# THỐNG KÊ ỨNG DỤNG

Giảng viên: TS. Bùi Thanh Hùng
Bộ môn Khoa học dữ liệu, Khoa Công nghệ thông tin
Đại học Công nghiệp thành phố Hồ Chí Minh
Email: buithanhhung@iuh.edu.vn
Website: https://sites.google.com/site/hungthanhbui1980/

**Mỗi bài tập đều làm theo 2 cách: Tự tính bằng tay và viết code cho máy tính tính và trực quan hóa kết quả đó.**

## Bài 1:

Những sự ảnh hưởng theo độ tuổi không chỉ xảy ra ở điều mà bạn xem thấy trên tivi mà còn ở nơi mà bạn xem tivi. Một nghiên cứu (Darnay, 1994, trang 784) đã cho thấy rằng những người Mỹ lớn tuổi hơn thì ít xem tivi trên giường ngủ hơn so với những người trẻ tuổi và thường xem tivi tại phòng ăn nhiều hơn. Với dữ liệu đã cho trong bảng, giả định rằng cỡ mẫu cho từng nhóm tuổi là 100.

| Khu vực | 25 đến 44 | 45 đến 69 | 60 trở lên |
|---|---|---|---|
| Phòng khách/phòng sinh hoạt chung/phòng làm việc | 95% | 95% | 93% |
| Phòng ngủ | 58% | 57% | 45% |
| Nhà bếp | 12% | 20% | 20% |
| Phòng ăn | 10% | 10% | 10% |

a) Tìm ước lượng khoảng tin cậy 95% về sự khác biệt trong các tỷ lệ của người Mỹ trong độ tuổi từ 45 đến 69 và những người từ 60 tuổi trở lên mà xem tivi tại phòng ăn.
b) Ước lượng sự khác biệt giữa các tỷ lệ của người Mỹ trong nhóm tuổi từ 25 đến 59 và những người trong nhóm độ tuổi từ 60 trở lên mà xem tivi tại phòng khách, phòng sinh hoạt chung hay phòng riêng làm việc và tìm biên sai số ước lượng. [Gợi ý:Tỷ lệ cho nhóm tuổi từ 25 đến 59 sẽ là bình quân giản đơn của các tỷ lệ riêng lẻ dựa trên cỡ mẫu là 2000]

## Bài 2:

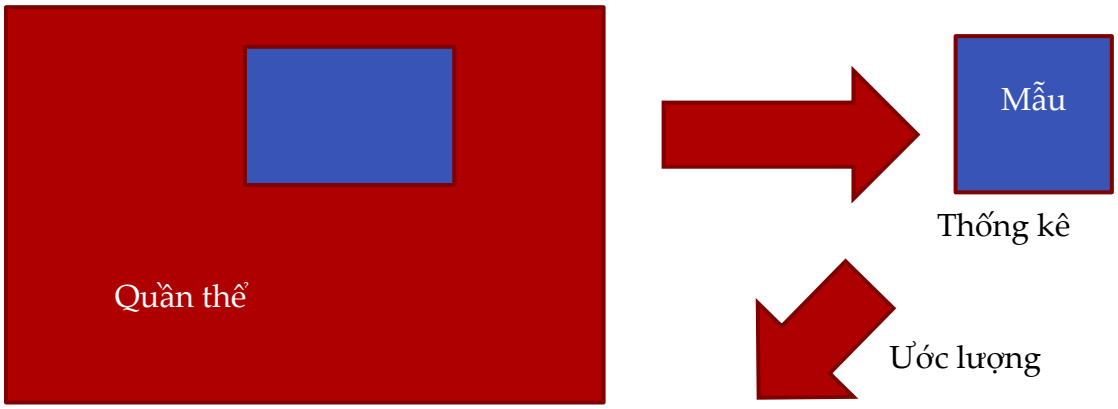
Một cuộc điều tra 100 đại lý mua hàng tạo ra một ước lượng về tỷ lệ những người bán buôn ống nước polyvinyl mà có kế hoạch gia tăng sự mua hàng của mình trong năm tới. Biên sai số, 0.096, là tương đối lớn. Giả định rằng tổ chức tiếp thị tiến hành cuộc điều tra này được yêu cầu phải thực hiện một cuộc điều tra mới và đạt được một ước lượng chính xác trong giới hạn 0.04 với xác suất bằng với 0.90. Xấp xỉ có bao nhiêu nhà bán buôn ắt đã phải tính đến trong cuộc điều tra này?

**Bài 3**:
Một giám đốc nhân sự mong muốn so sánh tính hiệu quả của hai phương pháp huấn luyện các nhân viên công nghiệp nhằm thực hiện một hoạt động lắp ráp nào đó. Một số lượng nhân viên được chia thành hai nhóm bằng nhau, n hóm thứ nhất nhận được phương pháp huấn luyện 1 và nhóm thứ hai được huấn luyện bằng phương pháp 2. Mỗi nhóm sẽ thực hiện hoạt động lắp ráp này, và độ dài của thời gian lắp ráp sẽ được ghi nhận. Người ta kỳ vọng rằng các đại lượng cho cả hai nhóm sẽ có một khoảng xấp xỉ 8 phút. Để cho ước lượng về sự khác biệt về thời gian trung bình để lắp ráp chính xác trong giới hạn 1 phút với xác suất bằng với 0.95, thì cần phải đưa bao nhiêu công nhân vào mỗi nhóm huấn luyện?

**Bài 4**:
Một công ty kiểm toán mong muốn ước lượng sai số trung bình mỗi tài khoản trong các khoản phải thu cho một công ty cung cấp hệ thống ống nước chính xác trong giới hạn $20 với xác suất bằng 0.99. Một mẫu nhỏ trước đó gợi ý rằng sai số mỗi tài khoản sở hữu một độ lệch chuẩn xấp xỉ bằng với $58. Nếu công ty kiểm toán đó mong muốn ước lượng sai số trung bình mỗi tài khoản chính xác trong giới hạn $20, thì có bao nhiêu tài khoản ắt sẽ phải được chọn mẫu? Mẫu này phải sở hữu (các) thuộc tính nào?

# ƯỚC LƯỢNG

Ước lượng tham số: sử dụng thống kê mẫu để ước lượng cho tham số quần thể




Ước lượng điểm (point estimation): xác định một giá trị số là giá trị ước lượng cho tham số quần thể
Ước lượng khoảng (interval estimation): xác định một khoảng giá trị có nhiều khả năng chứa giá trị tham số quần thể

1. Giá trị trung bình của mẫu là ước lượng điểm tốt nhất của trung bình quần thể μ.
2. Sử dụng dữ liệu mẫu để xây dựng một khoảng ước lượng cho giá trị trung bình quần thể.
3. Xác định kích thước mẫu cần thiết để ước lượng trung bình quần thể.

Khoảng ước lượng cho trung bình μ của quần thể là: $\bar{X} \pm E$ hay

$$\bar{X} - E < \mu < \bar{X} + E$$

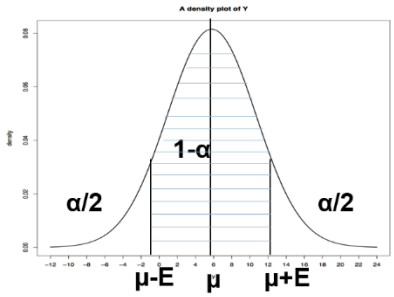
(trong đó $\bar{X}$ là trung bình của tập mẫu có kích thước n, E gọi là biên độ lỗi) .

E được tính dựa vào định lý giới hạn trung tâm.
Định lý giới hạn trung tâm: khi ta lấy mẫu ngẫu nhiên, kích thước tập mẫu càng lớn thì phân phối xác suất của đặc trưng trung bình của tập mẫu càng gần với phân phối chuẩn.

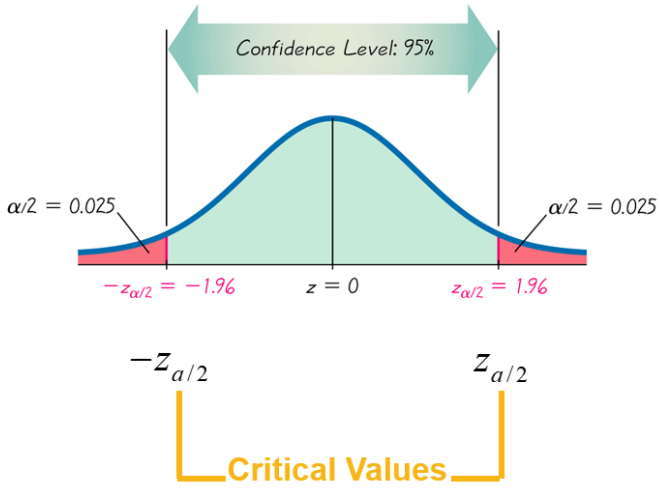
## Khoảng ước lượng cho trung bình μ

- Theo định lý giới hạn trung tâm, nếu n đủ lớn, $\overline{X}$ có phân phối chuẩn với kỳ vọng là μ, phương sai là $\frac{\sigma^2}{n}$
- Trong đó μ là trung bình của quần thể và $\sigma^2$ là phương sai của quần thể

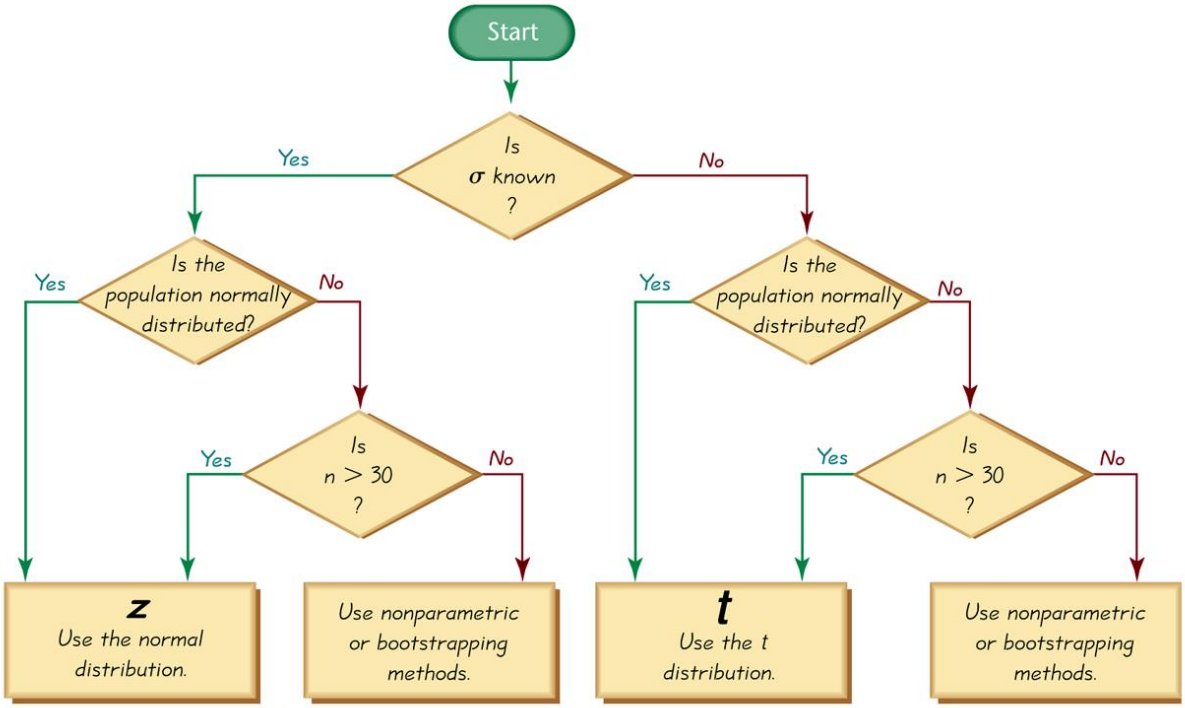
**Phân phối chuẩn của $\overline{X}$**




Độ tin cậy (confidence level) là xác suất 1 - α (thường được biểu thị bằng giá trị phần trăm tương đương) mà khoảng tin cậy thực sự chứa tham số quần thể, giả định rằng quá trình ước lượng được lặp lại một số lượng lớn lần. Độ tin cậy cũng được gọi là bậc tin cậy (degree of confidence), hoặc hệ số tin cậy (confidence coefficient).



## Chọn phân phối phù hợp




| | |
|---|---|
| Sử dụng phân phối chuẩn | $\sigma$ đã biết và quần thể có phân phối chuẩn hoặc $n > 30$ |

$$\bar{x} - z_{\alpha/2} * \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

| | |
|---|---|
| Sử dụng phân phối t | $\sigma$ chưa biết và quần thể có phân phối chuẩn hoặc $n > 30$ |

$$\bar{x} - t_{\alpha/2} * \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

| | |
|---|---|
| Sử dụng phương pháp phi tham số | Quần thể không có phân phối chuẩn và $n \leq 30$ |

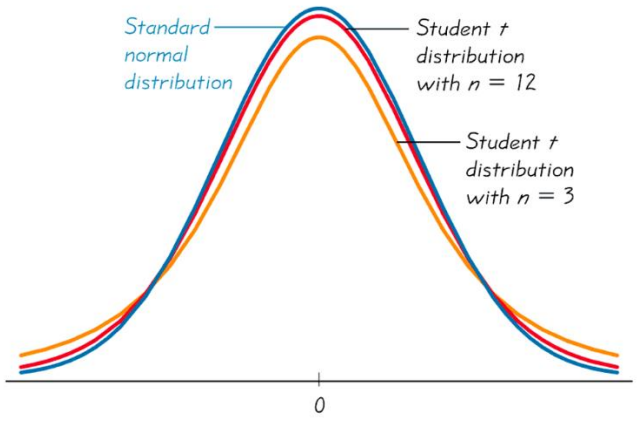
**Phân phối t**
Nếu phân phối của quần thể là phân phối chuẩn thì phân phối của

là một phân phối *t* cho tất cả các mẫu có kích thước. Nó thường được gọi là phân phối t và được sử dụng để tìm các critical values được biểu thị bằng .

**Bậc tự do**
Số bậc tự do của tập hợp dữ liệu mẫu là số các giá trị mẫu có thể khác nhau sau khi có ràng buộc cụ thể trên tất cả các giá trị dữ liệu.
Bậc tự do thường viết tắt là df.
df= $n$ – 1 (Đối với các phương pháp trong phần này)

Các thuộc tính quan trọng của phân phối t

- ✓ Phân phối t khác nhau đối với các cỡ mẫu khác nhau.
- ✓ Phân phối t có hình dạng chuông đối xứng chung giống như phân bố chuẩn nhưng nó phản ánh sự biến thiên lớn hơn (với phân bố rộng hơn) được mong đợi với các mẫu nhỏ.
- ✓ Phân phối t có giá trị trung bình là t = 0 (giống như phân bố chuẩn có giá trị trung bình là 0).
- ✓ Độ lệch chuẩn của phân phối t thay đổi theo cỡ mẫu và lớn hơn 1 (không giống như phân bố chuẩn, có σ= 1).
- ✓ Khi cỡ mẫu n lớn hơn, phân phối t sẽ gần hơn với phân bố chuẩn.



**Tìm kích thước mẫu để ước lượng trung bình quần thể**

$\mu$ = trung bình quần thể

$\sigma$ = độ lệch chuẩn của quần thể

$\bar{x}$ = trung bình mẫu

$E$ = biên độ lỗi mong đợi

$$n = \left[ \frac{(z_{a/2}) \cdot \sigma}{E} \right]^2$$

Nếu cỡ mẫu được tính n không phải là số nguyên, hãy làm tròn giá trị của n đến số nguyên lớn hơn tiếp theo.

**Tìm kích thước mẫu n khi σ không xác định**

1. Sử dụng quy tắc sau để ước tính độ lệch chuẩn:

2. Bắt đầu quá trình thu thập mẫu mà không biết σ, sử dụng một vài giá trị đầu tiên, tính toán độ lệch chuẩn mẫu và sử dụng nó thay cho σ. Giá trị ước tính của σ sau đó có thể được cải thiện khi thu được nhiều dữ liệu mẫu hơn và kích thước mẫu có thể được tinh chỉnh cho phù hợp.
3. Ước lượng giá trị của σ bằng cách sử dụng kết quả của một số nghiên cứu trước đó khác.

Tỉ lệ mẫu là ước lượng điểm tốt nhất của tỉ lệ quần thể.
Sử dụng tỉ lệ mẫu để xây dựng khoảng tin cậy để ước lượng giá trị đúng của tỉ lệ quần thể và chúng ta cũng nên biết cách diễn dịch ý nghĩa về khoảng tin cậy.
Xác định kích thước mẫu cần thiết để ước lượng tỉ lệ quần thể.

Tỉ lệ mẫu $\hat{p}$ là ước lượng điểm tốt nhất của ước lượng tỉ lệ quần thể *p*.

**Sử dụng tỉ lệ mẫu để ước lượng giá trị đúng của tỉ lệ quần thể:**

- Theo chương 6, phân phối tỉ lệ mẫu của biến X là phân phối nhị thức vì thỏa các điều kiện sau:
    - Số lần thí nghiệm của tiến trình ngẫu nhiên đang xét là cố định
    - Hậu quả của thí nghiệm chỉ có thể được phân thành 2 lớp (thành công hay thất bại)

    - Xác suất thành công trong mọi lần thí nghiệm là như nhau
    - Các lần thí nghiệm là độc lập nhau
    - X = số lần thí nghiệm thành công trong n lần thí nghiệm
- Trong pp nhị thức, tính xác suất khi số phép thử lớn (ví dụ như 100) là gần như không thể.
- Phân phối chuẩn có thể được dùng để xấp xỉ phân phối nhị thức khi n lớn.

    - Điều kiện để phân phối chuẩn có thể được dùng để xấp xỉ phân phối nhị thức khi n lớn:
        - np ≥ 5 & n(1 – p)≥ 5
        - *n càng lớn thì xấp xỉ càng tốt.*
    - Khi phân phối nhị thức xấp xỉ phân phối chuẩn, phân phối của tỉ lệ mẫu có:

$$\mu_{\hat{p}} = \hat{p} \qquad \sigma_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$\hat{p} - E \le p \le \hat{p} + E$$

$$\hat{p} - z_{\alpha/2} * \sqrt{\frac{\hat{p}\hat{q}}{n}} \le p \le \hat{p} + z_{\alpha/2} * \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Quy tắc làm tròn cho ước lượng khoảng tin cậy của p
**Làm tròn giới hạn khoảng tin cậy cho p đến ba chữ số có nghĩa**

Xác định kích thước mẫu

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**(solve for *n* by algebra)**

$$n = \frac{(z_{\alpha/2})^2 \hat{p}\hat{q}}{E^2}$$

## Kích thước mẫu để ước tính tỷ lệ p

**Khi ước lượng tỉ lệ $\hat{p}$ đã biết:**

$$n = \frac{(z_{a/2})^2 \hat{p}\hat{q}}{E^2}$$

**Khi ước lượng tỉ lệ $\hat{p}$ chưa biết:**

$$n = \frac{(z_{a/2})^2 0.25}{E^2}$$

**Nếu cỡ mẫu n được tính không phải là số nguyên, hãy làm tròn giá trị của n đến số nguyên lớn hơn tiếp theo.**