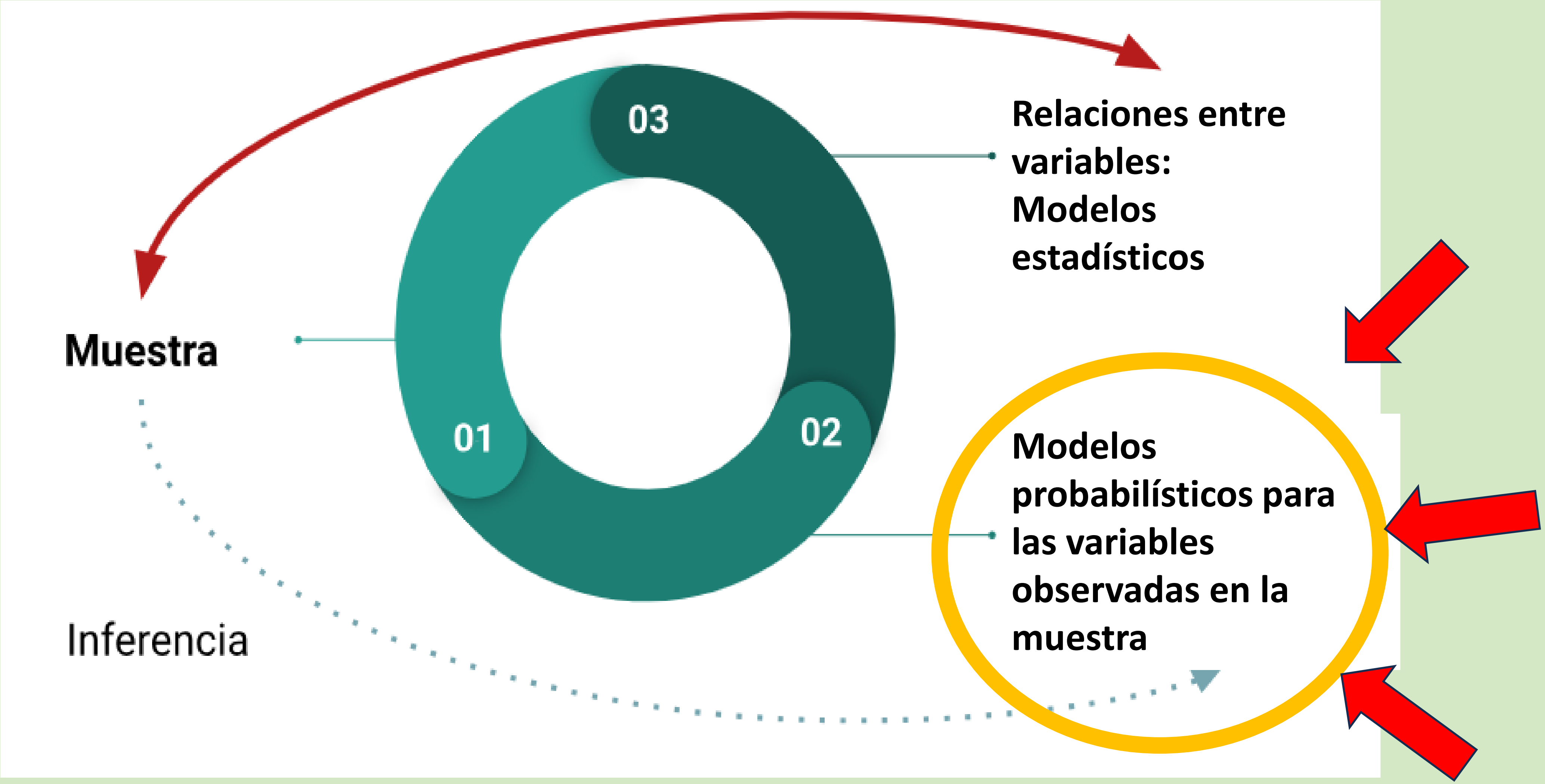
Fundamentos de Estadística

Silvia N. Pérez
Especialización en Ciencia de Datos - UNO

Recordemos: etapas del proceso estadístico



¿Qué es un modelo probabilístico?

- Primero tendremos que definir qué es una un espacio de probabilidad…
- Luego, hablar de variables aleatorias..
- Finalmente, podremos definir el concepto de “modelo” para la distribución de estas variables.

Espacios de probabilidad (en breve)

Dada una experiencia aleatoria, definimos el espacio muestral como el conjunto de todos los resultados posibles.

Un conjunto de resultados posibles será un “evento”, y sobre estos se define una medida de probabilidad, que serán números entre 0 y 1 cumpliendo algunas condiciones.

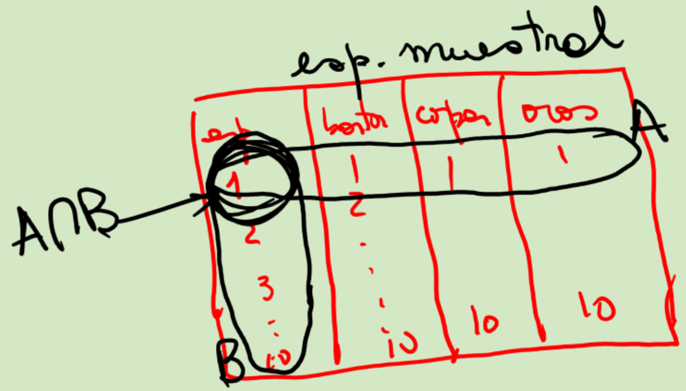
Ejemplo simple:

- Experiencia aleatoria: elijo una carta del mazo de truco
- Espacio muestral: {1,2,…..40}
- H= { sale ancho de espadas} es un evento
- P(H) = 1/40

Probabilidad condicional (en breve)

Dados dos eventos, se define la probabilidad de A si se sabe que ocurrió B:

$$\underbrace{\mathrm{P(A|B)}}_{\substack{\text{A dado B} \\ \text{A si B}}} = \frac{P(A\cap B)}{P(B)}$$



Ejemplo simple:

- A= { sale el número 1}   B= { sale carta de espadas}

$P(A) = \frac{4}{40} = \frac{1}{10}$

$P(B) = \frac{10}{40} = \frac{1}{4}$

- $\mathrm{P(A|B)} = \frac{P(\overbrace{A\cap B}^{\text{1 de esp}})}{P(B)} = \frac{1/40}{1/4} = \frac{1}{10}$ en este caso da igual que P(A)

pero esto no siempre es así!!

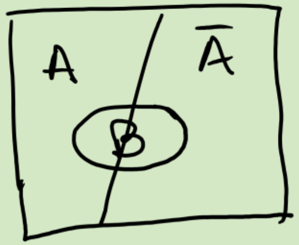
$P(A\cap B) = \frac{1}{40}$

$\boxed{\text{Cuando } P(A|B) = P(A) \text{ se dice que A y B son INDEP}}$

Fórmula de Bayes

a B

Dados dos eventos, se puede expresar la probabilidad de A condicional usando las condicionales en otro orden, lo que puede ser conveniente según la información disponible:

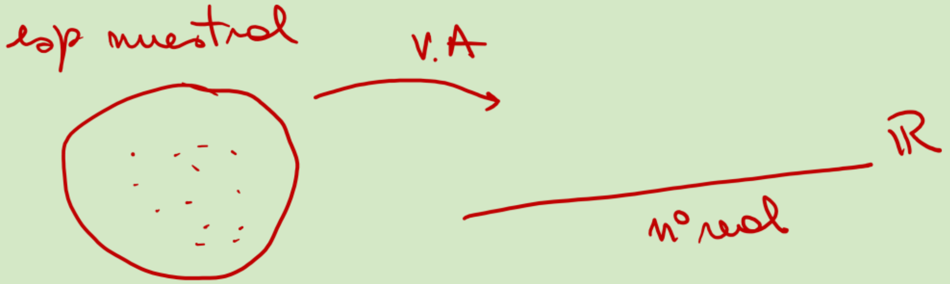


$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|\bar{A})P(\bar{A})}$$

Ejemplo simple 2: tengo 4 y 6 en una caja. Saco dos bolitas sin reponer.

- A_1= { sale amarilla la primera} R_2= { sale roja la segunda}
- $P(A_1|R_2) = \frac{P(R_2|A_1)\cdot P(A_1)}{P(R_2|A_1)P(A_1)+P(R_2|R_1)P(R_1)} = \frac{6/9 * 4/10}{6/9*\frac{4}{10} + 5/9*6/10} = \boxed{\frac{24}{54}}$

Variables aleatorias



Transforma (codifica) los resultados del experimento aleatorio a números reales.

Esto es, para cada resultado del espacio muestral hace corresponder un número.

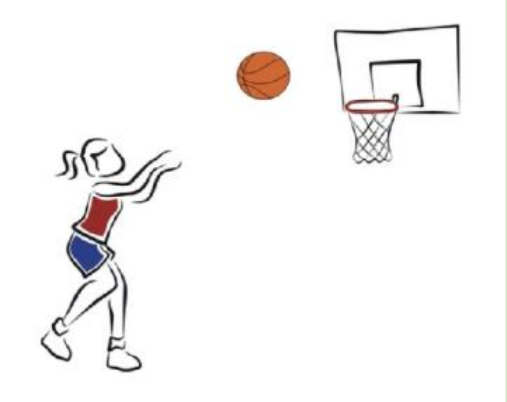
Ejemplos:

X: cantidad de bolas amarillas en dos extracciones

Y: tiempo hasta el primer gol del partido

Z: monto en caja de ahorro al día 30

W: cantidad tiros hasta el primer acierto



Variables aleatorias: tipos

- Discretas
- Continuas
- Mixtas

Abalone

pienso q' datos obs corresponden a V.A en la población

- .. N. anillos
- .. Long, Diam, Peso....

Premios

- ..
- ..

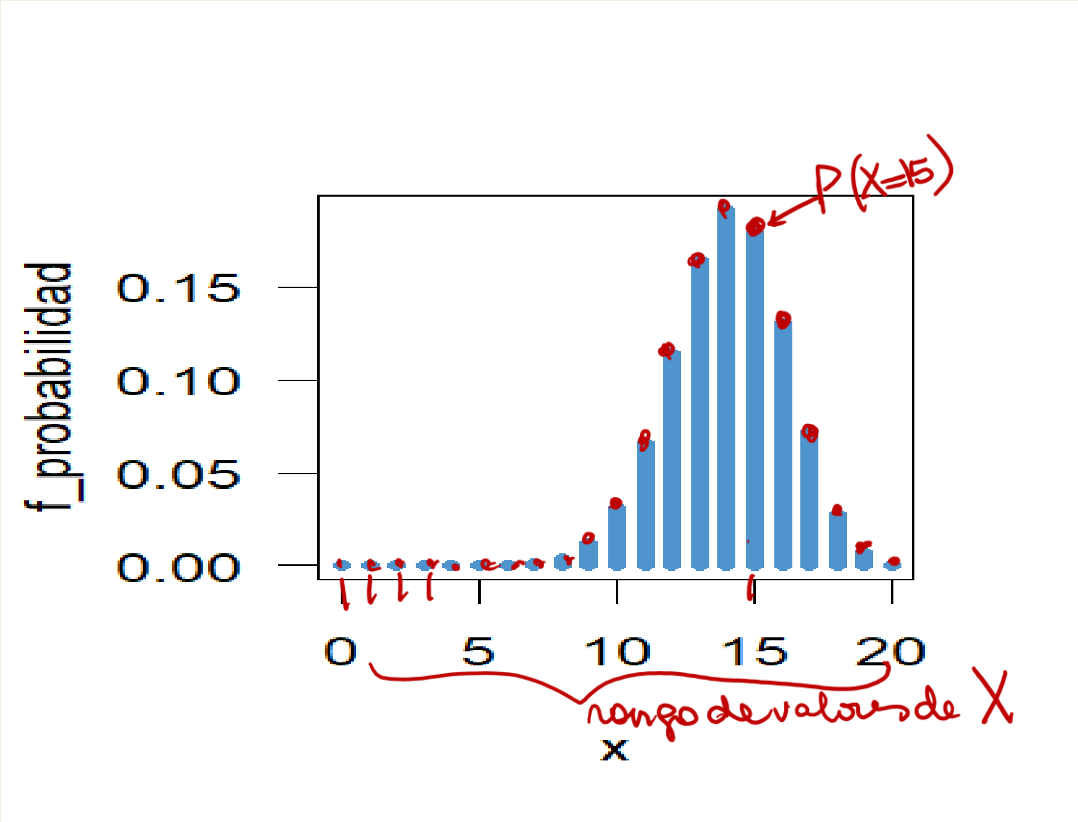
Variables aleatorias discretas

: toma valores en un conj. discreto de nº reales.

Se describen mediante una función de probabilidad P*(X = x)*

$\overbrace{\text{P(X=x)}}^{P_X(x)}$ debe cumplir:

- P(X=x) $\geq$ 0
- $\sum_{i=0}^{i=n}$ P(X=xi) =1



Parámetros básicos de una VAD

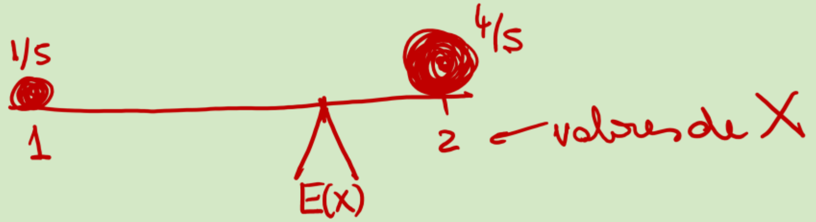
Media o esperanza o valor esperado:

Es una medida de tendencia central. Indica “alrededor” de que valor se encuentran distribuidos los valores de la variable aleatoria. Se mide en la misma unidad que la variable.

E(X) = m = $\sum_{i=1}^{i=n} \boldsymbol{xi} \,.\, \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{xi})$

o μ

Ej:

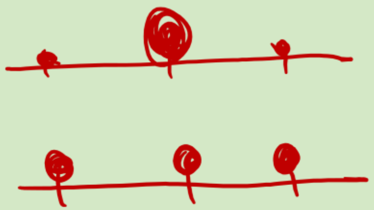


E(X) es el punto de equilibrio o centro de masa de las probabilid.

Varianza:

Es una medida de dispersión. Indica cuán alejados están los valores de la variable en relación con la media, Se mide la unidad que la variable al cuadrado.

proporcionados según la probabilidad



+ dispersión = + varianza

V(X) = σ^2x = $\sum_{i=1}^{i=n} (\boldsymbol{xi} - \mu)^2 \,.\, \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{xi})$

Desvio = $\sqrt{V(x)}$

Distribuciones discretas más conocidas

- **Binomial**: # éxitos en n repeticiones de experiencias dicotómicas indep

Ej: Observo en 20 personas en la guardia la cantidad de personas con gripe

- **Geométrica**: # experiencias hasta el primer éxito.

Ej: Cuento # de personas que llegan a guardia hasta el primero con gripe.

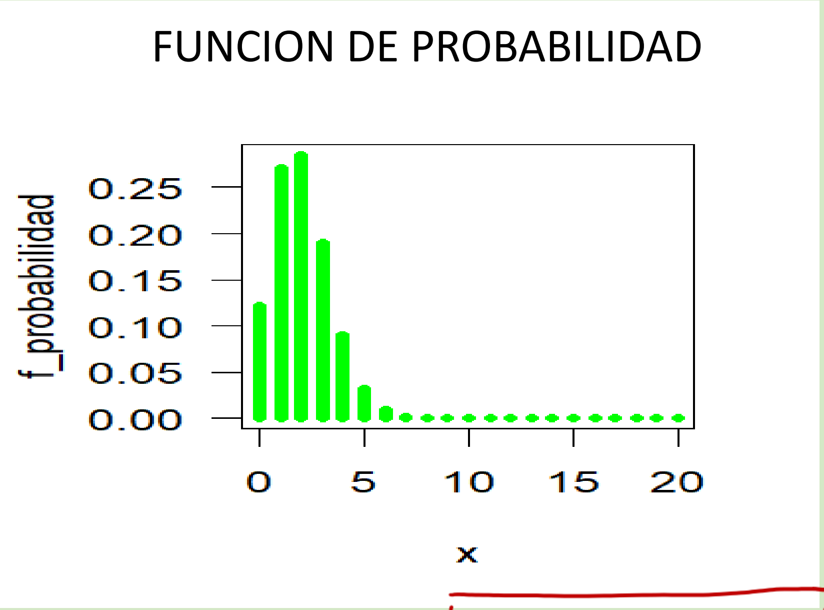
- **Hipergeométrica**: # éxitos en n extracciones sin reposición

- **Poisson**: # éxitos ocurridos en un intervalo fijo de tiempo (o longitud)

Ej: Observo la cantidad de personas que ingresa a guardia entre las 8 y las 9 hs.

Ejemplo Distribución Binomial

Se controla en 20 pacientes la cantidad que tiene gripe.
Se sabe que, en la población, la probabilidad de tener gripe es 0,1.
X: cantidad de pacientes con gripe en 20 que llegaron a guardia.



$\boxed{X \sim Bi(n,p)}$

X ~ Binomial (20; 0,1)

$E(X) = np$

P(X=x) → $P(X=x) = \binom{20}{x} 0,1^{x} (1-0,1)^{20-x} \quad x=0,\ldots\ldots 20$

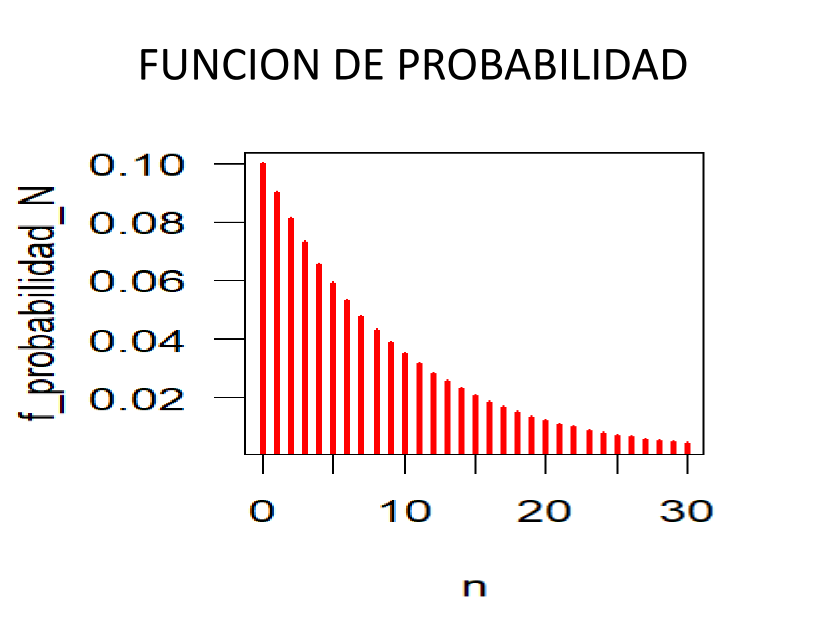
$V(X) = np(1-p)$

F(x) → FUNCIÓN DISTRIBUCIÓN : $F(x) = P(X \leq x)$

Ejemplo Distribución Geométrica

Se registran pacientes en la guardia hasta que aparece el primero que tiene gripe.
Se sabe que, en la población, la probabilidad de tener gripe es 0,1.
N: cantidad de pacientes que llegaron a guardia hasta el primero que tiene gripe.

$\boxed{N \sim Geo(p)}$



N ~ Geométrica(0,1)

P(X=x) → $P(X=x) = (1-p)^{x-1}\, p \qquad x = 1, 2, \ldots$

F(x) → $F(x) = \mathbb{P}(X \leq x)$

Ejemplo Distribución Poisson

Las personas ingresan a guardia según una tasa de arribos de 15 por hora.

K: cantidad de pacientes que llegaron a guardia entre las 8 y las 9 hs .

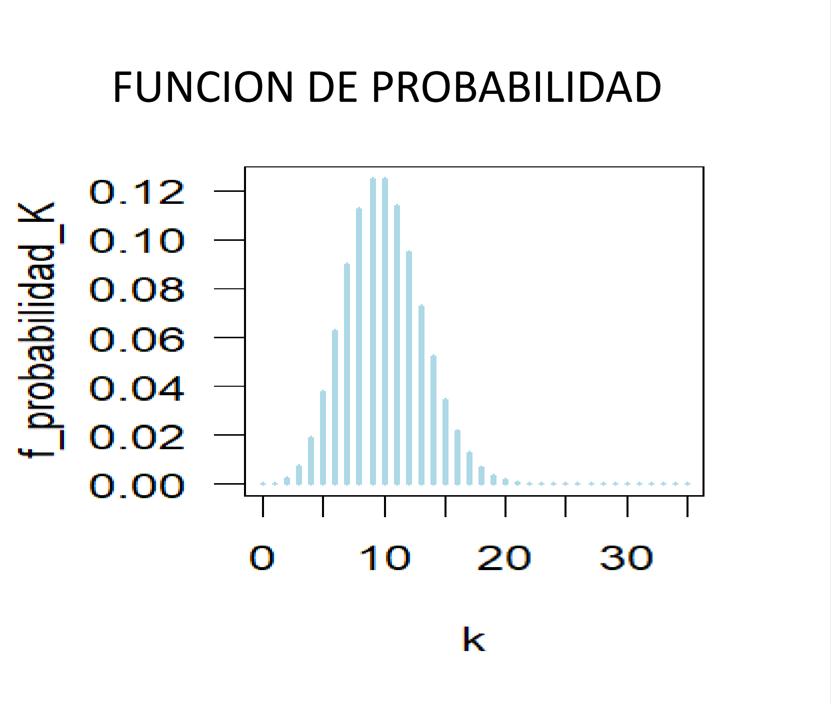
$\boxed{K \sim \text{Poisson}(\mu)}$

$E(X) = \mu$

K ~ Poisson(15 pers/hora)

P(X=x) →

F(x) →



Variables aleatorias en R

Para cada distribución:

dxxx(x, ...) # Función de probabilidad p(x) (para VAD) ó densidad f(x) (para VAC)

parámetros de la V.A.

pxxx(q, ...) # Función de distribución acumulada hasta q, F(q)= P(X <= q)

qxxx(p, ...) # Cuantil: valor **q** para el cual P(X <= q) es el valor de probabilidad p dada.

rxxx(n, ...) # Generador de números aleatorios de esa variable

Veamos en R!!

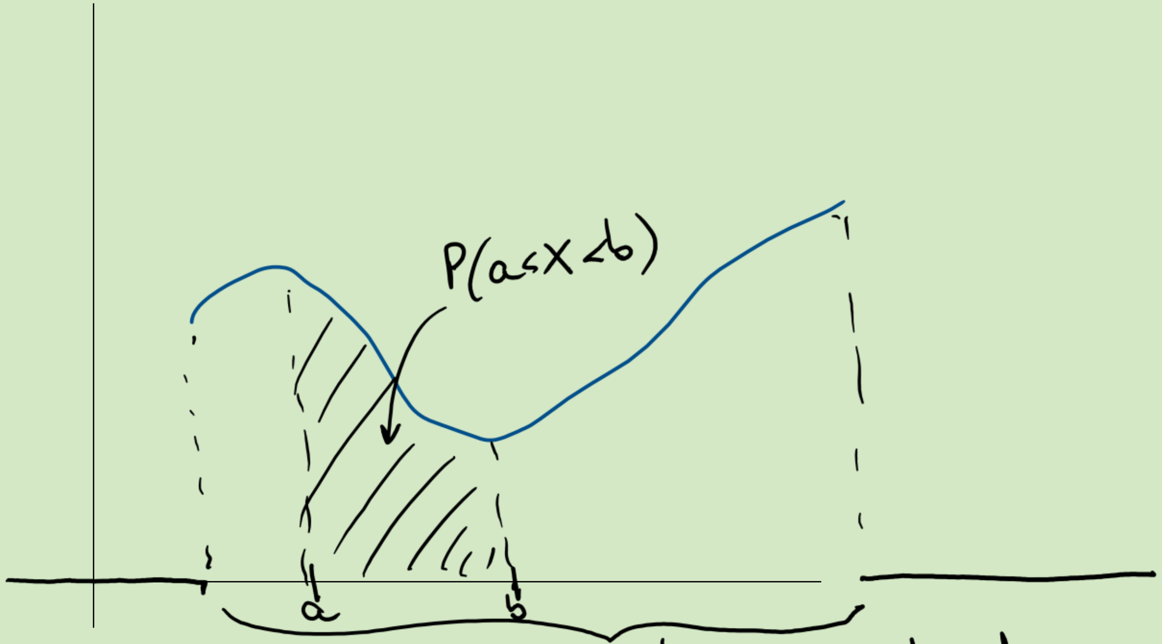
Variables aleatorias continuas

<u>Obs</u>: una VAC tiene rango o soporte en un intervalo (o unión de intervalos).

Se describen mediante una <u>función de densidad *f(x)*</u>

f(x) debe cumplir:

- $f(x) \geq 0$
- $\underbrace{\int_{-\infty}^{+\infty} f(x)dx}_{\text{área total bajo f vale} = 1} = 1$



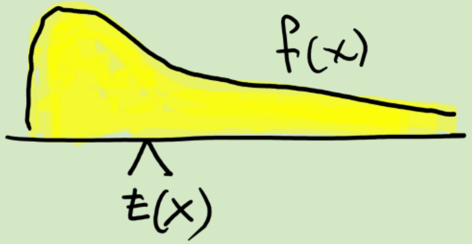
<u>PROPIEDAD</u>: $\mathrm{P}(\mathrm{a} < \mathrm{X} < \mathrm{b}) = \int_a^b f(x)dx = área\ bajo\ f\ en\ (a,b)$

Parámetros básicos de una VAC

Media o esperanza o valor esperado:

Es una medida de tendencia central. Indica “alrededor” de que valor se encuentran distribuidos los valores de la variable aleatoria. Se mide en la misma unidad que la variable.

$$\mathrm{E(X)} = \underset{\mu}{m} = \int_{-\infty}^{+\infty} x f(x)\,dx$$



Varianza:

Es una medida de dispersión. Indica cuán alejados están los valores de la variable en relación con la media. Se mide la unidad que la variable al cuadrado.

$$\mathrm{V(X)} = \sigma^2 x = \int_{-\infty}^{+\infty} (x-m)^2 f(x)\,dx$$

$$D(X) = \sqrt{V(X)}$$

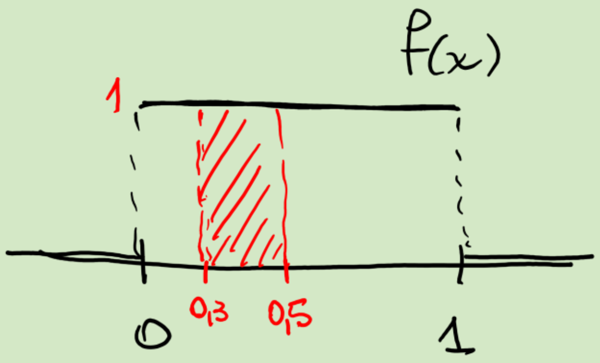
Distribuciones continuas más conocidas

- Uniforme
- Exponencial
- Normal
- T de Student
- Chi-cuadrado

Distribución Uniforme

Ejemplo: se elige un punto al azar en el intervalo (0,1)

X: número elegido



X ~ Uniforme$\overbrace{(0,1)}^{\text{intervalo}}$

Densidad → $f(x) = \begin{cases} 1 & \text{si } 0<x<1 \\ 0 & \text{otro caso} \end{cases}$

Función de distribución → $F(x) = P(X \le x)$

En gral: $X \sim U(a,b)$

$$E(X) = \frac{a+b}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$

OBS $P(X=0.3) = 0$

$P(0{,}3 < X < 0.5) = 0{,}2 * 1 = 0{,}2$

0,3 0,5

$\hookrightarrow P(X<0.5) - P(X<0.3) =$

$F(0{,}5)$ $= F(0.5) - F(0.3)$

0,3

Distribución Exponencial

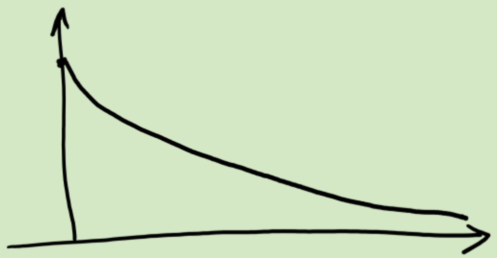
Ejemplo: se toma el tiempo (años) hasta que falle un artefacto. Se sabe que la tasa de fallas es 1 cada 5 años.

T: tiempo de vida en años

$T \sim \mathcal{E}(\lambda)$

λ = tasa de fallos
= intensidad de arribos

T ~ Exp(1/5)



$E(T) = \frac{1}{\lambda}$

Densidad → $f(t) = \lambda e^{-\lambda t}$

$V(T) = \frac{1}{\lambda^2}$

Función de distribución → $F(t) = \mathbb{P}(T \leq t)$

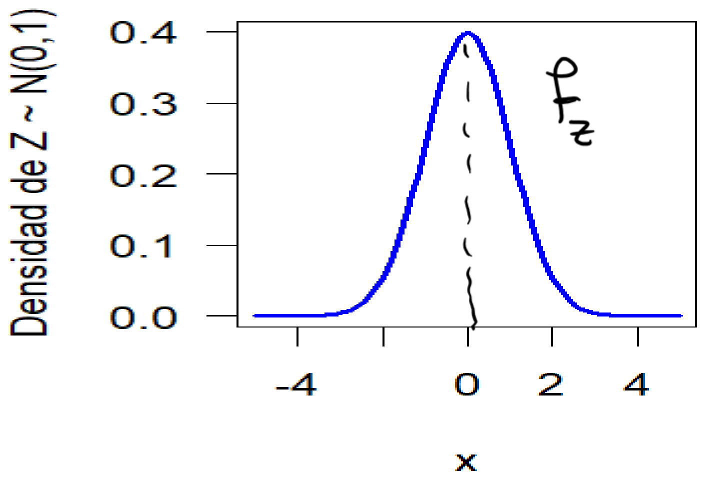
Distribución normal estándar $Z \sim N(0, 1)$

media $= E(Z)$

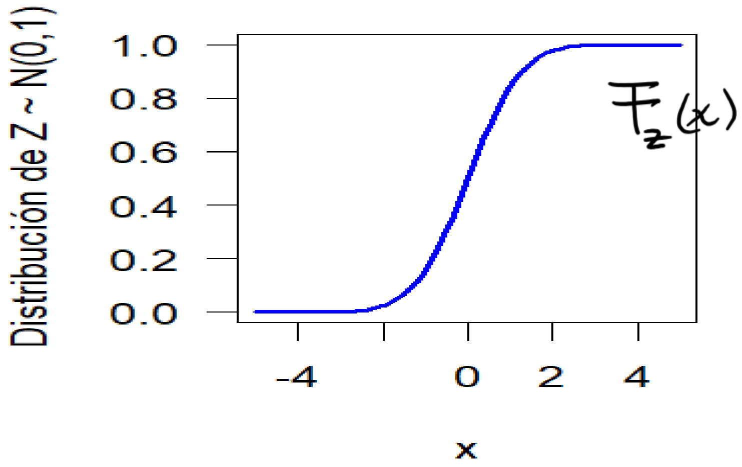
desvio $= \sqrt{V(Z)}$

Distribución Gaussiana

Densidad → $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$



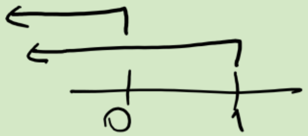
Función de distribución → $\mathbb{P}(Z \leq x)$



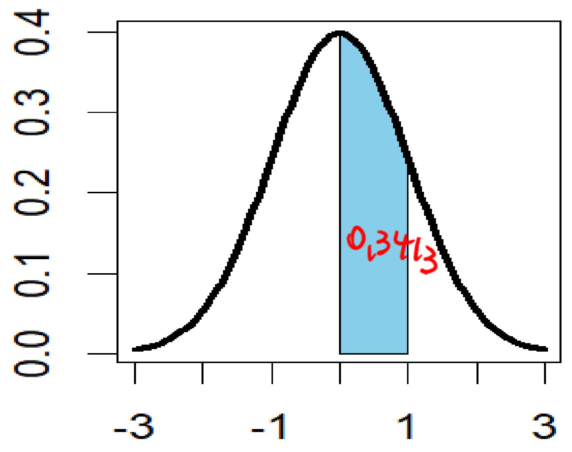
Ejercicio: se tiene una variable **Z ~ N(0, 1)**

1. A qué probabilidad corresponde la parte pintada de celeste?

$P(0<Z<1)$



OBS Z es VAC $\Rightarrow P(Z=a)=0$



2. Cuánto vale esta probabilidad?

$P(0<Z<1) = P(Z<1) - P(Z<0) = 0{,}8413477 - 0{,}5 =$

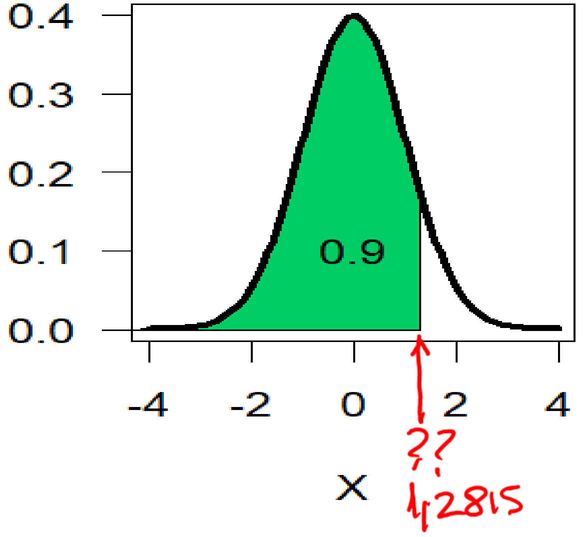
$= \underline{0{,}3413}$

3. Hallar el valor de z tal que P(Z<z) da la probabilidad pintada de verde.

Busco z tal que $P(Z<\overset{??}{z})=0.9$

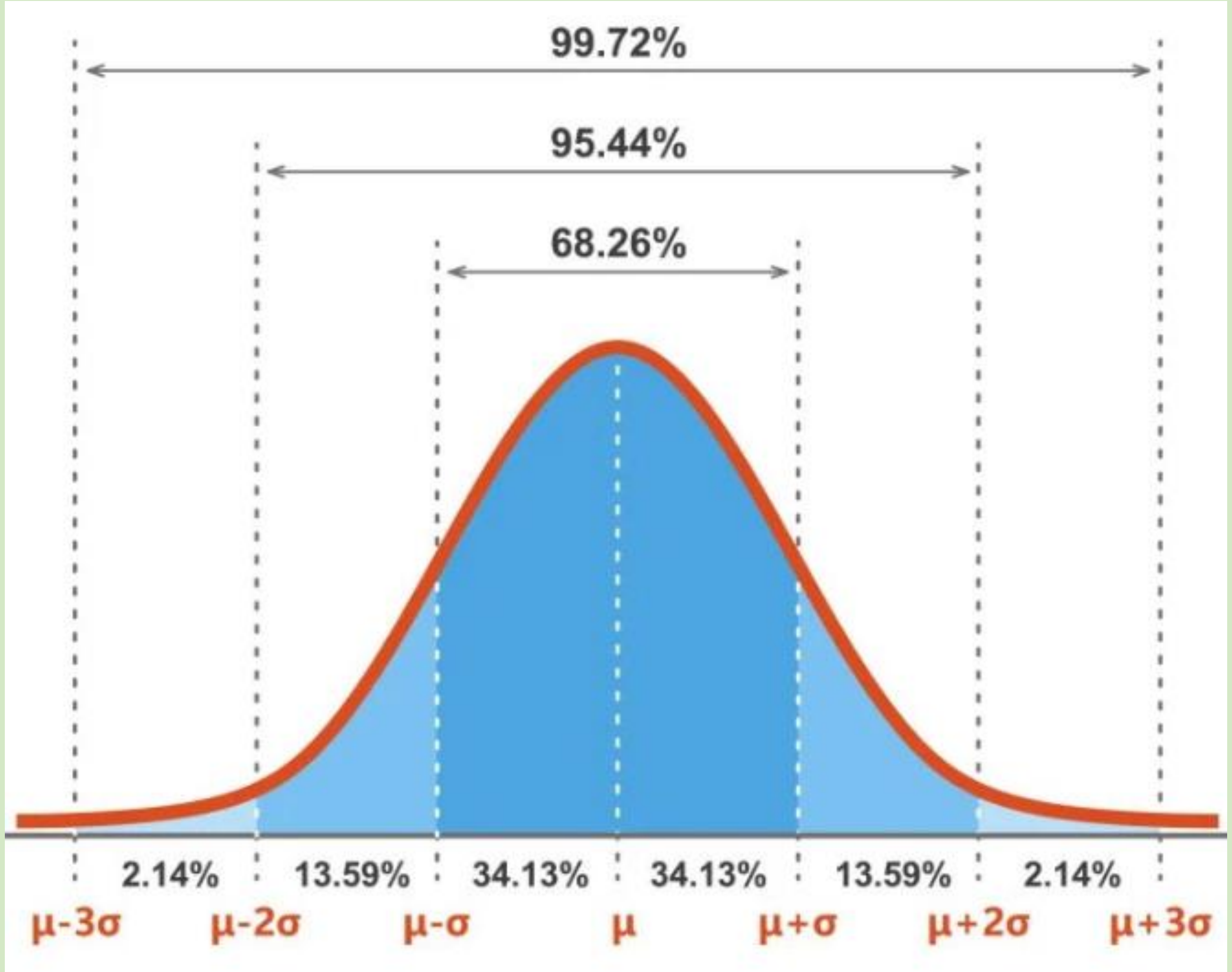
esto z se llama "cuantil 0.9"

Se calcula con qnorm(0.9, 0, 1) = 1,2815

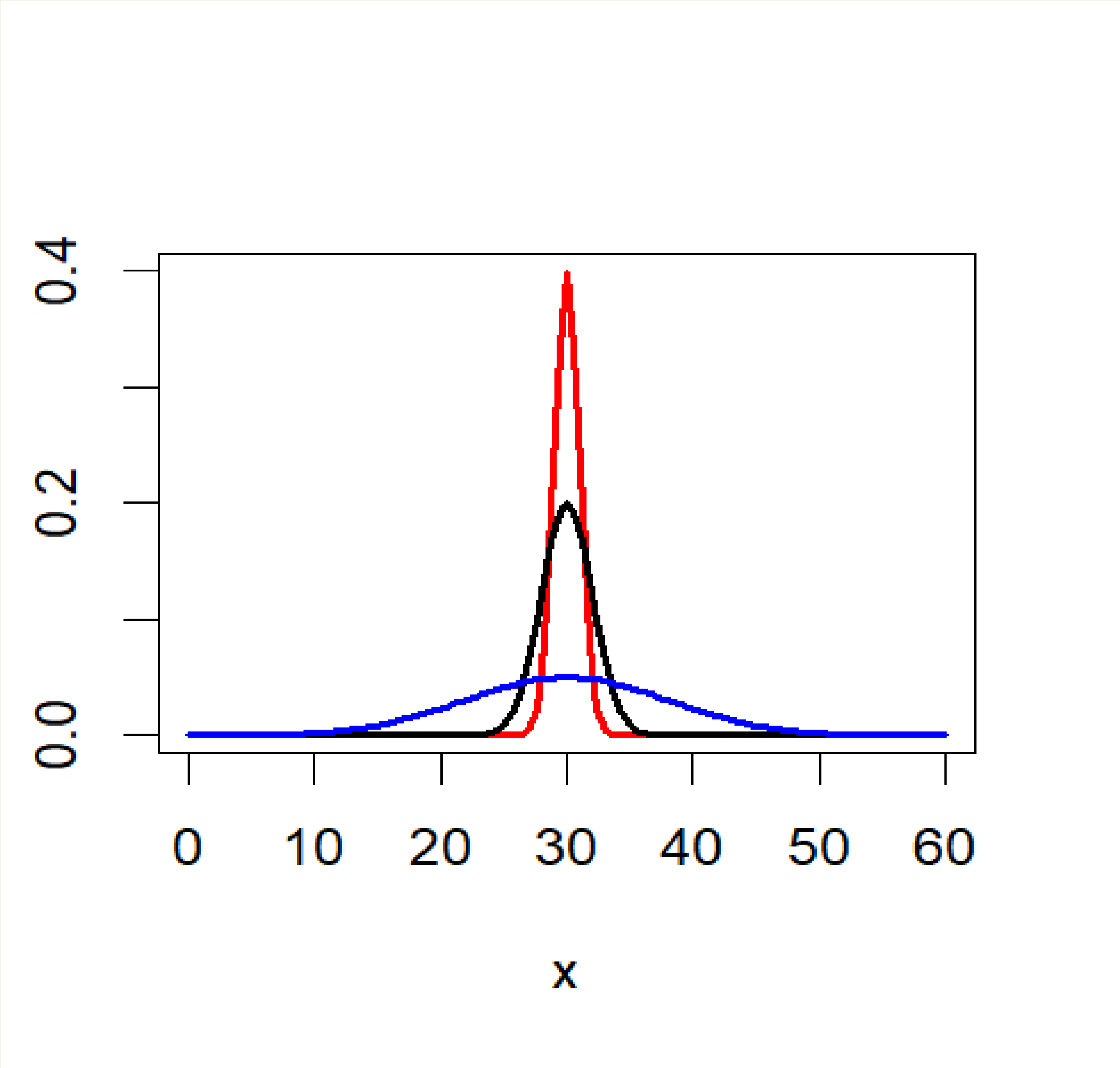


Distribución Normal $X \sim N(\mu, \sigma^2)$

Propiedad



¿Cómo afecta el desvío a la Distribución Normal?



Media= 30

En rojo: desvío =1

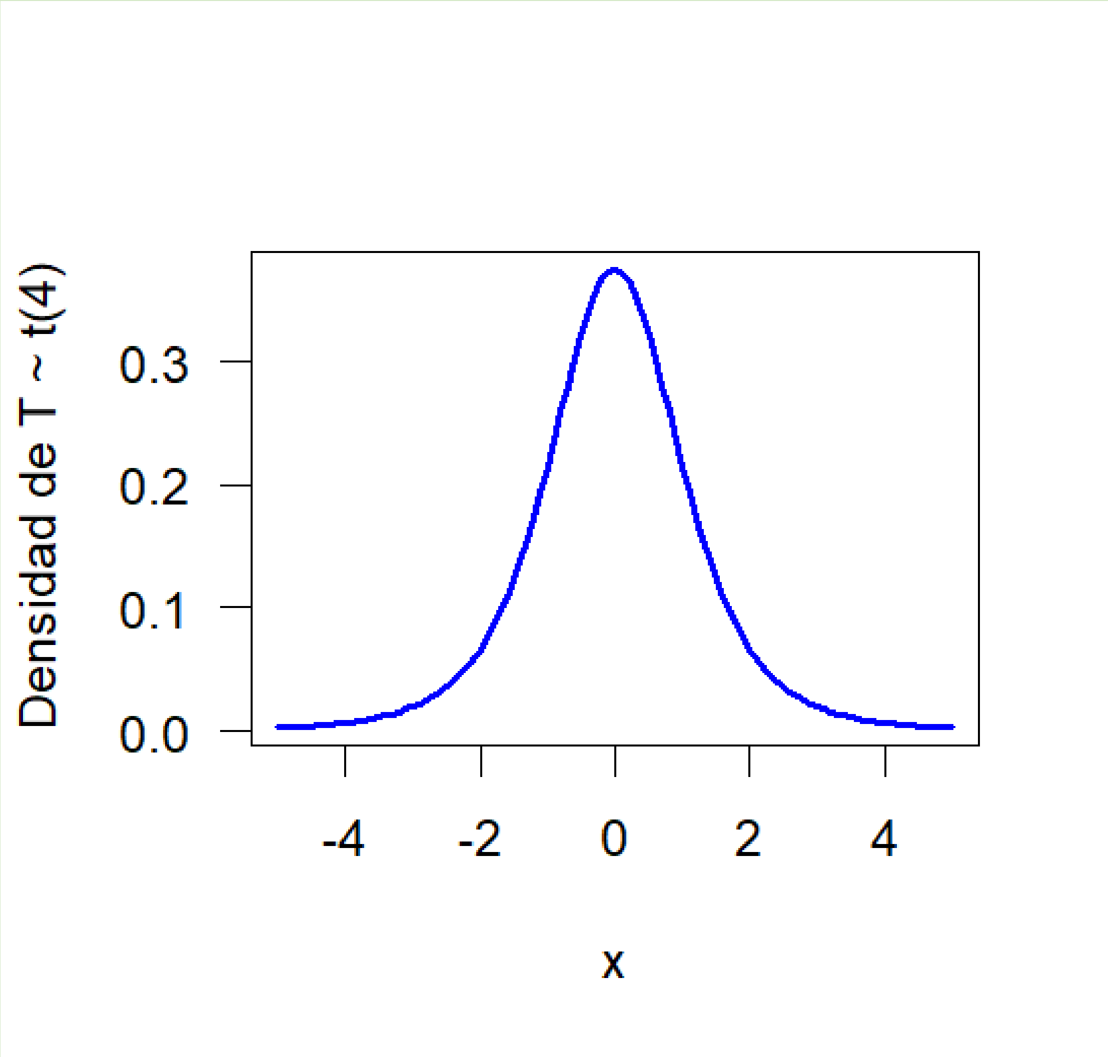
En negro: desvío = 2

En azul: desvío = 8

Distribución t de Student

T ~ t(v)

Densidad →

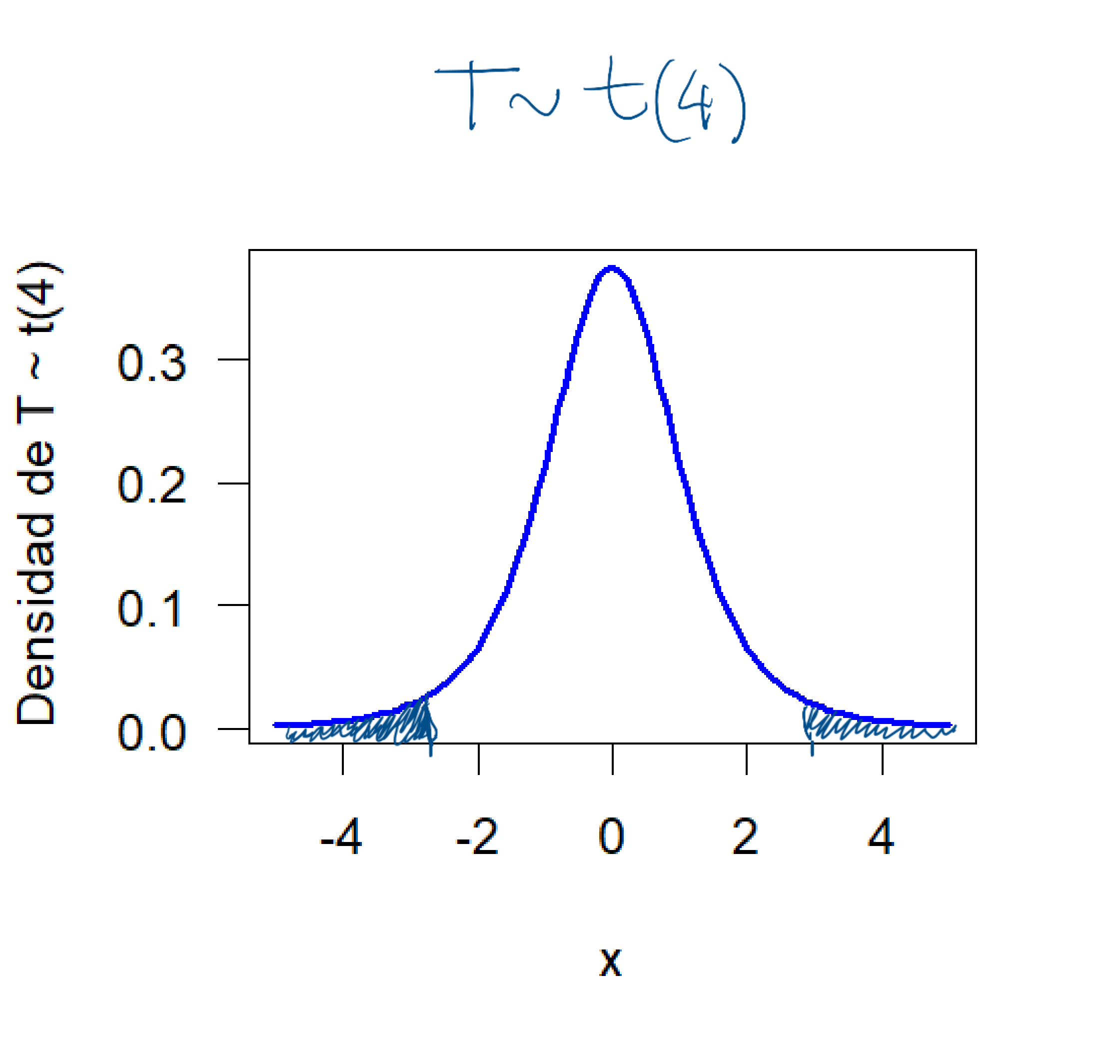


<u>Propiedad:</u>

Si v->infinito entonces T -> Z

Distribución t de Student

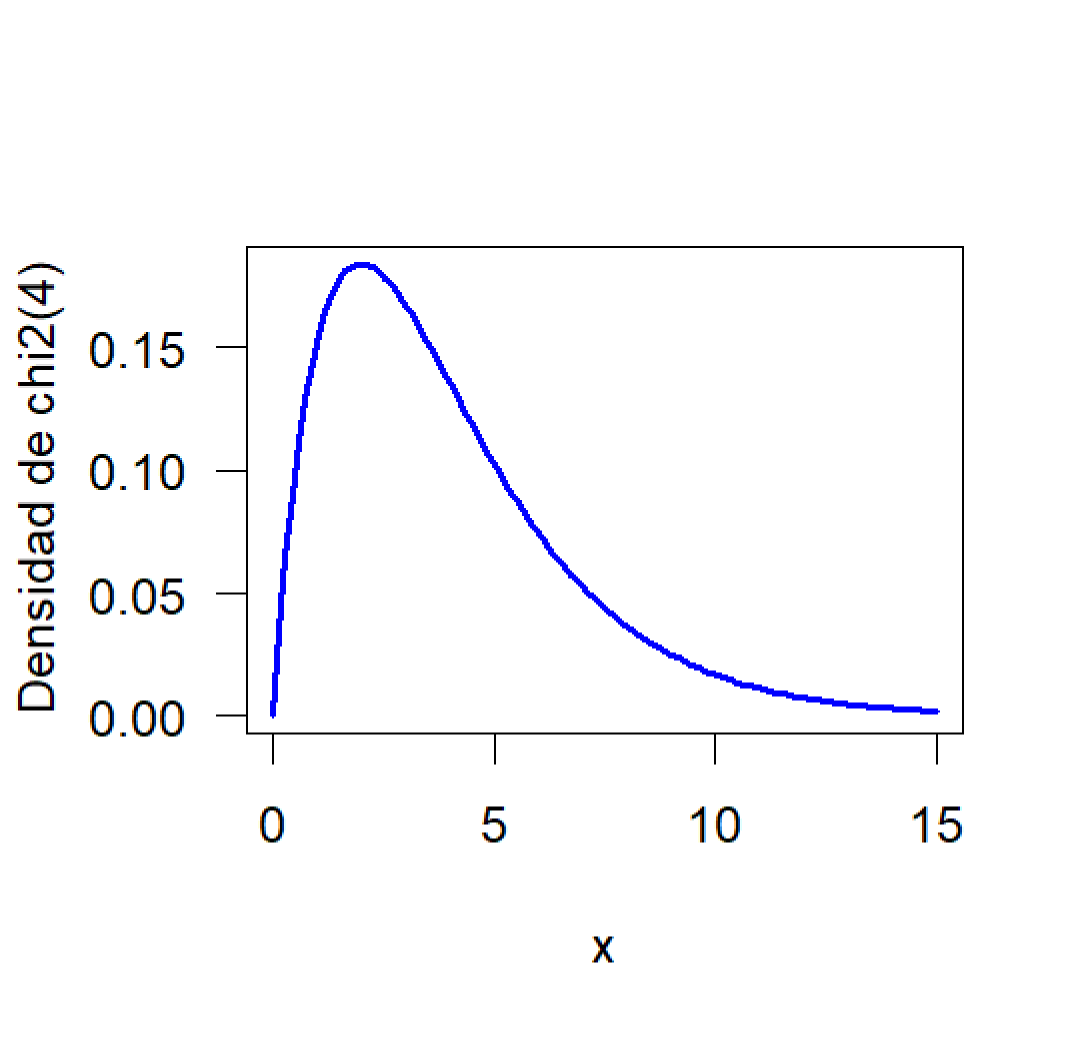
Pregunta:
La zona pintada corresponde a una probabilidad 0.05.
¿Dónde se ubica el valor -3 en el gráfico?



Distribución

X ~ χ^2 (v)

Densidad →



<u>Propiedad:</u>

χ^2 corresponde a un caso particular de distribución Gamma

Veamos en R!!