERCENT   NUMBER OF CASES CLASSIFIED INTO GROUP -
           CORRECT
                     VARON    MUJER
 VARON      95.0     38        2
 MUJER      95.9      2       47

TOTAL       95.5     40       49
```


260
240
220
200
180
longi
7.2
8.1
9.0
9.9
10.8
11.7
12.6
minmdd


260
240
220
200
180
longi
7.2
8.1
9.0
9.9
10.8
11.7
12.6
minmdd

VALIDACIÓN CRUZADA:

En este procedimiento se obtienen las funciones de clasificación con una muestra y se valida con otra muestra diferente. Los errores obtenidos son más reales.

El investigador puede dividir su muestra aleatoriamente en dos submuestras, no necesariamente del mismo tamaño.

Muestra base, diseño: usada para obtener las funciones.

Muestra test, de validación: usada para validar.

Cuando la muestra total es pequeña puede usarse el método llamado "***Jacknife***". En este método se excluye un caso y se calculan las funciones de clasificación con el resto. Entonces se clasifica la observación excluida. Este procedimiento se aplica para cada una de las observaciones.

Cuando se incorporan sucesivamente nuevas variables para obtener la funciones de clasificación el error de clasificación obtenido con toda la muestra disminuye pero el error obtenido por validación cruzada o Jacknife puede aumentar a partir de un número elevado de variables.