

BÀI 6 ƯỚC LƯỢNG THAM SỐ

Hướng dẫn học

Bài này tiếp tục với bài giảng trước về mẫu và tổng thể. Để có thể dùng mẫu phản ánh về tổng thể, không chỉ tính ra các giá trị thống kê mẫu. Cần kết hợp với lý thuyết xác suất để xây dựng các bài toán suy diễn cho tổng thể. Trong bài giảng này cần chú ý cách kết hợp giữa các thống kê đã tính toán trong bài trước và những giá trị tới hạn tra được, những quy luật phân phối xác suất đã đề cập ở bài giảng số 4.

Để học tốt bài này, sinh viên cần tham khảo các phương pháp học sau:

- Học đúng lịch trình của môn học theo tuần, làm các bài luyện tập đầy đủ và tham gia thảo luận trên diễn đàn.
- Đọc tài liệu: Giáo trình Lý thuyết xác suất và thống kê toán của NXB Đại học KTQD.
- Sinh viên làm việc theo nhóm và trao đổi với giảng viên trực tiếp tại lớp học hoặc qua email.
- Tham khảo các thông tin từ trang Web môn học.

Nội dung

- Khái niệm ước lượng;
- Ước lượng điểm và tính chất, ước lượng một số tham số tổng thể;
- Lý thuyết ước lượng khoảng;
- Ước lượng khoảng cho trung bình, phương sai, tỷ lệ tổng thể và ứng dụng;
- Bài toán xác định kích thước mẫu.

Mục tiêu

- Hiểu được khái niệm ước lượng;
- Tìm được ước lượng không chệch, hiệu quả trong số các ước lượng đã cho;
- Với số liệu mẫu, ước lượng được các tham số tổng thể và suy luận từ đó.

Tình huống dẫn nhập

Ước lượng các tham số về chi tiêu của khách hàng

Với tình huống trong bài giảng số 5, có số liệu về chi tiêu của 100 khách hàng cho trong bảng số liệu ở sau (đơn vị: nghìn đồng). Giả thiết chi tiêu là biến ngẫu nhiên phân phối chuẩn, với độ tin cậy 95%.

Chi tiêu	60–100	100–140	140–180	180–220	220–260	260–300	300–340
Số người	3	9	25	29	21	7	6

?

1. Người quản lý muốn ước lượng mức chi tiêu trung bình của tất cả khách hàng.
2. Người quản lý muốn đánh giá mức độ dao động của mức chi tiêu của khách hàng.
3. Nếu khách hàng chi tiêu từ 260 nghìn trở lên là khách hàng quan trọng thì tỷ lệ khách hàng loại này chiếm bao nhiêu phần trăm trong tổng thể khách hàng.

6.1. Lý thuyết ước lượng

6.1.1. Khái niệm ước lượng

Ước lượng nghĩa là tính toán một cách gần đúng nhất giá trị của một đại lượng chưa biết dựa trên những thông tin đã có. Ta sẽ ước lượng cho những đại lượng đo lường về độ lớn như trung bình; đại lượng đo lường về độ dao động như phương sai, độ lệch chuẩn; đại lượng đo lường về khả năng như xác suất, tỷ lệ tổng thể.

Đại lượng cần đo lường là các tham số tổng thể (trung bình, phương sai, tỷ lệ), là chưa biết. Chính vì chưa biết nhưng lại cần biết để phân tích các quyết định nên cần ước lượng. Thông tin mà ta có đều lấy từ mẫu.

Khái niệm: Ước lượng tham số là tính toán một cách gần đúng nhất giá trị của một tham số chưa biết trong tổng thể dựa trên thông tin từ một mẫu.

Có nhiều tham số trong tổng thể, nhưng trong bài trước chỉ đề cập đến ba tham số chính, vì vậy tại đây ta cũng sẽ tập trung vào ba tham số này, vì vậy ta có ba bài toán:

- Ước lượng trung bình tổng thể: μ
- Ước lượng phương sai tổng thể: σ^2
- Ước lượng tỷ lệ tổng thể: p

Tham số độ lệch chuẩn tổng thể σ luôn phải tính toán thông qua phương sai tổng thể σ^2 do đó không cần tách thành bài toán riêng.

Thay vì phải viết với ba tham số μ , σ^2 , p riêng biệt, tạm thời dùng ký hiệu chung là tham số θ (đọc là tê – ta). Khi viết tham số tổng quát θ thì ta hiểu có ba trường hợp chính là μ , σ^2 , p .

Khi ước lượng cho tham số θ dựa trên thông tin từ mẫu, có hai loại ước lượng là ước lượng điểm và ước lượng khoảng.

Ước lượng điểm: Trong thực tế ta thường dùng khái niệm ước lượng điểm như khi nói: “ước lượng cho lạm phát là 6,5%”; “ước lượng mức tăng trưởng kinh tế là 8%”, nghĩa là chỉ dùng một con số duy nhất để ước lượng.

Ước lượng điểm trong thống kê toán là tìm ra một giá trị, tính toán trên mẫu, do đó tùy thuộc mẫu mà kết quả sẽ có thể khác nhau.

Ước lượng khoảng: Bên cạnh ước lượng điểm “ước lượng lạm phát là 6,5%”, có cách nói thứ hai: “ước lượng mức lạm phát là trong khoảng 6 đến 7 phần trăm”. Đây là hình ảnh của một ước lượng khoảng. Khi ước lượng khoảng ta mong muốn khoảng đó chứa con số cần tìm là đúng với khả năng cao nhất. Từ đó có khái niệm ước lượng khoảng trong thống kê toán.

6.1.2. Khái niệm ước lượng điểm

Khái niệm: Ước lượng tham số bằng một giá trị tính toán trên mẫu gọi là ước lượng điểm cho tham số đó. Với mẫu ngẫu nhiên thì giá trị đó là một thống kê ngẫu nhiên, với mẫu cụ thể thì giá trị đó là một con số.

- Ký hiệu ước lượng điểm của tham số θ là $\hat{\theta}$ (tê – ta mũ), đây là quy ước quốc tế của ước lượng điểm.
- Với mẫu ngẫu nhiên (X_1, X_2, \dots, X_n) thì thống kê có dạng: $\hat{\theta} = f(X_1, X_2, \dots, X_n)$ là một hàm số trên mẫu.

- Với mẫu cụ thể (x_1, x_2, \dots, x_n) thì thống kê có dạng $\hat{\theta}_{qs} = f(x_1, x_2, \dots, x_n)$ và là một con số. Chữ qs viết tắt của quan sát. Giá trị tính trên mẫu gọi là *giá trị quan sát*.

Tuy nhiên với một tham số của tổng thể, có nhiều cách để tìm được một ước lượng điểm. Trong số nhiều ước lượng điểm, cần tìm ra ước lượng điểm chính xác nhất. Để lựa chọn ước lượng tốt, cần có tiêu chuẩn đặt ra.

6.1.3. Tiêu chuẩn lựa chọn ước lượng điểm

Một ước lượng điểm là tốt khi nó không có sai lầm mang tính hệ thống, và sai lầm ngẫu nhiên phải ở mức nhỏ nhất.

Ví dụ 6.1. Để ước lượng trung bình tổng thể m , xét một mẫu ngẫu nhiên gồm hai người sẽ điều tra, hay mẫu kích thước là 2: $W = (X_1, X_2)$. Với mẫu này, có ba ý kiến như sau:

Ý kiến 1: Cho rằng người được điều tra đầu tiên là quan trọng nhất, do đó dùng công thức:

$$\hat{m}_1 = \frac{2X_1 + X_2}{2}$$

Ý kiến 2: Cho rằng người được điều tra đầu tiên là quan trọng, nhưng khi chia cho tổng số thì phải coi người thứ nhất như hai người, do đó dùng công thức:

$$\hat{m}_2 = \frac{2X_1 + X_2}{3}$$

Ý kiến 3: Cho rằng hai người được điều tra quan trọng như nhau, nên dùng công thức:

$$\hat{m}_3 = \frac{X_1 + X_2}{2}$$

Trong ba công thức đó, công thức nào dùng là tốt nhất?

Để giải bài toán này, xét hai tiêu chuẩn của ước lượng điểm là tính *không chệch* và tính *hiệu quả*.

Tính không chệch

Định nghĩa – Tính không chệch: Thống kê $\hat{\theta}$ của mẫu gọi là *ước lượng không chệch* của tham số θ của tổng thể nếu kỳ vọng của nó bằng đúng giá trị tham số.

Vậy $\hat{\theta}$ là ước lượng không chệch của θ thì:

$$E(\hat{\theta}) = \theta \quad (6.1)$$

Nếu $E(\hat{\theta}) \neq \theta$ thì $\hat{\theta}$ là ước lượng chệch của θ . Ước lượng chệch sẽ dẫn đến những sai lệch mang tính hệ thống, ước lượng cao quá hoặc thấp quá giá trị cần ước lượng. Nếu ước lượng chệch được dùng trong các ước lượng tham số khác nữa, thì kết quả sẽ càng sai lầm.

Tính hiệu quả

Giả sử $\hat{\theta}_1, \hat{\theta}_2$ là các ước lượng không chệch của θ , nếu $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ thì ước lượng $\hat{\theta}_1$ được gọi là hiệu quả hơn ước lượng $\hat{\theta}_2$.

Ước lượng không chệch $\hat{\theta}^*$ được gọi là hiệu quả nhất nếu nó có phương sai nhỏ nhất trong số tất cả các ước lượng không chệch được xây dựng trên cùng một mẫu, tức là $V(\hat{\theta}^*) \leq V(\hat{\theta})$ với mọi $\hat{\theta}$ là ước lượng không chệch.

Định nghĩa – Tính hiệu quả: Thống kê $\hat{\theta}$ của mẫu gọi là ước lượng hiệu quả của tham số θ của tổng thể nếu $\hat{\theta}$ là ước lượng không chệch và có phương sai nhỏ nhất trong số các ước lượng không chệch của θ .

Như vậy ước lượng hiệu quả trước tiên phải là ước lượng không chệch. Ước lượng không chệch và hiệu quả được gọi là ước lượng tốt nhất.

Ví dụ 6.1 (tiếp). Tìm ước lượng không chệch, hiệu quả của trung bình tổng thể m trong ba ước lượng sau:

$$\hat{m}_1 = \frac{2X_1 + X_2}{2} \quad \hat{m}_2 = \frac{2X_1 + X_2}{3} \quad \hat{m}_3 = \frac{X_1 + X_2}{2}$$

Giải:

Cần nhớ lại kiến thức của bài trước: Biến ngẫu nhiên gốc X có $E(X) = m$ và $V(X) = \sigma^2$ thì với mẫu ngẫu nhiên thì mọi thành phần mẫu đều có: $E(X_i) = m$ và $V(X_i) = \sigma^2$. Ngoài ra còn cần các tính chất của kỳ vọng và phương sai đã học trong bài số 2.

Do đó để xét tính không chệch, ta tính kỳ vọng của các ước lượng.

$$E(\hat{m}_1) = E\left(\frac{2X_1 + X_2}{2}\right) = \frac{2E(X_1) + E(X_2)}{2} = \frac{2m + m}{2} = \frac{3}{2}m$$

$$E(\hat{m}_2) = E\left(\frac{2X_1 + X_2}{3}\right) = \frac{2E(X_1) + E(X_2)}{3} = \frac{2m + m}{3} = m$$

$$E(\hat{m}_3) = E\left(\frac{X_1 + X_2}{2}\right) = \frac{E(X_1) + E(X_2)}{2} = \frac{m + m}{2} = m$$

Vậy trong ba ước lượng thì ước lượng đầu tiên có kỳ vọng khác với m , nó là ước lượng chệch, không nên sử dụng.

Với hai ước lượng không chệch còn lại, xét tính hiệu quả qua phương sai. Nhớ rằng khi cho hằng số ra ngoài phương sai phải bình phương hằng số đó lên:

$$V(\hat{m}_2) = V\left(\frac{2X_1 + X_2}{3}\right) = \frac{2^2 V(X_1) + V(X_2)}{3^2} = \frac{4\sigma^2 + \sigma^2}{9} = \frac{5}{9}\sigma^2$$

$$V(\hat{m}_3) = V\left(\frac{X_1 + X_2}{2}\right) = \frac{V(X_1) + V(X_2)}{2^2} = \frac{\sigma^2 + \sigma^2}{4} = \frac{1}{2}\sigma^2$$

Ta thấy $V(\hat{m}_2) > V(\hat{m}_3)$ nên ước lượng thứ ba hiệu quả hơn.

Vậy trong ba ước lượng thì ước lượng thứ nhất là chệch, ước lượng thứ hai tuy không chệch nhưng không hiệu quả như ước lượng thứ ba. Ước lượng thứ ba là không chệch và hiệu quả nhất.

Ví dụ 6.2. Biến ngẫu nhiên gốc X có trung bình là m , phương sai là σ^2 . Với mẫu ngẫu nhiên kích thước là 3: $W = (X_1, X_2, X_3)$, cho biết trong số các thống kê sau, hàm nào là ước lượng không chệch, hàm nào là ước lượng hiệu quả hơn:

$$G_1 = \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3; G_2 = \frac{1}{2}X_1 + \frac{1}{3}X_2 + \frac{1}{6}X_3; G_3 = \frac{1}{3}(X_1 + X_2 + X_3)$$

Giải:

Để xét tính không chệch, tính trung bình (kì vọng) của các thống kê. Do mẫu là ngẫu nhiên nên ta có: $E(X_1) = E(X_2) = E(X_3) = m$ và $V(X_1) = V(X_2) = V(X_3) = \sigma^2$. Ta có:

$$\begin{aligned} E(G_1) &= E\left(\frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{1}{3}X_3\right) = \frac{1}{2}E(X_1) + \frac{1}{2}E(X_2) + \frac{1}{3}E(X_3) \\ &= \frac{1}{2}m + \frac{1}{2}m + \frac{1}{3}m = \frac{4}{3}m \neq m \end{aligned}$$

$$\text{Tương tự: } E(G_2) = \frac{1}{2}m + \frac{1}{3}m + \frac{1}{6}m = m; E(G_3) = \frac{1}{3}m + \frac{1}{3}m + \frac{1}{3}m = m$$

Trong các thống kê thì G_1 là ước lượng chệch, G_2, G_3 là các ước lượng không chệch.

Để xét tính hiệu quả, ta tính phương sai của các ước lượng không chệch. Ta có:

$$\begin{aligned} V(G_2) &= \frac{1}{4}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{36}\sigma^2 = \frac{7}{18}\sigma^2 \\ V(G_3) &= \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 + \frac{1}{9}\sigma^2 = \frac{1}{3}\sigma^2 \end{aligned}$$

Do $V(G_2) > V(G_3)$ nên G_3 là ước lượng có phương sai nhỏ nhất, là ước lượng hiệu quả nhất trong số ba ước lượng.

Qua hai ví dụ trên có thể rút ra nhận xét là: Ước lượng cho trung bình tổng thể trên một tổ hợp của các thành phần mẫu thì

- Tổ hợp đó là ước lượng không chệch khi tổng các hệ số bằng 1.
- Tổ hợp không chệch (tổng hệ số bằng 1) là hiệu quả nhất khi các hệ số bằng nhau.

Dựa vào các định lý và tính chất của biến ngẫu nhiên, người ta chứng minh được nếu biến ngẫu nhiên gốc phân phối chuẩn thì trung bình mẫu \bar{X} là ước lượng không chệch, hiệu quả của trung bình tổng thể m (cũng chính là tham số μ); phương sai S^2 là ước lượng không chệch của phương sai tổng thể σ^2 ; tỷ lệ mẫu f là ước lượng không chệch hiệu quả của tần suất tổng thể p . Với một mẫu cụ thể, các ước lượng điểm sẽ được tính là một giá trị cụ thể.

6.1.4. Khái niệm ước lượng khoảng

Trong phần trước, ước lượng điểm là một giá trị dùng để ước lượng cho tham số chưa biết. Trong nhiều trường hợp, ước lượng điểm là chưa đủ và có thể không đảm bảo độ chính xác cần thiết, khi đó cần tìm một khoảng giá trị để ước lượng cho tham số chưa biết. Với một khoảng giá trị dùng để ước lượng cho một tham số chưa biết, có thể không đảm bảo chính xác hoàn toàn, mà chỉ đúng với một xác suất đúng nhất định, và vẫn có khả năng sai.

Khái niệm – Ước lượng khoảng: Ước lượng tham số bằng một khoảng tính toán trên mẫu, sao cho xác suất để khoảng đó chứa con số cần tìm là một giá trị đủ lớn, gọi là ước lượng khoảng cho tham số đó.

Ước lượng khoảng cho tham số θ là tìm một khoảng (θ_1, θ_2) sao cho: $P(\theta_1 < \theta < \theta_2)$ là con số đủ lớn. Nếu ký hiệu xác suất cho phép sai là α thì xác suất yêu cầu đúng là $(1 - \alpha)$, ta có:

$$P(\theta_1 < \theta < \theta_2) = 1 - \alpha \quad (6.2)$$

Khi đó ta có các cách gọi như sau:

- Khoảng (θ_1, θ_2) gọi là khoảng tin cậy của tham số θ .
- Giá trị $(1 - \alpha)$ gọi là độ tin cậy của ước lượng.
- Đại lượng $I = \theta_2 - \theta_1$ gọi là độ dài khoảng tin cậy.

Ước lượng khoảng có độ dài khoảng tin cậy càng ngắn là càng tốt. Thông thường lấy độ tin cậy là 95% hay nói khác đi là xác suất đúng là 95%, cho phép sai là 5%.

6.1.5. Phương pháp tìm ước lượng khoảng

Phương pháp tìm ước lượng khoảng tổng quát tính trên các mẫu ngẫu nhiên sẽ dựa vào các quy luật phân phối xác suất liên hệ đã đề cập trong bài giảng số 5. Việc thực hiện chi tiết các biến đổi có thể xem trong giáo trình.

Với mẫu ngẫu nhiên, công thức ước lượng khoảng sẽ cho một khoảng ngẫu nhiên, và có thể viết dưới dạng xác suất. Với một mẫu cụ thể, thay số vào các đại lượng ngẫu nhiên, sẽ tính ra một khoảng cụ thể, và không gắn với xác suất. Để tránh nhầm lẫn, trong cả hai trường hợp mẫu ngẫu nhiên hay cụ thể đều không viết với xác suất. Người học sẽ sử dụng các công thức đã được chứng minh, thay giá trị bằng số để tính ra kết quả cuối cùng.

6.2. Ước lượng trung bình tổng thể

6.2.1. Ước lượng điểm trung bình tổng thể

Ước lượng điểm không chệch cho trung bình tổng thể chính là trung bình mẫu. Trong bài giảng số 5 ta đã có khu trung bình tổng thể là m thì $E(\bar{X}) = m$ nên \bar{X} là ước lượng không chệch của m . Phương sai của trung bình mẫu: $V(\bar{X}) = \frac{\sigma^2}{n}$.

Chứng minh được khi tổng thể phân phối Chuẩn thì \bar{X} cũng là ước lượng hiệu quả nhất, hay là ước lượng tốt nhất.

6.2.2. Ước lượng khoảng trung bình tổng thể phân phối Chuẩn

Ta chỉ xét bài toán ước lượng khoảng cho trung bình tổng thể khi tổng thể phân phối Chuẩn. Giả sử tổng thể có biến ngẫu nhiên gốc X phân phối chuẩn $X \sim N(\mu, \sigma^2)$, khi đó trung bình tổng thể sẽ được ký hiệu là μ , phương sai tổng thể cũng là phương sai biến ngẫu nhiên σ^2 .

Với mẫu W kích thước n , với độ tin cậy là $(1 - \alpha)$ cho trước

Với W , tính được các thống kê đặc trưng mẫu \bar{X} , S^2 .

Chúng ta xây dựng công thức ước lượng khoảng cho trung bình tổng thể như sau:

$$T = \frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim T(n-1)$$

Chứng minh được công thức sau:

$$P\left(\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)}\right) = 1 - \alpha \quad (6.3)$$

Như vậy xác suất để trung bình tổng thể μ rơi vào khoảng trên bằng $(1 - \alpha)$. Khoảng đó chính là khoảng tin cậy, hoặc ước lượng khoảng cho tham số μ .

$$\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \quad (6.4)$$

Khoảng tin cậy (6.3) đối xứng qua giá trị trung bình nên gọi là khoảng tin cậy đối xứng.

Với một mẫu cụ thể, thay các giá trị thống kê mẫu ngẫu nhiên bởi các con số, sẽ cho kết quả là một khoảng cụ thể. Khoảng cụ thể sẽ là:

$$\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu < \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \quad (6.5)$$

Ta có thể viết khoảng tin cậy dưới dạng:

$$\bar{x} - \varepsilon < \mu < \bar{x} + \varepsilon \quad (6.6)$$

Trong đó ε gọi là sai số, và sai số:

$$\varepsilon = \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \quad (6.7)$$

Sai số của ước lượng càng nhỏ, ta gọi là ước lượng càng chính xác. Khi nói ước lượng chính xác hơn tức là sai số giảm đi. Dựa trên công thức, có thể thấy khi muốn sai số giảm đi, có thể có các cách sau:

- Tăng kích thước mẫu: n tăng lên thì ε giảm.
- Giảm độ tin cậy: $(1 - \alpha)$ giảm thì α tăng và giá trị tới hạn giảm.

Trong trường hợp mà giữ nguyên độ tin cậy, muốn sai số của ước lượng ε không vượt quá một khoảng ε_0 cho trước thì kích thước mẫu tối thiểu cần điều tra được xác định xấp xỉ như sau:

$$n' \geq \frac{s^2}{\varepsilon_0^2} \left(t_{\alpha/2}^{(n-1)}\right)^2 \quad (6.8)$$

Con số n' trên là xấp xỉ, vì khi kích thước mẫu thay đổi thì giá trị tới hạn và s^2 cũng thay đổi, nhưng ta chấp nhận rằng sự thay đổi là không nhiều.

Giá trị tới hạn Student tra trong phụ lục 4, với bậc tự do lớn hơn 30 thì dùng xấp xỉ giá trị tới hạn Chuẩn, chính là dòng cuối của bảng.

Ví dụ 6.3. Khảo sát giá của một loại hàng thiết yếu trên thị trường tự do tại 20 cửa hàng thấy giá trung bình là 135,8 nghìn, với độ dao động đo bởi phương sai là 23,2 nghìn². Giả thiết giá loại hàng này là biến phân phối Chuẩn.

- Với độ tin cậy 95%, ước lượng khoảng giá trung bình thị trường.
- Với độ tin cậy 95%, nếu muốn sai số của ước lượng không quá 2 nghìn thì cần khảo sát thêm ít nhất bao nhiêu cửa hàng nữa?
- Với độ tin cậy 90% thì kết quả ước lượng khoảng như thế nào?

Giải:

Theo như đề bài, thị trường là một tổng thể gồm nhiều cửa hàng, mà ta chỉ điều tra 20 cửa hàng tức là chỉ là một mẫu.

Đặt X là giá của hàng hóa này trên thị trường, đơn vị là nghìn, thì theo giả thiết X phân phối Chuẩn: $X \sim N(\mu, \sigma^2)$.

Lưu ý rằng trong bài này μ là trung bình tổng thể là chưa biết, không phải $\mu = 135,8$ vì con số 135,8 chỉ là trung bình của 20 cửa hàng chứ không phải toàn bộ thị trường gồm rất nhiều cửa hàng. Các con số có trong bài đều là của một mẫu cụ thể.

Như vậy, mẫu có kích thước $n = 20$, trung bình mẫu $\bar{x} = 135,8$ và phương sai mẫu $s^2 = 23,3$; suy ra độ lệch chuẩn mẫu $s = \sqrt{23,3} = 4,827$.

(a) Độ tin cậy 95% tức là $(1 - \alpha) = 0,95$ hay $\alpha = 0,05$, ước lượng giá trung bình của thị trường tức là ước lượng cho trung bình tổng thể.

Công thức:

$$\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu < \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)}$$

Các giá trị \bar{x} và s đã có, cần tra giá trị $t_{\alpha/2}^{(n-1)}$ tại bảng giá trị tới hạn Student – Phụ lục 4. Tra bảng ta có: $t_{\alpha/2}^{(n-1)} = t_{0,05/2}^{(20-1)} = t_{0,025}^{(19)} = 2,093$.

Lưu ý không nhầm lẫn $n = 20$ và $\alpha = 0,05$ trong bài để tra sai thành $t_{0,05}^{(20)}$

Thay số vào ta có:

$$135,8 - \frac{4,827}{\sqrt{20}} \times 2,093 < \mu < 135,8 + \frac{4,827}{\sqrt{20}} \times 2,093$$

$$135,8 - 2,259 < \mu < 135,8 + 2,259$$

$$133,541 < \mu < 138,059$$

Vậy với độ tin cậy 95%, ước lượng khoảng, hay khoảng tin cậy cho giá trung bình của thị trường là (133,541 ; 138,059) nghìn đồng.

Sai số của ước lượng là $\varepsilon = 2,259$ nghìn đồng.

(b) Nếu muốn sai số của ước lượng không quá 2 nghìn đồng, hay $\varepsilon \leq 2$, theo công thức ta có:

$$n' \geq \frac{s^2}{\varepsilon_0^2} (t_{\alpha/2}^{(n-1)})^2 \Leftrightarrow n' \geq \frac{23,3}{2^2} (2,093)^2$$

$\Leftrightarrow n' \geq 25,5$ nhưng do n' là số tự nhiên nên $n' \geq 26$.

Vậy để sai số của ước lượng không quá 2 nghìn đồng thì cần khảo sát *thêm* ít nhất 6 cửa hàng nữa (vì ta đã có số liệu của 20 cửa hàng rồi).

(c) Khi ước lượng khoảng cho μ với độ tin cậy là 90%, đại lượng thay đổi là $t_{\alpha/2}^{(n-1)}$, được tính lại là: $t_{\alpha/2}^{(n-1)} = t_{0,1/2}^{(20-1)} = t_{0,05}^{(19)} = 1,729$

Do đó khi thay số vào công thức ta được:

$$135,8 - \frac{4,827}{\sqrt{20}} \times 1,729 < \mu < 135,8 + \frac{4,827}{\sqrt{20}} \times 1,729$$

$$135,8 - 1,866 < \mu < 135,8 + 1,866$$

$$133,934 < \mu < 137,666$$

Như vậy ta thấy khi độ tin cậy thay đổi thì với kích thước mẫu như cũ thì sai số của ước lượng cũng giảm đi. Tuy nhiên khi sai số ước lượng giảm đi mà không dựa trên việc khảo sát thêm thì độ tin cậy cũng phải giảm, và như vậy khả năng có sai lầm cũng tăng lên.

6.3. Ước lượng phương sai tổng thể

6.3.1. Ước lượng điểm

Phương sai tổng thể là $V(X) = \sigma^2$. Ước lượng không chệch cho phương sai tổng thể trong mẫu ngẫu nhiên chính là phương sai mẫu S^2 , vì ta đã có $E(S^2) = \sigma^2$.

Tuy nhiên việc chứng minh tính hiệu quả của phương sai mẫu không dễ dàng, kể cả khi biến ngẫu nhiên gốc phân phối Chuẩn, đặc biệt khi không có thông tin gì về trung bình tổng thể.

6.3.2. Ước lượng khoảng phương sai tổng thể phân phối Chuẩn

Khi X phân phối chuẩn, phương sai tổng thể σ^2 cần ước lượng, với độ tin cậy $(1 - \alpha)$, dựa trên quy luật của thống kê phân phối Khi – bình phương bậc tự do $(n - 1)$.

Từ đó có công thức ước lượng khoảng, hay khoảng tin cậy của phương sai tổng thể:

$$\frac{(n-1)S^2}{\chi_{\alpha/2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2}^{2(n-1)}} \quad (6.9)$$

Trong đó hai giá trị ở dưới mẫu là giá trị tới hạn Khi – bình phương bậc tự do $(n - 1)$ với mức là $(\alpha/2)$ và $(1 - \alpha/2)$.

Với mẫu cụ thể, thay S^2 bằng s^2 tính từ mẫu, nên khoảng tin cậy là:

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^{2(n-1)}} \quad (6.10)$$

Ví dụ 6.3 (tiếp). Với mẫu 20 cửa hàng khảo sát có trung bình mẫu là 135,8 nghìn và phương sai mẫu là 23,3 nghìn². Giả thiết giá phân phối Chuẩn.

(d) Với độ tin cậy 95% hãy ước lượng độ dao động của giá bán trên thị trường, đo bởi phương sai và độ lệch chuẩn.

(e) Với độ tin cậy 90%, hãy tìm khoảng tin cậy cho độ phân tán của giá bán trên thị trường.

Giải:

Với ví dụ này, $X \sim N(\mu, \sigma^2)$ với σ^2 là phương sai tổng thể chưa biết.

Mẫu cụ thể: $n = 20$, $\bar{x} = 135,8$ và $s^2 = 23,3$.

(d) Đề bài ghi rõ “độ dao động đo bởi phương sai và độ lệch chuẩn” nên ta hiểu đó chính là phương sai và độ lệch chuẩn của tổng thể. Trong nhiều trường hợp, như

câu hỏi (e), đề bài không nói trực tiếp mà dùng cụm từ “độ phân tán”, thì cũng chính là phương sai hoặc độ lệch chuẩn. Độ lệch chuẩn tính trực tiếp từ phương sai, do đó cần ước lượng phương sai:

Công thức ước lượng là:

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^{2(n-1)}}$$

Cần có hai giá trị tới hạn ở dưới mẫu:

$$\chi_{\alpha/2}^{2(n-1)} = \chi_{0,05/2}^{2(20-1)} = \chi_{0,025}^{2(19)} = 32,85$$

$$\chi_{1-\alpha/2}^{2(n-1)} = \chi_{1-0,05/2}^{2(20-1)} = \chi_{0,975}^{2(19)} = 8,907$$

Lưu ý: ký hiệu χ^2 thì số 2 là ký hiệu viết liền, do đó không nhân số 2 vào trong ngoặc tròn.

Thay số vào công thức ta được:

$$\frac{(20-1)23,3}{32,85} < \sigma^2 < \frac{(20-1)23,3}{8,907}$$

$$13,476 < \sigma^2 < 49,702$$

Để có giá trị độ lệch chuẩn, chỉ cần căn bậc hai của phương sai, ta có:

$$3,67 < \sigma < 7,05$$

Vậy với độ tin cậy 95%, ước lượng độ dao động của giá bán trên thị trường, đo bằng phương sai là (13,476; 49,702) nghìn², đo bằng độ lệch chuẩn là (3,67; 7,05) nghìn.

(e) Khi yêu cầu tìm khoảng tin cậy cho “độ phân tán”, ta sẽ hiểu đó là phương sai hoặc độ lệch chuẩn, vì đây là hai đại lượng đo độ phân tán. Với câu hỏi này chỉ có sự khác biệt là độ tin cậy 90% tương ứng với $\alpha = 0,9$. Tính lại hai giá trị tới hạn:

$$\chi_{\alpha/2}^{2(n-1)} = \chi_{0,1/2}^{2(20-1)} = \chi_{0,05}^{2(19)} = 30,14$$

Thay số:

$$\frac{(20-1) \times 23,3}{30,14} < \sigma^2 < \frac{(20-1) \times 23,3}{10,12}$$

$$14,688 < \sigma^2 < 43,745$$

$$3,832 < \sigma < 6,614$$

Vậy với độ tin cậy 95%, ước lượng độ phân tán của giá bán trên thị trường, đo bằng phương sai là (14,688; 43,745) nghìn², đo bằng độ lệch chuẩn là (3,832; 6,614) nghìn.

Ví dụ 6.4 (Tình huống dẫn nhập). Có số liệu về chi tiêu của 100 khách hàng cho trong bảng số liệu sau (đơn vị: nghìn đồng). Giả thiết chi tiêu là biến ngẫu nhiên phân phối chuẩn, với độ tin cậy 95%.

Chỉ tiêu	60 – 100	100 – 140	140 – 180	180 – 220	220 – 260	260 – 300	300 – 340
Số người	3	9	25	29	21	7	6

- (a) Ước lượng điểm và khoảng cho chỉ tiêu trung bình tất cả các khách hàng.
(b) Muốn sai số trong câu (a) còn một nửa thì cần điều ra bao nhiêu hóa đơn khách hàng?
(c) Ước lượng cho độ dao động của chỉ tiêu, đo bằng độ lệch chuẩn.

Giải:

Đặt chỉ tiêu là X , theo giả thiết X phân phối Chuẩn: $X \sim N(\mu, \sigma^2)$ với các tham số là chưa biết.

Từ số liệu của đề bài, thực hiện tính toán như trong bài giảng số 5, ta được các thông tin từ mẫu cụ thể này:

$$n = 100, \bar{x} = 200,4 \text{ (nghìn đồng)}, s^2 = 3086,71 \text{ (nghìn đồng)}^2, s = 55,558 \text{ (nghìn đồng)}$$

Độ tin cậy 95% nghĩa là $\alpha = 0,05$.

- (a) Ước lượng điểm cho chỉ tiêu trung bình tất cả các khách hàng chính là trung bình mẫu, và bằng 200,4 nghìn đồng.

Ước lượng khoảng theo công thức:

$$\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} < \mu < \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)}$$

$$\text{Giá trị tới hạn: } t_{\alpha/2}^{(n-1)} = t_{0,05/2}^{(100-1)} = t_{0,025}^{(99)} \approx 1,96$$

Thay số vào ta có:

$$200,4 - \frac{55,558}{\sqrt{100}} \times 1,96 < \mu < 200,4 + \frac{55,558}{\sqrt{100}} \times 1,96$$

$$200,4 - 10,89 < \mu < 200,4 + 10,89$$

$$189,51 < \mu < 211,29$$

Vậy với độ tin cậy 95%, ước lượng khoảng cho chỉ tiêu trung bình tất cả các khách hàng là trong khoảng (189,51; 211,29) nghìn đồng.

- (b) Muốn sai số trong câu (a) còn một nửa, tức là: $\varepsilon = \frac{10,89}{2} = 5,445$

$$\text{Theo công thức: } \varepsilon = \frac{s}{\sqrt{n}} t_{\alpha/2}^{(n-1)} \text{ nên suy ra: } n = \left(\frac{s}{\varepsilon} t_{\alpha/2}^{(n-1)} \right)^2$$

$$\text{Thay số ta có: } n = \left(\frac{55,558}{5,445} \times 1,96 \right)^2 \approx 400$$

Thực ra kết quả này có thể suy luận trực tiếp từ công thức về mối quan hệ giữa n và ε : đó là quan hệ n tỷ lệ nghịch với bình phương của ε , do đó muốn ε giảm đi một nửa thì phải tăng lên gấp bốn lần.

- (c) Ước lượng cho độ dao động đo bằng độ lệch chuẩn phải tính qua công thức ước lượng phương sai.

Công thức ước lượng là:

$$\frac{(n-1)s^2}{\chi_{\alpha/2}^{2(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^{2(n-1)}}$$

Cần có hai giá trị tới hạn ở dưới mẫu, lấy giá trị xấp xỉ khi trong bảng số không có giá trị chính xác:

$$\chi_{\alpha/2}^{2(n-1)} = \chi_{0,025}^{2(99)} \approx \chi_{0,025}^{2(100)} = 129,6$$

$$\chi_{1-\alpha/2}^{2(n-1)} = \chi_{0,975}^{2(99)} \approx \chi_{0,975}^{2(100)} = 74,22$$

Thay số vào công thức ta được:

$$\frac{(100-1) \times 308,71}{129,6} < \sigma^2 < \frac{(100-1) \times 308,71}{74,22}$$

$$2357,9 < \sigma^2 < 4117,28$$

$$48,558 < \sigma < 64,166$$

Vậy ước lượng cho độ lệch chuẩn của chi tiêu là trong khoảng (48,558 ; 64,166) nghìn đồng.

6.4. Ước lượng tỷ lệ tổng thể

6.4.1. Ước lượng điểm tỷ lệ tổng thể

Với một tổng thể kích thước N , có M phần tử chứa dấu hiệu A , thì $p = \frac{M}{N}$ là tỷ lệ tổng thể, hay tần suất tổng thể của dấu hiệu A . Nếu coi việc xuất hiện dấu hiệu A là một biến cố, thì p chính là xác suất của biến cố đó.

Với một mẫu kích thước n , tỷ lệ mẫu hay tần suất mẫu là f , chứng minh được f là ước lượng điểm không chệch và hiệu quả nhất của p .

6.4.2. Ước lượng khoảng tỷ lệ tổng thể

Cần ước lượng tần suất tổng thể p với độ tin cậy $(1 - \alpha)$ dựa trên một mẫu ngẫu nhiên kích thước n . Trong mẫu, tần suất mẫu của dấu hiệu A là $f = X_A/n$ với X_A là tần số ngẫu nhiên của dấu hiệu A trong mẫu.

Với mẫu ngẫu nhiên hoặc cụ thể, ký hiệu f dùng chung. Bài toán ước lượng khoảng cho tỷ lệ tổng thể thực hiện với mẫu có kích thước từ 100 trở lên: $n \geq 100$.

Ước lượng khoảng, hoặc khoảng tin cậy có dạng đối xứng:

$$f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} \quad (6.11)$$

Ước lượng khoảng này có thể viết dưới dạng:

$$f - \varepsilon < p < f + \varepsilon \quad (6.12)$$

Trong đó ε gọi là sai số, và

$$\varepsilon = \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} \quad (6.13)$$

Do đó nếu muốn sai số của ước lượng không vượt quá giá trị ε_0 cho trước thì kích thước mẫu tối thiểu là:

$$n' \geq \frac{f(1-f)}{\varepsilon_0^2} u_{\alpha/2}^2 \quad (6.14)$$

Ví dụ 6.5. Kiểm tra ngẫu nhiên 400 sản phẩm của một nhà máy sản xuất, thấy có 92 sản phẩm đạt chất lượng loại I. Với độ tin cậy 95%:

- Tỉ lệ sản phẩm loại I của nhà máy nằm trong khoảng nào?
- Nếu muốn sai số không quá 4% thì cần kiểm tra ít nhất bao nhiêu sản phẩm?
- Hãy ước lượng số sản phẩm loại I trong số 40 nghìn sản phẩm?

Giải:

Đặt p là tỉ lệ sản phẩm loại I của nhà máy, p là chưa biết.

Với mẫu, $n = 400$; số sản phẩm loại I, $k = 92$, đặt f là tỉ lệ sản phẩm loại I trong mẫu

$$f = \frac{k}{n} = \frac{92}{400} = 0,23$$

- Ước lượng cho p với độ tin cậy 95% hay $\alpha = 0,05$

Công thức:

$$f - \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2} < p < f + \frac{\sqrt{f(1-f)}}{\sqrt{n}} u_{\alpha/2}$$

Với $u_{\alpha/2} = u_{0,025} = 1,96$, thay số vào ta có:

$$0,23 - \frac{\sqrt{0,23 \times (1-0,23)}}{\sqrt{400}} \times 1,96 < p < 0,23 + \frac{\sqrt{0,23 \times (1-0,23)}}{\sqrt{400}} \times 1,96$$

$$0,23 - 0,0412 < p < 0,23 + 0,0412$$

$$0,1888 < p < 0,2712$$

Vậy tỉ lệ sản phẩm loại I nằm trong khoảng (18,88; 27,12) %

- Nếu muốn sai số không vượt quá 4% hay muốn $\varepsilon \leq 0,04$ thì:

$$n' \geq \frac{0,23 \times (1-0,23)}{0,04^2} \times 1,96^2$$

Vậy $n' \geq 425$; cần kiểm tra ít nhất 425 sản phẩm.

- Khi sản xuất 40000 sản phẩm, muốn ước lượng số sản phẩm loại I chỉ đơn giản là nhân giá trị tỷ lệ với con số 40000. Nếu đặt số sản phẩm loại I trong số 40000 sản phẩm là M thì:

$$M = 40000p$$

Do đó từ kết quả ước lượng p trong câu (a), có

$$0,1888 < p < 0,2712$$

$$0,1888 \times 40000 < 40000p < 0,2712 \times 40000$$

$$7552 < M < 10848$$

Vậy với độ tin cậy 95% có khoảng (7552 ; 10848) sản phẩm loại I.

Tóm lược cuối bài

- Ước lượng tham số là tính toán gần đúng nhất giá trị của tham số chưa biết trong tổng thể, gồm các tham số: μ , σ^2 , p , dựa trên thông tin từ mẫu. Có hai loại ước lượng là ước lượng điểm và ước lượng khoảng.
- Ước lượng điểm được gọi là không chệch khi kỳ vọng của nó bằng đúng tham số, và gọi là hiệu quả khi nó là ước lượng không chệch có phương sai nhỏ nhất.
- Ước lượng khoảng là tìm một khoảng sao cho xác suất đúng bằng một mức cho trước, mức đó gọi là độ tin cậy. Khoảng tìm được gọi là khoảng tin cậy, khoảng có độ dài càng ngắn được coi là càng tốt.
- Ước lượng khoảng cho trung bình tổng thể và tỷ lệ tổng thể có dạng đối xứng, sai số của ước lượng phản ánh độ chính xác. Muốn sai số giảm đi thì hoặc tăng kích thước mẫu hoặc giảm độ tin cậy.
- Từ ước lượng khoảng cho phương sai tổng thể sẽ tính được ước lượng khoảng cho độ lệch chuẩn tổng thể.

Câu hỏi ôn tập

1. Thế nào là ước lượng? Ta thường ước lượng những tham số nào của tổng thể?
2. Phân biệt ước lượng điểm và ước lượng khoảng?
3. Thế nào là ước lượng điểm không chệch? Thế nào là ước lượng điểm hiệu quả?
4. Ước lượng khoảng được thể hiện qua xác suất thế nào?
5. Tại sao gọi công thức ước lượng khoảng của trung bình và tỷ lệ tổng thể là khoảng tin cậy đối xứng?
6. Ước lượng khoảng của trung bình và tỷ lệ tổng thể sẽ thế nào nếu kích thước mẫu tăng lên?
7. Ước lượng khoảng của trung bình và tỷ lệ tổng thể sẽ thế nào nếu độ tin cậy tăng lên?
8. Muốn ước lượng khoảng cho độ lệch chuẩn tổng thể thì phải làm thế nào?