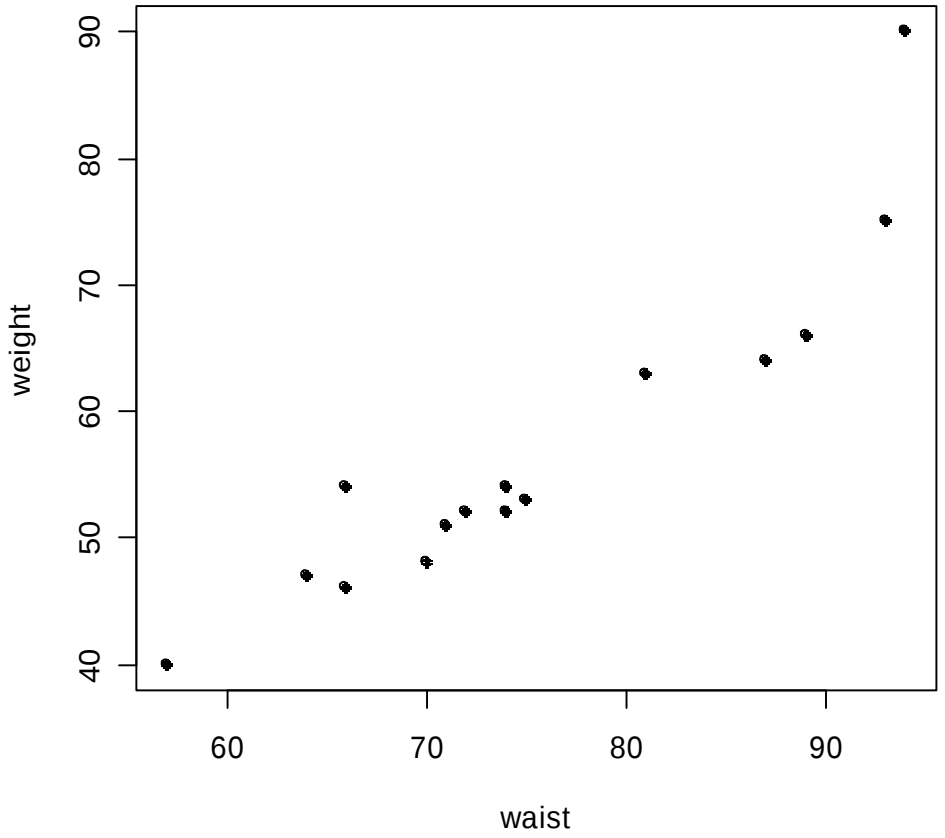
# Phân tích tương quan

# Ví dụ

**Cân nặng và vòng eo**. Số liệu sau đây được trích ra từ một nghiên cứu qui mô (trên 3000 người) ở Việt Nam về mối liên hệ giữa các chỉ số nhân trắc và bệnh tiểu đường. Trọng lượng và vòng eo của 15 đối tượng:

| Trọng lượng | Vòng eo |
|---|---|
| 51.0 | 71.0 |
| 66.0 | 89.0 |
| 47.0 | 64.0 |
| 54.0 | 74.0 |
| 64.0 | 87.0 |
| 75.0 | 93.0 |
| 54.0 | 66.0 |
| 52.0 | 74.0 |
| 53.0 | 75.0 |
| 52.0 | 72.0 |
| 48.0 | 70.0 |
| 46.0 | 66.0 |
| 63.0 | 81.0 |
| 40.0 | 57.0 |
| 90.0 | 94.0 |



# Vài thông số cơ bản

| | Cân nặng | Vòng eo |
|---|---|---|
| Trung bình | 57 kg | 75.5 cm |
| Phương sai (variance) | 163.6 | 122.6 |
| Độ lệch chuẩn | 12.8 | 11.1 |

Chúng ta cần một thông số để "nối kết" hai biến.

Thông số đó là "hiệp biến" (covariance).

Hiệp biến là thông số giao chéo (tích số) giữa hai biến sau khi điều chỉnh cho số trung bình.

$$\mathrm{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 130.8$$

# Hệ số tương quan

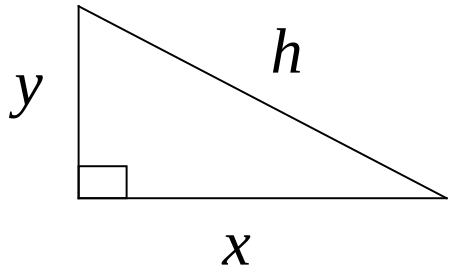
| | Cân nặng | Vòng eo |
|---|---|---|
| Trung bình | 57 kg | 75.5 cm |
| Phương sai (variance) | 163.6 | 122.6 |
| Độ lệch chuẩn | 12.8 | 11.1 |

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 130.8$$

$$r = \frac{Cov(x,y)}{s_x \times s_y}$$

$$r = \frac{130.8}{12.8 \times 11.1} = 0.92$$

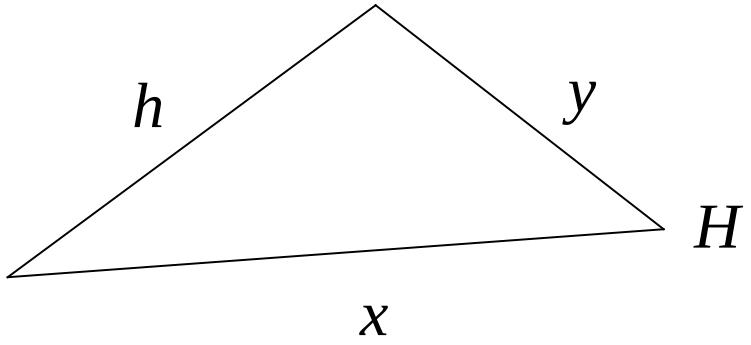
# Mối liên hệ giữa tương quan và hình học




$h^2 = x^2 + y^2$

**Tam giác vuông**

Nếu hai biến x và y đọc lập, hiệp biến = 0

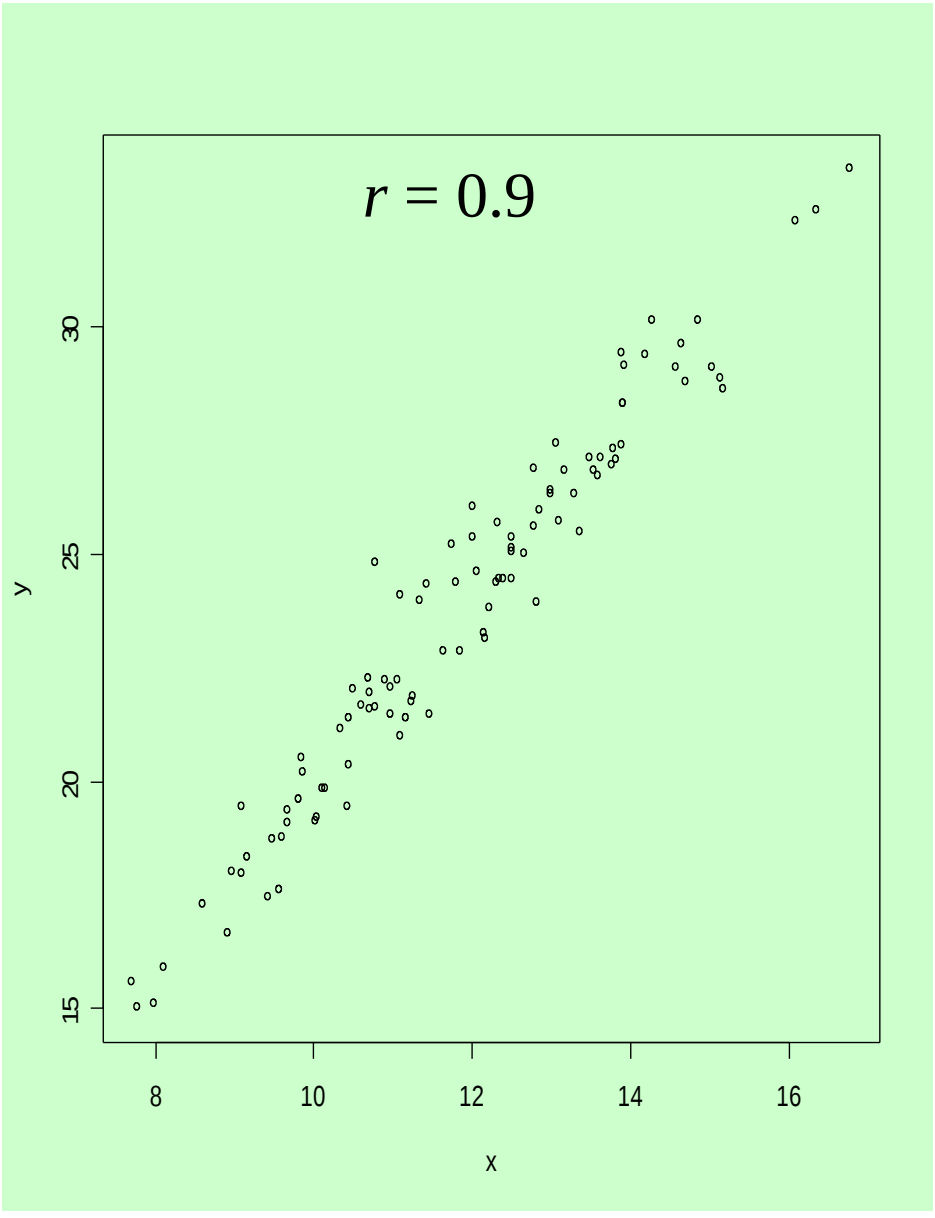


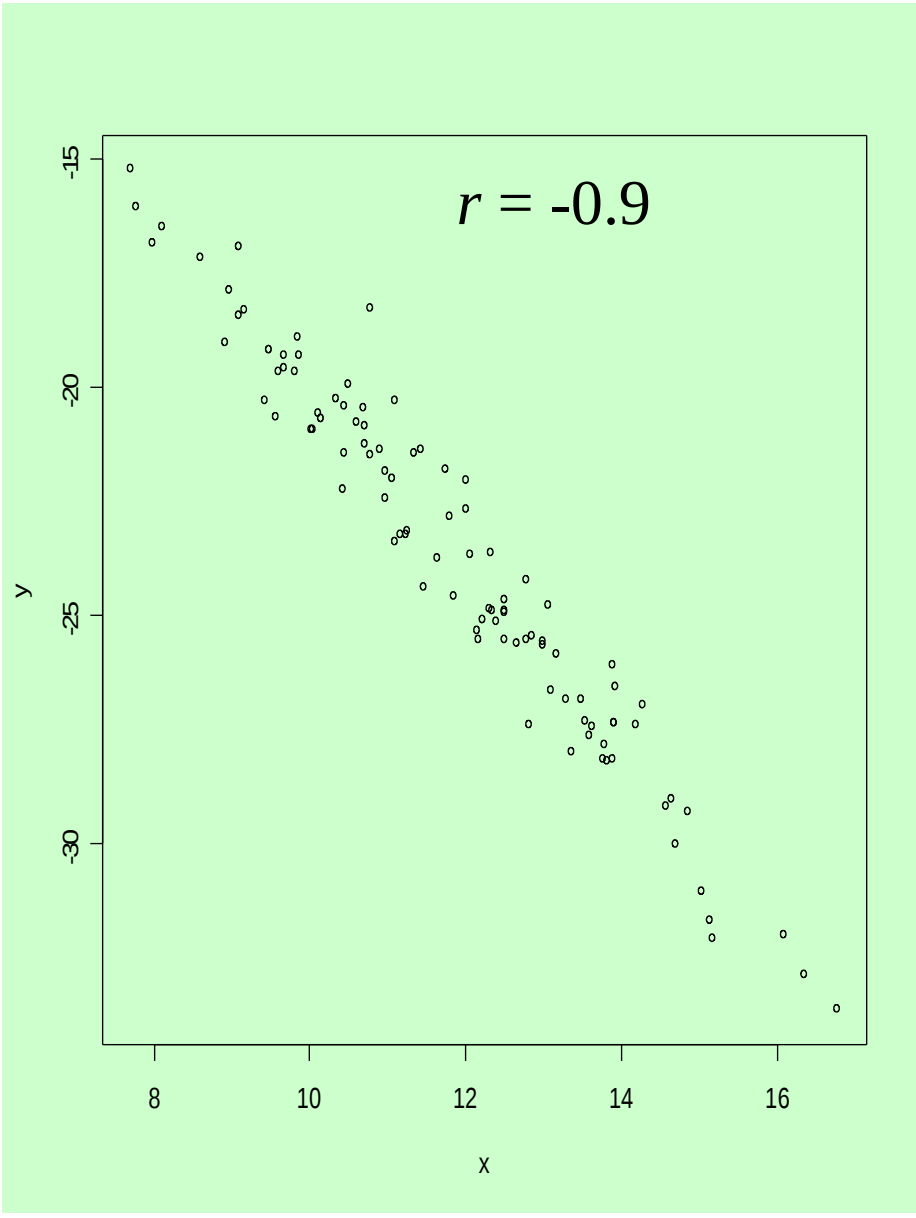

$h^2 = x^2 + y^2 - 2xy\cos(H)$

**Tam giác thường**

Nếu hai biến x và y phụ thuộc, hiệp biến không bằng 0

# Tương quan thuận và nghịch



# Ý nghĩa của hệ số tương quan

| **Hệ số tương quan** | **Ý nghĩa** |
|---|---|
| ±0.01 đến ±0.1 | Mối tương quan quá thấp, không đáng kể |
| ±0.2 đến ±0.3 | Mối tương quan thấp |
| ±0.4 đến ±0.5 | Mối tương quan trung bình |
| ±0.6 đến ±0.7 | Mối tương quan cao |
| ±0.8 trở lên | Mối tương quan rất cao |

# Ước tính khoảng tin cậy 95%

- Khó ước tính trực tiếp, nên phải thông qua phương pháp Fisher.
- Hoán chuyển *r* sang *z:*

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$$

- Tính sai số chuẩn của *z* theo công thức sau (chú ý: *n* là số cỡ mẫu)

$$SE_z = \frac{1}{\sqrt{n-3}}$$

- Tính khoảng tin cậy 95% của z = z $\pm$ 1.96 x $Se_z$
- Hoán chuyển ngược lại cho *r* theo công thức:

$$r = \frac{e^{2z}-1}{e^{2z}+1}$$

# Ví dụ ước tính khoảng tin cậy 95%

- Trong ví dụ 1, chúng ta có $n = 15$ và $r = 0.92$.
- Hoán chuyển $r$ sang $z$:

$$z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = 0.5\log\left(\frac{1+0.92}{1-0.92}\right) = 1.906$$

- Tính sai số chuẩn của $z$ theo công thức sau (chú ý: $n$ là số cỡ mẫu)

$$SE_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{15-3}} = 0.288$$

- Tính khoảng tin cậy 95% của z = 1.906 $\pm$ 1.96 x 0.288

  = 1.34 đến 2.47

- Hoán chuyển ngược lại cho $r$ theo công thức:

$$r = \frac{e^{2z}-1}{e^{2z}+1} = \frac{e^{2\times 1.34}-1}{e^{2\times 1.34}+1} = 0.87 \qquad r = \frac{e^{2z}-1}{e^{2z}+1} = \frac{e^{2\times 2.47}-1}{e^{2\times 2.47}+1} = 0.98$$

# Kiểm định giả thuyết

- Hệ số tương quan *r* là ước số của hệ số tương quan trong quần thể $\rho$.
- Chúng ta không biết giá trị của $\rho$, nhưng biết rằng nó dao động trong khoảng 0.87 và 0.98 với xác suất 95%.
- Giả thuyết đặt ra là $\rho = 0$ (không có mối liên hệ giữa cân nặng và vòng eo).
- Kiểm định giả thuyết là *t*: $t = z / Se_z$

  Trong ví dụ: $z = 1.906$, $SE_z = 0.288$

  Kiểm định: $t = 1.906 / 0.288 = 6.61$

  **Chúng ta có bằng chứng từ chối giả thuyết, và kết luận rằng có mối liên hệ giữa cân nặng và vòng eo.**

# Cẩn thận khi diễn dịch

- “Correlation is not causation” – tương quan không có nghĩa là nguyên nhân – hệ quả.
- r (trọng lượng và vòng eo) = 0.92 không có nghĩa là trọng lượng là nguyên nhân làm cho người ta có vòng eo rộng, hay vòng eo rộng là nguyên nhân làm cho người ta cân nặng.
- *r* = 0.92, hệ số bội $r^2 = (0.92)^2 = 0.846$. Điều này có nghĩa là “*vòng eo ‘giải thích’ khoảng 85% những khác biệt về cân nặng giữa các cá nhân*”

  hay “*khoảng 85% khác biệt về cân nặng giữa các cá nhân có thể giải thích qua vòng eo*.”