TEMA 0. Introducción

¿Qué es la Estadística?

Estadística es la ciencia de*:*

- Recolectar
- Describir
- Organizar
- Interpretar

Datos

con el fin de transformar dichos datos en información y conseguir una toma de decisiones más eficiente.

¿Para qué sirve la estadística?

- La Ciencia se ocupa en general de fenómenos observables
- La Ciencia se desarrolla observando hechos, formulando leyes que los explican y realizando experimentos para validar o rechazar dichas leyes
- Los modelos que crea la ciencia pueden ser o de tipo determinista o de tipo **aleatorio (estocástico)**
- La **Estadística** se utiliza como **tecnología al servicio** de las ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza.

Definición

La Estadística es la Ciencia de la

Descriptiva

- **Sistematización recogida, ordenación y presentación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de**

Probabilidad

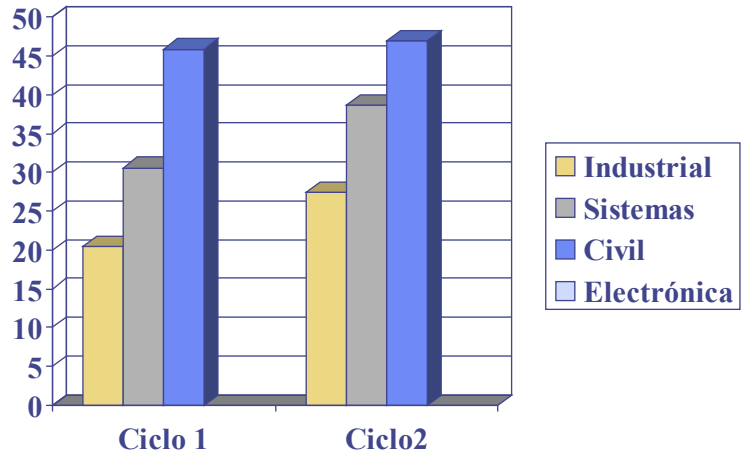
- **deducir las leyes que rigen esos fenómenos,**

Inferencia

- **y poder de esa forma hacer predicciones sobre los mismos, tomar decisiones u obtener conclusiones.**

SUBDIVISIONES DE LA ESTADÍSTICA

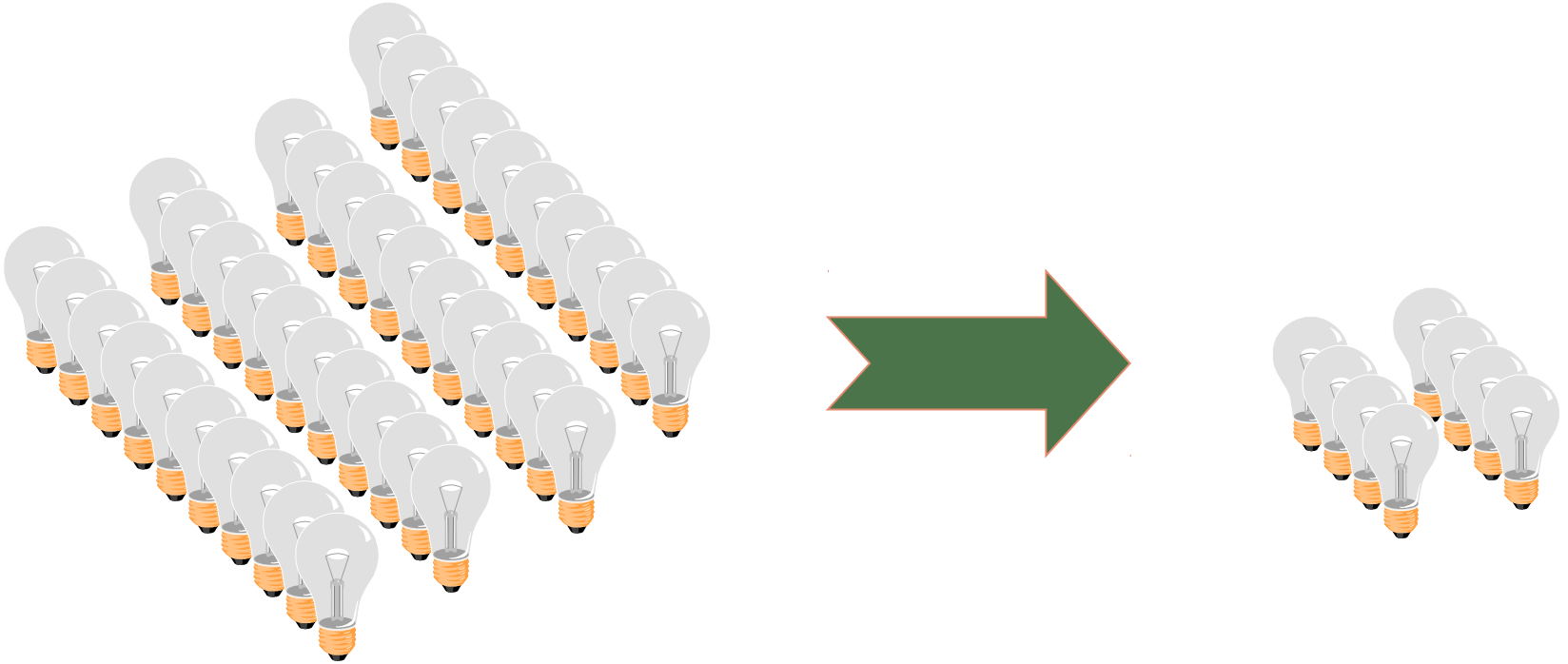
N° trabajadores que se ausentaron	N° días
0-4	4
5-9	10
10-14	8



$$\overline{X} = \frac{\sum_{i \in R_X} X_i}{n}, \quad S = \sqrt{\frac{\sum (X - \mu)^2}{n-1}}$$

Estadística Descriptiva:

Conjunto de métodos estadísticos que se relacionan con el ***resumen*** y ***descripción de datos***, como tablas, gráficas y el análisis mediante algunos cálculos. Trata con la enumeración, organización y representación gráfica de los datos.



<u>Estadística Inferencial</u>.- Conjunto de métodos cuya finalidad es hacer *generalizaciones* o inferencia sobre una población, utilizando la información de una parte de ella. **Está interesada en llegar a conclusiones de información incompleta, o sea, generalizado desde la muestra**

Algunos Ejemplos

- ¿Cuál es el número de llamadas telefónicas recibidas en una centralita durante un día? No existe un número fijo que pueda ser conocido a priori, sino un conjunto de posibles valores, cada uno de ellos con un cierto grado de certeza.
- ¿Cuál es el tamaño de un paquete de información que se transmite a través de HTTP? No existe un número fijo, sino que éste es desconocido a priori.
- ¿Cuál es la posición de un objeto detectado mediante GPS? Dicho sistema transmite una estimación de dicha posición, pero existen márgenes de error que determinan una región del plano donde el objeto se encuentra con alta probabilidad.
- ¿Qué ruido se adhiere a una señal que se envía desde un emisor a un receptor? Dependiendo de las características del canal, dicho ruido será más o menos relevante. Su presencia deberá ser diferenciada de la señal primitiva, sin que se conozca ésta, teniendo en cuenta que se trata de un ruido aleatorio.
- ¿Cuál fue el programa de televisión más visto la pasada noche? Los índices de audiencia se obtienen mediante estimaciones a partir de muestras representativas.

Conceptos fundamentales: Población y muestra

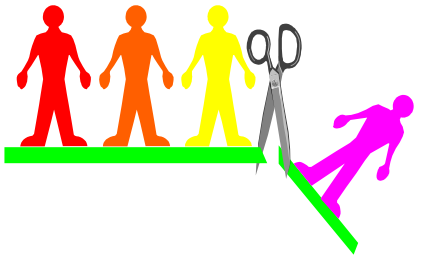
- **Población** es el conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencia).

  Observación: Normalmente es demasiado grande para poder abarcarlo.

- **Muestra** es un subconjunto de la población al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)

Observaciones: - Debería ser “representativo”

- Esta formado por miembros “seleccionados” de la población (individuos, unidades experimentales).



Ejercicio1.-Revisar todos los artículos fabricados que salen de una línea de ensamble con el fin de detectar defectos sería un procedimiento costoso que demandaría mucho tiempo. Un método económico y eficaz para determinar la cantidad de artículos defectuosos implica la selección y examen de una fracción de los artículos por parte de un ingeniero de control de calidad. Se calcula el porcentaje de los artículos examinados que salieron defectuosos y esta cifra se usa para estimar el porcentaje de todos los artículos fabricados en la línea que tienen defectos. Identifique la población, la muestra y el tipo de estadística que puede hacerse para este problema.

Pasos en un estudio estadístico

- **Recoger los datos (*muestreo*)**

¿Aleatorio?¿Estratificado?

- **Describir (resumir) los datos obtenidos**

- tiempo medio de sueño en usuarios y no usuarios (*estadísticos*)
    - % de horas de sueño por usuario, por nivel socioeconómico, por sexo (*frecuencias*)
- gráficos,...

- **Realizar una inferencia sobre la población**

-Los usuarios de msn duermen mínimo 3 horas menos en relación a la media de horas de sueño de los no usuarios

- **Cuantificar la confianza en la inferencia**

- *Nivel de confianza del 95%*

Variables

- Una **variable** es una **característica observable** *que varía entre los diferentes individuos* de una población. La información que disponemos de cada individuo es resumida en **variables**.

Dato

- Valor de la variable asociada a un elemento de una población o muestra. Este valor puede ser un número, una palabra o un símbolo
- Ejemplo: José Hernández ingresó a la universidad a la edad de “23” años, su cabello es “café”, mide “1.80m” y pesa “83 kg”.
- Estas cuatro piezas de datos son los valores de las cuatro variables aplicadas a José Hernández.

Tipos de variables

- **Cualitativas**

  Si sus valores no se pueden asociar naturalmente a un número.
  (no se pueden hacer operaciones algebraicas con ellos)

  - Nominales: Si sus valores no se pueden ordenar

Ejemplo: Sexo, Grupo Sanguíneo, Religión, Nacionalidad, Fumar (Sí/No)

  - Ordinales: Si sus valores se pueden ordenar

Ejemplo: Mejoría a un tratamiento, Grado de satisfacción, Intensidad del dolor

Tipos de variables

- **Cuantitativas o Numéricas**

  Si sus valores son numéricos
  (tiene sentido hacer operaciones algebraicas con ellos)

  - Discretas: Si toma valores enteros

Ejemplo: Número de hijos, Número de cigarrillos, Num. de “cumpleaños”

  - Continuas: Si entre dos valores, son posibles infinitos valores intermedios.

Ejemplo: Altura, Presión intraocular, Dosis de medicamento administrado, edad

Ejercicio2.-En una revista especializada, se informó de las dimensiones de desempeño de redes de distribución de agua en el área de Filadelfia. En una parte del estudio recabaron los siguientes datos para una muestra de secciones de tuberías de agua. Identifique los datos como cuantitativos o cualitativos.

1. **Diámetro de la tubería (pulgadas)**
2. **Material de la tubería.**
3. **Edad (año de instalación)**
4. **Ubicación.**
5. **Longitud de la tubería (pies)**
6. **Estabilidad del suelo circundante (inestable, moderadamente estable o estable)**
7. **Corrosividad del suelo circundante (corrosivo o no corrosivo)**

Ejercicio 3

- Identifique las siguientes expresiones como ejemplos de variables de atributos (cualitativas) o variables numéricas (cuantitativas)

  a) La resistencia a la rotura de un tipo de cuerda dado
  b) El color de cabello de los niños que se presentan a una audición
  c) El número de señales de alto que hay en poblaciones con menos de quinientos habitantes
  d) Si un grifo es o no defectuoso
  e) El tiempo necesario para contestar una llamada telefónica en cierta oficina de bienes raíces.

1.1. Primeros pasos en un análisis estadístico

Cuando disponemos de un conjunto de datos, debemos identificar:

1. La **característica** que representan dichos datos (**variable**).
2. La **población** de la que proceden los datos (conjunto total de individuos de interés).
3. La **naturaleza** de los datos:

   3.1. Variables cualitativas o atributos: Expresan una cualidad y no un valor numérico. Ejemplos: Sexo, Nacionalidad, Marcas de coche, Grado de Satisfacción con la Universidad, etc..

   3.2. Variables cuantitativas: Toma valores numéricos

   a) *Cuantitativas Discretas*: sólo pueden asumir ciertos valores y normalmente hay huecos entre ellos. Son conteos normalmente. Ejemplos: nº de asignaturas aprobadas, cantidad de hijos.

   b) *Cuantitativas Continuas*: puede asumir cualquier valor dentro de un intervalo. Normalmente representan magnitudes como longitud, superficie, volumen, peso, tiempo, dinero.

Formas de presentar y resumir la información de un conjunto de datos:

A) Tabla de frecuencias

A.1) Datos no agrupados

A.2) Datos agrupados

B) Descripción gráfica

B.1) Gráficos para v. cualitativas o cuantitativas discretas

B.2) Gráficos para v. cuantitativas continuas

B.3) Diagramas acumulados

B.4) Gráfico temporal

C) Descripción numérica

C.1) Medidas de localización o centralización

C.2) Medidas de dispersión o variabilidad

C.3) Medidas de forma

A) Tabla de Frecuencias

Intentan resumir la información recogida en la muestra, de forma que no se pierda nada de información (o poca).

- Frecuencias absolutas: Contabilizan el número de individuos de cada modalidad o **clase.**
- Frecuencias relativas (porcentajes): Es el cociente entre la frecuencia absoluta y el número total de datos. Contabilizan el porcentaje de individuos de cada modalidad.
- Frecuencias acumuladas: Contabilizan el número de individuos que toman un valor menor o igual que el dado en una modalidad. Sólo tienen sentido para variables cuantitativas (numéricas)
- **Ejemplos** de tablas de frecuencias para datos cualitativos y para datos cuantitativos discretos (transparencia 1)

Tabla de frecuencias

Suponga que estamos interesados en estudiar el número de niños en las familias viviendo en la comunidad. Los datos siguientes fueron reunidos basados en una muestra aleatoria de n=30 familias de la comunidad.

2, 2, 5, 3, 0, 1, 3, 2, 3, 4, 1, 3, 4, 5, 7, 3, 2, 4, 1, 0, 5, 8, 6, 5, 4 , 2, 4, 4, 7, 6

¡Organice estos datos en una tabla de frecuencias!

X=No. de niños x familia	Cuenta (Frecuencia)	Frecuencia relativa
0	2	2/30=0.067
1	3	3/30=0.100
2	5	5/30=0.167
3	5	5/30=0.167
4	6	6/30=0.200
5	4	4/30=0.133
6	2	2/30=0.067
7	2	2/30=0.067
8	1	1/30=0.033

Total es 30

Ejercicio:

Construir la distribución de frecuencias del número de trabajadores que se ausentaron en 25 días laborables:

2	3	3	0	1	2
1	2	2	1	3	3
2	1	0	1	2	3
4	3	2	4	2	1
					0

Los datos se pueden analizar de manera agrupada o no agrupada.

Hablamos de datos agrupados como aquellos que pertenecen a un tamaño de muestra definido. Su objetivo es resumir información.

Es necesario que se **ordenen y clasifiquen** en una tabla de frecuencias. Los datos se agrupan de forma lógica y coherente en **intervalos de clase**.

Tabla de frecuencias datos agrupados

Suponga que necesitamos construir una tabla de frecuencias similar para la edad de pacientes con problemas relacionados al corazón en una clínica.

Los siguientes datos han sido reunidos basados en una muestra aleatoria de n=30 pacientes quienes fueron a emergencias de la clínica por problemas relacionados al corazón.

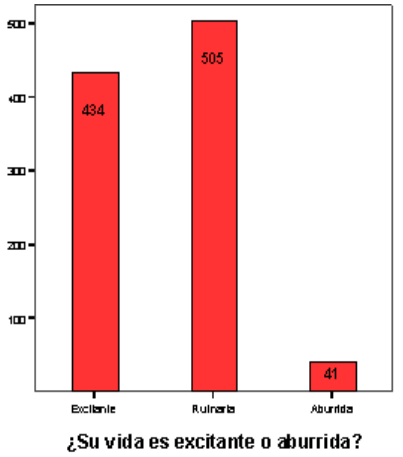
Las mediciones fueron: 42, 38, 51, 53, 40, 68, 62, 36, 32, 45, 51, 67, 53, 59, 47, 63, 52, 64, 61, 43, 56, 58, 66, 54, 56, 52, 40, 55, 72, 69.

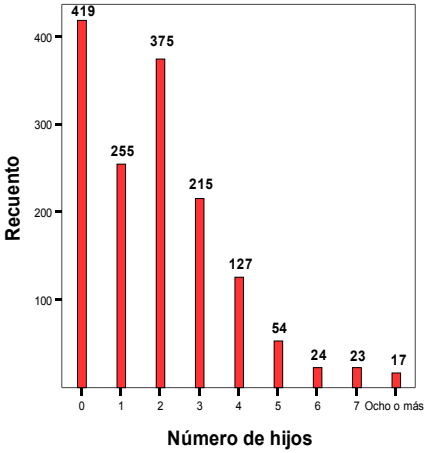
Grupos de edad	Frecuencia	Frecuencia relativa
32 -36.99	2	2/30=0.067
37- 41.99	3	3/30=0.100
42-46.99	4	4/30=0.134
47-51.99	3	3/30=0.100
52-56.99	8	8/30=0.267
57-61.99	3	3/30=0.100
62-66.99	4	4/30=0.134
67-72	3	3/30=0.100
Total	n=30	1.00

B) Descripción Gráfica

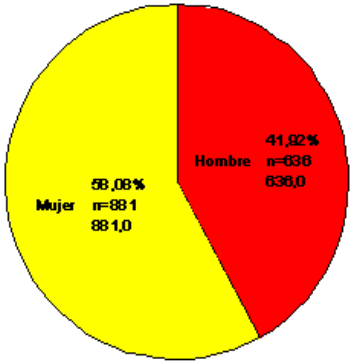
B.1) Gráficos para v. cualitativas o cuantitativas discretas

- Diagramas de barras
  - Alturas proporcionales a las frecuencias (abs. o rel.)





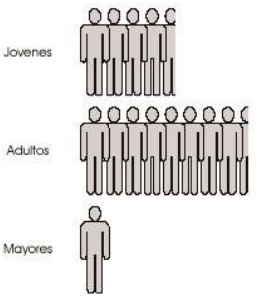
- Diagramas de sectores (tartas, polares)
  - El área de cada sector es proporcional a su frecuencia (abs. o rel.)



- Pictogramas
  - El área de cada modalidad debe ser proporcional a la frecuencia.

Pirámide de edad de una población de 31 millones de personas.

Edad (años)	0-19	20-64	Más de 64	Total
Porcentaje(%)	34,1	59,0	6,9	100,0



Pictograma de repetición

Pictograma de amplificación

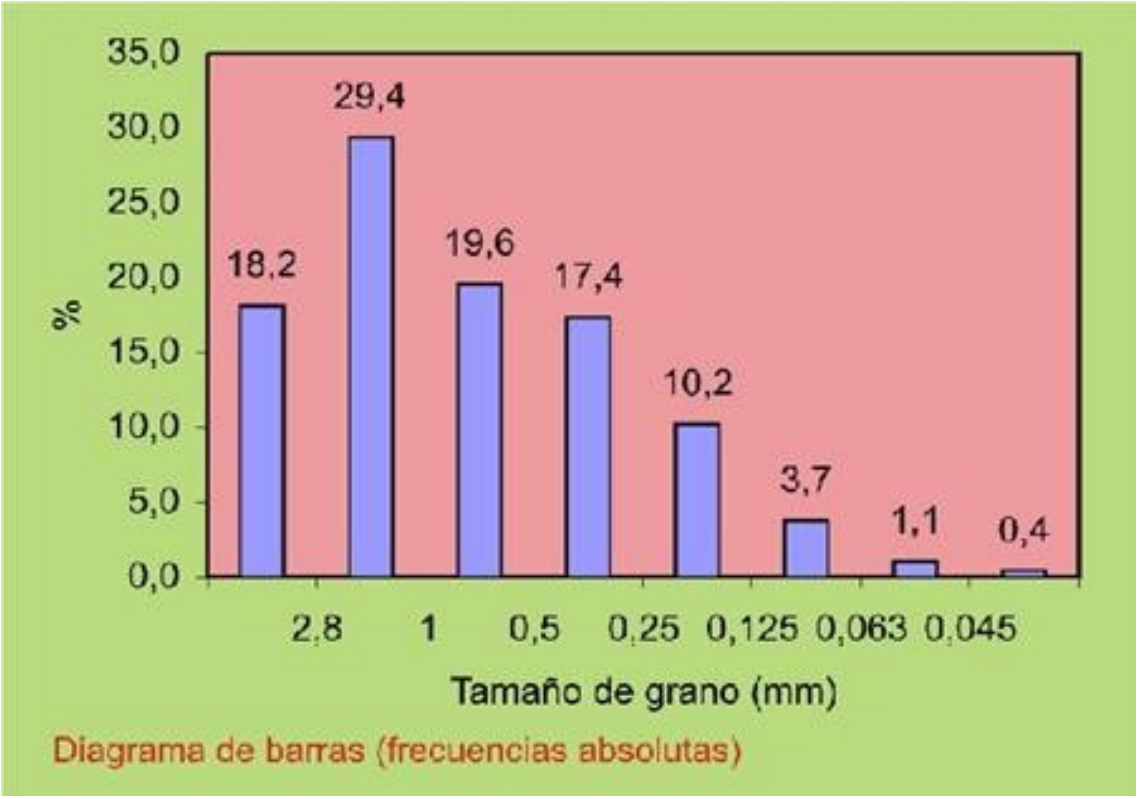
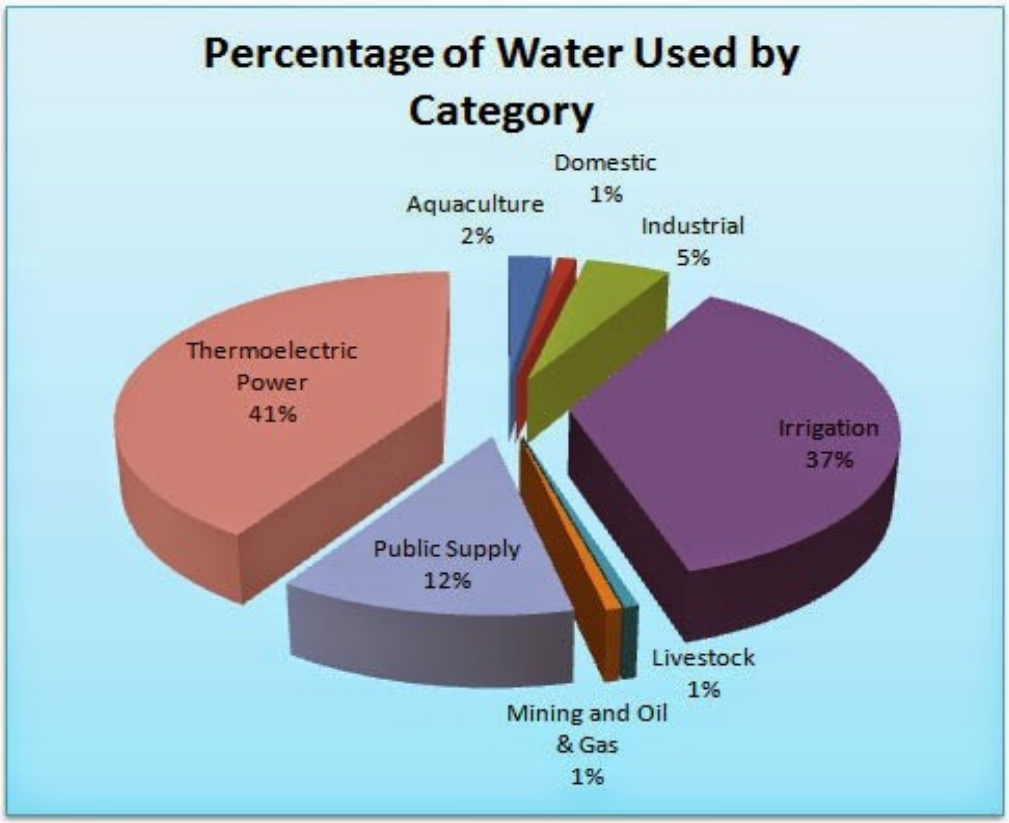
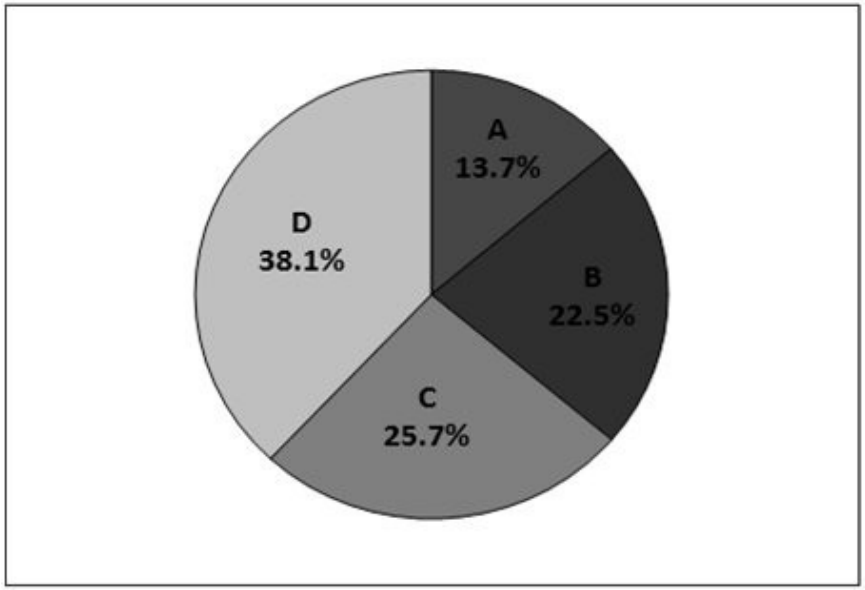


Diagrama de barras (frecuencias absolutas)


Percentage of Water Used by Category
Domestic 1%
Aquaculture 2%
Industrial 5%
Thermoelectric Power 41%
Irrigation 37%
Public Supply 12%
Livestock 1%
Mining and Oil & Gas 1%


A 13.7%
B 22.5%
C 25.7%
D 38.1%

Pictogram

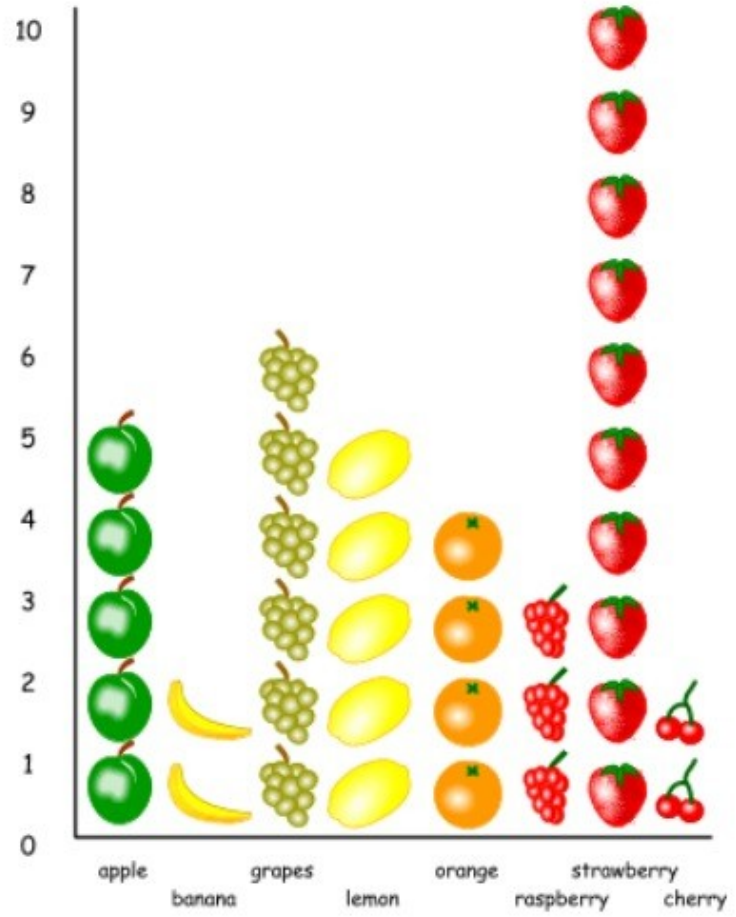
as of Aug 26, 2003

City	
Boston	
Dallas	
Los Angeles	
Orlando	
Seattle	
St Louis	

* Each TV equals 200000 units

What might the title of this chart be?

What might the label on this y-axis be?

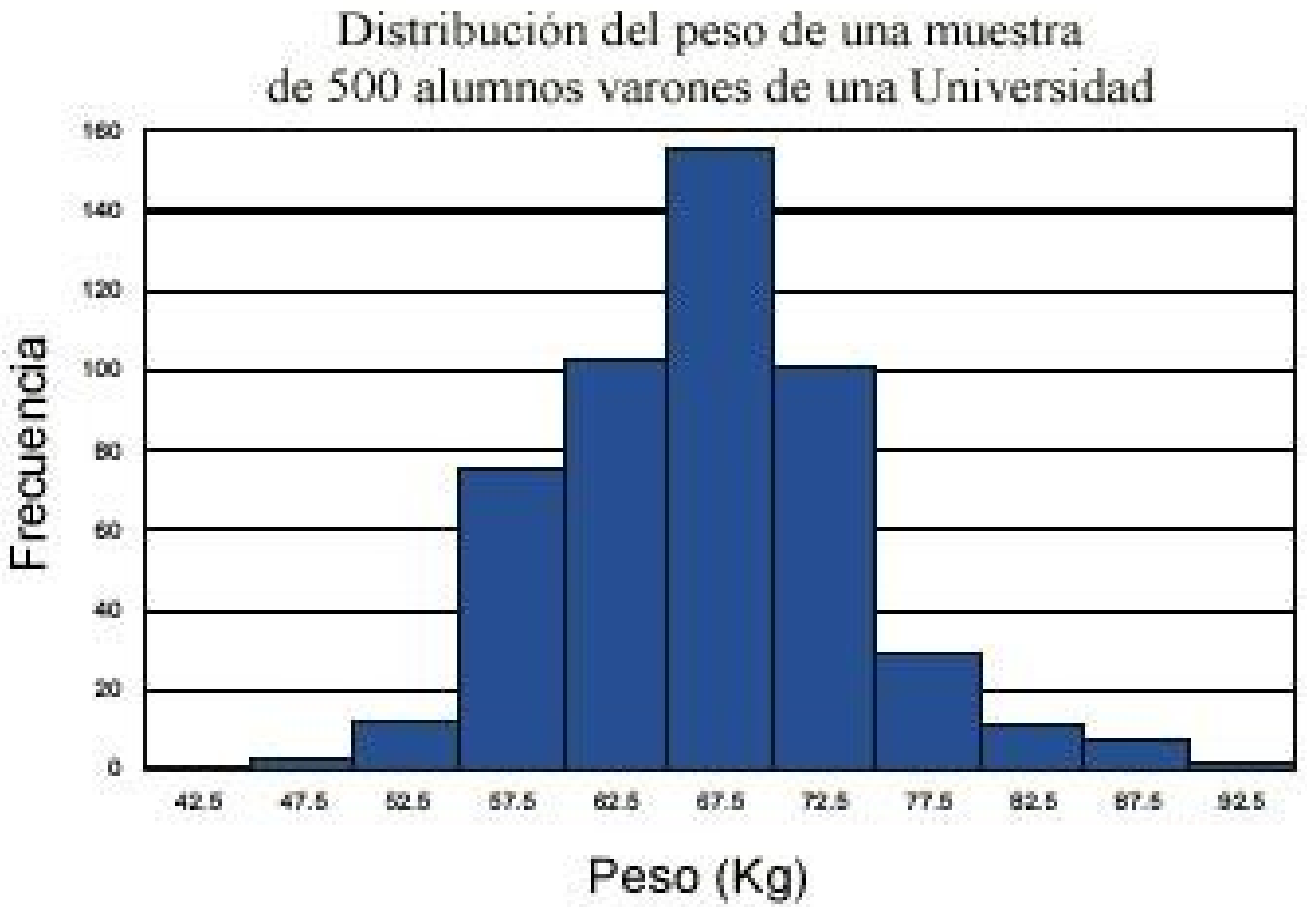


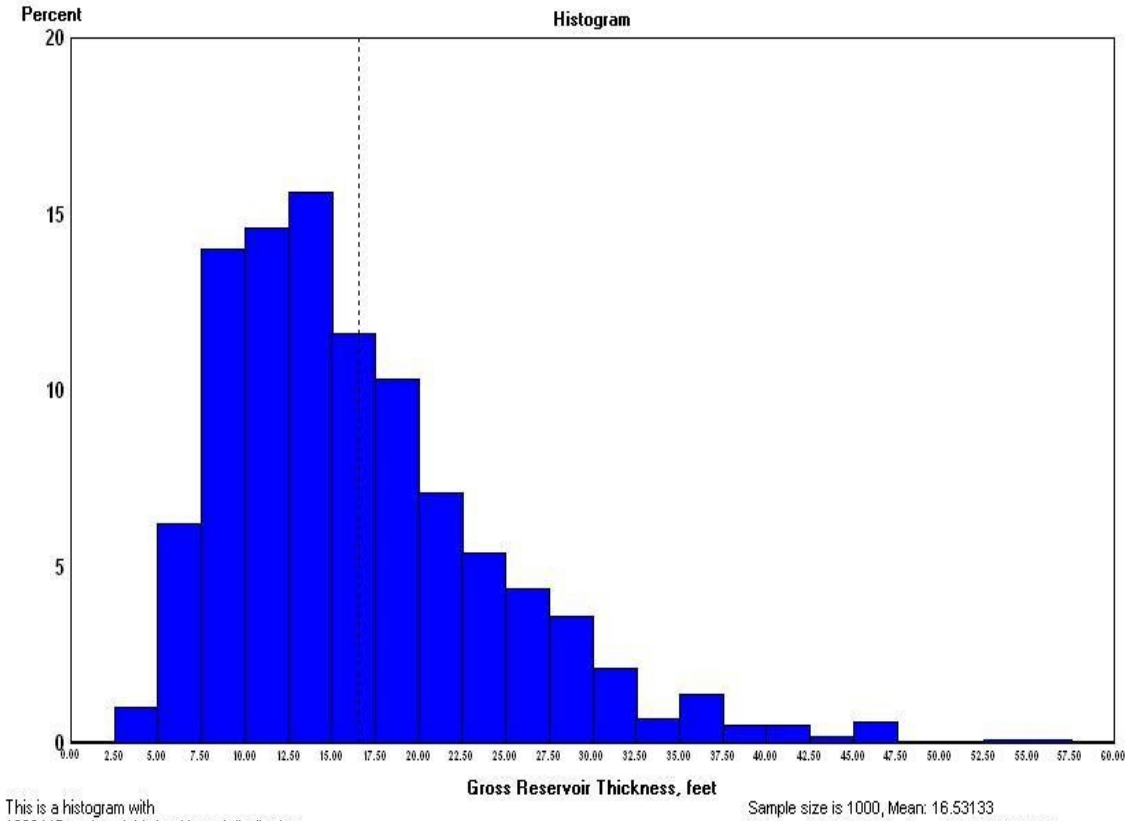
What might the label on this x-axis be?

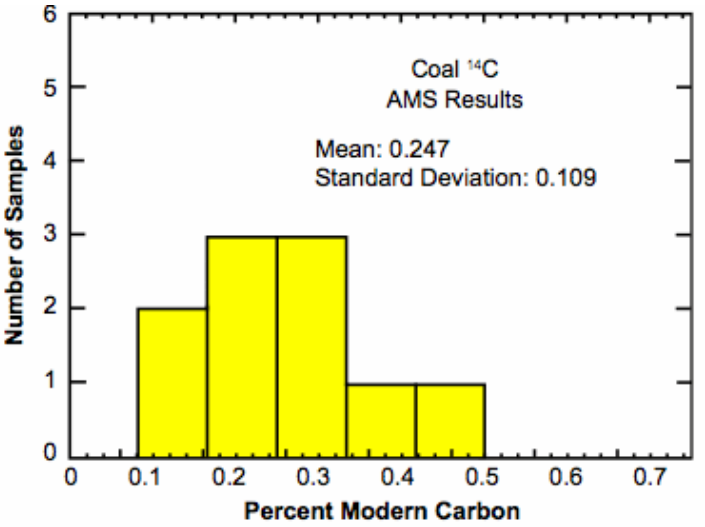
B.2) Gráficos para v. cuantitativas continuas

- **Histograma**

  - Para conjuntos con gran número de datos. Es la representación gráfica de la tabla de frecuencias para datos agrupados en clases. El área que hay bajo el histograma entre dos puntos cualesquiera indica la cantidad (porcentaje o frecuencia) de individuos en dicho intervalo.

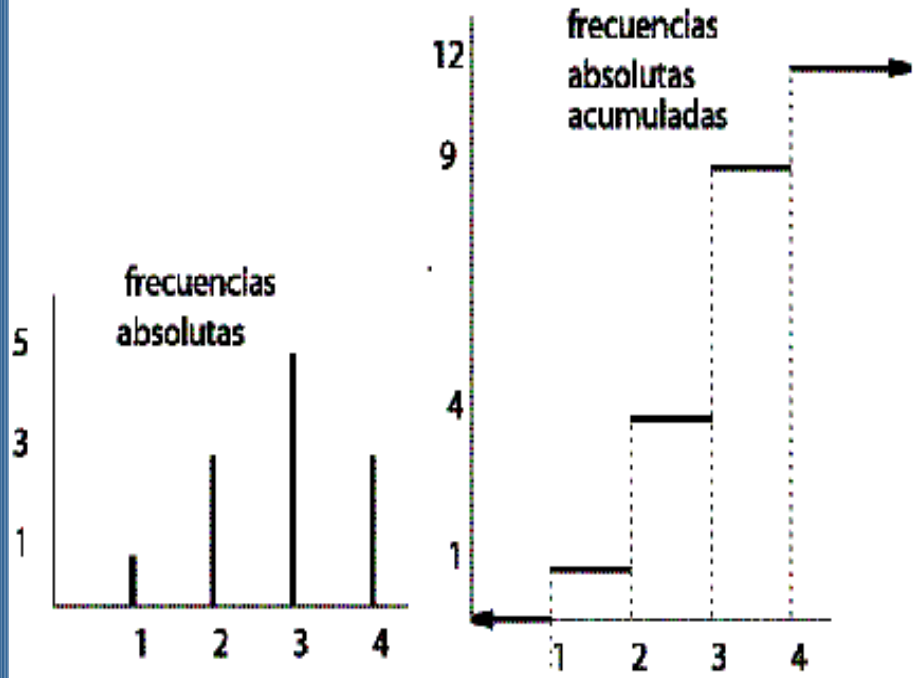



Percent
Histogram
20
15
10
5
0
0.00
2.50
5.00
7.50
10.00
12.50
15.00
17.50
20.00
22.50
25.00
27.50
30.00
32.50
35.00
37.50
40.00
42.50
45.00
47.50
50.00
52.50
55.00
57.50
60.00
Gross Reservoir Thickness, feet
This is a histogram with
1000 MC cycles of this LogNormal distribution
Sample size is 1000, Mean: 16.53133
Minimum: 3.81843, Maximum: 55.44495, POS: 1


Coal ^{14}C
AMS Results
Mean: 0.247
Standard Deviation: 0.109
Number of Samples
0
1
2
3
4
5
6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Percent Modern Carbon

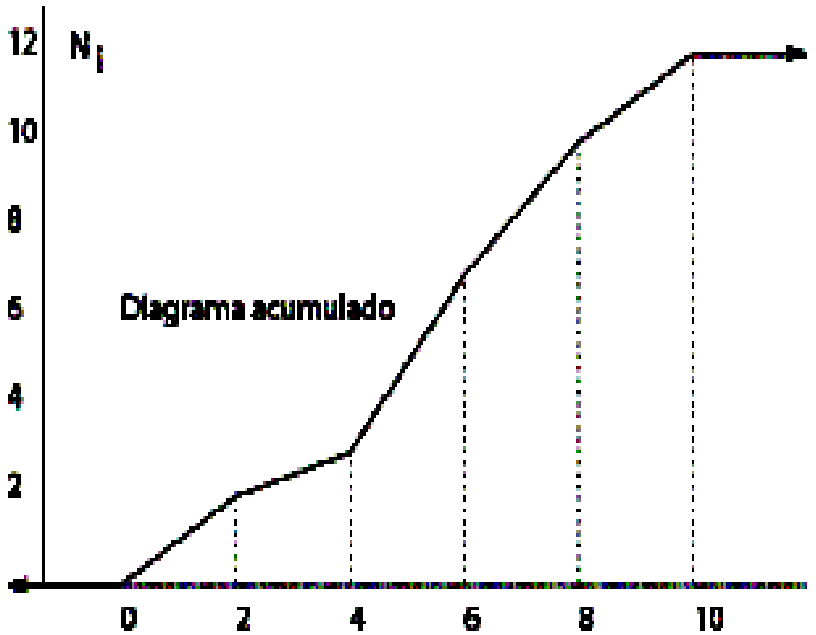
B.3) Diagramas acumulados

Algunos de los diagramas anteriores tiene su correspondiente diagrama acumulado. Se realizan a partir de las frecuencias acumuladas. Indican, para cada valor de la variable, la cantidad (frecuencia) de individuos que poseen un valor inferior o igual al mismo.

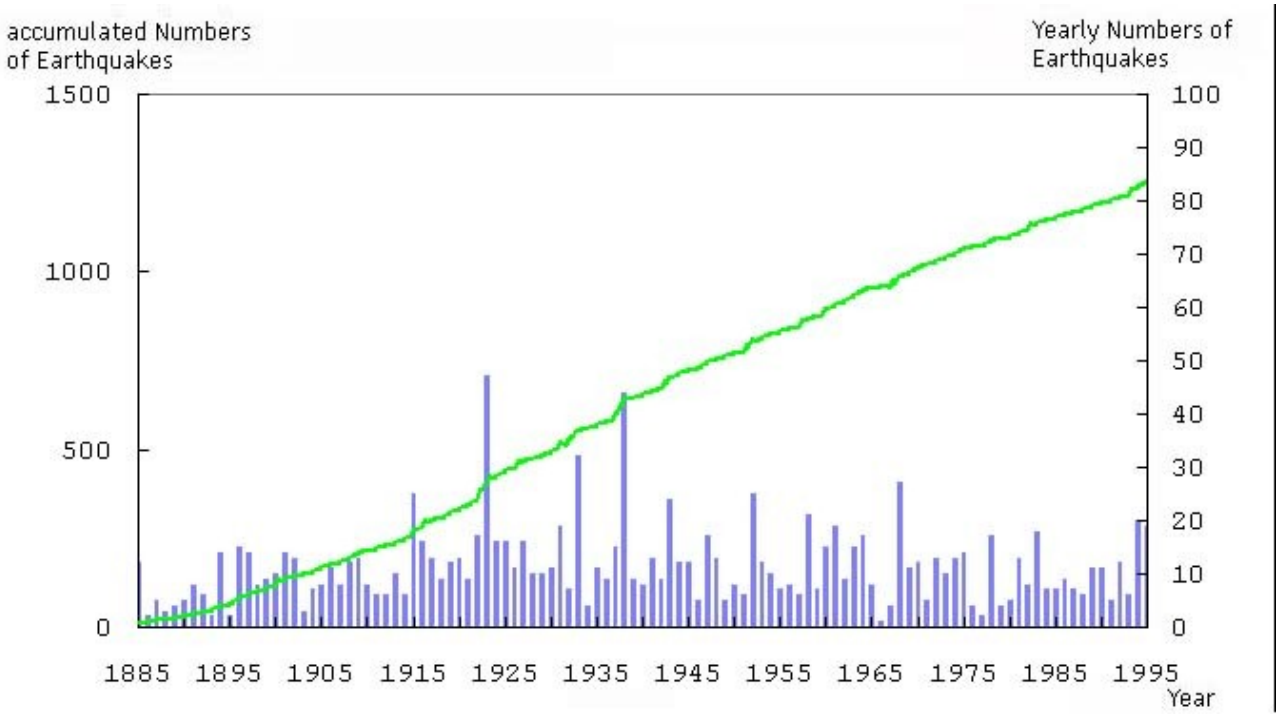


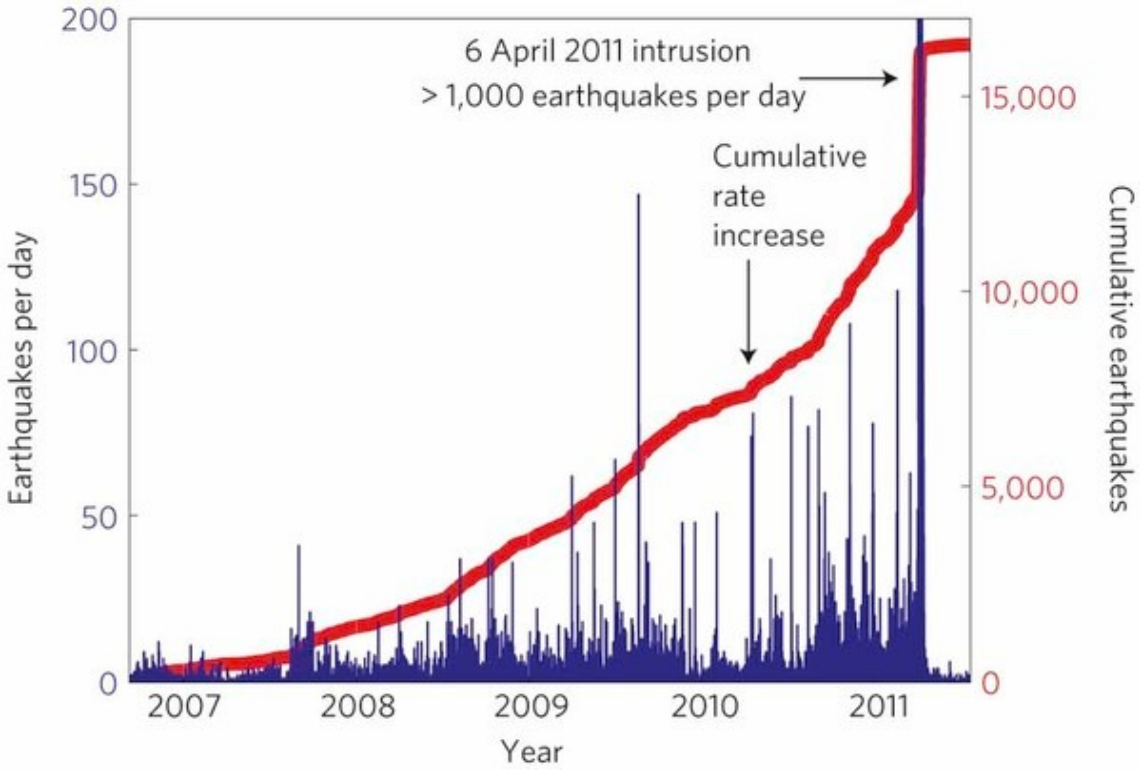
Frec. Absolutas

Frec. Abs Acumuladas



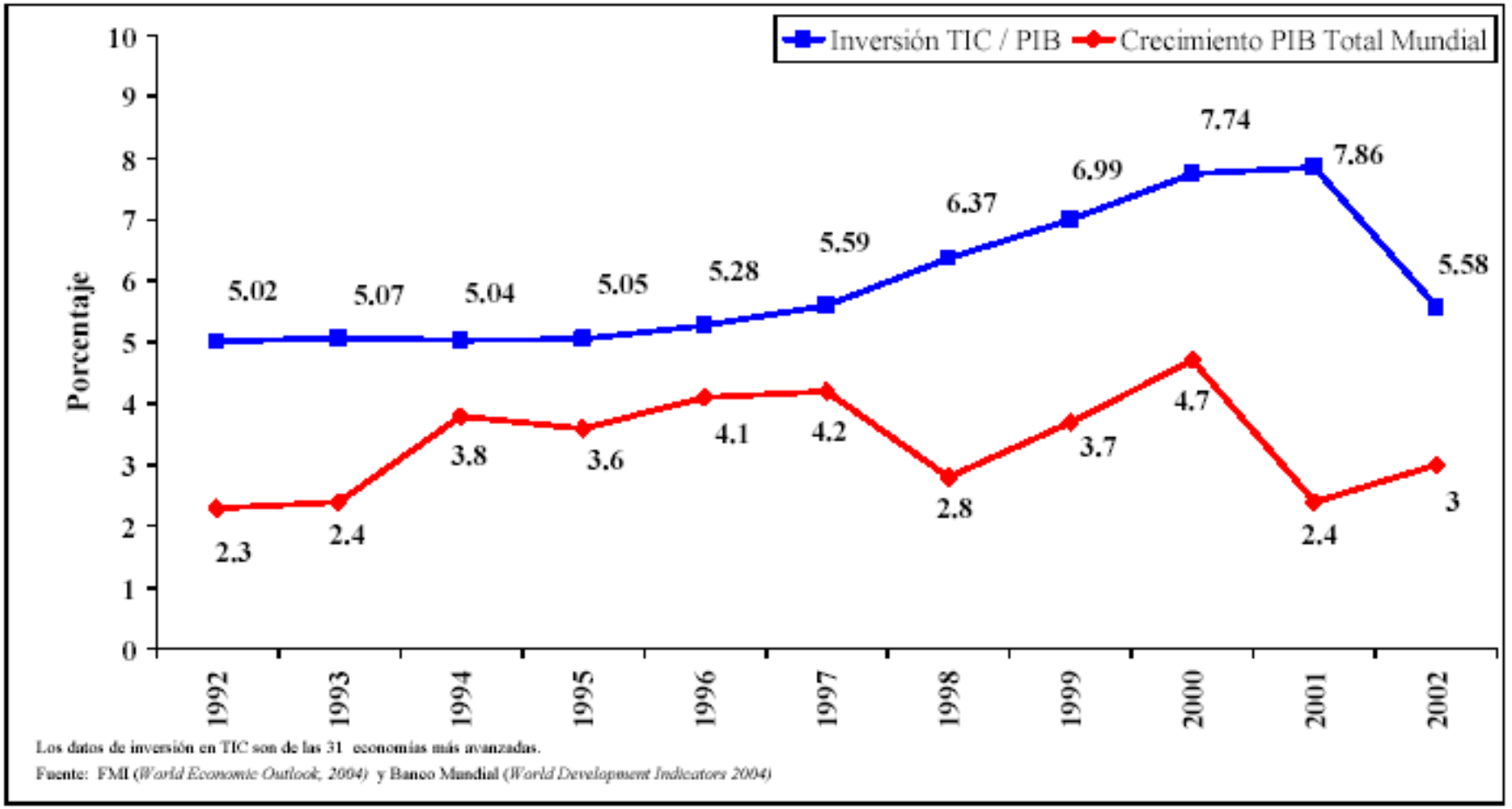
Polígono de Frec. Abs Acumuladas


accumulated Numbers
of Earthquakes
Yearly Numbers of
Earthquakes
1500
1000
500
0
100
90
80
70
60
50
40
30
20
10
0
1885 1895 1905 1915 1925 1935 1945 1955 1965 1975 1985 1995
Year


6 April 2011 intrusion
> 1,000 earthquakes per day
Cumulative
rate
increase
Earthquakes per day
Cumulative earthquakes
200
150
100
50
0
15,000
10,000
5,000
0
2007
2008
2009
2010
2011
Year

B.4) Gráfico temporal

Gráfico 1: Evolución del Sector TIC e IMPACTO en PIB Mundial 1992-2002*



- El crecimiento TICs va de la mano con crecimiento PIB.

C) Descripción Numérica

Objetivo: Resumir la información más relevante de la muestra o población en unos pocos números (parámetros).

C.1) Medidas de Centralización o Localización

- Indican valores con respecto a los que los datos parecen agruparse.
  - Media, mediana y moda

C.2) Medidas de Posición

- Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.
  - Cuantiles, percentiles, cuartiles, deciles,...

C.3) Medidas de Dispersión o Variabilidad

- Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
  - Rango, varianza, desviación típica, rango intercuartílico, coeficiente de variación

C.4) Medidad de Forma

- Indican la forma en que se distribuyen los datos
  - Coeficientes de asimetría y de apuntamiento o curtosis

Diferencia entre encuestas y experimentos

Datos de una encuesta representan observaciones de eventos o fenómenos sobre los cuales pocos o ningún, control se impone.

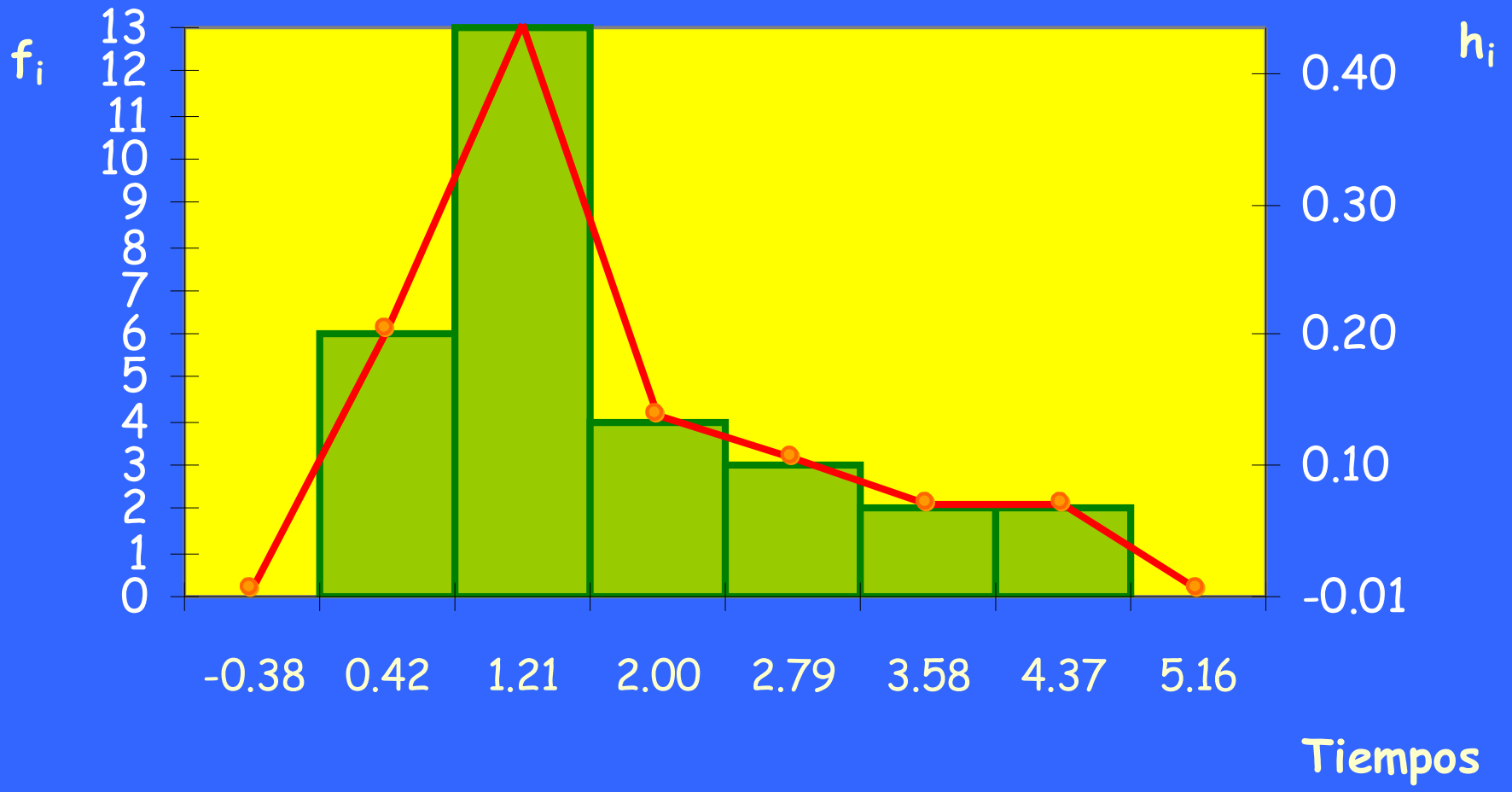
(e.g., evaluando la asociación entre diferentes estilos de vida y enfermedad cardiaca)

En un experimento diseñamos una investigación planeada a propósito para imponer controles sobre la cantidad de exposición (tratamiento) a una medicamento. (e.g., estudios clínicos)

Ejemplo. Los siguientes datos representan el tiempo (en segundos) que 30 trabajadores estuvieron al control de la unidad central de procesos (CPU) de una computadora mainframe grande.

0.02	0.75	1.16	1.38	1.94	3.07
0.15	0.82	1.17	1.4	2.01	3.53
0.19	0.84	1.19	1.42	2.16	3.76
0.47	0.92	1.22	1.59	2.41	4.50
0.71	0.96	1.23	1.61	2.59	4.75

Histograma de los tiempos



Ojiva de los tiempos