BIOESTADISTICA

Gabriel Cavada

Un poco de historia

- ¿Qué es Estadística?, etimológicamente el vocablo deriva de Estado y significa "contar los bienes del Estado“

- Los albores de esta disciplina se encuentran en la Antigüedad

  Las autoridades del Egipto faraónico contaban sus bienes y registraban la profundidad del río Nilo en cada estación del año

  Jesucristo nace en Belén, porque un edicto del emperador romano ordena un censo, para conocer el número y características de los habitantes del Imperio

En nuestros días ¿Qué es estadística?

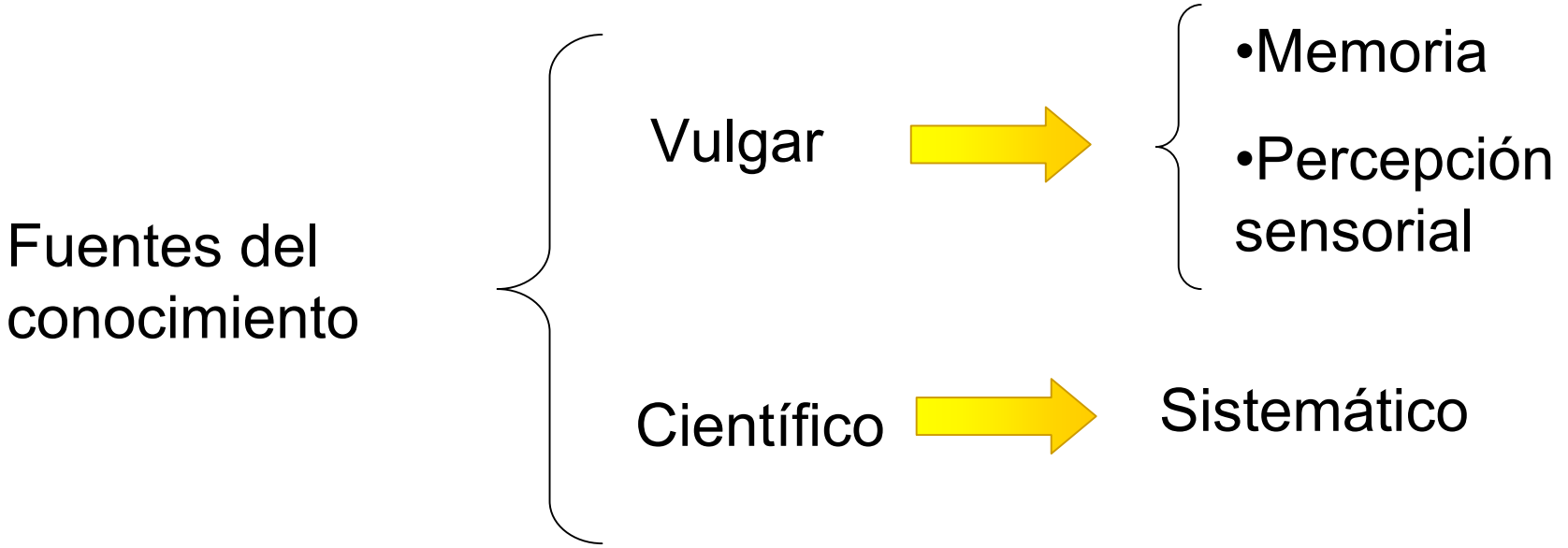
Estadística es la disciplina que se ocupa de:

"La recolección, organización y procesamiento de datos, para obtener inferencias hacia un universo cuando se observa sólo una parte de este universo"

"Identificar la variabilidad de un fenómeno y tratar de explicarla"

"Tomar desiciones bajo incertidumbre"

¿Cómo conocemos?



Método Científico

- Se sistematiza en cinco puntos:

1. Detección y Enunciado del Problema: Es la descripción de una situación problema o es el planteamiento de una pregunta.

2. Formulación de la Hipótesis: Es una respuesta o explicación al problema enunciado, que se hace en base al conocimiento científico existente.

3. Deducción de una consecuencia verificable: Como la hipótesis es una explicación general, a menudo ocurre que no se puede investigar directamente, luego se procede a deducir, lógicamente, consecuencias particulares de la hipótesis.

Método Científico

4. Verificación de la consecuencia: En ciencias exactas esto se realiza usando lógica pura, sin embargo en ciencias no exactas la verificación se hace a través de la recolección de información o la observación de los fenómenos, lo que hace necesario la aplicación de Procedimientos Estadísticos.

5. Conclusión: Consiste en la aceptación, modificación o total rechazo de la hipótesis planteada.

Método Estadístico

Método Estadístico es el que proporciona las técnicas necesarias para:

Recolectar y Analizar la información requerida.

El Método Estadístico distingue dos etapas: la Planificación y la Ejecución.

Método estadístico: Planificación

- Definición de objetivos: es la descripción formal del problema que da origen a la investigación. Se debe señalar detalladamente lo que se va a investigar, el qué, cómo, dónde, cuándo y por qué.

- Universo del estudio: es la definición del conjunto desde el cual se extraerá la información y hacia el que se generalizarán las conclusiones obtenidas.

- Diseño de la muestra: la Teoría de Muestreo garantiza que la información que generaremos nos permita proyectar válidamente las conclusiones al Universo de interés.

Método estadístico: Planificación

- Definición de las unidades de observación (que objetos observaremos), las escalas de clasificación y las unidades de medida.

- Preparación del Plan de Tabulación y Análisis de la información: aquí se determinan las formas de presentar y analizar la información recolectada

Método estadístico: Ejecución

En la fase de Ejecución se reconocer los siguientes aspectos:

- Recolección de la información
- Elaboración de la información
- Análisis de los resultados

Unidad de análisis y atributos

- **Unidad de análisis**: Una vez definido el problema que se va a investigar, se definen naturalmente los objetos que serán observados:
  - Seres humanos
  - Animales
  - Células
  - Órganos
  - Etcétera

Unidades de análisis

Unidad de análisis y atributos

- **Atributos:** Teniendo definidas las unidades de análisis, obviamente ellas presentan características que nos importan para nuestro estudio: Si nuestro estudio es antropométrico, podemos consignar algunas características esenciales tales como:
  - Sexo
  - Estatura
  - Raza
  - Peso

Variables

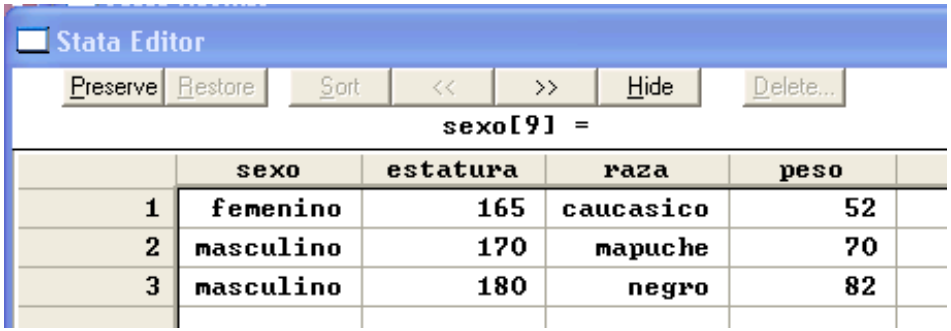
Variables: Cuando se han definido los atributos a estudiar, podemos ya observar unidades de análisis especificadas y los atributos quedan consignados como características únicas del objeto que estamos estudiando. Si observamos una persona en particular podemos consignar:

Sexo: Femenino

Estatura: 165 centímetros

Raza: Caucásico

Peso: 52 kilógramos



	sexo	estatura	raza	peso
1	femenino	165	caucasico	52
2	masculino	170	mapuche	70
3	masculino	180	negro	82

Cuando los atributos ya han sido evaluados, reciben el nombre de Variables del estudio.

Escalas de medida

Cuando procedemos a medir las variables del estudio, debemos tener presente que estamos consignando valores con unidades de medida y por consiguiente introduciendo escalas de medición. Estas escalas pueden ser: Nominales, Ordinales o Intervalares (o de Razón). Estas escalas tienen diferente Poder de Clasificación

Escalas de medida

Escala de medida	Capacidad
Nominal	Sólo es capaz de nombrar o etiquetar la unidad de análisis. Por ejemplo: Sexo, raza, nacionalidad
Ordinal	Es capaz de nombrar pero además introduce una jerarquía en las unidades observadas. Por ejemplo: Grado que se cursa en el sistema escolar básico, nivel económico, escala analógica para el dolor
Intervalar y de razón	Es capaz de nombrar, jerarquizar pero además permite hacer comparaciones matemáticas entre las unidades de análisis. Por ejemplo: Temperatura en grados Celcius (intervalar). Peso, estatura (de razón). **Las escalas de razón el cero indica ausencia de la variable.**

Escalas de medida

- **Las escalas de medida se pueden bajar pero nunca subir.** Es decir una variable en escala intervalar se puede dejar en escala ordinal y una en escala ordinal se puede dejar en escala nominal, pero una variable en escala nominal no se puede dejar en escala ordinal y una en escala ordinal no se puede dejar en escala intervalar

Escalas de medida

Las variables medidas en escala intervalar pueden ser:

Discretas: Asociadas a los números naturales, es decir sólo cuentan, por ejemplo: Número de hijos, células por campo

Continuas: Asociadas a los números reales, es decir miden, por ejemplo: Peso, temperatura, edad

Escalas de medida

- Una variable continua se puede discretizar, pero una variable discreta no se puede continuizar.

  – Por ejemplo: la edad es una medida de tiempo y de naturaleza continua, sin embargo se registra en años cumplidos que es de naturaleza discreta. Resulta poco cómodo registrar la edad de alguien como: 30.2130 años (30 años con 2 meses, 16 días, 16 horas y 19 minutos) es mejor contar la cantidad de velas que apagó en la torta en su último cumpleaños, 30 velas = 30 años

Escalas de medida

- La precisión con que se mide una variable va de acuerdo al interés de la investigación, como se estableció en el Método Estadístico.

Población y muestra

Población: Llamamos Población al Conjunto Universo de las unidades de análisis, la población puede ser de tamaño finito o infinito:

Si se desea averiguar el volumen de la cavidad craneana en humanos adultos, la población en estudio son todos los humanos vivos en este momento, esta población en la práctica es infinita.

Si se desea saber la edad de los sujetos VIH+ en Chile actualmente, la población es finita.

Población y muestra

MUESTRA:

Es un **SUBCONJUNTO FINITO y FACTIBLE** de la Población, que debe cumplir características ineludibles para lograr que las conclusiones estadísticas sean válidas.

Población y muestra

LAS CARACTERÍSTICAS DE UNA "BUENA MUESTRA" SON:

Aleatoria: garantiza que los elementos que componen la muestra fueron escogidos completamente al azar, es decir no hay predilección alguna por incluir o excluir determinada unidad de análisis (todos los sujetos de una población tienen la misma probabilidad de integrar la muestra)

El tamaño de la muestra, que es el número de unidades de análisis que se deben escoger, debe ser lo suficientemente grande como para garantizar la generalización de los resultados a la población.

La determinación del tamaño de una muestra no es un problema trivial y constituye una especialización de la estadística llamada Teoría del Muestreo.

Estadística Descriptiva

- Se llama estadística descriptiva, al conjunto de técnicas que permiten ordenar, resumir y representar la información recolectada.
- Esta sólo pretende hacer una descripción cuantitativa del fenómeno sin proyectar, aún, sus resultados a la universalidad del fenómeno.

Ordenación y representación de datos

- Obtenida la información que se desea analizar es necesario ordenarla, para ello utilizaremos técnicas que dependen de la naturaleza de la variable y su escala de medida

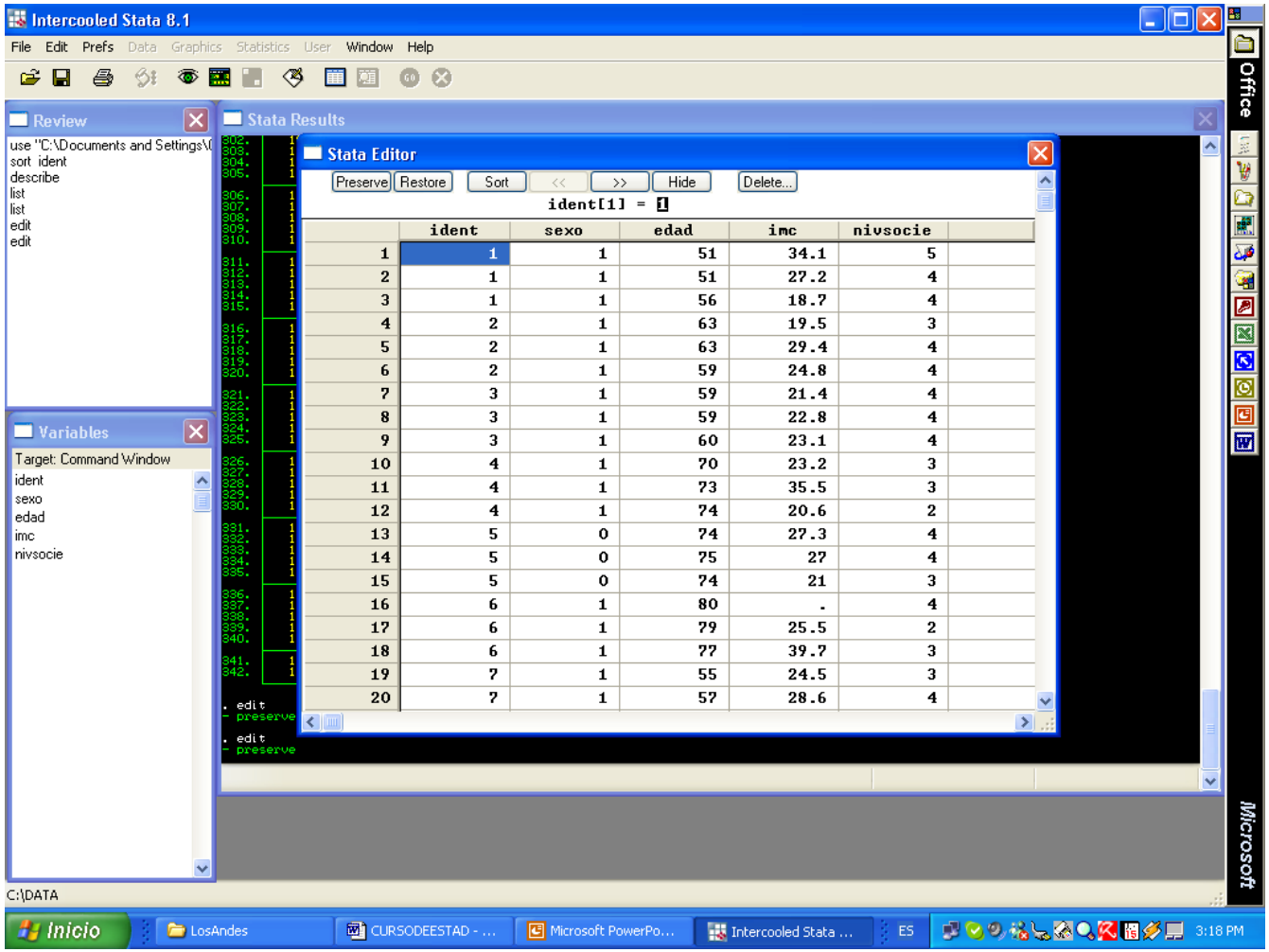
Ordenación y representación de datos

- Para desarrollar este capítulo nos referiremos a la base de datos AURI.dta, que contiene información de pacientes con cáncer vesicular confirmado por estudio histológico:

Ordenación y representación de datos

```
. describe

Contains data from F:\LosAndes\AURI.DTA
  obs:           342
 vars:             5                          16 Jul 2005 09:31
 size:         8,208 (99.2% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
ident           float  %9.0g
sexo            float  %9.0g                 0: hombre 1: mujer
edad            float  %9.0g
imc             float  %9.0g
nivsocie        float  %9.0g                 1:alto 2:medio alto 3:medio
                                                4:bajo 5:muy bajo
-------------------------------------------------------------------------------
Sorted by:
```

Ordenación y representación de datos



	ident	sexo	edad	imc	nivsocie
1	1	1	51	34.1	5
2	1	1	51	27.2	4
3	1	1	56	18.7	4
4	2	1	63	19.5	3
5	2	1	63	29.4	4
6	2	1	59	24.8	4
7	3	1	59	21.4	4
8	3	1	59	22.8	4
9	3	1	60	23.1	4
10	4	1	70	23.2	3
11	4	1	73	35.5	3
12	4	1	74	20.6	2
13	5	0	74	27.3	4
14	5	0	75	27	4
15	5	0	74	21	3
16	6	1	80	.	4
17	6	1	79	25.5	2
18	6	1	77	39.7	3
19	7	1	55	24.5	3
20	7	1	57	28.6	4

Ordenación y representación de datos

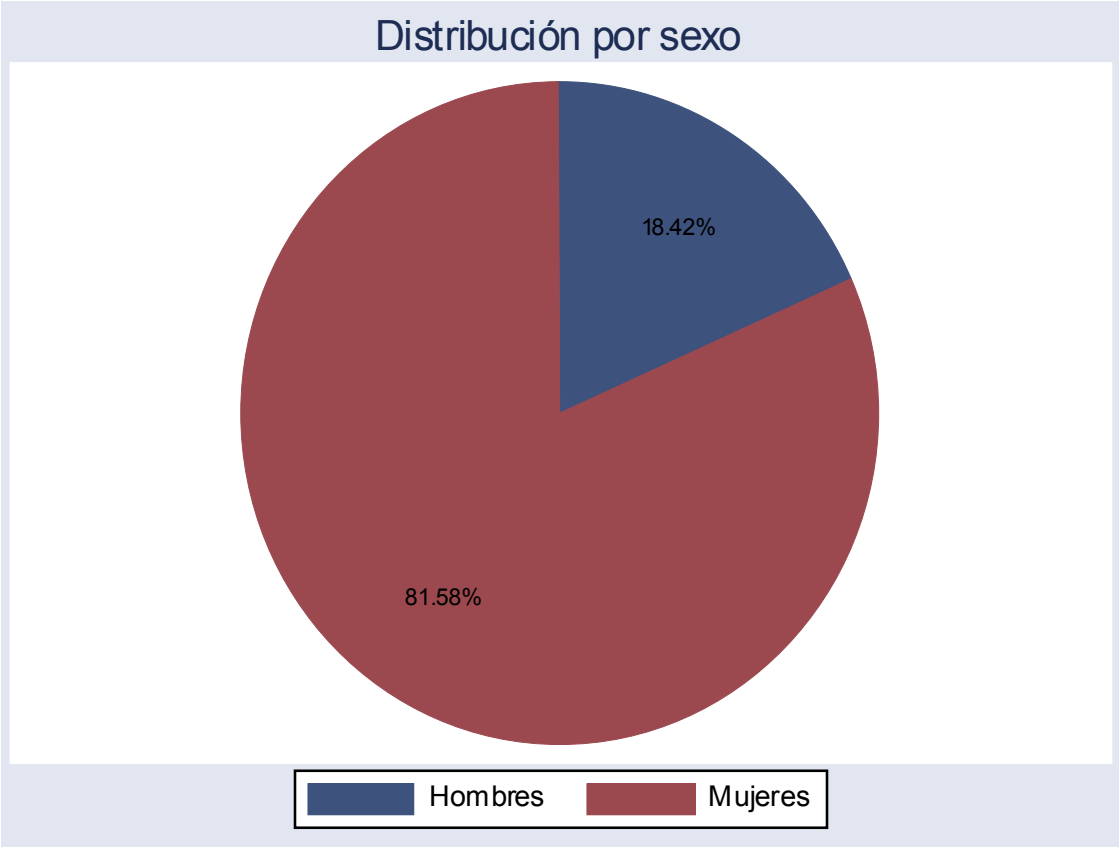
- Sexo : medida en escala nominal
- Edad : medida en escala de razón
- Imc : medida en escala de razón
- Nivsocie: medida en escala ordinal

Ordenación y representación de datos

- Representación de la variable Sexo:

```
. tab sexo

  0: hombre |
   1: mujer |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         63       18.42       18.42
          1 |        279       81.58      100.00
------------+-----------------------------------
      Total |        342      100.00
```

Ordenación y representación de datos

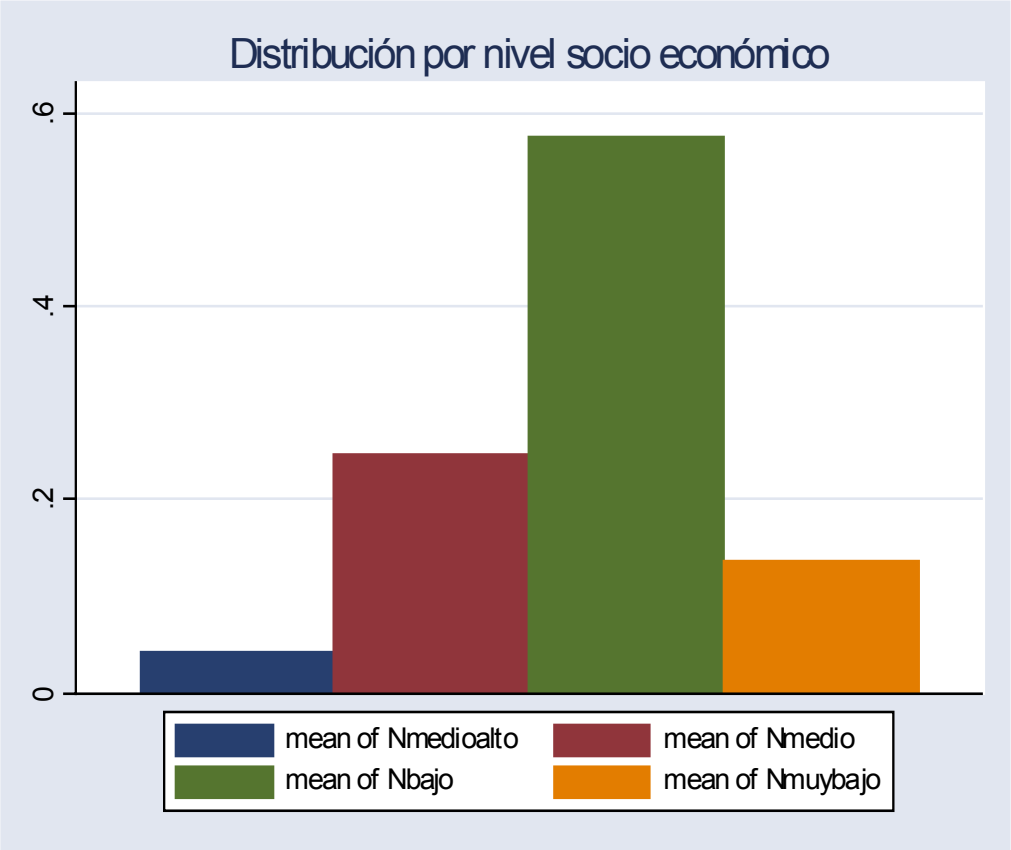


Ordenación y representación de datos

- Representación de la variable nivel socioeconómico:

```
. tab nivsocie, gen(Niv)

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```

Ordenación y representación de datos



Ordenación y representación de datos

- Variable Edad: Ordenación en tallo y hoja

```
. stem edad, line(2)

Stem-and-leaf plot for edad (años cumplidos)

  2* | 34
  2. |
  3* | 114
  3. | 556666688
  4* | 000011111122223333
  4. | 55555667777777788888999999999
  5* | 00000000000001111111111222222222223333334444
  5. | 55566666667777888888899999999
  6* | 000000000011111111222223333333333333333333344444444444444
  6. | 55555555555666666667777777778888888889999999999
  7* | 00000000000011111112222222233333333333334444444444444
  7. | 55555555555566666666777777777888889999
  8* | 000123344
  8. | 689
```

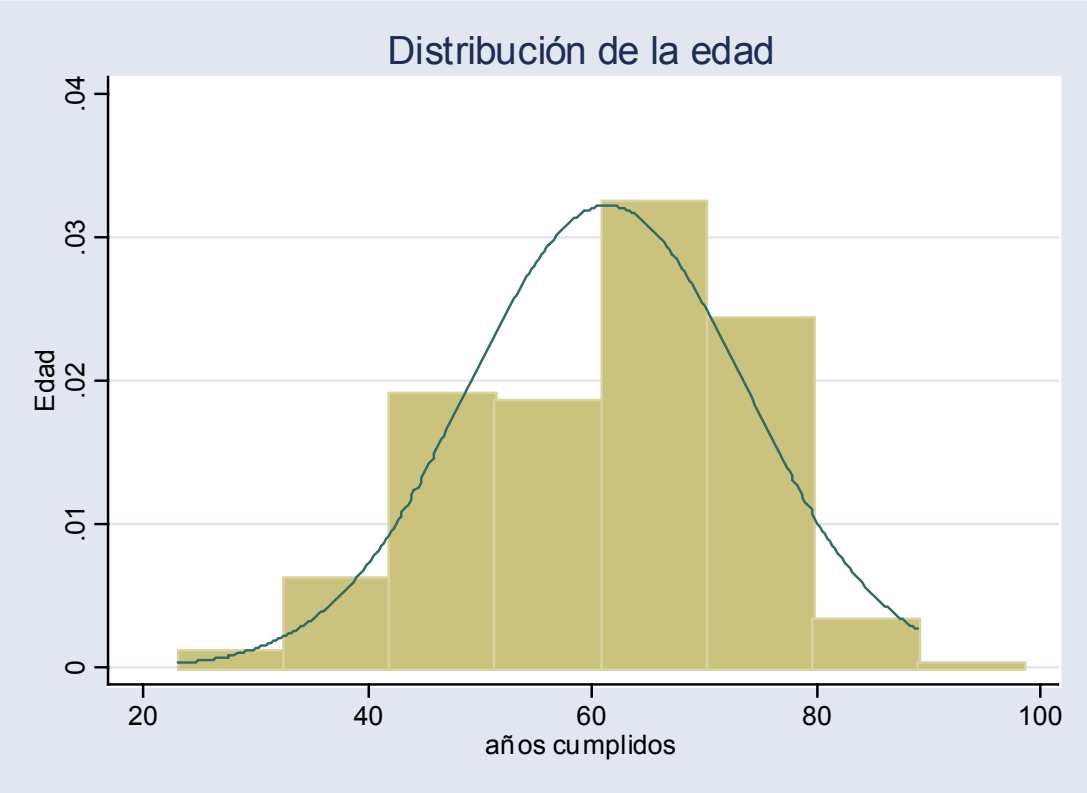
Ordenación y representación de datos

- Variable Edad: Ordenación tabulación

```
. tab  Edad

        Edad |      Freq.     Percent        Cum.
-------------+-----------------------------------
       20-30 |          2        0.58        0.58
       30-40 |         12        3.51        4.09
       40-50 |         49       14.33       18.42
       50-60 |         73       21.35       39.77
       60-70 |        103       30.12       69.88
       70-80 |         91       26.61       96.49
       80-90 |         12        3.51      100.00
-------------+-----------------------------------
       Total |        342      100.00
```

Ordenación y representación de datos

- Variable Edad: Histograma



Ordenación y representación de datos

- Frecuencias ajustadas: Para construir un histograma hay que considerar la siguiente regla:
- “La área de cada barra es proporcional a la frecuencia que representa”

Ordenación y representación de datos

- Cuando se desea construir un histograma en que la tabulación presenta intervalos de clase de distinta longitud, es necesario ajustar por dichos largos usando la siguiente fórmula:

$$f_k^* = \frac{f_k}{l_k}$$

Ordenación y representación de datos

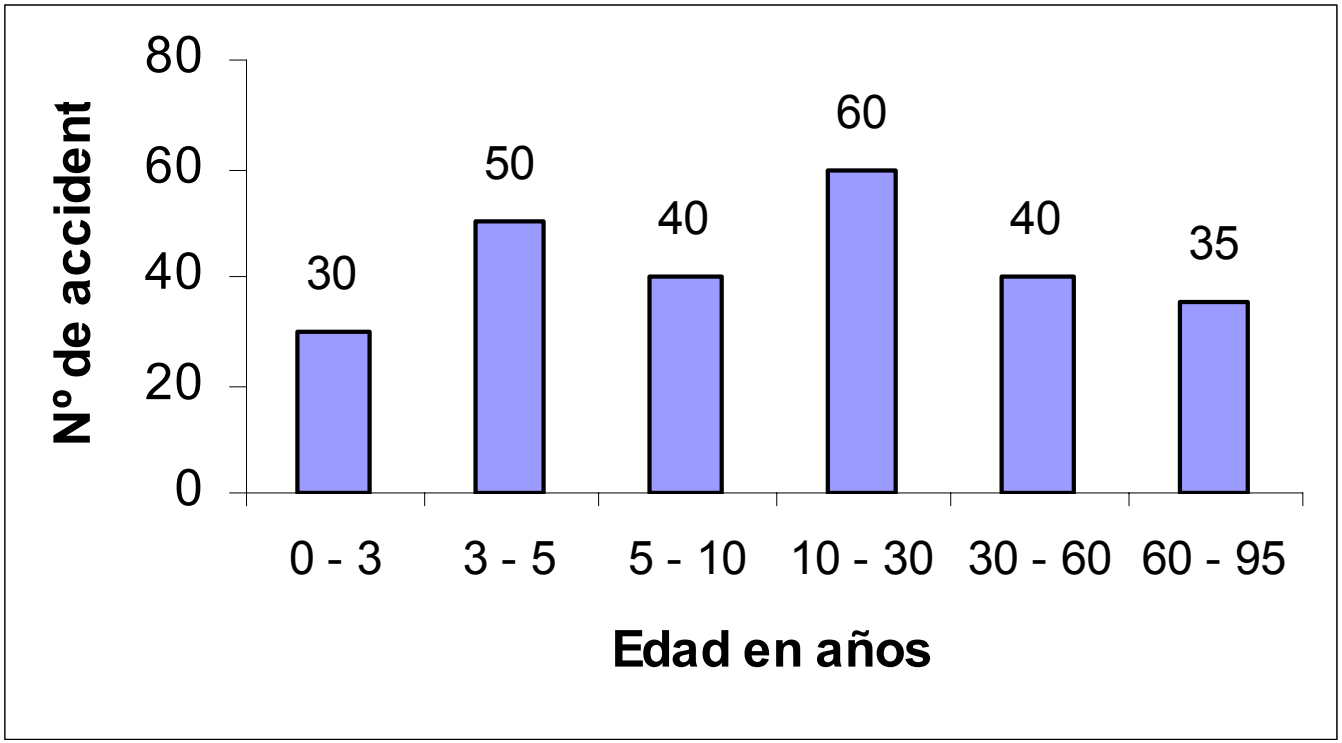
Revisemos el siguiente ejemplo: La siguiente tabla muestra la frecuencia de accidentes caseros por rangos de edad (Hospital Clinic BCN Dic. 2003):

Edad	frec.
0 - 3	30
3 - 5	50
5 - 10	40
10 - 30	60
30 - 60	40
60 - 95	35

Ordenación y representación de datos

Histograma:

Edad	frec.
0 - 3	30
3 - 5	50
5 - 10	40
10 - 30	60
30 - 60	40
60 - 95	35



Ordenación y representación de datos

Histograma:

Edad	frec.
0 - 3	30
3 - 5	50
5 - 10	40
10 - 30	
30 - 60	
60 - 95	35

¡ERROR!

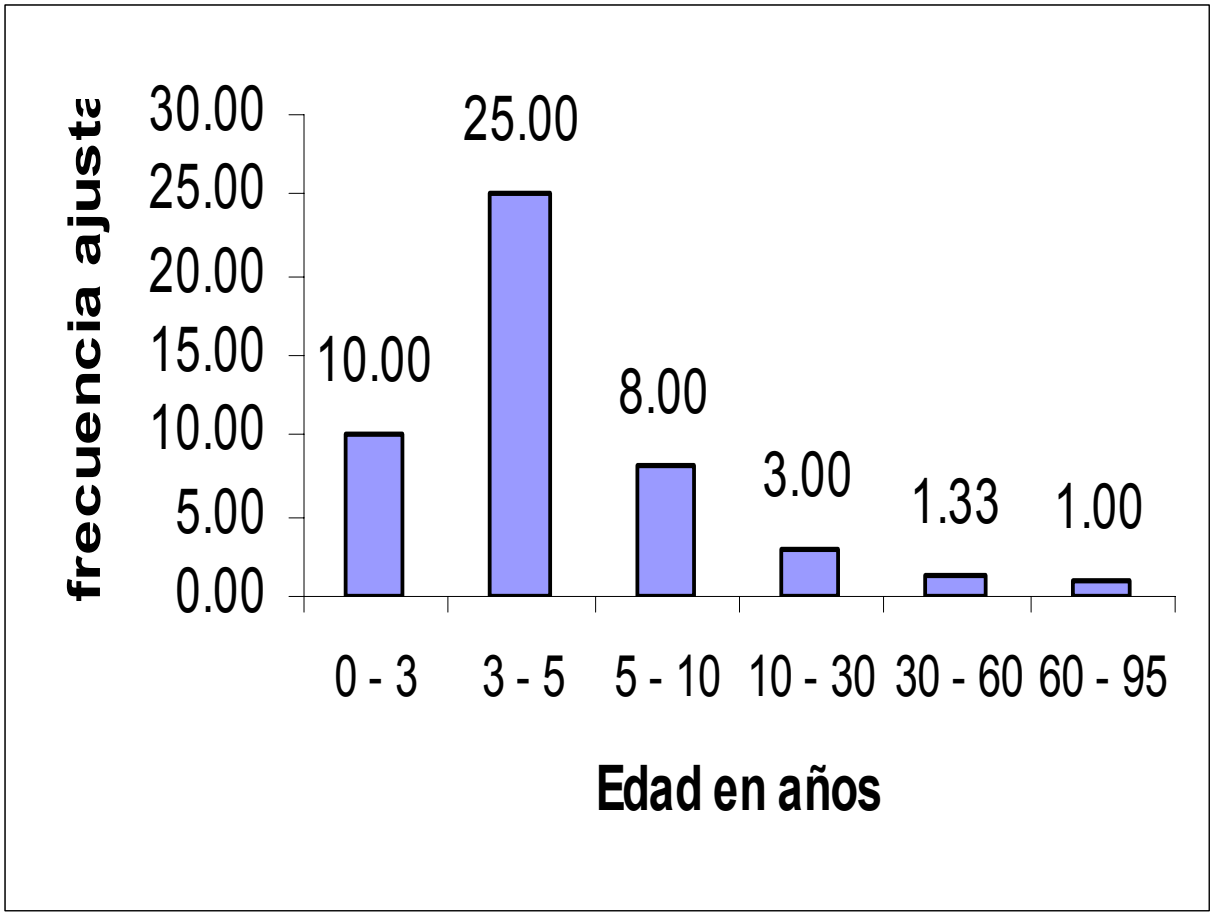
Los intervalos no tienen igual longitud

Ajustar las frecuencias !

Ordenación y representación de datos

Histograma:

Edad	frec.	largo	frec.*
0 - 3	30	3.0	10.00
3 - 5	50	2.0	25.00
5 - 10	40	5.0	8.00
10 - 30	60	20.0	3.00
30 - 60	40	30.0	1.33
60 - 95	35	35.0	1.00




Prueba Contundente del Calentamiento Global
Siglo XVIII
1900
1950
1970
1980
1990
2006

Estadígrafos o estadísticos

Estadígrafos: llamaremos estadígrafo o estadístico, a números resúmenes, que nos permiten establecer conclusiones a cerca de la estructura de una muestra, estos números son construidos considerando TODA la información que contiene dicha muestra, es decir consideran TODOS los datos que han sido recolectados.

Estadígrafos o estadísticos

Pueden construirse estadígrafos para distintos fines, sin embargo estudiaremos cuatro tipos de ellos, estadígrafos de:

- Posición
  - Tendencia central
    - Variabilidad o dispersión
      - Y de forma.

Estadígrafos o estadísticos

Cada vez que la muestra de datos, medidos en al menos en escala ordinal, ha sido ordenada, se establece un Ranking para cada una de las observaciones, este ranking, indica en que posición, en dirección ascendente, se encuentra el dato respecto a la muestra.

Estadígrafos o estadísticos

Este ranking se denota por un subíndice encerrado entre paréntesis. Por ejemplo si se tienen los datos:

12, 7, 15 y 13

al ordenarlos se tiene:

7, 12, 13 y 15

es decir el primer dato ordenado es 7, el segundo es 12 etc. Este hecho lo anotamos simbólicamente como sigue:

$$X_{(1)}=7,\ X_{(2)}=12,\ X_{(3)}=13 \text{ y } X_{(4)}=15$$

Estadígrafos o estadísticos

De este modo la muestra la podemos visualizar sobre un eje ordenado:

Así $X_{(1)}=\text{mín}(X_1,X_2...X_n)$ y $X_{(n)}=\text{máx}(X_1,X_2...X_n)$

Estadígrafos o estadísticos

Estadígrafos de posición: son aquellos que dan información a cerca del orden en la estructura de una muestra.

Ya hemos mencionado dos de ellos que aparecen en forma instantánea al ordenar la muestra, nos referimos al máximo, $X_{(n)}$, y al mínimo, $X_{(1)}$.

Percentiles

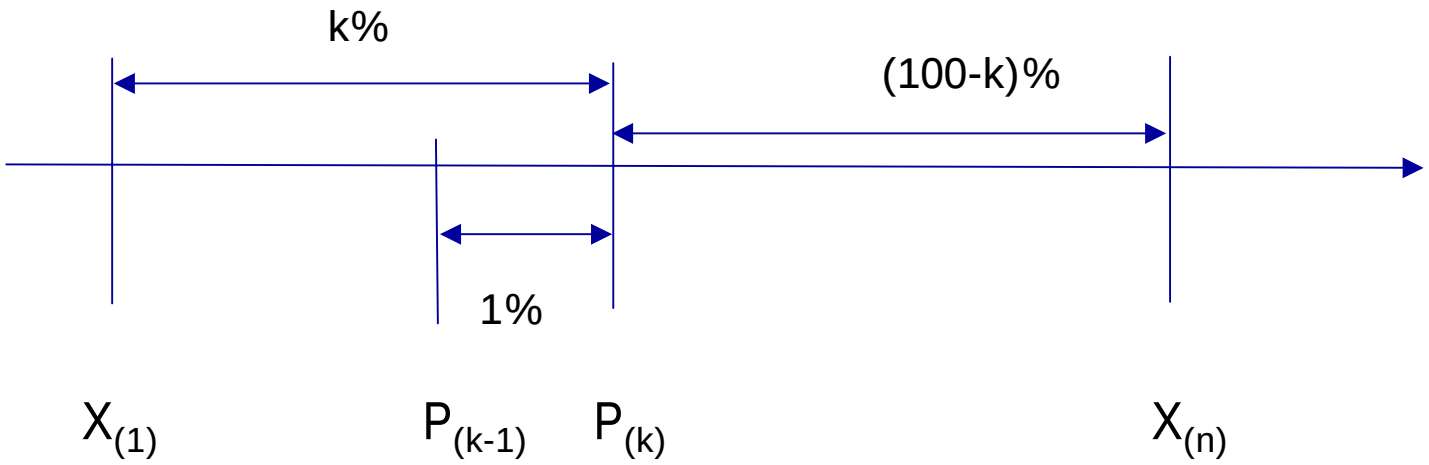
Llamaremos PERCENTILES, a cada uno de los números que dividen la muestra en 100 partes iguales.

- Hay 99 percentiles, y se denotan por $P_{(k)}$, donde k es el orden del percentil indicado.

- Dado el percentil $P_{(k)}$, este divide la muestra en dos partes, la inferior que contiene el k% inferior de las observaciones y la superior que contiene el (100-k)% de las observaciones.

- Entre dos percentiles consecutivos está contenido un 1% de la muestra

Percentiles



Percentiles

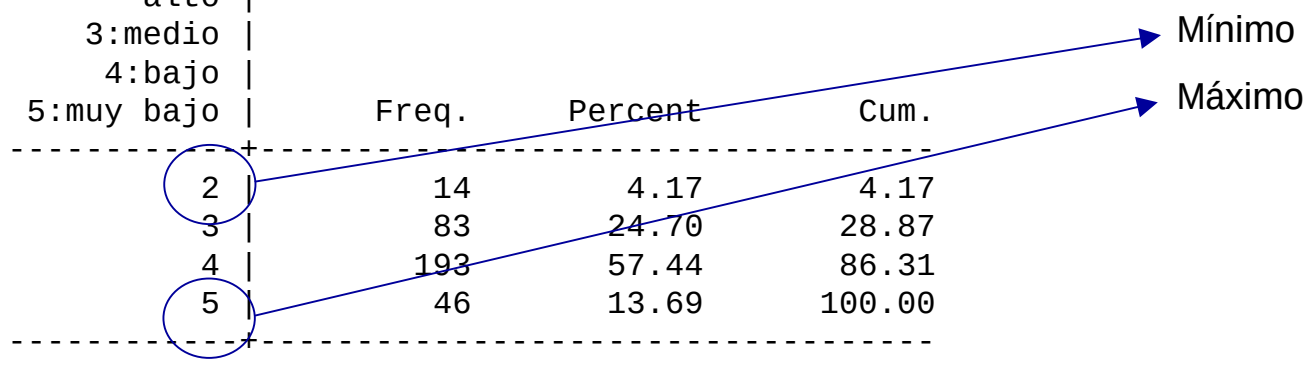
Cálculo de los percentiles para variables medidas en escala ORDINAL o variables de RAZON DISCRETAS:

P_k es el valor de la variable para el cual la frecuencia acumulada IGUALA o SUPERA por primera vez el orden del percentil buscado.

Percentiles

En la base AURI.dta tabulamos la variable nivel social:

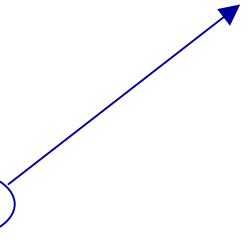
```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```



Percentiles

En la base AURI.dta tabulamos la variable nivel social:

```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```



4.17% supera o iguala por primera vez los órdenes 1,2,3 y 4 %

Percentiles

En la base AURI.dta tabulamos la variable nivel social:

```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```

4.17% supera o iguala por primera vez los órdenes 1,2,3 y 4 %

P_1,P_2,P_3 y P_4 son iguales a 2

Percentiles

Busquemos P_{25}, P_{50} y P_{75}:

```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```

28.87% supera o iguala por primera vez el orden 25%, luego $P_{25}=3$

Percentiles

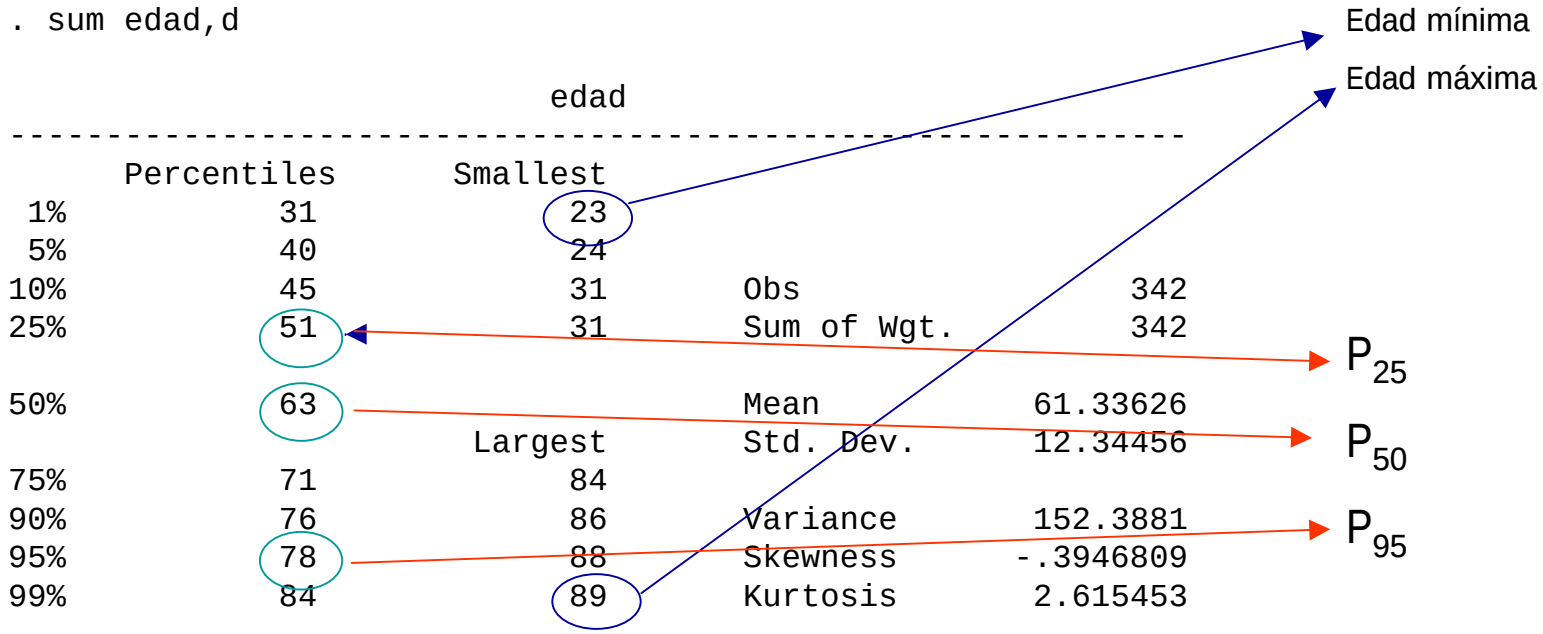
Busquemos P_{25}, P_{50} y P_{75}:

```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```

86.31% supera o iguala los órdenes 50% y 75% por primera vez, luego $P_{50}=4$ y $P_{75}=4$

Percentiles

Si la variable es de naturaleza continua hay que pensar menos. Por ejemplo los percentiles de la Edad en AURI.dta:



Percentiles

Hay percentiles, que por la popularidad de interpretación que tienen, reciben nombre propio: entre ellos están:

- Los Cuartiles: son tres, denotados por Q_1, Q_2 y Q_3 , que corresponden respectivamente a los percentiles P_{25}, P_{50} y P_{75}, ellos dividen la muestra en cuatro partes iguales.

- **Los quintiles: son cuatro, denotados por C_1, C_2, C_3 y C_4, que corresponden respectivamente a los percentiles P_{20}, P_{40}, P_{60} y P_{80}, ellos dividen la muestra en cinco partes iguales.**

- Los deciles: son nueve, denotados por D_1, D_2,...,D_9, que corresponden respectivamente a los percentiles P_{10}, P_{20},..., P_{90}, ellos dividen la muestra en diez partes iguales.

Estadígrafos de centralización

Cada vez que se observa un fenómeno cuantitativo, nos interesa saber si los datos recolectados se aglutinan en torno a ciertos valores representativos que son propios del fenómeno estudiado:

Por ejemplo si pensamos en la Edad de los jugadores profesionales de fútbol, la experiencia nos dice que sus edades varían entre los 17 y 35 años, siendo raro pero no imposible, encontrar jugadores con mas de 35 años o menores de 17 años, además sabemos que la gran mayoría de estos jugadores tienen entre 23 y 30 años.

Ahora la pregunta general se hace obvia, dada una colección de datos, ¿es posible saber a que valores tienden dichos datos?, la respuesta la entregan los llamados estadígrafos de tendencia central.

Estadígrafos de centralización

- En consecuencia llamamos estadísticos de tendencia central a aquellos valores hacia los cuales tienden a aglomerarse los datos de una muestra. Los mas utilizados son:

MODA

MEDIANA

PROMEDIO O MEDIA

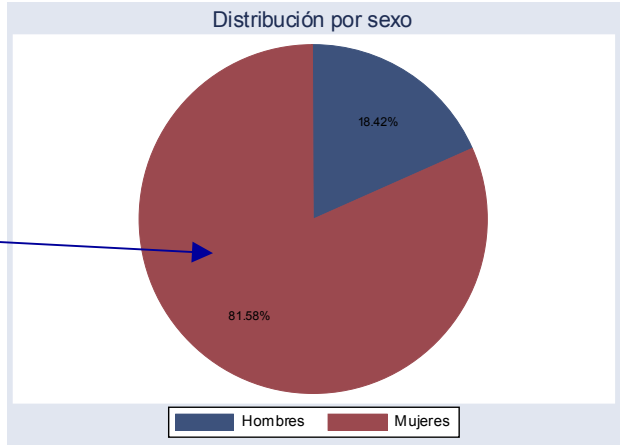
Moda

MODA: es el dato con mayor frecuencia de aparición, apropiada para describir datos medidos en escala **NOMINAL, ORDINAL o DE RAZON PERO DISCRETOS**

Moda en una variable nominal:

¡ Aquí está de MODA ser mujer !

```
. tab sexo

  0: hombre |
   1: mujer |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         63       18.42       18.42
          1 |        279       81.58      100.00
------------+-----------------------------------
      Total |        342      100.00
```

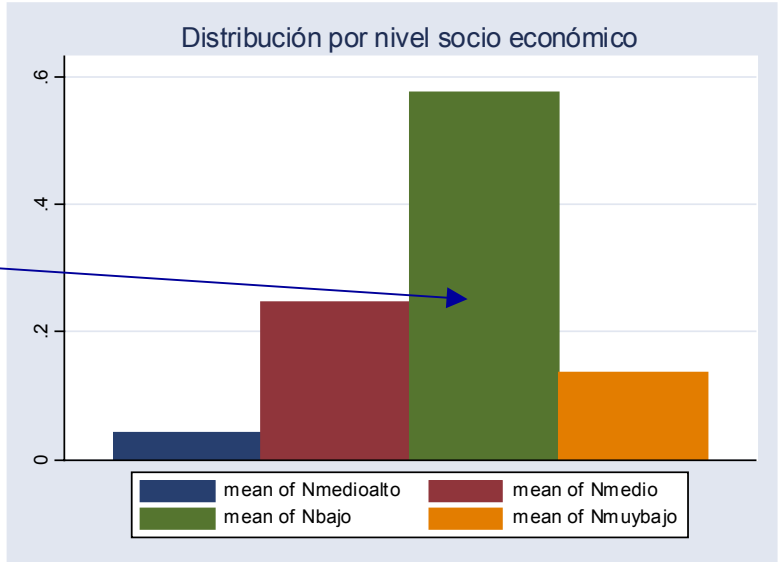


Moda

Moda en una variable ordinal:

```
. tab  nivsocie

     1:alto |
    2:medio |
       alto |
    3:medio |
     4:bajo |
 5:muy bajo |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |         14        4.17        4.17
          3 |         83       24.70       28.87
          4 |        193       57.44       86.31
          5 |         46       13.69      100.00
------------+-----------------------------------
      Total |        336      100.00
```

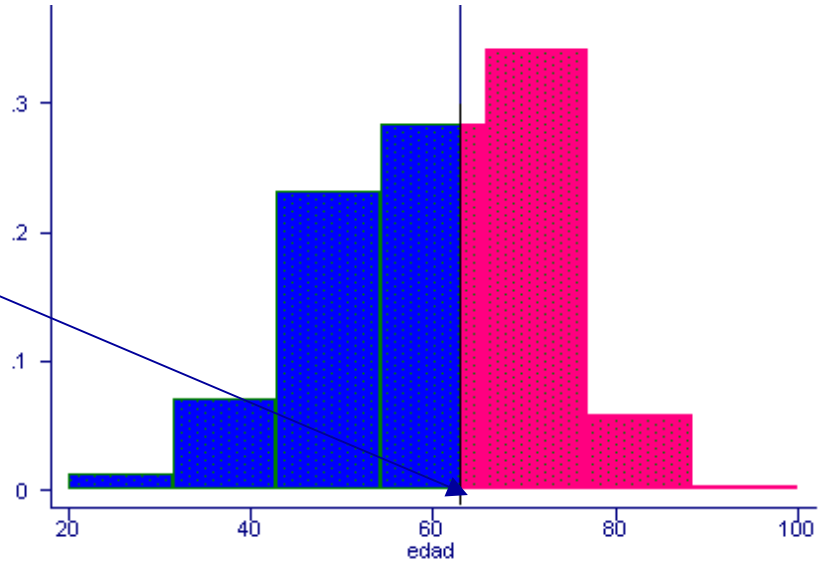
¡ Aquí está de MODA ser de nivel social bajo !



Mediana

MEDIANA: es el punto que divide a la muestra en dos partes iguales, se trata en consecuencia del P_{50} o Q_2 , es apropiada para describir datos medidos en escala **ORDINAL o DE RAZON ya sean discretos o contínuos. La forma de calcularla ya fue revisada.**

```
. tabstat edad, stat(n min q max)

    variable |         N       min       p25       p50       p75       max
-------------+--------------------------------------------------------------
        edad |       342        23        51        63        71        89
----------------------------------------------------------------------------
```



Promedio o media aritmética

MEDIA: es el punto en donde se ubica el centro de masas de la muestra. Es el estadígrafo de tendencia central mas conocido, usado y abusado y se calcula según la fórmula:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Se interpreta como el valor al cual se pueden asimilar todos y cada uno de los datos, es decir, describe bien sólo si la muestra es homogénea y/o simétrica.

Sólo puede ser calculada en variables medidas en escalas **intervalares o de razón**. (Jamás sobre variables medidas en escala ordinal)

Promedio o media aritmética

Ante cambios de escala, tiene las siguientes propiedades:

$$\overline{X \pm a} = \overline{X} \pm a$$

$$\overline{a \cdot X} = a \cdot \overline{X}$$

$$\overline{a} = a$$

Promedio o media aritmética

Ejemplo 1: Promedio de la Edad en AURI.dta

```
. sum edad, d

                            edad
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453
```

Promedio de Edad 61.3 años

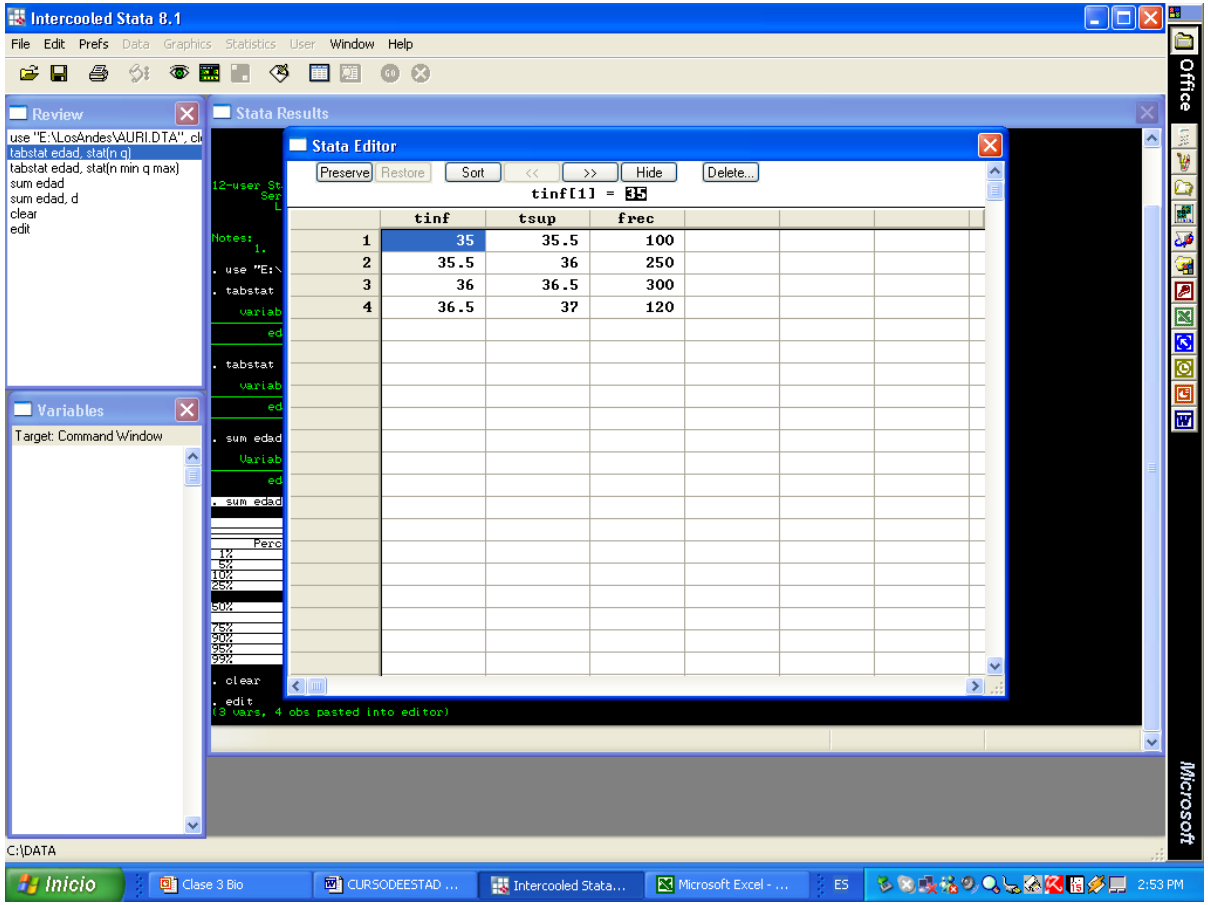
Promedio o media aritmética

Ejemplo 2: Calcular el promedio de temperatura, a partir de la siguiente tabla:

temperatura			frecuencia
35.0	-	35.5	100
35.5	-	36.0	250
36.0	-	36.5	300
36.5	-	37.0	120

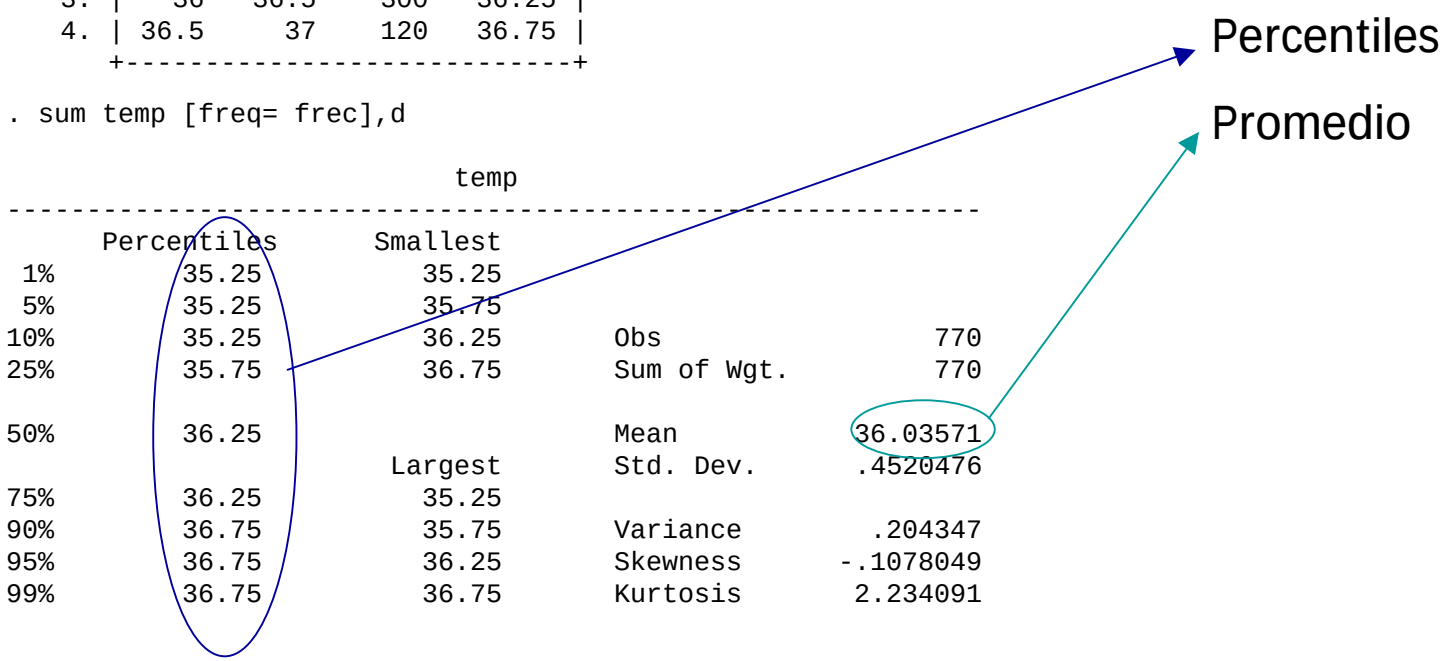
En STATA ingresamos:

Promedio o media aritmética



Promedio o media aritmética

```
. gen temp=( tinf+ tsup)/2

. list

     +---------------------------+
     | tinf   tsup   frec    temp |
     |---------------------------|
  1. |   35   35.5    100   35.25 |
  2. | 35.5     36    250   35.75 |
  3. |   36   36.5    300   36.25 |
  4. | 36.5     37    120   36.75 |
     +---------------------------+

. sum temp [freq= frec],d

                            temp
-------------------------------------------------------------
      Percentiles      Smallest
 1%         35.25          35.25
 5%         35.25          35.75
10%         35.25          36.25       Obs                 770
25%         35.75          36.75       Sum of Wgt.         770

50%         36.25                      Mean           36.03571
                          Largest       Std. Dev.      .4520476
75%         36.25          35.25
90%         36.75          35.75       Variance        .204347
95%         36.75          36.25       Skewness       -.1078049
99%         36.75          36.75       Kurtosis        2.234091
```



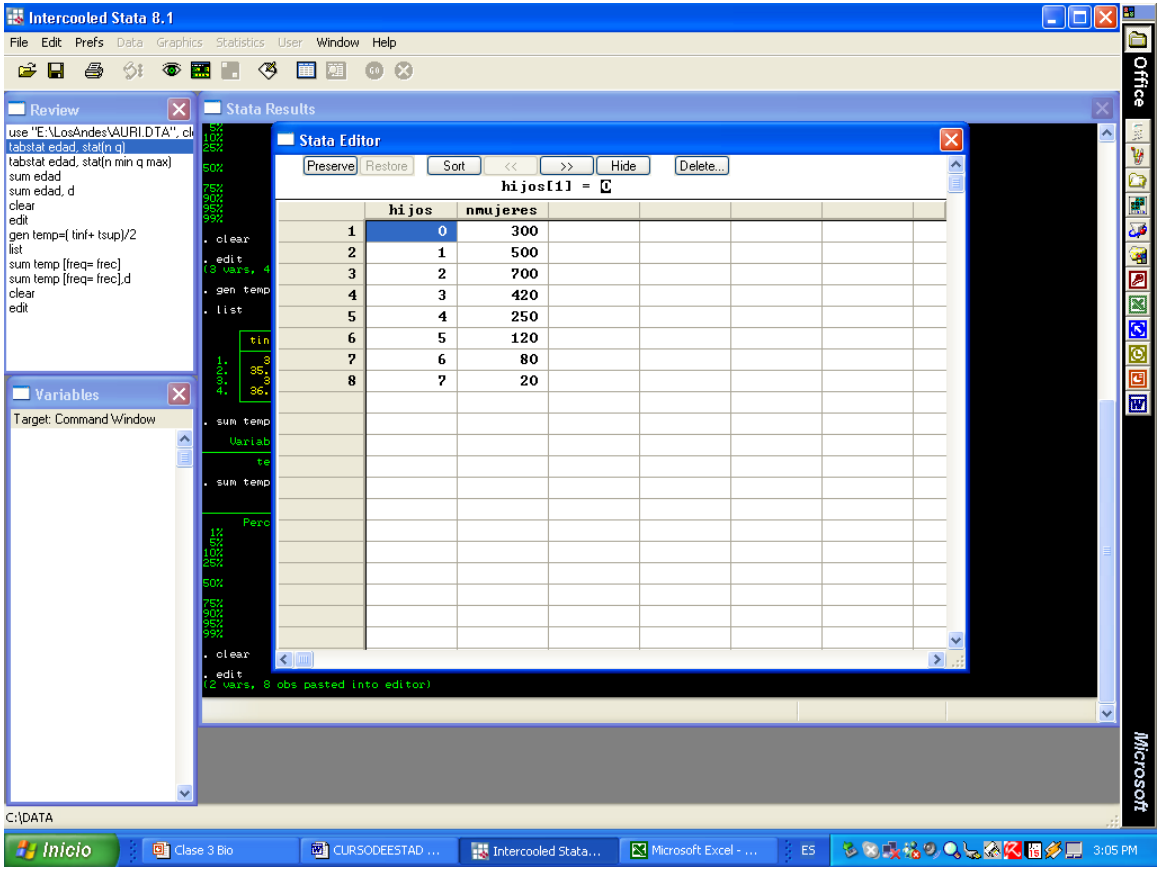
Promedio o media aritmética

Ejemplo 3: Calcular el promedio hijos, a partir de la siguiente tabla:

hijos	nmujeres
0	300
1	500
2	700
3	420
4	250
5	120
6	80
7	20

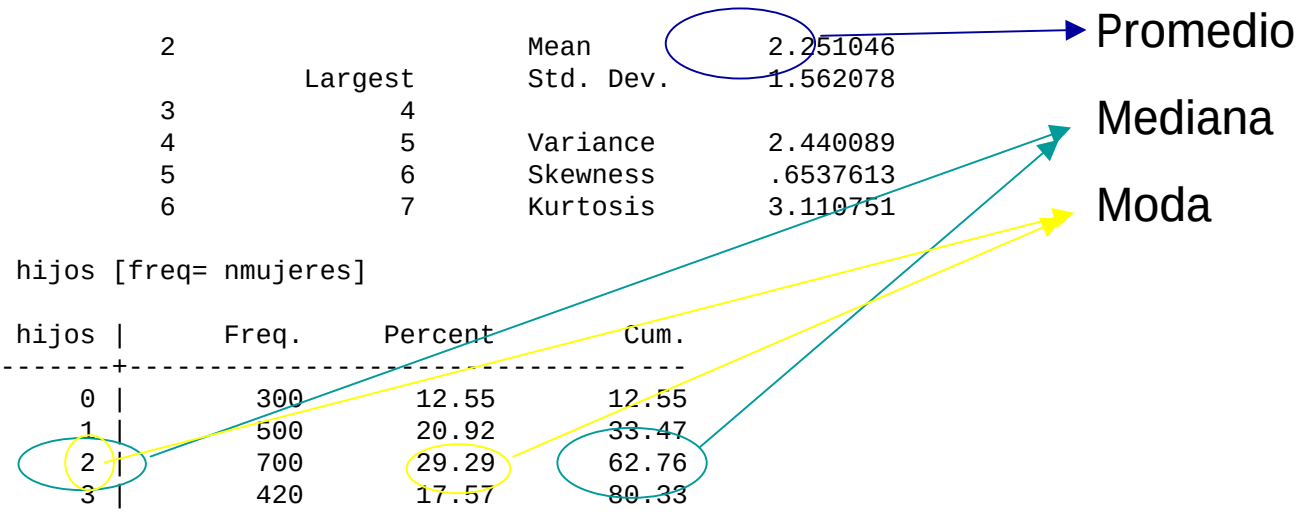
Promedio o media aritmética

Ingresamos en STATA:



Promedio o media aritmética

```
. sum hijos [freq= nmujeres],d

                          hijos
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0             0
 5%             0             1
10%             0             2       Obs                2390
25%             1             3       Sum of Wgt.        2390

50%             2                     Mean           2.251046
                        Largest       Std. Dev.      1.562078
75%             3             4
90%             4             5       Variance       2.440089
95%             5             6       Skewness       .6537613
99%             6             7       Kurtosis       3.110751

. tab hijos [freq= nmujeres]

      hijos |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        300       12.55       12.55
          1 |        500       20.92       33.47
          2 |        700       29.29       62.76
          3 |        420       17.57       80.33
          4 |        250       10.46       90.79
          5 |        120        5.02       95.82
          6 |         80        3.35       99.16
          7 |         20        0.84      100.00
------------+-----------------------------------
      Total |      2,390      100.00
```



Promedio o media aritmética

¿Cómo saber cuando describir con el promedio?

Estadígrafos de variabilidad

Consideremos las calificaciones en bioestadística de dos alumnos: Pedro y Pablo

Alumno								Promedio
Pedro	4.0	4.0	4.0	4.0	4.0	4.0	4.0	**4.0**
Pablo	2.0	3.0	7.0	4.0	6.0	5.0	1.0	**4.0**

Como es observa, tanto Pedro como Pablo tienen idéntico rendimiento promedio. Sin embargo ¿quién tiene rendimiento mas homogéneo?

La respuesta la encontramos en los estadígrafos de variabilidad o dispersión

Estadígrafos de variabilidad

Estudiaremos tres de ellos :

- Recorrido
- Recorrido intercuartílico
- Varianza y desviación estándar

Recorrido

Se llama recorrido de una variable a la diferencia entre el MAXIMO y el MINIMO :

$$\text{Recorrido} = X_{(n)} - X_{(1)}$$

Alumno								Promedio
Pedro	4.0	4.0	4.0	4.0	4.0	4.0	4.0	**4.0**
Pablo	2.0	3.0	7.0	4.0	6.0	5.0	1.0	**4.0**

Aquí:

Recorrido(Pedro)=4.0-4.0=0

Recorrido(Pablo)=7.0-1.0=6

Recorrido

- El recorrido se puede calcular si la variable está medida en a lo menos escala ordinal
- Puede ser una medida muy exagerada de variabilidad

Alumno								Promedio
Pedro	1.0	4.0	4.0	4.0	4.0	4.0	7.0	**4.0**
Pablo	2.0	3.0	7.0	4.0	6.0	5.0	1.0	**4.0**

Aquí:

Recorrido(Pedro)=7.0-1.0=6

Recorrido(Pablo)=7.0-1.0=6

Sin embargo Pedro sigue teniendo un rendimiento mas homogeneo

Recorrido

```
. sum edad,d

                        años cumplidos
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453
```

R(Edad)=89-23=66 años

Recorrido intercuartílico

Se llama recorrido intercuartílico de una variable a la diferencia entre los CUARTILES TERCERO y PRIMERO :

$$\text{Recorrido intercuartílico} = Q_{(3)} - Q_{(1)}$$

Alumno								Promedio
Pedro	1.0	4.0	4.0	4.0	4.0	4.0	7.0	**4.0**
Pablo	2.0	3.0	7.0	4.0	6.0	5.0	1.0	**4.0**

Aquí:

RIC(Pedro)=4.0-4.0 = 0

RIC(Pablo)=6.0-2.0 = 4

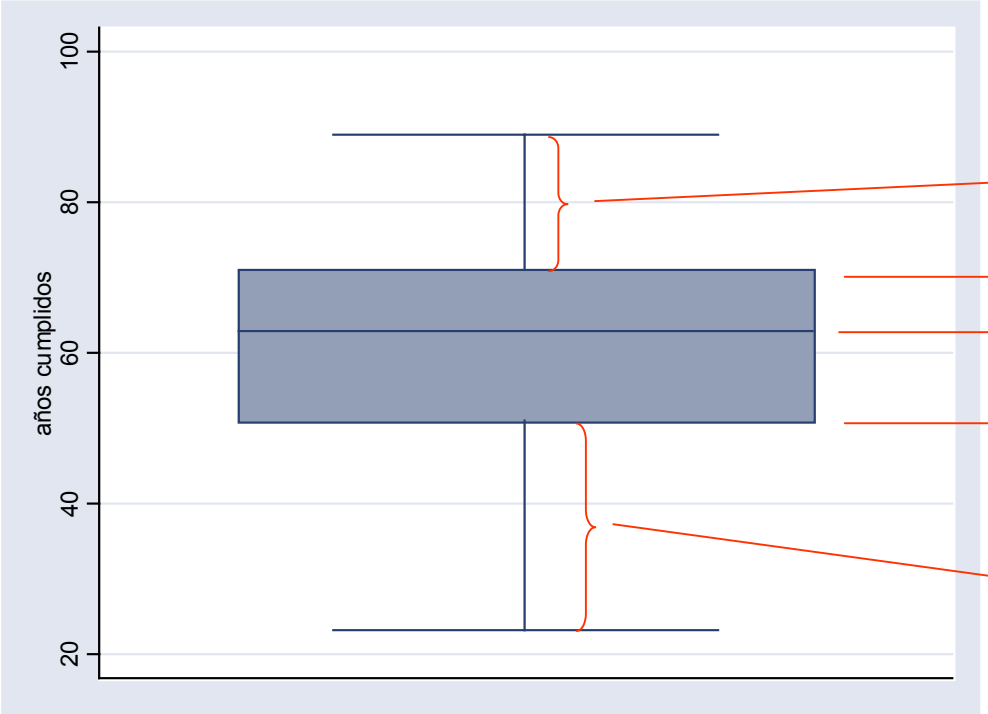
Recorrido intercuartílico

```
. sum edad,d

                        años cumplidos
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453
```

RIC(Edad)=71-51=20 años

Recorrido intercuartílico

Un gráfico muy informativo que relaciona el concepto de cuartil y recorrido intercuartílico, es el llamado CAJON CON BIGOTES (Box plot)



$Q_3+1.5 \cdot RIC$ (lo mas próximo)

Q_3

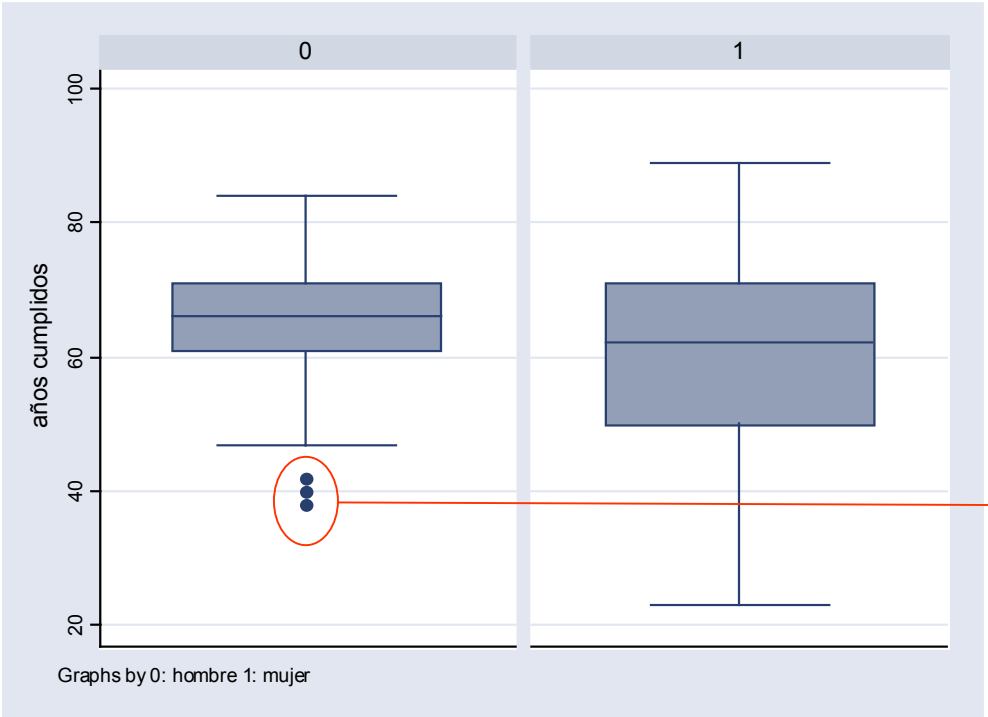
Q_2

Q_1

$Q_1-1.5 \cdot RIC$ (lo mas próximo)

Recorrido intercuartílico

- El CAJON con BIGOTES permite comparar una variable desagregada por otra variable nominal



Varianza

- Llamaremos desvío del i-ésimo dato respecto al promedio a la expresión:

$$d_i = X_i - \overline{X}$$

Es decir la distancia dirigida entre el dato y el promedio

Varianza

- Llamaremos VARIANZA a la expresión:

$$S_x^{\,2} = Var(X) = \frac{d_1^2 + d_2^2 + \ldots + d_n^2}{n-1} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

La varianza sólo se puede calcular para variables medidas en escala intervalar o de razón

Varianza

- La VARIANZA ante cambios de escala tiene las siguientes propiedades:

$$Var(X \pm a) = Var(X)$$

$$Var(a \cdot X) = a^2 \cdot Var(X)$$

$$Var(a) = 0$$

Desviación estándar

- Llamamos DESVIACION ESTANDAR a la RAIZ CUADRADA de la VARIANZA:

$$S_x = \sqrt{\frac{d_1^2 + d_2^2 + \ldots + d_n^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

Desviación estándar

- Cuando la distribución de los datos se acerca a una distribución normal, la mayoría de los datos (alrededor del 67%) está contenido entre:

"EL PROMEDIO MENOS LA DESVIACION y EL PROMEDIO MAS LA DESVIACION"

Desviación estándar

```
. sum edad,d

                        años cumplidos
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453
```

Desviación estándar

Varianza

Coeficiente de variabilidad

- Llamamos COEFICIENTE de VARIABILIDAD a la expresión:

$$C.V. = \frac{S_x}{\overline{X}} \cdot 100\%$$

Coeficiente de variabilidad

```
. sum  edad,d

                        años cumplidos
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453

. display r(sd)/r(mean)*100
20.126034
```

20.1% de variabilidad

Coeficiente de variabilidad

El C.V. sirve para comparar descriptivamente las dispersiones de una variable desagregada por otra.

```
. sum  edad if sexo==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        edad |        63    64.95238    9.258699         38         84

. display r(sd)/r(mean)*100
14.254595

. sum  edad if sexo==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        edad |       279    60.51971    12.81294         23         89

. display r(sd)/r(mean)*100
21.171511
```

La edad de los hombres es mas homogénea

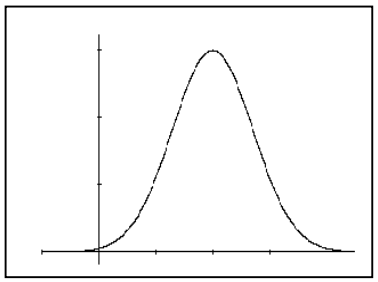
Estadígrafos de forma

- Son aquellos números resúmenes, que indican la morfología de la distribución de los datos, es decir de la simetría y apuntamiento que tiene el histograma de la variable en estudio. Sólo se pueden calcular en variables medidas en escala intervalar y de razón.
- Son el SESGO y la CURTOSIS

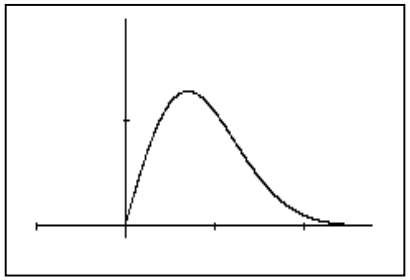
Sesgo

Sesgo: mide el grado de asimetría, respecto de la moda (el máximo del perfil del histograma), que tienen los datos.

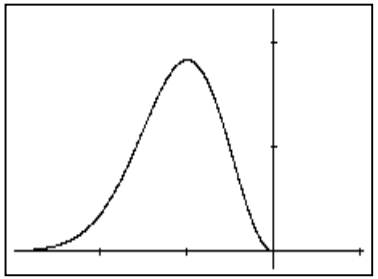
Sesgo = 0



Sesgo > 0



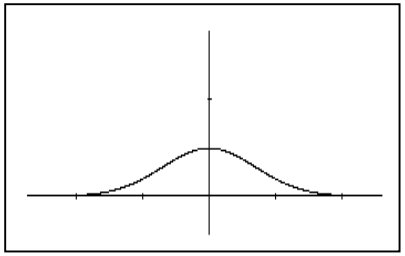
Sesgo < 0



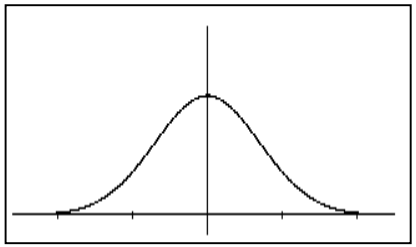
Curtosis

- Curtosis: mide el grado de apuntamiento que tienen los datos

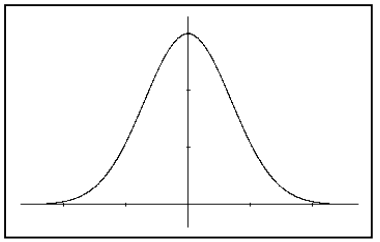
K baja (<3)



K normal (=3)



K alta (>3)



Sesgo y Curtosis

```
. sum  edad,d

                        años cumplidos
-------------------------------------------------------------
      Percentiles      Smallest
 1%           31             23
 5%           40             24
10%           45             31       Obs                 342
25%           51             31       Sum of Wgt.         342

50%           63                      Mean           61.33626
                        Largest       Std. Dev.      12.34456
75%           71             84
90%           76             86       Variance       152.3881
95%           78             88       Skewness      -.3946809
99%           84             89       Kurtosis       2.615453
```

Sesgo

Curtosis

Cálculo de probabilidades

- **Introducción**

  El cálculo de probabilidades tiene su origen en la época pos renacentista, nace del estudio de los juegos de azar, del deseo de poder cuantificar las posibilidades de ganar o perder que se tienen ante una mano de naipes, el lanzamiento de un dado o lanzar una moneda al aire. Sin embargo este interés lúdico inicial trascendió en la historia del pensamiento, pues un análisis mas fino de cualquier situación real nos lleva a considerar una porción de azar (imponderables) que está presente en la misma.

Cálculo de probabilidades

- **¿De qué estamos seguros?**

  Sólo de nuestra muerte “biológica”, la mayoría de las veces cuando decimos que algo será “seguro” en realidad estamos diciendo que es altamente probable que ocurra.

Cálculo de probabilidades

- Al estudiar la realidad podemos distinguir dos tipos de experimentos: Los determinísticos y los probabilísticos.
- Los experimentos determinísticos son aquellos que tienen sólo un resultado posible y además este es predecible.
- Los experimentos probabilísticos son aquellos que tienen mas de un resultado posible y cada resultado no es predecible.

Cálculo de probabilidades

- Dado un experimento cualquiera, que denotaremos por E, llamamos ESPACIO MUESTRAL, denotado por Ω, al conjunto de todos los posibles resultados de E. Como ejemplos tenemos:

  a) E: Se lanza una moneda al aire
  $\Omega = \{$ cara, sello $\}$
  b) E: Se lanza un dado
  $\Omega = \{ 1,2,3,4,5,6 \}$
  c) E: Se juega una cartilla de Loto
  $\Omega_1 = \{$ se gana premio, no se gana premio $\}$ ó
  $\Omega_2 = \{ 0,1,2,3,4,5,6 \}$

Cálculo de probabilidades

- Se llama suceso o evento a cualquier subconjunto de Ω.
- Los sucesos se denotan por letras mayúsculas: A, B,…
- El hecho que A sea un suceso de Ω, lo denotamos por $A \subset \Omega$.
- El conjunto vacío ($\varnothing$) es un suceso, pues $\varnothing \subset \Omega$ y le llamamos suceso vacío o suceso imposible.
- Como $\Omega \subset \Omega$, Ω también es un suceso que llamamos suceso seguro.

Cálculo de probabilidades

Si jugamos al Cara y Sello y previamente se nos pregunta por la “probabilidad” de sacar Cara, seguramente diremos que es de 50%, pues diremos que hay sólo dos posibles resultados, pero además hemos supuesto que las posibilidades de obtener Cara son idénticas a las de obtener Sello, este concepto se denomina EQUIPROBABILIDAD

Cálculo de probabilidades

- Se llama medida de un conjunto a algún número que nos indique el tamaño del conjunto, la medida del conjunto A se denota por m(A).
- Si el conjunto es finito y se pueden contar sus elementos, la medida natural que aparece es m(A)=“número de elementos del conjunto”.
- Si el conjunto es un intervalo de la recta real o una porción del plano cartesiano puede considerarse como m(A)=”longitud del intervalo” o m(A)=”área de la porción del plano cartesiano” según sea el caso.

Cálculo de probabilidades

- **Definición clásica de probabilidad**
- Introducido el concepto de medida, podemos dar una definición de probabilidad del un suceso A como: “medida de A dividido por medida de Ω”, en símbolos:

$$P(A) = \frac{m(A)}{m(\Omega)}$$

Cálculo de probabilidades

- De esta definición aparecen dos resultados fundamentales:
  - $P(\varnothing)=0$, la probabilidad del suceso imposible es nula.
  - $P(\Omega)=1$, la probabilidad del espacio muestral es 1.

Cálculo de probabilidades

- Dos sucesos A y B se dicen excluyentes, si es IMPOSIBLE que ocurran juntos (al mismo tiempo), en símbolos $A \cap B = \varnothing$.
- Por ejemplo se lanza un dado y el dado muestra “un número par e impar” a la vez.

Cálculo de probabilidades

- Hechas las consideraciones anteriores, enunciamos los AXIOMAS del cálculo de probabilidades:
  1. $0 \leq P(A) \leq 1$
  2. Si $A \cap B = \varnothing$ entonces $P(A \cup B) = P(A) + P(B)$

Cálculo de probabilidades

- Para enfrentar un problema de cálculo de probabilidades, se debe poner especial cuidado en definir los sucesos de interés. Ejemplifiquemos con algunas situaciones elementales del experimento “lanzar un dado”:

  E: Se lanza un dado, así: $\Omega=\{1,2,3,4,5,6\}$

  Definamos los sucesos siguientes y calculemos sus probabilidades de ocurrencia:

1. A: “el dado muestra as”, así: $A=\{1\}$ y m(A)=1, con lo que: $P(A)=\frac{1}{6}$
2. B: “el dado muestra un número impar”, así $B=\{1,3,5\}$ y m(B)=3, con lo que:

$$P(B)=\frac{3}{6}=\frac{1}{2}$$

Cálculo de probabilidades

- La realidad presenta sucesos compuestos, los que se forman uniéndolos , intersectándolos y complementándolos.
- Dados los sucesos A y B se tiene:
  - $A \cap B$: sucede A y sucede B (suceden ambos a la vez)
  - $A \cup B$: sucede A ó B, así $P(A \cup B)=P(A) + P(B) – P(A \cap B)$
  - A^c : no sucede A, así $P(A^c)=1 – P(A)$

Cálculo de probabilidades

Decimos de los sucesos A y B son INDEPENDIENTES, si la ocurrencia de uno de ellos no altera la ocurrencia o no ocurrencia del otro, la hipótesis de independencia se expresa así:

$$P(A \cap B) = P(A) \cdot P(B)$$

Cálculo de probabilidades

Además la realidad presenta abundantemente SUCESOS CONDICIONALES, es decir sucesos que condicionan su ocurrencia a la presencia de otros, así podemos preguntarnos por la probabilidad de que ocurra un evento DADO EL HECHO que ocurrió tal o cual evento.

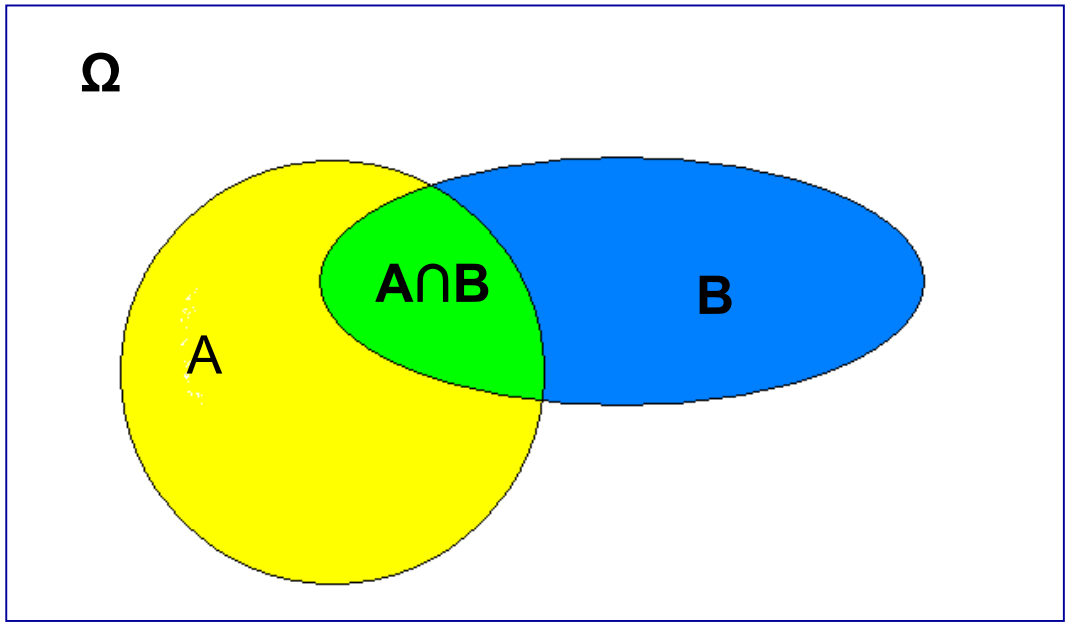
Cálculo de probabilidades

Si consideramos los sucesos A y B, de modo que B condiciona la ocurrencia de A entonces la probabilidad de que “ocurra A dado el hecho que ocurrió B” es:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Cálculo de probabilidades

Condicionar el suceso A al suceso B, es reducir el espacio muestral a B.



Cálculo de probabilidades

- De la fórmula: $$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Tener presente que:

- $P(A \cap B) = P(A|B) \cdot P(B)$
- $P(A|B) \neq P(B|A)$

Cálculo de probabilidades

- Ejemplo: Considerar la siguiente tabla:

	Sano	Enfermo	
Mujer	6	2	**8**
Hombre	8	4	**12**
	14	**6**	**20**

Aquí se pueden distinguir cuatro sucesos, de los cuales dos son fundamentales:

A : la persona es MUJER

B : la persona está SANA

A^c : la persona es HOMBRE

B^c : la persona está ENFERMA

Cálculo de probabilidades

	Sano (B)	Enfermo (B^c)	
Mujer (A)	6	2	**8**
Hombre (A^c)	8	4	**12**
	14	**6**	**20**

-P(A) = 8/20= 0.40, la probabilidad de ser mujer.

- P(B) = 14/20=0.60, la probabilidad de estar sano.

-$P(A \cap B^c)$=2/20=0.10, la probabilidad de ser mujer y estar enfermo.

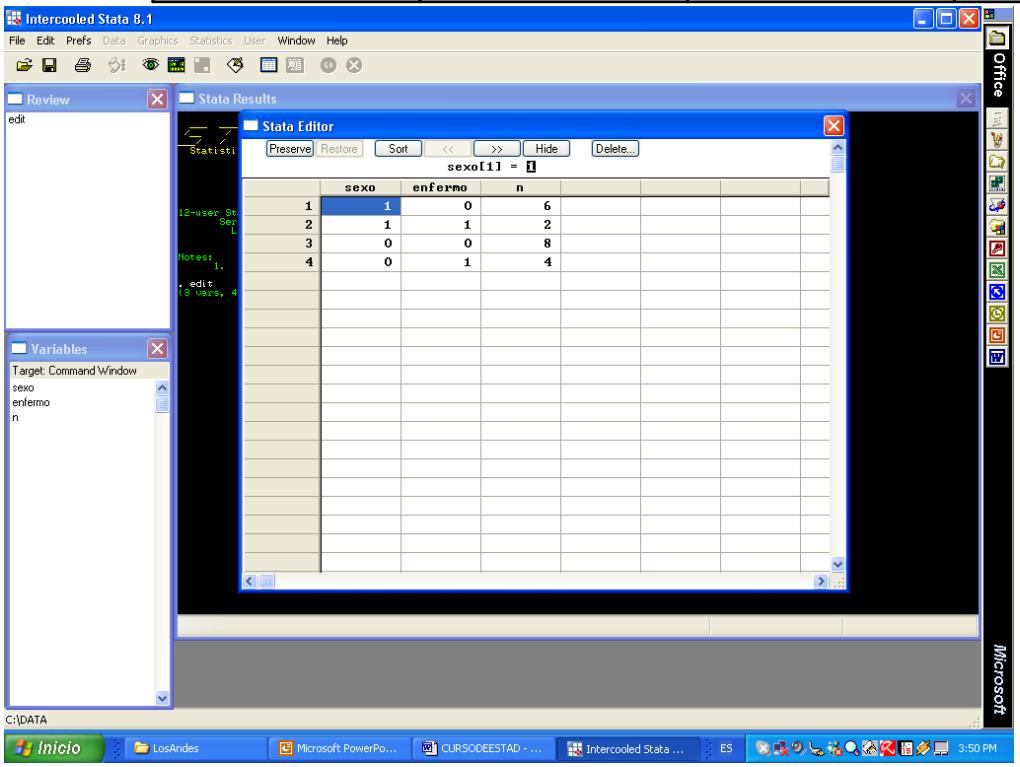
-P(A|B)=6/14=0.43, la probabilidad de ser mujer dado que está sano.

P(B|A)=6/8=0.75, la probabilidad de estar sano dado que es mujer.

Cálculo de probabilidades

- En STATA

	Sano (B)	Enfermo (B^c)	
Mujer (A)	6	2	**8**
Hombre (A^c)	8	4	**12**
	14	**6**	**20**



Cálculo de probabilidades

- En STATA

```
. tab sexo enfermo [freq=n],  row col cell

+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

  0:hombre |   0:sano 1:enfermo
   1:mujer |         0          1 |     Total
-----------+----------------------+----------
         0 |         8          4 |        12
           |     66.67      33.33 |    100.00
           |     57.14      66.67 |     60.00
           |     40.00      20.00 |     60.00
-----------+----------------------+----------
         1 |         6          2 |         8
           |     75.00      25.00 |    100.00
           |     42.86      33.33 |     40.00
           |     30.00      10.00 |     40.00
-----------+----------------------+----------
     Total |        14          6 |        20
           |     70.00      30.00 |    100.00
           |    100.00     100.00 |    100.00

           |     70.00      30.00 |    100.00
```

Probabilidad de estar sano dado que se es hombre = 8/12

Probabilidad de ser hombre dado que se está sano = 8/14

Probabilidad de ser hombre y estar sano = 8/20

Probabilidad de ser mujer = 8/20

Probabilidad de estar enfermo = 6/20

Cálculo de probabilidades

- En múltiples oportunidades la ocurrencia de un suceso principal A se debe a la ocurrencia previa de causas, que también son sucesos, de modo que en el cálculo de la probabilidad de la ocurrencia de A las probabilidades de los sucesos causales deben ser incluidas según la ponderación o influencia que tengan sobre A.

  Si el suceso principal A se debe a las causas E_1, E_2,...,E_n , entonces:

Cálculo de probabilidades

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \ldots + P(A|E_n)P(E_n)$$

$$=$$

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i)$$

Esta fórmula recibe el nombre de TEOREMA DE LA PROBABILIDAD TOTAL

Cálculo de probabilidades

Ejemplo: En un hospital hay tres servicios: Urgencia, Cirugía y Medicina. El porcentaje de hospitalizados por servicio es: Urgencia 30%, Cirugía 20% y Medicina 50%. Si la mortalidad en cada servicio es 10%, 5% y 3% respectivamente. ¿Cuál es la probabilidad de que un paciente hospitalizado muera?

Suceso principal, A : el paciente muere

Causas : E_1: el paciente está en urgencia

E_2: el paciente está en cirugía

E_3: el paciente está en medicina

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + P(A|E_3)P(E_3)$$

$$P(A) = 0.1 \cdot 0.3 + 0.05 \cdot 0.2 + 0.03 \cdot 0.5 = 0.055$$

Cálculo de probabilidades

- En ocasiones es necesario calcular la probabilidad de que una determinada causa haya producido el suceso principal. Es decir necesitamos saber $P(E_k|A)$.
- En el ejemplo: Si se nos comunica que ha ocurrido una muerte, ¿Cuál es la probabilidad que haya ocurrido en Urgencia?
  Suceso principal, A : el paciente muere.

  Causas:E_1: el paciente está en Urgencia;E_2: el paciente está en Cirugía; E_3: el paciente está en Medicina
  Es decir se pide:

$$P(E_1 \mid A) = \frac{P(E_1 \cap A)}{P(A)} = \frac{P(A \cap E_1)}{P(A)} = \frac{P(A \mid E_1) \cdot P(E_1)}{P(A \mid E_1) \cdot P(E_1) + P(A \mid E_2) \cdot P(E_2) + P(A \mid E_3) \cdot P(E_3)}$$

Cálculo de probabilidades

Suceso principal, A : el paciente muere

Causas : E_1: el paciente está en urgencia

E_2: el paciente está en cirugía

E_3: el paciente está en medicina

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + P(A|E_3)P(E_3)$$

$$P(A) = 0.1 \cdot 0.3 + 0.05 \cdot 0.2 + 0.03 \cdot 0.5 = 0.055$$

$$P(E_1 \mid A) = \frac{P(E_1 \cap A)}{P(A)} = \frac{P(A \cap E_1)}{P(A)} = \frac{P(A \mid E_1) \cdot P(E_1)}{P(A \mid E_1) \cdot P(E_1) + P(A \mid E_2) \cdot P(E_2) + P(A \mid E_3) \cdot P(E_3)}$$

$$P(E_1 \mid A) = \frac{0.1 \cdot 0.3}{0.055} = 0.545$$

Cálculo de probabilidades

- Generalizando el resultado anterior

$$P(E_k \mid A) = \frac{P(A \mid E_k) \cdot P(E_k)}{P(A \mid E_1) \cdot P(E_1) + P(A \mid E_2) \cdot P(E_2) + \ldots + P(A \mid E_n) \cdot P(E_n)}$$

ó

$$P(E_k \mid A) = \frac{P(A \mid E_k) \cdot P(E_k)}{\sum_{i=1}^{n} P(A \mid E_i) \cdot P(E_i)}$$

Fórmula que es conocida como el TEOREMA DE BAYES

Variables aleatorias

- **Introducción**

  Una variable aleatoria, en general, es una codificación numérica de los posibles resultados que contiene el espacio muestral de un experimento, dicha codificación puede ser arbitraria, sin embargo, si el espacio muestral tiene algún orden jerárquico específico, este mismo orden sugiere la codificación. El empleo de variables aleatorias permite descubrir nuevas propiedades del experimento que se está estudiando.

Variables aleatorias

- Como ejemplo, retomemos el experimento de lanzar un dado, para lo cual tenemos:

  E: Se lanza un dado

  $\Omega = \{1,2,3,4,5,6\}$

el espacio muestral, tiene seis sucesos fundamentales:

A1: el dado muestra as

A2: el dado muestra 2

.....

A6: el dado muestra 6

Variables aleatorias

Sin embargo, estos sucesos pueden ser codificados por la variable X, que tal que: X=1 si ocurre A1 , X=2 si ocurre A2 etc. Hecha esta codificación, podemos hacer una descripción completa del experimento, pues las probabilidades asociadas con cada suceso se pueden representar por una función, que recibe el nombre de FUNCION DE CUANTIA DE PROBABILIDADES

X	P(X)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6
Total	1

Observamos que la suma de las P(X) es igual a 1, pues es la probabilidad del espacio muestral completo.

Variables aleatorias

En general una Función de Cuantía de Probabilidades, es una función, cuyo dominio es un subconjunto, *A,* de los Naturales, que cumple las siguientes propiedades:

$$P(X): A \subseteq N \rightarrow R$$

$$P(X) \geq 0$$

$$\sum_{A} P(X) = 1$$

Se llama FUNCION DE DISTRIBUCION DE PROBABILIDADES a la expresión:

$$F(X_j) = P(X \leq X_j) = \sum_{min(A)}^{j} P(X_i)$$

Variables aleatorias

Dada una función de cuantía, podemos definir el valor promedio de ella, que llamaremos esperanza o valor esperado de la variable, que denotamos por E(X) o μ , y la definimos por:

$$E(X) = \mu = X_1 \cdot P(X_1) + X_2 \cdot P(X_2) + ... + X_n \cdot P(X_n) = \sum_{i=1}^{n} X_i \cdot P(X_i)$$

Ante cambios de escala, E(X) tiene las siguientes propiedades:

$$E(X \pm a) = E(X) \pm a$$

$$E(aX) = aE(X)$$

$$E(a) = a$$

Generalizando este resultado, se llama MOMENTO DE ORDEN K respecto del origen a la expresión:

$$E(X^k) = \sum_{i=1}^{n} X_i^k \cdot P(X_i)$$

Variables aleatorias

También podemos medir la variabilidad de la variable aleatoria X, mediante el cálculo de lo que llamaremos VARIANZA de X, que denotaremos por V(X), y definimos como sigue:

$$V(X) = (X_1 - \mu)^2 \cdot P(X_1) + (X_2 - \mu)^2 \cdot P(X_2) + ... + (X_n - \mu)^2 \cdot P(X_n) = \sum_{i=1}^{n} (X_i - \mu)^2 \cdot P(X_i)$$

ó

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

A la raíz cuadrada de la varianza se le llama desviación estándar de X, la denotamos por σ y junto con μ tienen la misma interpretación que en la estadística descriptiva.

Ante cambios de escala, V(X) tiene las siguientes propiedades:

$$V(X \pm a) = V(X)$$

$$V(aX) = a^2 V(X) \rightarrow \sigma(aX) = |a|\sigma(X)$$

$$V(a) = 0$$

La distribución uniforme

Función de cuantía:

X	P(X)
1	$\frac{1}{n}$
2	$\frac{1}{n}$
...	$\frac{1}{n}$
n	$\frac{1}{n}$

$$E(X) = \frac{n+1}{2}$$

$$V(X) = \frac{n^2 - 1}{12}$$

La distribución geométrica

El modelo geométrico permite calcular la probabilidad de “OBTENER ÉXITO POR PRIMERA VEZ EN EL K-ESIMO INTENTO”, así si X es la variable que denota el intento donde se produce el éxito por primera vez, X puede tomar valores desde 1 al infinito, pues el éxito podría aparecer en el primer intento o bien podríamos pasarnos la vida completa esperando que se produzca el éxito, así:

$$P(X = k) = q^{k-1} \cdot p, X = 1,2,3,....$$

$$E(X) = \frac{1}{p}$$

$$V(X) = \frac{q}{p^2}$$

Distribución Binomial

El modelo binomial, permite calcular la probabilidad de tener k éxitos en n intentos, si tenemos n intentos la cantidad de éxitos que podríamos obtener van desde 0 a n, es decir X=0, 1, 2, ..., n . En este contexto:

$$P(X=k)=\binom{n}{k}\cdot q^{n-k}\cdot p^{k}, X=0,1,2,3...,n$$

$$m_x(t)=(q+pe^t)^n$$

$$E(X)=n\cdot p$$

$$V(X)=n\cdot p\cdot q$$

Distribución de Poisson

El modelo probabilístico de Poisson, calcula la probabilidad de ocurrencia de fenómenos de rara ocurrencia ya sea por: unidad de tiempo, de longitud de área etcétera. Dado un fenómeno de rara ocurrencia por alguna unidad de medida, es posible, por la experiencia acumulada, establecer una tasa de ocurrencia que llamaremos λ. En estas condiciones la variable X es la cantidad de veces que aparece el fenómeno en un período, así X puede tomar valores desde 0 al infinito, con lo que:

$$P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, X = 0,1,2,3....$$

$$E(X) = \lambda$$

$$V(X) = \lambda$$

Distribución Hipergeométrica

Supongamos ahora que tenemos un conjunto con N de objetos, de los cuales r son de una determinada característica de interés, por lo tanto N-r no tienen la característica de interés. Si de este conjunto de N objetos sacamos una muestra de tamaño n, nos interesa la probabilidad de que en dicha muestra hayan k objetos de interés, así esta probabilidad está dada por:

$$P(k) = \frac{\binom{r}{k} \cdot \binom{N-r}{n-k}}{\binom{N}{n}}$$

$$E(k) = \frac{nr}{N}$$

$$V(k) = \frac{nr(N-r)(N-n)}{N^2(N-1)}$$

Experimentos de Bernoulli

- Antes de continuar revisando otras importantes funciones de cuantía de probabilidad, definamos lo que entenderemos por EXPERIMENTOS DE BERNOULLI. En efecto es una secuencia de experimentos que tiene las siguientes características:
- El experimento tiene sólo dos posibles resultados, que llamaremos éxito y fracaso.
- Cada vez que se repite el experimento, la probabilidad de aparición del éxito (y de fracaso) se mantiene constante.
- Cada ensayo es independiente de otro.

Experimentos de Bernoulli

- Si llamamos p a la probabilidad del éxito, obviamente la probabilidad del fracaso es 1-p al que llamaremos q, es decir q=1-p o bien p+q=1

Distribución de Bernoulli

- En una población que esta dicotomizada respecto de un determinado atributo (los elementos que poseen el atributo versus el resto de la población), en que la proporción con el atributo es p y q=1-p la proporción que no lo posee, se realiza el experimento de extraer un elemento y se observa la presencia del atributo, podemos asumir la codificación:

  X=0, si el objeto no tiene el atributo

  X=1, si el objeto tiene el atributo,

  con lo que se obtiene la siguiente función de cuantía:

Distribución de Bernoulli

X	P(X)
0	q
1	p

O bien: $P(X=x)=p^{x}q^{1-x}, x=0,1$ así se tiene que:

$$E(X)=p$$

$$m_x(t)=q+pe^{t}$$

$$V(X)=pq$$

La distribución normal

- **Introducción**

  Es la distribución mas querida usada y abusada por los usuarios de la estadística.

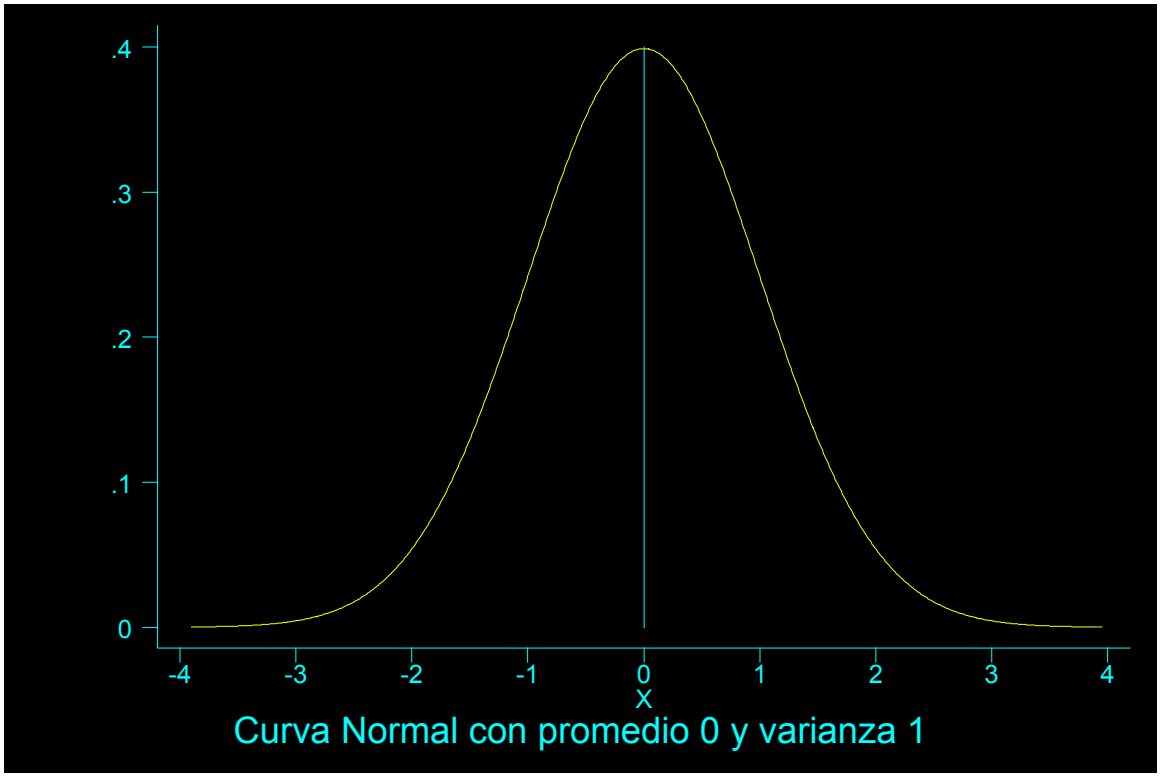
  Decimos que la variable aleatoria, X, sigue una distribución normal con promedio (o esperanza) μ y varianza σ^2, si la función densidad de probabilidades (curva perfil del histograma) está dada por:

La distribución normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in R, \mu \in R, \sigma > 0$$

La distribución normal

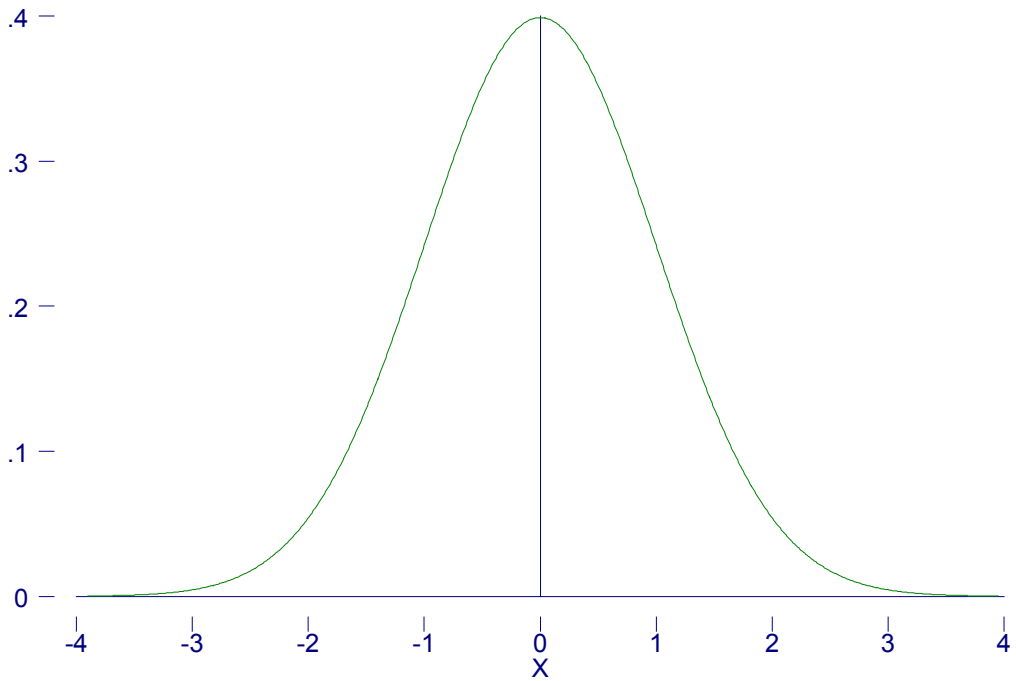
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in R, \mu \in R, \sigma > 0$$



La distribución normal

El gráfico de esta curva es tal que:

- **Tiene un máximo en $x=\mu$**
- **Es simétrica respecto a la vertical $x=\mu$**
- **Tiene puntos de inflexión en $x=\mu - \sigma$ y $x=\mu + \sigma$**
- **Se aproxima asintóticamente al eje X, lo que se refleja en la relación:** $f(\mu - 3\sigma) = f(\mu + 3\sigma) = \frac{1}{100} f(\mu)$

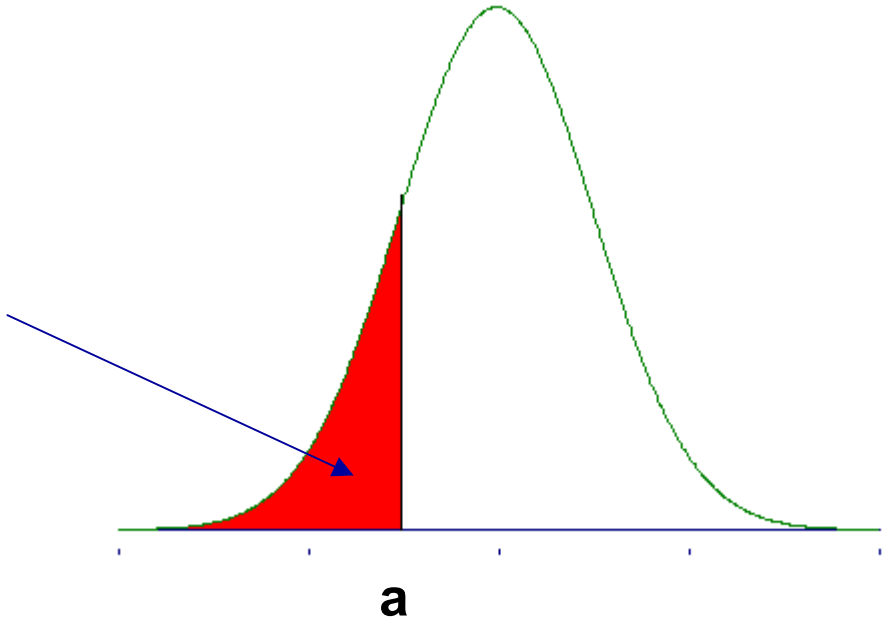


Curva Normal con promedio 0 y varianza 1

La distribución normal

La probabilidad, P(X<a) está dada por:

$$P(X < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

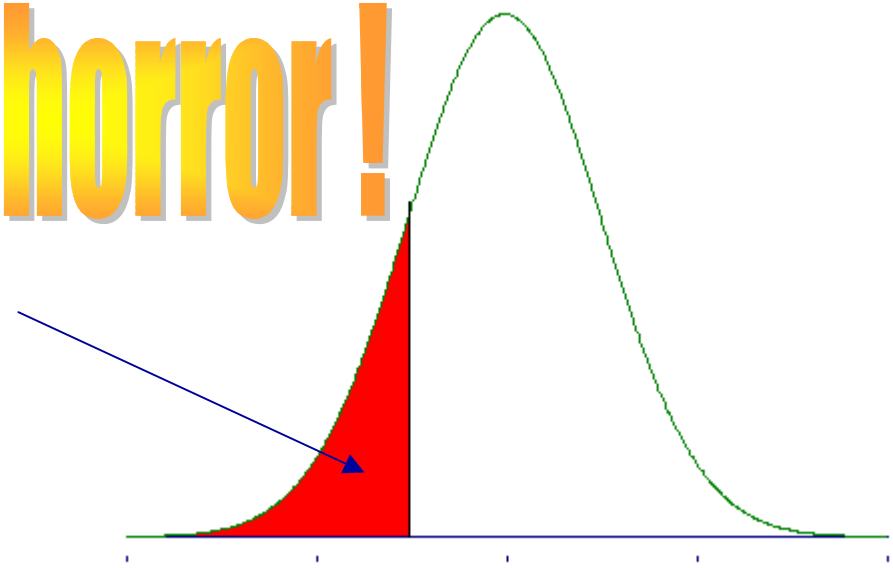


La distribución normal

La probabilidad, P(X<a) está dada por:

¡Que horror!

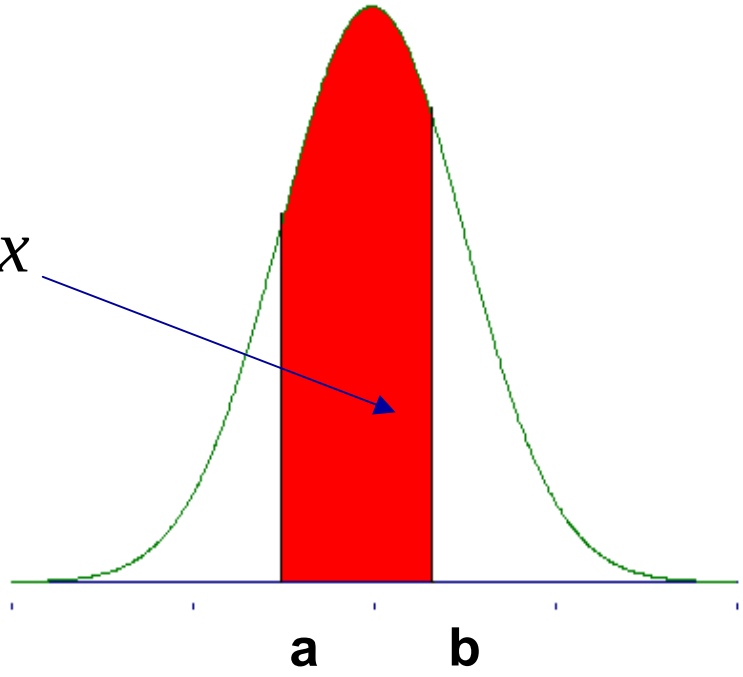
$$P(X < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



La distribución normal

¡Calma, los cálculos serán muy simples!

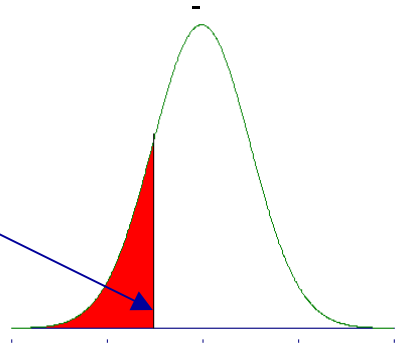
$$P(a<X<b)=\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$



La distribución normal

Si μ=0 y σ=1 se habla de una distribució estándar, típica o reducida.

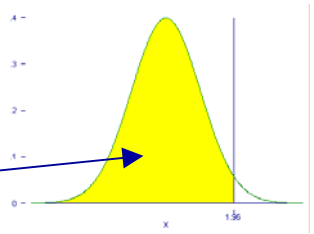
En la normal estándar a la P(Z<z)=Φ(z)



Valor que despliega STATA:

P(Z<1.96)=Φ(1.96)

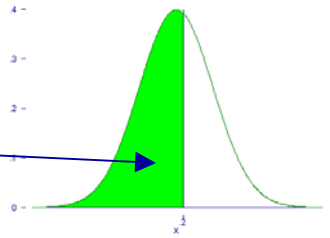
```
 display norm(1.96)
.9750021
```



z

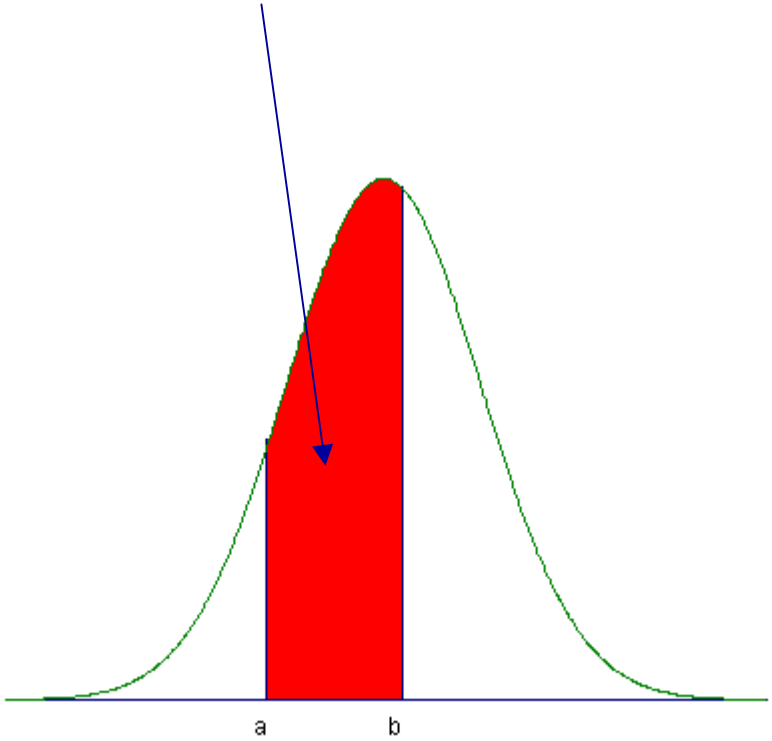
P(Z<0.2)=Φ(0.2)

```
. display norm(0.2)
.57925971
```



La distribución normal

La probabilidad $P(a<Z<b)=\Phi(b)-\Phi(a)$



La distribución normal

Ejemplo: P(1.2<Z<2.5)=Φ(2.5)- Φ(1.2)

```
display norm(2.5)-norm(1.2)
.10886
```

La probabilidad P(Z>a)=1- Φ(a)

Ejemplo: P(Z>1.5)=1-Φ(1.5)

```
display 1-norm(1.5)
.0668072
```

La distribución normal

Si $X \sim N(\mu, \sigma^2)$ entonces

$Z=(X-\mu)/\sigma \sim N(0,1)$ es decir normal estándar.

La distribución normal

Ejemplo:Si la temperatura,T, de una persona sana sigue una distribución normal con media de 36.5º y desviación estándar 0.1º, calcular:

La distribución normal

- P(T<36.3)=Φ((36.3-**36.5**)/**0.1**)

```
display norm((36.3-36.5)/0.1)
.02275013
```

- P(36.4<T<36.8)=Φ((36.8-**36.5**)/**0.1**)- Φ((36.4-**36.5**)/**0.1**)

```
display norm((36.8-36.5)/0.1)-norm((36.4-
36.5)/0.1)
.83999485
```

- P(T>36.9)=1-Φ((36.9-**36.5**)/**0.1**)

```
display 1-norm((36.9-36.5)/0.1)
.00003167
```

La distribución normal

¿Cuál es el percentil 75 de las temperaturas?

Es decir para que valor de t se tiene:

$P(T<t)=0.75$

```
display 36.5 + 0.1*invnorm(.75)
36.567449
```

¿Y el percentil 99?

```
display 36.5 + 0.1*invnorm(.99)
36.732635
```

La distribución normal

Inferencia estadística

Supongamos una población de tamaño N=10.000, de personas adultas en que se les ha medido sus estaturas en metros, los parámetros poblacionales son:

```
. sum X,d

                          Estatura
-------------------------------------------------------------
      Percentiles      Smallest
 1%     119.1348       86.62138
 5%     131.7351        88.6692
10%     139.2241       91.11671       Obs                 10000
25%     151.5324       93.84512       Sum of Wgt.         10000

50%      165.226                      Mean                164.988
                        Largest       Std. Dev.          20.01317
75%     178.3917       230.2685
90%     190.5882       233.3427       Variance            400.527
95%      197.993       234.4337       Skewness          -.0088502
99%     211.7315       237.8318       Kurtosis           2.982042
```

μ

σ^2

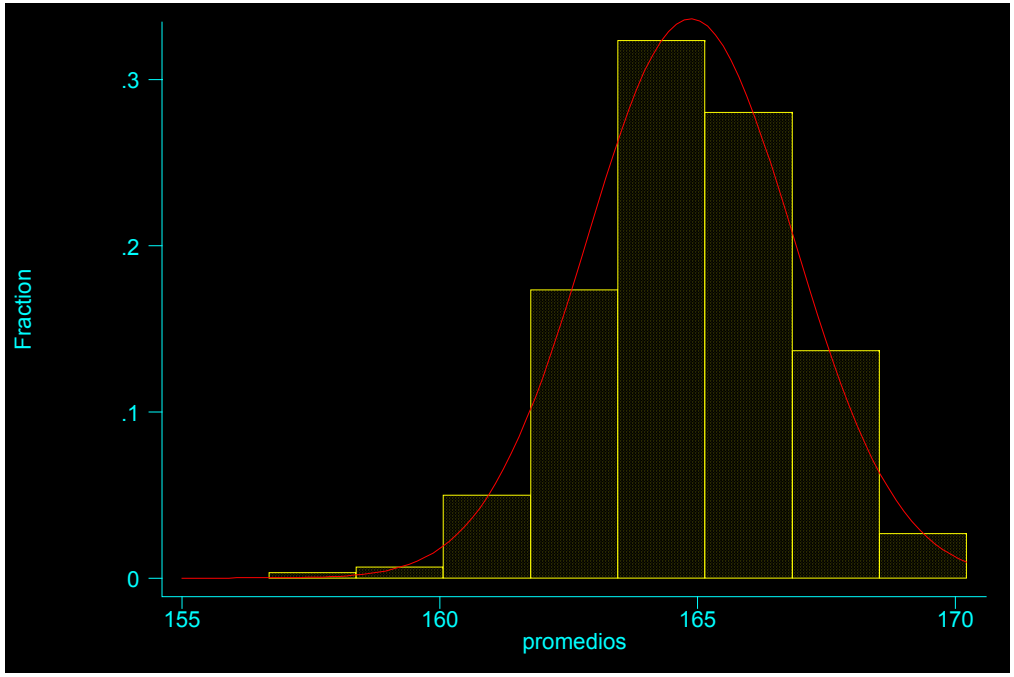
Inferencia estadística

El siguiente programa STATA, extraerá 300 muestras de tamaño 100 de la Población y en cada una de ellas se calculará el promedio y la desviación estándar:

```
* Programa TCL
set more off

          local iterate = 1
          while `iterate' <= 300   {

use "C:\Documents and Settings\Gabriel Cavada\Escritorio\LosAndes\TCL.dta", clear
sample 1
sum X
clear
local iterate =`iterate' + 1
          }
```

Inferencia estadística

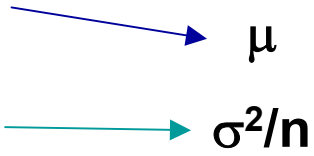
Al registrar el promedio en cada muestra tenemos "una muestra de promedios" es decir el promedio muestral es una variable aleatoria, con la siguiente distribución:



Inferencia estadística

Las estadísticas descriptivas de estos "promedios" son:

```
. sum  promedios,d

                          promedios
-------------------------------------------------------------
      Percentiles      Smallest
 1%     160.0712       157.0182
 5%     161.6588       159.0572
10%      162.374        159.708       Obs                 300
25%     163.5157       160.4345       Sum of Wgt.         300

50%     164.7495                      Mean           164.8657
                        Largest       Std. Dev.      2.004618
75%     166.2547       169.1602
90%     167.4832       169.2505       Variance       4.018492
95%     168.3237       169.7911       Skewness      -.1519181
99%     169.2054       170.2179       Kurtosis       3.343029
```



Hemos probado empíricamente que $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

Cuando n es muy grande

Inferencia estadística

Este resultado permite construir un intervalo de confianza para μ.

En la realidad no se dispone del valor de σ^2 y/o de muestras muy grandes, así entonces la distribución de probabilidades no es normal sino t-Student.

Basta tomar una muestra, de tamaño n, en la población para hacer inferencias acerca del promedio poblacional:

Inferencia estadística

```
. ci  X

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
           X |     100    163.7289    1.993126        159.7741    167.6837
```

Apostamos con 95% de certeza que el promedio poblacional está comprendido entre 157.8 y 167.7 metros

Inferencia estadística

Lo mismo ocurre con una proporción. En nuestra población de tamaño 10.000, la proporción de enfermos es:

```
. tab enfermo

    enfermo |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |       8014       80.14       80.14
          1 |       1986       19.86      100.00
------------+-----------------------------------
      Total |      10000      100.00
```

Inferencia estadística

Al extraer una muestra de tamaño 500, encontramos un intervalo de confianza para la prevalencia de:

```
. sample 5
(9500 observations deleted)

. ci enfermo,bin

                                                    -- Binomial Exact --
    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     enfermo |     500        .202    .0179553       .1676573    .2399116
```

Inferencia estadística

Ejemplo ilustrativo (diseño antes después): A 20 mujeres obesas se les registra el peso en Kgs. Luego se les somete a una dieta hipocalórica y al cabo de un mes son evaluadas. Los datos se muestran a continuación:

Inferencia estadística

. list

	id	pesoini~l	pesofinal
1.	1	92.9	74.8
2.	2	91.1	88
3.	3	86.5	82.4
4.	4	80.1	79.9
5.	5	84.3	92.8
6.	6	84.8	68.9
7.	7	96.3	71.6
8.	8	97.8	74.4
9.	9	90.6	85.3
10.	10	89.4	76.5
11.	11	80	78.6
12.	12	94.6	88.6
13.	13	92.9	85
14.	14	107.5	64.7
15.	15	83	80.4
16.	16	96.3	93.1
17.	17	94.8	84.8
18.	18	86.3	86.1
19.	19	75	86
20.	20	96.3	95.2

¿Es efectiva la dieta?

¿Qué tan efectiva es la dieta?

Inferencia estadística

```
. gen dif= pesofinal- pesoinicial
. list
```

	id	pesoini~l	pesofinal	dif
1.	1	92.9	74.8	-18.1
2.	2	91.1	88	-3.099998
3.	3	86.5	82.4	-4.099998
4.	4	80.1	79.9	-.1999969
5.	5	84.3	92.8	8.5
6.	6	84.8	68.9	-15.9
7.	7	96.3	71.6	-24.7
8.	8	97.8	74.4	-23.4
9.	9	90.6	85.3	-5.299995
10.	10	89.4	76.5	-12.9
11.	11	80	78.6	-1.400002
12.	12	94.6	88.6	-6
13.	13	92.9	85	-7.900002
14.	14	107.5	64.7	-42.8
15.	15	83	80.4	-2.599998
16.	16	96.3	93.1	-3.200005
17.	17	94.8	84.8	-10
18.	18	86.3	86.1	-.2000046
19.	19	75	86	11
20.	20	96.3	95.2	-1.100006

Inferencia estadística

```
. sum dif, d

                            dif
-------------------------------------------------------------
      Percentiles      Smallest
 1%        -42.8          -42.8
 5%       -33.75          -24.7
10%       -24.05          -23.4       Obs                  20
25%        -14.4          -18.1       Sum of Wgt.          20

50%    -4.699997                      Mean           -8.170001
                        Largest       Std. Dev.       12.28555
75%    -1.250004      -.2000046
90%     4.150002      -.1999969       Variance        150.9349
95%         9.75            8.5       Skewness       -1.084781
99%           11             11       Kurtosis        4.396793

. ci  dif

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
         dif |      20    -8.170001    2.747134      -13.91982   -2.420184
```

Inferencia estadística

Ejemplo ilustrativo (diseño antes después): A 50 hombres con dolor lumbar se les da un tratamiento anti inflamatorio, de ellos 38 mejoran. ¿En qué porcentaje es efectivo el tratamiento?

```
. list

      mejora
 1.        0
 2.        1
 3.        0
 4.        1
 5.        1

47.        0
48.        1
49.        0
50.        1
```

```
. tab mejora

     mejora |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         12       24.00       24.00
          1 |         38       76.00      100.00
------------+-----------------------------------
      Total |         50      100.00

. ci mejora,bin

                                                    -- Binomial Exact --
    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
      mejora |      50        .76    .0603987        .6183118    .8693945
```

Dócimas de hipótesis

- **Introducción**

  Hipótesis estadística es una afirmación respecto de una característica poblacional (forma de ella o valor de sus parámetros); esta sentencia puede ser “docimada” (probada) en base a una muestra aleatoria extraída de esa población.

Dócimas de hipótesis

En muchas ocasiones es necesario decidir entre una afirmación de la forma $\theta = \theta_0$ (Hipótesis nula) u otra que puede tomar las siguientes formas $\theta \neq \theta_0$, $\theta > \theta_0$ ó $\theta < \theta_0$ (Hipótesis alternativa). En símbolos:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

$$ó$$

$$H_1 : \theta > \theta_0$$

$$ó$$

$$H_1 : \theta < \theta_0$$

Dócimas de hipótesis

Necesitamos desarrollar un procedimiento que nos permita tomar una decisión acerca de H_0, como esta decisión es tomada en base a información muestral está sujeta a errores probables, debido a que no se sabe como es realmente la naturaleza y sólo tenemos una percepción de ella. Cruzando este efecto con la decisión tenemos:

Dócimas de hipótesis

		Estado de la naturaleza	
		H_o es Verdad	H_o es Falsa
Percepción de la naturaleza	Se rechaza H_o	Error tipo I	Decisión correcta
	No se rechaza H_o	Decisión correcta	Error tipo II

Dócimas de hipótesis

Deseamos que los errores no se cometan, pero como la decisión será tomada bajo incertidumbre, sólo podemos pedir que la probabilidad de cometerlos sea pequeña.

La filosofía para docimar consiste en suponer que H_o es verdadera, hasta encontrar evidencia muestral suficiente que permita decir lo contrario, si esta evidencia no existe no podemos dudar de la afirmación contenida en H_o. Así el error mas grave que se puede cometer es el Error tipo I, que es el que intentamos de controlar.

Dócimas de hipótesis

Llamamos:

α=P(Rechazar H_o | H_o es Verdad) , tamaño del Error tipo I

β=P(No rechazar H_o | H_o es Falsa) , tamaño del Error tipo II

nos interesa que α sea pequeño (generalmente 5% o menos).

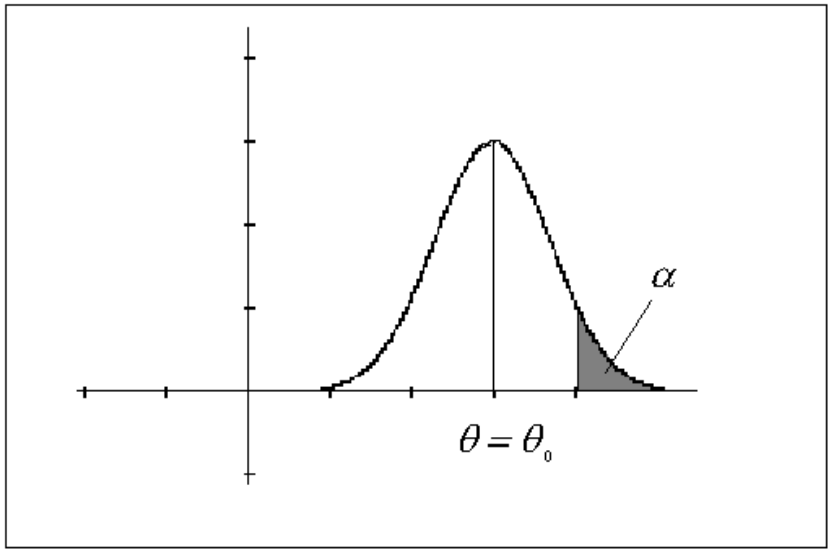
α se llama significación de la dócima y 1-β se llama potencia de la dócima, la potencia depende de la hipótesis alternativa que estemos proponiendo.

Dócimas de hipótesis

Se llama estadística de prueba, E, a una función que contenga el parámetro de interés (que se desea docimar) y toda la información muestral. Deseablemente la estadística de prueba, bajo la hipótesis nula, debe seguir una distribución de probabilidades conocida.

Dócimas de hipótesis

Se llama región crítica o de rechazo, aquella porción de los reales para la cual la probabilidad de que E esté en ella, considerando la veracidad de H_0, sea menor que α

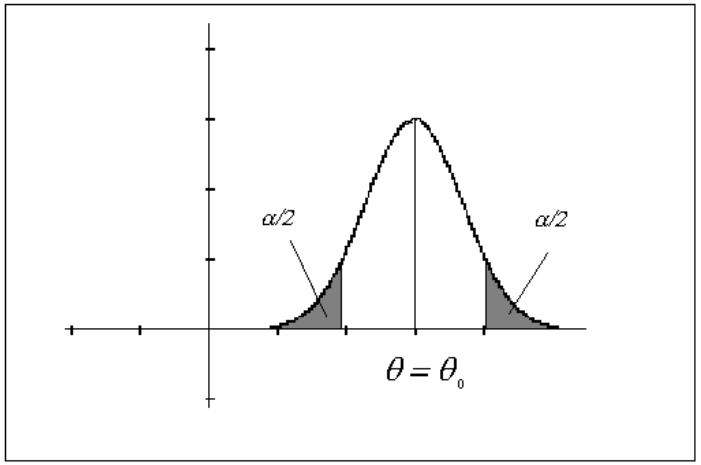


Dócimas de hipótesis

- **Una dócima de la forma:** $H_0 : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

se llama de "dos colas" pues la región de rechazo, se compone de dos porciones de los reales inconexas, que se muestran en el siguiente gráfico:



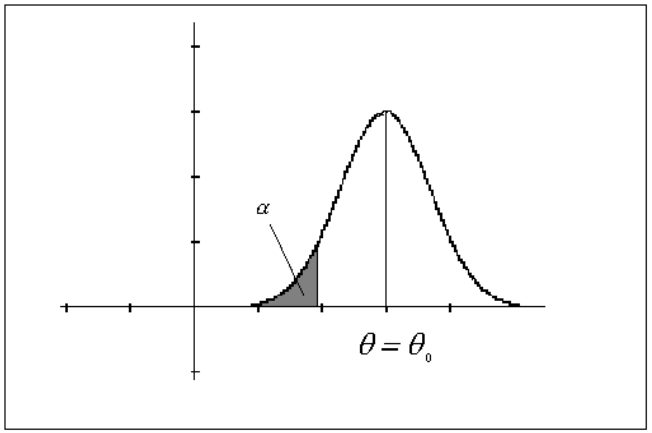
Dócimas de hipótesis

- **Una dócima de la forma:**

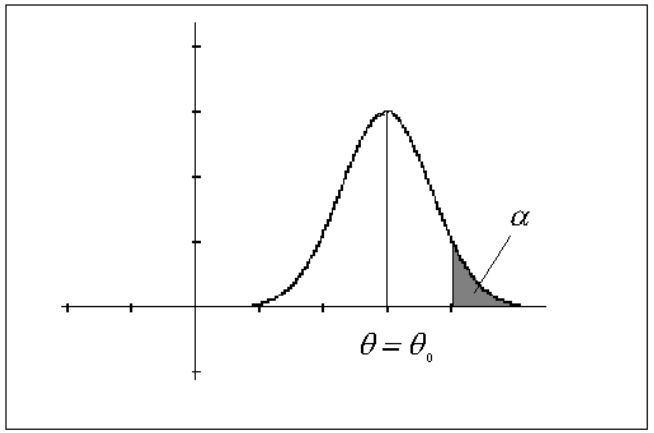
$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$
$$\acute{o}$$
$$H_1 : \theta < \theta_0$$

se llama de “una cola” pues la región de rechazo, se compone de una porción de los reales conexa, como se muestra a continuación:

Dócimas de hipótesis



$$H_1 : \theta < \theta_0$$

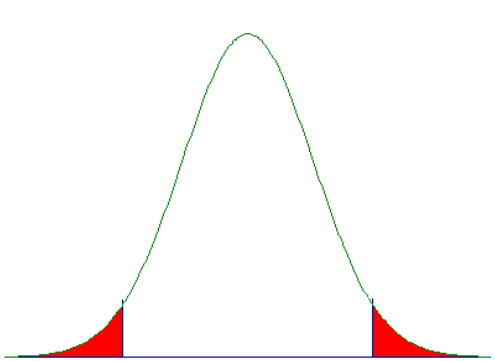


$$H_1 : \theta > \theta_0$$

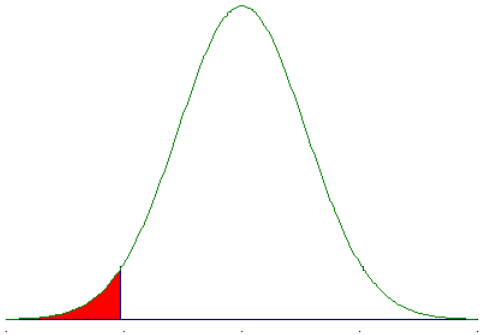
Dócimas de hipótesis

Como deseablemente la estadística de prueba, E, tiene una distribución de probabilidades conocida, se pueden calcular las siguientes probabilidades llamadas

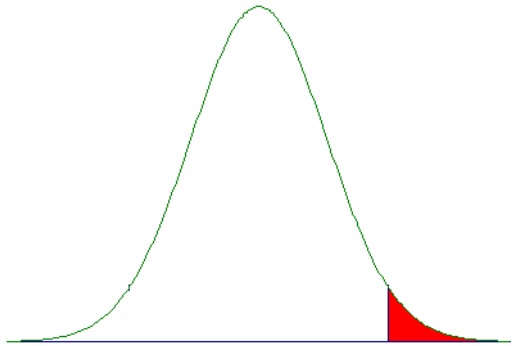
P-VALUES, el P-VALUE es el tamaño del Error I:



$P(E<-E_0 \text{ U } E>E_0)=\alpha$



$P(E<-E_0)=\alpha$



$P(E>E_0)=\alpha$

Dócimas respecto de promedios

El caso de una muestra y de dos muestras pareadas

Hipótesis Nula	Estadística de Prueba	Distribución de la estadística de prueba
$H_0 : \mu = \mu_0$	$\frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$	$t(n-1)$

Dócimas respecto de promedios

Ejemplo ilustrativo (diseño antes después): A 20 mujeres obesas se les registra el peso en Kgs. Luego se les somete a una dieta hipocalórica y al cabo de un mes son evaluadas. Los datos se muestran a continuación:

Dócimas respecto de promedios

```
. gen dif= pesofinal- pesoinicial
. list
```

	id	pesoini~l	pesofinal	dif
1.	1	92.9	74.8	-18.1
2.	2	91.1	88	-3.099998
3.	3	86.5	82.4	-4.099998
4.	4	80.1	79.9	-.1999969
5.	5	84.3	92.8	8.5
6.	6	84.8	68.9	-15.9
7.	7	96.3	71.6	-24.7
8.	8	97.8	74.4	-23.4
9.	9	90.6	85.3	-5.299995
10.	10	89.4	76.5	-12.9
11.	11	80	78.6	-1.400002
12.	12	94.6	88.6	-6
13.	13	92.9	85	-7.900002
14.	14	107.5	64.7	-42.8
15.	15	83	80.4	-2.599998
16.	16	96.3	93.1	-3.200005
17.	17	94.8	84.8	-10
18.	18	86.3	86.1	-.2000046
19.	19	75	86	11
20.	20	96.3	95.2	-1.100006

Dócimas respecto de promedios

```
. sum dif, d

                            dif
-------------------------------------------------------------
      Percentiles      Smallest
 1%        -42.8          -42.8
 5%       -33.75          -24.7
10%       -24.05          -23.4       Obs                  20
25%        -14.4          -18.1       Sum of Wgt.          20

50%    -4.699997                      Mean           -8.170001
                        Largest       Std. Dev.       12.28555
75%    -1.250004      -.2000046
90%     4.150002      -.1999969       Variance        150.9349
95%         9.75            8.5       Skewness       -1.084781
99%           11             11       Kurtosis        4.396793

. ci  dif

    Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+--------------------------------------------------------------
         dif |      20    -8.170001    2.747134      -13.91982   -2.420184
```

Dócimas respecto de promedios

Otras preguntas relevantes

¿Es posible afirmar que la dieta en promedio permite bajar 5 Kgs.?

```
. ttest  dif=-5

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     dif |      20   -8.170001    2.747134    12.28555   -13.91982   -2.420184
------------------------------------------------------------------------------
Degrees of freedom: 19

                     Ho: mean(dif) = -5

     Ha: mean < -5               Ha: mean ~= -5               Ha: mean > -5
       t =  -1.1539                t =  -1.1539                 t =  -1.1539
   P < t =   0.1314            P > |t| =   0.2628           P > t =   0.8686
```

Dócimas respecto de promedios

¿Es posible afirmar que la dieta en promedio permite bajar 15 Kgs.?

```
. ttest  dif=-15

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     dif |      20   -8.170001    2.747134    12.28555   -13.91982   -2.420184
------------------------------------------------------------------------------
Degrees of freedom: 19

                     Ho: mean(dif) = -15

    Ha: mean < -15               Ha: mean ~= -15               Ha: mean > -15
       t =   2.4862                  t =   2.4862                  t =   2.4862

   P < t =   0.9888           P > |t| =   0.0224           P > t =   0.0112
```

Dócimas respecto de promedios

La pregunta mas relevante:

¿La dieta en promedio es efectiva?

```
. ttest  dif=0

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     dif |      20   -8.170001    2.747134    12.28555   -13.91982   -2.420184
------------------------------------------------------------------------------
Degrees of freedom: 19

                     Ho: mean(dif) = 0

     Ha: mean < 0                 Ha: mean ~= 0                 Ha: mean > 0
       t =  -2.9740                 t =  -2.9740                 t =  -2.9740
   P < t =   0.0039         P > |t| =   0.0078         P > t =   0.9961
```

Dócimas respecto de promedios

Otra forma de verlo

¿La dieta en promedio es efectiva?

```
. ttest   pesoinicial= pesofinal

Paired t test

------------------------------------------------------------------------------
Variable     |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
-------------+------------------------------------------------------------------
pesoiniciall |      20      90.025    1.706472    7.631574    86.45331    93.59669
pesofinall   |      20      81.855    1.843459    8.244199     77.9966     85.7134
-------------+------------------------------------------------------------------
    diff     |      20    8.170001    2.747134    12.28555    2.420184    13.91982
------------------------------------------------------------------------------

             Ho: mean(pesoinicial - pesofinal) = mean(diff) = 0

  Ha: mean(diff) < 0         Ha: mean(diff) ~= 0         Ha: mean(diff) > 0
       t =   2.9740               t =   2.9740               t =   2.9740
   P < t =   0.9961           P > |t| =   0.0078           P > t =   0.0039
```

Dócimas respecto de promedios

El caso de dos muestras independientes

Hipótesis Nula	Estadística de Prueba	Distribución de la estadística de prueba
$H_0 : \mu_x - \mu_y = 0$	$\frac{\overline{X} - \overline{Y}}{S_c \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$ $S_c = \sqrt{\frac{(n_x - 1)S_x^2 + (n_y - 1)S_y^2}{n_x + n_y - 2}}$	$t(n_x + n_y - 2)$

Dócimas respecto de promedios

- Ejemplo: Se aleatorizan 40 niños afiebrados a causa de una gripe común, para probar la efectividad de dos antipiréticos A y B, se desea probar que el antipirético B es mejor que A. Los datos se muestran a continuación:

Dócimas respecto de promedios

```
. list
```

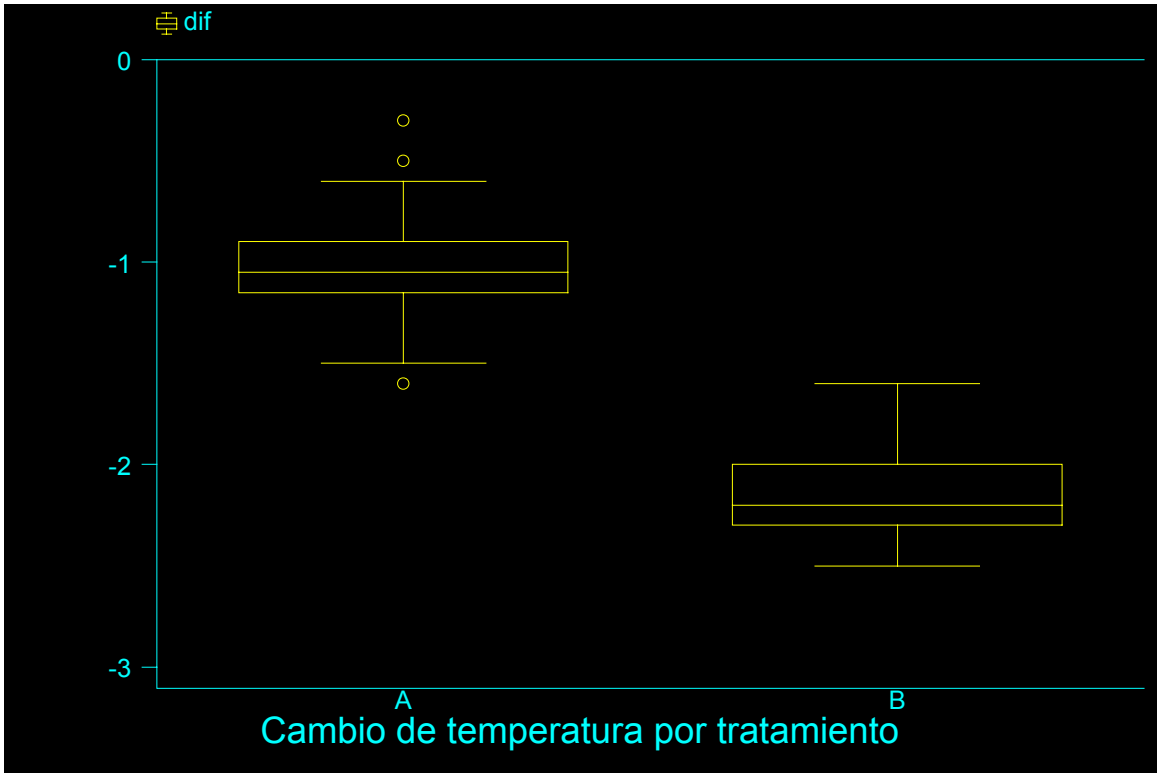
	id	tratami~o	tinicial	tfinal
1.	1	B	38.9	36.7
2.	2	A	39.3	38
3.	3	B	38.9	36.8
4.	4	A	39	37.7
5.	5	B	38.9	36.8
6.	6	A	39	37.9
7.	7	B	38.7	37.1
8.	8	B	39.2	36.7
9.	9	A	39.3	37.8
10.	10	A	38.9	37.9
11.	11	B	38.8	36.8
12.	12	A	39.2	38.1
13.	13	B	39	37
14.	14	A	38.8	38.2
15.	15	A	39.1	37.9
16.	16	A	39.4	37.8
17.	17	B	38.7	36.7
18.	18	A	39.1	38.1
19.	19	B	39.3	36.8
20.	20	B	39	36.6

	id	tratami~o	tinicial	tfinal
21.	21	B	39.1	37.1
22.	22	B	39.1	36.9
23.	23	A	39.1	38.1
24.	24	A	38.8	37.8
25.	25	A	38.6	38.1
26.	26	A	38.8	38
27.	27	A	38.9	38.1
28.	28	B	38.6	36.8
29.	29	A	39.1	38
30.	30	B	39.1	36.8
31.	31	B	39.1	36.8
32.	32	A	39.2	38.1
33.	33	B	39	36.8
34.	34	B	38.9	36.6
35.	35	A	38.5	38.2
36.	36	B	39.1	36.9
37.	37	A	39	37.9
38.	38	A	38.9	37.9
39.	39	B	38.9	36.7
40.	40	B	38.9	36.7

Dócimas respecto de promedios

```
. gen dif= tfinal- tinicial

. sort  tratamiento

. by  tratamiento: sum dif

-------------------------------------------------------------------------------
-> tratamiento = A

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         dif |      20       -1.02   .3122079  -1.600002  -.2999992

-------------------------------------------------------------------------------
-> tratamiento = B

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         dif |      20      -2.155   .2187883       -2.5  -1.600002
```

Dócimas respecto de promedios



Dócimas respecto de promedios

```
. ttest  dif, by( tratamiento)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       A |      20       -1.02    .0698118    .3122079   -1.166118   -.8738821
       B |      20      -2.155    .0489226    .2187883   -2.257396   -2.052604
---------+--------------------------------------------------------------------
combined |      40     -1.5875    .1001402     .633342   -1.790053   -1.384948
---------+--------------------------------------------------------------------
    diff |               1.135    .0852473                .9624262    1.307575
------------------------------------------------------------------------------
Degrees of freedom: 38

                    Ho: mean(A) - mean(B) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  13.3142                 t =  13.3142                 t =  13.3142
   P < t =   1.0000             P > |t| =   0.0000             P > t =   0.0000
```

Dócimas de proporciones

Dócima de una proporción en el caso de dos muestras.

Recordemos la base de datos practico1.dta, en la cual se dispone información de pacientes con una determinada enfermedad renal. Al recordar la variable sexo, podemos hacernos algunas preguntas:

Dócimas de proporciones

Por ejemplo:

- ¿La proporción poblacional de hombres es igual a la de mujeres?
- ¿Es esta una enfermedad que afecta en proporción 3:1 a hombres respecto de mujeres?

Dócimas de proporciones

Estas hipótesis pueden plantearse mediante la siguiente dócima:

Hipótesis Nula	**Estadística de Prueba**	Distribución de la estadística de prueba
$H_0 : P = P_0$	$\frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}}$	$N(0,1)$

Dócimas de proporciones

¿La proporción poblacional de hombres es igual a la de mujeres?

$$H_0 : P = 0.5$$

Donde P es la proporción poblacional de mujeres

Dócimas de proporciones

```
. tab sexo

   0:hombre |
    1:mujer |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        161       80.50       80.50
          1 |         39       19.50      100.00
------------+-----------------------------------
      Total |        200      100.00

. prtest  sexo=0.5

One-sample test of proportion                  sexo: Number of obs =      200

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    sexo |      .195    .0280156    6.9604   0.0000       .1400904    .2499096
------------------------------------------------------------------------------

                       Ho: proportion(sexo) = .5

     Ha: sexo < .5           Ha: sexo ~= .5           Ha: sexo > .5
          z = -8.627              z = -8.627               z = -8.627
      P < z = 0.0000          P > |z| = 0.0000          P > z = 1.0000
```

Dócimas de proporciones

¿Es esta una enfermedad que afecta en proporción 3:1 a hombres respecto de mujeres?

$$H_0 : P = 0.25$$

Donde P es la proporción poblacional de mujeres

Dócimas de proporciones

```
. prtest  sexo=0.25

One-sample test of proportion                  sexo: Number of obs =      200

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    sexo |      .195    .0280156    6.9604   0.0000       .1400904    .2499096
------------------------------------------------------------------------------

                     Ho: proportion(sexo) = .25

     Ha: sexo < .25          Ha: sexo ~= .25          Ha: sexo > .25
          z = -1.796               z = -1.796               z = -1.796
      P < z = 0.0362          P > |z| = 0.0724          P > z = 0.9638
```

Dócimas de proporciones

Para comparar proporciones en dos muestras independientes, usamos:

Hipótesis Nula	**Estadística de Prueba**	Distribución de la estadística de prueba
$H_0 : P_x - P_y = 0$	$\frac{p_x - p_y}{\sqrt{PQ(\frac{1}{n_x} + \frac{1}{n_y})}}$	$N(0,1)$

$$P = \frac{n_x p_x + n_y p_y}{n_x + n_y}$$

Dócimas de proporciones

Ejemplo: Se aleatorizan 60 pacientes en dos tratamientos (0 y 1) y se registra la condición de mejoría (0: no mejora 1: mejora)

pac	trat	mejora	pac	trat	mejora	pac	trat	mejora
1	0	1	21	0	1	41	0	1
2	0	1	22	1	1	42	0	1
3	1	1	23	0	1	43	0	1
4	1	1	24	0	1	44	1	1
5	1	1	25	1	1	45	0	0
6	0	0	26	0	1	46	1	1
7	1	1	27	1	1	47	0	0
8	1	1	28	1	1	48	1	1
9	0	1	29	0	0	49	1	0
10	0	0	30	1	1	50	0	0
11	1	1	31	0	1	51	0	1
12	0	1	32	1	1	52	0	0
13	0	1	33	0	1	53	0	1
14	1	1	34	1	1	54	0	0
15	1	1	35	1	1	55	1	1
16	1	1	36	0	0	56	1	1
17	1	1	37	1	1	57	0	0
18	1	0	38	0	0	58	1	1
19	0	1	39	0	0	59	1	1
20	1	1	40	0	1	60	1	1

Dócimas de proporciones

```
. by trat: tab mejora

-------------------------------------------------------------------------------
-> trat = 0

     mejora |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         12       40.00       40.00
          1 |         18       60.00      100.00
------------+-----------------------------------
      Total |         30      100.00

-------------------------------------------------------------------------------
-> trat = 1

     mejora |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          2        6.67        6.67
          1 |         28       93.33      100.00
------------+-----------------------------------
      Total |         30      100.00
```

Dócimas de proporciones

```
. prtest  mejora, by(trat)

Two-sample test of proportion                     0: Number of obs =       30
                                                  1: Number of obs =       30

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |        .6    .0894427    6.7082   0.0000       .4246955    .7753045
       1 |  .9333333     .045542   20.4939   0.0000       .8440726    1.022594
---------+--------------------------------------------------------------------
    diff | -.3333333    .1003697                         -.5300543   -.1366124
         |  under Ho:   .1092059  -3.05234   0.0023
------------------------------------------------------------------------------

               Ho: proportion(0) - proportion(1) = diff = 0

       Ha: diff < 0            Ha: diff ~= 0             Ha: diff > 0
        z = -3.052               z = -3.052               z = -3.052
    P < z =  0.0011        P > |z| =  0.0023          P > z =  0.9989
```

Dócima de independencia entre dos variables nominales

Interesa averiguar si dos variables cualitativas X e Y están vinculadas. En cada unidad de observación se registra un par (x,y) de valores observados, en consecuencia a partir de lo obtenido en **n** unidades de observación, se obtiene una Tabla de Contingencia de s×r (tabla observada):

	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{...X_r}$	Total
$\mathbf{Y_1}$	O_{11}	O_{12}	O_{1r}	$\mathbf{n_{1.}}$
$\mathbf{Y_2}$	O_{21}	O_{22}	O_{2r}	$\mathbf{n_{2.}}$
$\mathbf{...Y_s}$	O_{s1}	O_{s2}	O_{sr}	$\mathbf{n_{s.}}$
Total	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	$\mathbf{n_{.r}}$	**n**

Dócima de independencia entre dos variables nominales

Bajo la Hipótesis de independencia, estas frecuencias se pueden recalcular, creándose una Tabla Esperada:

	X_1	X_2	$...X_r$	Total
$\mathbf{Y}_1$	E_{11}	E_{12}	E_{1r}	
Y_2	E_{21}	E_{22}	E_{2r}	
$...Y_s$	E_{s1}	E_{s2}	E_{sr}	
Total				n

Dócima de independencia entre dos variables nominales

Donde :

$$E_{ij} = \frac{n_{.j} \cdot n_{i.}}{n}$$

En estas condiciones, podemos plantear la Hipótesis Nula:

H_0 : X es independiente de Y

Contrastada con la Hipótesis alternativa:

H_1 : X está asociada con Y

Dócima de independencia entre dos variables nominales

La estadística de prueba es:

$$\sum_{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

que sigue una distribución

$$\chi^2_{((s-1)\cdot(r-1))}$$

¡ Siempre es una dócima de una cola !

Dócima de independencia entre dos variables nominales

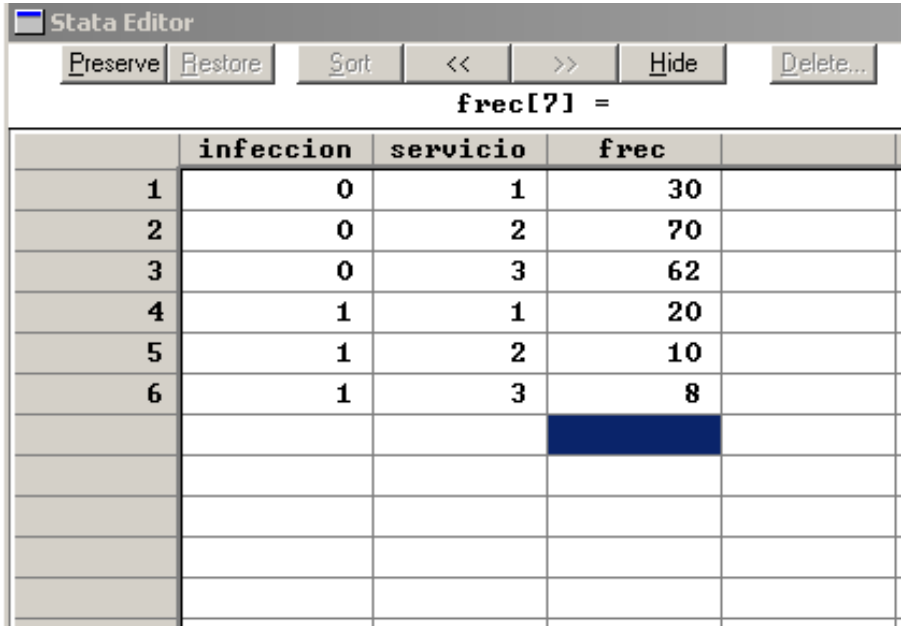
Ejemplo: Se cree que la cantidad de casos con cierta infección intra hospitalaria está asociada al servicio hospitalario, para probar dicha hipótesis se dispone de la siguiente información:

	Cirujía (1)	Medicina (2)	Urgencia (3)	
sin infección (0)	30	70	62	162
con infección(1)	20	10	8	38
	50	80	70	200

H_0: La condición de infectado es independiente del servicio

H_1: La condición de infectado está asociado al servicio

Dócima de independencia entre dos variables nominales



	infeccion	servicio	frec
1	0	1	30
2	0	2	70
3	0	3	62
4	1	1	20
5	1	2	10
6	1	3	8

Dócima de independencia entre dos variables nominales

```
. tab  infeccion servicio [freq=frec], chi2

           |             servicio
 infeccion |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |        30         70         62 |       162
         1 |        20         10          8 |        38
-----------+---------------------------------+----------
     Total |        50         80         70 |       200

          Pearson chi2(2) =  19.1312   Pr = 0.000
```

Dócima de independencia entre dos variables nominales

```
. tab  infeccion servicio [freq=frec], chi2 col

           |             servicio
 infeccion |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |        30         70         62 |       162
           |     60.00      87.50      88.57 |     81.00
-----------+---------------------------------+----------
         1 |        20         10          8 |        38
           |     40.00      12.50      11.43 |     19.00
-----------+---------------------------------+----------
     Total |        50         80         70 |       200
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =  19.1312   Pr = 0.000
```

Dócima de independencia entre dos variables nominales

¿Cuáles de estas proporciones difieren?

```
. prtesti 50 0.4 80 0.125

Two-sample test of proportion                     x: Number of obs =       50
                                                  y: Number of obs =       80

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |        .4     .069282    5.7735   0.0000       .2642097    .5357903
       y |      .125    .0369755   3.38062   0.0007       .0525294    .1974706
---------+--------------------------------------------------------------------
    diff |      .275    .0785314                           .1210812    .4289188
         |  under Ho:   .0759555   3.62054   0.0003
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0           Ha: diff ~= 0             Ha: diff > 0
        z =  3.621              z =  3.621               z =  3.621
    P < z =  0.9999       P > |z| =  0.0003         P > z =  0.0001
```

Dócima de independencia entre dos variables nominales

¿Cuáles de estas proporciones difieren?

```
. prtesti 50 0.4 70 0.1143

Two-sample test of proportion                     x: Number of obs =       50
                                                  y: Number of obs =       70

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |        .4     .069282    5.7735   0.0000       .2642097    .5357903
       y |     .1143    .0380292   3.00558   0.0027       .0397641    .1888359
---------+--------------------------------------------------------------------
    diff |     .2857     .079033                          .1307981    .4406019
         |  under Ho:   .0783166   3.64801   0.0003
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

      Ha: diff < 0            Ha: diff ~= 0             Ha: diff > 0
       z =  3.648               z =  3.648               z =  3.648
   P < z =  0.9999       P > |z| =  0.0003         P > z =  0.0001
```

Dócima de independencia entre dos variables nominales

¿Cuáles de estas proporciones difieren?

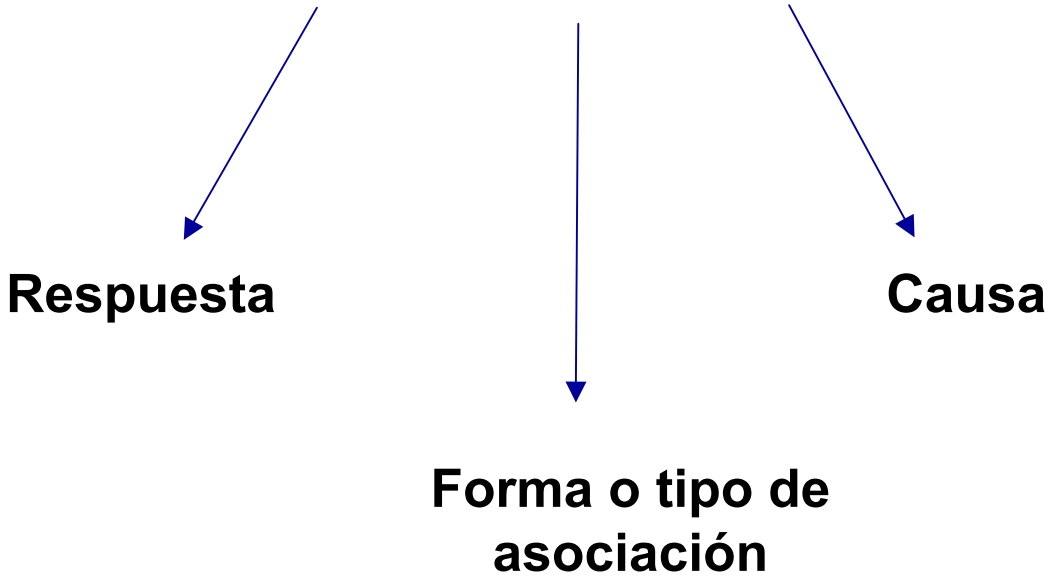
```
. prtesti 80 0.125 70 0.1143

Two-sample test of proportion                     x: Number of obs =       80
                                                  y: Number of obs =       70

------------------------------------------------------------------------------
Variable |      Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      .125    .0369755   3.38062   0.0007       .0525294    .1974706
       y |     .1143    .0380292   3.00558   0.0027       .0397641    .1888359
---------+--------------------------------------------------------------------
    diff |     .0107    .0530416                         -.0932596    .1146596
         |  under Ho:   .0531856   .201182   0.8406
------------------------------------------------------------------------------

               Ho: proportion(x) - proportion(y) = diff = 0

       Ha: diff < 0            Ha: diff ~= 0             Ha: diff > 0
        z =  0.201               z =  0.201                z =  0.201
    P < z =  0.5797       P > |z| =  0.8406          P > z =  0.4203
```

Regresión lineal simple

Como ya hemos visto cuando se observa una causa, buscamos la o las causas que lo produjeron. Al simplificar esta estructura cognoscitiva podemos pensar que una respuesta es generada por una causa; lo que podemos representar, cuando causa y efecto son medibles numéricamente, por una relación funcional:

Regresión lineal simple

$$Y = f(X)$$



Regresión lineal simple

Particularmente, nos interesa modelar la respuesta cuando la relación funcional entre la respuesta y la causa es lineal, es decir, de la forma:

$$Y = \alpha + \beta \cdot X$$

Regresión lineal simple

Obviamente antes de ajustar un modelo como el propuesto es necesario saber si la variable respuesta se asocia linealmente con la variable independiente, cuando ambas se miden en n unidades de análisis, esto es, cuando se tiene una muestra de la forma:

Observación	X	Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
...	...	...
...	...	...
...	...	...
n	x_n	y_n

Regresión lineal simple

Para ello, definimos el Coeficiente de Correlación entre X e Y como:

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

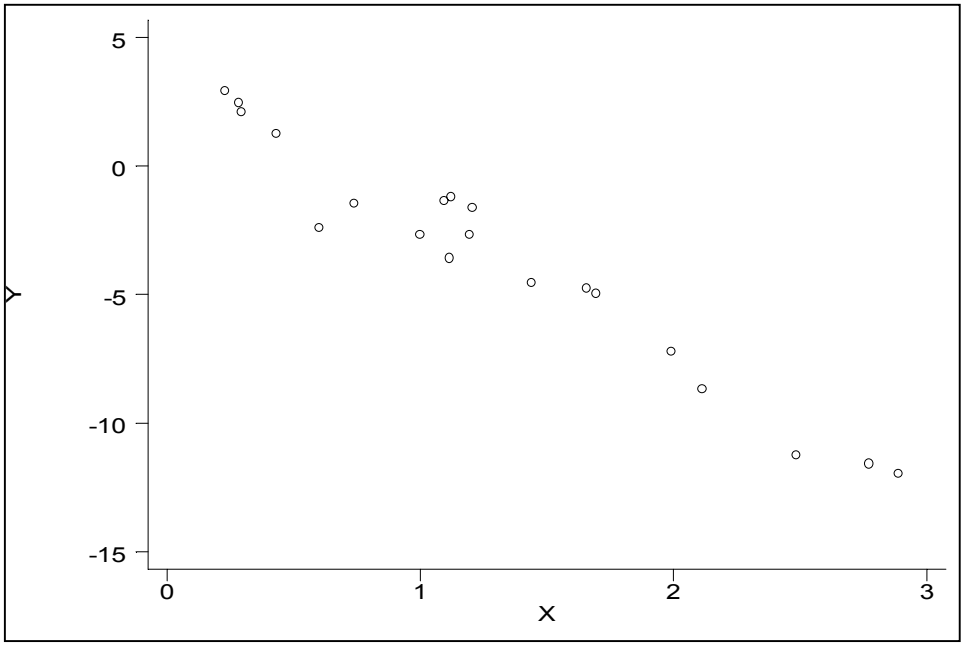
r_{xy} *mide el grado de asociación lineal entre X e Y*, puede demostrarse que:

$$-1 \leq r_{xy} \leq 1$$

Regresión lineal simple

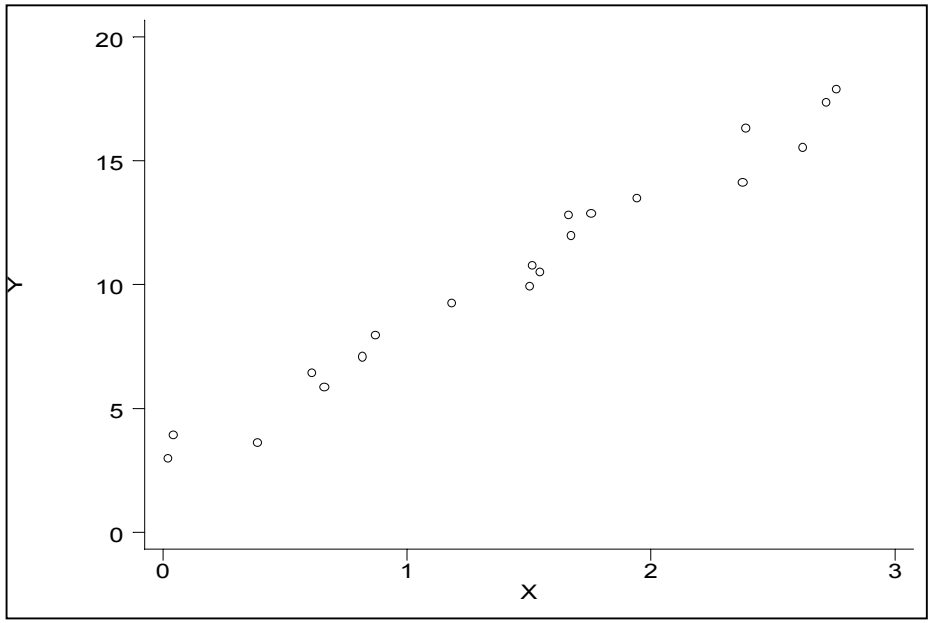
- r_{xy} tiende a 1 la asociación es directa
- r_{xy} tiende a -1 la asociación es inversa
- r_{xy} tiende a 0 no existe asociación lineal

Regresión lineal simple



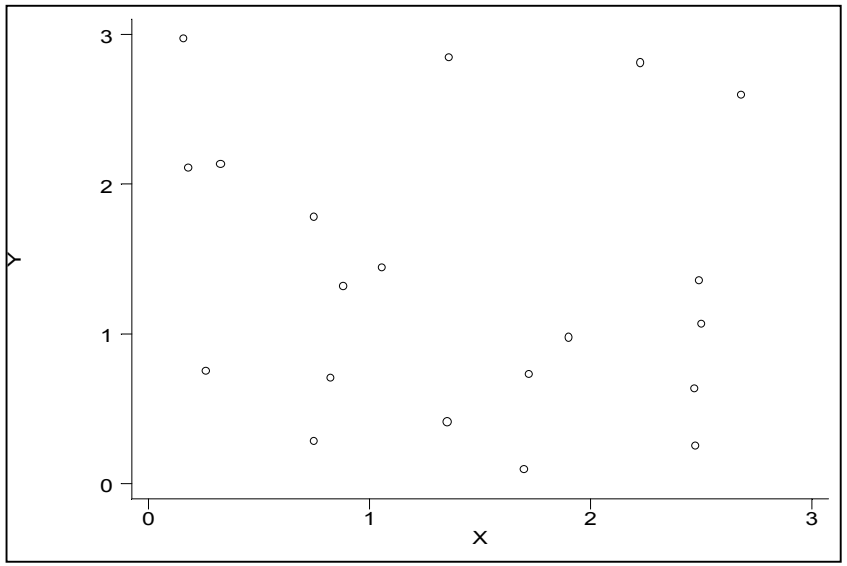
$$r_{xy} \rightarrow -1$$

Regresión lineal simple



$$r_{xy} \rightarrow 1$$

Regresión lineal simple



$$r_{xy} \rightarrow 0$$

Regresión lineal simple

Para ajustar un modelo de la forma : $Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$

consideramos la función: $\varepsilon^2(\alpha, \beta) = \sum \left(Y_i - \alpha - \beta \cdot X_i\right)^2$

El procedimiento consiste en encontrar los valores de α y β que hagan mínima la función:

$$\varepsilon^2(\alpha, \beta)$$

Estos valores se llaman estimadores mínimo cuadráticos y los denotamos por:

$$\hat{\alpha} \quad y \quad \hat{\beta}$$

Regresión lineal simple

Interpretación de β (pendiente de la recta):

Como:

$$Y(X) = \alpha + \beta \cdot X$$

Se tiene:

$$Y(X+1) = \alpha + \beta \cdot (X+1) = \alpha + \beta \cdot X + \beta$$

Luego:

$$Y(X+1) - Y(X) = \beta$$

β Representa el cambio de Y por unidad de X

Regresión lineal simple

Mediante cálculo diferencial bivariado se encuentra:

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$\hat{\alpha} = \overline{Y} - b \cdot \overline{X}$$

También puede demostrarse que:

$$\hat{\beta} = \frac{S_y}{S_x} r_{xy}$$

Regresión lineal simple

En consecuencia, ajustado el modelo, se tiene la siguiente tabla:

Observación	X	Y Valor observado de Y	$\hat{Y} = a + b \cdot X$ Valor estimado de Y	$\varepsilon = Y - \hat{Y}$ Residuo o Error
1	x_1	y_1	$\hat{Y}_1$	ε_1
2	x_2	y_2	$\hat{Y}_2$	ε_2
3	x_3	y_3	$\hat{Y}_3$	ε_3
4	x_4	y_4	$\hat{Y}_4$	ε_4
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
n	x_n	y_n	$\hat{Y}_n$	ε_n

Regresión lineal simple

Si deseamos hacer inferencias relativas al modelo ajustado, es necesario agregar los siguientes supuestos:

- $\varepsilon_i \sim N(0, \sigma^2)$
- Los X_i son independientes entre si, por lo tanto los ε_i también son independientes entre si (no correlacionados).

Regresión lineal simple

Una vez ajustado un modelo de regresión, es necesario conocer la calidad del mismo, para ello la variabilidad total de Y, que no depende del modelo ajustado, puede descomponerse del siguiente modo:

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

SCTotal = SCError + SCRegresión

Varianza Total = Varianza no explicada + Varianza explicada

Regresión lineal simple

Notar que de la identidad algebraica:

SCTotal = SCError + SCRegresión

Podemos escribir:

$$1 = \frac{SC\,\mathrm{Re}\,g}{SCTotal} + \frac{SCError}{SCTotal}$$

Se define el coeficiente de determinación como:

$$R^2 = \frac{\sum\left(\hat{Y} - \overline{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

que en el caso de la regresión lineal simple coincide con r^2_{xy}.

Regresión lineal simple

La descomposición de la variabilidad es posible resumirla en la conocida Tabla de Análisis de la Varianza (Tabla ANOVA)

Fuente de Variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
Regresión	1	$\sum(\hat{Y}-\overline{Y})^2$	$CMreg=\frac{\sum(\hat{Y}-\overline{Y})^2}{1}$	$F=\frac{CMreg}{CMres}$
Residuo	n-2	$\sum(Y-\hat{Y})^2$	$CMres=\frac{\sum(Y-\hat{Y})^2}{n-2}$	
Total	n-1	$\sum(Y-\overline{Y})^2$		

Regresión lineal simple

Asociada a la descomposición de la variabilidad y por ende a la calidad del modelo, se tiene la siguiente dócima:

$$H_0 : \Re^2 = 0$$

$$H_1 : \Re^2 > 0$$

Cuya estadística de prueba es:

$$F = \frac{CMreg}{CMres} \quad \sim F(1, n-2)$$

Regresión lineal simple

La estimación de la varianza del error es: $S^2 = CMres = \frac{\sum (Y - \hat{Y})^2}{n-2}$

Dócimas e intervalos de confianza:

	Estadística de prueba		Intervalo de confianza
$H_0 : \alpha = \alpha_0$	$\frac{a - \alpha_0}{S\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2}}}$	~t(n-2)	$a \pm t_{(n-2)} S \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum (X - \overline{X})^2}}$
$H_0 : \beta = \beta_0$	$\frac{b - \beta_0}{S} \sqrt{\sum (X - \overline{X})^2}$	~t(n-2)	$b \pm t_{(n-2)} \frac{S}{\sqrt{\sum (X - \overline{X})^2}}$

Intervalo de confianza para la predicción: $\hat{Y}_0 \pm t_{(n-2)} S \sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X - \overline{X})^2}}$

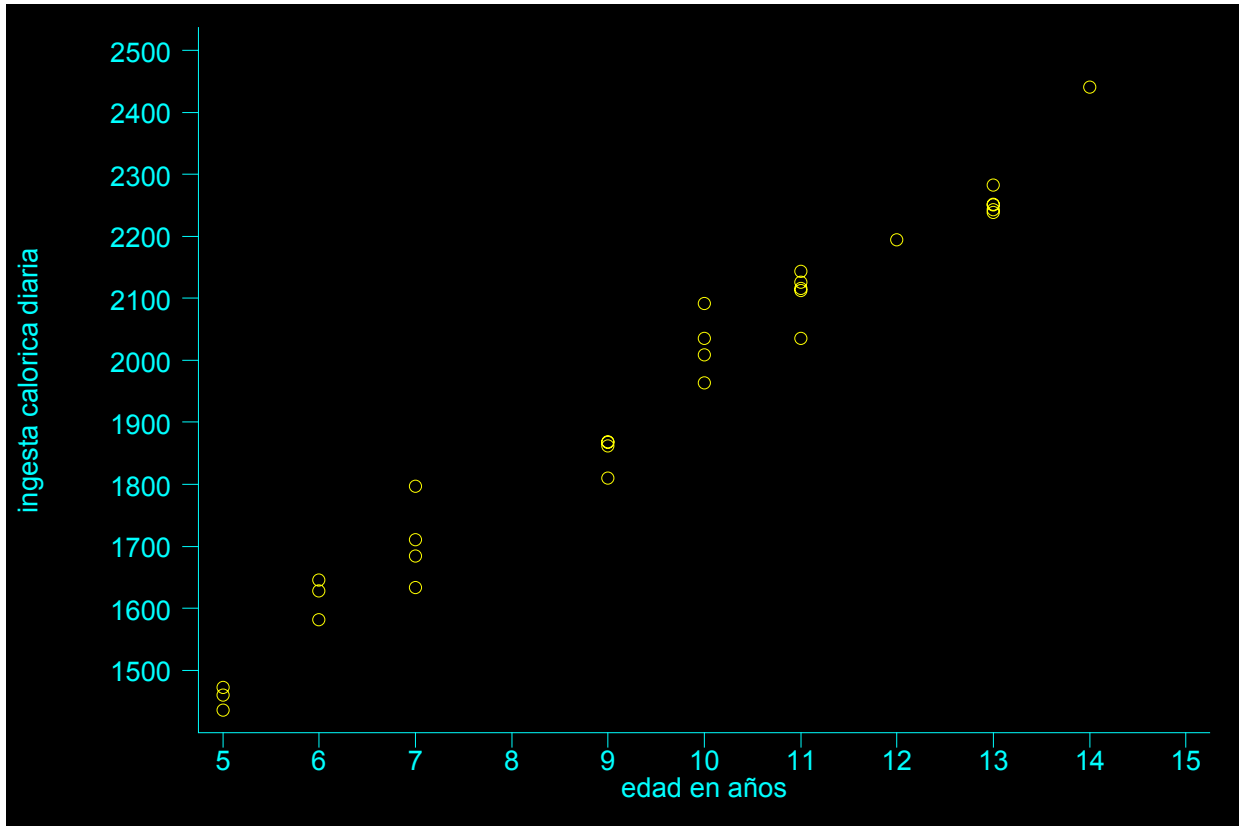
Regresión lineal simple

Ejemplo: Un nutriólogo, desea probar la hipótesis que afirma que la ingesta calórica diaria en niños varones no obesos entre los 5 y 15 años de edad aumenta con esta, para probar dicha hipótesis dispone de la siguiente información:

id	edad	cal	id	edad	cal
1	6	1628	16	13	2283
2	11	2126	17	7	1684
3	10	1963	18	10	2092
4	11	2035	19	7	1710
5	11	2112	20	14	2441
6	6	1581	21	7	1633
7	11	2143	22	6	1645
8	5	1436	23	9	1868
9	10	2009	24	12	2194
10	13	2238	25	13	2243
11	7	1797	26	9	1862
12	5	1460	27	9	1810
13	9	1867	28	13	2252
14	13	2251	29	5	1472
15	10	2035	30	11	2116

Regresión lineal simple

```
. graph7  cal edad, xlabel(5,6 to 15) ylabel(1500,1600 to 2500)
```




```
. corr  cal edad
(obs=30)

             |      cal     edad
-------------+------------------
         cal |   1.0000
        edad |   0.9858   1.0000
```

Regresión lineal simple

```
. reg  cal edad

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =  961.65
       Model |  2208153.27     1  2208153.27           Prob > F      =  0.0000
    Residual |  64294.1928    28  2296.22117           R-squared     =  0.9717
-------------+------------------------------           Adj R-squared =  0.9707
       Total |  2272447.47    29  78360.2575           Root MSE      =  47.919

------------------------------------------------------------------------------
         cal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edad |   99.42725   3.206252    31.01   0.000     92.85954     105.995
       _cons |   994.9363   31.48555    31.60   0.000     930.4411    1059.432
------------------------------------------------------------------------------
```

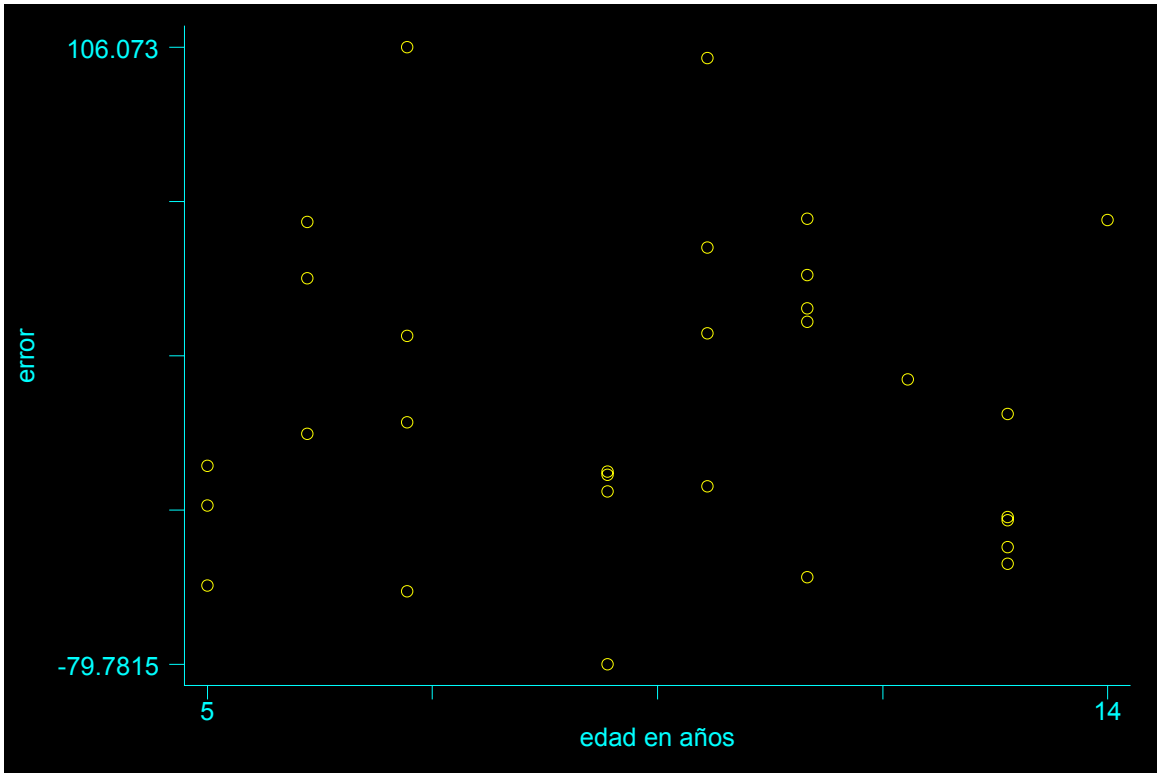
$$\text{Calorias} = 994.9363 + 99.42725 \cdot \text{Edad}$$

Regresión lineal simple

Prueba de los supuestos del modelo:

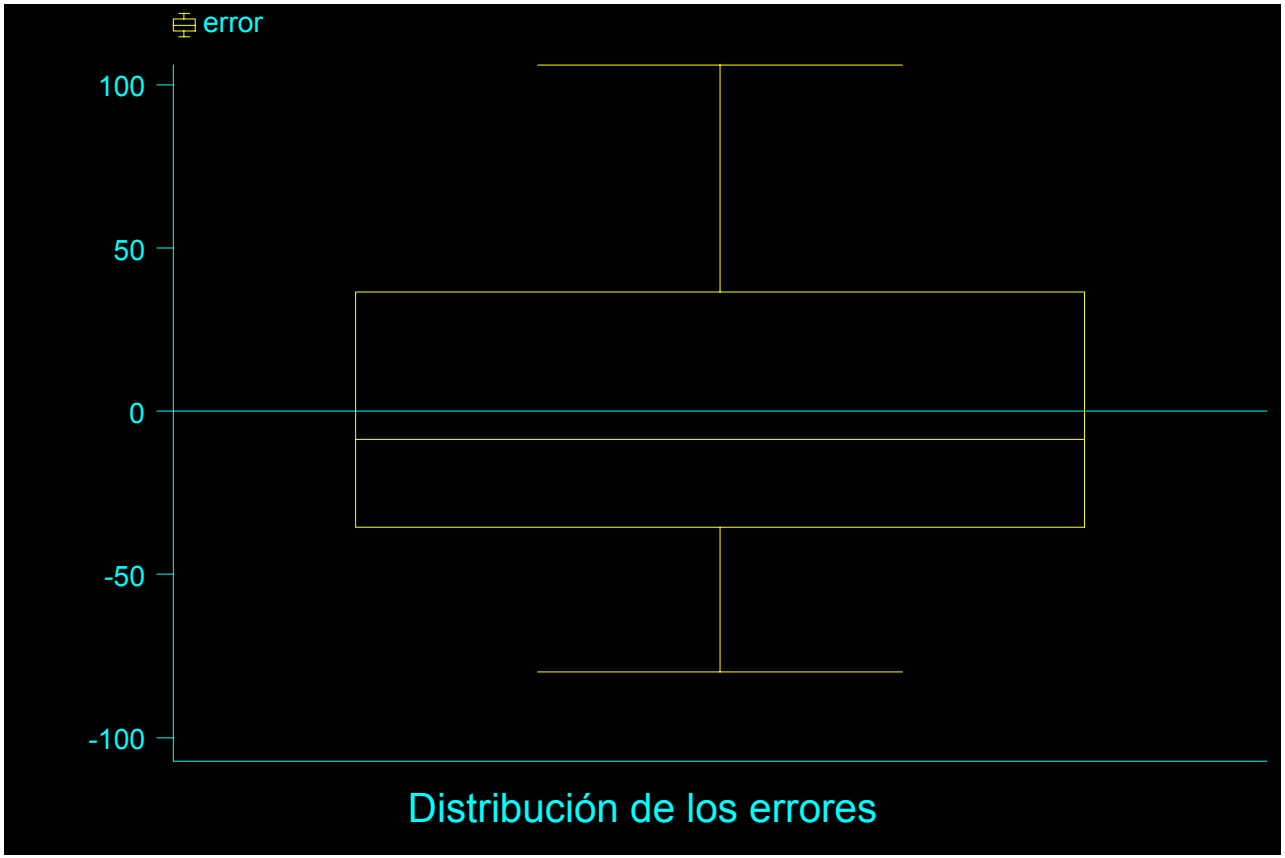
```
. predict calhat
(option xb assumed; fitted values)
. gen error= cal- calhat
. swilk  error
                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z     Prob>z
-------------+-------------------------------------------------
       error |     30    0.95642      1.385      0.674  0.25031
```

Regresión lineal simple



Si al graficar los errores versus la variable independiente (edad) no se encuentra un patrón de comportamiento los errores no están correlacionados

Regresión lineal simple



Regresión lineal simple

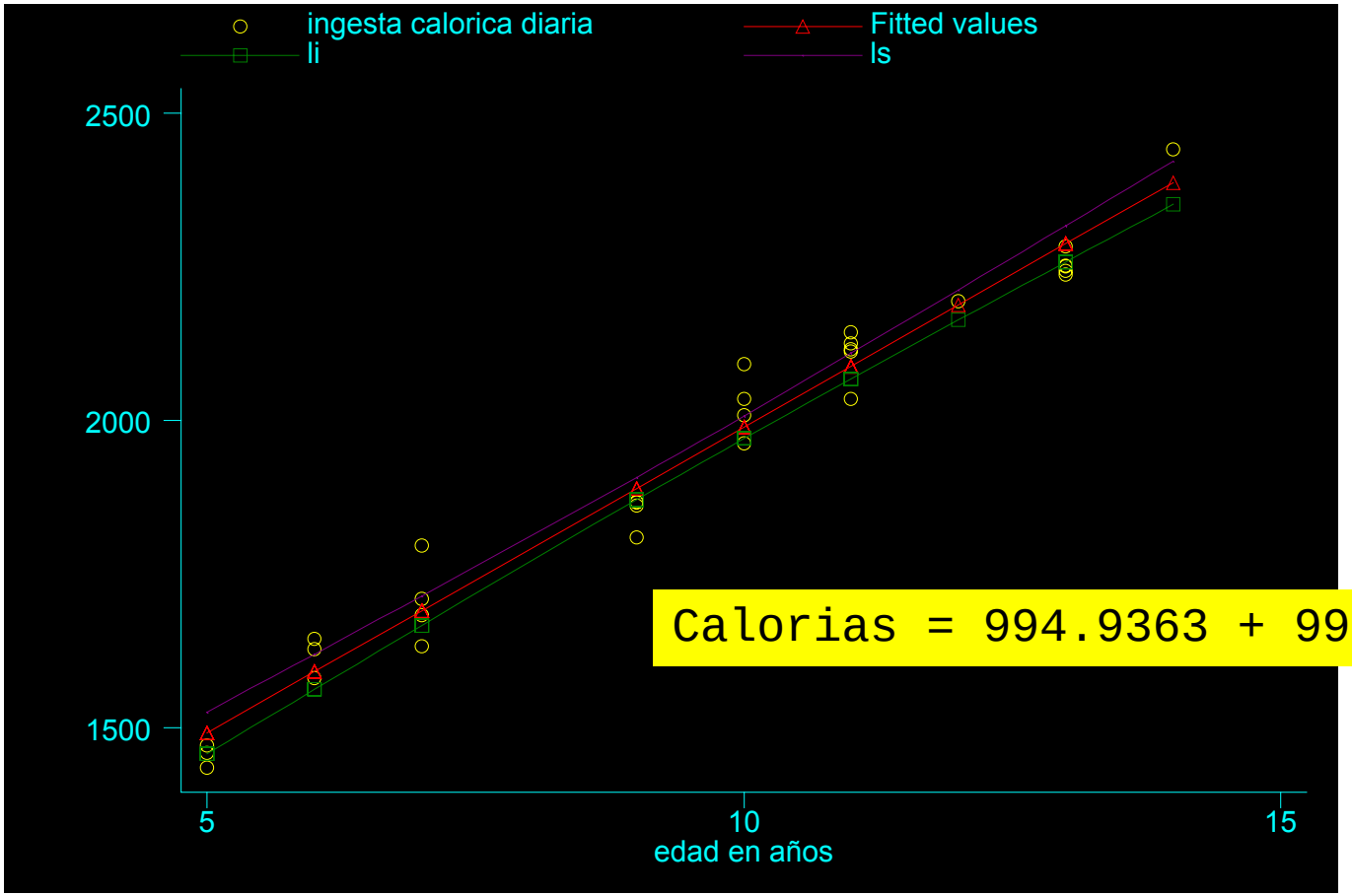
Gráfico de las observaciones, los valores predichos y sus intervalos de confianza:

```
. dis invttail(28, 0.025)
2.0484071
. predict es, stdp
. gen li= calhat-2.0484071* es
. gen ls= calhat+2.0484071* es
```

id	edad	cal	calhat	error	es	li	ls
1	6	1628	1591.5	36.50024	14.06128	1562.697	1620.303
2	11	2126	2088.636	37.36401	10.08824	2067.971	2109.301
3	10	1963	1989.209	-26.20874	8.935421	1970.905	2007.512
4	11	2035	2088.636	-53.63599	10.08824	2067.971	2109.301
5	11	2112	2088.636	23.36401	10.08824	2067.971	2109.301
6	6	1581	1591.5	-10.49976	14.06128	1562.697	1620.303
7	11	2143	2088.636	54.36401	10.08824	2067.971	2109.301
8	5	1436	1492.073	-56.07251	16.691	1457.883	1526.262
9	10	2009	1989.209	19.79126	8.935421	1970.905	2007.512
10	13	2238	2287.49	-49.49048	14.39842	2257.997	2316.984
11	7	1797	1690.927	106.073	11.7222	1666.915	1714.939
12	5	1460	1492.073	-32.07251	16.691	1457.883	1526.262

Regresión lineal simple

```
graph7   cal calhat  li ls edad, sort c(.lss)    xlabel ylabel
```



$$\text{Calorias} = 994.9363 + 99.42725 \cdot \text{Edad}$$


¡Por fin: el FIN!