

Distribución normal

Fórmula de distribución normal

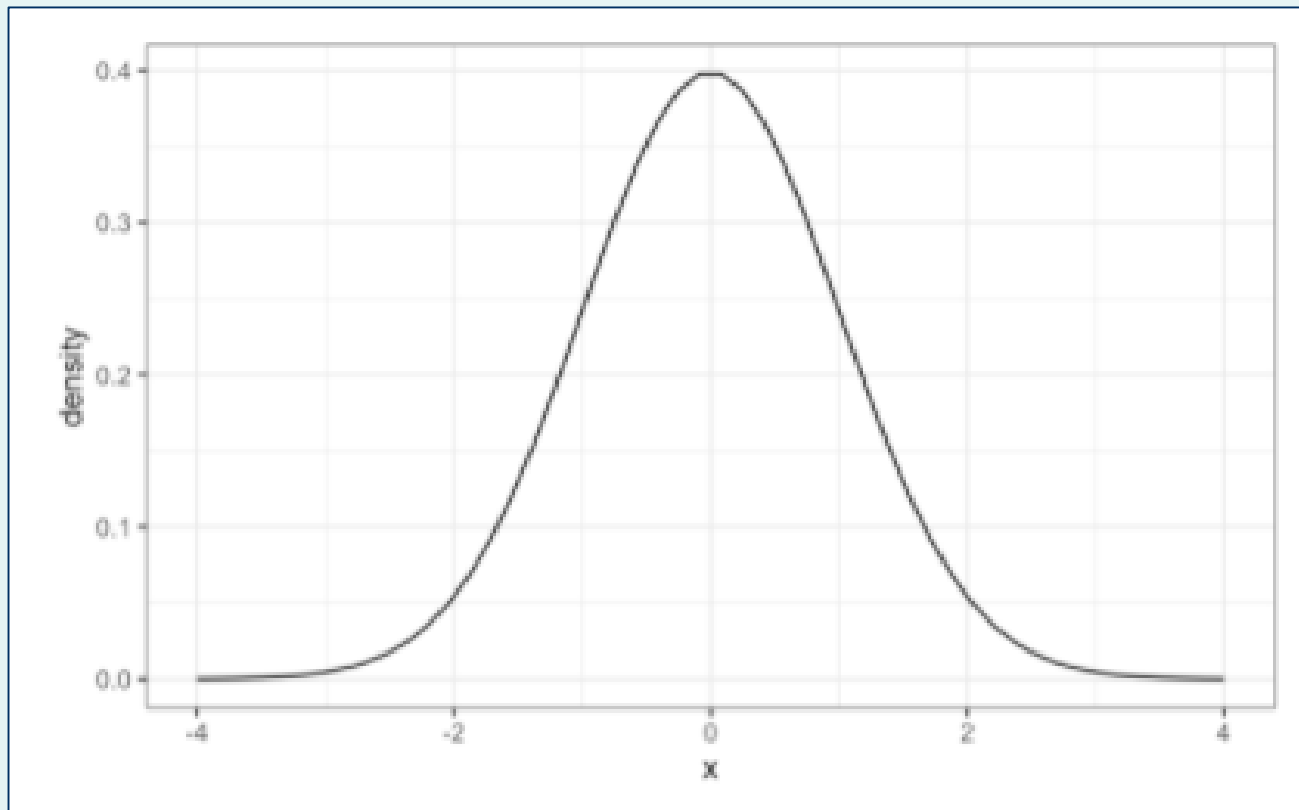
- Para un intervalo $a - b$ la proporción de valores en ese intervalo se calcula:

$$\Pr(a < x < b) = \int_a^b \frac{1}{\sqrt{2\pi s}} e^{-\frac{1}{2} \left(\frac{x-m}{s} \right)^2} dx$$

- Siendo m es el promedio-mean, y s es la desviación estándar-standard deviation.

Distribución normal

Resumen de tablas de datos en:
Promedio \pm desviación estándar ($m \pm s$)



Valores importantes para una lista arbitraria de números

- Para una lista de números de un vector x
 - Promedio (mean) - m :

```
m <- sum(x) / length(x) → mean(x)
```

- Desviación estándar (sd) - s :

```
s <- sqrt(sum((x-mu)^2) / length(x)) → sd(x)
```

Se interpreta como la distancia promedio entre los valores y su promedio

Para nuestros datos de alturas

```
index <- heights$sex == "Male"
```

```
x <- heights$height[index]
```

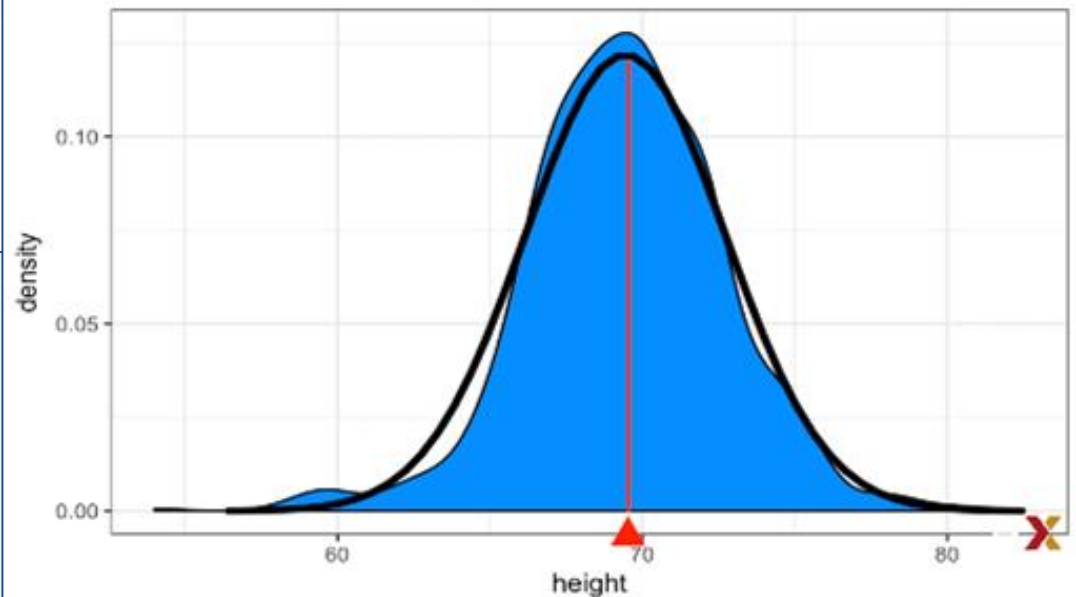
```
m <- mean(x)
```

```
s <- sd(x)
```

```
c(average = m, sd = s)
```

```
#> average      sd
```

```
#> 69.31      3.61
```



Unidades estándar

La unidad estándar de un valor nos dice cuán lejos, en desviaciones estándar está del promedio.

Para un valor x :

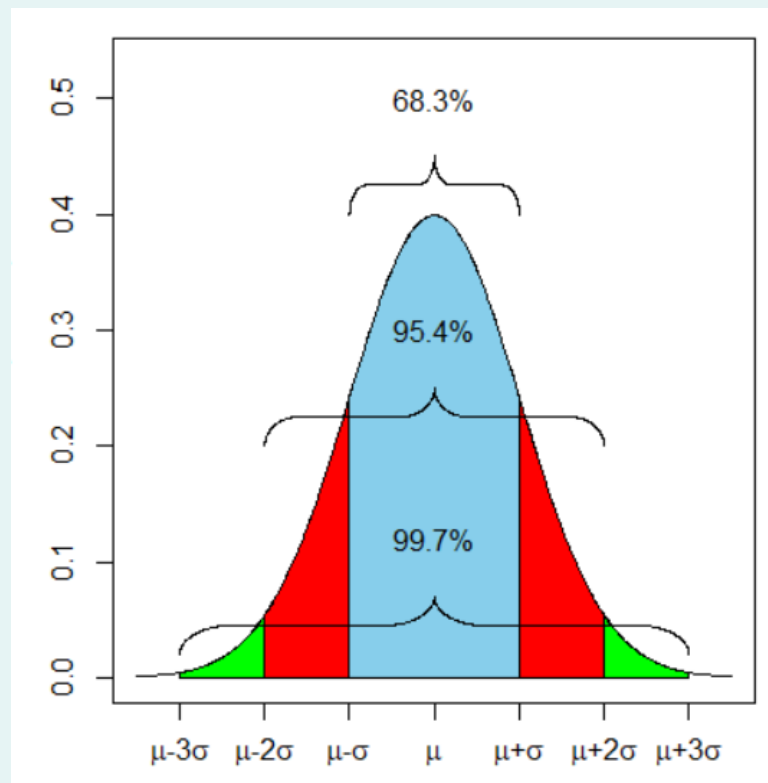
$$z = (x - m) / s \rightarrow \text{scale}(x)$$

Siendo s y m la desviación estándar y el promedio respectivamente.

- En alturas {
- Persona de altura promedio $z=0$
 - Uno de los más altos $z \approx 2$
 - Uno de los más pequeños $z \approx -2$
 - Una altura muy rara $z > 3$ o $z < -3$

Distribución normal estándar

- Distribución con $m=0$ y $s=1$
- Se describe con la regla de 68-95-99.7



Distribución normal

Decimos que los datos son normalmente distribuidos con promedio m y desviación estándar s si la distribución de probabilidades se define por:

$$F(a) = \text{pnorm}(a, m, s)$$

Estimamos probabilidad de que un hombre sea más alto que 70.5inches

```
library(tidyverse)

library(dslabs)

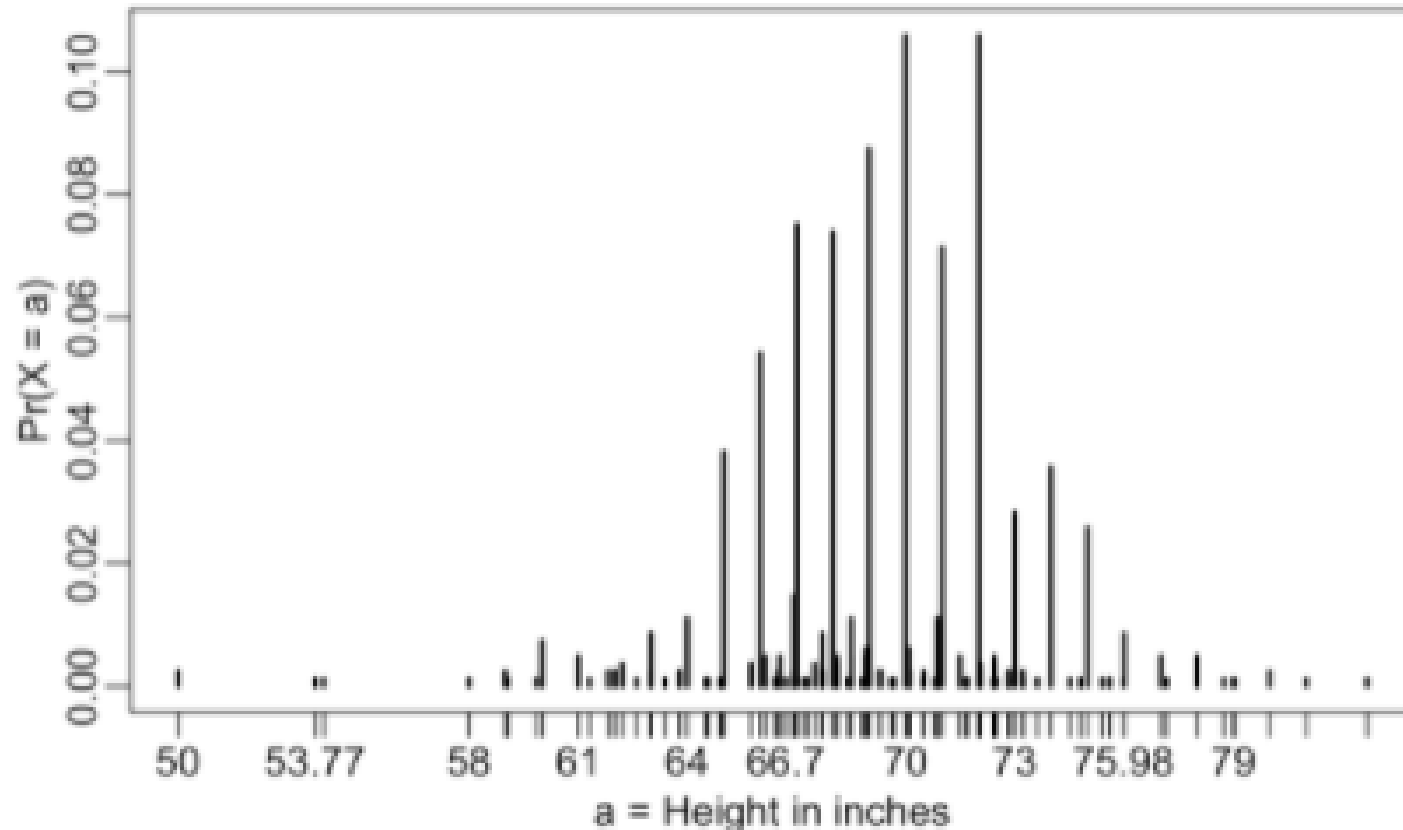
data(heights)

x <- heights %>% filter(sex=="Male") %>% pull(height)

1 - pnorm(70.5, mean(x), sd(x))
```


Discretización y aproximación normal

Si para nuestros datos de altura consideramos cada altura como una categoría única y graficamos la proporción de estudiantes que tiene cada altura.



Discretización de datos

Rango alturas $x1 \leq X \leq x2$	Proporción de estudiantes $\text{mean}(x \leq x2) - \text{mean}(x \leq x1)$	Aprox. a distribución normal $\text{pnorm}(x2, m, s) - \text{pnorm}(x1, m, s)$
$67.5 \leq X \leq 68.5$	0.115	0.103
$68.5 \leq X \leq 69.5$	0.119	0.11
$69.5 \leq X \leq 70.5$	0.122	0.108
$70.1 \leq X \leq 70.9$	0.0222	0.0836