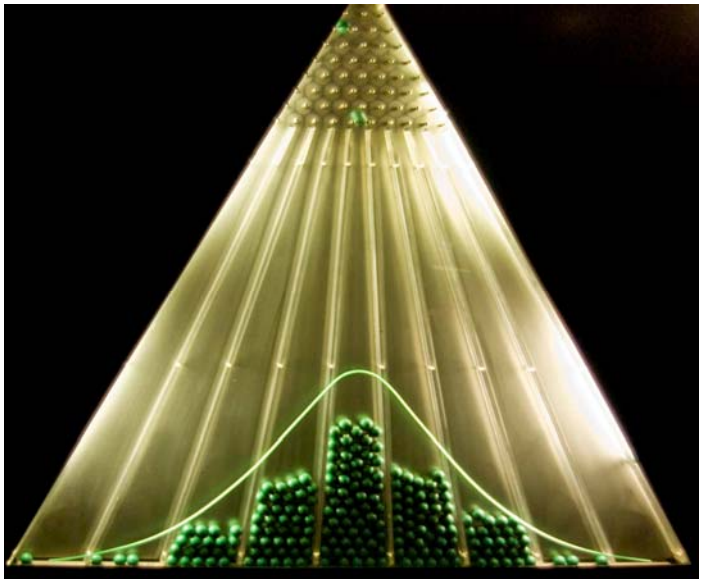
Tema 5 Leyes de los grandes números

1. Muestreo. Inferencia muestral.
2. Ley de los grandes números.
3. Teorema del límite central.
4. Significado del teorema del límite central.
5. Generador de números aleatorios gaussiano.



1. Muestreo. Inferencia muestral

En Física se asocian ***funciones densidad de probabilidad*** (pdf) a sistemas físicos específicos. La función ***pdf*** representa el resultado de todas las medidas concebibles sobre dicho sistema si las medidas se pudieran repetir infinitamente bajo las mismas condiciones experimentales.

Una pdf $f(x)$ describe las propiedades de una ***población*** o ***universo***

Población o universo.- Conjunto total de todos los resultados de todas las medidas repetidas infinitamente. Idealización inalcanzable en la práctica.

Espacio muestral.- Conjunto de todos los posibles resultados del experimento. Resultados idénticos se representan por un solo miembro del conjunto.

Muestra (sample).- Conjunto de los resultados de un experimento en particular. Puede haber resultados idénticos.

Cuando realizamos un ***experimento***, asociamos un ***observable*** a una ***variable aleatoria*** y obtenemos como resultado un conjunto finito de valores o medidas de una cierta cantidad:

$$x_1, x_2, \cdots, x_n$$

Se trata de un subconjunto o ***muestra*** de la ***población padre*** ("***parent distribution***") o ***universo***. Hablamos de una muestra de tamaño *n* extraída del universo.

1. Muestreo. Inferencia muestral

Supongamos una muestra de tamaño n: $x_1, x_2, \cdots, x_n$. *La muestra se puede caracterizar por las siguientes cantidades:*

Media muestral

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

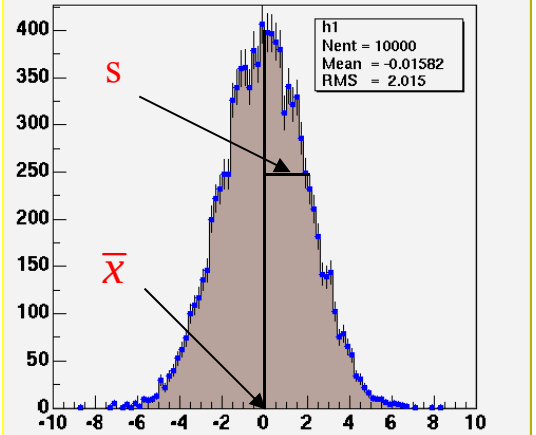
Varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

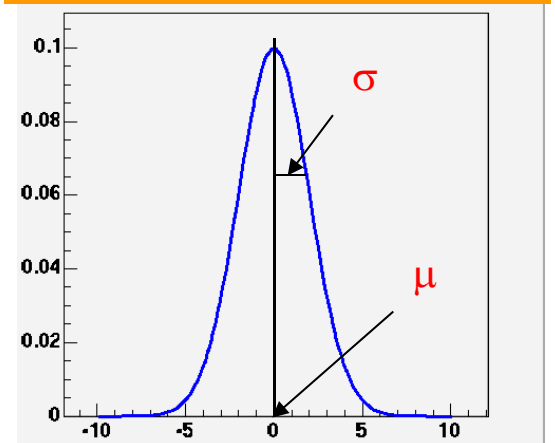
Tanto $\bar{x}, s^2$ *son funciones de variables aleatorias (**estadísticos**), por tanto son variables aleatorias también con su propia distribución.*

Inferencia muestral *Averiguar algo de la "realidad" (distribución **verdadera**) a partir de un número restringido de observaciones. Suponemos que la muestra es representativa de la población.*

Muestra de datos



Distribución "*verdadera*"



$\bar{x}$ – media de la ***muestra***

s – Desviación estándar de la ***muestra***

μ – media de la dist. ***verdadera***

σ– Desviación estándar de la dist. ***verdadera***

1. Muestreo. Inferencia muestral

Al realizar un experimento estamos interesados en la ***distribución verdadera*** que define nuestro ***sistema*** o ***universo***

Al medir, lo que hacemos es sacar conclusiones ("***inferir***") sobre dicho "***universo***" o "***sistema***" a partir de un número restringido de observaciones

Una medida del valor medio de la población vendrá dada por:

$$\hat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\overline{x}$ es una estimación del la media de la población μ

Una medida de la varianza de la población vendrá dada por:

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

s^2 es una estimación del la varianza de la población σ^2

El hecho de poner *(n-1)* en vez de *n* en el estimador de la varianza se hace para asegurar que se trata de un estimador sin sesgo (Tema 7).

1. Muestreo. Inferencia muestral

Teoremas de Convergencia

Cuando $n \to \infty$ las propiedades de la muestra se aproximan a las propiedades de la población

$$\hat{\mu} = \overline{x} \to \mu \qquad \hat{\sigma}^2 = s^2 \to \sigma^2$$

Explicación intuitiva basada en los valores esperados de los estimadores

Valor medio

$$E[\overline{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$V[\overline{x}] = V\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[x_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

La dispersión de $\hat{\mu} = \overline{x}$ en torno a μ disminuye a medida que aumenta *n*

Varianza

$$E[s^2] = \sigma^2$$

$$V(s^2) = \frac{1}{n}(\mu_4 - \mu_2^2) + \frac{2}{(n-1)n}\mu_2^2$$

La dispersión de $\hat{\sigma}^2 = s^2$ en torno a σ^2 disminuye a medida que aumenta *n*

2. Ley de los grandes números

Teorema de Chebysev

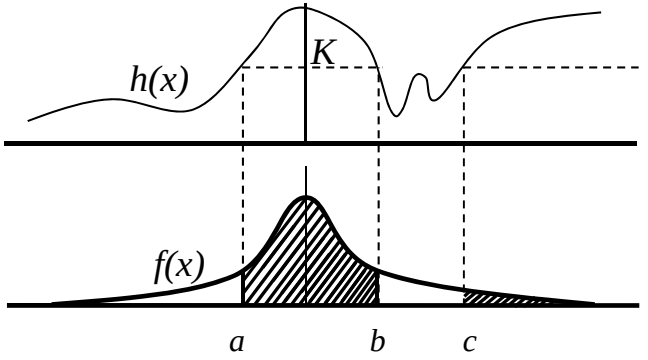
Sea $h(x)$ una función no negativa de una variable aleatoria x. Siempre se puede encontrar un límite superior a la probabilidad de que $h(x)$ exceda un determinado valor K >0

$$P\{h(x) \geq K\} \leq \frac{E[h(x)]}{K}$$

Demostración

$$E[h(x)] = \int_{\Omega} h(x)\, f(x)\, dx \geq 0$$

Puesto que ambas son funciones no negativas



Definimos R como la región de Ω tal que $h(x) \geq K$

$$\text{Si } x \in R \rightarrow h(x) \geq K \quad \Rightarrow \quad P[h(x) \geq K] = \int_{R} f(x)\, dx$$

$$R = [a,b] \cup [c,+\infty)$$

$$E[h(x)] = \int_{\Omega} h(x)\, f(x)\, dx \geq \int_{R} h(x)\, f(x)\, dx \geq \int_{R} Kf(x)\, dx = KP[h(x) \geq K]$$

$R \subset \Omega$

$K \leq h(x) \forall x \in R$

$$P[h(x) \geq K] \leq \frac{E[h(x)]}{K}$$

2. Ley de los grandes números

Desigualdad de Chebysev-Bienayme

Supongamos que x es una variable aleatoria con:

- Valor esperado $E[x]$
- Varianza $V[x]$

$$P\{h(x) \geq K\} \leq \frac{E[h(x)]}{K}$$

$$h(x) \rightarrow \left(x - E[x]\right)^2 \geq 0$$

$$K \rightarrow k^2 V(x)$$

$$P\left[\left(x - E[x]\right)^2 \geq k^2 V(x)\right] \leq \frac{E\left[\left(x - E[x]\right)^2\right]}{k^2 V(x)} = \frac{1}{k^2}$$

Tomando raíces cuadradas

$$P\left[\left|x - E[x]\right| \geq k V^{1/2}(x)\right] \leq \frac{1}{k^2}$$

Llamando

$$\mu = E[x]$$

$$\sigma^2 = V(x)$$

$$P\left[\left|x - \mu\right| \geq k\sigma\right] \leq \frac{1}{k^2}$$

k	$1/k^2$	%
1	1	100
2	1/4	25
3	1/9	11
4	1/16	6.3

La probabilidad de que una variable aleatoria x tome un valor fuera del intervalo $[\mu - k\sigma, \mu + k\sigma]$ es menor que $1/k^2$

- Independiente de cual sea la pdf
- Siempre que σ sea conocida.

2. Ley de los grandes números

Ley de los grandes números

Si aplicamos la desigualdad anterior a valores medios

$$\bar{x} = \frac{1}{N}\sum_i x_i$$

$$P\left[\left|x-\mu\right| \geq k\sigma\right] \leq \frac{1}{k^2}$$

$$P\left[\left|\bar{x}-\mu\right| \geq \frac{k\sigma}{\sqrt{N}}\right] \leq \frac{1}{k^2}$$

Mediante el cambio:

$$\varepsilon = \frac{k\sigma}{\sqrt{N}}$$

$$P\left[\left|\bar{x}-\mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2}{N\varepsilon^2} = \delta$$

$$x \to \bar{x}$$

$$E\left[\bar{x}\right] = \mu$$

$$V\left(\bar{x}\right) = \frac{\sigma^2}{N}$$

Dado un ε > 0 existe un δ > 0 y un *N* tal que la probabilidad de que $\bar{x}$ difiera de μ en una cantidad mayor que ε es menor que δ para todo *n* > *N*

Tomando límites →

$$\lim_{N\to\infty} P\left[\left|\bar{x}-\mu\right| \geq \varepsilon\right] = 0$$

"La media de n variables aleatorias converge a su valor esperado"

2. Ley de los grandes números

Generalización

En el caso más general tendremos medias de funciones de variables aleatorias, por lo que podemos generalizar del siguiente modo:

Supongamos que tenemos n variables aleatorias $x_1, x_2, \cdots, x_N$ todas ellas con valor medio μ y varianza σ^2

Supongamos que tenemos una función *g(x)* de las variables aleatorias de manera que: $\begin{cases} \text{Valor medio} & E[g] \\ \text{Varianza} & V(g) \end{cases}$

Definimos el valor medio G como: $G = \frac{1}{N}\sum_i g(x_i)$

$$E[G] = \frac{1}{N}\sum E[g] = E[g] \qquad V(G) = V\left(\frac{1}{N}\sum g(x_i)\right) = \frac{1}{N^2}\sum V(g(x_i)) = \frac{V(g)}{N}$$

$$P\left[\left|\overline{x} - \mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2}{N\varepsilon^2} \Rightarrow \begin{cases} \overline{x} \to G \\ \mu = E[G] \\ V(G) = \frac{V(g)}{N} \end{cases} \Rightarrow P\left[\left|G - E[G]\right| \geq \varepsilon\right] \leq \frac{V(g)}{N\varepsilon^2}$$

Las medias tienden a sus valores esperados

$$G = \frac{1}{N}\sum g(x_i) \to E[G]$$

2. Ley de los grandes números

Ejemplo. Integración Montecarlo

Queremos calcular una estimación de la integral definida:

$$I = \int_a^b g(x)\,dx$$

Supongamos que tenemos N números x_i *independientes y aleatorios distribuidos uniformemente, en* $[a,b]$ *según la pdf:*

$$f(x) = \frac{1}{(b-a)} \qquad a \le x \le b$$

El valor esperado de la función $g(x)$ *respecto a la pdf* $f(x)$ *viene dado por:*

$$E[g] = \int_a^b g(x)\, f(x)\,dx = \frac{1}{(b-a)} \int_a^b g(x)\,dx = \frac{I}{(b-a)}$$

Si tomamos N números aleatorios X_i *y calculamos el valor medio G como:*

$$G = \frac{1}{N}\sum g(x_i) \rightarrow E[G] = E[g] = \frac{I}{(b-a)} \quad \Rightarrow \quad \theta = \frac{(b-a)}{N}\sum g(x_i) \rightarrow \int_a^b g(x)\,dx$$

La ley de los grandes números nos dice que el estimador θ converge a la solución exacta I cuando la muestra aleatoria tiende a infinito

El valor de θ es una estimador de la integral I

La probabilidad de obtener una estimación de la integral que difiera del valor exacto en una cantidad dada, disminuye a medida que aumenta el número de puntos

3. Teorema del límite central

Ley de los grandes números → *La media de una variable aleatoria, o de una función de variables aleatorias tiende a su valor esperado cuando* $n \to \infty$

¿Cómo de rápido ? ⟸ **Teorema del Límite Central**

Supongamos que tenemos n variables aleatorias e independientes $x_1, x_2, \cdots, x_N$ *con:* $x_i \to \mu_i, \sigma_i^2$

Definimos una nueva variable s de la siguiente forma:

$$s = \sum x_i \quad \Longrightarrow \quad \mu_s = \sum \mu_i \;\; ; \;\; \sigma_s^2 = \sum \sigma_i^2$$

Independientemente de la distribución original de cada x_i *y siempre que las variables sean independientes (**Véase Tema 3** → Valor esperado y varianza de una combinación lineal de variables aleatorias)*

¿Cuál es la distribución de la nueva variable s?

"Teorema del Límite Central (forma condensada)"

"Sean $x_1, x_2, \cdots, x_N$ *un conjunto de variables aleatorias independientes de manera que cada* x_i *está distribuida con un valor medio* μ_i *y una varianza finita* σ_i^2*. Entonces la variable*

$$z = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \to N(0,1)$$

tiene como distribución límite $(n \to \infty)$ *una distribución normal centrada en 0 y con varianza 1 "*

3. Teorema del límite central

Demostración (I)

Supongamos para simplificar que todas las variables tienen la misma pdf, el mismo valor medio y la misma varianza:

$$\mu_i = \mu \;;\quad \sigma_i^2 = \sigma^2 \quad \forall\, i = 1, \cdots, n \quad \Rightarrow \quad z = \frac{\sum x_i - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} = \frac{\sum x_i - \sum \mu}{\sqrt{\sum \sigma^2}} = \frac{\sum (x_i - \mu)}{\sigma\sqrt{n}} = \sum \frac{x_i - \mu}{\sigma\sqrt{n}}$$

Cada variable aleatoria x_i tiene su función característica

$$\Phi(t) = E\left[e^{itx_i}\right] = 1 + itE\left[x_i\right] + \frac{(it)^2}{2!}E\left[x_i^2\right] + \cdots = 1 + it\mu + \frac{(it)^2}{2!}\left(\mu^2 + \sigma^2\right) + \cdots$$

Si las variables x_i son independientes también lo son los términos $\dfrac{x_i - \mu}{\sigma\sqrt{n}}$

La función característica de z será entonces:

$$\Phi_z(t) = E\left[e^{itz}\right] = E\left[e^{it\sum \frac{x_i - \mu}{\sigma\sqrt{n}}}\right] = E\left[e^{it\frac{x_1 - \mu}{\sigma\sqrt{n}}} \cdot e^{it\frac{x_2 - \mu}{\sigma\sqrt{n}}} \cdots e^{it\frac{x_n - \mu}{\sigma\sqrt{n}}}\right] = \left(E\left[e^{it\frac{x_i - \mu}{\sigma\sqrt{n}}}\right]\right)^n =$$

$$= \left(e^{-\frac{it\mu}{\sigma\sqrt{n}}} E\left[e^{i\frac{t}{\sigma\sqrt{n}}x_i}\right]\right)^n = \left(e^{-\frac{it\mu}{\sigma\sqrt{n}}} \Phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n \quad \Rightarrow \quad \Phi_z(t) = \left[e^{-\frac{it\mu}{\sigma\sqrt{n}}} \Phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

3. Teorema del límite central

Demostración (II)

$$\ln(1+x) = x - \frac{1}{2}x^2 + \cdots$$

Tomando logaritmos:

$$\ln \Phi_z(t) = n\left\{-\frac{it\mu}{\sigma\sqrt{n}} + \ln \Phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right\} = n\left\{-\frac{it\mu}{\sigma\sqrt{n}} + \ln\left[1 + \underbrace{\left(\frac{it}{\sigma\sqrt{n}}\right)\mu + \left(\frac{it}{\sigma\sqrt{n}}\right)^2 \frac{\left(\mu^2+\sigma^2\right)}{2!} + \cdots}_{x}\right]\right\} =$$

$$= n\left\{-\frac{it\mu}{\sigma\sqrt{n}} + \frac{it\mu}{\sigma\sqrt{n}} + \frac{1}{2!}\left(\frac{it}{\sigma\sqrt{n}}\right)^2\left(\mu^2+\sigma^2\right) + \cdots - \frac{1}{2}\left(\frac{it}{\sigma\sqrt{n}}\right)^2\mu^2 + \cdots\right\} = -\frac{t^2}{2} + \vartheta\left(\frac{1}{\sqrt{n}}\right)$$

$$\ln \Phi_z(t) = -\frac{t^2}{2} + \vartheta\left(\frac{1}{\sqrt{n}}\right) \xrightarrow[n \to \infty]{} \Phi_z(t) = e^{-\frac{t^2}{2}}$$

Función característica de una distribución normal con media 0 y varianza unidad

En el límite de grandes números z se distribuye según una gaussiana $N(0,1)$

4. Significado teorema del límite central

¿Por qué la mayoría de las medidas que hacemos responden a distribuciones gaussianas?

*La mayoría de las cosas que medimos dependen de multitud de efectos y factores perturbadores e incontrolables. Lo que medimos es lo que queremos medir ("**efecto verdadero**") más un gran número de pequeñas contribuciones ("**error total**"):*

Medida = "Efecto verdadero" + "Error total"

$$\text{"Error Total"} = \sum_i (\text{factores})_i$$

Cada factor responde a una distribución o pdf desconocida.

Combinación de un gran número de errores (variables aleatorias) independientes

Cuando $n \to \infty$ *el error total se distribuye* ***normalmente***

¿Cómo se distribuyen los valores medios?

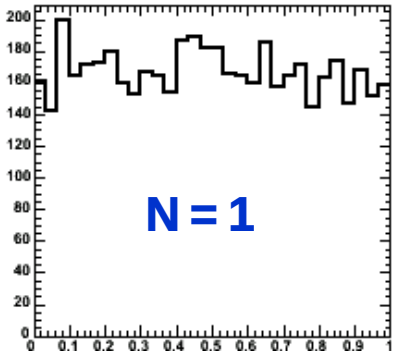
Según el TLC, los valores medios se distribuyen de acuerdo a una distribución normal con valor medio, el valor medio original μ *y desviación estándar:*

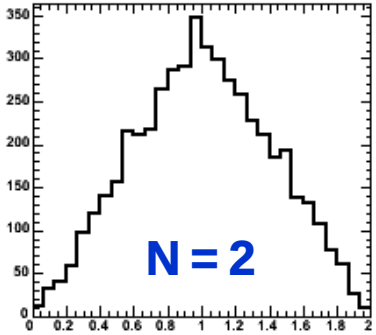
$$\frac{\sigma}{\sqrt{n}}$$

¡ El error de la media se reduce a medida que realizamos más medidas en la media !

5. Generador de números aleatorio gaussiano

5000 números aleatorios distribuidos uniformemente entre [0,1]
Media ½, Varianza 1/12



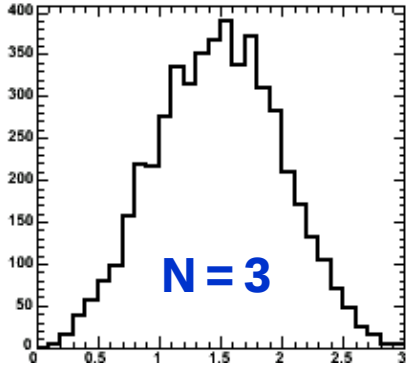


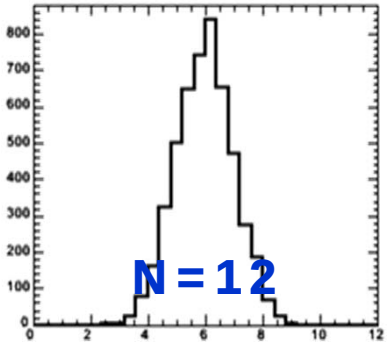
5000 números aleatorios cada uno de los cuales es la suma de dos números aleatorios uniformes distribuidos entre [0,1]

$$x = x_1 + x_2$$

5000 números aleatorios, suma de tres números aleatorios uniformes distribuidos entre [0,1]

$$x = x_1 + x_2 + x_3$$





5000 números aleatorios, suma de 12 números aleatorios uniformes distribuidos entre [0,1]