# Áreas bajo curvas y distribuciones de probabilidad con WxMaxima

Si $X$ es una variable aleatoria, y $f(x)$ su función de probabilidad (o densidad), entonces

$$\text{Prob}\,\big(X \in [a,b]\big) \;=\; \int_a^b f(x)\,dx\,,$$

que corresponde con el área bajo la curva $f(x)$ cuando $x \in [a,b]$.

En Estadística se utilizan varias distribuciones de probabilidad, siendo las más frecuentes

- **Distribución normal estándar,** $X \sim N(0,1)$

$$f(x) \;=\; \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}, \quad x \in \mathbb{R}$$

- **Distribución normal general (de media $\mu$ y DT $\sigma$),** $X \sim N(\mu,\sigma)$

$$f_{\mu,\sigma}(x) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\,e^{-(\frac{x-\mu}{\sigma})^2/2}, \quad x \in \mathbb{R}$$

- **Distribución $t$-Student de grado $m$,** $X \sim \mathtt{T}_m$

$$f_m(x) \;=\; \frac{c_m}{(1+\frac{x^2}{m})^{\frac{m+1}{2}}}, \quad x \in \mathbb{R},$$

  donde la constante $c_m > 0$ es tal que $\int_{-\infty}^{\infty} f_m(x)\,dx = 1$.

En **WxMaxima**, el paquete `distrib` contiene estas y otras distribuciones

`load(distrib)`

Los comandos más usuales (para normales y t-Student) son los siguientes

- Función de densidad $f(x)$

  `pdf_normal`$(x,0,1)$, `pdf_normal`$(x,\mu,\sigma)$, `pdf_student_t`$(x,m)$

- Función de densidad acumulada: $F(x) = \text{Prob}\;(X \le x) = \displaystyle\int_{-\infty}^{x} f(y)dy$

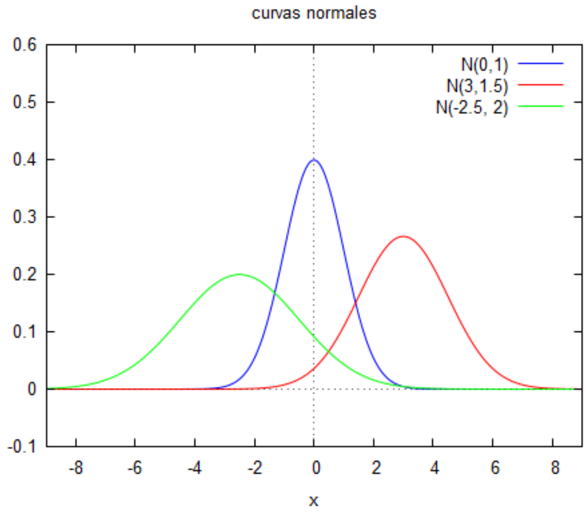
  `cdf_normal`$(x,\mu,\sigma)$, `cdf_student_t`$(x,m)$

- Percentil-$\alpha$: es el número $x_\alpha$ que cumple $\text{Prob}\;(X \le x_\alpha) = \displaystyle\int_{-\infty}^{x_\alpha} f(y)dy = \alpha$

  `quantile_normal`$(\alpha,\mu,\sigma)$, `quantile_student_t`$(\alpha,m)$

**Ejemplo 1:** Dibuja las curvas normales $N(0,1), N(3,1'5), N(-2'5,2)$

```
(%i1)   load(distrib);
(%i2)   N(x,mu,sigma):= pdf_normal(x,mu,sigma);
(%i3)   wxplot2d([N(x,0,1),N(x,3,1.5), N(x,-2.5,2)], [x,-9,9], [y, -.1,.6],
            [legend, "N(0,1)", "N(3,1.5)", "N(-2.5, 2)"],
             [title, "curvas normales"]);
```




**Ejemplo 2:** Para la distribución normal $X \sim N(0,1)$, calcula las probabilidades

$$\text{Prob }(X \leq 2), \qquad \text{Prob }(-1 \leq X \leq 1), \qquad \text{Prob }(X \geq 0'5)$$

```
(%i1)   load(distrib); numer:true
(%i2)   cdf_normal(2,0,1);
        cdf_normal(1,0,1) - cdf_normal(-1,0,1);
        1-cdf_normal(0.5,0,1);
```

```
(% o2) 0.9772498680518208
(% o3) 0.6826894921370859
(% o4) 0.3085375387259869
```

**Ejemplo 3:** Para la distribución $t$-Student $\mathtt{T}_6$, calcula los percentiles cuando

$$\alpha = 0'30 \qquad \alpha = 0'975 \qquad \alpha = 0'995$$

```
(%i1)   load(distrib); numer:true
(%i2)   quantile_student_t(0.30, 6);
        quantile_student_t(0.975, 6);
        quantile_student_t(0.995, 6);
```

```
(% o2) -0.553380923551519
(% o3) 2.446911851144877
(% o4) 3.707428021260888
```

## Cálculo de intervalos de confianza

Cuando $X \sim N(\mu, \sigma)$, y el valor de $\mu$ es desconocido, es posible estimarlo a partir de una muestra aleatoria $\{x_1, \ldots, x_n\}$ de $X$. El estimador habitual es la media muestral

$$\hat{\mu} = \bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

Sin embargo, esta elección de $\hat{\mu}$ cambia con la muestra...

> ¿Es posible dar un intervalo $\bar{x}\pm$ error, que podamos afirmar que contiene el valor real de $\mu$ con una *confianza del* 95 %?

La respuesta es positiva, si $\{x_1, \ldots, x_n\}$ es una muestra aleatoria *independiente*. Para ello se utilizan los percentiles

$$z_\alpha = \texttt{quantile\_normal}(\alpha, 0, 1), \qquad t_\alpha^{(m)} = \texttt{quantile\_student\_t}(\alpha, m)$$

**Caso 1:** $X \sim N(\mu, \sigma)$, con $\sigma$ conocido. Entonces $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, y se cumple

$$\text{Prob}\left( - z_{\frac{1+\alpha}{2}} \ \leq \ \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ \leq \ z_{\frac{1+\alpha}{2}} \right) = \alpha\,.$$

En particular, si $\alpha = 0'95$,

$$\boxed{\bar{x} \pm z_{0{,}975} \frac{\sigma}{\sqrt{n}}}$$

es un *intervalo de confianza para* $\mu$ *al* 95 %.

**Caso 2:** Si $\sigma$ es desconocido, entonces debemos estimarlo con la *desviación típica muestral*

$$\hat{\sigma} = s_n = \sqrt{\tfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

En ese caso, $T = \frac{\overline{X}-\mu}{S_n/\sqrt{n}} \sim \mathtt{T}_{n-1}$, y se cumple

$$\text{Prob}\left( - t^{(n-1)}_{\frac{1+\alpha}{2}} \ \leq \ \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ \leq \ t^{(n-1)}_{\frac{1+\alpha}{2}} \right) = \alpha\,.$$

En particular, si $\alpha = 0'95$,

$$\boxed{\bar{x} \pm t^{(n-1)}_{0{,}975} \frac{s_n}{\sqrt{n}}}$$

es un *intervalo de confianza para* $\mu$ *al* 95 %.

**Ejemplo 1:** *La cantidad de Na en un medicamento se distribuye como una normal con $\sigma = 0'6$, y $\mu$ a determinar. Extraemos una muestra de $n = 36$ pastillas al azar y obtenemos*

$$\bar{x} = 2'2$$

*Encontrar intervalos de confianza para $\mu$ al 95 % y al 99 %. Si el fabricante afirma que sus pastillas tienen $\mu = 2'5$, ¿podría tener razón?*
Solución: Utilizando WxMaxima obtenemos

1. Intervalo al 95 %

$$\bar{x} \pm z_{0,975}\frac{\sigma}{\sqrt{n}} \;=\; 2'2 \pm 1'96 \cdot \tfrac{0'6}{\sqrt{36}} \;\approx\; 2'2 \pm 0'2 \;=\; \big[2, 2'4\big]$$

2. Intervalo al 99 %

$$\bar{x} \pm z_{0,995}\frac{\sigma}{\sqrt{n}} \;=\; 2'2 \pm 2'58 \cdot \tfrac{0'6}{\sqrt{36}} \;\approx\; 2'2 \pm 0'26 \;=\; \big[1'94, 2'46\big]$$

En particular, con un 99 % de confianza podemos garantizar que el fabricante **no** tiene razón cuando afirma que sus pastillas tienen $\mu = 2'5$.

**Ejemplo 2:** *La concentración de flúor en la pasta de dientes sigue una distribución $N(\mu, \sigma)$. La normativa exige que para uso infantil no se superen las 400 ppm de Fl. Un inspector extrae $n = 15$ muestras aleatorias de un fabricante, y obtiene*

$$\bar{x} = 425, \qquad\qquad s_n = 30$$

*Encontrar intervalos de confianza para $\mu$ al 95 % y al 99 %. ¿Tiene el inspector evidencia suficiente para afirmar que se superan los 400 ppm de Fl?*
Solución: Como $\mu, \sigma$ son ambos desconocidos, utilizamos intervalos con la $t$-Student

1. Intervalo al 95 %

$$\bar{x} \pm t^{(14)}_{0,975}\frac{s_n}{\sqrt{n}} \;=\; 425 \pm 2'145 \cdot \tfrac{30}{\sqrt{15}} \;\approx\; 425 \pm 16'6 \;\approx\; \big[408, 442\big]$$

2. Intervalo al 99 %

$$\bar{x} \pm t^{(14)}_{0,995}\frac{s_n}{\sqrt{n}} \;=\; 425 \pm 2'977 \cdot \tfrac{30}{\sqrt{15}} \;\approx\; 425 \pm 23 \;\approx\; \big[402, 448\big].$$

En particular, con un 99 % de confianza podemos garantizar que para esa pasta de dientes $\mu$ supera la cantidad de 400 ppm.

**Ejemplo 3:** *Para obtener bandera azul en una playa, la concentración de E. Coli debe ser inferior a 500 UFC/100 ml. La concentración en muestras aleatorias se comporta como una $N(\mu, \sigma)$. Un inspector toma $n = 8$ muestras y obtiene*

$$\bar{x} = 515, \qquad\qquad s_n = 20$$

*Encontrar un intervalo de confianza para $\mu$ al 95 %. El Ayuntamiento afirma que $\mu = 500$. ¿Tiene el inspector evidencia suficiente para rechazar la bandera azul?*
Solución: De nuevo utilizamos intervalos con la $t$-Student

1. Intervalo al 95 % para $\mu$

$$\bar{x} \pm t^{(7)}_{0,975}\frac{s_n}{\sqrt{n}} \;=\; 515 \pm 2'365 \cdot \tfrac{20}{\sqrt{8}} \;\approx\; 515 \pm 16'72 \;\approx\; \big[498, 532\big]$$

Como 500 pertenece al intervalo, no tenemos evidencia suficiente (al 95 %) para concluir que el Ayuntamiento no tenga razón. Por tanto, debemos mantener la bandera azul.