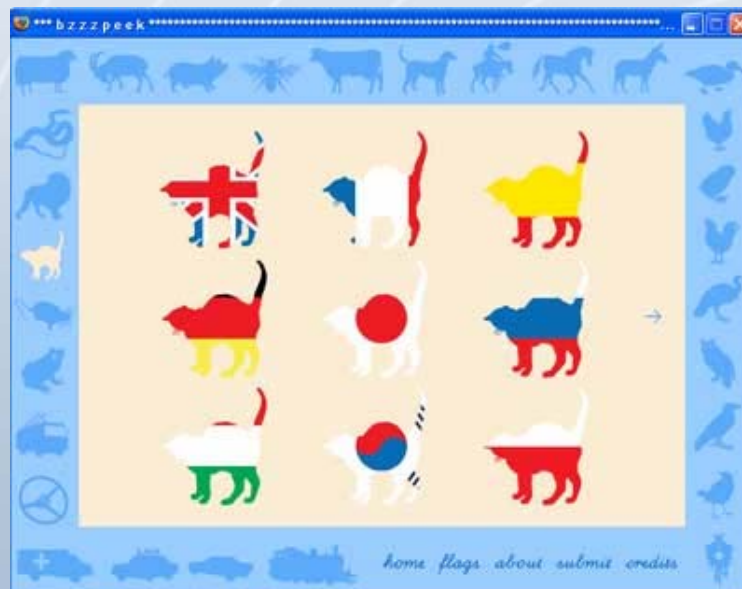


Aprendiendo idiomas III

María Jesús Vázquez Gallo



La web www.bzzzpeek.com está dedicada a las onomatopeyas y su expresión en distintos países



1. **Introducción: Estimación y contraste**
2. **Estimación no paramétrica.**
3. **Continuará...**



¿RECUERDAS?

- Estadística DESCRIPTIVA
Recolección, descripción, visualización y resumen de datos.
- INFERENCIA Estadística
Generación de modelos y predicciones, considerando la aleatoriedad del fenómeno observado.

Aprendiendo idiomas I

Aprendiendo idiomas II

Ejemplo de aplicación de la inferencia → Control de calidad

1. Estimación: aproximar datos numéricos desconocidos de la población a partir de la muestra.

Ejemplo: aproximar el porcentaje de piezas defectuosas en un lote.

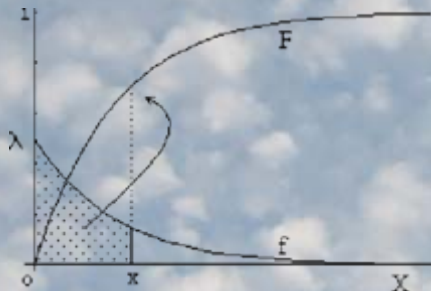
2. Contraste de hipótesis: comprobar alguna información sobre la población a partir de la muestra.

Ejemplo: comprobar que menos del 5% de las piezas de un lote son defectuosas.



Estimación
y
contraste:• **Paramétricos**

Se **estiman** características numéricas que describen la población: los **parámetros**; y se **contrastan hipótesis sobre ellos**, suponiendo que los datos siguen cierta distribución.

• **No Paramétricos**

Se **busca la distribución** a la que mejor se ajustan los datos en vez de suponer una distribución de partida; y se realizan **contrastos de hipótesis no basados en parámetros**.



¿Y qué significa aquí una **distribución** de los datos?

Los **datos** nos los proporciona una **muestra** de una **población** en la que se quiere estudiar una **variable aleatoria concreta** (¿**recuerdas** de **Aprendiendo idiomas I** la variable número de cervezas consumido por los habitantes de Dakota del Norte?)

Cada **variable aleatoria** posee una **distribución de probabilidad** que describe su comportamiento.

¿Cómo lo hace?

Función de densidad → le asigna a cada valor posible de la variable la probabilidad de que ocurra.

Función de distribución → le asigna a cada valor posible de la variable la probabilidad de que la variable no lo supere (expresa probabilidades acumuladas).



Ejemplo: Tiramos una moneda 3 veces.

La **probabilidad** de sacar, por ejemplo, tres caras es $1/8=(1/2)^3$, ya que la probabilidad de sacar cara en una tirada es $1/2$ y las tiradas se consideran **independientes**.

Definimos la **variable aleatoria** “número de caras”, que puede tomar los valores $\{0, 1, 2, 3\}$.

La **función de densidad** asignaría $1/8$ al valor 0 (la probabilidad de que haya 0 caras es $1/8$ porque sucede en 1 de los 8 casos posibles: xxx), $3/8$ al valor 1 (cxx, xcx, xxc), $3/8$ al valor 2 (ccx, cxc, xcc) y $1/8$ al valor 3 (ccc).

La **función de distribución** iría acumulando estas probabilidades y asignaría $1/8$ al valor 0, $4/8$ al valor 1, $7/8$ al valor 2 y $1(=8/8)$ al valor 3.

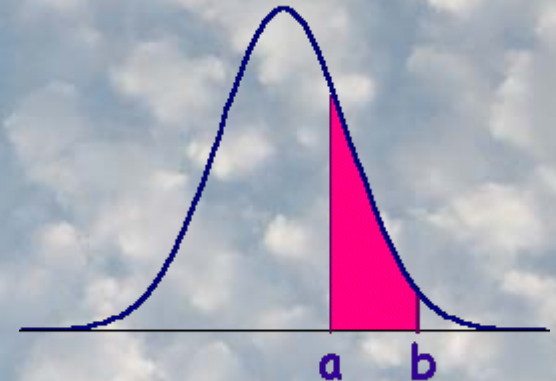


Si recordamos algo sobre funciones e integrales...

La integral de una función positiva en un intervalo es el área que queda bajo la curva que corresponde a la gráfica de la función.

→ Para variables aleatorias continuas:

La probabilidad de que la variable tome un valor cualquiera de un intervalo $[a,b]$ es el área bajo la curva que representa a la función de densidad en ese intervalo. ¿Sabrías explicar entonces por qué el área total bajo la curva debe ser 1?

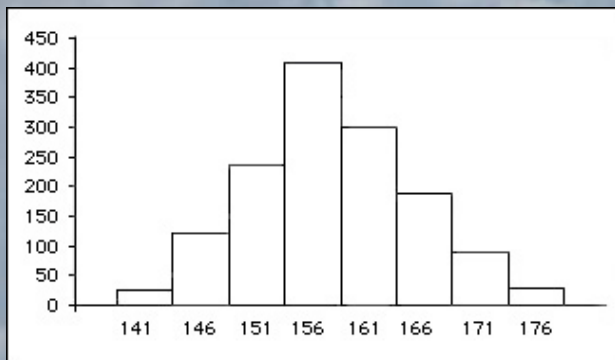


La función de distribución le asigna a cada valor posible de la variable aleatoria, la integral de la función de densidad en un intervalo que comprende todos los valores posibles hasta ese valor concreto.

1. Introducción

Es interesante construir **modelos de distribuciones de probabilidad** que representen el comportamiento teórico de los diversos **fenómenos aleatorios** que aparecen en el mundo real.

Muchos de ellos: el tener cierta talla o peso en un conjunto de individuos, el cociente intelectual, los resultados de los exámenes, el consumo de ciertos productos por individuos del mismo grupo, la resistencia a la rotura de ciertas piezas, etc. se ajustan a la famosa **distribución normal**



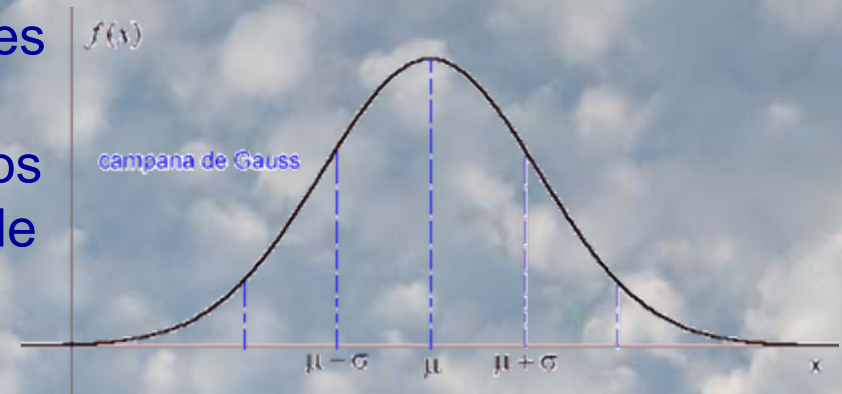
Altura de 1400 mujeres

Hay pocos individuos en los extremos y el aumento es gradual hasta llegar a la parte central, donde están la mayoría.

¿Te suena de algo la campana de Gauss?

1. Introducción

La gráfica de su **función de densidad** es simétrica respecto a μ (la *media*), con máximo en la media, y tiene dos puntos de inflexión, situados a ambos lados de la media y a distancia σ (*desviación típica*) de ella.



La distribución normal tiene **muy buenas propiedades** y por ello, en la estadística paramétrica, se supone con frecuencia que variables aleatorias con distribuciones desconocidas, siguen una normal, especialmente en Física y Astronomía.

Gauss la usó a principios del S.XIX para estudiar datos astronómicos.

Con frecuencia, esta aproximación es buena debido al **Teorema del Límite Central** que afirma que la distribución de la media de la muestra se acerca a una normal cuando el tamaño de la muestra crece.

Para saber más sobre la distribución normal:

2. Estimación no paramétrica

Continuamos con la **estimación no paramétrica**.

La **estadística no paramétrica**, frente a la paramétrica, no supone que los datos se ajustan a una distribución de probabilidad concreta.

Sus métodos son más sencillos de aplicar y más **robustos** (cambian menos al variar las condiciones de partida).

Se utiliza con frecuencia cuando:

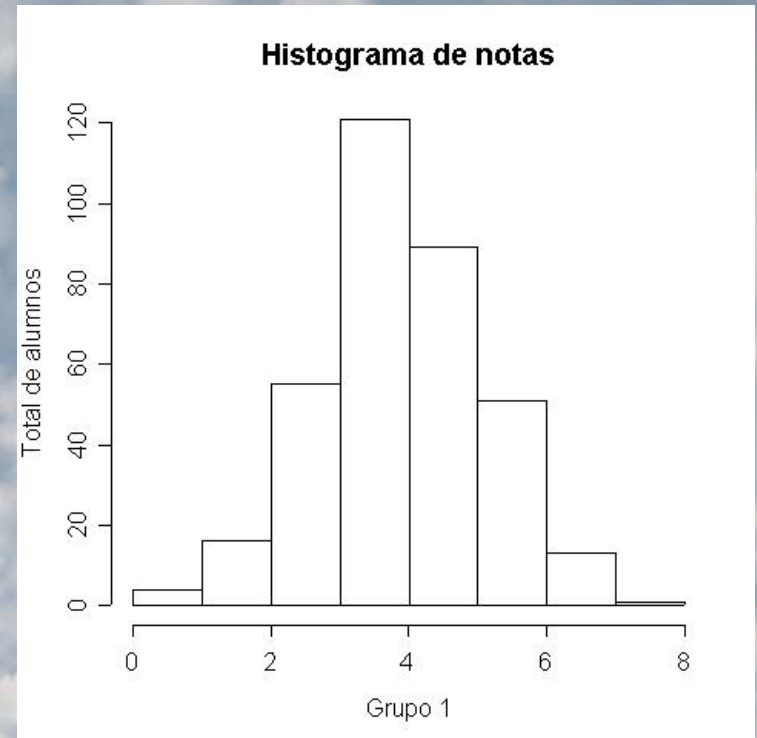
- los datos presentan **relaciones de orden** pero no tienen una interpretación numérica clara (por ejemplo, la asignación de preferencias: conceder de 1 a 5 estrellas a las películas en cartel);
- no se tiene mucha información sobre los datos;
- se tienen pocos datos.



¿Recuerdas lo que es un **histograma**?

Aquí tienes uno

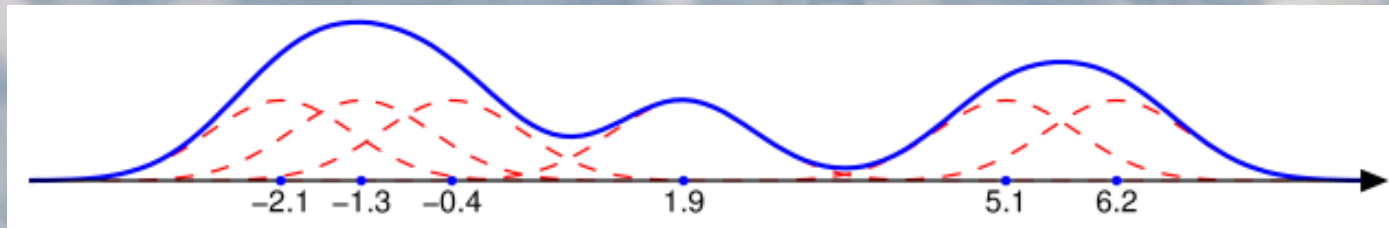
Es un **diagrama de frecuencias** que muestra qué proporción de casos encajan en una serie de categorías, a través de rectángulos adyacentes cuyas anchuras representan la clasificación por categorías y cuyas alturas representan las correspondientes frecuencias.



→ Un histograma es una **estimación no paramétrica** de la **función de densidad** de una **distribución de probabilidad desconocida**.

Pero el contorno de un histograma da saltos, no corresponde a una función de densidad “continua”...

→ Otra estimación no paramétrica de la función de densidad de una distribución de probabilidad desconocida es la llamada estimación de densidad núcleo.



→ En vez de agrupar las observaciones como hace un histograma, se coloca una pequeña montaña en cada observación, determinada por la función núcleo (que suele ser una función de densidad de una distribución normal). El estimador consiste en una suma de esas montañas y el contorno resulta ser más suave.



Otra técnica de suavizado es la de los **promedios móviles** conocidos como **MA (moving average)**.

→ Un promedio móvil es un conjunto de números, cada uno es el promedio del subconjunto correspondiente de un conjunto de datos.

Ejemplo: Tenemos 500 datos. El primer valor del promedio móvil puede ser la media aritmética de los datos del 1 al 50. El siguiente la de los datos del 2 al 51, y así hasta el último valor que será la media de los datos 451 al 500.

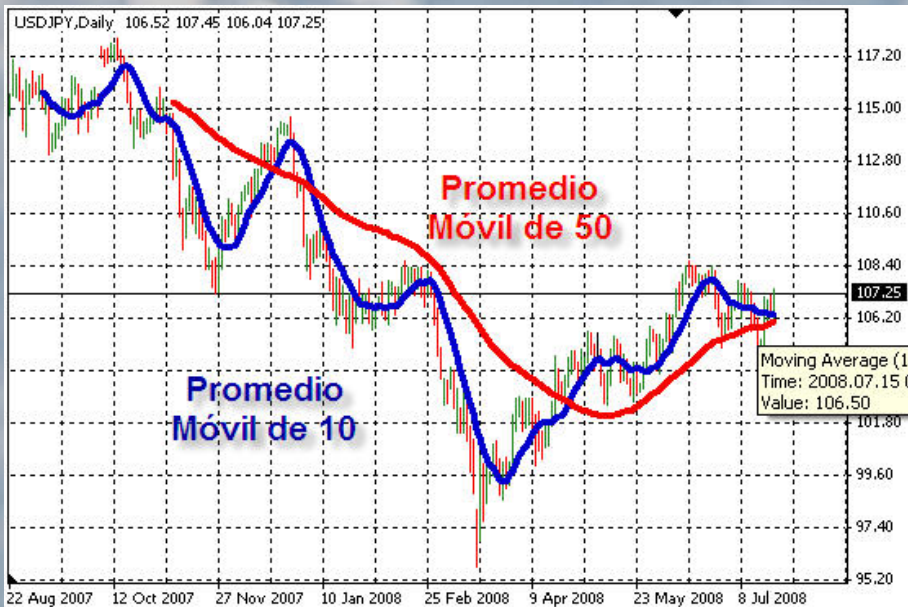


→ Un promedio móvil se puede aplicar a cualquier conjuntos de datos, pero suele utilizarse con datos de **series temporales** (¿las recuerdas de la entrega **Series temporales**?) para suavizar fluctuaciones a corto plazo y mostrar tendencias o ciclos a largo plazo.



Los promedios móviles se utilizan mucho en el análisis de los mercados bursátiles, para estudiar el precio de una acción y sus tendencias.

→ Un promedio móvil suavizará la curva de precio, es decir, producirá una línea que sigue la misma dirección que el precio pero de manera más suave, con lo cual será más fácil identificar las tendencias.



→ Un promedio móvil no anticipa lo que sucede en el mercado, pero sí puede servir para confirmar los cambios.

¿Período de tiempo adecuado para tomar los promedios?

