# Taller Iniciación Estadística Aplicada a la Investigación

## Normalidad

Hugo J. Bello

2024/04

# La distribución de los datos

Cuando tenemos una muestra $x_1, \ldots, x_n$, a menudo nos hacemos preguntas cómo

- ¿Qué lugares o intervalos son más probables que otros?
- ¿Cuál es el *lugar central* en los datos?
- ¿Cómo de *esparcidos* están los datos?

Hemos venido resppondiendo a estas preguntas mediante medidas de concentración y dispersión. Pero hay un modelo matemático más preciso para esto y son las **distribuciones de probabilidad**

# Distribuciones

Las funciones de distribución $f(x)$, podemos entenderlas como curvas con las que podemos calcular la probabilidad de encontrar datos en cualquier intervalo usando precisamente el area bajo la curva en ese intervalo.

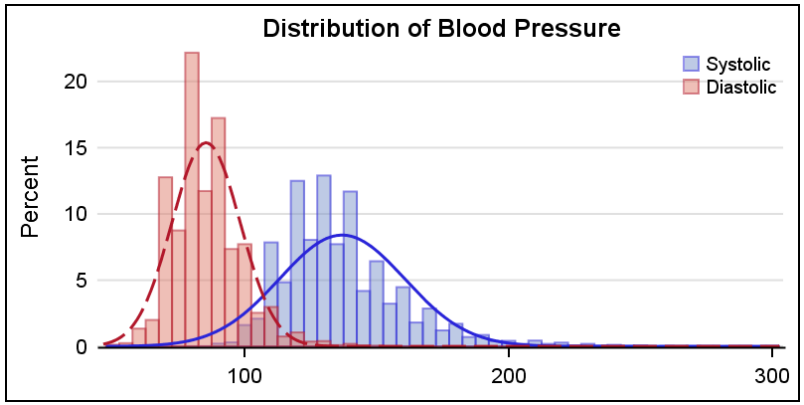
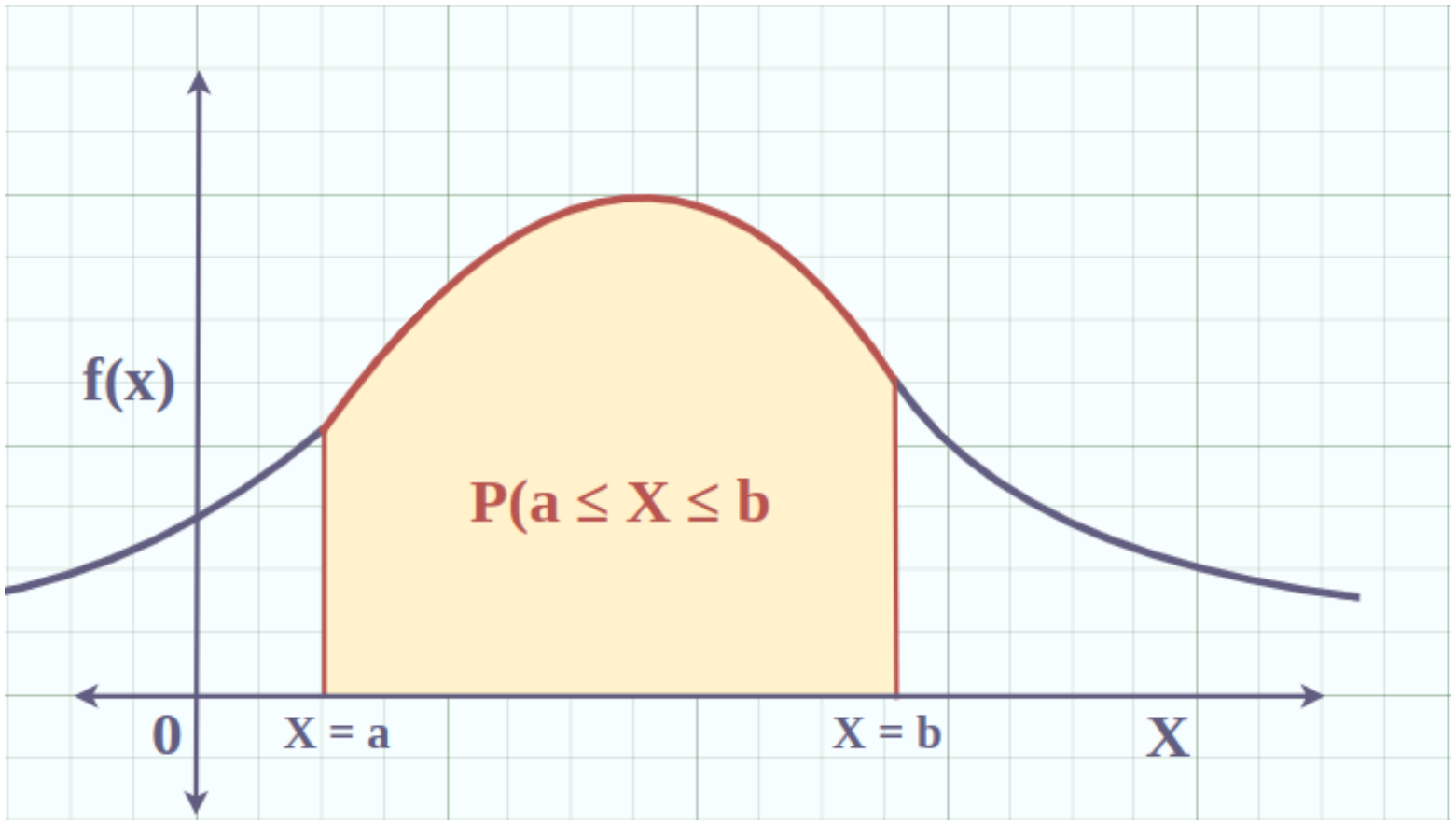



Figure: Distribuciones

Algunas propiedades:

- La gráfica de la función de distribución que gobierna los datos se asemejará al histograma.
- Si conocemos la distribución $f(x)$ podemos calcular probabilidades usando $P(a \leq x \leq b) = \int_a^b f(x)dx$ esto es, el area bajo la curva en el intervalo estudiado.
- El lugar central ($\int xf(x)dx$) de la gráfica se le llama **esperanza o media** y suele coincidir con la media de los datos.


f(x)
P(a ≤ X ≤ b
0
X = a
X = b
X

# Distribución Normal

La **distribución normal** es una de las distribuciones de probabilidad de variable continua que con más frecuencia aparece en estadística y en la teoría de probabilidades.

La gráfica de su función de densidad tiene una forma acampanada y es simétrica respecto de un determinado parámetro estadístico.

La curva viene dada por la función (de densidad)

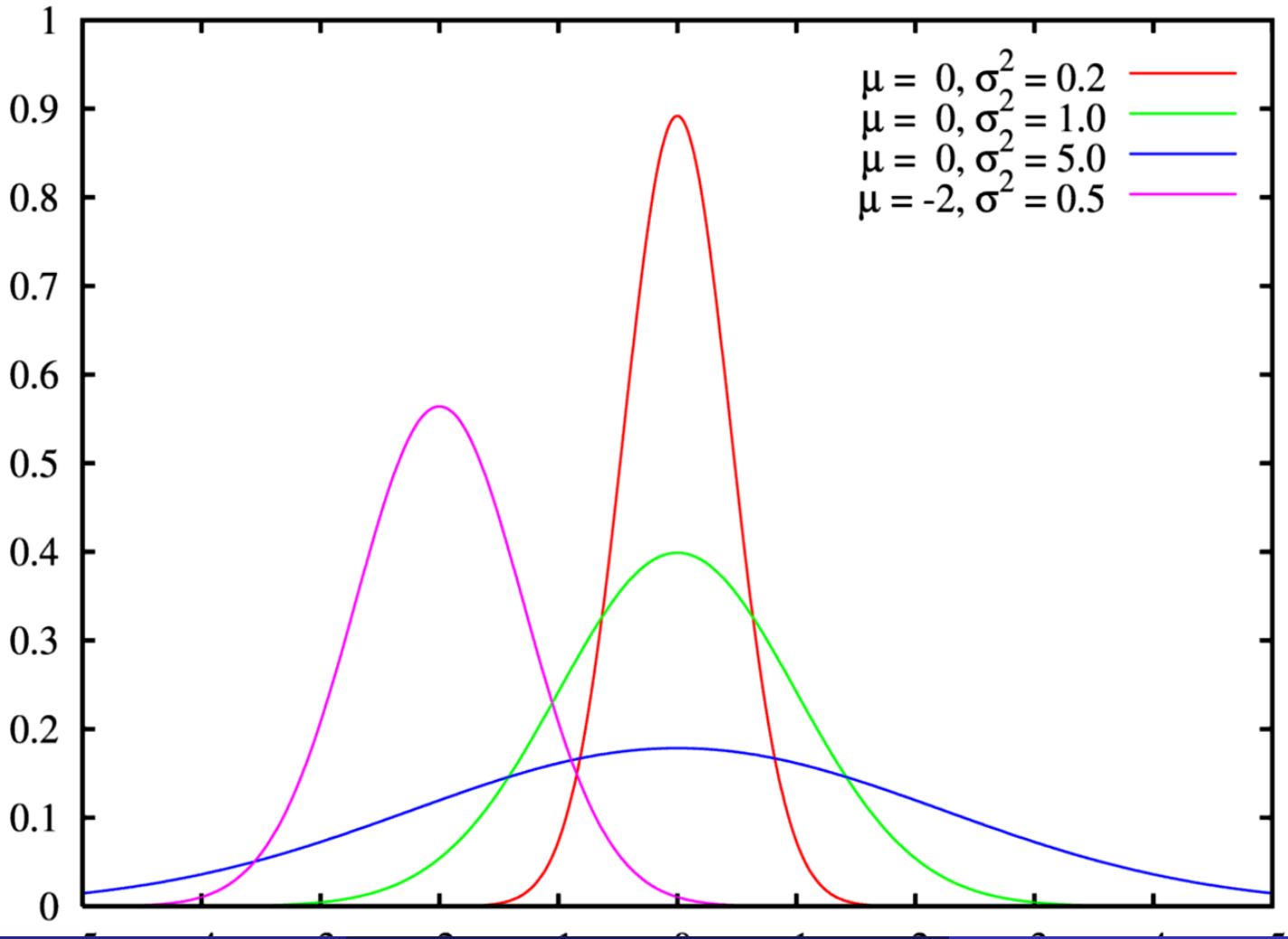
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde

- $\mu$ es la media (también puede ser la mediana, la moda o el valor esperado, según aplique)
- $\sigma$ es la desviación típica

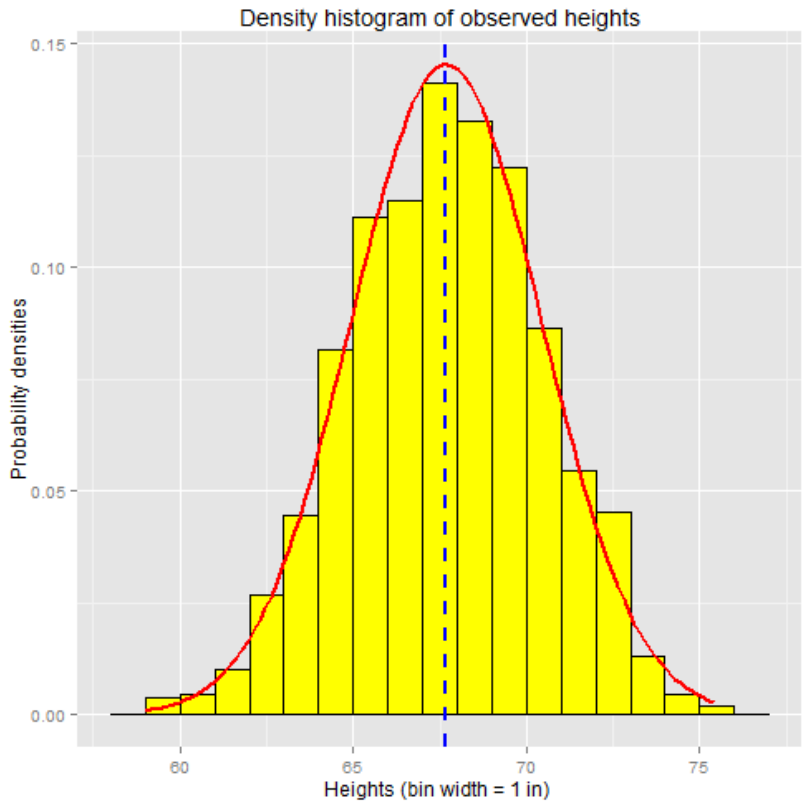
# Gráfica

Su gráfica es la siguiente



# Variables que siguen la distribución normal

Si una variable estudiada sigue la distribución normal su histograma se asemejará a la curva normal para un cierto $\mu$ (que coincidirá con su media muestral) y un cierto $\sigma$ (que coincidirá con la desviación típica).



# Importancia de la Normalidad

Muchos análisis estadísticos (tests, intervalos de confianza, modelos...) se fundamentan en que algún aspecto de los datos siguen una distribución normal. Por lo tanto es **de vital importancia conocer si los datos siguen tal distribución** (pues de lo contrario esas herramientas no funcionarían).

En el caso de no encontrar normalidad deberemos usar otras herramientas denominadas **no paramétricas** que en ocasiones son menos efectivas.

Uno de los errores más comunes en los análisis estadísticos es asumir la normalidad cuando no la hay.

# Importancia de la Normalidad

Muchos análisis estadísticos (tests, intervalos de confianza, modelos...) se fundamentan en que algún aspecto de los datos siguen una distribución normal. Por lo tanto es **de vital importancia conocer si los datos siguen tal distribución** (pues de lo contrario esas herramientas no funcionarían).

En el caso de no encontrar normalidad deberemos usar otras herramientas denominadas **no paramétricas** que en ocasiones son menos efectivas.

Uno de los errores más comunes en los análisis estadísticos es asumir la normalidad cuando no la hay.

# Cuándo se da

Por suerte la teoría estadística nos asegura situaciones en las que la normalidad se da (para muestras grandes la media $\overline{X}$ sigue una normal para ciertos parámetros $\mu$, $\sigma$, esto es el **teorema del límite central**).

Además para averiguar si unos datos siguen una distribución normal tenemos varias herramientas como son los **los histogramas, los qqplots y los test de normalidad**. De momento nos centraremos en los dos primeros.

# Meme

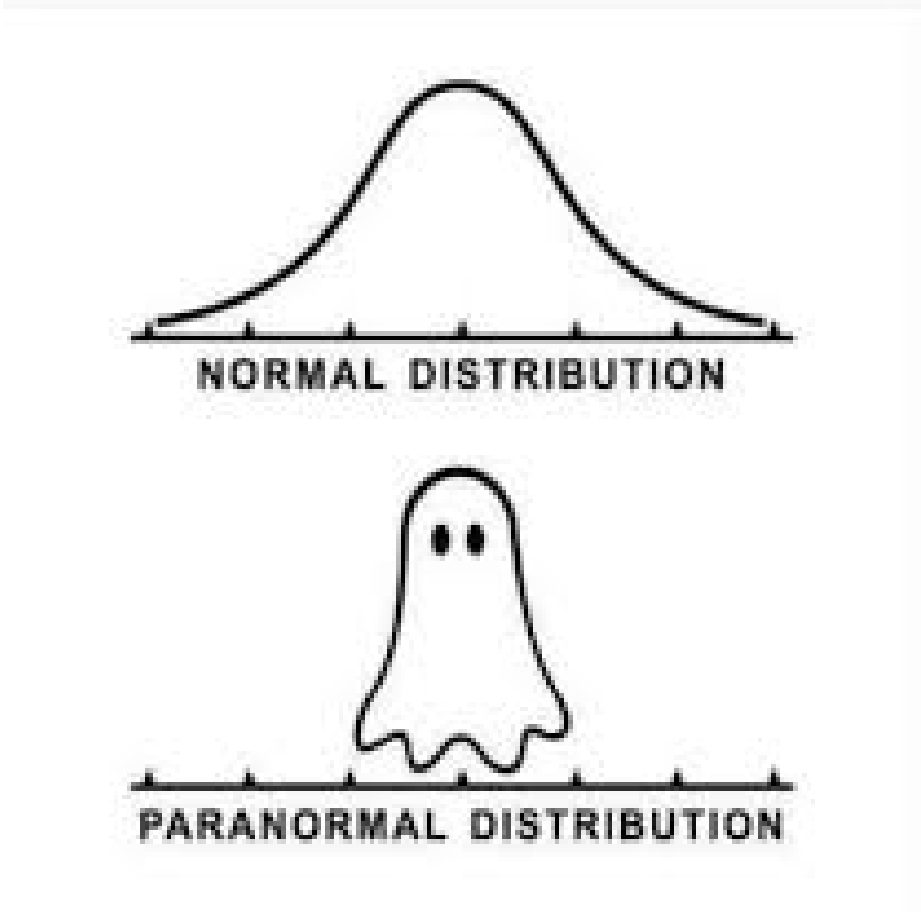



Figure:

# Los histogramas y la normalidad

Los histogramas pueden utilizarse como herramienta para visualizar la normalidad. Si al hacer un histograma

- Los datos están concentrados entorno a un valor central. Este valor central es la media y la mediana a la vez.
- La gráfica es simétrica en torno a la media (más o menos mismo peso a ambos lados de la media)
- Se podría dibujar una campana de gauss en torno al histograma

se trata de indicios de que la variable es de tipo normal

# Ejemplo: Tiempo de ingreso

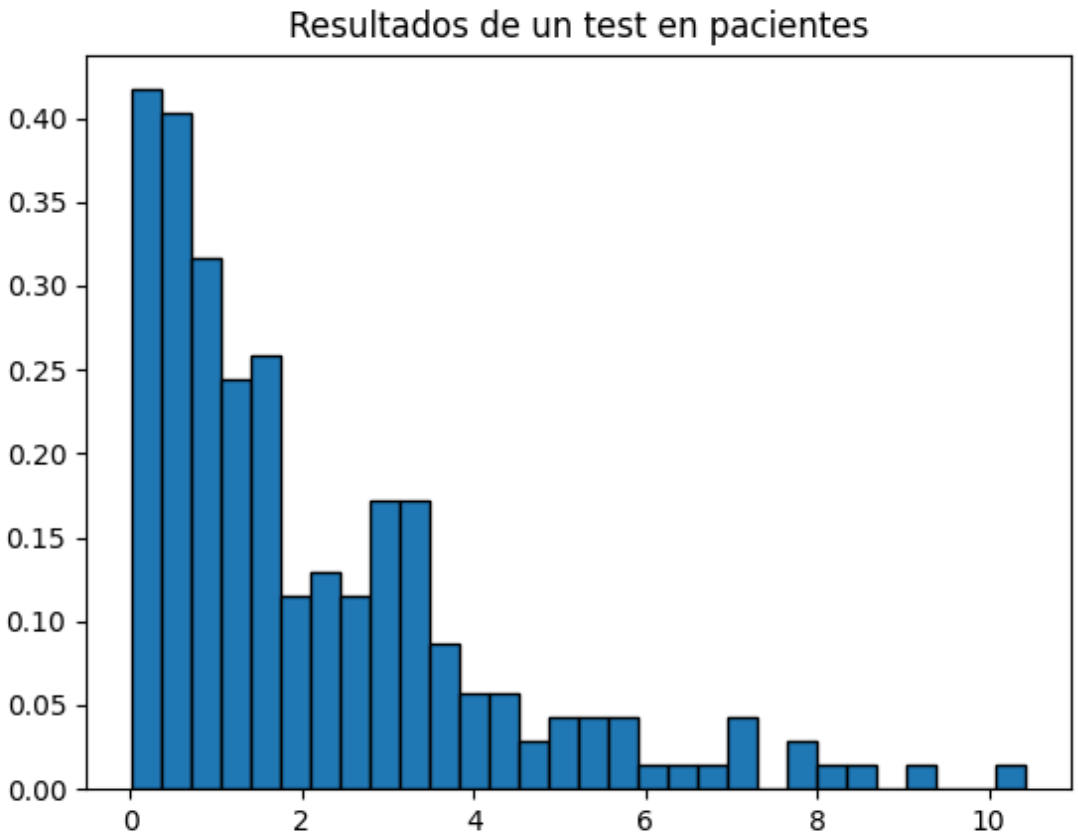



Figure: resultados de un test sobre de 200 pacientes

# Ejemplo: Pesos de pacientes

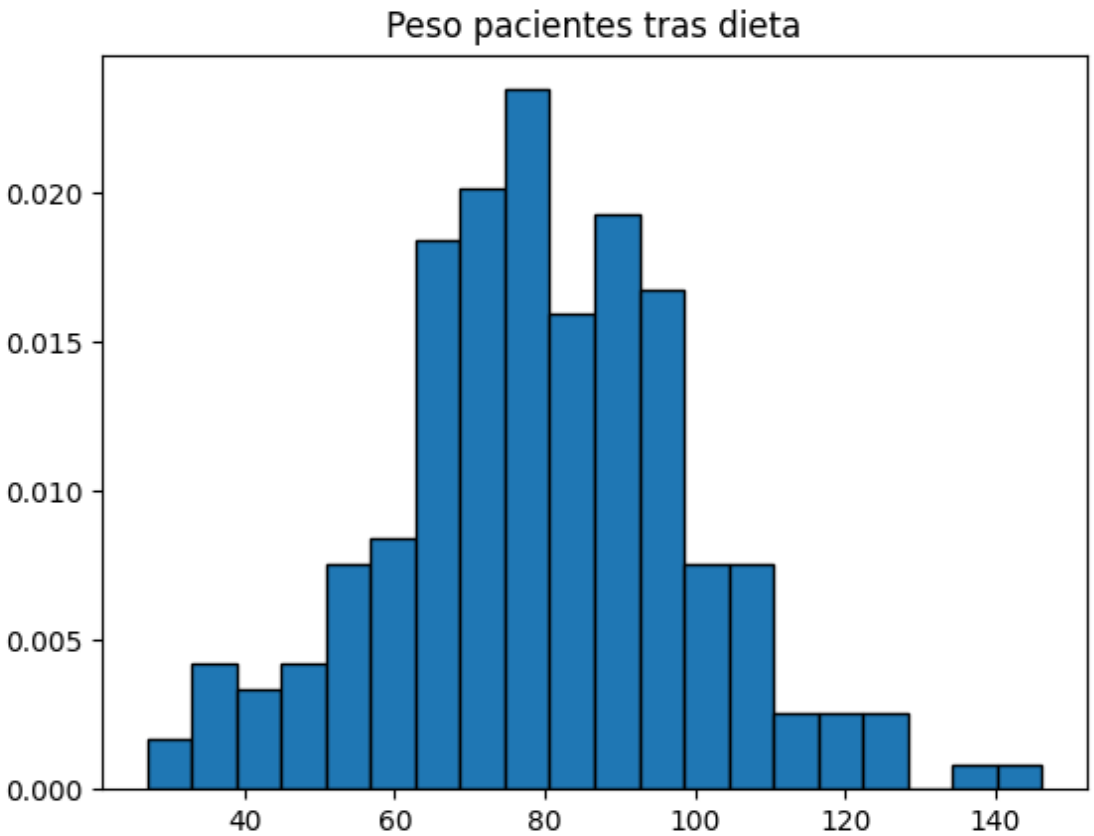



Figure: Pesos de 200 pacientes

# Interpretación

- La primera gráfica no cumple las condiciones para que provenga de una variable normal.
- La segunda sí. Es simétrica, centrada en torno al valor central y con forma de campana.

# Los qqplots y la normalidad

Un gráfico Q-Q es un método gráfico para determinar si una muestra sigue una normal. Consiste en colocar en el eje X los cuantiles de la normal y el eje $Y$ los de la muestra.

**Si los datos siguen una distribución normal se obtendrá, aproximadamente, una línea recta, especialmente cerca de su centro**

# Ejemplo anterior: días de ingreso

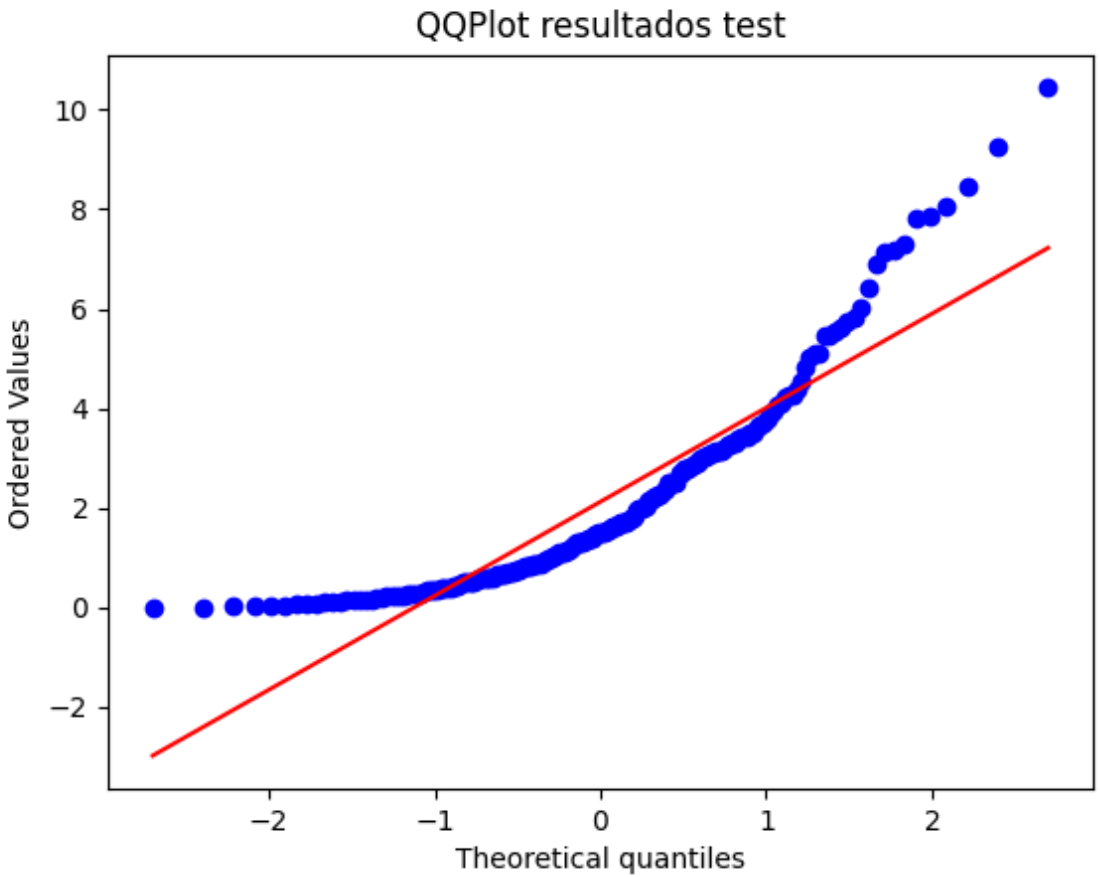



Figure: QQ plot días de ingreso

# Ejemplo anterior: pesos

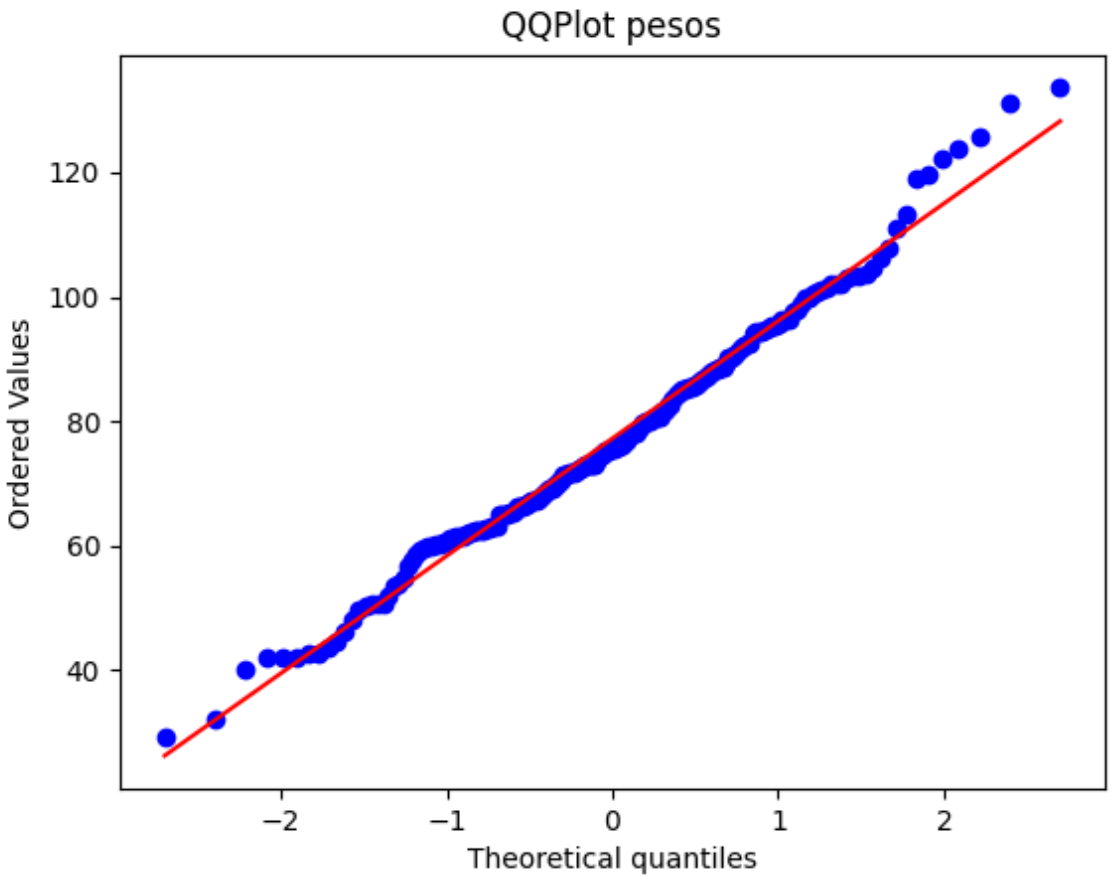
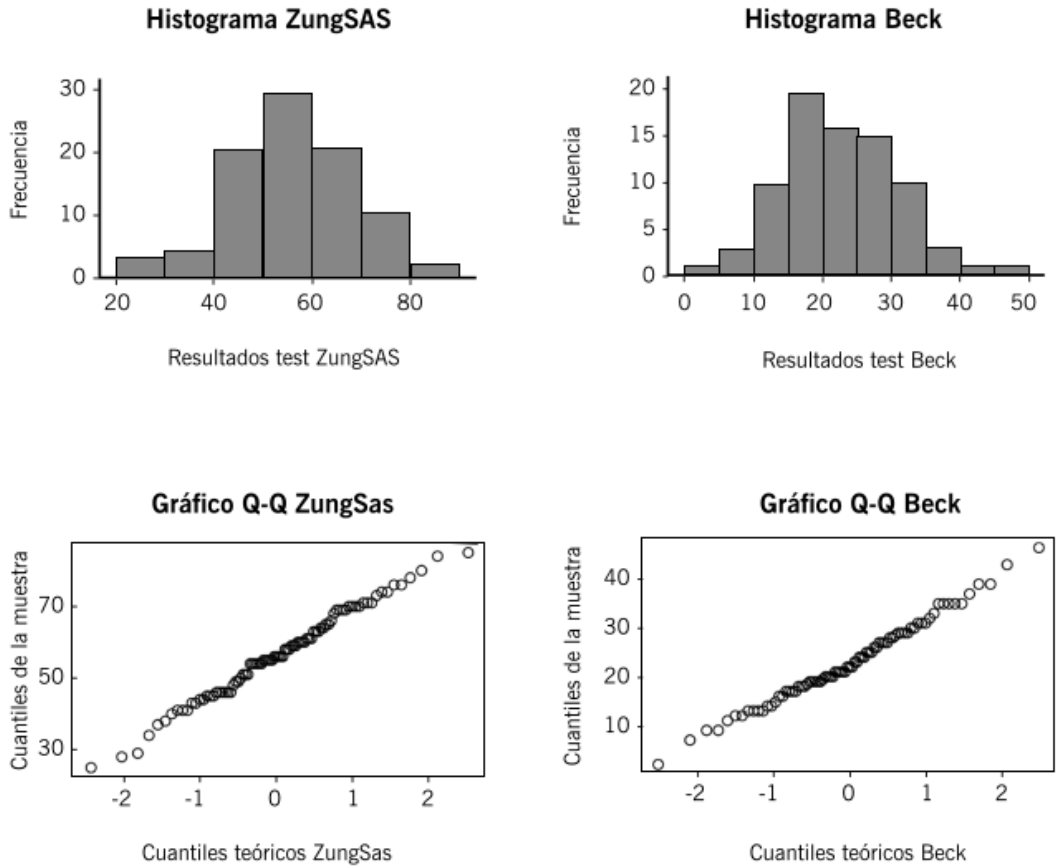



Figure: QQ plot pesos

# Ejemplo real

fuente: *Valoración forense del riesgo psicológico inicial en víctimas de violencia de género*



En el gráfico anterior los autores utilizan los qqplots para concluir que las variables estudiadas **siguen una distribución normal**

Para interpretar los qqplots no hay un criterio inequívoco. La idea es observar cuanto se desvían los puntos de la recta, especialmente en los extremos. En los casos en que no se presente una distribución normal esta desviación será evidente.