Introducción a distribuciones

Ana Cecilia Plavan, Andrés Andruskevicius

- Podemos pensar la distribución de datos como una manera compacta de describir una lista con muchos elementos.
- Con datos categóricos la distribución describe la proporción de datos en cada categoría (ej. 23%mujeres, 77%hombres).
- CDF (Distribución acumulada): Definir la distribución de una variable numérica es reportar la proporción de los datos menores a un valor X.

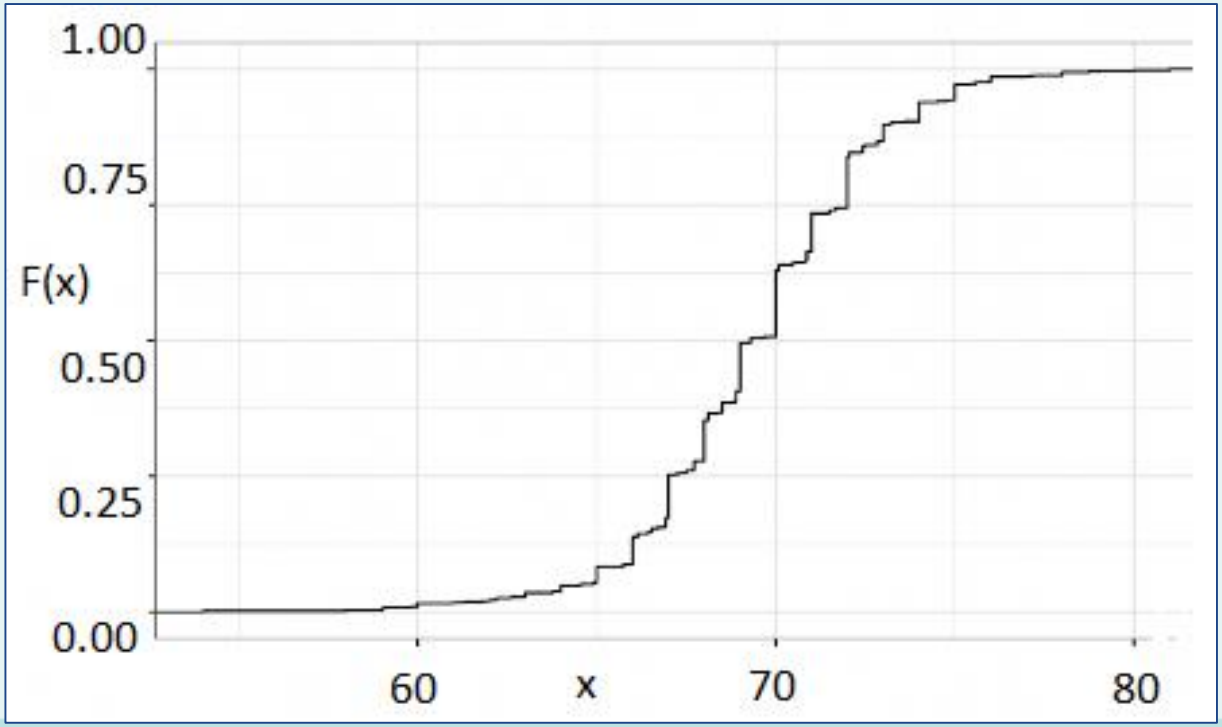
$$F(x) = \mathrm{P}[X \leq x]$$

- Tabla de distribución de datos categóricos

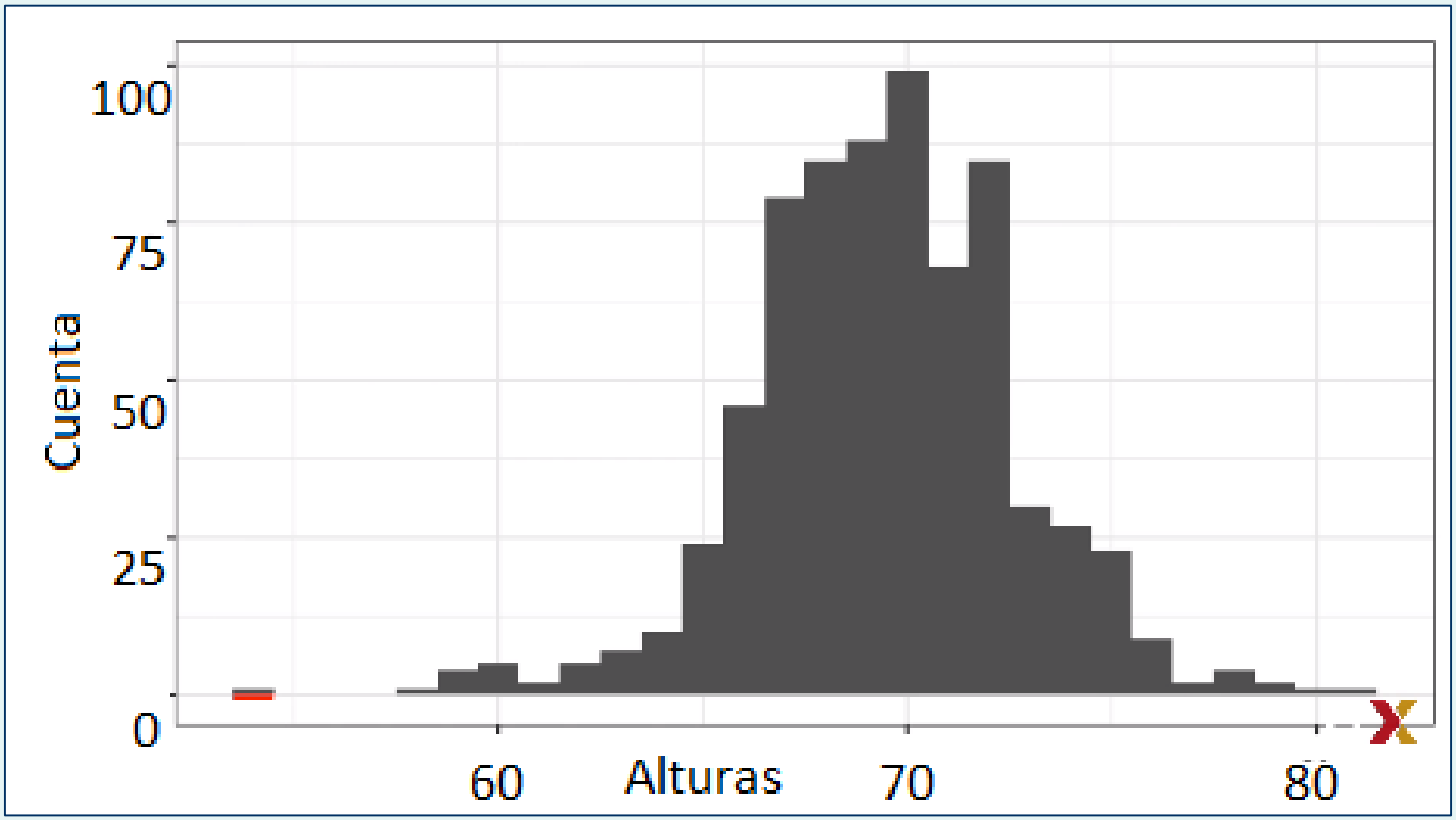
```
> prop.table(table(heights$sex))

    Female       Male
0.2266667 0.7733333
```

- Cabecera de datos y gráfico de distribución de alturas

```
> head(heights)
     sex height
1   Male     75
2   Male     70
3   Male     68
4   Male     74
5   Male     61
6 Female     65
```



Histograma de alturas



Calculamos CDF

```
# Definimos el rango de valores
a <- seq(min(my_data), max(my_data),
length = 100)

cdf_function <- function(x) {
# Calculamos la probabilidad de un valor
    mean(my_data <= x)
}
#Ejecutamos la función y graficamos
cdf_values <- sapply(a, cdf_function)
plot(a, cdf_values)
```

Resumiendo

- La CDF define una proporción de datos menores a un valor X o a.
- Para definir la proporción de datos mayores a X calculamos:

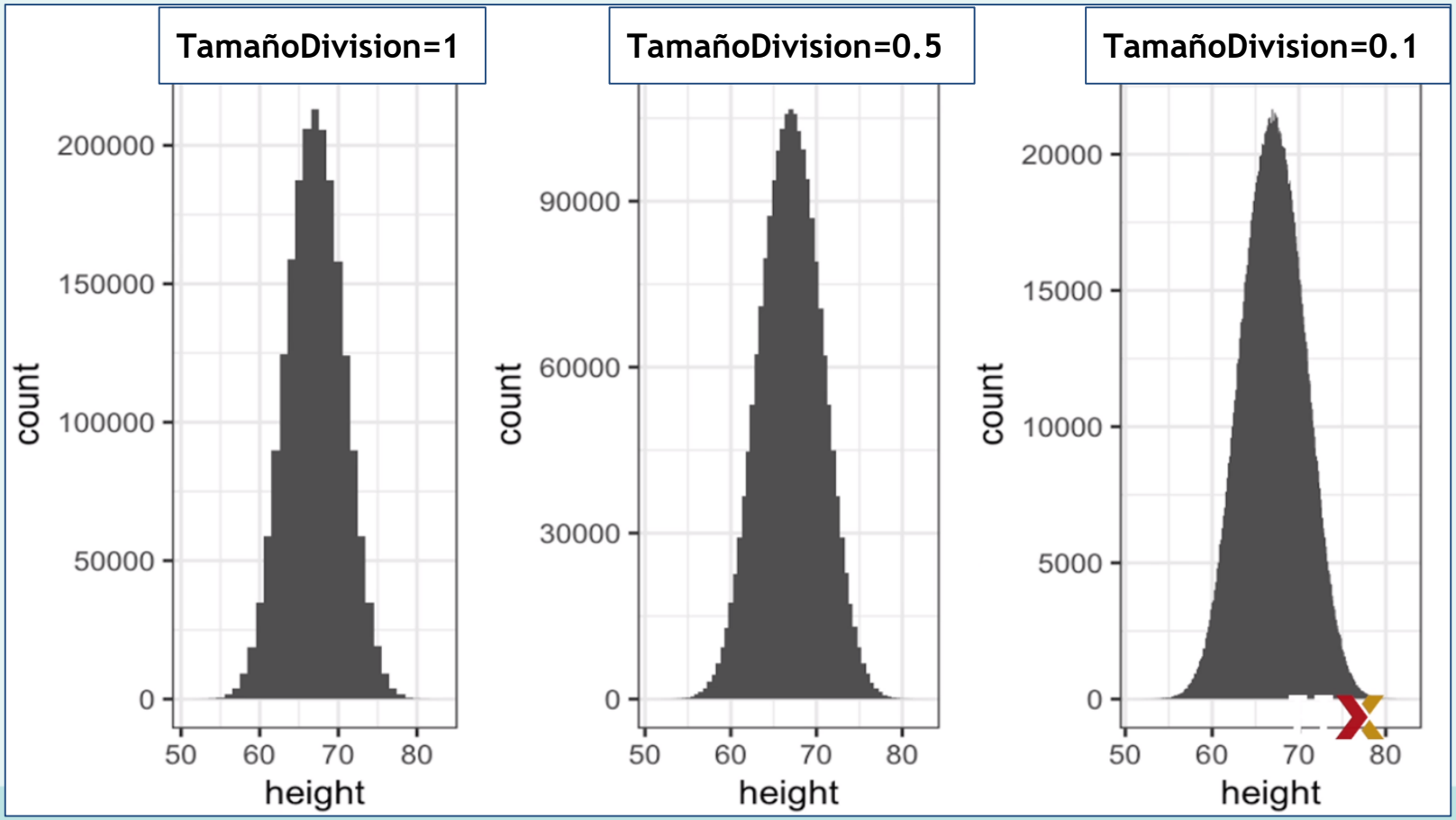
$$1 - F(a)$$

- Para definir la proporción de valores entre a y b calculamos:

$$F(b) - F(a)$$

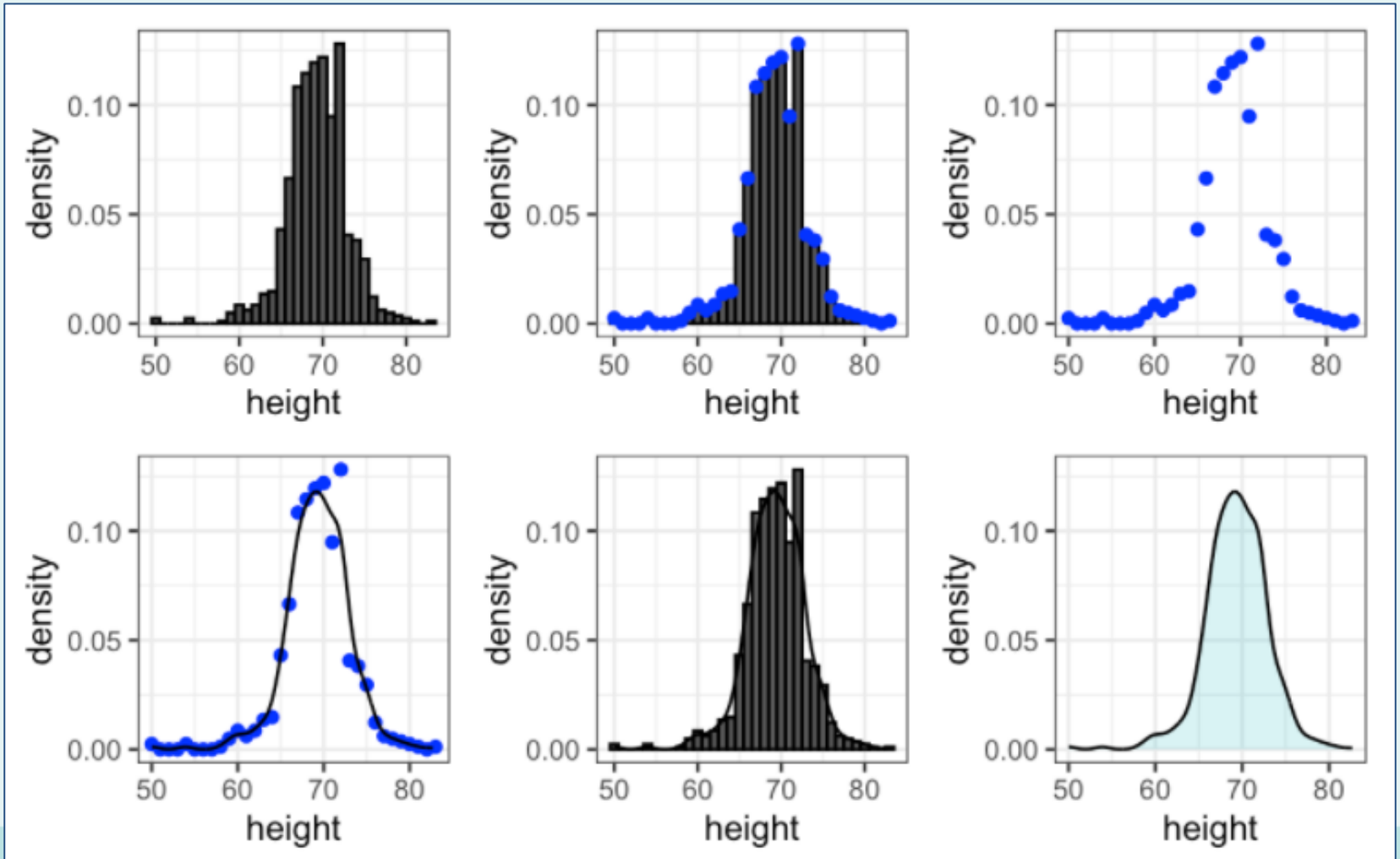
Gráfico de dispersión - distribución

- Gráfico que traza la línea entre las barras del histograma cuando el ancho de cada barra es muy pequeño.
- Aumentando la cantidad de medidas de las alturas hasta tomar un millon de valores obtenemos:

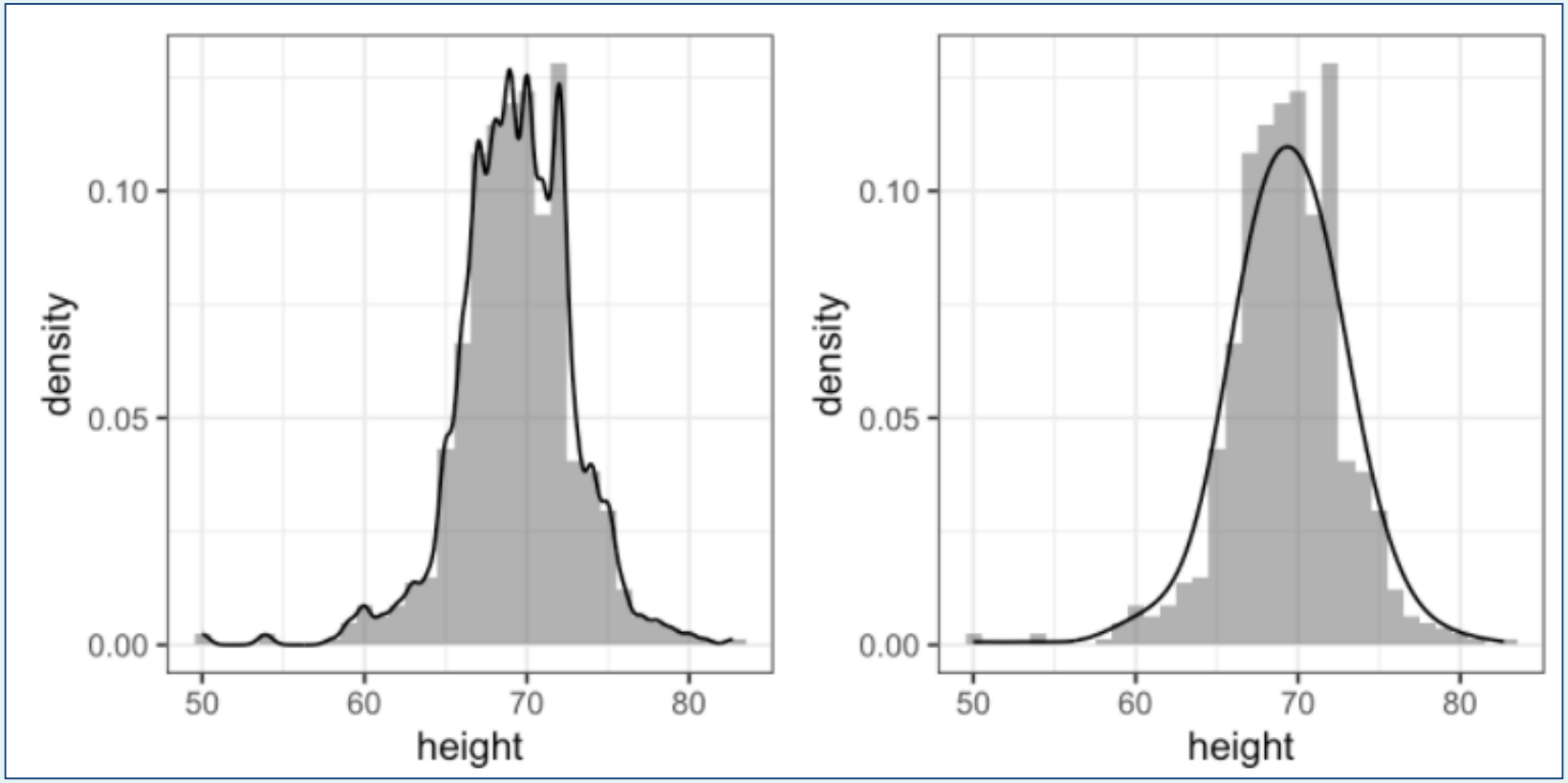


El gráfico de dispersión es un gráfico estimativo

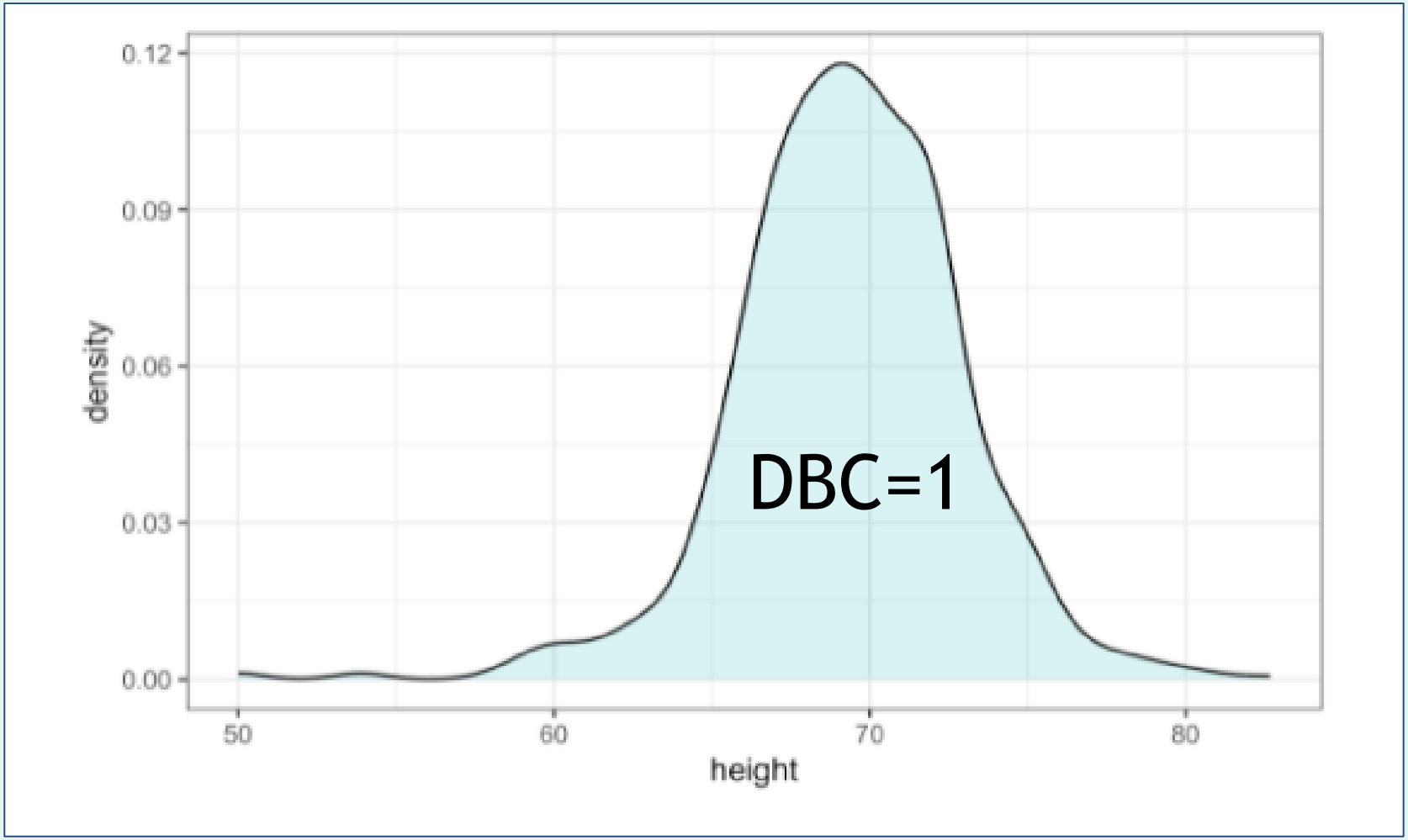
1. Realizamos un histograma con tamaño de rango apropiado para nuestros datos calculando la frecuencia en lugar de la cuenta.
2. Dibujamos la curva que pasa por los puntos de las alturas.



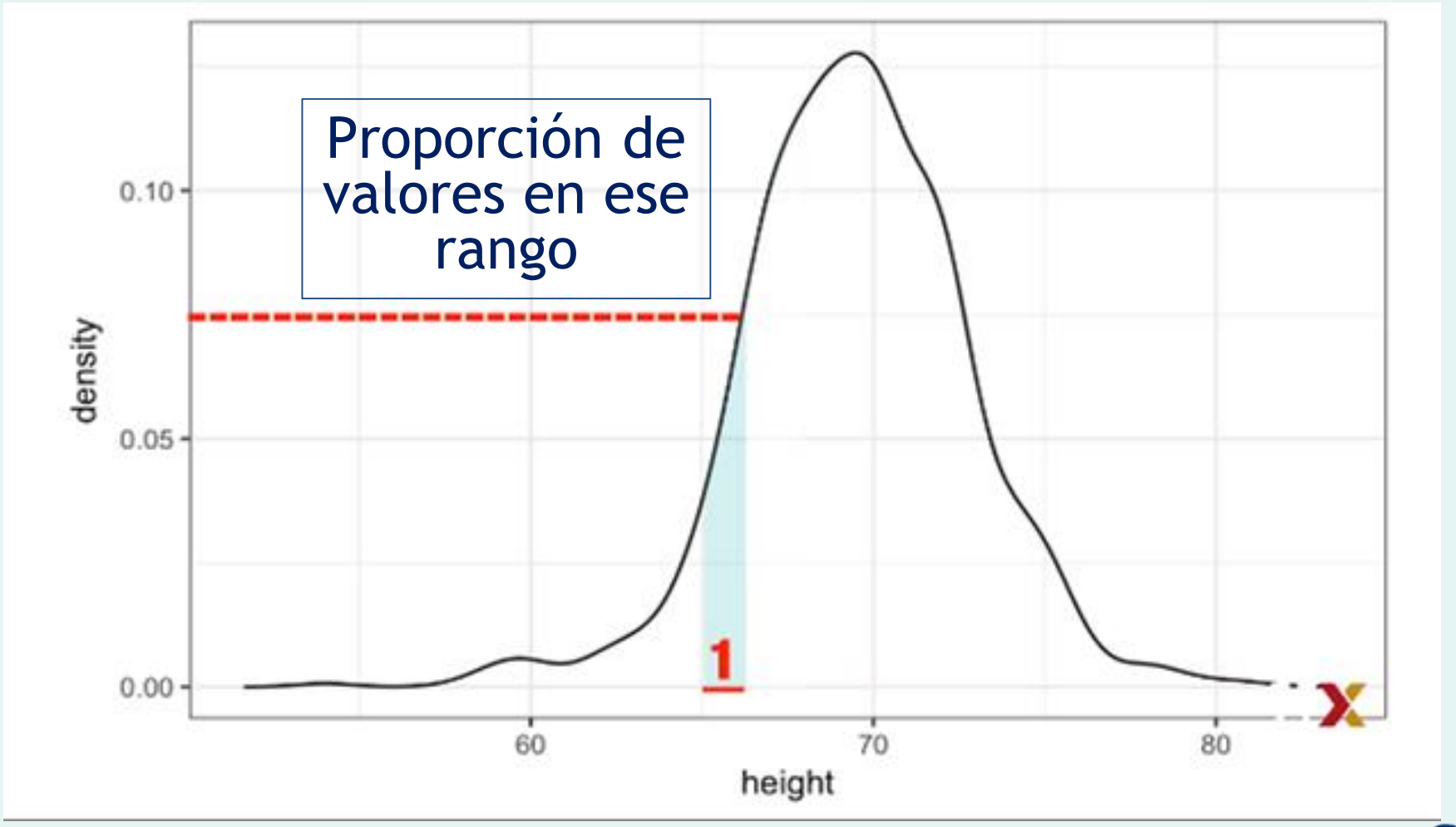
Función de distribución de datos



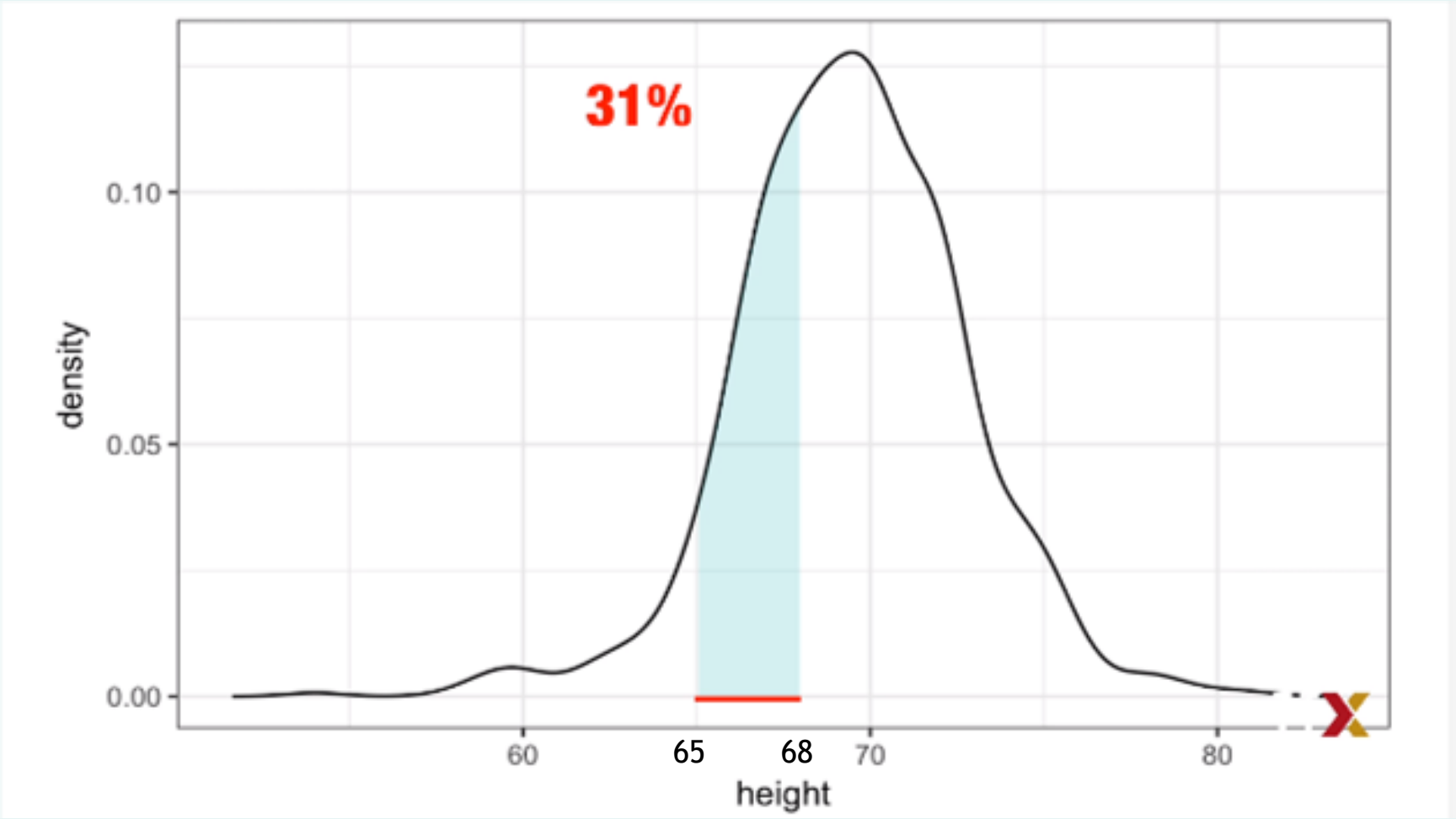
Eje Y en gráfico de distribución



Eje Y en gráfico de distribución



Análisis eje Y



Comparativa