Phân tích tương quan

Tuan V. Nguyen

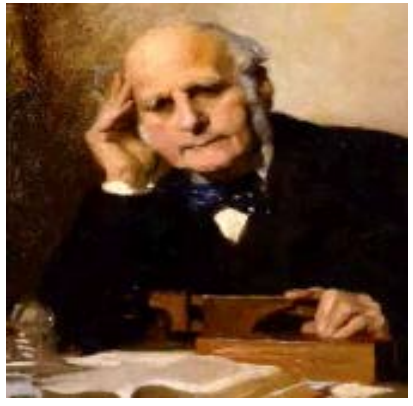
Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam

Sir Francis Galton (16/2/1822 – 17/1/1911)



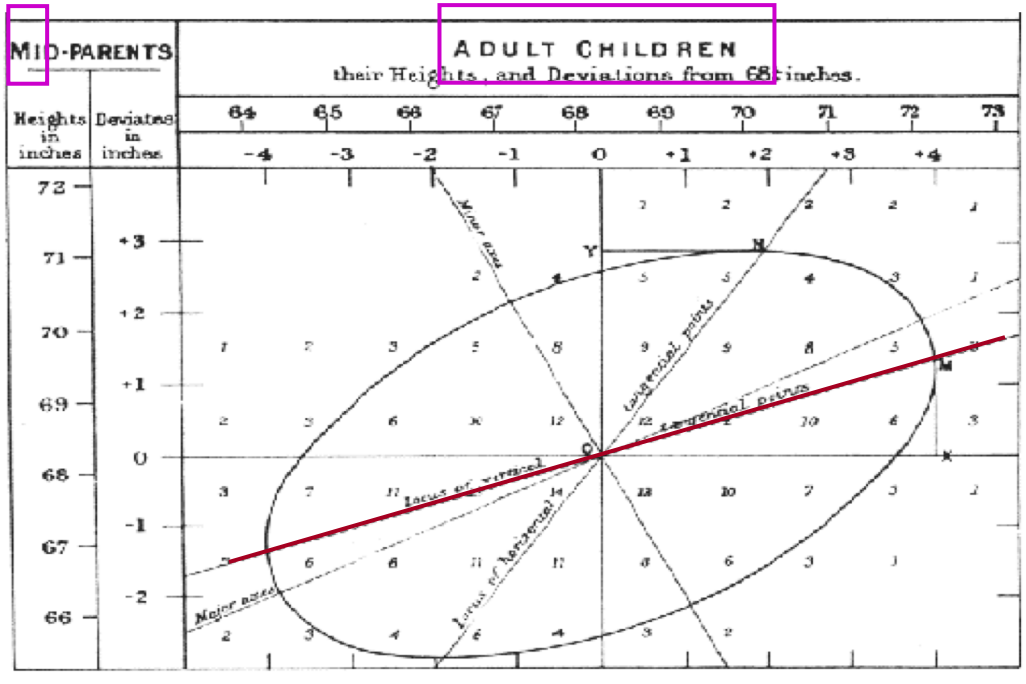
Research interest:

"Those qualifications of intellect and disposition which … lead to reputation"

Galton's conclusions:

- **Nature dominates:** "families of reputation were much more likely than ordinary families to produce offspring of ability"
- **Recommended "judicious marriages during several generations" to "produce a highly gifted race of men"**
- His "genetic utopia": **"Bright, healthy individuals were treated and paid well, and encouraged to have plenty of children. Social undesirables were treated with reasonable kindness so long as they worked hard and stayed celibate."**

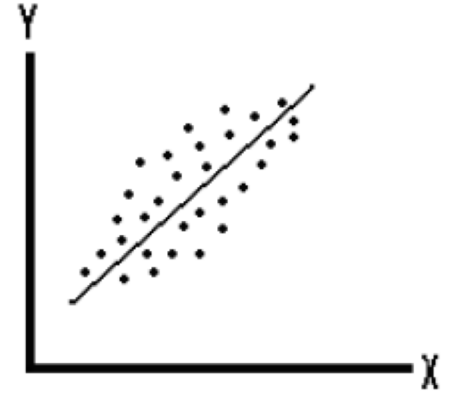
Didn't have data on "intelligence" so instead studied HEIGHT



- Although a self-proclaimed genius, who wrote that he could read @2½, write/do arithmetic @4, and was comfortable with Latin texts @8, **he couldn't figure out how to model these data(!)**
- He went to JD Dickson, a mathematician at Cambridge, who formalized the relationship by developing what we now know as **linear regression**

Khái niệm tương quan (correlation)

- Khi hai biến số (*x* và *y*) có liên quan với nhau
- Mối liên quan có thể cùng chiều hay nghịch đảo
- Ví dụ: mối liên quan giữa tiêu thụ chocolate và giải Nobel (?)



Positive Correlation

Negative Correlation

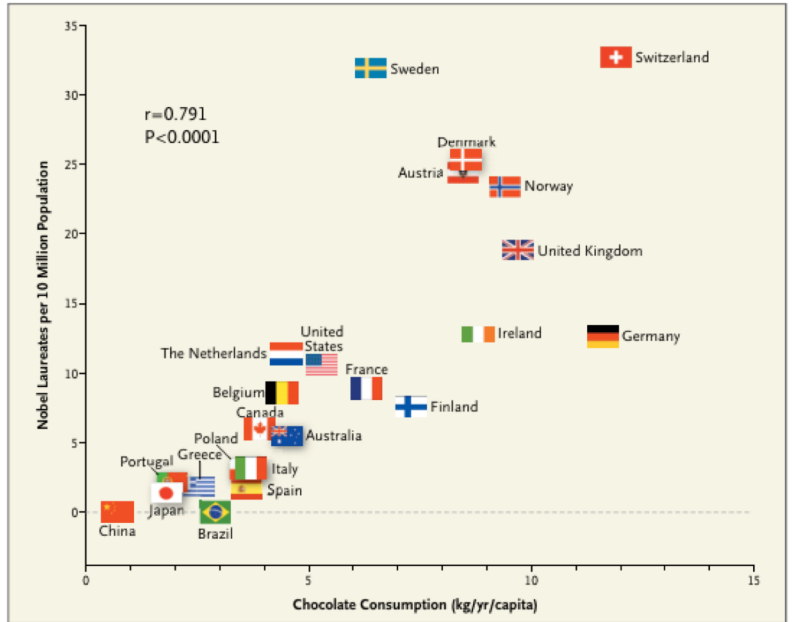


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Định lượng mối tương quan

- Gọi X và Y là 2 biến liên tục
- Phương sai của X

$$s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

- Phương sai của Y

$$s_y^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

- Chúng ta cần một chỉ số đo lường *hiệp phương sai giữa X và Y*
- Covariance

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

Ý nghĩa của phương sai và hiệp biến

- Phương sai lúc nào cũng là số dương
- Hiệp biến có thể dương hay âm (vì là bình phương của tích số X và Y)
  - Nếu covariance = 0, *X và Y độc lập*
  - Nếu covariance > 0, X và Y biến thiên cùng chiều
  - Nếu covariance < 0, X và Y biến thiên nghịch chiều
- **Covariance = thước đo độ liên quan**

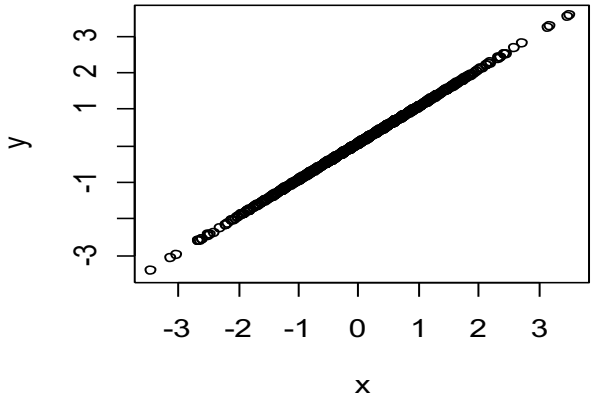
Hệ số tương quan

- Covariance có đơn vị đo lường (X * Y).
- **Coefficient of correlation** (*r*) giữa *X* và *Y* là một **standardized covariance – không có đơn vị đo lường**
- *r* định nghĩa như sau:

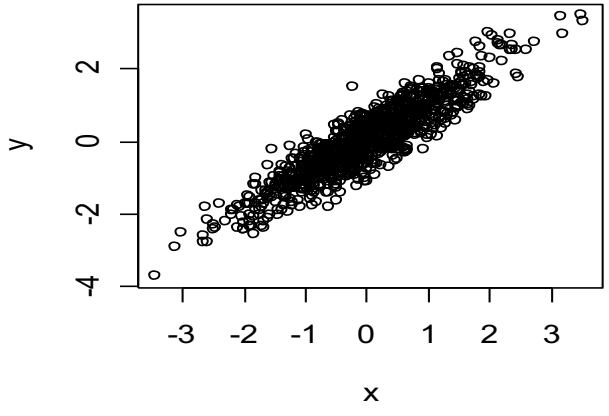
$$r = \frac{cov}{\sqrt{s_x^2 . s_x^2}} = \frac{cov}{s_x . s_y}$$

Vài hệ số tương quan

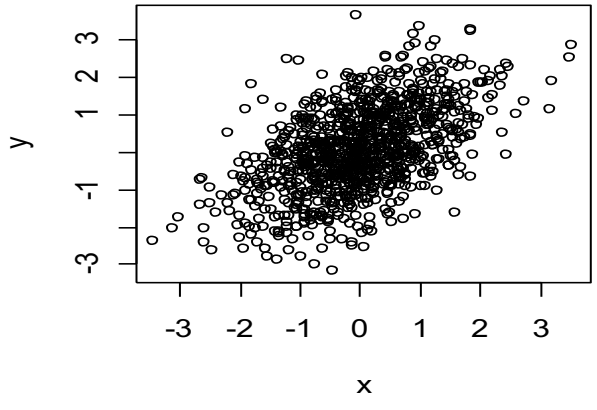
r = 0.99



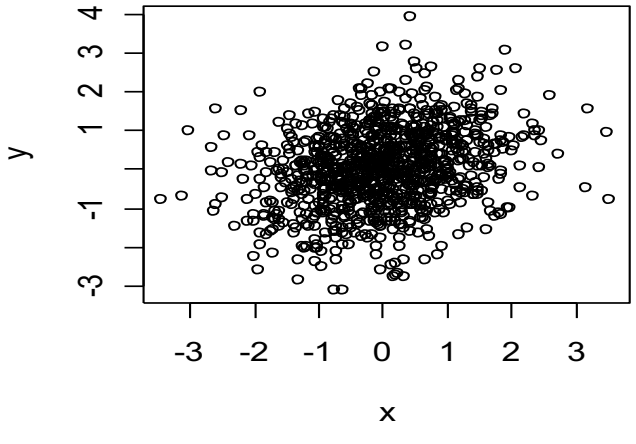
r = 0.90



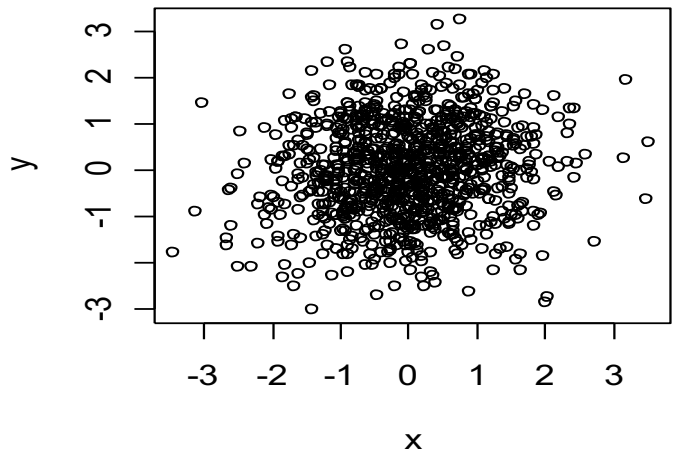
r = 0.50



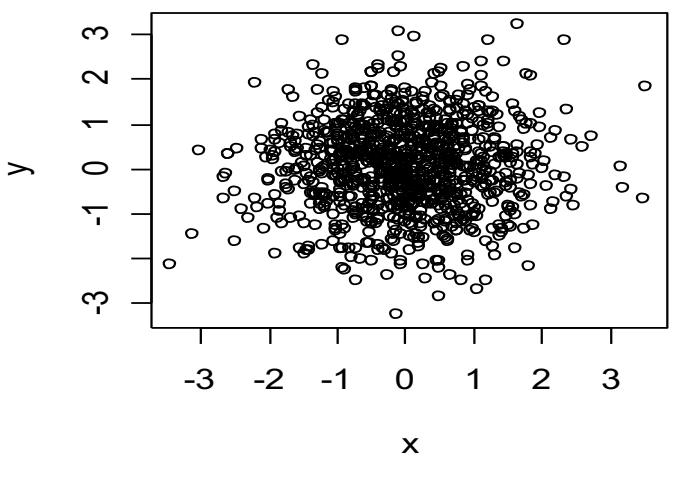
r = 0.25



r = 0.10



r = 0.01



Hàm cor

- Nếu không có missing values

```
cor(x, y)
```

- Nếu có missing values

```
cor(x, y, use="pairwise.complete.obs")
```

```
cor(x, y, use="complete.obs")
```

Hệ số tương quan Pearson với R

```
iq = read.csv("iqsize.csv")

> cor(iq$PIQ, iq$Brain)
[1] 0.3778155

> cor.test(iq$PIQ, iq$Brain)

        Pearson's product-moment correlation

data:  iq$PIQ and iq$Brain
t = 2.4484, df = 36, p-value = 0.01935
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06611792 0.62233266
sample estimates:
      cor
0.3778155
```

r and R^2

- r là hệ số tương quan
- R^2 là hệ số xác định (coefficient of determination)

  phản ảnh phần trăm phương sai của y có thể giải thích bởi biến x
- r(X, Y) = 0.33 có nghĩa là $R^2 = (0.33)^2 = 0.11$.

  11% độ khác biệt về X có thể giải thích bằng những khác biệt về Y

Giả định

- *x* và *y* tuân theo (hay xấp xỉ) luật phân bố chuẩn
- Phương sai của *y* không thay đổi theo giá trị của *x*
- Mối liên quan giữa *x* và *y* phải là tuyến tính
- Không có giá trị ngoại (outliers) trong dữ liệu *x* và *y*

Hệ số tương quan Spearman

- Khi các giả định không đáp ứng ?
- Giải pháp: Hệ số tương quan Spearman
  - Hoán chuyển X và Y sang *rank,* tính d_i = rank(Y_i) – rank(X_i)
  - Tính hệ số tương quan trên rank

$$\rho = \frac{6 \sum_{i=1}^{n} d_i}{n(n^2 - 1)}$$

Hệ số tương quan Spearman với R

```
iq = read.csv("iqsize.csv")

rankIQ = rank(iq$PIQ)
rankBrain = rank(iq$Brain)

> cor(rankIQ, rankBrain)
[1] 0.4126407

> cor(iq$PIQ, iq$Brain, method="spearman")
[1] 0.4126407
```

Hệ số tương quan

- Hệ số tương quan Pearson và Spearman
- Một chỉ số thống kê dùng để đo lường mối tương quan giữa hai biến liên tục
- Hệ số tương quan phản ảnh mối tương quan thống kê (statistical relationship), không hẳn là mối liên hệ nhân quả (causal relationship)

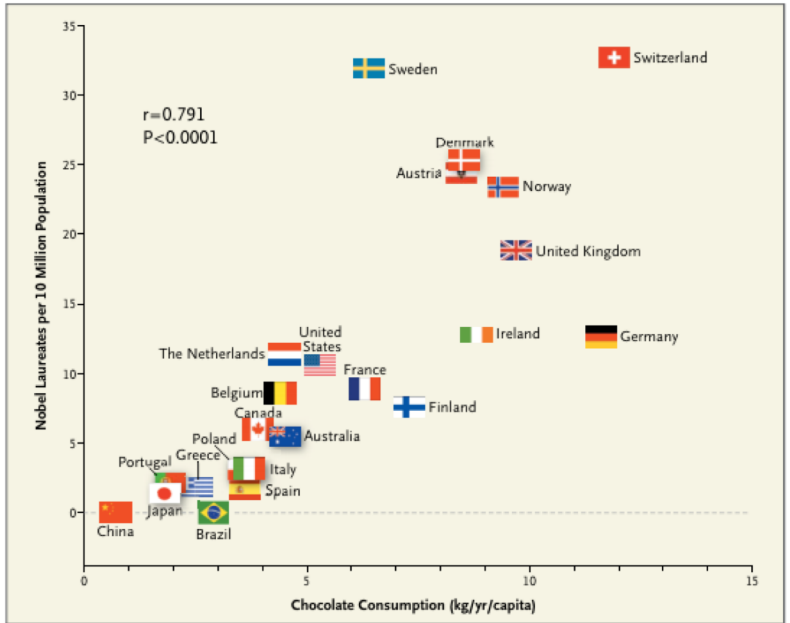


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.