Análisis estadístico en diseño experimental

La estadística es la primera de las ciencias inexactas
Edmond Gouncourt

La estadística es una ciencia que demuestra que si mi vecino tiene dos coches y yo ninguno, los dos tenemos uno
George Bernard Shaw

Las cosas complejas y estadísticamente improbables, son por naturaleza más difíciles de explicar que las cosas simples y estadísticamente probables.
Richard Dawkins

Hay tres clases de mentiras: las mentiras, las malditas mentiras y las estadísticas.
Mark Twain

¿Qué significa “ESTADÍSTICA”?

La palabra “estadística” tiene varios significados:

1. **Es usada frecuentemente al referirnos a datos registrados.**
2. **Denota características calculadas para un grupo de datos, (media, mediana…)**
3. **Técnicas y procedimientos para el diseño de experimentos (colección, organización, análisis de la información contenida en un grupo de datos) con la finalidad de hacer inferencias acerca de los parámetros de la población.**

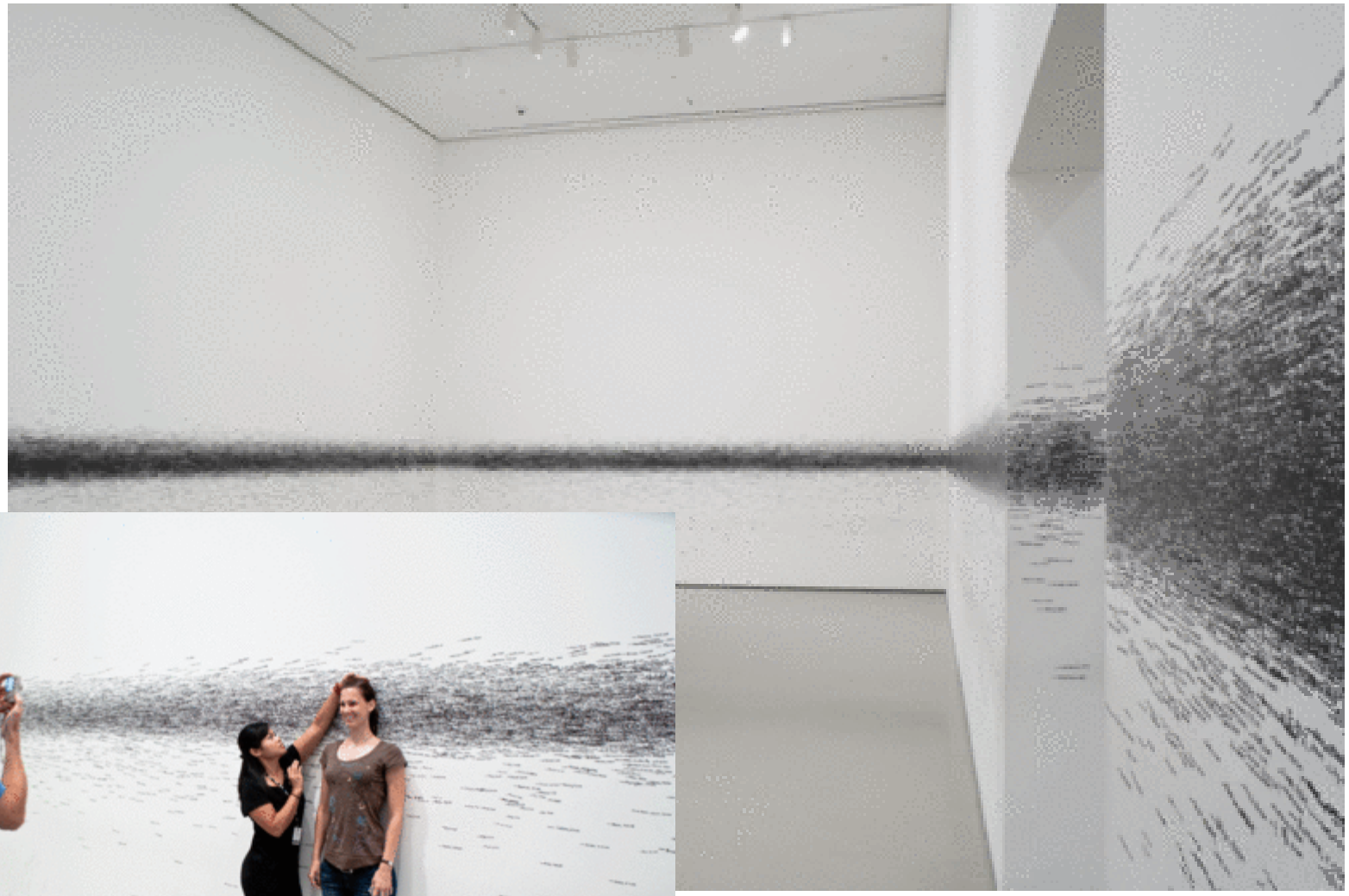
¿Qué hacen los estadísticos?

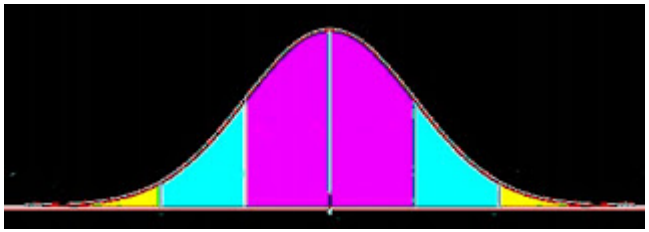
1. **Guiar el diseño de un experimento o encuesta antes de la colección de datos.**

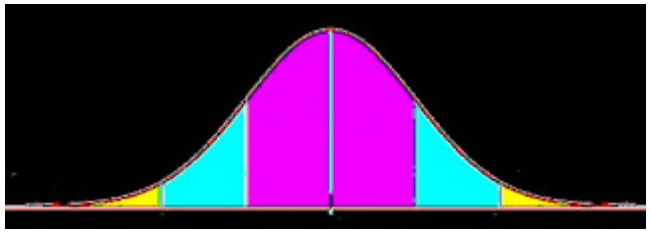
2. **Analizar datos usando los procedimientos y técnicas estadísticos adecuados.**

3. **Presentar e interpretar resultados a los investigadores u otros (instituciones, empresas…)**

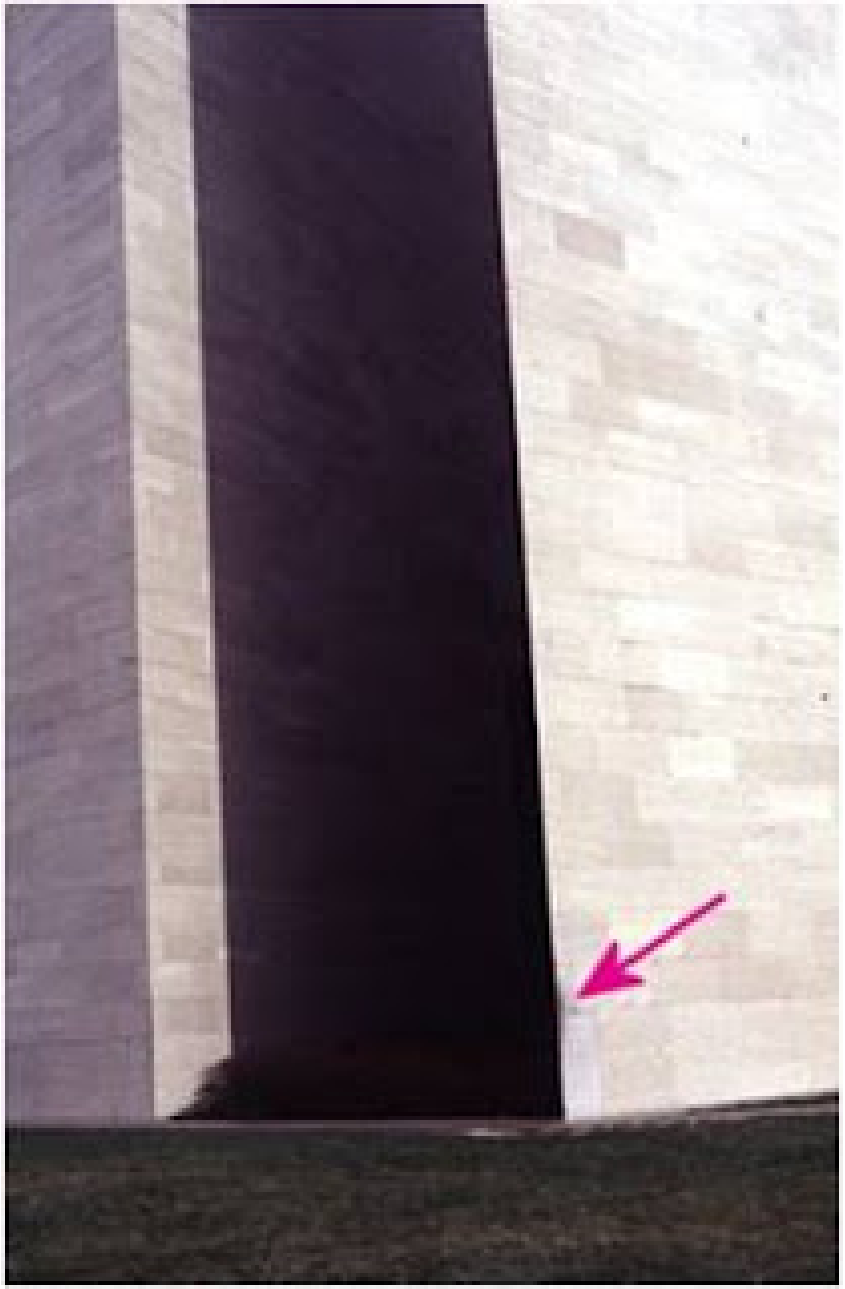
Measuring the Universe (2007), by Roman Ondák



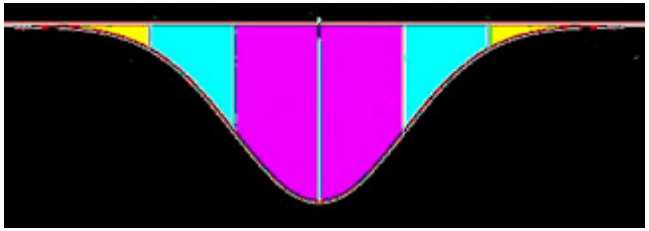




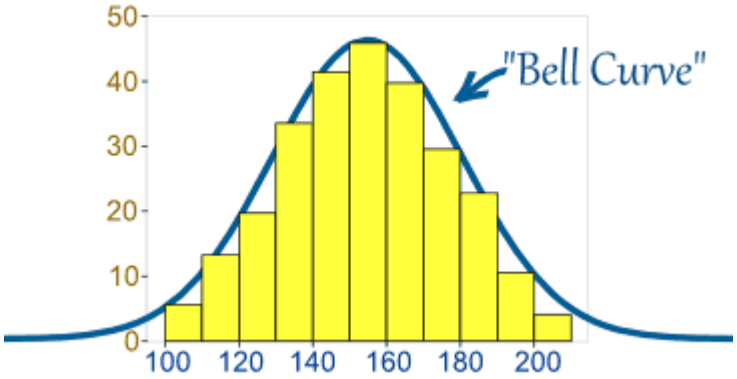
The National Gallery of Art, Washington, USA



Wells Cathedral, England



Marcas en las puertas de los servicios en un bar.
¿Dónde entrarías tú?



Desviación estandar de la población

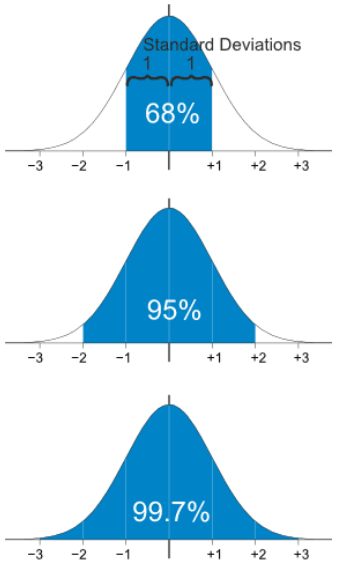
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

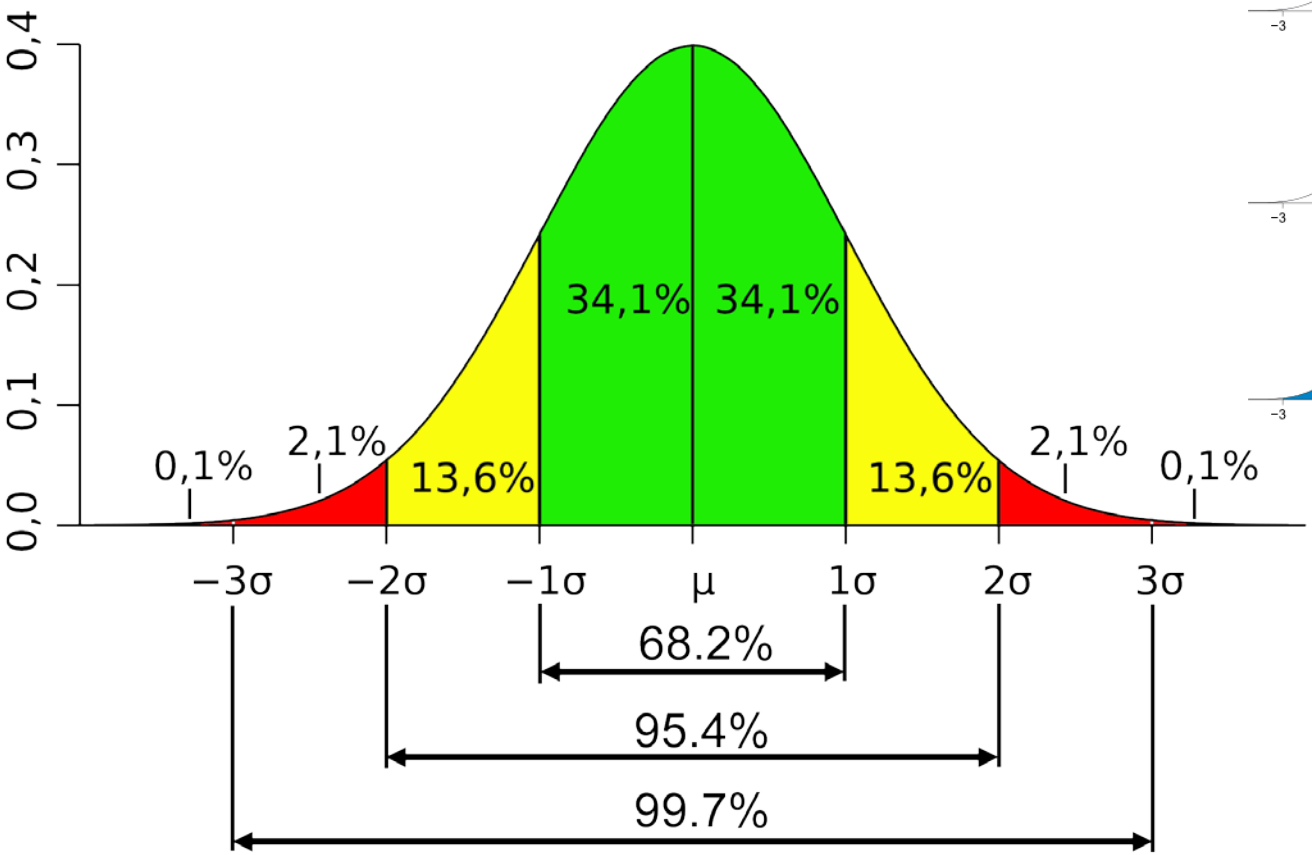
Desviación estandar de la muestra

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Corrección de Bessel.

La **desviación estandar** y la **distribución normal**





Distribución normal: N(μ, σ^2).
Es **simétrica** respecto de su media, μ;
La **moda y la mediana son iguales a la media,** μ.
Los **puntos de inflexión** de la curva se dan para x = μ − σ y x = μ + σ.

Desviación estandar de la población

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Error estandar de la población

$$SE = \frac{\sigma}{\sqrt{n}}$$

Desviación estandar de la muestra

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Error estandar de la muestra

$$SE = \frac{s}{\sqrt{n}}$$

Muestra A

95	104	95	113
120	117	111	101
107	84	91	103
113	108	109	98
102	109	109	97
99	96	120	114
95	94	116	101
92	92	108	80
84	96	102	87
104	101	88	101
117	86	116	85
89	94	99	94
97	120	114	102
85	87	105	114
88	97	119	100
114	110	104	100
109	92	102	103
102	101	100	101
99	93	99	100
109	86	86	81
104	110	103	108
103	105	93	109
94	123	110	102
98	119	103	88
91	98	114	100

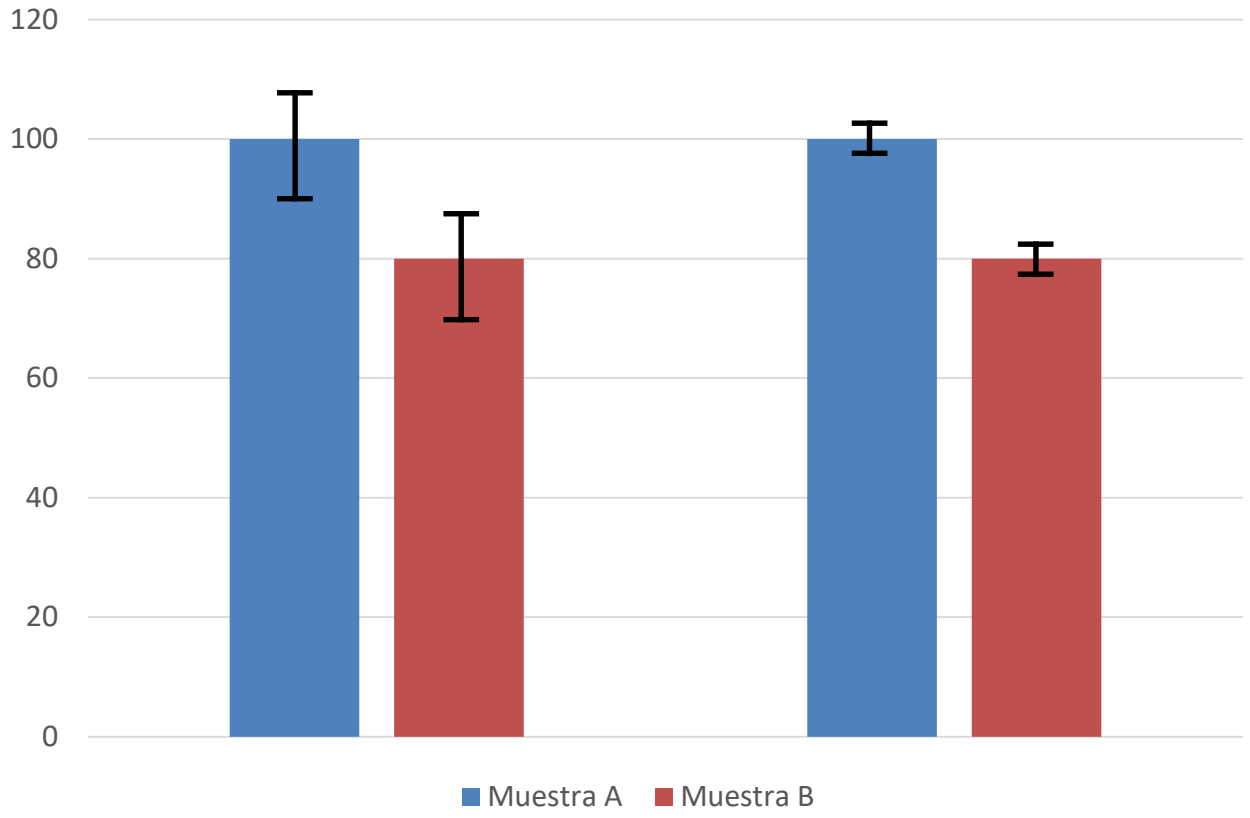
Muestra B

78	67	73	83
78	84	76	89
86	63	73	72
74	64	80	81
65	61	85	93
77	76	69	77
75	83	62	103
60	91	64	59
62	87	75	79
74	86	90	77
96	74	86	87
82	66	79	81
87	78	67	85
70	90	90	82
85	92	88	101
92	81	96	75
72	98	66	76
78	91	89	88
78	76	87	59
78	79	81	75
59	88	67	88
65	77	92	90
81	82	67	73
72	84	70	83
81	91	78	77

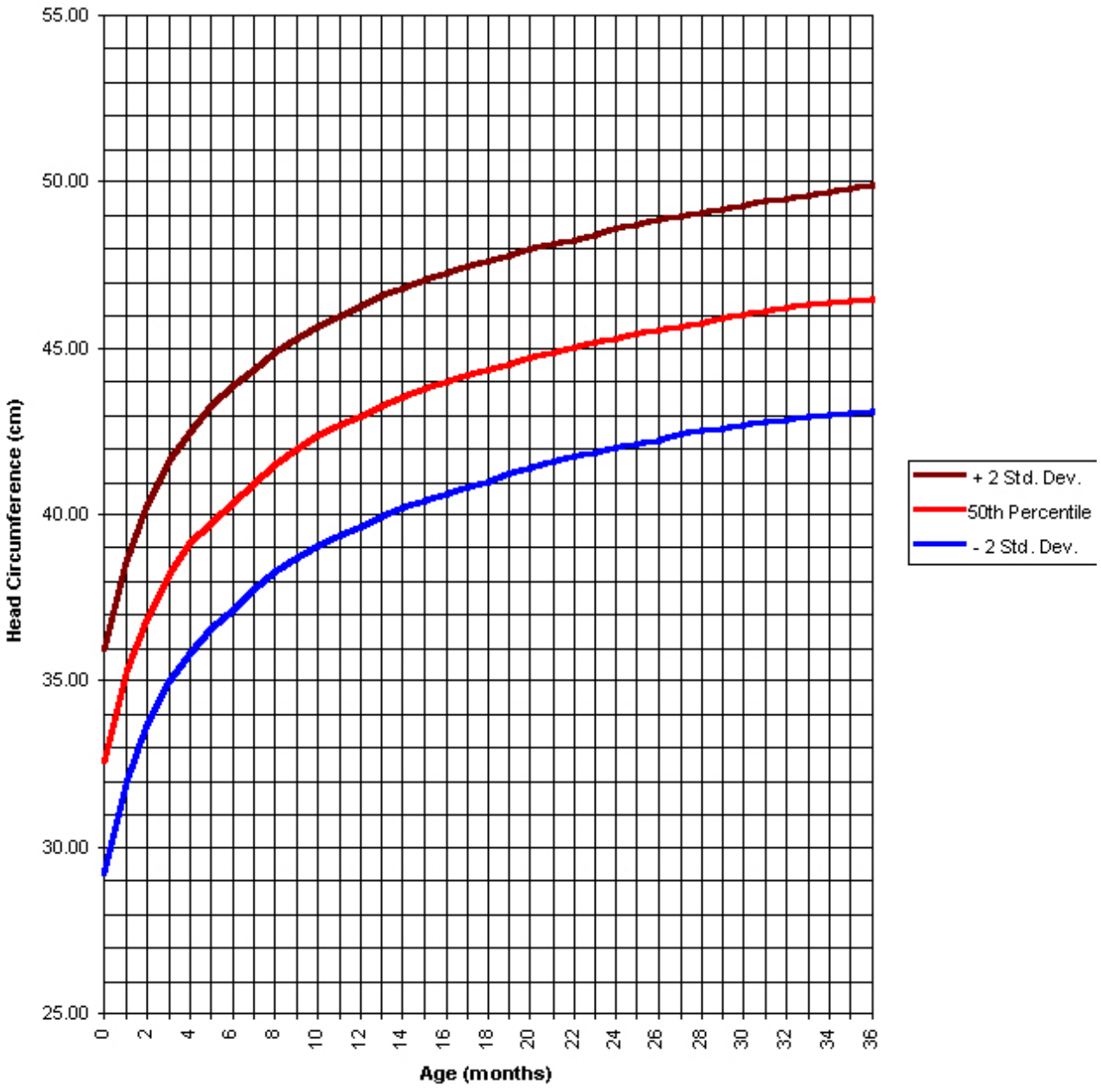
Media	100,1	78,79
Standard Deviation (of sample):	10,0378	10,0488
Standard Error (of sample):	1,0038	1,0049
N	100	100

Desviación estandar

Error estandar

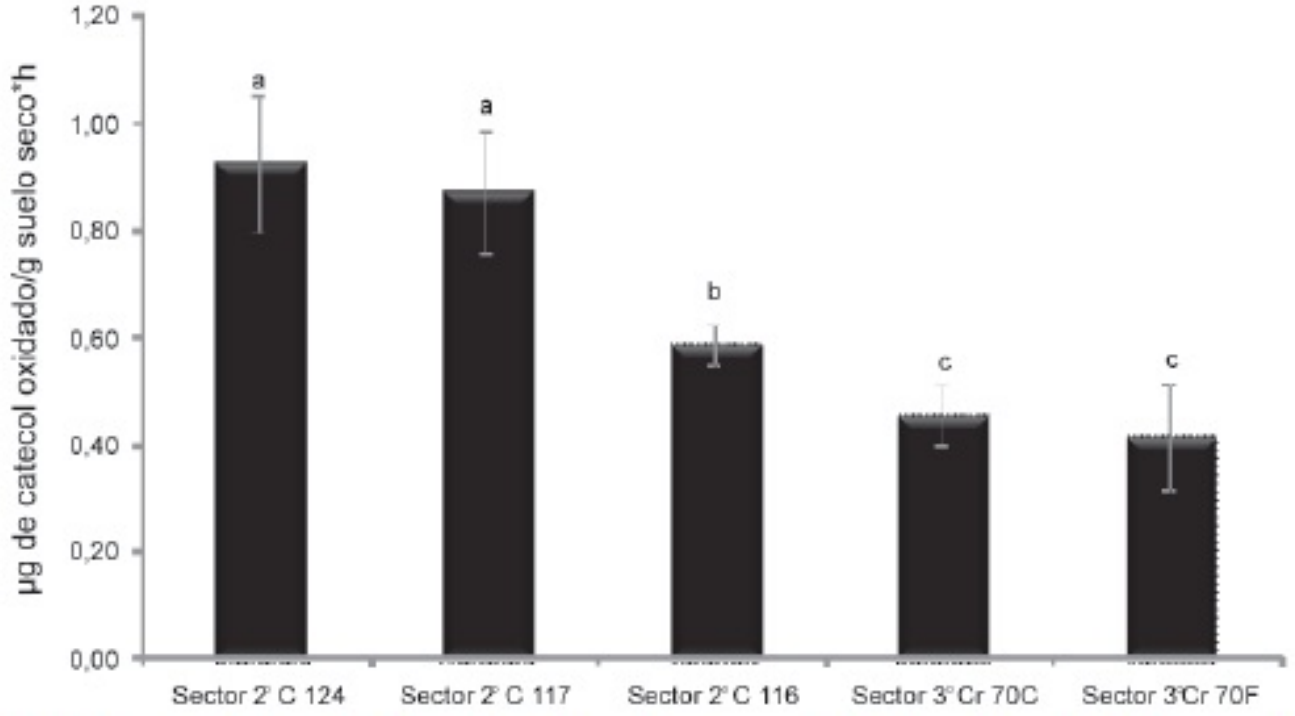


Media	100,1	78,79
Standard Deviation (of sample):	10,0378	10,0488
Standard Error (of sample):	1,0038	1,0049
N	100	100


55.00
50.00
45.00
40.00
35.00
30.00
25.00
Head Circumference (cm)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
Age (months)
+ 2 Std. Dev.
50th Percentile
- 2 Std. Dev.

Ver artículo en Scielo

Cerón Rincón, Laura Emilia, & Ramírez Valencia, Eduardo. (2011). ACTIVIDAD MICROBIANA EN SUELOS Y SEDIMENTOS EN EL SISTEMA CÓRDOBA JUAN AMARILLO, BOGOTÁ D.C.. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, *35*(136), 349-361.

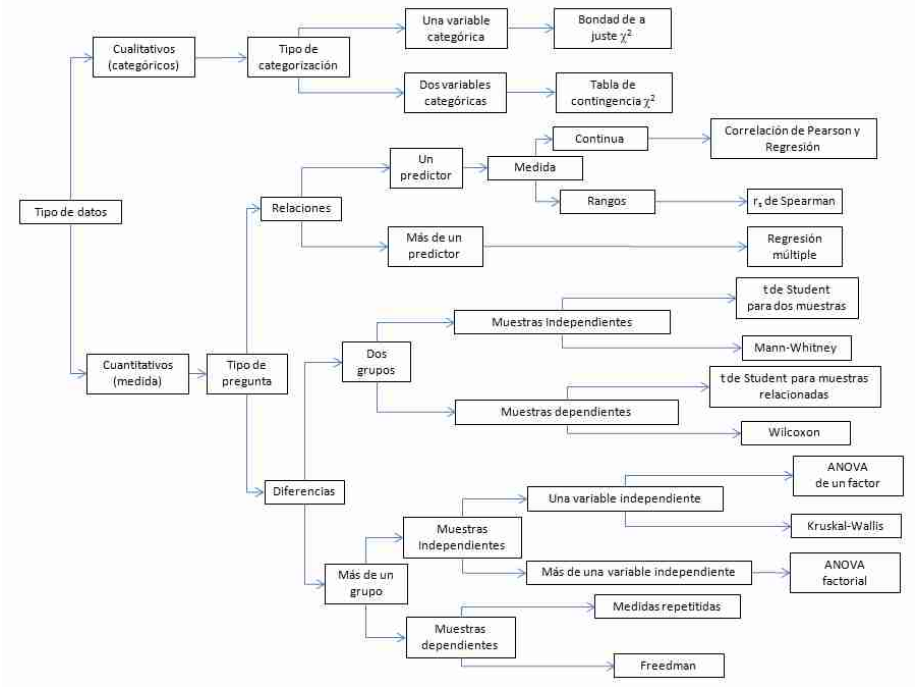


Gráfica 7. Actividad o-difenol oxidasa (µmol catecol oxidado/g suelo seco*h) en suelos aledaños al Humedal de Córdoba. Las columnas corresponden al promedio de tres réplicas y las barras a su desviación estándar, letras iguales indican que no hay diferencias significativas.

Comparación de poblaciones con distribución normal

2 poblaciones $\rightarrow$ T de student

>2 poblaciones $\rightarrow$ ANOVA (*Análisis of Variance*)



http://stats.testak.org/?flujograma

There was a significant increase in production of alcohol between
experiment (M = 8.7, SD = 3.1) and
control (M = 3.2, SD = 1.5),
$t(52) = 4.8$, $p < .001$.

Técnicas inmunológicas en microbiología

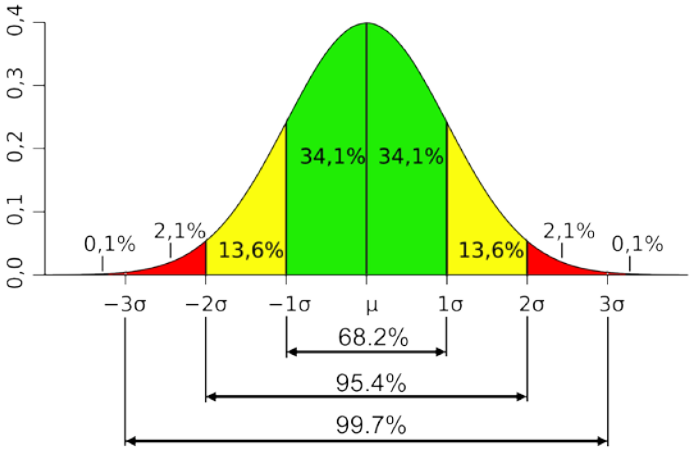
ENZIMOINMUNOENSAYO (E.L.I.S.A.)

RESULTADOS

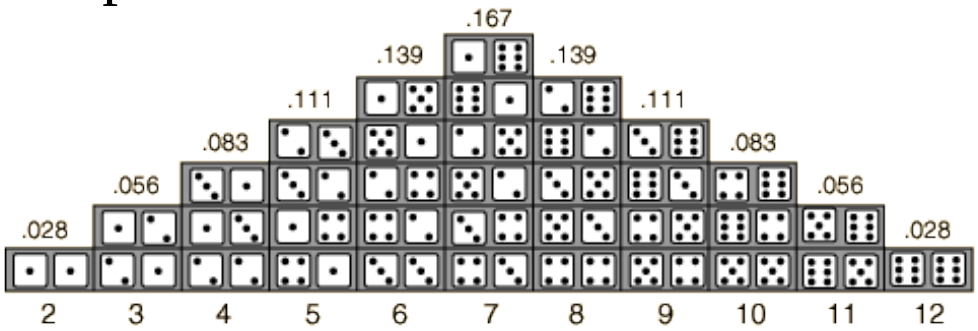
Leer las placas en un espectrofotómetro a 405 nm.

A partir de los valores obtenidos para los **3 controles negativos** calcularemos el cut off o punto de corte, valor por encima del cual consideramos que una muestra es positiva.

cut off = media + 3SD



36 posibles combinaciones

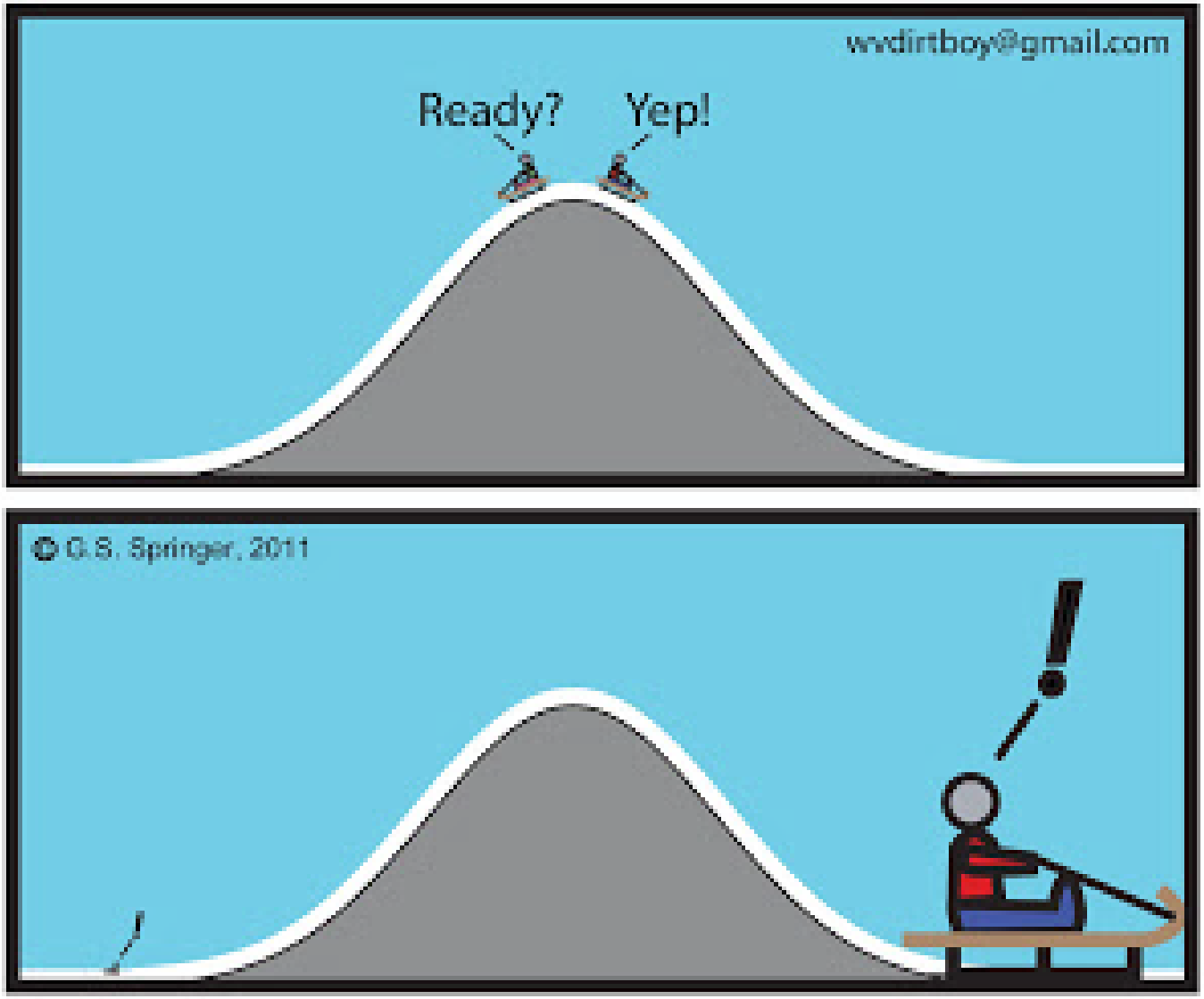


Descripción de diversos parámetros para toda la población

Cuando solamente se tiran los dados 10 veces (una muestra) ¿los parámetros serán diferentes a los anteriores?

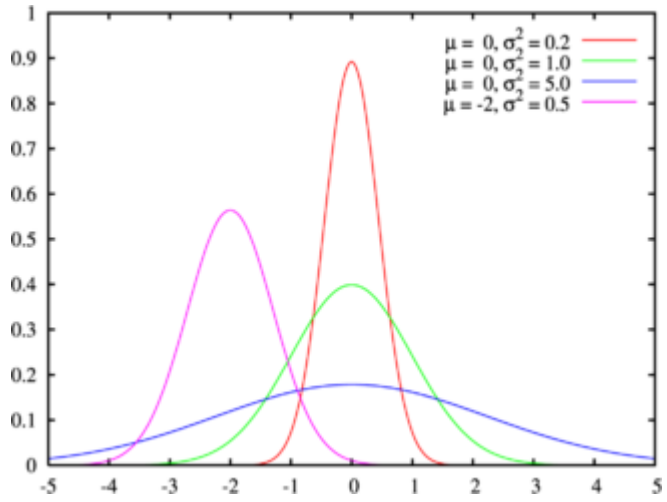
http://stats.testak.org/index.php?tirar_dados

¿Es una distribución normal?


wwdirtboy@gmail.com
Ready?
Yep!
© G.S. Springer, 2011
!
!

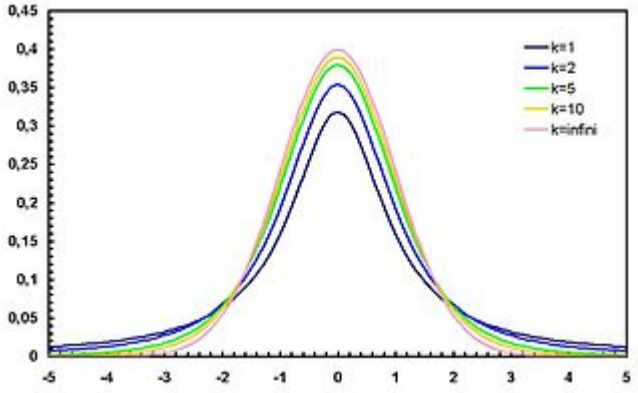
Distribuciones

Distribución normal, distribución de Gauss o distribución gaussiana

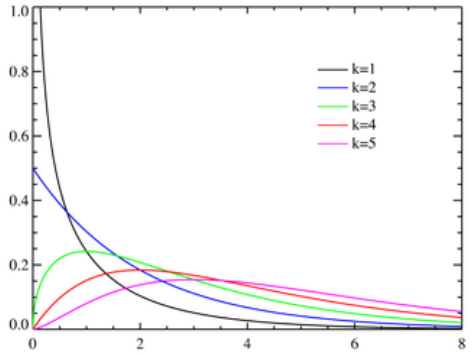


$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\,du$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du, \quad x \in \mathbb{R}$$

Distribución t (de Student)



Distribución de Pearson, ji cuadrado o chi cuadrado (χ^2)



Definición de población y muestra

Una población es un grupo de medidas de interés para un investigador.

Ejemplos:

1. Ingreso de familias viviendo en Karachi
2. Número de niños en familias viviendo en Pakistán
3. Status de salud de adultos en una comunidad

Un subgrupo de la población es llamado muestra.

Una muestra es usualmente seleccionada de tal forma que es representativa de la población.

Estadística descriptiva e inferencial

1. **Estadística descriptiva trata con la enumeración, organización y representación gráfica de los datos**
2. **Estadística inferencial está interesada en llegar a conclusiones de información incompleta, o sea, generalizado desde la muestra**

Un ejemplo de estadística inferencial incluye el uso de información disponible acerca del status de salud de las personas en una muestra para extraer inferencias acerca de la población de la cual se obtuvo la muestra

Estadística inferencial

El objetivo de la estadística inferencial es hacer inferencias acerca de los parámetros de la población basada en la información obtenida de la muestra.

1. **Estimación (e.g., estimando la prevalencia de hipertensión entre adultos viviendo en Vitoria)**
2. **Probando hipótesis (e.g., probando la efectividad de un nuevo medicamento para reducir los niveles de colesterol)**

Fuentes de datos

Los datos pueden obtenerse de diferentes fuentes:

1. Sistemas de vigilancia (e.g., osakidetza, Carlos III…)
2. **Encuestas** planeadas (Gobierno, universidades, ONG)
3. **Experimentos** (Compañías farmacéuticas)
4. Organizaciones de salud (Grupo de datos administrativos)
5. Sector privado (Bancos, compañías, etc)
6. Gobierno (Todas las agencias gubernamentales)

Diferencia entre encuestas y experimentos

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

(ejemplo: evaluar 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)

Variables cualitativas y cuantitativas

Ejemplos de variables cualitativas son ocupación, sexo, estado civil, etc.

Variables que producen observaciones que pueden medirse, se considera que son variables cuantitativas. Ejemplos de variables cuantitativas son peso, estatura, edad.

Variables cuantitativas pueden clasificarse en discretas o continuas

Tipos de variables

1. **Variables categóricas (e.g., Sexo, estado civil, categoría de ingreso)**
2. **Variables continuas (e.g., edad, ingreso, peso, estatura, tiempo en lograr un resultado)**
3. **Variables discretas (e.g. número de niños en una familia)**
4. **Variables dicotómicas o binarias (e.g., respuesta sí o no)**

Escala de datos

1. **Nominal: estos datos no representan una cantidad (e.g., estado civil, sexo)**

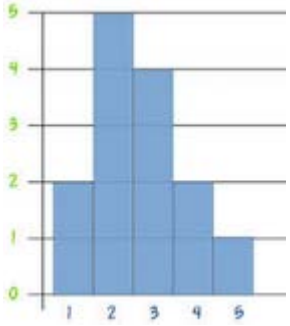
2. **Ordinal: estos datos representan una serie de datos ordenados (e.g., nivel de educación)**

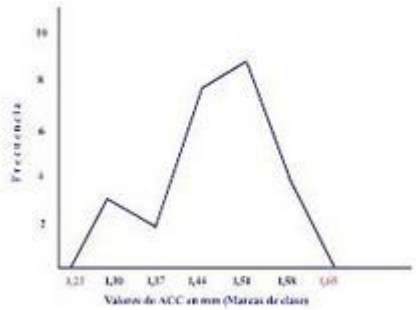
3. **Intervalo: estos datos son medidos en una escala de intervalo teniendo iguales unidades pero teniendo un 0 arbitrario (e.g.: temperatura en ° Fahrenheit)**

4. **Razón de intervalo: variable como peso para el cual podemos comparar significativamente un peso contra otro (digamos, 100 Kg es dos veces 50 Kg)**

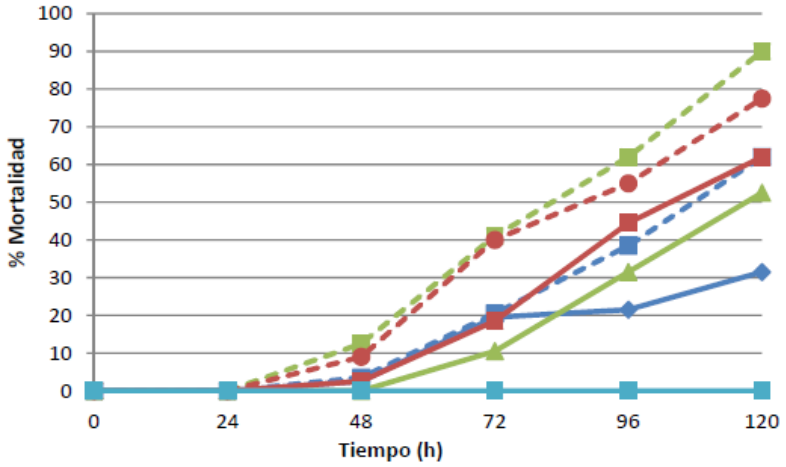
Organizando los datos

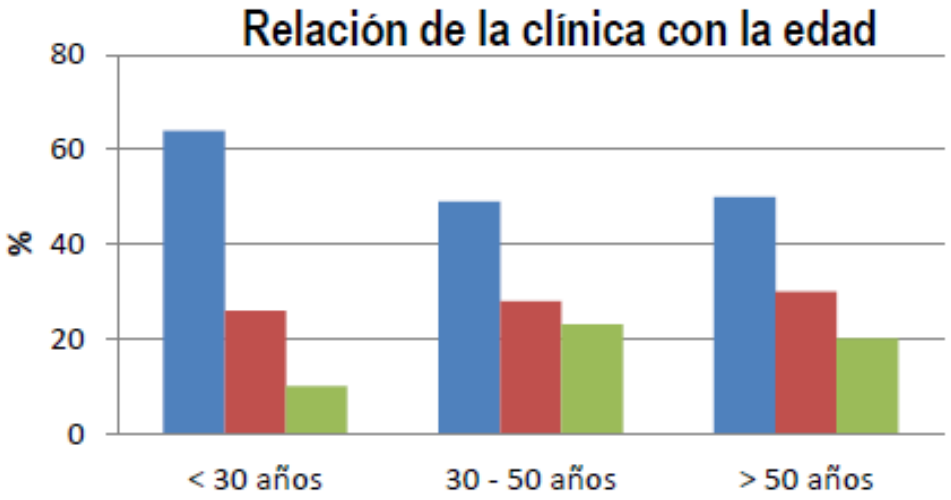
Tabla de frecuencias
Histograma de frecuencias
Histograma de frecuencias relativas
Polígono de frecuencias
Polígono de frecuencia relativa
Barras
Pastel
…



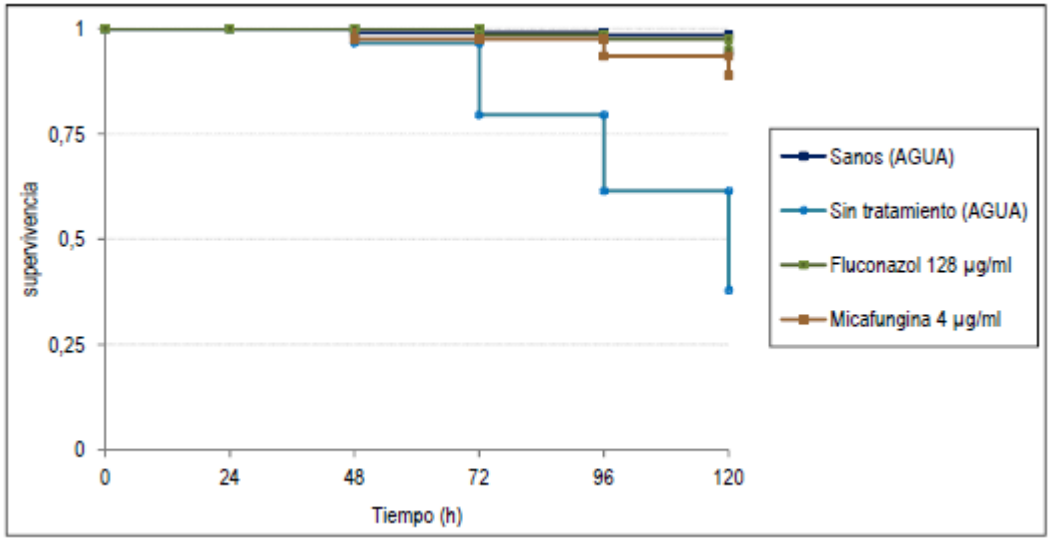


Porcentajes de mortalidad de *X infectados por las diversas especies de microorganismos.*





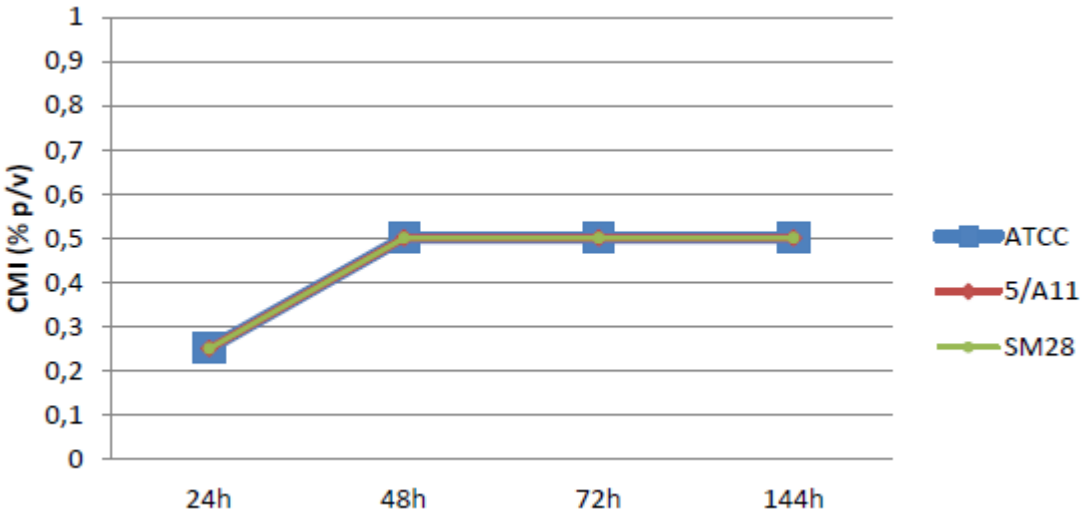
Curvas de supervivencia acumulada



Se han analizado 151 muestras, de las cuales 136 pertenecían a mujeres (90,10%) y 15 a hombres (9,90%). La edad media de las mujeres fue de 30,28 años ± 8,16 años (con un rango de 16 a 72 años) y de los hombres 41,33 años ± 12,40 años (con un rango de 24 a 62 años).

Métodos estadísticos para comparar datos de sensibilidad a antimicrobianos …

Las variables continuas se compararon mediante la prueba t de student o ANOVA ; las variables categóricas se compararon mediante la prueba de la ji cuadrado con la corrección de Yates para las categorías con menos de cinco datos.



Problema

Menores de 20 años		
	Hombre	Mujer
Si fuma	11	3
No fuma	1	7
Entre 20 y 50 años		
	Hombre	Mujer
Si fuma	1	4
No fuma	2	5
Mayores de 50 años		
	Hombre	Mujer
Si fuma	13	3
No fuma	2	6

¿Son independientes la edad y ser fumador?

Hipótesis: Edad y Fumador son independientes.

Se colapsa la tabla anterior

	Si fuma	No fuma
Menores de 20 años	14	8
Entre 20 y 50 años	5	7
Mayores de 50 años	14	8

http://stats.testak.org/programs/chisq.htm

¿Son independientes la edad y ser fumador?

Hipótesis: Edad y Fumador son independientes.

Hombres

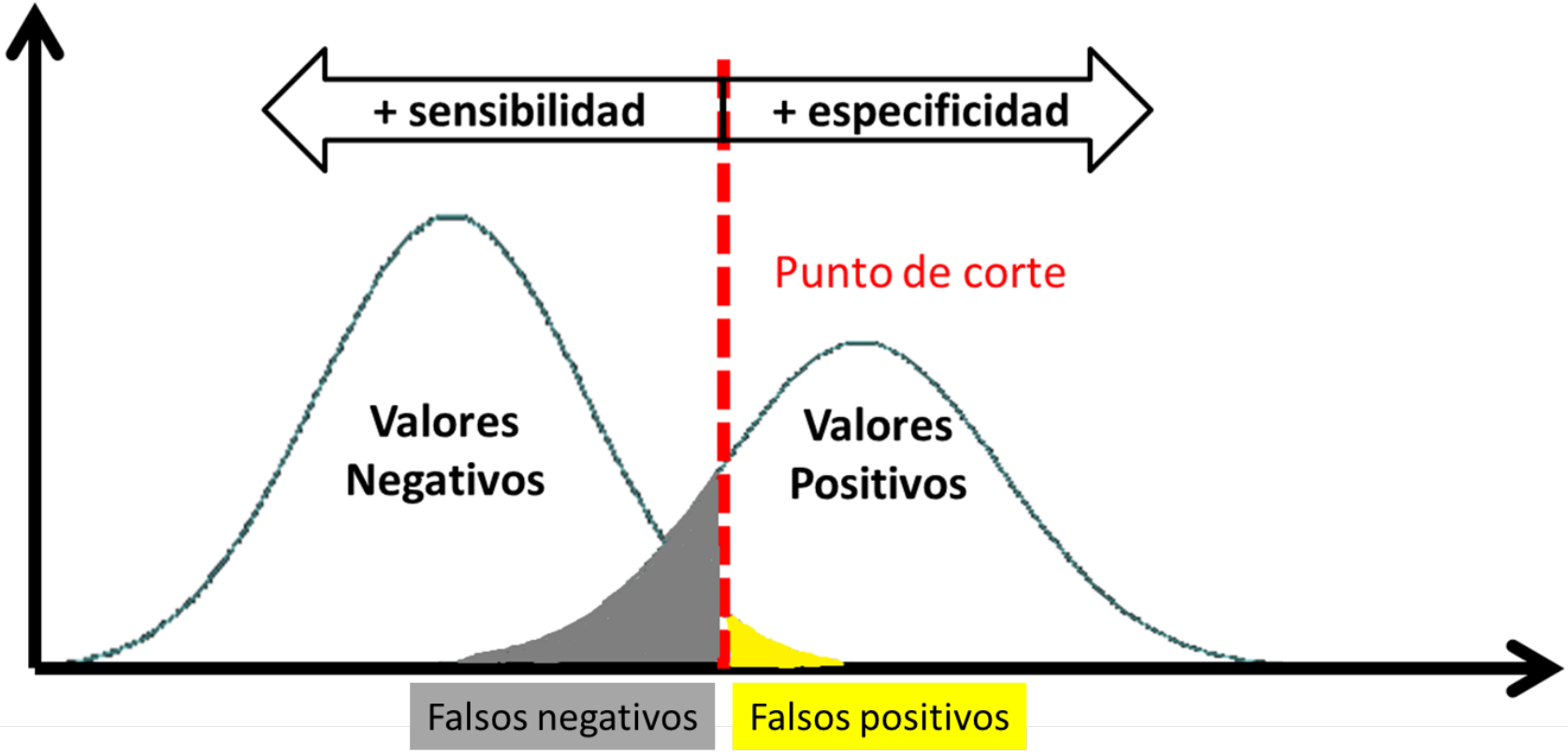
	Si fuma	No fuma
Menores de 20 años	11	1
Entre 20 y 50 años	1	2
Mayores de 50 años	13	2

Mujeres

	Si fuma	No fuma
Menores de 20 años	3	7
Entre 20 y 50 años	4	5
Mayores de 50 años	1	6

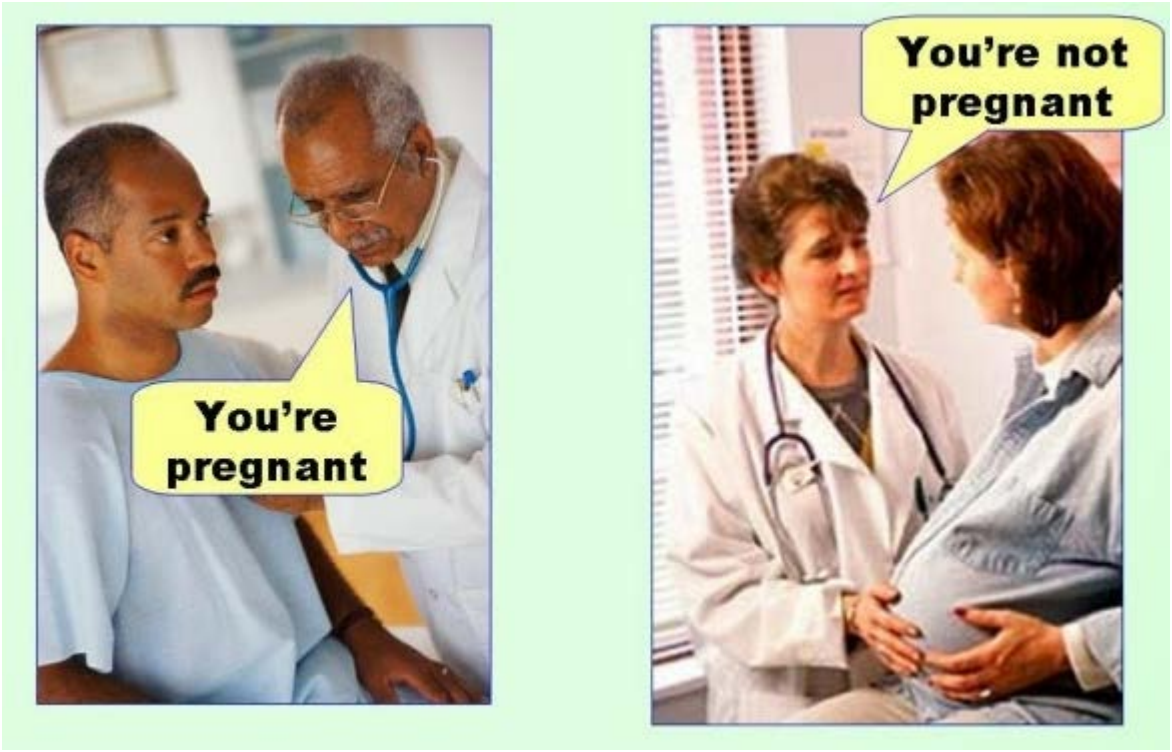
¿Existen independencias condicionadas?

Sensibilidad vs. espeficidad



False positive

False negative



Errores habituales en estadística

Comparar dos poblaciones con T-student

Las poblaciones deben de cumplir lo siguiente

1. Las observaciones deben de ser **independientes** entre si.

2. Las variables deben **ser cuantitativas continuas**.

3. Deben ser **poblaciones distribuidas normalmente**.

4. **Las varianzas deben de ser iguales**.
   (o, en casos especiales deben tener una proporción de varianzas conocidas).

Díaz, 2009
http://dx.doi.org/10.4067/S0717-75182009000400011

Problema

Se estudio el crecimiento de microorganismos mesófilos a partir de la superficie del dedo índice de los alumnos de la escuela de hostelería.

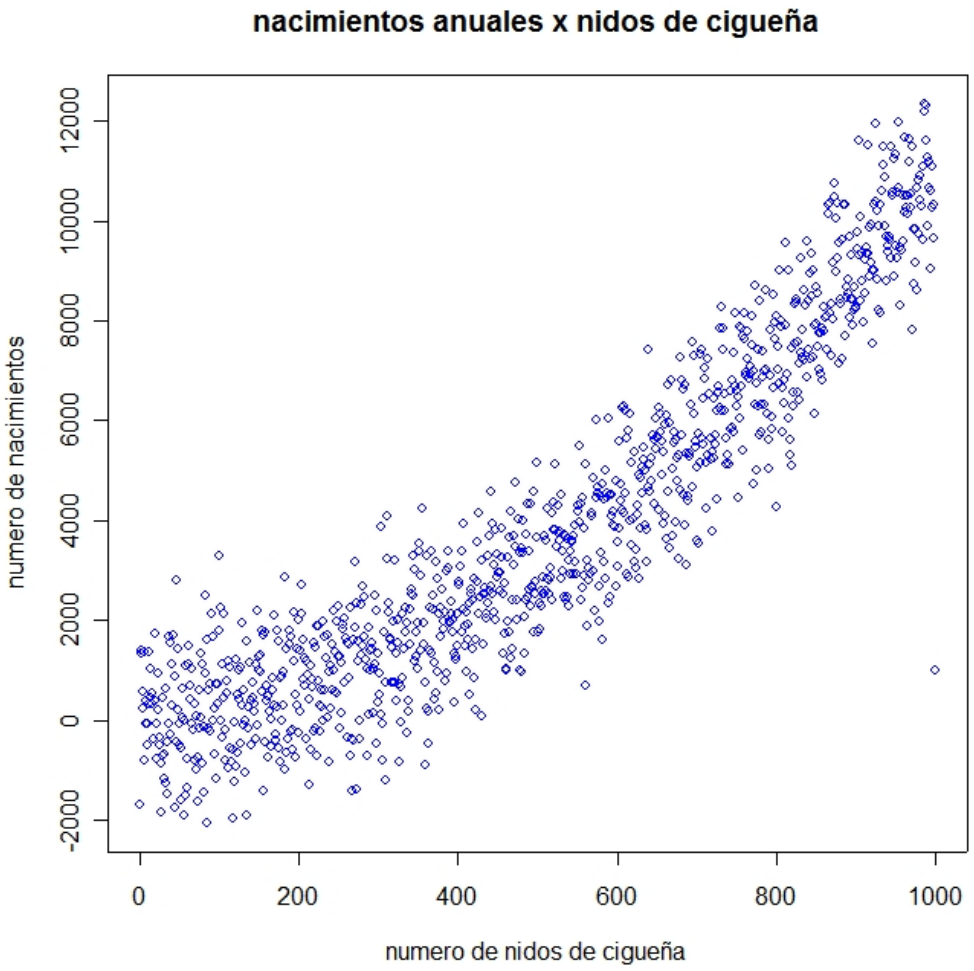
Se registro el número de unidades formadoras de colonias tras 24 h de cultivo.

Los resultados se muestran a continuación:

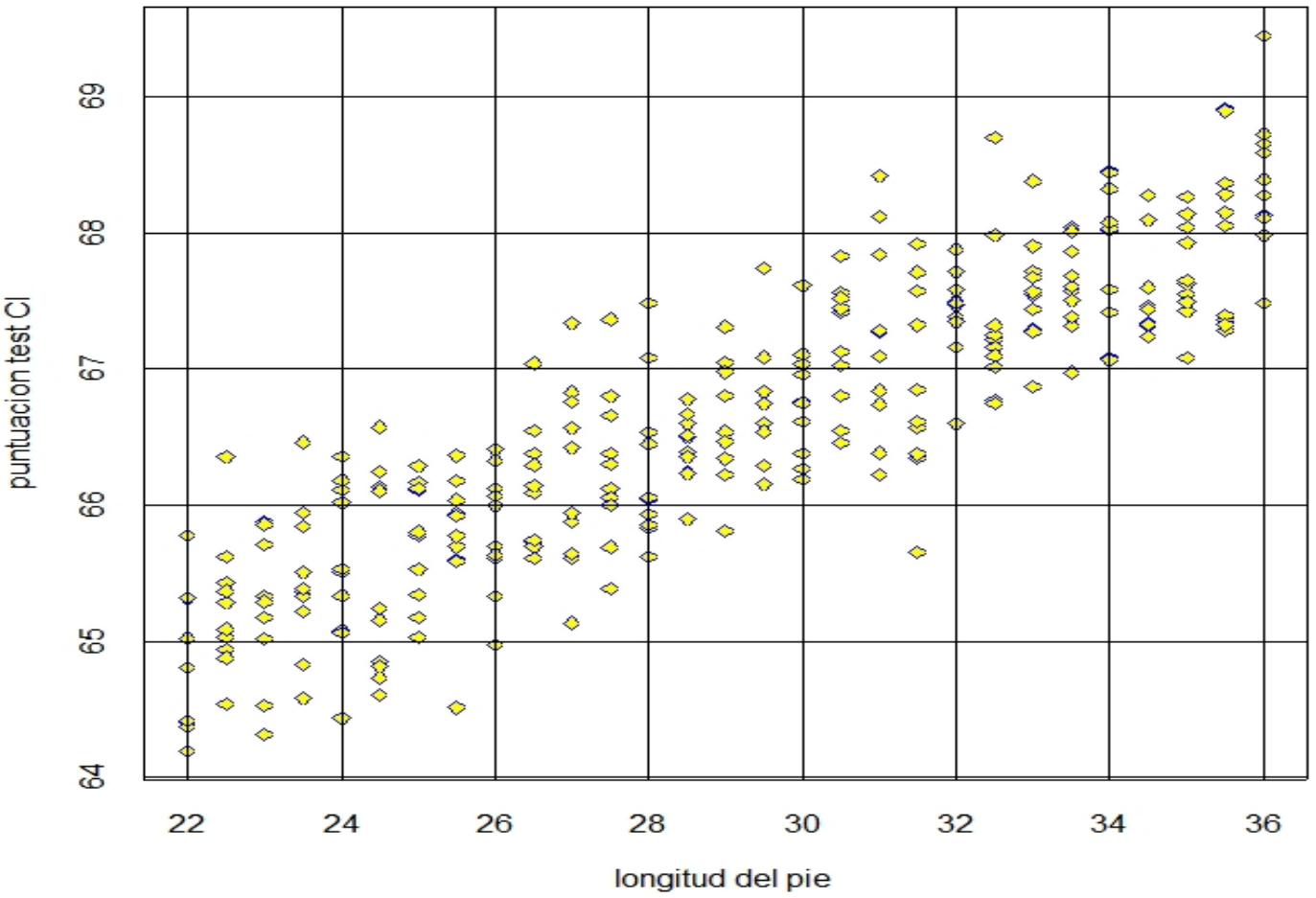
Intervalo al 95% de confianza para la media (ufc):

Límite inferior:	3
Límite superior:	27

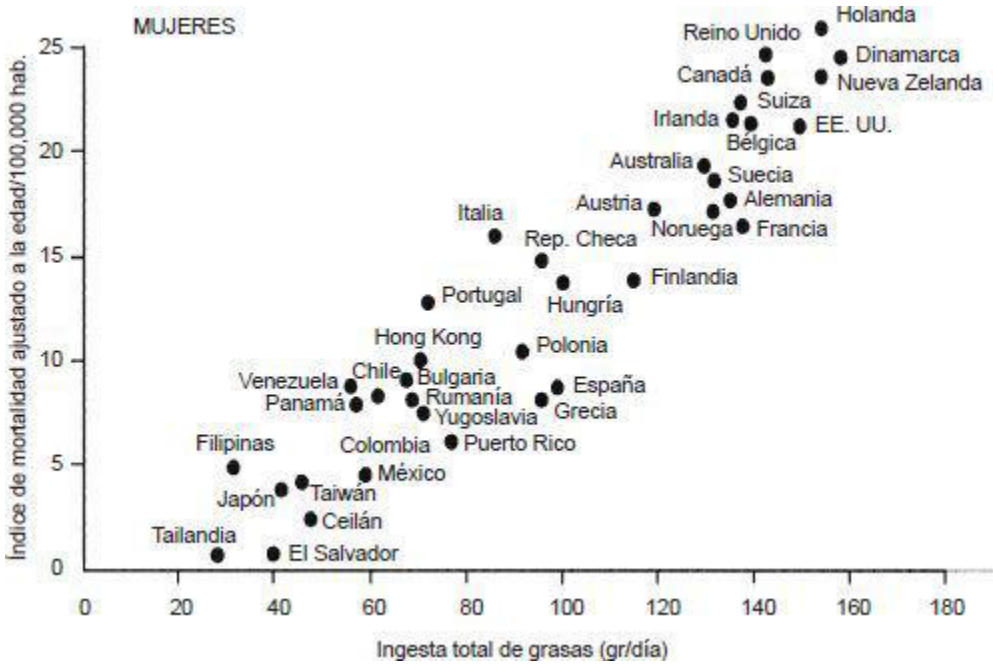
Grafica 22. Número de nidos de cigüeña y número de nacimientos en diversas municipios europeos.

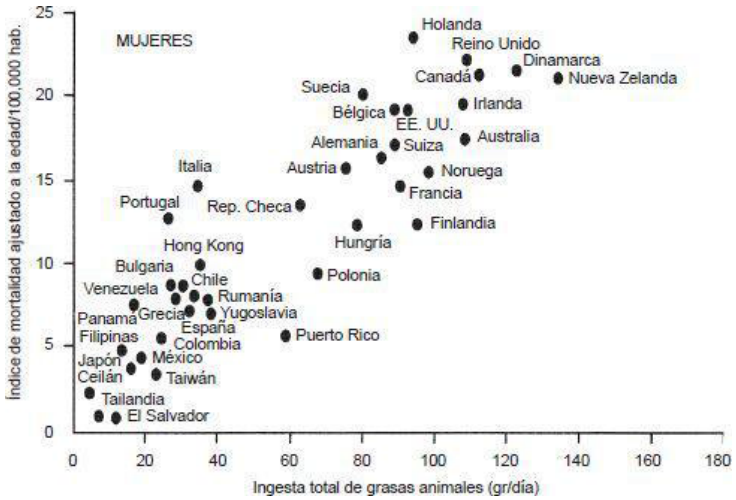


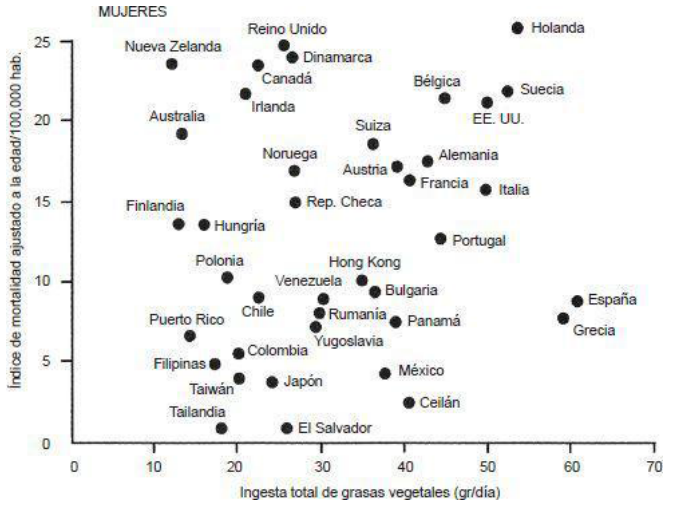
Coeficiente intelectual y número de pie en alumnos de primaria. ¿Hay relación?



¿Correlación entre ingesta de grasas y cáncer de mama?

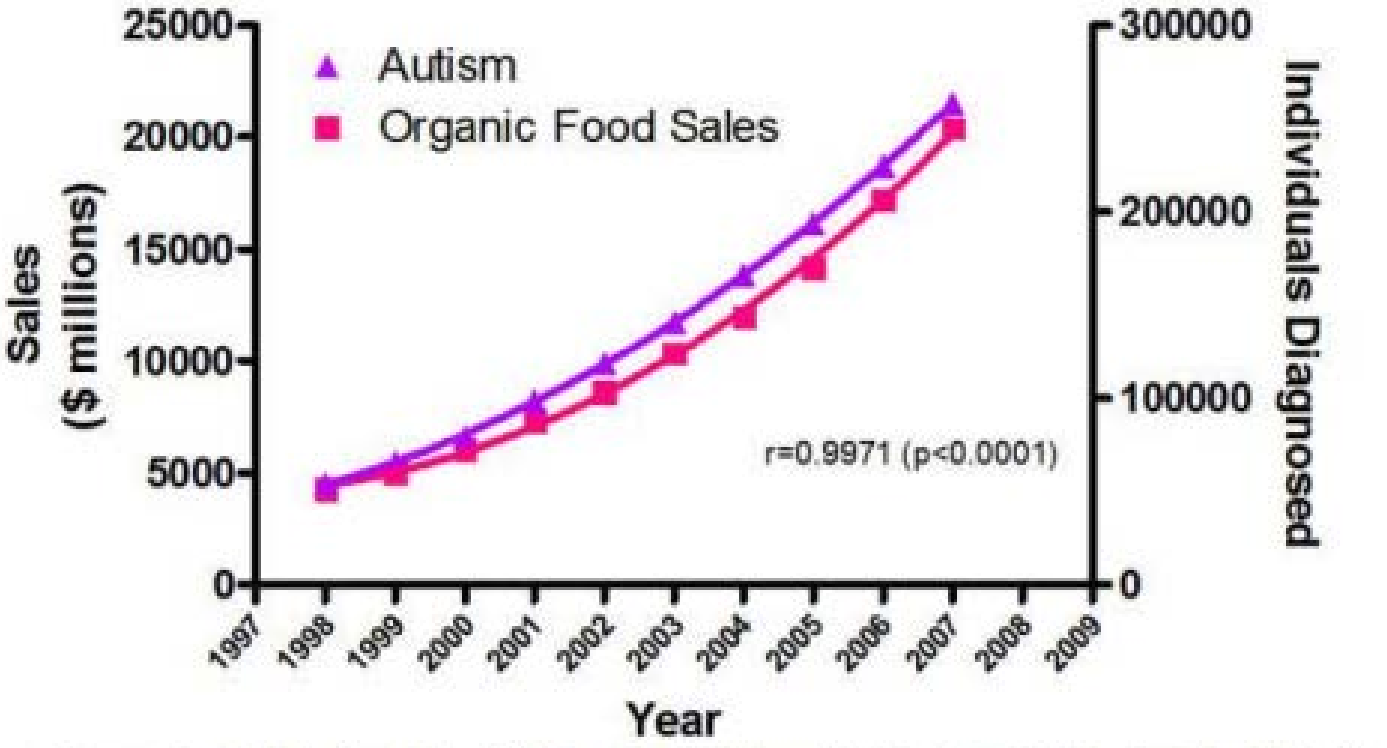






Carrol *et al.*, 1986

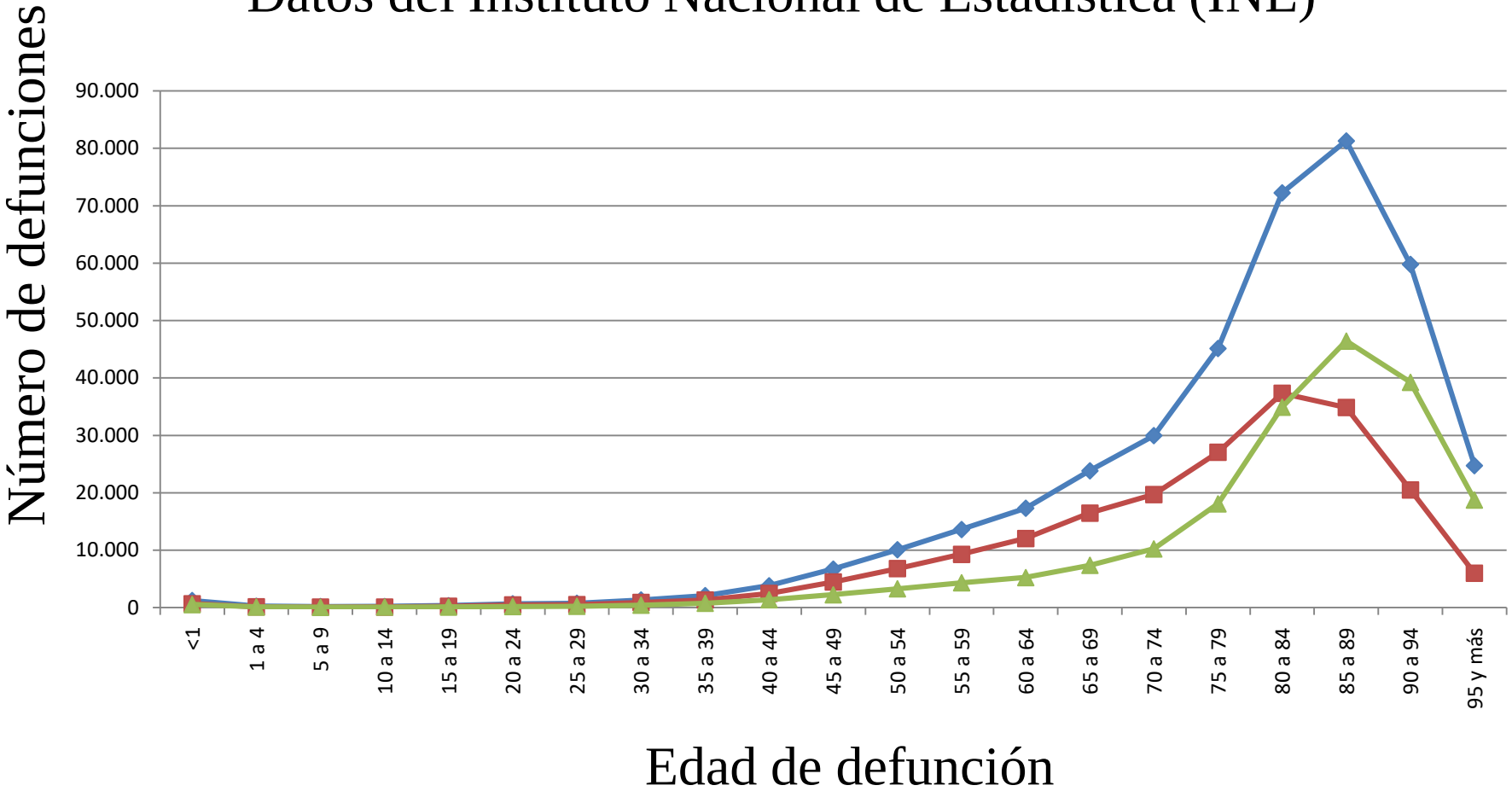
The real cause of increasing autism prevalence?



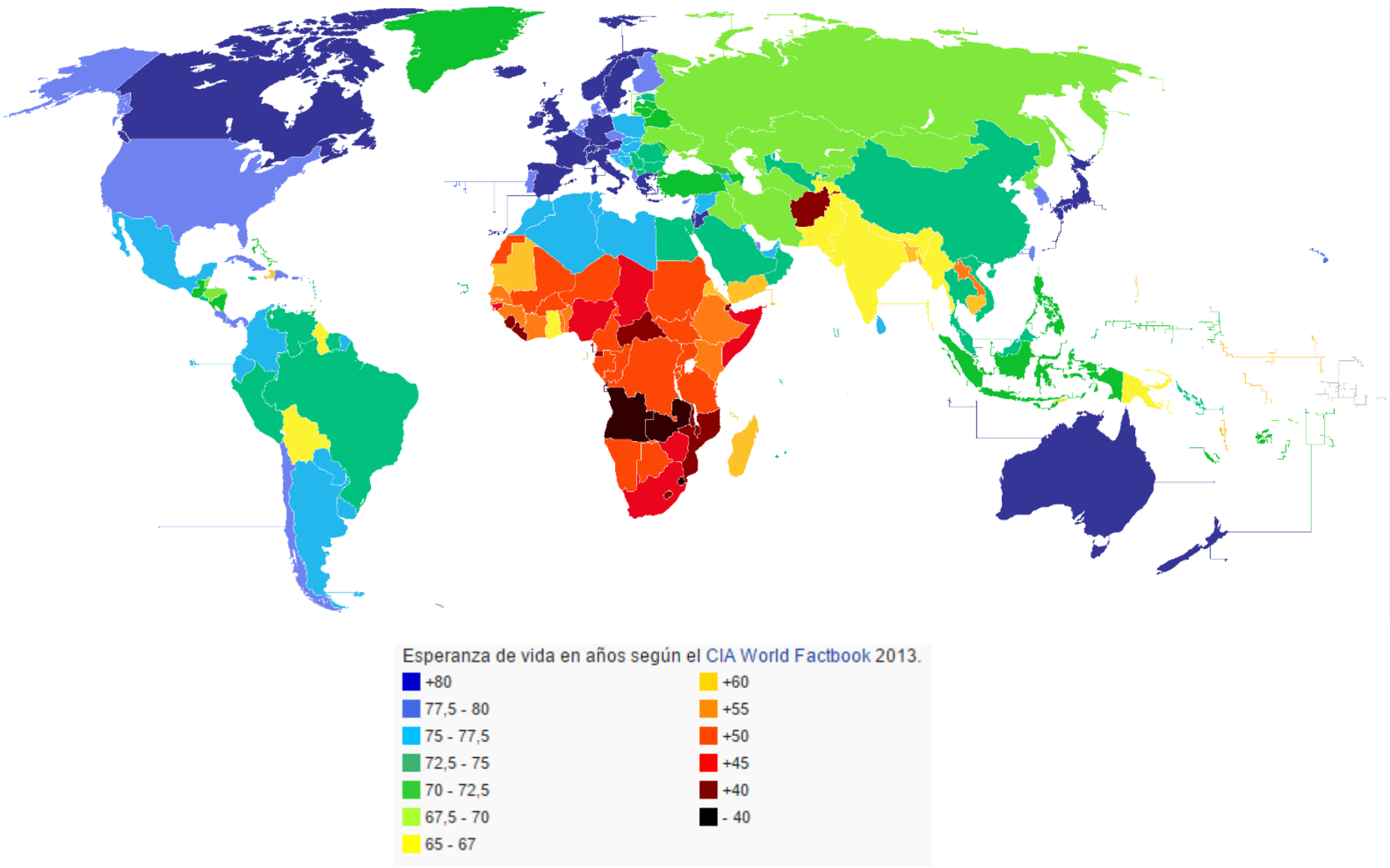
Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

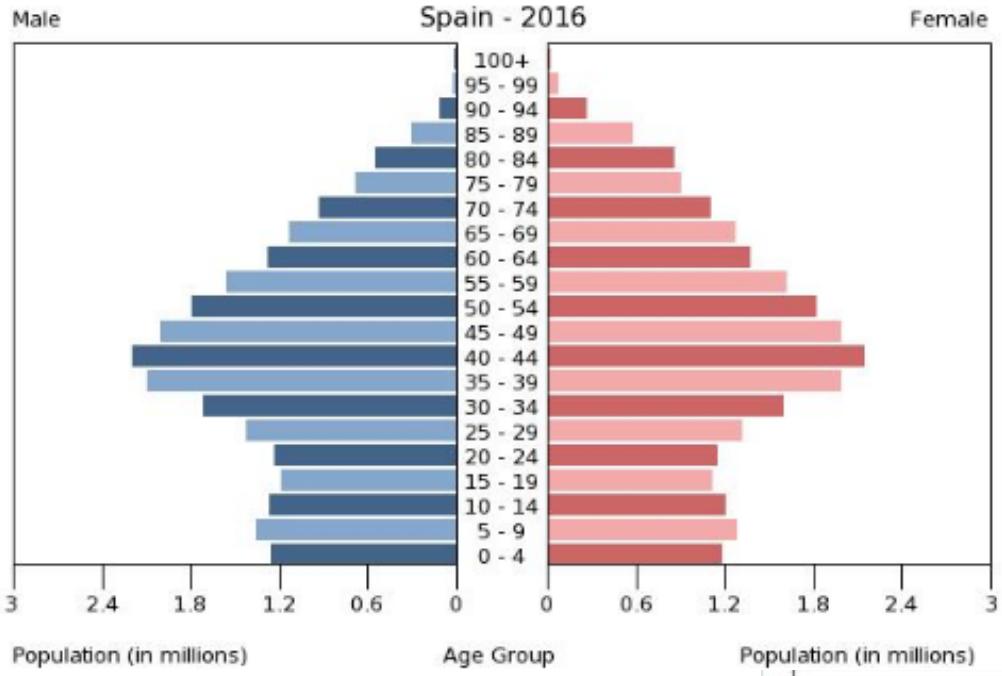
Defunciones según la edad en 2014

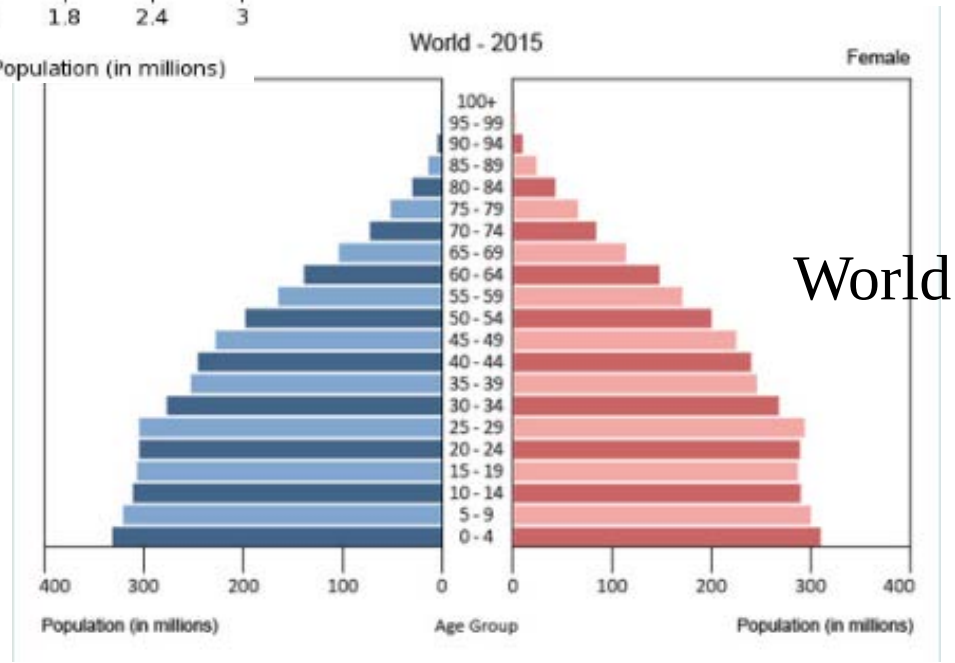
Datos del Instituto Nacional de Estadística (INE)



Gráfica generada a partir de datos del INE

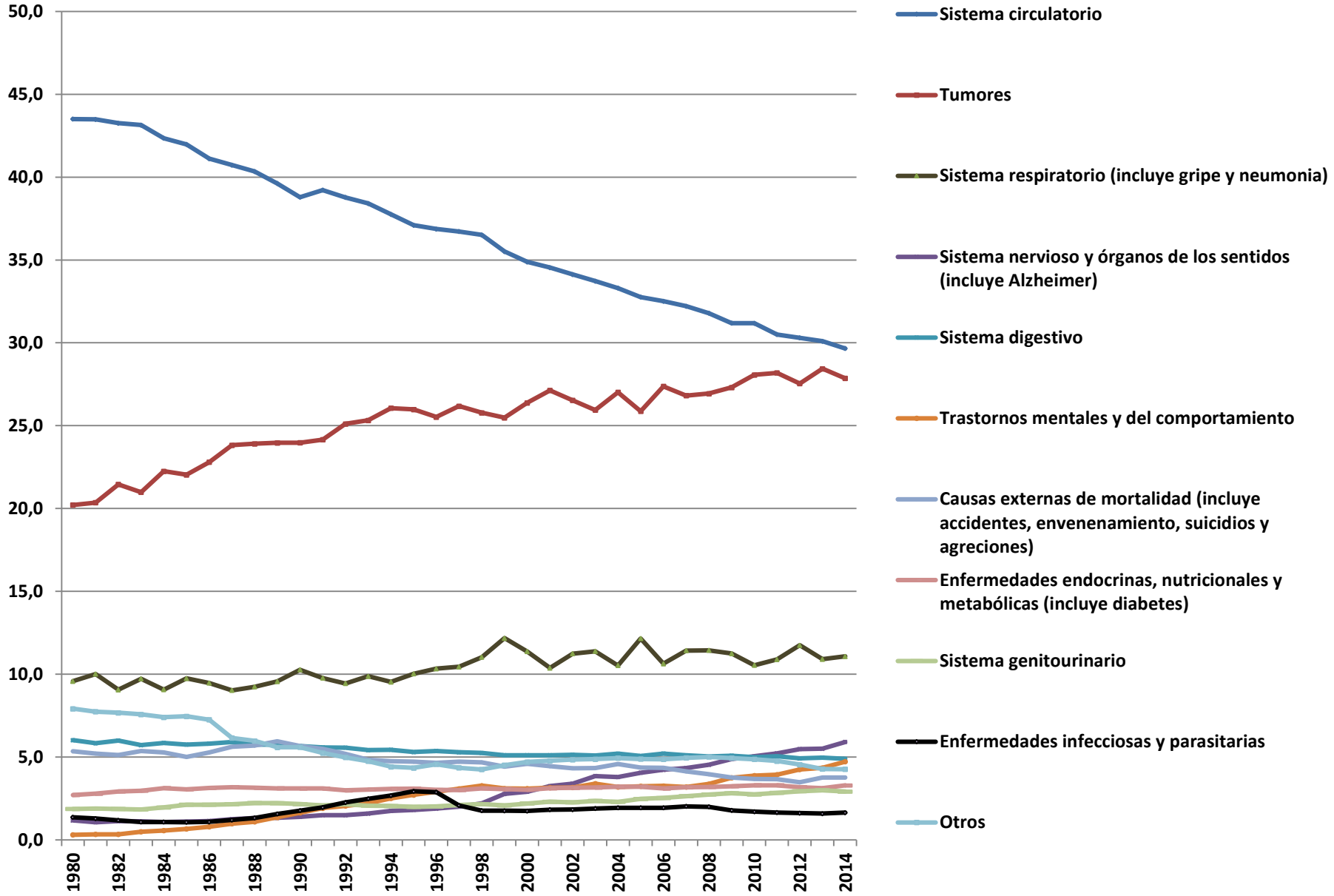

Esperanza de vida en años según el CIA World Factbook 2013.
+80
77,5 - 80
75 - 77,5
72,5 - 75
70 - 72,5
67,5 - 70
65 - 67
+60
+55
+50
+45
+40
- 40





World

Source: https://www.cia.gov/library/publications/resources/the-world-factbook/geos/sp.html



Gráfica generada a partir de datos del INE

Open

European Journal of Clinical Nutrition (2010) 64, 669–677

0954-3007/10
www.nature.com/ejcn

ORIGINAL ARTICLE

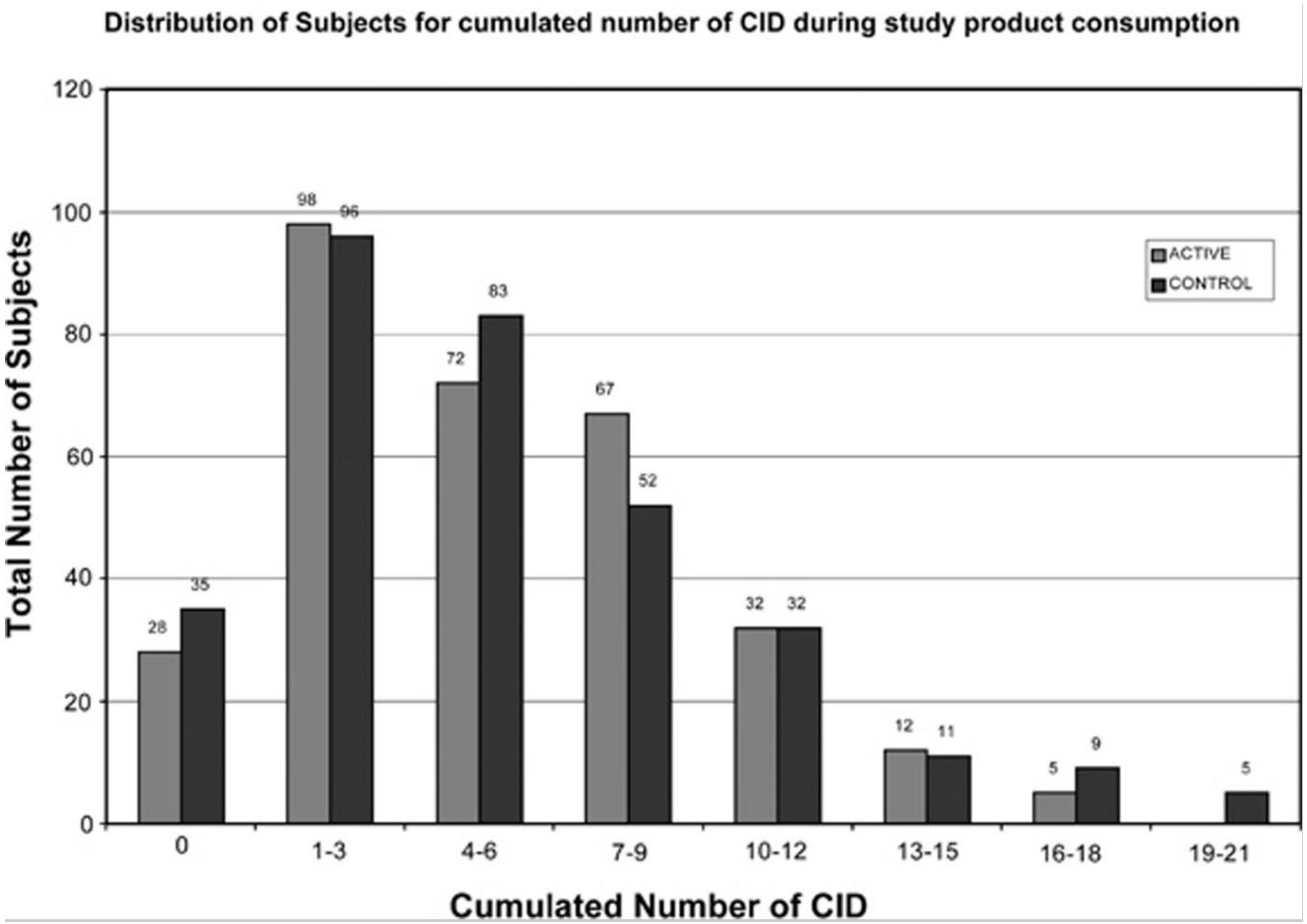
Use of a fermented dairy probiotic drink containing *Lactobacillus casei* (DN-114 001) to decrease the rate of illness in kids: the DRINK study

A patient-oriented, double-blind, cluster-randomized, placebo-controlled, clinical trial

D Merenstein[1], M Murphy[1], A Fokar[2], RK Hernandez[2], H Park[1], H Nsouli[2], ME Sanders[3], BA Davis[4], V Niborski[5], F Tondu[5] and NM Shara[2,6]

[1]*Department of Family Medicine, Georgetown University Medical Center, Washington, DC, USA;* [2]*Medstar Research Institute, Hyattsville, MD, USA;* [3]*Dairy & Food Culture Technologies, Centennial, CO, USA;* [4]*The Dannon Company, Inc., White Plains, NY, USA;* [5]*Danone Research, Palaiseau, France and* [6]*Department of Medicine, Georgetown University Medical Center, Washington, DC, USA*

Background: To evaluate whether a fermented dairy drink containing the probiotic strain *Lactobacillus casei* DN-114 001 could reduce the incidence of common infectious diseases (CIDs) and the change of behavior because of illness in children.
Subjects/Methods: We conducted a double-blinded, randomized, placebo-controlled allocation concealment clinical trial in the Washington, DC metropolitan area. Participants were 638 children 3–6 years old in daycare/schools. The intervention was a fermented dairy drink containing a specific probiotic strain or matching placebo with no live cultures for 90 consecutive days. Two primary outcomes were assessed: incidence of CIDs and change of behavior because of illness (both assessed by parental report).
Results: The rate of change of behavior because of illness was similar among active and control groups. However, the incidence rate for CIDs in the active group (0.0782) is 19% lower than that of the control group (0.0986) (incidence rate ratio = 0.81, 95% CI: 0.65, 099) $P = 0.046$.
Conclusions: Daily intake of a fermented dairy drink containing the probiotic strain *L. casei* DN-114 001 showed some promise in reducing overall incidence of illness, but was primarily driven by gastrointestinal infections and there were no differences in change of behavior.
European Journal of Clinical Nutrition (2010) **64**, 669–677; doi:10.1038/ejcn.2010.65; published online 19 May 2010


Distribution of Subjects for cumulated number of CID during study product consumption
Total Number of Subjects
120
100
80
60
40
20
0
98
96
72
83
67
52
32
32
28
35
12
11
5
9
5
ACTIVE
CONTROL
0
1-3
4-6
7-9
10-12
13-15
16-18
19-21
Cumulated Number of CID