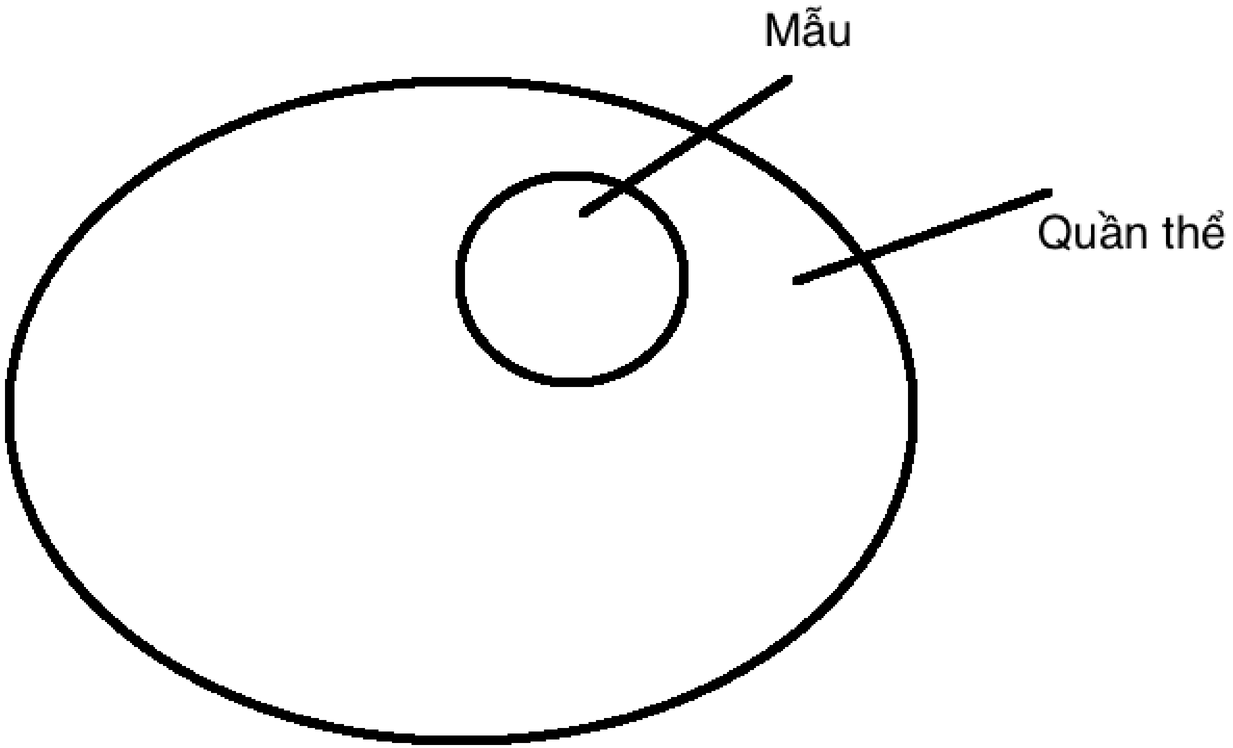
Giới thiệu về thống kê; Ước lượng tham số

Giảng viên: Hoàng Thị Điệp
Khoa CNTT – Đại học Công Nghệ

Nội dung

» Lấy mẫu
» Ước lượng điểm
» Ước lượng khoảng

Quần thể và mẫu



- Tổng thể/Quần thể (population): Tập hợp tất cả các đối tượng mà chúng ta muốn tiến hành nghiên cứu.
- Mẫu (sample): Một tập hợp con các đối tượng trong quần thể mà chúng ta tiến hành thu thập dữ liệu.

Ví dụ

» Khi tiến hành nghiên cứu số lượng bia trung bình 1 người đàn ông VN uống 1 năm, quần thể chúng ta quan tâm nghiên cứu là toàn bộ đàn ông VN.

» Để tiến hành nghiên cứu số lượng bia trung bình 1 người đàn ông VN uống 1 năm, người ta có thể chọn ngẫu nhiên một mẫu gồm 1000 người đàn ông ở các tỉnh, các độ tuổi khác nhau.

» Lưu ý: Số phần tử trong tập mẫu gọi là kích thước mẫu.

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 kê 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 nhiên 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 nhiên 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.

Mẫu ngẫu nhiên/mẫu bị thiên lệch

Để điều tra mức lương ra trường trung bình của sinh viên Trường ĐHCN. Tiến hành lấy mẫu 100 sinh viên như sau:

» Chọn 50 sinh viên khoa cơ, 50 sinh viên khoa CNTT.

» Chọn ngẫu nhiên 100 sinh viên ra trường đang làm việc tại Hà Nội.

» Chọn ngẫu nhiên 100 sinh viên ra trường đang làm việc tại 5 công ty tại Hà Nội.

» Chọn ngẫu nhiên 100 sinh viên mới ra trường, trong đó có 70 sinh viên có điểm học trung bình >2.75.

Phân bố của mẫu và định lí giới hạn trung tâm
Central limit theory

Giả sử $\{X_1, X_2, \ldots, X_n\}$ là một dãy các biến ngẫu nhiên độc lập có cùng phân bố với kì vọng μ và phương sai σ^2. Trung bình cộng

$$\bar{X} = (X_1 + X_2 + \ldots + X_n) / n$$

Theo luật số lớn $\bar{X}$ sẽ tiến gần đến μ theo xác suất. $\bar{X}$ có phân bố chuẩn với kì vọng μ và phương sai σ^2/n.

Ước lượng các tham số tổng thể

Có 2 loại ước lượng:

» **Ước lượng điểm** của một tham số tổng thể là cách thức tính toán một giá trị đơn lẻ của tham số tổng thể dựa trên dữ liệu mẫu.

» **Ước lượng khoảng** của một tham số tổng thể là cách thức tính toán 2 giá trị dựa trên dữ liệu mẫu, từ đó tạo nên một khoảng được kỳ vọng chứa tham số thống kê của tổng thể

Ước lượng điểm:
Ước lượng kì vọng và phương sai từ tập mẫu

Giả sử $S = \{x_1, x_2, \dots, x_n\}$ là một mẫu

Kì vọng μ và phương sai σ^2 có thể được ước lượng như sau:

- Ước lượng kì vọng của quần thể

$$\mu \cong \overline{x} = (x_1 + x_2 + \cdots + x_n)/n$$

- Ước lượng phương sai của quần thể

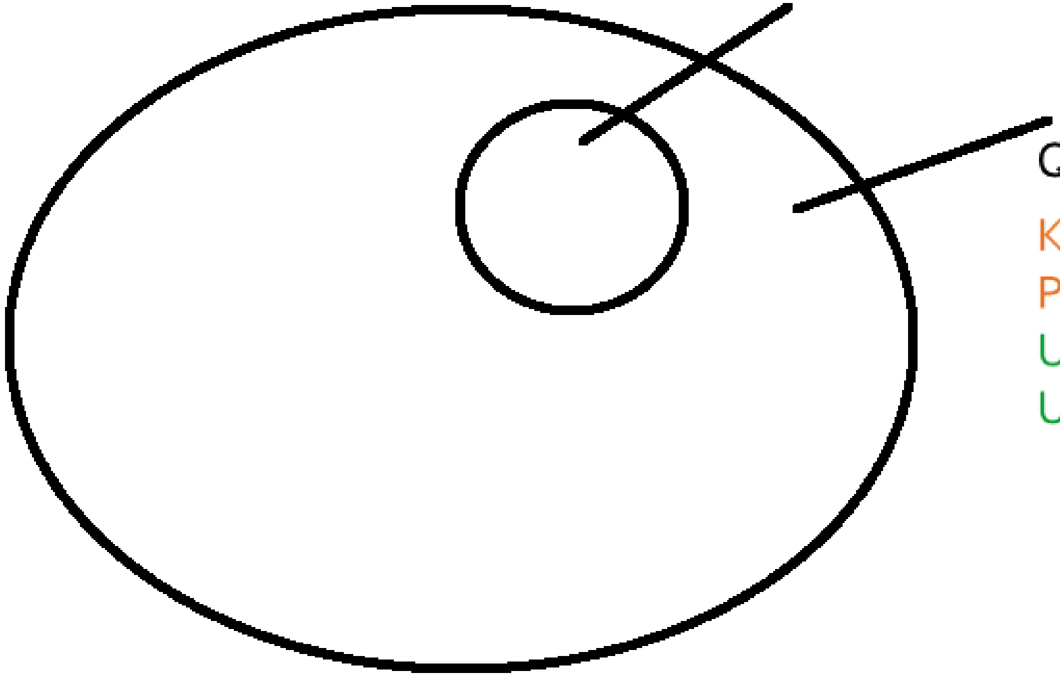
$$\sigma^2 \cong s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Kí hiệu

Kì vọng bnn trung bình mẫu: $\mu_{\bar{X}}$
Phương sai bnn trung bình mẫu: $\sigma^2_{\bar{X}}$

Trung bình mẫu: $\overline{x}$
Phương sai mẫu: s^2

Mẫu



Quần thể

Kì vọng: μ
Phương sai: σ^2
Ước lượng kì vọng: $\overline{x}$
Ước lượng phương sai: s^2

Ví dụ

- Một phương pháp điều trị mới đang được xem xét để đánh giá tính hiệu quả của nó. Một chỉ tiêu đánh giá là số ngày trung bình từ lúc điều trị đến lúc bệnh nhân khỏi bệnh. Một mẫu ngẫu nhiên gồm 11 bệnh nhân được theo dõi và số ngày điều trị cho tới khi khỏi được ghi lại như sau: 4, 4, 3, 8, 5, 6, 7, 12, 5, 3, 8.
- Tìm trung bình mẫu và phương sai mẫu cho số ngày điều trị cho tới khi khỏi.

$$\mu \cong \bar{x} = (x_1 + x_2 + \cdots + x_n)/n$$

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

Ước lượng khoảng:
Khoảng tin cậy cho kì vọng

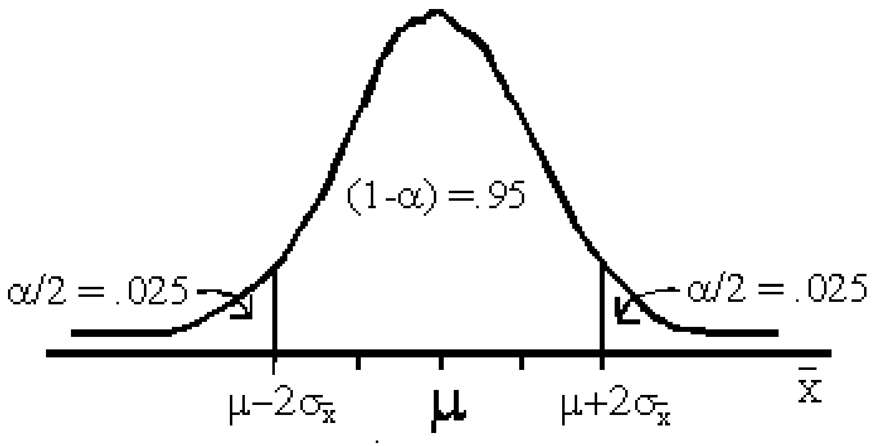
Giả sử $S = \{x1, x2, \dots, xn\}$ là một mẫu (n>=30), kì vọng μ của quần thể

$$\mu \cong (x1 + x2 + \cdots + xn)/n$$

Câu hỏi: Ước lượng khoảng tin cậy β% cho kì vọng μ?

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

The 95% confidence interval for μ



Khoảng tin cậy cho kì vọng

Các trường hợp

» Biết phương sai tổng thể

» Không biết phương sai tổng thể

  - $n >= 30$
  - $n < 30$

Khoảng tin cậy cho kì vọng

Đoạn [a, b] sẽ có dạng

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

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

Ví dụ

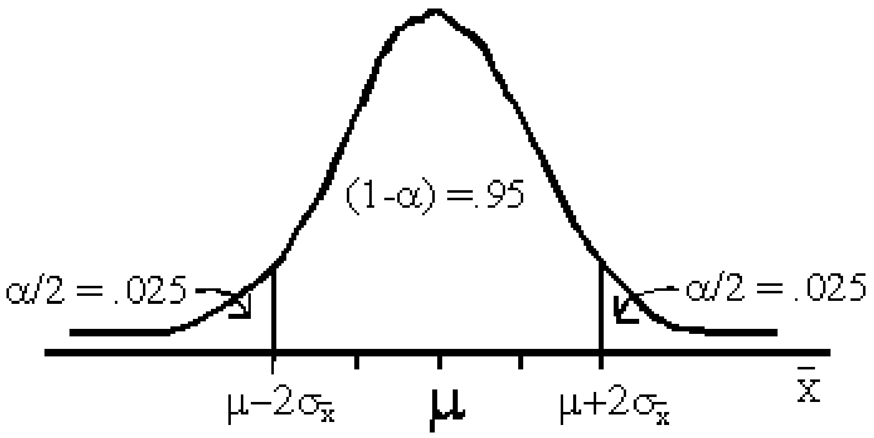
» β = 90%, thì u_β = 1.64

» β = 95%, thì u_β = 1.96

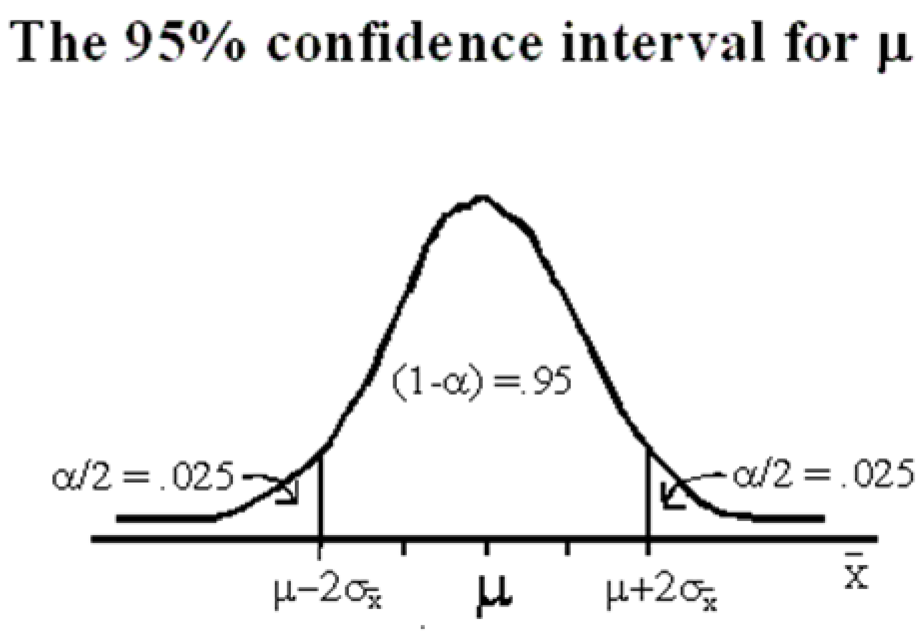
» β = 98%, thì u_β = 2.33

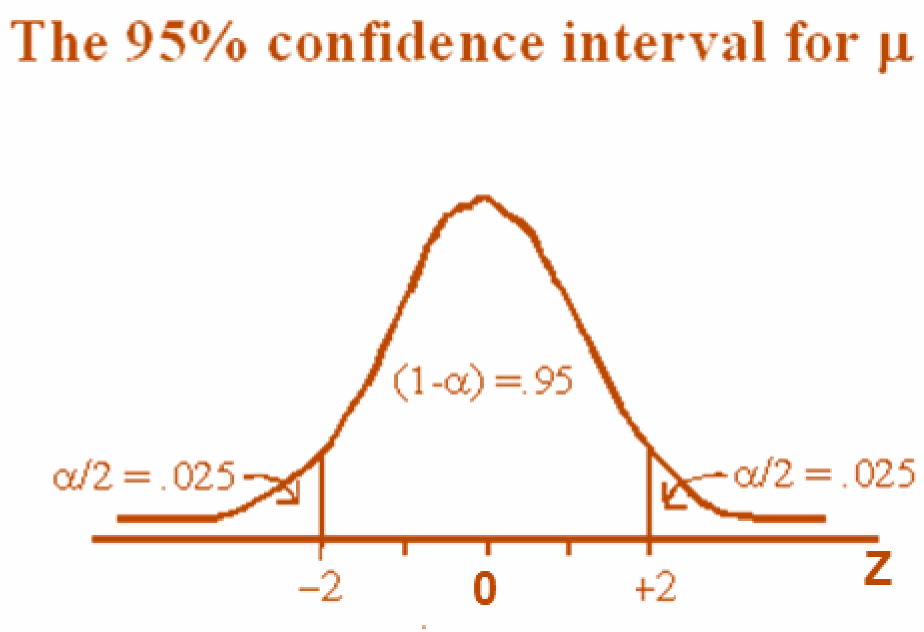
» β = 99%, thì u_β = 2.58

The 95% confidence interval for μ



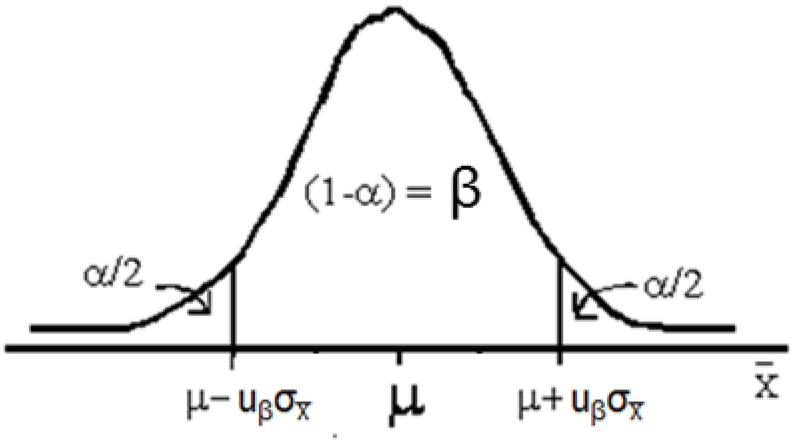
Khoảng tin cậy 95%



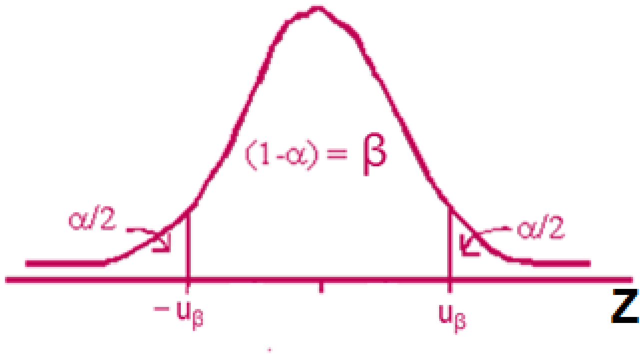


Khoảng tin cậy β

The β confidence interval for μ

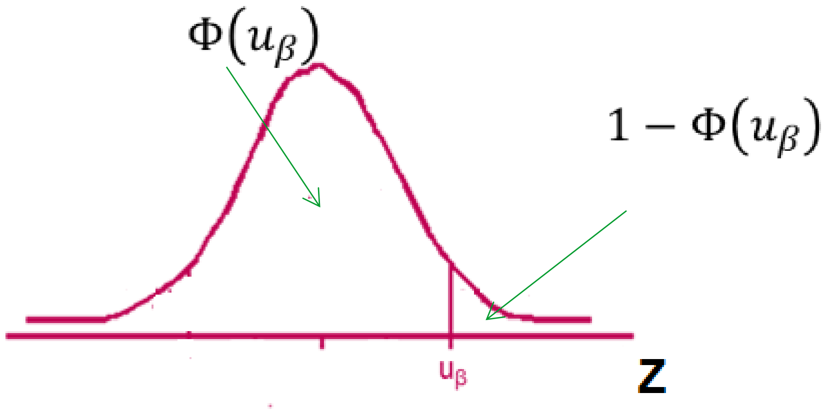


The β confidence interval for μ



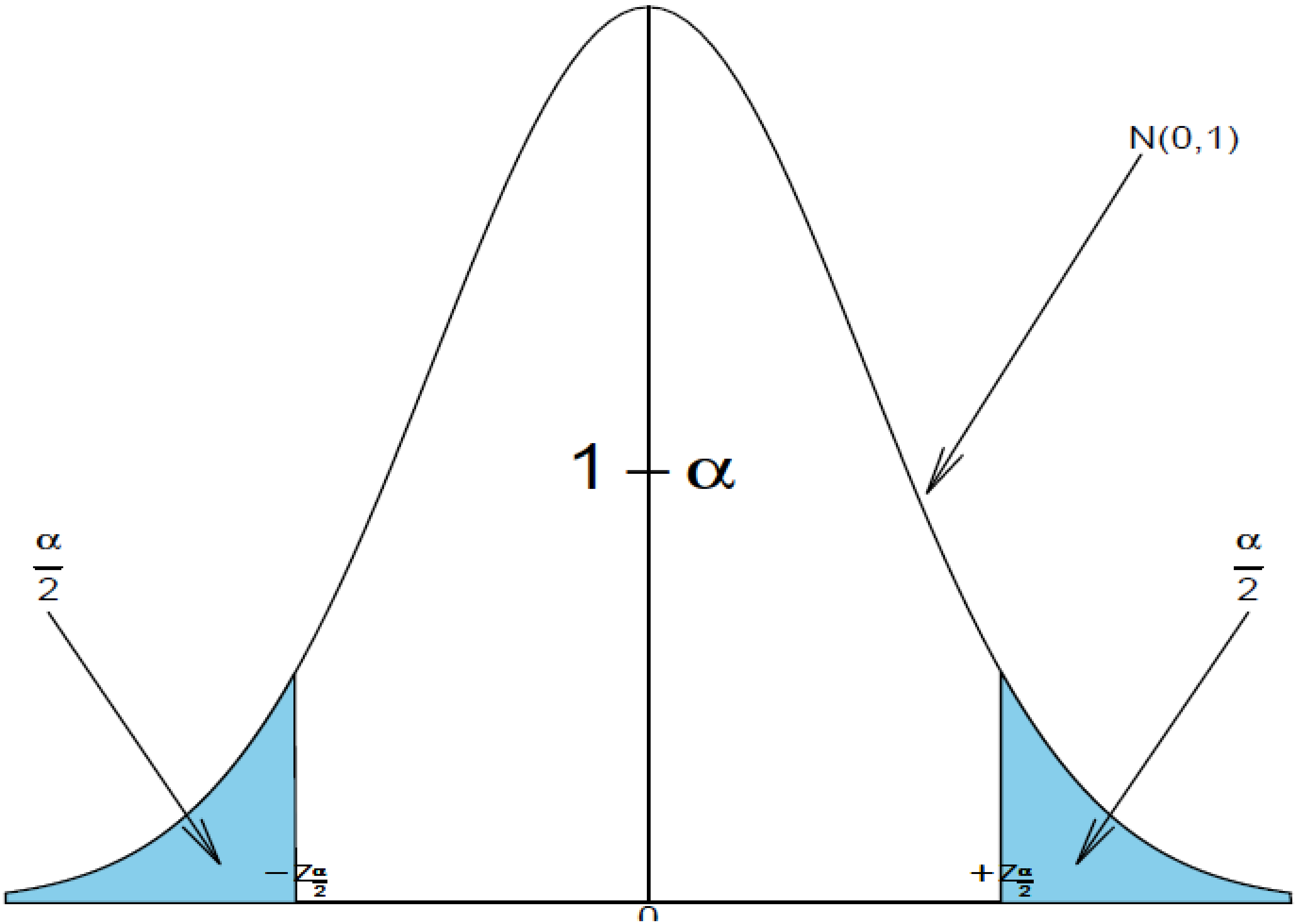
Nhắc lại

- Hàm phân bố tích lũy $F(x) = P(X < x)$
- Với biến ngẫu nhiên phân bố chuẩn tắc, kí hiệu $F(x)$ bằng $\Phi(x)$



- $\Phi(1.96) = 0.975$

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$



Ví dụ

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

β = 90%, thì u_β= 1.64
β = 95%, thì u_β= 1.96
β = 98%, thì u_β= 2,33
β = 99%, thì u_β= 2,58

Chiều cao trung bình của 50 sinh viên ĐHCN là 160 cm. Giả sử độ lệch chuẩn của chiều cao người lớn là 5cm.

a) Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 90%
b) Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 95%
c) Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 99%
d) Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 80%
e) Giải các câu trên với kích thước mẫu là 30, 70, 100, 200 sinh viên.

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?

$$|\overline{X} - \mu| \leq u_\beta \frac{\sigma}{\sqrt{n}}$$

$$hay$$

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

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

Ví dụ

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

Trong một nghiên cứu về chiều cao trung bình của sinh viên ĐHCN, giả sử biết độ lệch chuẩn của chiều cao người lớn là 5cm.

a) 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%.

b) 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%.

c) 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%.

d) 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%.

Phương sai chưa biết và n>=30

» Nếu n>=30, ta có thể xấp xỉ phương sai quần thể bằng:

$$\sigma^2 \cong s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \Rightarrow \sigma^2_{\bar{X}} = s^2 / n$$

» **Ví dụ:** Một trường đại học tiến hành một nghiên cứu xem trung bình một sinh viên tiêu hết bao nhiêu tiền gọi điện thoại trong một tháng. Một mẫu ngẫu nhiên gồm 59 sinh viên được chọn và kết quả như 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, 37, 60, 41, 35, 26,
20, 58, 33, 23, 35.

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

Phương sai chưa biết và n<30

» Trung bình mẫu:

$$\bar{x} = (x_1 + x_2 + \cdots + x_n)/n$$

» Ước lượng phương sai tổng thể:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

» Khi kích thước mẫu nhỏ (n<30), thì $\bar{X}$

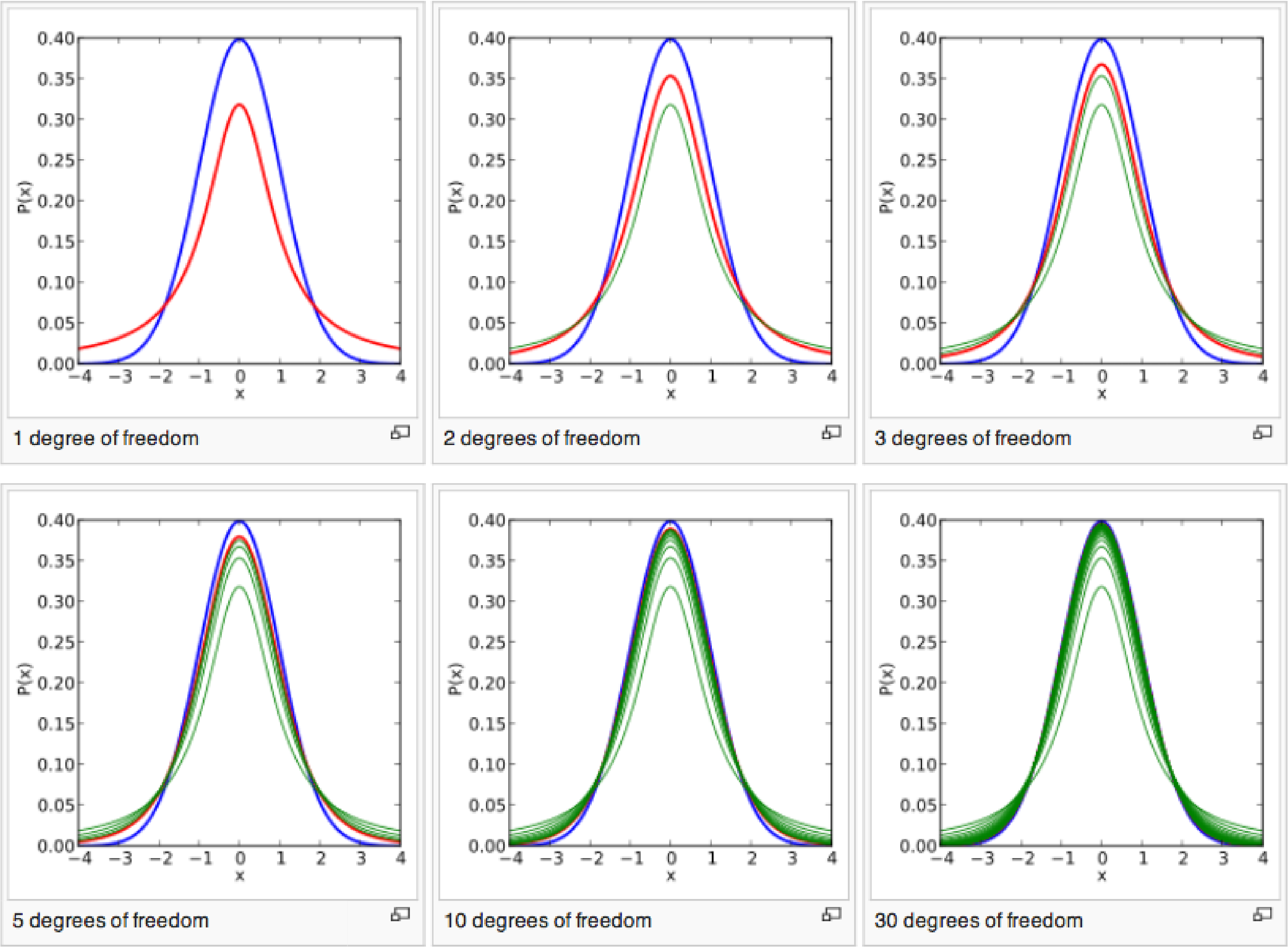
- có phân bố Student (t-distribution) với (n-1) bậc tự do;
- có kì vọng μ và phương sai $\sigma^2_{\bar{x}} = s^2/n$.

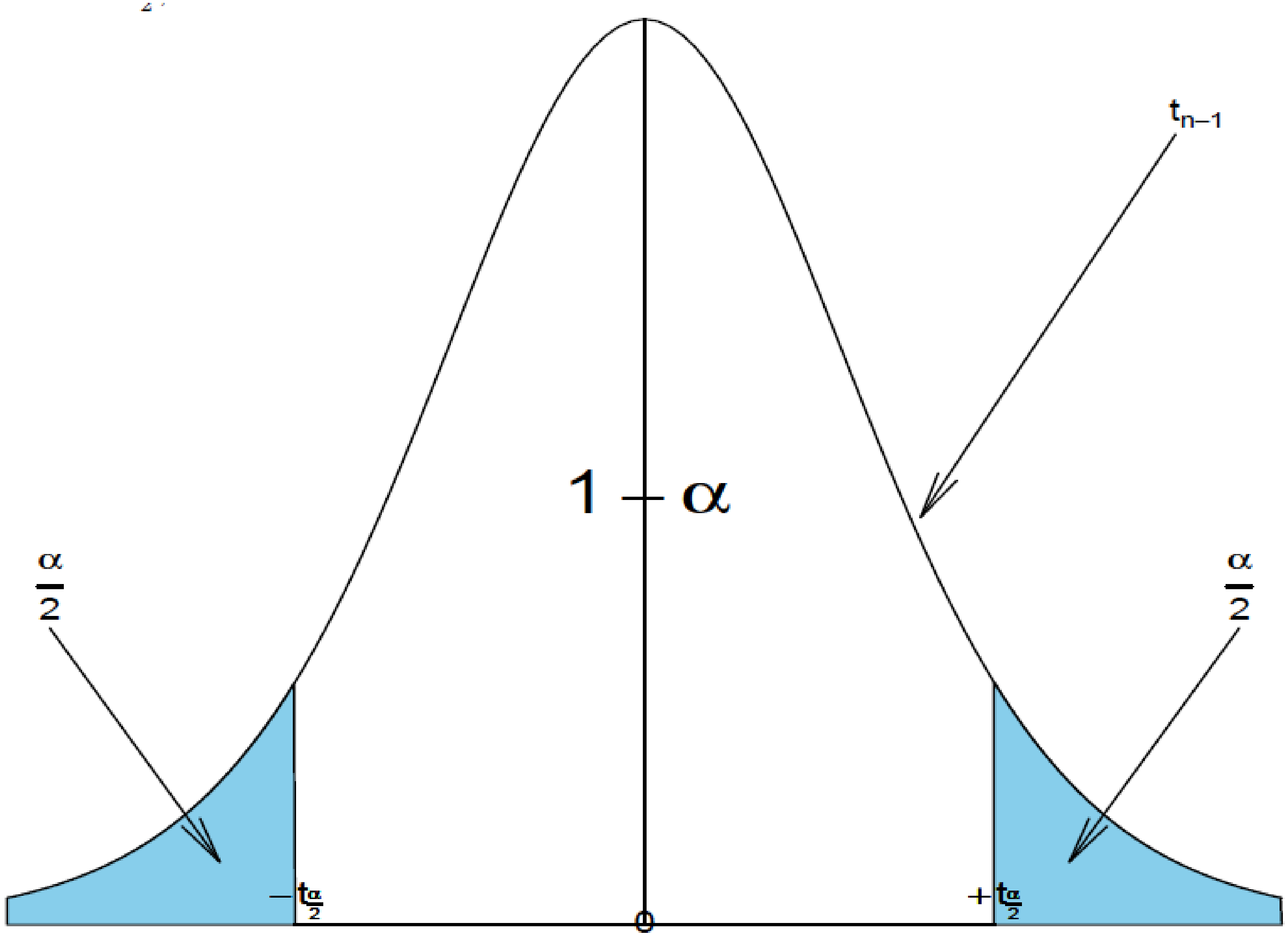
$$[\bar{X} - u_\beta \sigma_{\bar{x}}, \bar{X} + u_\beta \sigma_{\bar{x}}]$$

Phân bố Student

Density of the t-distribution (red) for 1, 2, 3, 5, 10, and 30 degrees of freedom compared to the standard normal distribution (blue).

Previous plots shown in green.





t_{n-1}
$1-\alpha$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$-t_{\frac{\alpha}{2}}$
0
$+t_{\frac{\alpha}{2}}$

Ví dụ

Để ước lượng chiều cao trung bình μ của nữ sinh ĐHCN, một mẫu ngẫu nhiên 16 người được chọn như sau:

162, 155, 170, 165, 160, 165, 158, 164, 168, 150, 165, 167, 164, 159, 152, 154.

Sử dụng phân phối Student để tìm:

a) Hãy tìm khoảng tin cậy cho μ với độ tin cậy β=90%.
b) Hãy tìm khoảng tin cậy cho μ với độ tin cậy β=95%.
c) Hãy tìm khoảng tin cậy cho μ với độ tin cậy β=99%.

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

Sửa bảng ở trang 267

Bậc tự do	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
...				
15	1.753	2.131	2.606	2.947

Ví dụ

Để ước lượng chiều cao trung bình μ của nữ sinh ĐHCN, một mẫu ngẫu nhiên 25 người được chọn như sau:

162, 155, 170, 165, 160, 165, 158, 164, 168, 150, 165, 167, 164, 159, 152, 152, 160, 154, 170, 164, 160, 165, 167, 164, 157.

Sử dụng phân phối chuẩn và phân phối Student để tìm:

a) Hãy tìm khoảng tin cậy cho μ với độ tin cậy $\beta=90\%$.
b) Hãy tìm khoảng tin cậy cho μ với độ tin cậy $\beta=95\%$.
c) Hãy tìm khoảng tin cậy cho μ với độ tin cậy $\beta=99\%$.

Vậy thì sản lượng trung bình của cây sau khi bón phân là bao nhiêu ? Bài toán nảy sinh tiếp theo là bài toán ước lượng giá trị trung bình μ của tổng thể. Một cách hợp lí, trung bình mẫu $\overline{X}$ được dùng làm ước lượng cho μ. Ta hãy tính xác suất để sai số $|\overline{X} - \mu|$ bé hơn ε. Ta đã biết với mẫu lớn thì $\overline{X}$ có phân bố xấp xỉ phân bố chuẩn với kì vọng μ và phương sai $\frac{\sigma^2}{n}$.

Thành thử

$$P\{|\overline{X} - \mu| < \varepsilon\}\} = P\{\mu - \varepsilon < \overline{X} < \mu + \varepsilon\} =$$

$$= \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - \Phi\left(-\frac{\varepsilon\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1.$$

Ấn định xác suất này bằng 0,95 ta có

$$2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 1 = 0{,}95 \Rightarrow \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = \Phi(1{,}96)$$

$$\Rightarrow \varepsilon = \frac{1{,}96\sigma}{\sqrt{n}}.$$

Vậy ta đi tới kết luận : có 95% mẫu ngẫu nhiên thỏa mãn bất đẳng thức

$$\overline{X} - \frac{1{,}96\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{1{,}96\sigma}{\sqrt{n}}.$$

180

Khoảng tin cậy cho tỉ lệ

» 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

Đoạn [a, b] sẽ có dạng

$$[f - u_\beta \sigma_F \, , \, f + u_\beta \sigma_F]$$

- DF = f(1-f)/n
- với điều kiện nf>10 và n(1-f)>10.

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 cho Hilary Clinton

a) Với độ tin cậy 90%
b) Với độ tin cậy 95%
c) Với độ tin cậy 99%

Đoạn [a, b] sẽ có dạng

$$[f - u_\beta \sigma_F , f + u_\beta \sigma_F]$$

DF = f(1-f)/n

với điều kiện nf>10 và n(1-f)>10.

Ví dụ

Một mẫu ngẫu nhiên gồm 100 người dùng xe máy có 30 người dùng xe Honda. Tìm khoảng tin cậy cho tỉ lệ người dùng xe Honda với

a) Độ tin cậy 90%
b) Độ tin cậy 95%
c) Độ tin cậy 96%
d) Độ tin cậy 99%

Đoạn [a, b] sẽ có dạng

$[f - u_\beta \sigma_F , f + u_\beta \sigma_F]$

DF = f(1-f)/n

với điều kiện nf>10 và n(1-f)>10.

Ví dụ

Kiểm tra ngẫu nhiên 300 người có 10 người mắc bệnh tim. Tìm khoảng tin cậy cho tỉ lệ mắc bệnh tim trong toàn dân số với

a) Độ tin cậy 90%
b) Độ tin cậy 95%
c) Độ tin cậy 99%

Bài tập

Bài 1: Để ước lượng chiều cao trung bình μ của nữ sinh ĐHCN, một mẫu ngẫu nhiên 16 người được chọn như sau:

160, 155, 170, 160, 162, 165, 158, 164, 168,

152, 160, 167, 164, 159, 148, 156.

Tìm khoảng tin cậy cho μ với độ tin cậy β=95%.

Bài 2: Chiều cao trung bình của 30 sinh viên ĐHCN là 165 cm. Giả sử độ lệch chuẩn của chiều cao người lớn là 10cm. Tính khoảng tin cậy chiều cao trung bình sinh viên ĐHCN với độ tin cậy 95%

Đề thi mẫu

» Một nhà sản xuất tuyên bố rằng loại pin mới được cải tiến của ông ta tuổi thọ dài hơn loại pin cũ. Biết rằng, loại pin cũ có tuổi thọ tuân theo phân bố chuẩn với kì vọng toán là 150 giờ và phương sai là 16 giờ^2. Để kiểm tra, người ta đo tuổi thọ của 9 pin loại mới được chọn một cách ngẫu nhiên và tính được trung bình mẫu là 153 giờ. Giả sử rằng phương sai của loại pin mới không thay đổi so với loại pin cũ.

a) Hãy xây dựng khoảng tin cậy 90% cho tuổi thọ trung bình của loại pin mới.

b) Độ rộng của khoảng tin cậy 95% bằng bao nhiêu?

Biết $z_{0.05}=1.645$; $z_{0.025}=1.96$; $z_{0.01}=2.326$; $z_{0.005}=2.576$

$t_{0.05;8}=1.860$; $t_{0.025;8}=2.306$; $t_{0.01;8}=2.896$; $t_{0.005;8}=3.355$

Chuẩn bị bài tới

- Đọc chương 3 và làm bài tập cuối chương (Giáo trình *Thống kê và Ứng dụng*)
- Hoàn thành bài tập gửi qua email