3.3. ƯỚC LƯỢNG ĐIỂM

3.3.1. Ước lượng tham số

Ước lượng tham số là một trong những bài toán cơ bản của thống kê toán học. Khi nghiên cứu đặc tính A của mỗi cá thể của tổng thể, nếu xác định được quy luật xác suất của A thì việc đưa ra các đánh giá cũng như các dự báo về sự biến động của tổng thể liên quan đến đặc tính này sẽ chính xác và khách quan. Tuy nhiên không phải lúc nào chúng ta cũng xác định được quy luật xác suất của A. Trong một số trường hợp, ta chỉ biết được dạng toán học của hàm phân phối hoặc hàm mật độ của biến định lượng A mà chưa biết các tham số có mặt trong chúng. Vì vậy để xác định quy luật xác suất của A trước hết phải đưa ra những đánh giá về các tham số này. Bài toán ước lượng tham số sẽ giúp ta giải quyết vấn đề trên.

Bài toán ước lượng tham số có thể phát biểu tổng quát như sau: Cho BNN X của tổng thể có luật phân phối xác suất đã biết nhưng chưa biết tham số θ nào đó, ta phải xác định giá trị của θ dựa trên các thông tin thu được từ một mẫu quan sát $x_1, \cdots, x_n$ của X. Quá trình xác định một tham số θ chưa biết được gọi là quá trình ước lượng tham số. Giá trị tìm được trong quá trình ấy, kí hiệu là $\hat{\theta}$, được gọi là ước lượng của θ. Vì $\hat{\theta}$ là một giá trị số nên nó được gọi là *ước lượng điểm*, sau này ta còn có *ước lượng khoảng* (hay *khoảng tin cậy*).

Rõ ràng, $\hat{\theta}$ là một hàm số n biến $g(X_1, \cdots, X_n)$ nào đó hay là một thống kê, nhận đầu vào là các mẫu thực nghiệm $(x_1, \cdots, x_n)$ của X và đầu ra là giá trị ước lượng của θ. Điều chúng ta muốn có là sai số $|\hat{\theta} - \theta|$ giữa ước lượng $\hat{\theta}$ và giá trị thật của θ càng nhỏ càng tốt. Vì vậy ta phải đưa ra các tiêu chuẩn để đánh giá chất lượng của thống kê $\hat{\theta}$ như là một xấp xỉ tốt nhất của θ. Những tiêu chuẩn như vậy cho ta các nguyên lí thống kê khác nhau.

3.3.2. Các tính chất của ước lượng điểm

Với một mẫu ngẫu nhiên có thể xây dựng nhiều thống kê $\hat{\theta}$ khác nhau để ước lượng cho tham số θ. Vì vậy ta cần lựa chọn thống kê tốt nhất để ước lượng cho tham số θ dựa vào các tiêu chuẩn sau đây.

a) Ước lượng không chệch

Thống kê $\hat{\theta} = g(X_1, \cdots, X_n)$ là một hàm của các BNN $X_1, \cdots, X_n$ nên cũng là một BNN. Do đó ta có thể xét các đặc trưng của thống kê này.

Định nghĩa 3.2. Thống kê $\hat{\theta}$ được gọi là ước lượng không chệch của tham số θ của tổng thể nếu $E(\hat{\theta}) = \theta$.

Từ định nghĩa trên ta thấy $E(\hat{\theta} - \theta) = 0$, điều đó có nghĩa là trung bình độ lệch của ước lượng so với giá trị thật bằng 0. Nếu độ lệch có trung bình khác 0, ta có *ước lượng chệch*. Một sai số nào đó có trung bình khác không sẽ được gọi là *sai số hệ thống*; ngược lại sẽ là *sai số ngẫu nhiên*. Như vậy một ước lượng sẽ được gọi là không chệch khi độ lệch so với giá trị thật (sai số ước lượng) là sai số ngẫu nhiên.

Nhận xét: Giả sử BNN X của tổng thể có kì vọng $E(X) = \mu$, phương sai $D(X) = \sigma^2$. Khi đó:

(i) Từ công thức *(3.1)*, thống kê trung bình mẫu $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là một ước lượng không chệch của μ.

(ii) Từ công thức *(3.2)*, thống kê phương sai mẫu hiệu chỉnh S^2 là một ước lượng không chệch của σ^2, thống kê phương sai mẫu $\hat{S}^2$ là một ước lượng chệch của σ^2. Điều này giải thích vì sao chúng ta hay dùng công thức phương sai mẫu hiệu chỉnh S^2 thay vì $\hat{S}^2$ khi nói về phương sai của một mẫu.

(iii) Từ công thức *(3.4)*, thống kê tần suất mẫu $f = \overline{X}$ là một ước lượng không chệch của xác suất xuất hiện sự kiện A nào đó (nếu X có phân phối Bernoulli và việc lấy mẫu có hoàn lại).

(iv) Phương sai $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ là một ước lượng không chệch của σ^2.

b) Ước lượng hiệu quả

Giả sử $\hat{\theta}$ là ước lượng không chệch của θ, tức là $E(\hat{\theta}) = \theta$. Theo bất đẳng thức Chebyschev (Định lí 2.2), ta có

$$P\big(|\hat{\theta} - E(\hat{\theta})| < \varepsilon\big) \geq 1 - \frac{VD(\hat{\theta})}{\varepsilon^2} \Leftrightarrow P\big(|\hat{\theta} - \theta)| < \varepsilon\big) \geq 1 - \frac{D(\hat{\theta})}{\varepsilon^2}.$$

Nhận thấy $D(\hat{\theta})$ càng nhỏ thì $P\big(|\hat{\theta} - \theta)| < \varepsilon\big)$ càng gần 1. Do đó ta sẽ chọn $\hat{\theta}$ sao cho $D(\hat{\theta})$ nhỏ nhất. Từ đó ta có định nghĩa sau.

Định nghĩa 3.3. Ước lượng không chệch $\hat{\theta}$ được gọi là ước lượng hiệu quả của tham số θ nếu $\hat{\theta}$ có phương sai $D(\hat{\theta})$ nhỏ nhất trong các ước lượng không chệch khác được xây dựng trên cùng một mẫu ngẫu nhiên của θ.

Như vậy, để xét xem ước lượng không chệch $\hat{\theta}$ có phải là ước lượng hiệu quả của θ hay không ta cần phải tìm một cận dưới của phương sai của các ước lượng không chệch và so sánh phương sai của $\hat{\theta}$ với cận dưới này. Điều này được giải quyết bằng bất đẳng thức Cramér-Rao phát biểu như sau.

Định lý 3.1. (Cramér-Rao) *Giả sử BNN X có hàm mật độ xác suất $f(x, \theta)$ trong đó θ là 1 đặc số (trung bình, phương sai, độ lệch chuẩn...) của X và $\hat{\theta}$ là 1 ước lượng không chệch của θ, khi đó*

$$D(\hat{\theta}) \geq \frac{1}{n.E\left(\dfrac{\partial \ln f(x,\theta)}{\partial \theta}\right)^2}. \qquad (3.5)$$

Chú ý:

(i) Bất đẳng thức *(3.5)* được gọi là bất đẳng thức Cramér-Rao, cho biết cận dưới của phương sai các ước lượng không chệch.

(ii) Nếu $\hat{\theta}$ là ước lượng không chệch của θ, $\hat{\theta}$ có phương sai thỏa mãn dấu bằng trong bất đẳng thức *(3.5)* thì $\hat{\theta}$ là ước lượng hiệu quả của θ.

(iii) Nếu BNN của tổng thể $X \sim \mathcal{N}(\mu; \sigma^2)$ thì trung bình mẫu $\overline{X}$ là ước lượng hiệu quả của kì vọng $E(X) = \mu$.

(iv) Nếu BNN của tổng thể $X \sim \mathcal{B}(1; p)$ thì tần suất mẫu $f = \overline{X}$ là ước lượng hiệu quả của tần suất p của tổng thể.

c) Ước lượng vững

Một trong những đặc tính ưa chuộng của ước lượng là khi kích thước mẫu n đủ lớn, ước lượng sẽ có độ tin cậy đủ tốt.

Định nghĩa 3.4. Thống kê $\hat{\theta}$ được gọi là một ước lượng vững của tham số θ nếu với mọi $\varepsilon > 0$ cho trước ta có $\lim\limits_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$.

Như vậy, nếu thống kê $\hat{\theta}$ là một ước lượng vững của θ thì khi n lớn (kích thước mẫu lớn) sự sai khác giữa $\hat{\theta}$ và θ là không đáng kể.

Ví dụ 3.9. Cho tổng thể là BNN X với kì vọng $E(X) = \mu$ và $D(X) = \sigma^2$. Khi đó:

(i) Trung bình mẫu $\overline{X}$ là một ước lượng vững của μ. Thật vậy, $\forall \varepsilon > 0$, áp dụng bất đẳng thức Chebyschev (Định lí 2.2), ta có

$$P\big(|\overline{X} - \mu| < \varepsilon\big) \geq 1 - \frac{D(\overline{X})}{\varepsilon^2} = 1 - \frac{\sigma^2}{n\varepsilon^2} \to 1, \quad khi\ n \to \infty.$$

(ii) Tần suất mẫu $f = \overline{X}$ là ước lượng vững của xác suất p xuất hiện sự kiện A nào đó (nếu X có phân phối Bernoulli).

3.4. ƯỚC LƯỢNG KHOẢNG

Ước lượng điểm có một nhược điểm cơ bản là không thể biết được độ chính xác cũng như xác suất để ước lượng đó chính xác, nhất là khi kích thước mẫu nhỏ, sự sai lệch của ước lượng so với giá trị thật là khá lớn. Để khắc phục các hạn chế đó, người ta dựa vào khái niệm ước lượng bằng một khoảng giá trị. Tất nhiên một khoảng ước lượng vẫn có thể sai giống như mọi ước lượng khác, nhưng khác với ước lượng điểm, xác suất sai lầm có thể biết và trong chừng mực nào đó hy vọng có thể kiểm soát được. Nói như vậy không có nghĩa là không nên dùng ước lượng điểm nữa, nó vẫn cho ta một thông tin quan trọng và ước lượng khoảng sẽ được xây dựng xung quanh ước lượng điểm.

3.4.1. Khoảng tin cậy và độ tin cậy

Định nghĩa 3.5. Giả sử $\hat{\theta}_1$ và $\hat{\theta}_2$ là hai thống kê có từ mẫu ngẫu nhiên $(X_1, X_2.\cdots, X_n)$ và θ là một trong các đặc số của BNN X của tổng thể. Khi đó $[\hat{\theta}_1, \hat{\theta}_2]$ được gọi là khoảng tin cậy của θ với độ tin cậy β nếu

$$P\big(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\big) = \beta.$$

- Độ dài $\hat{\theta}_2 - \hat{\theta}_1$ được gọi là bề rộng của khoảng tin cậy.
- Hệ số $\alpha := 1 - \beta$ được gọi là mức ý nghĩa.

Trong thực tế, khi bề rộng của khoảng tin cậy giảm thì độ tin cậy β cũng giảm theo và ngược lại. Vì vậy, trong thống kê người ta thường cố định độ tin cậy β và tìm một khoảng tin cậy $[\hat{\theta}_1, \hat{\theta}_2]$ ứng với độ tin cậy này sao cho nó có bề rộng càng nhỏ càng tốt. Thông thường, người ta chọn độ tin cậy β trong đoạn $[0,95; 0,999]$, khả năng mắc sai lầm khi dùng các ước lượng khoảng là α.

Để tìm $\hat{\theta}_1$ và $\hat{\theta}_2$ ứng với độ tin cậy β, ta thực hiện theo các bước sau:

Bước 1: Tìm một thống kê $\hat{\theta}$ sao cho phân phối xác suất của $\hat{\theta}$ xác định hoàn toàn (không chứa đặc số θ).

Bước 2: Với độ tin cậy β cho trước, ta tìm cặp số dương α_1 và α_2 thỏa mãn $\alpha_1+\alpha_2=\alpha$ và tương đương với chúng là các phân vị $\hat{\theta}_{\alpha_1}$, $\hat{\theta}_{1-\alpha_2}$ thỏa mãn điều kiện

$$P(\hat{\theta}<\hat{\theta}_{\alpha_1})=\alpha_1 \quad \text{và} \quad P(\hat{\theta}>\hat{\theta}_{1-\alpha_2})=1-P(\hat{\theta}<\hat{\theta}_{1-\alpha_2})=\alpha_2.$$

Khi đó

$$P(\hat{\theta}_{\alpha_1}<\hat{\theta}<\hat{\theta}_{1-\alpha_2})=1-\alpha_2-\alpha_1=1-\alpha=\beta. \qquad (3.6)$$

Bước 3: Bằng các phép biến đổi tương đương ta đưa bất đẳng thức trong *(3.6)* về dạng $\hat{\theta}_1<\theta<\hat{\theta}_2$ và $P(\hat{\theta}_1<\theta<\hat{\theta}_2)=\beta$, đó chính là khoảng tin cậy cần tìm.

3.4.2. Khoảng tin cậy cho kì vọng

Giả sử BNN của tổng thể là $X\sim\mathcal{N}(\mu;\sigma^2)$ với tham số kì vọng μ chưa biết và mẫu ngẫu nhiên $(X_1,\cdots,X_n)$. Bài toán đặt ra là tìm khoảng tin cậy cho $E(X)=\mu$ với độ tin cậy β cho trước.

a) Bài toán 1 (phương sai σ^2 đã biết)

Chọn thống kê $\hat{\mu}:=Z=\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}$. Từ giả thiết phân phối chuẩn của X (hoặc theo Định lí 2.6) ta có $Z\sim\mathcal{N}(0;1)$. Theo *(3.6)*, ta cần tìm các phân vị z_{α_1} và $z_{1-\alpha_2}$ thỏa mãn:

$$\begin{aligned}P(z_{\alpha_1}<Z<z_{1-\alpha_2})=\beta &\Leftrightarrow P(z_{\alpha_1}<\frac{\overline{X}-\mu}{\sigma}\sqrt{n}<z_{1-\alpha_2})=\beta\\ &\Leftrightarrow P(\overline{X}-z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}-z_{\alpha_1}\frac{\sigma}{\sqrt{n}})=\beta.\end{aligned}$$

Do phân vị của phân phối chuẩn có tính chất $-z_{\alpha_1}=z_{1-\alpha_1}$ nên từ đẳng thức trên ta thu được khoảng tin cậy cần tìm là

$$\overline{X}-z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+z_{1-\alpha_1}\frac{\sigma}{\sqrt{n}}. \qquad (3.7)$$

Từ Định nghĩa 2.19 về giá trị tới hạn mức α của phân phối chuẩn tắc, ta có biểu thức *(3.7)* tương đương với

$$\overline{X}-U_{\alpha_2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+U_{\alpha_1}\frac{\sigma}{\sqrt{n}}. \qquad (3.8)$$

Như vậy đối với độ tin cậy β cho trước, ta sẽ có vô số cặp α_1,α_2 thỏa mãn $\alpha_1+\alpha_2=\alpha$ và tương ứng có vô số khoảng tin cậy. Một số trường hợp đặc biệt:

(i) Khoảng tin cậy đối xứng: Nếu ta chọn $\alpha_1=\alpha_2=\dfrac{\alpha}{2}$ thì từ *(3.8)* ta có

$$\overline{X}-U_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu<\overline{X}+U_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Đại lượng $\varepsilon=U_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ được gọi là độ chính xác (hay sai số) của ước lượng, nó phản ánh độ lệch của trung bình mẫu so với kì vọng lí thuyết với độ tin cậy β.

Với độ chính xác ε_0 và độ tin cậy β cho trước thì kích thước mẫu cần thiết là số tự nhiên n nhỏ nhất thỏa mãn: $n\geq\dfrac{\sigma^2U_{\alpha/2}^2}{\varepsilon_0^2}$.

(ii) Khoảng tin cậy phải: Nếu chọn $\alpha_1 = 0$ và $\alpha_2 = \alpha$ thì $U_0 = +\infty$ và khoảng tin cậy là

$$\left(\overline{X} - U_\alpha \frac{\sigma}{\sqrt{n}}; +\infty\right).$$

(iii) Khoảng tin cậy trái: Nếu chọn $\alpha_1 = \alpha$ và $\alpha_2 = 0$ thì $U_0 = +\infty$ và khoảng tin cậy là

$$\left(-\infty; \overline{X} + U_\alpha \frac{\sigma}{\sqrt{n}}\right).$$

Nếu không nói rõ tìm khoảng tin cậy bên phải hay bên trái thì ta quy ước là cần tìm khoảng tin cậy đối xứng.

Ví dụ 3.10. Khối lượng sản phẩm là BNN X có luật phân phối chuẩn, biết rằng phương sai $\sigma^2 = 4g^2$. Kiểm tra 25 sản phẩm, tính được khối lượng trung bình là 20g.

a) Tìm khoảng tin cậy 95% cho khối lượng trung bình của sản phẩm.

b) Nếu sai số ước lượng $\varepsilon = 0,4$g thì độ tin cậy của ước lượng là bao nhiêu?

c) Với $\varepsilon < 0,4$g, muốn độ tin cậy 95% thì phải kiểm tra ít nhất mấy sản phẩm?

Lời giải. Thông tin đầu vào gồm: $\overline{x} = 20$, $\sigma = 2$, $n = 25$.

a) Ta chọn khoảng tin cậy đối xứng, tức là cần tính sai số $\varepsilon = U_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Với độ tin cậy 95% thì $\alpha = 0,05$, tra Bảng II phần Phụ lục ta được

$$\Phi(U_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0,975 \Leftrightarrow U_{\alpha/2} = 1,96.$$

Vậy khoảng ước lượng trung bình khối lượng sản phẩm với độ tin cậy 95% là

$$\left(20 - 1,96 \cdot \frac{2}{5}; 20 + 1,96 \cdot \frac{2}{5}\right) \text{ hay } (19,216; 20,784).$$

b) Với $\varepsilon = 0,4$, khi đó

$$U_{\alpha/2} = \frac{\varepsilon\sqrt{n}}{\sigma} = \frac{0,4.5}{2} = 1.$$

Tra Bảng II ta được $\Phi(1) = 0,8413 = 1 - \frac{\alpha}{2}$ nên $\alpha = 0,3174$. Vậy độ tin cậy là $1 - 0,3174 = 0,6826$ hay $68,26\%$.

c) Với $\varepsilon < 0,4$ và $U_{0,05/2} = 1,96$ thì

$$U_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < 0,4 \Leftrightarrow n > U_{\alpha/2}^2 \cdot \frac{\sigma^2}{(0,4)^2} = (1,96)^2 \cdot \frac{2^2}{(0,4)^2} = 96,04.$$

Vì n là số nguyên nên $n \geq 97$ hay phải kiểm kiểm tra ít nhất 97 sản phẩm.

Chú ý: Công thức sai số ε cho thấy: độ tin cậy $1 - \alpha$ càng lớn thì sai số ε càng lớn, do đó khoảng ước lượng $(\overline{X} - \varepsilon; \overline{X} + \varepsilon)$ cho giá trị thông tin thấp. Kết quả câu b) cho thấy nếu giảm sai số ε thì khoảng ước lượng $(\overline{X} - \varepsilon; \overline{X} + \varepsilon)$ có giá trị thông tin cao nhưng độ tin cậy của ước lượng giảm xuống. Như vậy, muốn có sai số ε nhỏ và độ tin cậy $1 - \alpha$ lớn thì tăng kích thước mẫu n, tương tự kết quả câu c).

b) Bài toán 2 (phương sai σ^2 chưa biết, kích thước mẫu $n \geq 30$)

Trong nhiều bài toán thực tế, ta không biết phương sai σ^2 của BNN tổng thể X. Nhưng nếu kích thước mẫu n đủ lớn ($n \geq 30$), ta có thể xấp xỉ độ lệch chuẩn σ bởi độ lệch chuẩn mẫu hiệu chỉnh S (vì S^2 là ước lượng vững, không chệch của σ^2). Khi đó khoảng tin cậy của tham số μ với độ tin cậy $\beta = 1 - \alpha$ bao gồm:

(i) Khoảng tin cậy đối xứng: $\left(\overline{X} - U_{\alpha/2}\frac{S}{\sqrt{n}}; \overline{X} + U_{\alpha/2}\frac{S}{\sqrt{n}}\right)$.

(ii) Khoảng tin cậy phải: $\left(\overline{X} - U_{\alpha}\frac{S}{\sqrt{n}}; +\infty\right)$.

(iii) Khoảng tin cậy trái: $\left(-\infty; \overline{X} + U_{\alpha}\frac{S}{\sqrt{n}}\right)$.

c) Bài toán 3 (phương sai σ^2 chưa biết, kích thước mẫu $n < 30$)

*- Phân phối **khi bình phương** n bậc tự do $\chi^2(n)$:*

Định nghĩa 3.6. Xét n BNN độc lập $X_i \sim \mathcal{N}(0;1), i = \overline{1,n}$. Khi đó BNN sau có phân phối *khi bình phương* n bậc tự do

$$Z_n = X_1 + X_2 + \cdots + X_n \sim \chi^2(n). \qquad (3.9)$$

Rõ ràng *(3.9)* cho ta cách nhận biết một BNN có phân phối *khi bình phương* xuất phát từ n biến độc lập cùng phân phối chuẩn tắc. Các đặc số quan trọng của phân phối *khi bình phương* gồm: $E(Z_n) = n$, $D(Z_n) = 2n$.

Các tính chất của phân phối χ^2:

(i) Nếu $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ và độc lập thì $X + Y \sim \chi^2(n+m)$.

(ii) BNN $\dfrac{Z_n - n}{\sqrt{2n}} \sim \mathcal{N}(0;1)$ khi $n \to \infty$.

(iii) Giả sử n BNN độc lập $X_i \sim \mathcal{N}(\mu, \sigma^2), i = \overline{1,n}$ và $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ thì $\sum\limits_{i=1}^{n}\left(\dfrac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2(n-1)$.

(iv) Giá trị tới hạn *khi bình phương* n bậc tự do mức α, kí hiệu $\chi^2_\alpha(n)$, được định nghĩa là $P\big(\chi^2 > \chi^2_\alpha(n)\big) = \alpha$.

(v) Bảng các giá trị tới hạn $\chi^2_\alpha(n)$ cho trong Bảng III phần Phụ lục.

*- Phân phối **Student** n bậc tự do $T(n)$*

Định nghĩa 3.7. Cho $X \sim \mathcal{N}(0;1)$ và $Y \sim \chi^2(n)$ là hai BNN độc lập. Khi đó BNN sau có phân phối *Student* n bậc tự do

$$T_n = \frac{X}{\sqrt{Y/n}} \sim T(n).$$

Các tính chất của phân phối Student:

(i) $E(T_n) = 0$, $(n > 1)$ và $D(T_n) = \dfrac{n}{n-2}$, $(n > 2)$.

(ii) Khi n khá lớn thì quy luật Student $T(n)$ hội tụ khá nhanh về phân phối $\mathcal{N}(0;1)$. Trong thực tế, nếu $n > 30$ ta có thể xem thống kê Student xấp xỉ $\mathcal{N}(0;1)$.

(iii) Giá trị tới hạn mức α của phân phối Student n bậc tự do, kí hiệu $t_n(\alpha)$ thỏa mãn

$$P\big(T > t_\alpha(n)\big) = P\big(T < -t_\alpha(n)\big) = P\big(|T| > t_{\alpha/2}(n)\big) = \alpha.$$

(iv) Bảng tính các giá trị tới hạn $t_\alpha(n)$ cho trong Bảng IV phần Phụ lục.

Theo tính chất *(iii)* của phân phối χ^2, ta có

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{\sigma}\right)^2 \sim \chi^2(n-1),$$

khi đó, thống kê

$$T = \frac{\overline{X}-\mu}{S}\sqrt{n} = \frac{\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\dfrac{(n-1)S^2}{\sigma^2(n-1)}}} = \frac{\dfrac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\chi^2/(n-1)}} \sim T(n-1). \qquad (3.10)$$

Tương tự cách xây dựng đối với trường hợp *Bài toán 1*, ta nhận được các khoảng tin cậy của tham số μ với độ tin cậy $\beta = 1-\alpha$ bao gồm:

(i) Khoảng tin cậy đối xứng: $\left(\overline{X} - t_{\alpha/2}(n-1)\cdot\dfrac{S}{\sqrt{n}}; \overline{X} + t_{\alpha/2}(n-1)\cdot\dfrac{S}{\sqrt{n}}\right)$.

(ii) Khoảng tin cậy phải: $\left(\overline{X} - t_{\alpha}(n-1)\cdot\dfrac{S}{\sqrt{n}}; +\infty\right)$.

(iii) Khoảng tin cậy trái: $\left(-\infty; \overline{X} + t_{\alpha}(n-1)\cdot\dfrac{S}{\sqrt{n}}\right)$.

Ví dụ 3.11. Để đánh giá nhiệt độ lớn nhất trung bình ở tỉnh Khánh Hòa vào ngày 5 tháng 9 (giả sử nhiệt độ tuân theo luật chuẩn), người ta lấy số liệu ở 5 vùng của tỉnh đo được trong ngày là 29, 31, 33, 35 và 36 độ C. Xác định khoảng tin cậy 95% cho nhiệt độ cao nhất trung bình trong ngày đang xét.

Lời giải. Gọi X là nhiệt độ cao nhất ở Khánh Hòa vào ngày 05/09, theo giả thiết $X \sim \mathcal{N}(\mu;\sigma^2)$. Từ số liệu đã cho ta có bảng sau

x_i	29	31	33	35	36	$\overline{x} = \frac{164}{5} = 32,8$
$x_i - \overline{x}$	-3,8	-1,8	0,2	2,2	3,2	
$(x_i - \overline{x})^2$	14,44	3,24	0,04	4,48	10,24	$s^2 = \frac{32,8}{4} = 8,2$

Với độ tin cậy 95%, tra Bảng IV phần Phụ lục ta có $t_{0,025}(4) = 2,776$. Vậy khoảng tin cậy là

$$\left(32,8 - 2,776\cdot\sqrt{\frac{8,2}{5}}; 32,8 + 2,776\cdot\sqrt{\frac{8,2}{5}}\right) \approx (29,245; 36,355).$$

Để ý đây là khoảng tin cậy 95% tính trên bộ số liệu cụ thể của ví dụ, nó hoàn toàn không có nghĩa là xác suất để trung bình thật rơi vào khoảng tin cậy trên là 0,95. Bởi vậy không nên quên rằng độ tin cậy 95% của một khoảng nào đó được

hiểu theo nghĩa thông kê (tức là nếu cứ làm thí nghiệm 100 lần với các khoảng tin cậy 95% thì có khoảng 95 lần giá trị trung bình thật nằm trong khoảng đó).

Nhận xét: Nếu BNN gốc không tuân theo luật phân phối chuẩn, việc xác định khoảng tin cậy cho $E(X)$ sẽ rất phức tạp và đòi hỏi các kỹ thuật hiện đại hơn. Tuy nhiên trong trường hợp n đủ lớn, cả hai thống kê $Z = \dfrac{\overline{X} - \mu}{\sigma}\sqrt{n}$ và $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$ đều có phân phối xấp xỉ $\mathcal{N}(0;1)$. Do đó các thủ tục ước lượng khoảng làm giống như *Bài toán 1.*

3.4.3. Khoảng tin cậy cho tỉ lệ

Giả sử ta cần nghiên cứu tính chất A nào đó của tổng thể. Nếu cá thể trong tổng thể có tính chất A thì nhận giá trị 1, trường hợp ngược lại nhận giá trị 0. Khi đó tính chất được nghiên cứu có thể xem là BNN X có quy luật phân phối Bernoulli, tỉ lệ phần tử có tính chất A là p chưa biết. Bài toán đặt ra là ước lượng tỉ lệ cá thể có tính chất A trong khoảng $(f_1; f_2)$ sao cho $P(f_1 < p < f_2) = 1 - \alpha = \beta$.

Lấy mẫu ngẫu nhiên $X_1, \cdots, X_n$ là các BNN độc lập có cùng phân phối Bernoulli với $E(X_i) = p$ và $D(X_i) = p(1-p)$, $i = \overline{1,n}$. Từ *(3.4)*, tần suất mẫu $f = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ có $E(f) = p$ và $D(f) = \dfrac{p(1-p)}{n}$. Theo Định lí Giới hạn trung tâm 2.6 thì

$$\frac{f-p}{\sqrt{p(1-p)}}\sqrt{n} = \frac{(X_1 + \cdots + X_n) - np}{n\sqrt{p(1-p)}} \sim \mathcal{N}(0;1) \text{ khi } n \text{ đủ lớn.}$$

Tuy nhiên vì p chưa biết trong khi tần suất mẫu f là ước lượng không chệch, vững và hiệu quả của tỉ lệ tổng thể p, vì vậy khi n đủ lớn, ta có thể thay p bằng f trong tính toán.

Với điều kiện $\begin{cases} nf > 5 \\ n(1-f) > 5 \end{cases}$, lập luận tương tự *Bài toán 1* ta suy ra các khoảng tin cậy cho tỉ lệ p của tổng thể với độ tin cậy $\beta = 1 - \alpha$ là:

(i) Khoảng tin cậy đối xứng: $\left(f - U_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}}; f + U_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}}\right)$.

(ii) Khoảng tin cậy phải: $\left(f - U_{\alpha}\sqrt{\dfrac{f(1-f)}{n}}; +\infty\right)$.

(iii) Khoảng tin cậy trái: $\left(-\infty; f + U_{\alpha}\sqrt{\dfrac{f(1-f)}{n}}\right)$.

Độ chính xác (sai số) của khoảng tin cậy là $\varepsilon = U_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}}$. Với độ tin cậy β và sai số ε_0 cho trước, kích thước mẫu cần thiết là $n \in \mathbb{N}^*$ nhỏ nhất thỏa mãn

$$n \geq f(1-f)\left(\frac{U_{\alpha/2}}{\varepsilon_0}\right)^2, \qquad (3.11)$$

với f là tần suất của mẫu ngẫu nhiên nào đó.

Ví dụ 3.12. Phỏng vấn 400 người ở một khu vực 300 000 người thấy có 240 người ủng hộ dự luật A.

a) Với độ tin cậy 0,95 hãy ước lượng số người ít nhất ủng hộ dự luật A.

b) Nếu muốn độ chính xác của ước lượng không vượt quá 0,02 thì cần phỏng vấn tối thiểu bao nhiêu người.

Lời giải. Gọi p là tỷ lệ người ủng hộ dự luật A. Tổng thể nghiên cứu là tập hợp 300 nghìn người. Dấu hiệu nghiên cứu là những người sẽ bỏ phiếu ủng hộ dự luật A, có thể xem là BNN có phân phối Bernoulli tham số p.

Theo đề bài ta có $f = \dfrac{240}{400} = 0,6$ thỏa mãn điều kiện $nf = 240 > 5$ và $n(1-f) = 160 > 5$; với $\alpha = 0,05$, tra Bảng II phần Phụ lục ta được $U_{\alpha/2} = 1,96$.

a) Độ chính xác của ước lượng là: $\varepsilon = U_{\alpha/2}\sqrt{\dfrac{f(1-f)}{n}} \approx 0,048$.

Khoảng tin cậy $(0,6 - 0,048; 0,6 + 0,048) = (0,552; 0,648)$. Do đó số người ít nhất ủng hộ dự luật A là $300\ 000.0,552 = 165\ 600$.

b) Theo *(3.11)* thì $n \geq 0,6.0,4.\left(\dfrac{1,96}{0,02}\right)^2 = 2304,96$. Vậy cần phỏng vấn ít nhất 2305 người.