Thống kê
(Statistics)

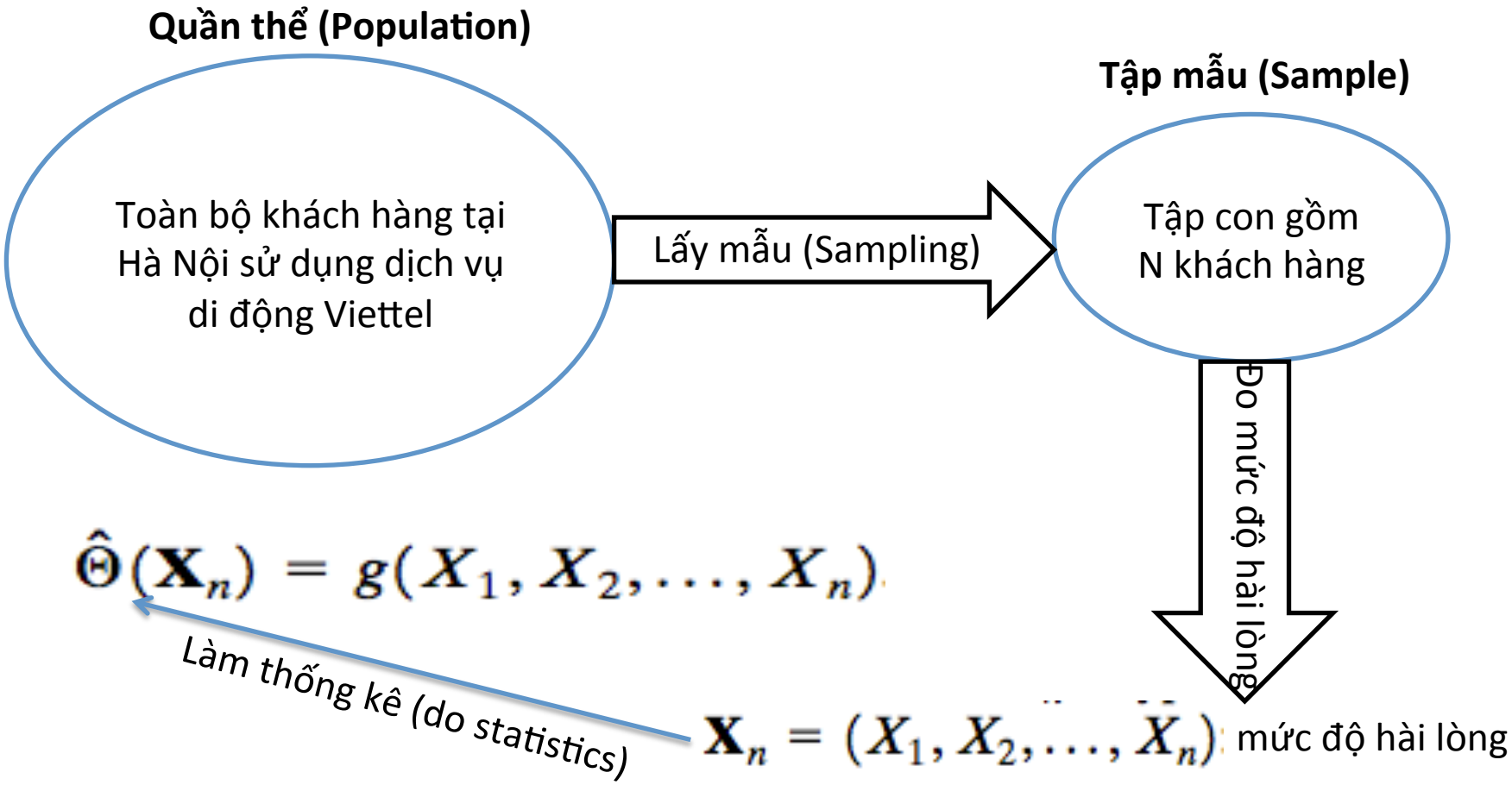
Đặng Thanh Hải (Ph.D)
School of Engineering and Technology, VNUH
Email: hai.dang@vnu.edu.vn

Tại sao cần Thống kê?

- What are the values of parameters, e.g., mean and variance, of a random variable of interest?
- Are the observed data consistent with an assumed distribution?
- Are the observed data consistent with a given parameter value of a random variable?

Thống kê (Statistics)

Đánh giá mức độ hài lòng (X) của khách hàng dịch vụ di động Viettel tại Hà Nội



Mẫu ngẫu nhiên/mẫu bị thiên lệch

- Để tập mẫu phản ánh được tổng thể, tập mẫu cần được lấy ngẫu nhiên từ tổng thể.

- Mẫu bị thiên lệch (biased sample) sẽ làm cho kết quả thống kê thu được từ mẫu không phản ánh được bản chất của tổng thể.

Ví dụ: Để thống kế số lượng bia trung bình 1 người đàn ông VN uống, người ta tiến hành lấy mẫu như sau:

- Chọn ngẫu nhiên 1000 người đàn ông uống bia tại quán bia Lan Chín, Cầu Giấy vào 4 ngày thứ bảy của tháng 6.
- Chọn ngẫu nhiễn 1000 người đàn ông uống bia ở 20 quán bia khác nhau tại Hà Nội vào các ngày bất kì từ tháng 6 đến tháng 10.
- Chọn ngẫu nhiễn 1000 người đàn ông uống bia ở 20 quán bia khác nhau tại 10 tỉnh/thành phố vào các ngày bất kì từ tháng 1 đến tháng 12.

Ước lượng điểm kì vọng và phương sai

Giả sử S={X1, X2,…,Xn} là một mẫu, kì vọng μ và phương sai σ^2 có thể được ướng lượng như sau:

- Ước lượng điểm kì vọng cho quần thể

  $\mu \cong \overline{x} = (X1+X2+\ldots+Xn)/n$

- Ước lượng điểm phương sai cho quần thể

$$\sigma^2 \cong s^2 = \frac{\sum (xi - \overline{x})^2}{n-1}$$

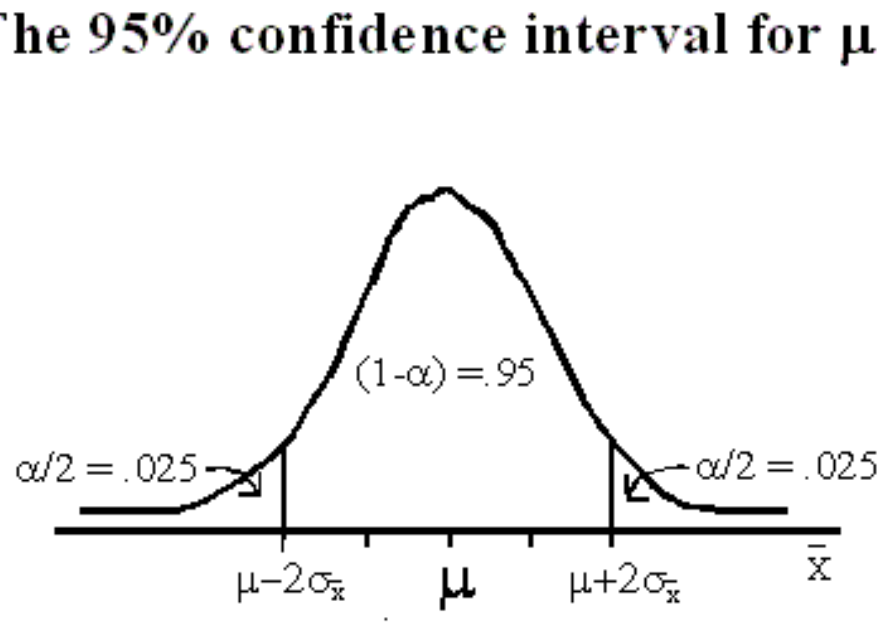
Khoảng tin cậy (Confidence Interval)

Giả sử S={X1, X2,…,Xn} là một mẫu (n>=30), kì vọng μ của quần thể

$$\mu \cong (X1+X2+\ldots+Xn)/n$$

Câu hỏi: Ước lượng khoảng tin cây cho kì vọng μ khi đã biết phương sai σ?

Hay ta muốn tìm 1 đoạn [a,b] để μ thuộc đoạn trên với xác suất β%.



Khoảng tin cậy cho kỳ vọng với phương sai biết trước

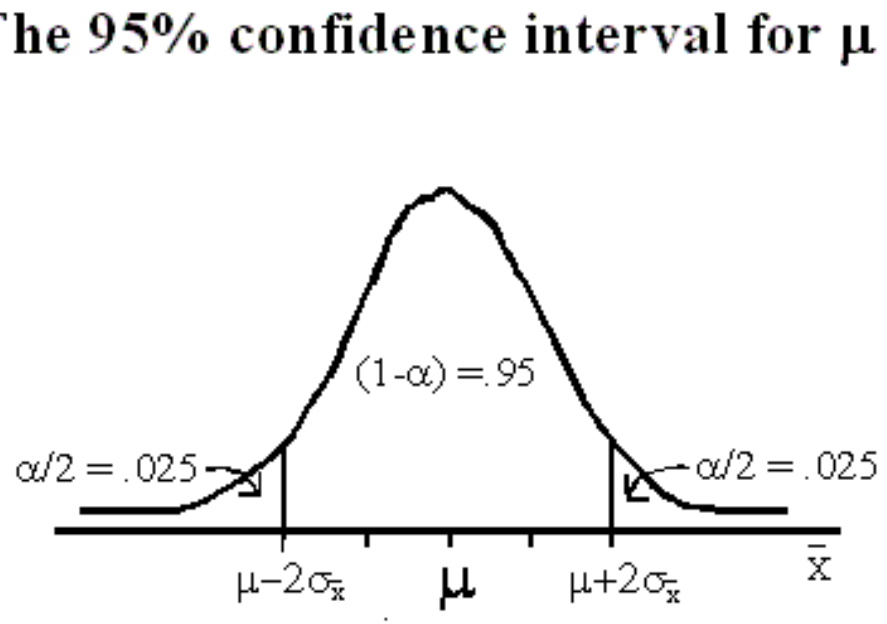
Đoạn [a, b] sẽ có dạng

$$[\overline{x} - u_\beta \sigma_{\overline{x}} , \overline{x} + u_\beta \sigma_{\overline{x}}]$$

Trong đó u_β là số lần độ lệch chuẩn; $\sigma^2_{\overline{x}} = \sigma^2/n$ là phương sai của $\overline{X}$.

Ví dụ

- $\beta = 90\%$, thì $u_\beta = 1.64$
- $\beta = 95\%$, thì $u_\beta = 1.96$
- $\beta = 98\%$, thì $u_\beta = 2.33$
- $\beta = 99\%$, thì $u_\beta = 2.58$



Ví dụ

Chiều cao trung bình của n=36 sinh viên là 66 inches. Giả sử độ lệch chuẩn σ của chiều cao người lớn là 3 inches.

- Tính khoảng tin cậy chiều cao trung bình sinh viên với độ tin cậy 95%
- Tính khoảng tin cậy chiều cao trung bình sinh viên với độ tin cậy 99%

β=95% → u_β = 1.96

σ=3; n=36 → $\sigma^2_{\bar{x}} = \sigma^2/n = 3^2/36$ → $\sigma_{\bar{x}} = 3/6 = 0.5$

$$x \pm u\downarrow\beta\, \sigma\downarrow x = 66 \pm 1.96 \times 0.5$$

Như vậy, với mức độ tin cậy 95%, chiều cao trung bình μ nằm giữa 65.02 và 66.98

β=99% → u_β = 2.58

σ=3; n=36 → $\sigma^2_{\bar{x}} = \sigma^2/n = 3^2/36$ → $\sigma_{\bar{x}} = 3/6 = 0.5$

$$x \pm u\downarrow\beta\, \sigma\downarrow x = 66 \pm 2.58 \times 0.5$$

Như vậy, với mức độ tin cậy 99%, chiều cao trung bình μ nằm giữa 64.71 và 67.29

Nhận xét: Trên cùng một kích thước mẫu, nếu độ tin cậy càng lớn thì độ dài khoảng tin cậy càng lớn.

Xác định kích thước mẫu

- Với một độ tin cậy β% cho trước, khoảng tin cậy [a,b] phụ thuộc vào kích thước mẫu. Kích thước mẫu càng lớn thì khoảng tin cậy càng hẹp và ngược lại.

- **Câu hỏi**: Giả sử muốn ước lượng μ với sai số không quá ε cho trước với độ tin cậy β, thì chúng ta phải tiến hành lấy tối thiểu bao nhiêu mẫu?

- $[\overline{x} - u_\beta \sigma_{\overline{x}} , \overline{x} + u_\beta \sigma_{\overline{x}}]$, trong đó u_β là số lần độ lệch chuẩn; $\sigma^2_{\overline{x}} = \sigma^2/n$ là phương sai của $\overline{X}$.

$$u_\beta \frac{\sigma}{\sqrt{n}} \leq \varepsilon$$

$$\Rightarrow n \geq \left(\frac{\sigma u_\beta}{\varepsilon} \right)^2$$

Ví dụ

Giả sử độ lệch chuẩn của chiều cao người lớn là 5cm.

- Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 2cm với độ tin cậy 90%.

- Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 5cm với độ tin cậy 90%.

- Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 2cm với độ tin cậy 95%.

- Tính số sinh viên phải lấy mẫu để tính chiều cao trung bình sinh viên ĐHCN với sai số không quá 5cm với độ tin cậy 95%.

Khoảng tin cậy cho kỳ vọng với phương sai biết trước

Sử dụng phương sai mẫu (ước lượng điểm cho phương sai): $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2$.

Tính thống kê $T = \frac{\overline{X}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$. T tuân theo **phân bố t của Student với n-1 bậc tự do**

Khoảng tin cậy $(1-\alpha)\times 100\%$: $\left[\overline{X}_n - t_{\alpha/2,\,n-1}\hat{\sigma}_n/\sqrt{n},\ \overline{X}_n + t_{\alpha/2,\,n-1}\hat{\sigma}_n/\sqrt{n}\right]$

Người ta tiến hành nghiên cứu ở một trường đại học xem trong một tháng trung bình một sinh viên tiêu hết bao nhiêu tiền gọi điện thoại. Lấy một mẫu ngẫu nhiên gồm 59 sinh viên thu được kết quả sau:

14	18	22	30	36	28	42	79	36	52	15	47
95	16	27	111	37	63	127	23	31	70	27	11
30	147	72	37	25	7	33	29	35	41	48	15
29	73	26	15	26	31	57	40	18	85	28	32
22	36	60	41	35	26	20	58	33	23	35	

Hãy ước lượng khoảng tin cậy 95% cho số tiền gọi điện thoại trung bình hàng tháng của một sinh viên.

Khoảng tin cậy cho tỉ lệ (1)

Nghiên cứu một quần thể mà mỗi cá thể có thể có hoặc không có một thuộc tính A nào đó.

- P là tỉ lệ cá thể có thuộc tính A trong quần thể
- f = k/n là tỉ lệ (tần suất) cá thể có thuộc tính A trong mẫu nghiên cứu

Câu hỏi: Ước lượng khoảng tin cậy cho tỉ lệ p dựa vào tần suất f.

Định lí: Tần suất f là một ĐLNN có phân bố xấp xỉ phân bố chuẩn với kì vọng Ef = p và phương sai Df = p(1-p)/n với điều kiện np>5 và n(1-p)>5.

Do không biết p, cho nên Df có thể được xấp xỉ bằng

Df = f(1-f)/n

với điều kiện nf>10 và n(1-f)>10.

Khoảng tin cậy cho tỉ lệ p với độ tin cậy β

$$f-u_{\downarrow\beta}\sqrt{f(1-f)/n} \leq p \leq f+u_{\downarrow\beta}\sqrt{f(1-f)/n}$$

Ví dụ

Trước ngày bầu cử tổng thống, ta lấy ngẫu nhiên 100 người để hỏi ý kiến thì có 60 người ủng hộ Hilary Clinton. Tìm khoảng tin cậy tỉ lệ cử tri bỏ phiếu chi Hilary Clinton

- Với độ tin cậy 90%
- Với độ tin cậy 95%
- Với độ tin cậy 99%

n=100; k=60; f=k/n=0.6; β=90% → u_β = 1.64
Kiểm tra điều kiện: nf=100 x (0.6) = 60 > 10; n(1-f)=100 x (0.4)=40 > 10
Như vậy f có phân bố xấp xỉ chuẩn với Ef=p và Df=f(1-f)/n

$$f-u{\downarrow}\beta\sqrt{f(1-f)/n} = 0.6-1.64\times\sqrt{0.6\times0.4/100} \le p \le f+u{\downarrow}\beta\sqrt{f(1-f)/n} = 0.6+1.64\times\sqrt{0.6\times0.4/100}$$

$$0.52\le p\le 0.68$$

Ví dụ

Khảo sát có 150 người nghiện thuốc,

trung bình một người hút 97 điếu trong 1 tuần, độ lệch chuẩn 36

Tìm khoảng tin cậy 99% cho số điếu thuốc hút trong 1 tuần của người nghiện.

Tìm khoảng tin cậy 90% cho tỉ lệ p dựa trên các mẫu sau

1) n=100; k=25;
2) n=150; k=50

Ví dụ

1) Khảo sát 2074 gia đình có 373 gia đình có máy tính ở nhà. Tìm khoảng tin cậy 95% cho tỉ lệ những gia đình có máy tính ở nhà

2) Người ta muốn tìm khoảng tin cậy 95% cho tỉ lệ những gia đình có máy giặt với độ chính xác 0.04. Mẫu điều tra sơ bộ cho thấy f=0.72. Tính kích thước mẫu

3)

Người ta đo chiều sâu của biển, sai lệch ngẫu nhiên được giả thiết phân phối theo qui luật chuẩn với độ lệch tiêu chuẩn là $20m$. Cần đo bao nhiêu lần để xác định chiều sâu của biển với sai lệch không quá $15m$ và độ tin cậy đạt được 95%?

Khoảng tin cậy cho phương sai

Sử dụng phương sai mẫu (ước lượng điểm cho phương sai): $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$

Tính thống kê $\chi^2 = \frac{(n-1)\hat{\sigma}_n^2}{\sigma_X^2} = \frac{1}{\sigma_X^2}\sum \sum_{j=1}^{n}(X_j - \overline{X}_n)^2$, tuân theo **phân bố Chi-bình phương với n-1 bậc tự do**

Khoảng tin cậy $(1-\alpha)\times 100\%$: $\left[\frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)\hat{\sigma}_n^2}{\chi^2_{1-\alpha/2,n-1}}\right].$

TABLE 8.3 Critical values for chi-square distribution, $P[\chi^2 > \chi^2_{\alpha,n-1}] = \alpha$.

$n\backslash\alpha$	0.995	0.975	0.95	0.05	0.025	0.01	0.005
1	3.9271E-05	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
11	2.6032	3.8157	4.5748	19.6752	21.9200	24.7250	26.7569
12	3.0738	4.4038	5.2260	21.0261	23.3367	26.2170	28.2997