PROBABILIDAD Y ESTADÍSTICA

Curso 2023-2024

Mar Angulo Martínez

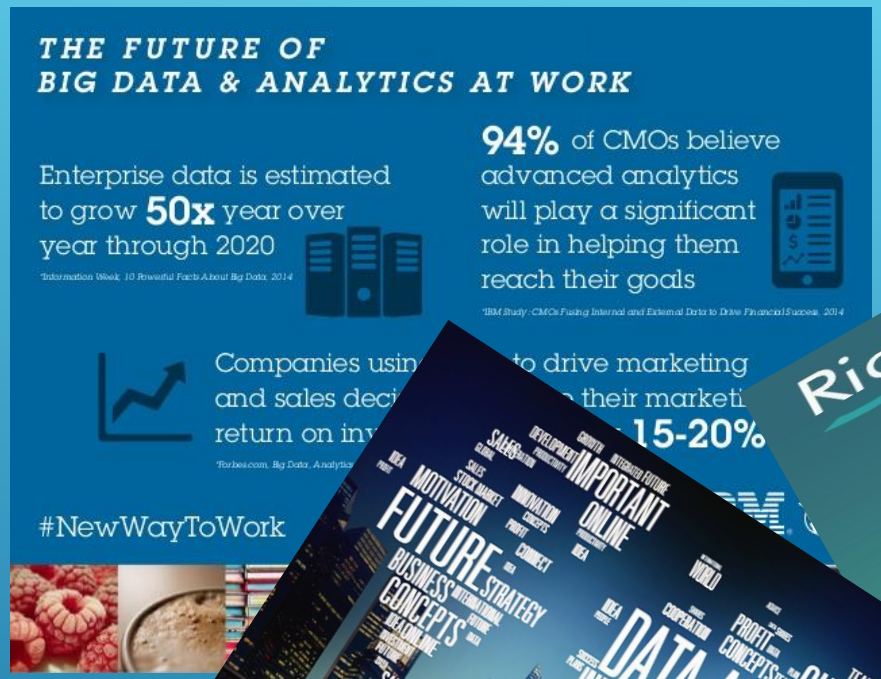
U-tad

En Estadística aprenderemos técnicas para tomar decisiones informadas y emitir juicios inteligentes en presencia de incertidumbre

¿PARA QUÉ APRENDER ESTADÍSTICA?

- "En los tiempos antiguos no tenían estadísticas por lo que tuvieron que recurrir a la mentira". Stephen Leacock
- **La estadística va de entender datos. Y el mundo de hoy está hecho fundamentalmente de datos.**

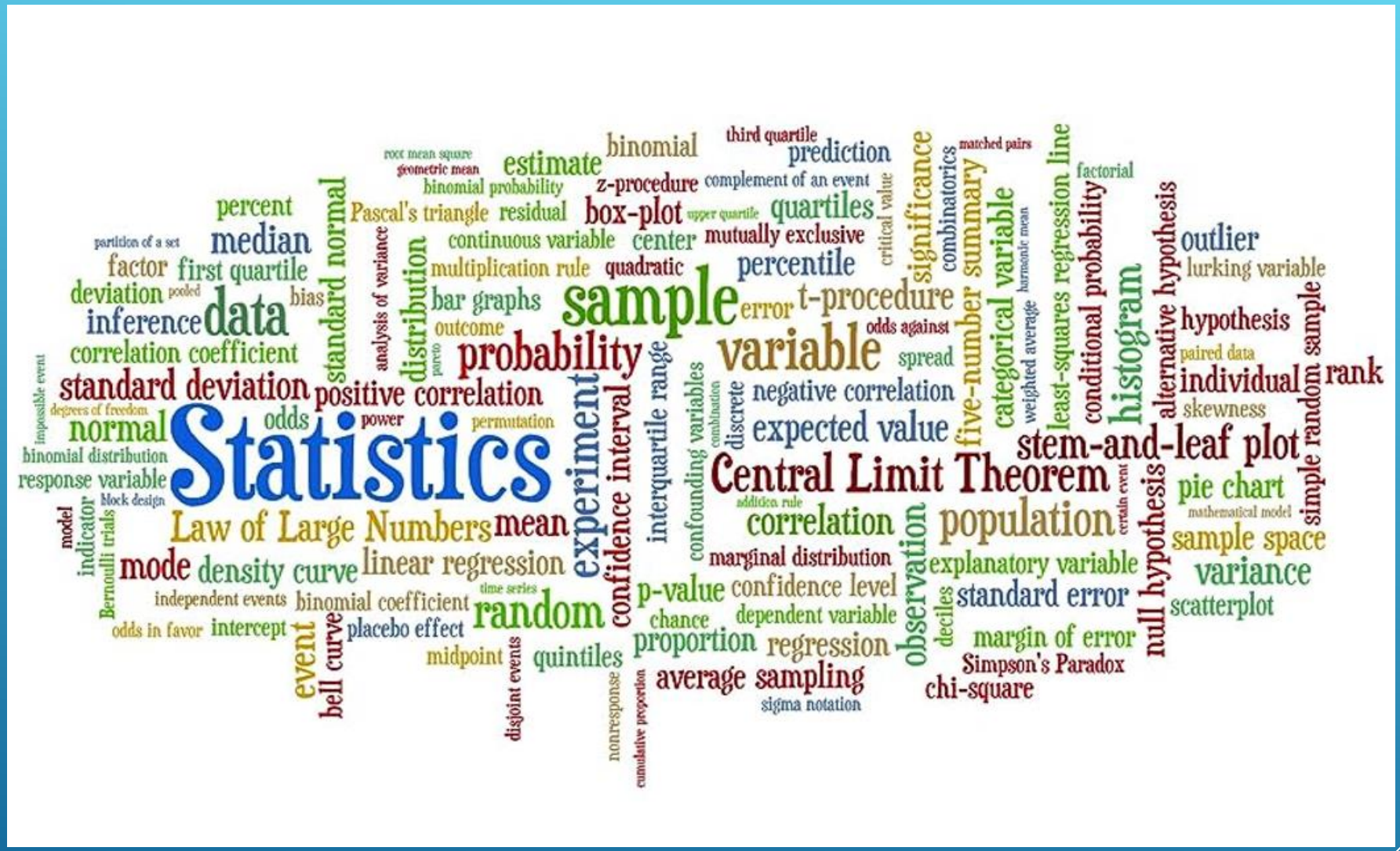
- Ricardo Galli, "las estadística es **una herramienta fundamental para analizar y entender los problemas** en un mundo tan complejo" como el actual.


THE FUTURE OF
BIG DATA & ANALYTICS AT WORK
Enterprise data is estimated to grow 50x year over year through 2020
*Information Week: 10 Powerful Facts About Big Data, 2014
94% of CMOs believe advanced analytics will play a significant role in helping them reach their goals
*IBM Study: CMOs Fusing Internal and External Data to Drive Financial Success, 2014
Companies usin to drive marketing and sales deci their marketi
15-20%
return on inv
*Forbes.com, Big Data, Analytics
#NewWayToWork


DATA ANALYSIS
INFORMATION
COMMUNICATION
FUTURE
STRATEGY
BUSINESS
CONCEPTS
IMPORTANT
ONLINE
QUALITY
PEOPLE
EDUCATIONAL
PLANS


Ride the Big Data Wave
By 2020
35 Zettabytes worth of data will be created
*35 Zettabytes = 35,000,000,000,000,000,000,000 bytes
90%
of the world's existing data has been created in the last
2 YEARS
1/3 of that data will be stored in or will have passed through the cloud
80,000 km
is the distance past the moon the current global storage capacity for digital information stored on a CD stack would stretch
10,000
Libraries of Congress worth of data are created annually by U.S. companies
There are actually more pieces of digital information than stars in the universe
Every 2 days as much information is created as from the dawn of civilization up until 2003
EVERY DAY
2.5 Quintillion
8 Years
YouTube
of video is uploaded to YouTube
1 Billion
bytes of data are generated by people
pieces of content are shared via Facebook's Open Graph
De-risk Your Database Vendor Selections by Aligning with Market Trends
See the Market Trends with the Progress 2014 Data Connectivity Outlook
An Uninformed Database Vendor Selection can be Costly to Your Business
Full Paper: http://goo.gl/ynp9DI
*Source: http://goo.gl/amfUuY

U-tad


Statistics
data
sample
variable
probability
Central Limit Theorem
Law of Large Numbers
standard deviation
positive correlation
normal
inference
median
mean
mode
density curve
linear regression
random
experiment
confidence interval
standard normal
distribution
bar graphs
percent
factor
first quartile
deviation
bias
correlation coefficient
odds
power
permutation
binomial distribution
response variable
block design
indicator
model
Bernoulli trials
independent events
binomial coefficient
odds in favor
intercept
placebo effect
event
bell curve
midpoint
quintiles
disjoint events
nonresponse
cumulative proportion
average
sampling
sigma notation
proportion
regression
chance
dependent variable
p-value
confidence level
time series
correlation
observation
population
marginal distribution
explanatory variable
standard error
margin of error
Simpson's Paradox
chi-square
deciles
null hypothesis
alternative hypothesis
hypothesis
paired data
individual
rank
skewness
stem-and-leaf plot
pie chart
mathematical model
sample space
variance
scatterplot
simple random sample
outlier
lurking variable
histogram
conditional probability
factorial
least-squares regression line
five-number summary
categorical variable
weighted average
harmonic mean
significance
combinatorics
matched pairs
critical value
upper quartile
mutually exclusive
center
continuous variable
multiplication rule
quadratic
percentile
t-procedure
error
odds against
spread
negative correlation
expected value
discrete
combination
confounding variables
interquartile range
addition rule
certain event
analysis of variance
pareto
outcome
z-procedure
complement of an event
box-plot
quartiles
Pascal's triangle
residual
binomial probability
root mean square
geometric mean
estimate
binomial
prediction
third quartile
partition of a set
pooled
degrees of freedom
impossible event
permutation

LA ESTADÍSTICA EN LA CIENCIA DE DATOS

La estadística permite diseñar los modelos y patrones sobre los datos, y permite inferir conclusiones sobre poblaciones a partir del análisis detallado de muestras de datos más pequeños

En todos los campos…

USOS Y APLICACIONES DE LA ESTADÍSTICA

Agricultura /Ecología

Sociología/Psicología/ Pedagogía

Gobierno/ Instituciones/ Economía

Física

Medicina y Salud/Biología

Educación

Ingeniería

Deportes

- *Dependencia entre meteorología y rendimientos*
- *Efectos de distintos tipos de fertilizantes, semillas, insecticidas,*
- *Distribución por zonas de las distintas especies*
- *Sistemas de vigilancia de la salud forestal por los efectos del cambio climático*

- *Comparación de conductas*
- *Análisis de comportamiento en distintos grupos sociales, culturales…*
- *Mejora de sistemas educativos, de evaluación*
- *Análisis de factores que intervienen en la inteligencia*

- *Encuestas, estudios de opinión, de mercado, análisis de clientes…*
- *Diseño de nuevas políticas, construcción de infraestructuras…*
- *Business Analytics: modelos data driven*
- *Índices económicos: inflación, desigualdad, competitividad, productividad…*

- *Física cuántica*
- *Radiactividad, energía atómica: procesos estocásticos*

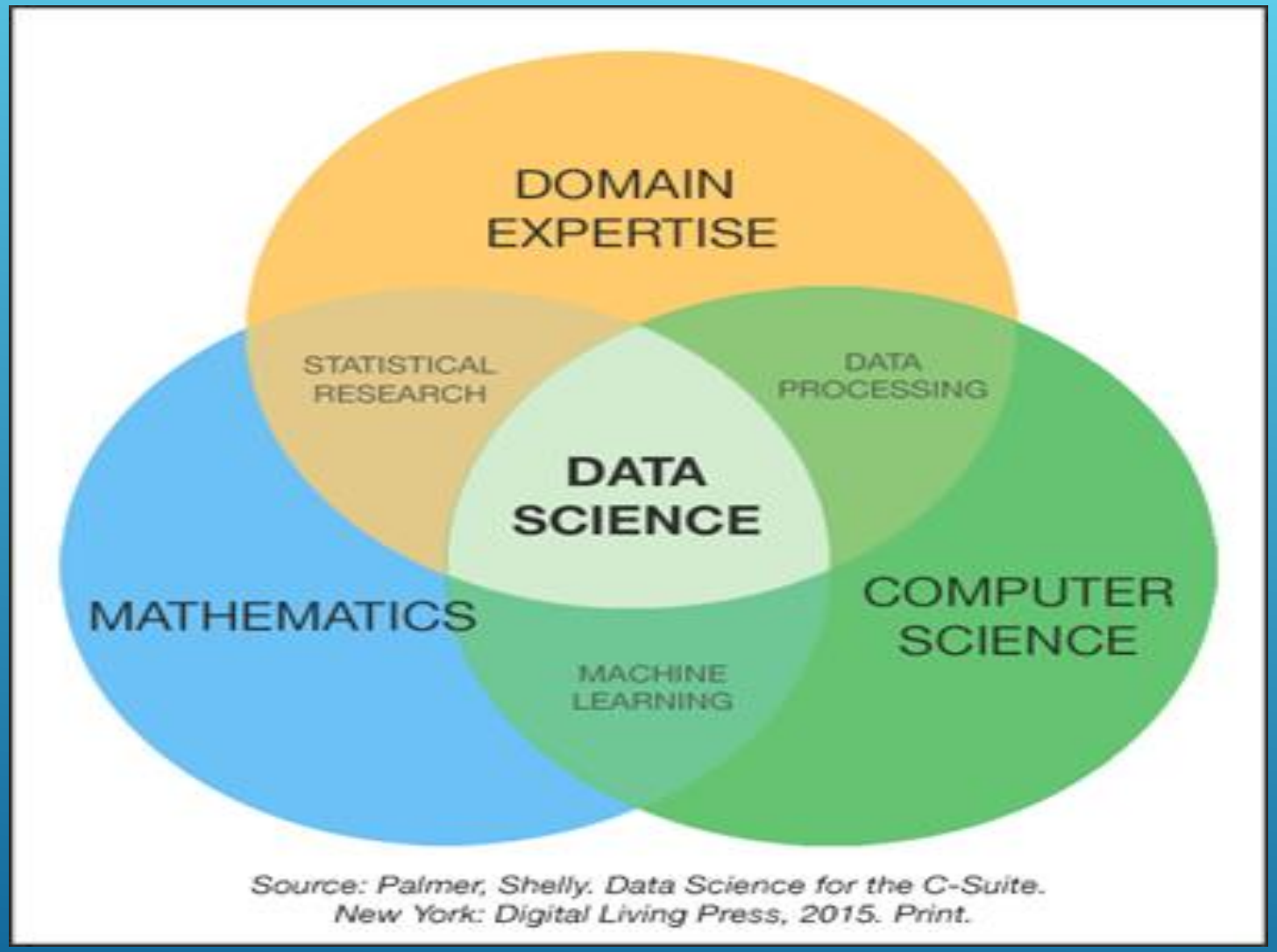
https://www.census.gov/programs-surveys/sis/resources/videos/why-statistics.html

- *Estudios sobre efectividad y comparación de fármacos*
- *Análisis de pruebas diagnósticas*
- *Clasificación de individuos (grupos de riesgo, tratamientos experimentales…)*
- *Planificación de infraestructura hospitalaria, equipos…*
- *Investigación del cáncer y análisis de enfermedades recurrentes (migraña): los datos ofrecen información sobre la incidencia, gravedad, costos médicos asociados…*

- *Técnicas de control de calidad*
- *Efectos de los esfuerzos en elementos estructurales*
- *Propiedades de tratamientos en materiales*
- *Software Big Data*

  - *media, desviación típica, tamaño muestral, p-valor, regresión…*
  - *Aprendizaje automático, Redes neuronales, algoritmos genéticos, reglas de asociación, análisis de series temporales…*

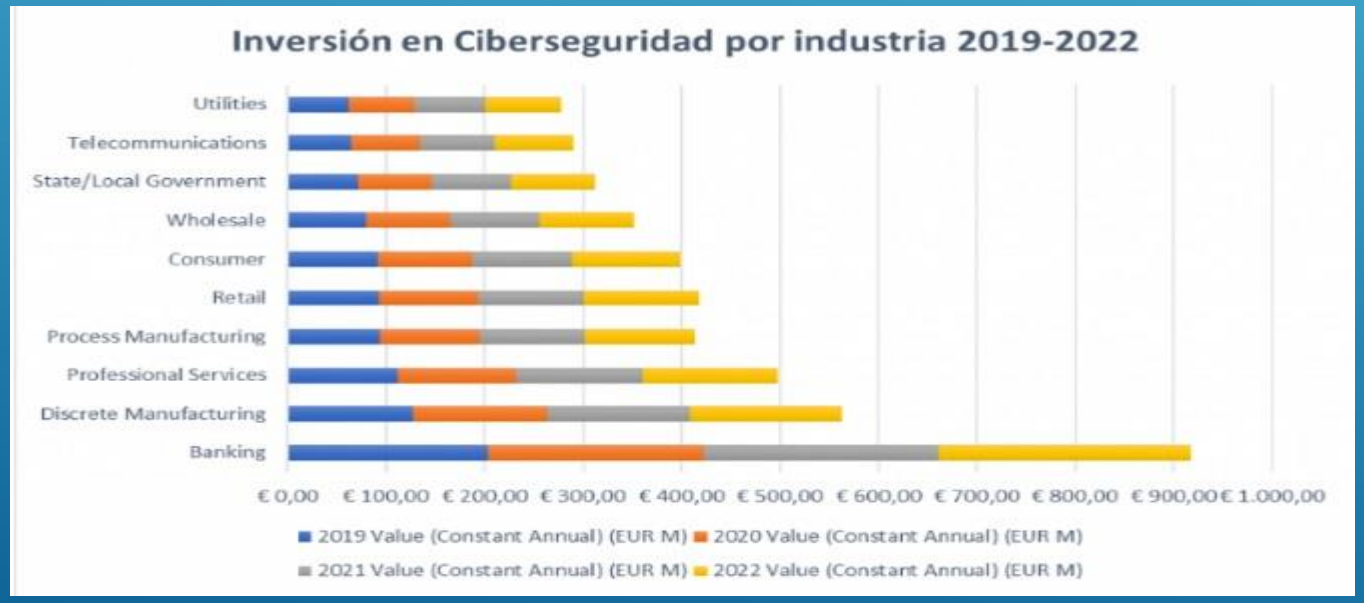
https://www.amstat.org/asa/publications/Journals.aspx



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Big Data : Los datos son oro para las empresas

- Tomar mejores decisiones
- Resolver problemas (Ej: descubrir en qué momento del proceso se produce un fallo de rendimiento)
- Conocer mejor el rendimiento y el negocio
- Mejorar los procesos (producción, marketing, RRHH…)
- Conocer las expectativas de los clientes



Qué es….

EL MÉTODO ESTADÍSTICO

https://www.census.gov/programs-surveys/sis/resources/videos/why-statistics.html

ESTADÍSTICA: DATOS Y AZAR

Es la rama de las Matemáticas que permite organizar, resumir, representar, analizar e interpretar los datos de una o varias características de una población

¿CENSOS O MUESTRAS?

U-Tad

- ESTADÍSTICA DESCRIPTIVA
- RECOGIDA, ORDENACIÓN, TABULACIÓN Y REPRESENTACIÓN DE DATOS DE UNA MUESTRA
- **PROBABILIDAD**
- INFERENCIA ESTADÍSTICA
- TÉCNICAS PARA EXTRAER CONCLUSIONES SOBRE LA POBLACIÓN

U-tad

Tema 1.-Conceptos generales

1.1. Fundamentos de la Estadística.

1.2. Población y muestra.

1.3. Estadística Descriptiva y Estadística Inferencial

1.4. El método estadístico

ESTADÍSTICA DESCRIPTIVA: ORGANIZAR, RESUMIR, REPRESENTAR….

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n}$$

POBLACION

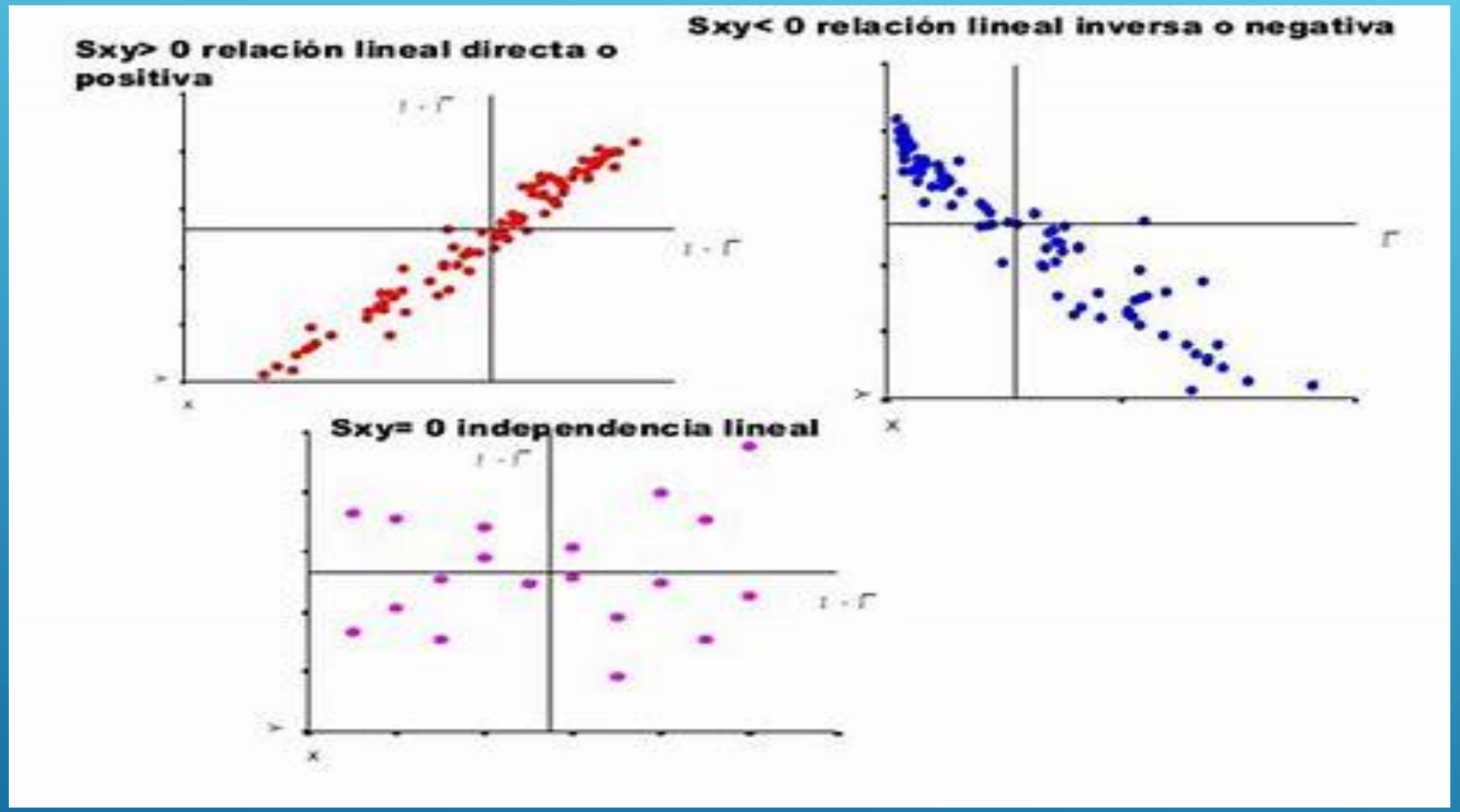
CP	PROVINCIA	2009p	2008p	2007p	2006p	2005p	2004p	2003p
01	Álava	313819	309635	305459	301926	299957	295905	294360
02	Albacete	400891	397493	392110	387658	384640	379448	376556
03	Alicante/Alacant	1917012	1891477	1825264	1783565	1732389	1657040	1632349
04	Almería	684426	667635	646833	635850	612315	580077	565310
05	Ávila	171680	171815	169638	167818	167032	166108	165480
06	Badajoz	688777	685246	678459	673474	671299	663896	663142
07	Balears (Illes)	1095426	1072844	1030650	1001062	983131	955045	947361
08	Barcelona	5487935	5416447	5332513	5309404	5226354	5117885	5052686
09	Burgos	375563	373672	365972	363874	361021	356437	355205
10	Cáceres	413633	412498	411531	412899	412580	411390	410762
11	Cádiz	1230594	1220467	1207343	1194062	1180817	1164374	1155724
12	Castellón/Castelló	602301	594915	573282	559761	543432	527345	518239
13	Ciudad Real	527273	522343	510122	506884	500060	492314	487670
14	Córdoba	803998	798822	792182	788287	784376	779870	775944
15	Coruña (A)	1145488	1139121	1132792	1129141	1126707	1121344	1120814
16	Cuenca	217363	215274	211375	208616	207974	204546	202982
17	Girona	747782	731864	706185	687331	664506	636198	619692
18	Granada	907428	901220	884099	876184	860898	841687	828107
19	Guadalajara	246151	237787	224076	213505	203737	193913	185474
20	Guipúzcoa	705698	701056	694944	691895	688708	686513	684416
21	Huelva	513403	507915	497671	492174	483792	476707	472446
22	Huesca	228409	226271	220107	218023	215864	212901	211286
23	Jaén	669782	667438	664742	662751	660284	654458	651565
24	León	500169	500200	497387	498223	495902	492720	495998
25	Lleida	436402	426872	414015	407496	399439	385092	377639
26	Rioja (La)	321702	317501	308968	306377	301084	293553	287390
27	Lugo	355195	355549	355176	356595	357625	358452	360512

2.-ESTADÍSTICA DESCRIPTIVA I. DISTRIBUCIONES UNIDIMENSIONALES

- 2.1. Tipos de datos. Clasificación de variables estadísticas
- 2.2. Organización de los datos. Tablas estadísticas
- 2.3. Visualización de los datos. Gráficas estadísticas
- 2.4. Distribuciones unidimensionales de frecuencias. Análisis de los datos
  - 2.4.1. Medidas de tendencia y posición
  - 2.4.2. Medidas de variabilidad
  - 2.4.3. Medidas de simetría
  - 2.4.4. Medidas de forma

3. ESTADÍSTICA DESCRIPTIVA II. DISTRIBUCIONES BIDIMENSIONALES

- 3.1. **Distribuciones bidimensionales y de frecuencias.**
- **3.2. Organización de los datos. Tablas de contingencia**
- 3.3. Visualización de los datos. Gráficas de datos bidimensionales
- 3.4. Distribuciones marginales y condicionadas
- 3.5. Dependencia funcional y estadística. Covarianza. Correlación lineal
- 3.6. El modelo de regresión lineal simple. Ajuste por mínimos cuadrados
- 3.7. Otros modelos de regresión
- 3.8. Regresión múltiple



Lineal?

Cuadrática?

Hiperbólica?

Logística?

.....

TEMA 4. CÁLCULO DE PROBABILIDADES

- Sucesos aleatorios. Espacio muestral
- Definiciones de probabilidad. Propiedades
- Probabilidad condicionada. Independencia de sucesos
- Teorema de Probabilidades Totales
- Teorema de Bayes

TEMA 5. MODELOS PROBABILÍSTICOS DISCRETOS

- Variable aleatoria.
- Variables aleatorias discretas.
- Función de cuantía y función de distribución
- Modelos unidimensionales discretos
- Distribución Binomial
- Distribución de Poisson
- Distribución Geométrica
- Distribución Hipergeométrica
- Distribución Binomial Negativa

TEMA 6. MODELOS PROBABILÍSTICOS CONTINUOS

- Variables aleatorias continuas.
- Función de densidad y función de distribución
- La Distribución Normal
- Otros modelos unidimensionales continuos
- Distribución Uniforme
- Distribuciones Gamma y Exponencial

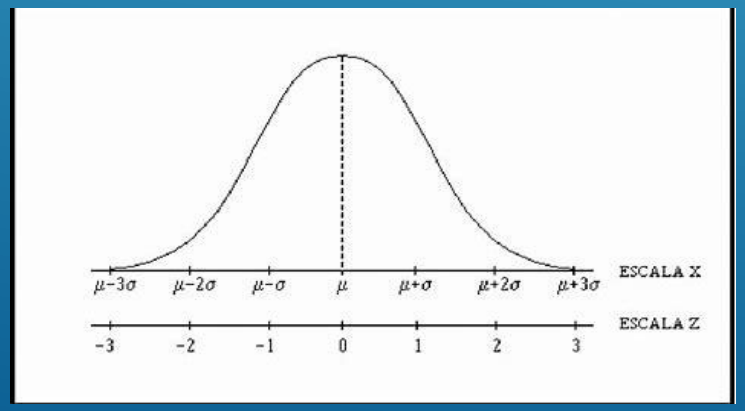
7.- Distribuciones de muestreo fundamentales

- 7.1. Modelos bidimensionales. Distribución conjunta
- 7.2.Distribución Normal multivariante.
- 7.3. Muestreo aleatorio.
- 7.4. El Teorema del límite central
- 7.5.Distribuciones asociadas a poblaciones normales
  - Distribución X^2 de Pearson
  - Distribución t de Student
  - Distribución F de Snedecor
- 7.6. Distribuciones de estadísticos en el muestreo
- 7.5.1. Distribución muestral de medias
- 7.5.2. Distribución muestral de varianzas
- 7.5.3. Distribución muestral de proporciones

INFERIR….CONCLUSIONES SOBRE LA POBLACIÓN

Población

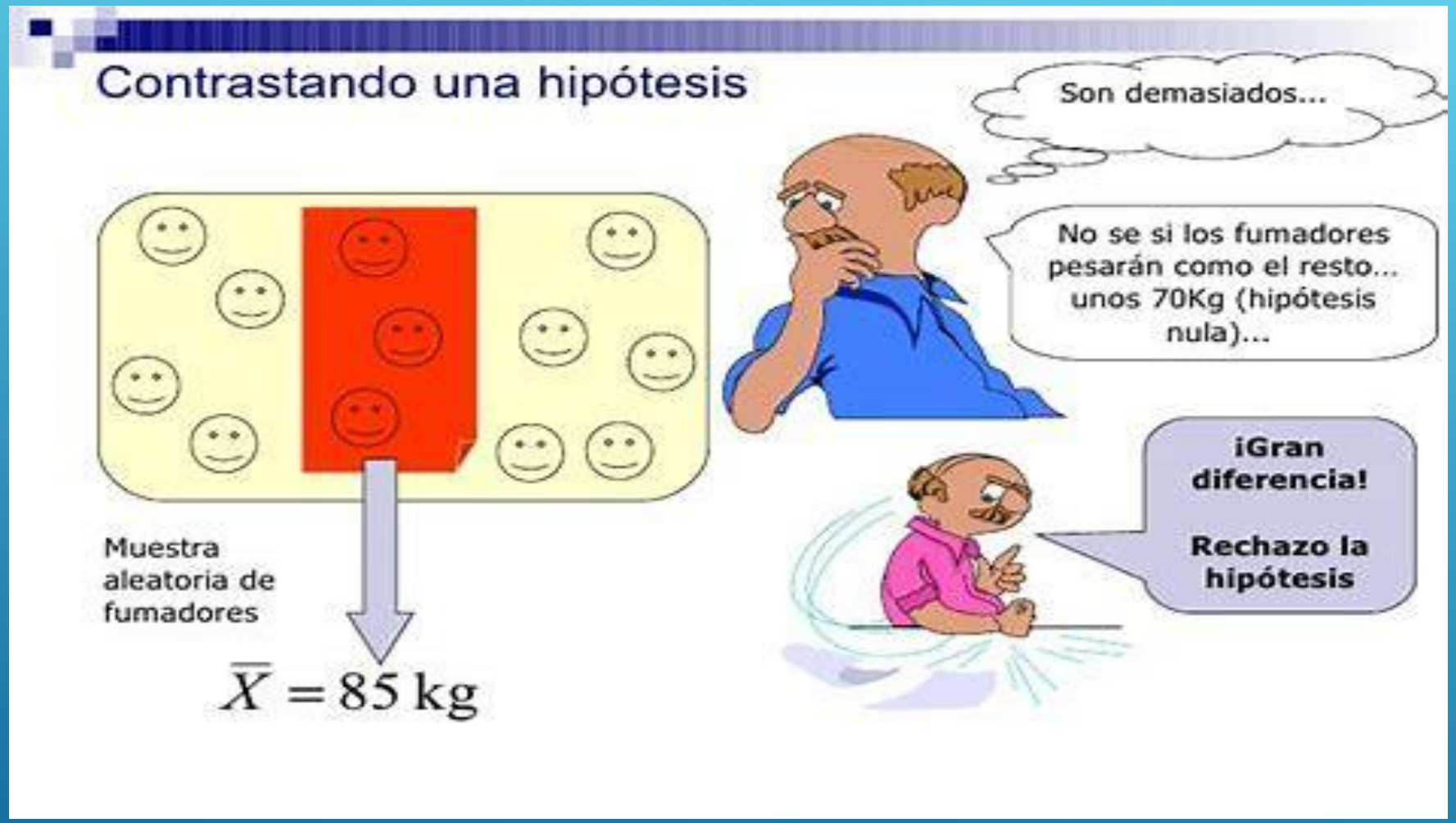
Muestra



TEMA 8. INTRODUCCIÓN A LA INFERENCIA. ESTIMACIÓN

- Estimación puntual y estimación por intervalo
- Error máximo de estimación. Determinación del tamaño muestral
- Estimación de la media de una población normal
- Estimación de la varianza de una población normal
- Estimación de una proporción poblacional
- Estimación de la diferencia entre dos medias
- Estimación de la diferencia entre dos varianzas
- Estimación de la diferencia entre dos proporciones

… Y LOS TESTS DE HIPÓTESIS



TEMA 9. PRUEBAS DE HIPÓTESIS

- Prueba de hipótesis: conceptos generales.
- El p-valor. Aplicación en la toma de decisiones
- Errores de tipo I y II en una prueba de hipótesis. Potencia del test
- Prueba de hipótesis sobre una media poblacional
- Prueba de hipótesis sobre una varianza poblacional
- Prueba de hipótesis sobre una proporción poblacional
- Prueba de hipótesis sobre una diferencia entre dos medias
- Prueba de hipótesis sobre la diferencia entre dos varianzas
- Prueba de hipótesis para la diferencia entre dos proporciones

TÉCNICAS DE MUESTREO

Métodos probabilísticos

Equiprobabilidad: Todos los individuos de la población tienen la misma probabilidad de ser seleccionados para la muestra

1.-Muestreo aleatorio simple (m.a.s.). Es un "sorteo puro y duro": se eligen n individuos de forma totalmente aleatoria

2.-Muestreo aleatorio sistemático. Se elige sólo un individuo i: el resto son i+k, i+2k, …i+(n-1)k

3.-Muestreo aleatorio estratificado. Individuos parecidos dentro del mismo estrato y diferencias importantes entre diferentes "estratos"

4.-Muestreo aleatorio por conglomerados. Tiene ventajas cuando la población es muy grande y dispersa

Métodos no probabilísticos

Sin Equiprobabilidad: los individuos de la población **NO** tienen la misma probabilidad de ser seleccionados para la muestra

1.-Muestreo intencional o de conveniencia. Se elige a los individuos por la mayor facilidad de acceso

2.-Muestreo por cuotas. El encuestador realiza la selección respetando unas cuotas determinadas

3.-Muestreo según el criterio. Se selecciona a los individuos que se "cree" que son más representativos

4.-Muestreo de bola de nieve. Se "encuentra" a unos individuos que a su vez nos conducen a otros

U-Tad

LAS 10 GRANDES TÉCNICAS DE LA CIENCIA DE DATOS

- ***Media aritmética: promedia los valores de los datos de una muestra.*** *Es la primera medida de posición de los datos.*
- ***Desviación típica o standard.*** *Es una medida de la dispersión de los datos*
- ***Cálculo del tamaño de una muestra.*** *Elegir el tamaño óptimo para mejorar la precisión de las estimaciones*
- ***Regresión.*** *Permite identificar tendencias en un fenómeno y predecir el valor de una variable respuesta en función de otras variables explicativas*
- ***Contraste de hipótesis.*** *Permite concluir, a partir de los datos observados de una muestra, si una premisa se puede admitir como cierta en el conjunto de la población.*
- ***Aprendizaje automático:*** *aprender a reconocer automáticamente* ***patrones*** *complejos y tomar decisiones inteligentes basadas en datos.*
- ***Redes neuronales:*** *Técnicas que permiten encontrar patrones no lineales en los datos. Por ejemplo, identificar clientes en riesgo de abandono.*
- ***Aprendizaje de reglas de asociación.*** *Técnicas para descubrir relaciones entre las variables de grandes bases de datos y resolver problemas de optimización a partir de dichas relaciones*
- ***Algoritmos genéticos.*** *Se basan en la "supervivencia del más fuerte". Se utilizan por ejemplo para optimizar el rendimiento de una cartera de inversiones o para optimizar una programación de tareas*
- ***Análisis de series temporales****. Para explicar el comportamiento futuro de un fenómeno a partir de datos del pasado.*

U-tad