Estadísticos muestrales (media, vza.)

Antonio Sala Piqueras

Dept. Ing. Sistemas y Automatica (DISA)
Universitat Politècnica de València (UPV)

Video-presentación disponible en:

http://personales.upv.es/asala/YT/V/**muest**.html

Presentación

Motivación:

En una aplicación suele ser necesario indicar alrededor de qué valores se situa una variable aleatoria, y cómo de "dispersas" están sus muestras. Con un "modelo", se calculan con sumas o integrales... ¿qué se hace con datos experimentales? Se estiman a partir de funciones (estadísticos) del conjunto de muestras.

Objetivos:

Comprender la relación de media, varianza y desviación típica de una variable aleatoria con sus estimados muestrales.

Contenidos:

Preliminares. Moda, media y varianza. Estadísticos muestrales. Ley de grandes números y resultados relacionados. Conclusiones.

Esperanza, Media y Varianza de una variable aleatoria

Sean una variable aleatoria $x \in \Omega \subset \mathbb{R}$, con una cierta función de densidad $f(x)$, o una variable discreta con un conjunto de probabilidades puntuales $p(x_1), p(x_2), \ldots$.

- **Esperanza matemática** de $g : \Omega \mapsto \mathbb{R}$:
  $E(g(\cdot)) = \int_\Omega g(x) f(x)\, dx$ Discreto: $\mu := \sum_{x\in\Omega} g(x)p(x)$

- **Media:** $\mu := E(x) = \int_\Omega x \cdot f(x)\, dx$ Discr.: $\mu := \sum_{x\in\Omega} p(x)x$

- **Varianza:**
  $\sigma^2 := E((x-\mu)^2) = \int_\Omega (x-\mu)^2 f(x)\, dx$ Discr.: $\sigma^2 = \sum_{x\in\Omega} p(x)\cdot(x-\mu)$

  Desviación típica: **raíz cuadrada** de varianza.

Media, Varianza y estadísticos **muestrales**

Si no se dispone de un **modelo** –*fórmula explícita de $f(x)$*–, puede obtenerse información **experimental** de una variable aleatoria x a base de **muestreo** (repetido).

A partir de muestras de datos $\xi^T := (x(1), x(2), x(3), x(4), \ldots, x(N))$, de N repeticiones independientes con la misma distribución de probabilidad:

- Media muestral: $\mu_N := \frac{1}{N}\sum_{i=1}^{N} x(i)$.
- Varianza muestral: $\sigma_N^2 := \frac{1}{N-1}\sum_{i=1}^{N}(x(i) - E_N(x))^2$. Con media muestral cero (vbles. incrementales), resulta $\sigma_N^2 = \frac{1}{N-1}\xi_x^T\xi_x$.
- Valor esperado muestral: $E_N(g(x)) := \frac{1}{N}\sum_{i=1}^{N} g(x(i))$

En general, un "estadístico muestral" es una función $T(\xi) = T(x(1), \ldots, x(N))$ que devuelve "números" que estiman algún parámetro de la distribución de probabilidad que ha generado la muestra ξ.

Exactidud según número de muestras N

Los **estadísticos muestrales** (media, varianza, . . .) son ellos mismos **variables aleatorias**: si se repite el experimento se obtienen resultados diferentes. ¿Cómo se "parecen" a los **valores ideales** $\int \_\_ \cdot f(x)dx$ de media μ, varianza σ^2 o esperanza matemática $E(g(x))$?

[1] **Ley de los grandes números:** muestras infinitas

"*las integrales teóricas se pueden aproximar promediando muchas muestras*"

$$\lim_{N\to\infty} E_N(x) = E(x), \quad \lim_{N\to\infty} E_N((\Delta x)^2) = \sigma^2, \quad \lim_{N\to\infty} E_N(g(x)) = E(g(x))$$

*Pero en la vida real, N nunca es infinito; a veces es bastante "pequeño"...

[2] **Leyes de muestras finitas genéricas** (o "casi"):

Cotas de Chebyshev, Chernoff, Hoeffding, Bernstein. $P(|\mu_N - \mu| \leq \epsilon) \geq \alpha(N, \sigma, \epsilon)$

[3] **Leyes de muestras finitas adaptadas a distribuciones de probabilidad concretas:** p.ej. intervalos de confianza y tests estadísticos en distribución normal.

Conclusiones

- En variables aleatorias numéricas, se pueden definir media y varianza.
- A partir de múltiples muestras **independientes**, se estiman con operaciones sobre secuencias de datos.
  - Tomando muchas muestras, los valores muestrales convergen a los parámetros ideales de la distribución subyacente (ley grandes números).
  - Existen fórmulas que calculan número de muestras N necesario dada exactitud deseada ϵ y confianza estadística α –equivalentemente, probabilidad de equivocarse (1-α)–.
- Aplicaciones:
  - **validar modelos:** ¿Está, de verdad, la media de lo que está generando estos datos entre 7 y 8?
  - **identificación:** ajuste por mínimos cuadrados da $K = 3$ N/m de cte. de muelle; usados 300 datos con ruido. Si el muelle fuera realmente lineal, ¿Es la '*probabilidad de que me esté equivocando en más de 0.3 N/m*' menor del 5%?.