Điều tra chọn mẫu

Nội dung trình bày

- Khái niệm, Phân loại, phạm vi áp dụng
- Chọn mẫu ngẫu nhiên
- Sai số, phạm vi sai số
- Các dạng bài toán cơ bản

Điều tra chọn mẫu

- Là tiến hành điều tra thu thập thông tin trên một số đơn vị của tổng thể chung theo phương pháp khoa học sao cho các đơn vị này phải đại diện cho cả tổng thể chung đó.
- Kết quả điều tra dùng để suy rộng cho cả tổng thể chung.
- Ví dụ: Điều tra mức sống hộ gia đình, điều tra năng suất, diện tích, sản lượng cây trồng trong nông nghiệp, điều tra thị trường sữa trẻ em...
- Đây là hình thức điều tra phổ biến nhất trong thực tế và rất phù hợp với các tổng thể tiềm ẩn.

3

Ưu điểm của điều tra chọn mẫu

- Tiết kiệm hơn cả về mặt thời gian lẫn chi phí so với điều tra toàn bộ.
- Có thể mở rộng nội dung điều tra đi sâu nghiên cứu chi tiết nhiều mặt của hiện tượng.
  - Tài liệu có độ chính xác cao hơn do giảm được sai số phi chọn mẫu:
  - Do phạm vi điều tra nhỏ hơn nên được chuẩn bị kỹ và kiểm tra kỹ lưỡng tỷ mỷ hơn
  - Do số đơn vị điều tra ít nên cần ít điều tra viên, do đó có điều kiện chọn được người có trình độ chuyên môn cao.
  - Dựa trên cơ sở khoa học của lý thuyết xác xuất thống kê và quy luật số lớn nên có thể tính được sai số và độ tin cậy của tài liệu.
- Tiến hành nhanh gọn, bảo đảm tính kịp thời của số liệu thống kê

Nhược điểm của điều tra chọn mẫu

- Không cho biết thông tin đầy đủ, chi tiết về từng đơn vị tổng thể, không cho biết qui mô tổng thể.
- Chắc chắn không tránh khỏi sai số khi suy rộng.
- Kết quả điều tra chọn mẫu không thể tiến hành phân nhỏ theo mọi phạm vi và tiêu thức nghiên cứu như điều tra toàn bộ

Áp dụng điều tra chọn mẫu khi nào?

- Sử dụng để thay thế điều tra toàn bộ trong trường không cho phép điều tra toàn bộ, hoặc do quy mô điều tra toàn bộ quá lớn, cần thu thập nhiều chỉ tiêu nhưng không đủ kinh phí và nhân lực
- Kết hợp với điều tra toàn bộ để mở rộng nội dung điều tra và đánh giá kết quả của điều tra toàn bộ.
- Sử dụng để tổng hợp nhanh tài liệu của điều tra toàn bộ phục vụ kịp thời yêu cầu thông tin cho các đối tượng sử dụng.
- Sử dụng trong trường hợp muốn so sánh các hiện tượng với nhau hoặc muốn đưa ra một nhận định nào đó mà chưa có tài liệu cụ thể.

Phân loại điều tra chọn mẫu

- Chọn ngẫu nhiên: là phương pháp chọn hoàn toàn ngẫu nhiên không phụ thuộc vào ý muốn chủ quan của con người. Khi đó người ta gọi là điều tra chọn mẫu ngẫu nhiên.
  - *Ví dụ:* rút thăm, dùng bảng số ngẫu nhiên.
- Chọn phi ngẫu nhiên: là phương pháp chọn đơn vị điều tra phụ thuộc vào ý muốn chủ quan của người chọn. Khi đó, ta có điều tra chọn mẫu phi ngẫu nhiên.
  - *Ví dụ:* chọn đơn vị trung bình, chọn chuyên gia.
  - Chọn mẫu phi ngẫu nhiên được sử dụng trong trường hợp việc chọn mẫu ngẫu nhiên gặp khó khăn như những cuộc điều tra mới hoàn toàn chưa có một thông tin tiên nghiệm nào về đối tượng điều tra, hoặc có những hiện tượng kinh tế phức tạp, sự phân tán không ổn định, biến động thất thường hoặc nhiều tầng lớp

Chọn mẫu ngẫu nhiên

- Chọn hoàn lại (chọn lặp, chọn nhiều lần): Mỗi khi đơn vị được chọn ra để điều tra sau đó sẽ được trả lại tổng thể chung và có cơ hội được chọn lại.
  - Qui mô của tổng thể chung không thay đổi, xác suất được chọn của mỗi đơn vị là như nhau (đều bằng 1/N). Số mẫu có thể hình thành: $k = N^n$
- Chọn không hoàn lại (chọn một lần): Mỗi khi các đơn vị được chọn ra để điều tra sau đó sẽ được xếp riêng ra không trả lại tổng thể chung và không có cơ hội được chọn lại.
  - Quy mô tổng thể giảm trong quá trình chọn. Số đơn vị trong tổng thể mẫu là hoàn toàn khác nhau và xác suất được chọn của các đơn vị là hoàn toàn khác nhau và tăng dần trong quá trình chọn. Số mẫu có thể hình thành:

$$k = C_N^n = \frac{N!}{n!(N-n)!}$$

Sai số trong điều tra chọn mẫu

- Sai số chọn mẫu là phần chênh lệch giữa kết quả thu được qua điều tra và giá trị thực tế của nó trong tổng thể chung.

$$\theta = \theta' \pm \varepsilon_M \pm \varepsilon_{0M}$$

- Trong đó:
  - θ - các tham số của tổng thể chung
  - θ' - các thống kê của tổng thể mẫu
  - ε_M- sai số chọn mẫu (sai số do tính đại biểu)
  - ε_{0M}- sai số phi chọn mẫu (do ghi chép)

9

Sai số chọn mẫu

- Sai số phi chọn mẫu: sai số này xảy ra ở tất cả các cuộc điều tra: do cân đong đo đếm sai, ghi chép sai, đơn vị điều tra cung cấp sai sự thật. Sai số này không bao giờ xoá bỏ được mà chỉ giảm
- Sai số chọn mẫu: sai số này chỉ xảy ra trong điều tra chọn mẫu, do chỉ điều tra một số ít đơn vị nhưng kết quả lại ước lượng cho cả tổng thể.
  - Sai số hệ thống: xảy ra do vi phạm nguyên tắc chọn, chọn một số đơn vị không đủ lớn để đảm bảo tính chất đại biểu, chọn mẫu không khách quan.
  - Sai số ngẫu nhiên: chỉ xuất hiện trong trường hợp các đơn vị của tổng thể chung được chọn theo nguyên tắc ngẫu nhiên. Nó không lường trước được lệch về hướng nào, nhiều hơn hay ít hơn so với thực tế. Sai số này được giảm dần khi điều tra một số đủ lớn các đơn vị.

Các yếu tố ảnh hưởng đến sai số chọn mẫu

- Số đơn vị tổng thể mẫu n: Khi số đơn vị điều tra tăng lên, tổng thể mẫu sẽ gần với tổng thể chung, sai số chọn mẫu sẽ giảm.

Khi n -> N thì $|\mu - \bar{x}| \to 0$, $|p - f| \to 0$.

- Phương pháp tổ chức chọn mẫu: Các phương pháp chọn mẫu khác nhau, tính đại diện của mẫu chọn ra khác nhau cũng sẽ dẫn đến những sai số chọn mẫu khác nhau.
- Độ đồng đều của tổng thể chung: nếu tổng thể có độ đồng đều cao tức phương sai tổng thể σ^2 tương đối nhỏ thì sai số chọn mẫu sẽ nhỏ.

Sai số bình quân chọn mẫu

- Sai số bình quân chọn mẫu là một trị số sai số chọn mẫu đại diện cho các giá trị của sai số chọn mẫu hay nói cách khác đó là bình quân của tất cả các sai số chọn mẫu do việc lựa chọn mẫu có kết cấu thay đổi.

Công thức tính sai số bình quân chọn mẫu

Suy rộng \ Cách chọn	Hoàn lại (chọn nhiều lần)	Không hoàn lại (chọn một lần)
Bình quân	$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}}$ hoặc $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_0^2}{n-1}}$	$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}(1-\frac{n}{N})}$ hoặc $\sigma_{\bar{x}} = \sqrt{\frac{\sigma_0^2}{n-1}(1-\frac{n}{N})}$
Tỷ lệ	$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$ hoặc $\sigma_p = \sqrt{\frac{f(1-f)}{n-1}}$	$\sigma_p = \sqrt{\frac{p(1-p)}{n}(1-\frac{n}{N})}$ hoặc $\sigma_p = \sqrt{\frac{f(1-f)}{n-1}(1-\frac{n}{N})}$

Phạm vi sai số chọn mẫu

- Phạm vi sai số chọn mẫu hay độ chính xác của suy rộng là chênh lệch giữa các chỉ tiêu của tổng thể mẫu và các chỉ tiêu tương ứng của tổng thể chung với độ tin cậy nhất định.
- Để ước lượng được tham số của tổng thể chung từ tham số của tổng thể mẫu thì phải xác định được phạm vi sai số chọn mẫu.
- Phạm vi sai số chọn mẫu được ký hiệu là ε và được xác định bằng công thức:

$$\varepsilon = z.\ \sigma$$

- *Trong đó:*

  σ: sai số bình quân chọn mẫu

  z: hệ số tin cậy

- Theo chứng minh của toán học thì z tương ứng với hàm xác suất $\phi(z)$ đã được Liapunốp tính sẵn thành bảng riêng. Ý nghĩa của hàm xác suất này được biểu hiện như sau:

$$P(|\mu - \bar{x}| \le \varepsilon_{\bar{x}}) = \Phi(z) = 1 - \alpha$$

$$P(|p - f| \le \varepsilon_p) = \Phi(z) = 1 - \alpha$$

- Với α là xác suất sai lầm; (1 - α) gọi là độ tin cậy ước lượng.
- Sau đây là một vài trị số tiêu biểu của z:

  z = 1 thì $\phi(z)$ = 0.6827

  z = 2 thì $\phi(z)$ = 0.9545

  z = 3 thì $\phi(z)$ = 0.9973

Ý nghĩa:

- Với xác suất là 0.6827 các tham số của tổng thể chung và tổng thể mẫu chênh lệch nhau không quá ±σ, hay $\theta = \theta' \pm \sigma$;
- Với xác suất là 0.9545 các tham số của tổng thể chung và tổng thể mẫu chênh lệch nhau không quá ±2σ, hay $\theta = \theta' \pm 2\sigma$;
- Với xác suất là 0.9973 các tham số của tổng thể chung và tổng thể mẫu chênh lệch nhau không quá ±3σ, hay $\theta = \theta' \pm 3\sigma$.

Suy rộng kết quả điều tra chọn mẫu

- Qua điều tra chọn mẫu chúng ta tính toán được các tham số , f, σ_0^2 của tổng thể mẫu. Nhưng mục đích của chúng ta là các tham số của tổng thể chung μ, p, σ^2. Do đó ta phải ước lượng, nghĩa là từ các tham số của tổng thể mẫu suy ra các tham số của tổng thể chung.
- Khi kích thước mẫu nhỏ ước lượng điểm cho sai số lớn và thường không đánh giá được khả năng mắc sai lầm khi ước lượng.
- phương pháp được sử dụng phổ biến nhất để ước lượng kết quả điều tra là phương pháp ước lượng khoảng.

- Dựa vào tham số của tổng thể mẫu và phạm vi sai số chọn mẫu tính toán được, các tham số của tổng thể chung được thống kê toán ước lượng như sau:
- Khi ước lượng số bình quân

$$\mu = \bar{x} \pm \varepsilon_x = \bar{x} \pm z.\sigma_x \qquad \bar{x} - z.\sigma_x \le \mu \le \bar{x} + z.\sigma_x$$

- Khi ước lượng tỷ lệ theo một tiêu thức nào đấy:
- $p = f \pm \varepsilon_p = f \pm z.\ \sigma_p$ hay $f - z.\ \sigma_p \le p \le f + z.\ \sigma_p$

. Xác định số đơn vị mẫu cần điều tra

- có rất nhiều phương pháp xác định cỡ mẫu khác nhau tuỳ thuộc vào từng điều kiện cụ thể.
- người ta thường căn cứ vào độ chính xác khi suy rộng kết quả điều tra chọn mẫu. Biểu hiện tập trung nhất của độ chính xác là phạm vi sai số chọn mẫu

Từ công thức tính phạm vi sai số chọn mẫu, suy ra công thức tính số đơn vị mẫu cần điều tra như sau

Cách chọn / Suy rộng	Chọn hoàn lại (chọn nhiều lần)	Chọn không hoàn lại (chọn một lần)
Bình quân	$n = \frac{z^2\sigma^2}{\varepsilon_x^2}$	$n = \frac{Nz^2\sigma^2}{N\varepsilon_x^2 + z^2\sigma^2}$
Tỷ lệ	$n = \frac{z^2p(1-p)}{\varepsilon_p^2}$	$n = \frac{Nz^2p(1-p)}{N\varepsilon_p^2 + z^2p(1-p)}$

Tính toán các tham số trong công thức

- Thực tế điều tra, người ta thường cho trước hệ số tin cậy và phạm vi sai số chọn mẫu nhưng phương sai thì chưa biết. Khi đó có thể xác định phương sai của tổng thể chung bằng một trong những cách sau:
  - Lấy phương sai lớn nhất trong những lần điều tra trước (nếu có) hoặc chọn p nào gần với 0.5 nhất.
  - Lấy phương sai của các cuộc điều tra khác có tính chất tương tự.
  - Điều tra thí điểm để xác định phương sai.
  - Ước lượng phương sai nhờ khoảng biến thiên. Thống kê toán đã chứng minh trong trường hợp hiện tượng phân phối chuẩn thì:
- CT: $$\sigma = \frac{R}{6} = \frac{x_{max} - x_{min}}{6}$$

Các nhân tố tác động đến kích thước mẫu điều tra

- Hệ số tin cậy z: Nếu yêu cầu trình độ tin cậy của ước lượng là lớn, tức hệ số tin cậy z lớn thì số đơn vị mẫu điều tra nhiều và ngược lại.
- Độ đồng đều của tổng thể chung (σ^2): Nếu tổng thể biến thiên lớn thì σ^2 tính ra lớn vì thế số đơn vị mẫu điều tra nhiều và ngược lại.
- Phạm vi sai số chọn mẫu ε: Nếu phạm vi sai số chọn mẫu lớn thì số đơn vị mẫu điều tra nhỏ và ngược lại.
- Trong trường hợp chọn không hoàn lại, qui mô của tổng thể chung có thể ảnh hưởng đến cỡ mẫu khi mà qui mô tổng thể không lớn.

Ví dụ: Doanh nghiệp A có 3000 lao động. Người ta tiến hành chọn ngẫu nhiên 300 lao động theo cách chọn không lặp để điều tra về năng suất lao động bình quân của công nhân trong doanh nghiệp và thu được kết quả sau:

NSLĐ (1000 đ)	Số LĐ (Người)	x_i	$x_i f_i$	$x_i^2 f_i$
40-50	25	45	1125	50625
50-60	40	55	2200	121000
60-70	70	65	4550	295750
70-80	85	75	6375	478125
80-90	60	85	5100	433500
90 trở lên	20	95	1900	180500
Tổng	**300**		**21250**	**1559500**

a.) Tính năng suất lao động bình quân của công nhân toàn doanh nghiệp với xác suất bằng 0,9544.

- NSLĐ bình quân của công nhân trong mẫu điều tra.

$$\bar{x} = \frac{\sum x_i f_i}{\sum x_i} = \frac{21250}{300} = 70.83$$

- Phương sai của mẫu

$$\sigma_0^2 = \frac{\sum x_i^2 f_i}{\sum f_i} - \left(\frac{\sum x_i f_i}{\sum f_i}\right)^2 = \frac{1559500}{300} - (70.83)^2 = 181.44$$

- Sai số bình quân chọn mẫu theo cách chọn không lặp:

$$\sigma_x = \sqrt{\frac{\sigma_0^2}{n-1}(1-\frac{n}{N})} = \sqrt{\frac{181.44}{300-1}(1-\frac{300}{3000})} = 0.739$$

- Công thức để suy rộng năng suất lao động bình quân của công nhân toàn doanh nghiệp với xác suất bằng 0,9544 hay hệ số tin cậy z=2.

$$\bar{x} - z.\sigma_x \le \mu \le \bar{x} + z.\sigma_x$$

- Thay số vào

$$70.83 - 2.0.739 \le \mu \le 70.83 + 2.0.739$$

$$69.352 \le \mu \le 72.308$$

- Vậy năng suất lao động bình quân của công nhân toàn doanh nghiệp nằm trong khoảng 69,352 đến 72,308 (nghìn đồng) với xác suất 0,9544.

b) Tính xác suất khi suy rộng tài liệu về năng suất lao động bình quân một công nhân trong doanh nghiệp với phạm vi sai số chọn mẫu không vượt quá 2,22 nghìn đồng

$$z = \frac{\varepsilon_x}{\sigma_x} = \frac{2.22}{0.739} = 3$$

- Với z=3 thì xác suất Φ(z)=0,9973.

c) Với xác suất bằng 0,9544 và phạm vi sai số chọn mẫu khi suy rộn về NSLĐ bình quân không vượt quá 2 nghìn đồng. Tính số công nhân cần điều tra theo cách chọn không lặp

- Áp dụng công thức tính số mẫu cần điều tra khi ước lượng số trung bình và theo cách chọn không lặp:

$$n = \frac{Nz^2\sigma^2}{N\varepsilon_x^2 + z^2\sigma^2} = \frac{3000.2^2.181{,}44}{3000.2^2 + 2^2.181{,}44} = 171{,}1$$ hay 172 (người)

- (trong trường hợp này lấy phương sai của tổng thể chung là phương sai của mẫu trong lần điều tra trước bằng 181.44 vừa tính được ở câu a).
- Vậy số công nhân cần điều tra là 172 người.

d) Với xác suất 0,9545, hãy xác định tỷ lệ công nhân có mức NSLĐ từ 80 nghìn đồng trở lên trong doanh nghiệp.

- Tỷ lệ công nhân có mức NSLĐ từ 80 nghìn đồng trở lên ở mẫu điều tra là:

$$f = \frac{60+20}{300} = 0.267$$

- Sai số bình quân chọn mẫu khi ước lượng tỷ lệ theo cách chọn không lặp:

$$\sigma_p = \sqrt{\frac{f(1-f)}{n-1}(1-\frac{n}{N})} = \sqrt{\frac{0.267(1-0.267)}{300-1}(1-\frac{300}{3000})} = 0.024$$

- Công thức để suy rộng tỷ lệ công nhân có mức NSLĐ từ 80 nghìn đồng trở lên trong toàn doanh nghiệp với hệ số tin cậy z=2:

$$f - z.\sigma_p \le p \le f + z.\sigma_p$$

- Thay số: 0,267-2.0,024≤ p ≤ 0,267+2.0,024

0,219≤ p ≤ 0,315 (lần)

e) Tính xác suất khi suy rộng tài liệu về tỷ lệ số công nhân có mức NSLĐ từ 80 nghìn đồng trở lên với phạm vi sai số chọn mẫu không vượt quá 7,21%

$$z = \frac{\varepsilon_p}{\sigma_p} = \frac{0.721}{0.024} = 3$$

- Vậy xác suất khi suy rộng là 0.9973.

- f) Với xác suất bằng 0,9545 và phạm vi sai số chọn mẫu không vượt quá 5% khi suy rộng về tỷ lệ số công nhân có mức NSLĐ từ 80 nghìn đồng trở lên, hãy tính số công nhân cần điều tra theo cách chọn không lặp.

$$n = \frac{Nz^2p(1-p)}{N\varepsilon_p^2 + z^2p(1-p)} = \frac{3000.2^2.0{,}267(1-0{,}267)}{3000.0{,}05^2 + 2^2.0{,}267(1-0{,}267)} = 283{,}54$$

- *<u>Lưu ý:</u>* Khi tính cỡ mẫu, phải luôn làm tròn lên.