Thiết Kế Thí Nghiệm (Experimetnal Design)

Dữ Liệu Thí Nghiệm và Dữ Liệu Quan Sát (Experimental and Observation Data)

- Dữ liệu từ thí nghiệm (experimental data): là dữ liệu thu thập được từ các thí nghiệm (experiments) khoa học, trong đó các yếu tố ảnh hưởng có thể được kiểm soát để tìm hiểu ảnh hưởng của tác động nhân quả cần nghiên cứu.
  - Ví dụ: thu thập dữ liệu thử nghiệm vaccine Covid-19 trên người.
- Dữ liệu quan sát (observational data): là dữ liệu được thu thập mà nhà nghiên cứu không thể tác động gì lên hiện tượng tạo ra dữ liệu.
  - Ví dụ: thu thập dữ liệu tiền lương, việc làm dân cư.

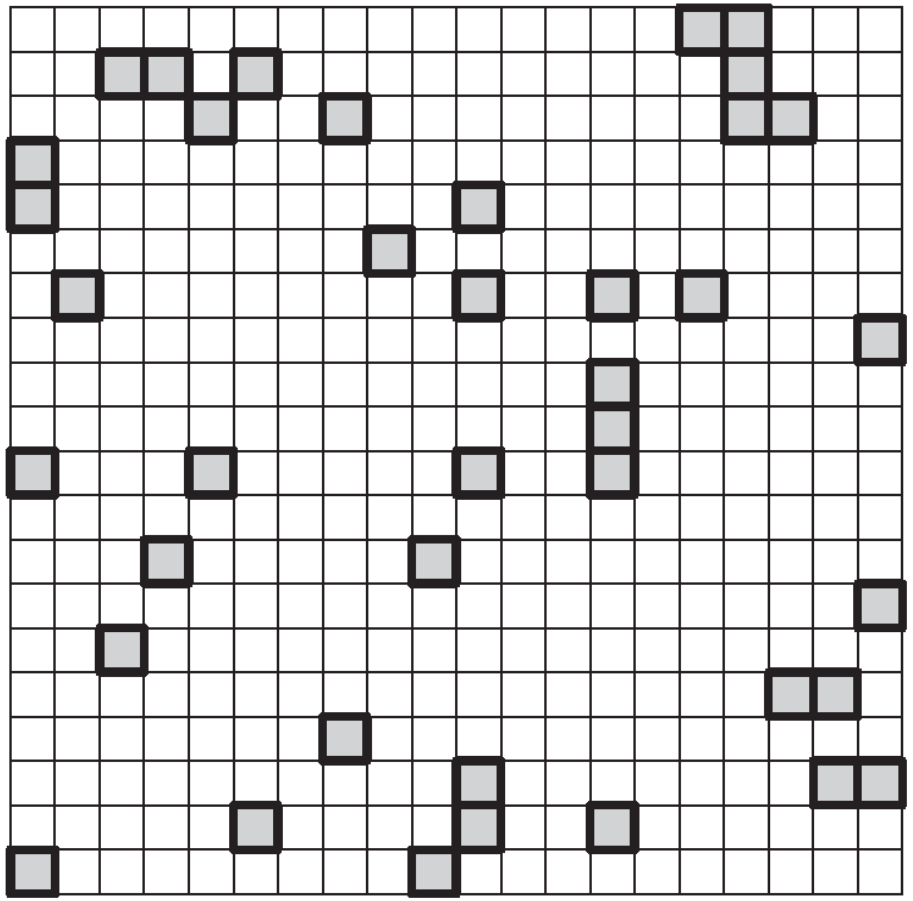
Đánh Giá Tác Động

- Khi muốn đánh giá tác động của một can thiệp hay tác động (manipulation, treatment, or intervention) đến quần thể, chúng ta cần phải thiết kế thí nghiệm trên một mẫu nhỏ hơn.
  - Ví dụ: đánh giá chất lượng phân bón với năng suất cây trồng, đánh giá tác động hỗ trợ vốn với tỷ lệ hộ nghèo.
- Để thực hiện thí nghiệm, nhà nghiên cứu chia mẫu thánh hai nhóm:
  - Nhóm hưởng lợi (treatment group): nhóm mà can thiệp được thực hiện.
  - Nhóm đối chứng (control group): nhóm không can thiệp để so sánh kết quả với nhóm điều tra.

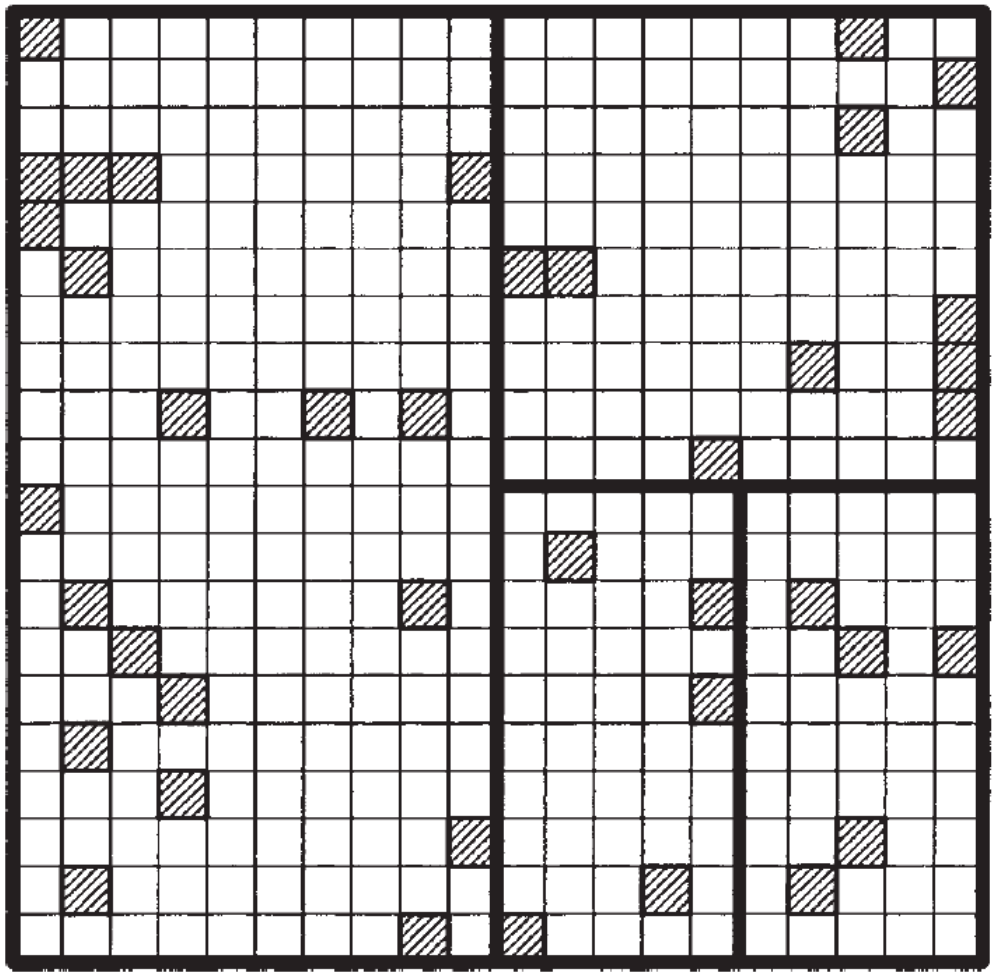
Ngẫu Nhiên Hóa (1)

- Ngẫu nhiên hóa: lựa chọn các quan sát ngẫu nhiên trong hai nhóm can thiệp và đối chứng để đảm bảo các đặc tính liên quan hoàn toàn giống nhau trước khi tác động được thực hiện.
- Đây gọi là randomized control trials (RCT).
- Ví dụ:
  - Chọn các nhiều khoảng đất nhỏ ngẫu nhiên để đảm bảo các yếu tố: ánh sáng, sâu bệnh, nguồn nước… đồng đều giữa hai nhóm thí nghiệm.
  - Chọn ngẫu nhiên nhiều hai nhóm người dân tại cùng khu vực địa lý (huyện, tỉnh) để các yếu tố nghề nghiệp, xã hội… không có ảnh hưởng đến tác động.

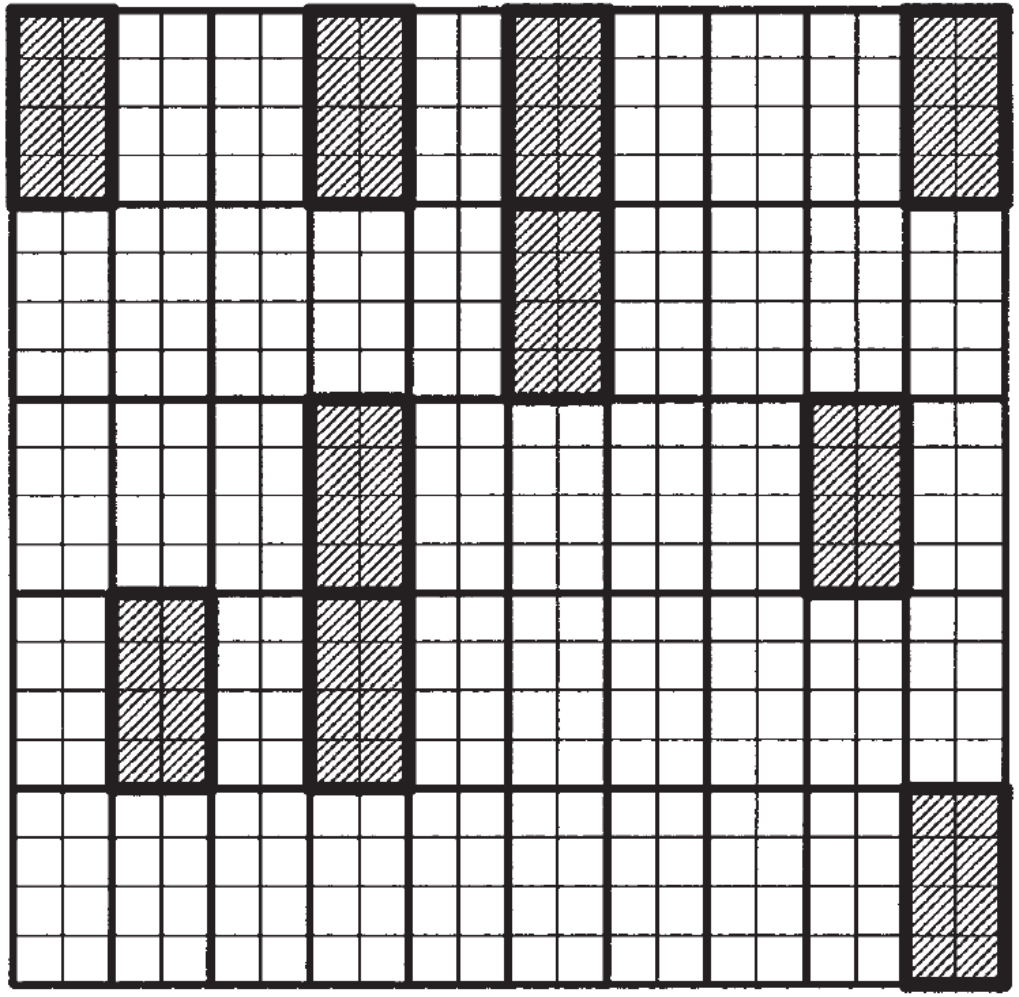
Random sampling



Stratified random sampling



Cluster sampling



Suy Luận Thống Kê Không Chính Xác

- Một nghiên cứu đánh giá tác động giảm nghèo từ một chính sách hỗ trợ vốn vay, kí hiệu *T*, ở một huyện X một tỉnh ở Tây Nguyên. Tác động này được đánh giá kĩ lưỡng theo phương pháp RCT. Kết quả chỉ ra rằng, chính cách *T* có **tác động tích cực** mang ý nghĩa thống kê (ví dụ ở 5%).
- Tuy nhiên sau khi áp dụng diện rộng ở một tỉnh khác ở ĐBSCL thì so sánh kết quả chỉ ra tác động không có hiệu quả. Lý do vì sao?
  - Điều này diễn ra do ngẫu nhiên?
  - Quá trình ngẫu nhiên hóa không thực hiện hoàn toàn? Các hộ tham gia vay vốn là những có đặc tính cá nhân tốt hơn (ý chí vươn lên) so với các hộ không tham gia.
  - Điều kiện kinh tế, xã hội ở ĐBSCL khác với ở Tây Nguyên? Hai bối cảnh khác nhau nên khó so sánh?

Ngẫu Nhiên Hóa (2)

- Tuy nhiên, trong khoa học xã hội thì quá trình ngẫu nhiên hóa cho các thí nghiệm thường không hoàn chỉnh vì các lý do kĩ thuật, đạo đức, và chính trị.
- Ví dụ: khi muốn đánh giá tác động của một chính sách hỗ trợ vốn vay tới người nghèo, những người chọn vay vốn thường là những người có ý chí hoặc hoàn cảnh tốt hơn so với những người không vay vốn (the self-selection problem).
- Hai nhóm can thiệp và đối chứng sẽ khác nhau về các nhân tố có liên quan.
  - Gọi Y_i là thu nhập cho từng cá nhân, sử dụng giá trị trung bình $E(Y_i|\text{vay vốn})$ và $E(Y_i|\text{không vay vốn})$ sẽ tạo ra sai lệch (biasness).

Ngẫu Nhiên Hóa (3)

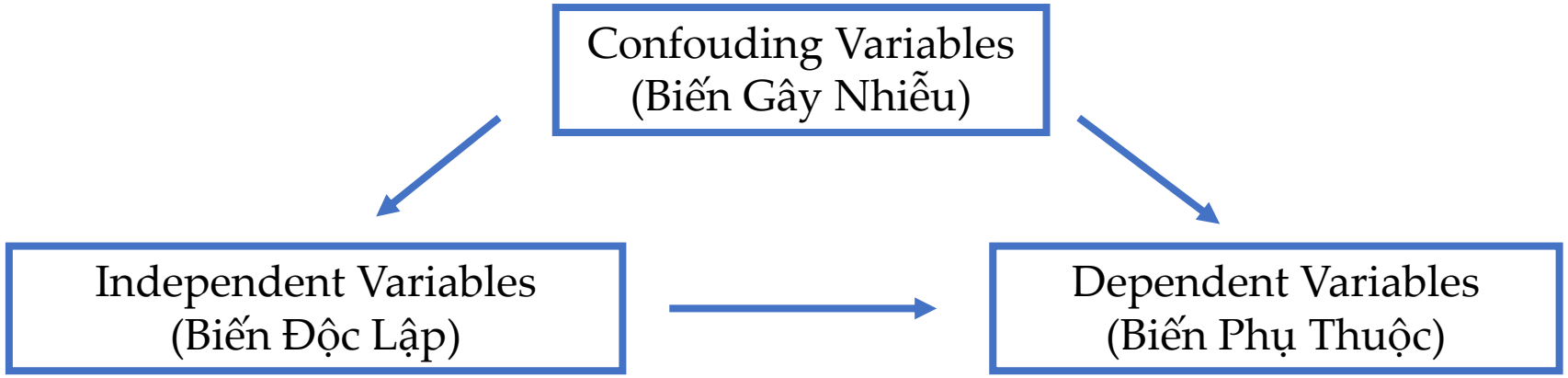
- Gọi:
  - Y_i là thu nhập cho từng cá nhân.
  - T_i là biến có 2 giá trị $T = 0$ nếu vay vốn và $T = 1$ nếu không vay vốn.
  - X_i là biến đo lường ý chí của người tham gia (giả định đo lường được).
- Chúng ta sử dụng mô hình hồi quy sau đây để đánh giá chính xác hơn tác động.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + \varepsilon_i$$

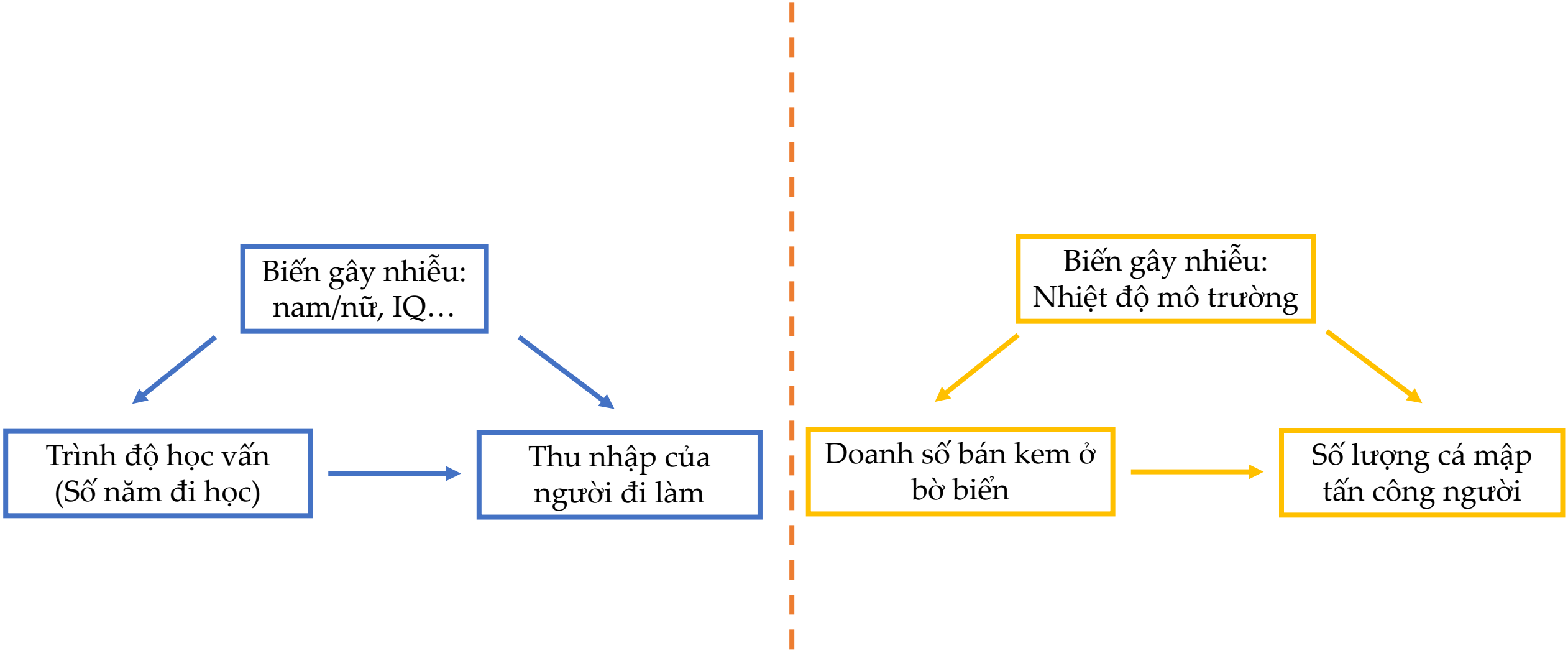
- Nhớ lại mô hình ANOVA 2 chiều: $Y_{i,j} = \mu_j + \gamma_j + (\mu\gamma)_{i,j} + \varepsilon_{i,j,k}$

Minh Họa Bằng Biểu Đồ

- Biến phụ thuộc: trạng thái nghèo/ thoát nghèo (hoặc thu nhập) của hộ (Y).
- Biến độc lập: can thiệp chính sách hỗ trợ vốn vay (T).
- Biến gây nhiễu: mức độ ý chí của hộ dân (X).
  - Ý chí có liên quan với việc tham gia vay vốn.
  - Ý chí là động lực giúp hộ vượt qua nghịch cảnh thoát nghèo.



Ví dụ:


Biến gây nhiễu:
nam/nữ, IQ…
Trình độ học vấn
(Số năm đi học)
Thu nhập của
người đi làm
Biến gây nhiễu:
Nhiệt độ mô trường
Doanh số bán kem ở
bờ biển
Số lượng cá mập
tấn công người

Đối Dữ Liệu Quan Sát

- Dữ liệu quan sát khó đảm bảo tương đồng về các tính chất liên quan giữa hai hay nhiều nhóm cần được so sánh.
- Các mô hình hồi quy sẽ giúp chúng ta phân tách các tính chất đó.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} + \varepsilon_i$$
$$E(\varepsilon_i | X_i = x_i) = 0$$

- Trường hợp khác, do giới hạn của quá trình thu thập dữ liệu hoặc sự khó quan sát, nhiều đặc điểm sẽ không có sẵn trong bộ dữ liệu.
  - Ví dụ: yếu tố di truyền trong việc tính toán chiều cao hoặc thu nhập.
- Các phương pháp về thiết lập mối quan hệ nhân quả sẽ được học ở môn học Định Lượng 2.