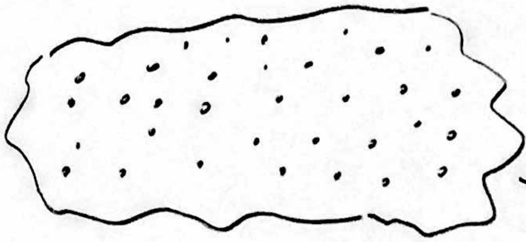
INTERVALOS DE CONFIANZA

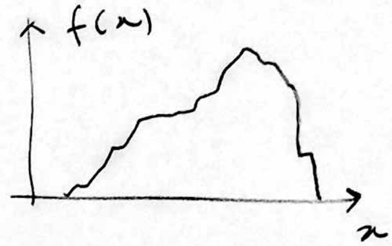


UNIVERSO
(ej. Estudiantes universitarios)

→ Cada muestra puede ser medida con una variable x (ej. Edad)

x : variable aleatoria con función de densidad de probabilidad $f(x)$, i.e.,

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1$$



$f(x)$ tiene una media $\mu = \int_{-\infty} x f(x) dx$

y una varianza $\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$

¿Cómo estimar μ?
¿Qué tan buena es su estimación?

TEOREMA DEL LIMITE CENTRAL

Si se toman N muestras del universo: $x_1, x_2, \ldots x_N$

se puede calcular $\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$, si este proceso lo

repetimos M veces (sacando N muestras aleatorias y promediándolas cada vez) entonces tendremos una nueva variable aleatoria $\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_M$ cuya media y varianza:

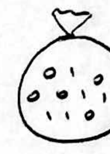
$$\tilde{\mu} = \frac{1}{M} \sum_{k=1}^{M} \hat{\mu}_k \quad \text{y} \quad \tilde{\sigma}^2 = \frac{1}{M-1} \sum_{k=1}^{M} (\hat{\mu}_k - \tilde{\mu})^2$$

Si M es suficientemente grande, la variable aleatoria tiene una distribución normal con media $\tilde{\mu} = \mu$ y varianza

$$\tilde{\sigma}^2 = \frac{\sigma^2}{N}.$$

→ PATØS_TeoremaLimCentral.m

(Si M es pequeño, e.g. $M < 20$, la distribución es t-Student con $M-1$ grados de libertad).



Ejemplo

En una bolsa hay millones de "1" y "0" escritos en bolas

¿Cuál es la probabilidad de sacar x "1" al sacar N bolas?

Supongamos que la probabilidad de sacar un "1" es p (y la de sacar un "0" es $q = 1-p$).

bolas:

$$2^N \text{ posibilidades} \left\{ \begin{array}{ccccc} N & N-1 & \cdots & 2 & 1 \\ 0 & 0 & & 0 & 0 \\ 0 & 0 & & 0 & 1 \\ 0 & 0 & & 1 & 0 \\ 0 & 0 & & 1 & 1 \\ & & \vdots & & \\ 1 & 1 & & 1 & 1 \end{array} \right.$$

Existen $\frac{N!}{x!\,(N-x)!}$ posibilidades (en las 2^N posibles combinaciones) de sacar exactamente x "1".

$$p(x) = \frac{N!}{x!\,(N-x)!}\, p^x q^{N-x}$$

Distribución de Bernoulli con $\mu = p$ $\sigma^2 = pq$

Ejemplo MATLAB:

```
M = 1000000   p=0.75   q=1-p
T = rand(M,1) > q
mean(T) ≃ 0.75 (p)
std(T) ≃ √pq
```

Si repetimos el experimento M veces, i.e. sacamos M veces N bolas

$$M \left\{ \begin{array}{l} \overbrace{0\ 1\ 0\ \ldots\ 0\ 1}^{N} \rightarrow \hat{\mu}_1 \\ 0\ 0\ 1\ \ldots\ 1 \rightarrow \hat{\mu}_2 \\ \vdots \\ 1\ 0\ 1\ \ldots\ 0\ 0 \rightarrow \hat{\mu}_M \end{array} \right.$$

Por el teorema del lím central ←

$$\tilde{\mu} = \frac{1}{M}\sum \hat{\mu}_k \simeq \mu = p$$

$$\tilde{\sigma}^2 \simeq \frac{\sigma^2}{N} = \frac{\sqrt{pq}}{N}$$

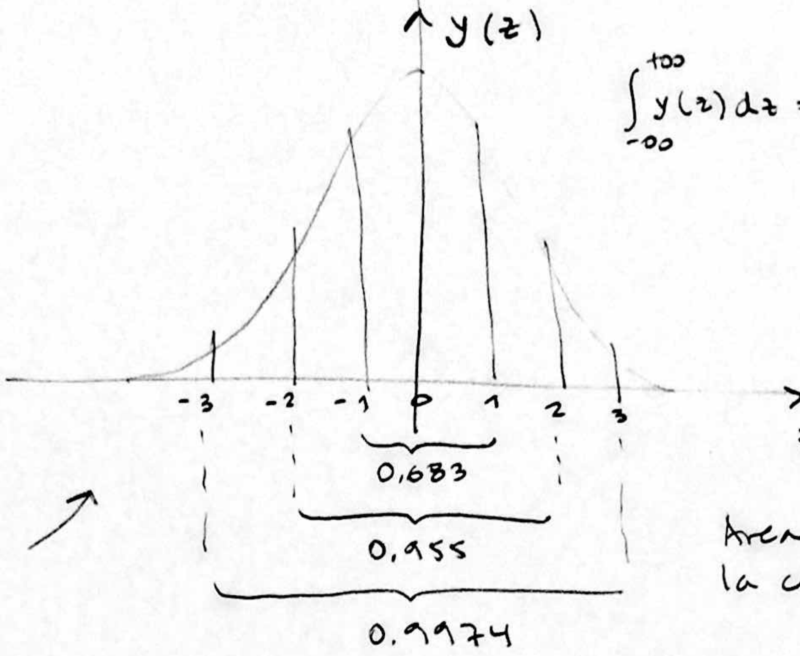
→ PATOS_Bernoulli.m

DISTRIBUCIÓN NORMAL

$$\int_{-\infty}^{+\infty} y(z)\,dz = 1$$

Si tenemos una distribución normal con media $\tilde{\mu}$ y varianza $\tilde{\sigma}^2$ podemos hacer un cambio de coordenadas

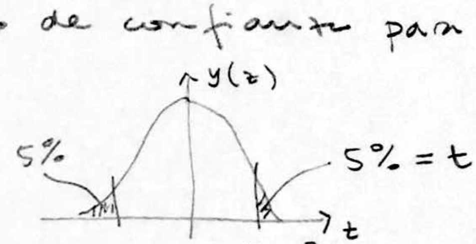
$$z = \frac{x - \tilde{\mu}}{\tilde{\sigma}}$$ y obtenemos



Área bajo la curva

Si se tiene una variable x de un universo (ej. edad de estudiantes) y calculamos μ_k como el promedio de N muestras aleatorias, y repetimos este experimento M veces, tendremos una estimación de $\tilde{\mu}$ que tiene una varianza $\tilde{\sigma}$. Sabemos entonces que con un 95.5% de probabilidad esta estimación de $\hat{\mu}$ está entre $z = -2$ y $z = 2$, es decir entre $\mu \pm 2\sigma$. Este es el intervalo de confianza de la estimación

Ejemplo: ¿Cómo calcular el intervalo de confianza para una probabilidad del 90%?



1) Se define $t = \frac{1-c}{2}$ $(c = 0.90)$

2) Se busca la inversa de la normal

$z_0 = \text{norminv}(1-t)^*$;

ojo: $\tilde{\sigma} = \frac{\sigma}{\sqrt{N}}$

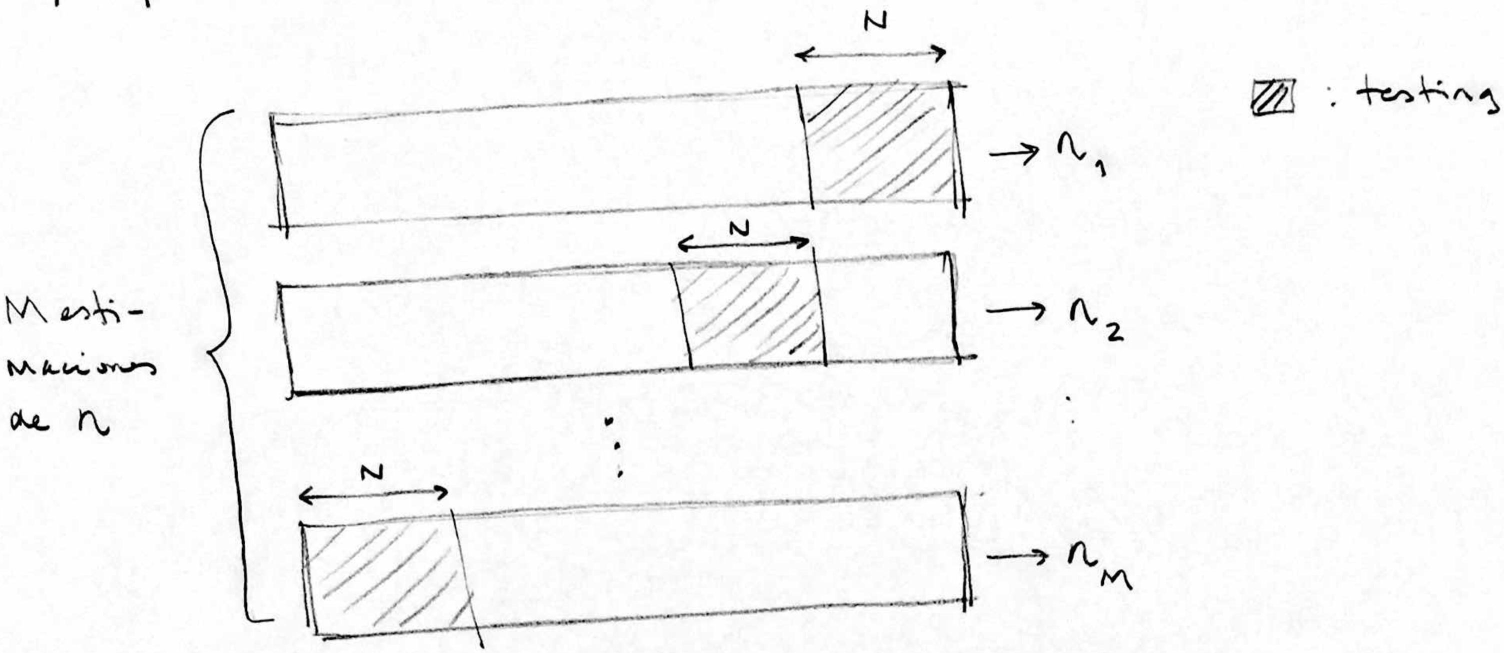
3) Intervalo de confianza:

$$\tilde{\mu} - \tilde{\sigma} z_0 \leq \tilde{\mu} \leq \tilde{\mu} + \tilde{\sigma} z_0$$

---

* Para t-student $z_0 = \text{tinv}(1-t, M-1)$;

Ejemplo de Intervalos de Confianza en Validación Cruzada:



N muestras de testing en cada experimento

0	1	0	...	1	$\rightarrow r_1$
0	0	1	...	1	$\rightarrow r_2$
1	1	1	...	0	⋮
1	0	1	...	0	$\rightarrow r_M$

(N columnas, M filas)

"1" bien clasificado

"0" mal clasificado

$\Longrightarrow$ Distribución de Bernoulli $\rightarrow \tilde{\mu} = \frac{1}{M}\sum_{k=1}^{M} r_k \simeq p$

$$\tilde{\sigma}^2 = \frac{\sigma^2}{N} \simeq \frac{p\overbrace{(1-p)}^{q}}{N}$$

¿Cómo se calculan los intervalos de confianza?

0. Definir la probabilidad del intervalo de confianza c, calcular $t = (1-c)/2$.
1. Realizar M experimentos con N datos c/u (ver arriba)
2. Estimar r_k for $k = 1..M$
3. Calcular $p = \frac{1}{M}\sum r_k$; $q = 1-p$; $\tilde{\sigma} = \sqrt{pq}/N$
4. If $M > 20$ $z_0 = \text{norminv}(1-t)$ else $z_0 = \text{tinv}(1-t, M-1)$
5. Intervalo de confianza $\tilde{r} - \Delta r \le \hat{r} = p \le \hat{r} + \Delta r$ con $\Delta r = z_0 \tilde{\sigma}$.

DISTRIBUCIONES :

NORMAL :

$$y(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

t-Student :

$$y(z) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{z^2}{\nu}\right)^{-(\nu+1)/2}$$

ν : grados de libertad

$\Gamma(\nu) = (\nu - 1)!$

$\longrightarrow$ DATOS_normal_vs_t.m