

Phân phối chọn mẫu của \bar{X}

Tổng thể $X \sim$ có phân phối chuẩn và phương sai σ^2 đã biết.

Chúng ta biết rằng $\bar{X} \sim N(\mu, \sigma_{\bar{X}}^2)$, trong đó $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ khi tổng thể này tuân theo phân phối chuẩn. Giá trị của thông tin này cho phép chúng ta hình thành một trình bày xác suất cho trung bình mẫu.

Ví dụ 1: Giả sử $X \sim N(100, 16)$. Nếu chúng ta rút ra một mẫu có độ lớn $n = 4$, xác suất để trung bình mẫu sẽ có giá trị lớn hơn hoặc bằng 102 là bao nhiêu?

$$\text{Đầu tiên tính } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{16}{4} = 4$$

Sau đó, viết trình bày xác suất:

$$P(102 \leq \bar{X}) = P\left(\frac{102 - \mu}{2} \leq \frac{\bar{X} - \mu}{2}\right) = P\left(\frac{102 - 100}{2} \leq Z\right) = P(1 \leq Z)$$

Ví dụ 2: Giả sử tổng thể tuân theo phân phối chuẩn, nhưng trung bình tổng thể chưa xác định. Tuy nhiên, giả sử rằng phương sai tổng thể được biết là $\sigma^2 = 16$.

Quy mô mẫu phải là bao nhiêu để chênh lệch giữa trung bình mẫu và trung bình tổng thể nằm trong phạm vi một đơn vị với xác suất 95%?

Trong trường hợp này, chúng ta viết trình bày xác suất như sau:

$$P(-1 \leq \bar{X} - \mu \leq 1) = P\left(\frac{-1}{4/\sqrt{n}} \leq \frac{\bar{X} - \mu}{4/\sqrt{n}} \leq \frac{1}{4/\sqrt{n}}\right) = P\left(\frac{-1}{4/\sqrt{n}} \leq Z \leq \frac{1}{4/\sqrt{n}}\right) = 0,95$$

Để trình bày xác suất này đúng, chúng ta biết rằng:

$$Z_{0,025} = -1,96 = \frac{-1}{4/\sqrt{n}} \quad \text{và} \quad Z_{0,975} = 1,96 = \frac{1}{4/\sqrt{n}}$$

Kết quả là $n = 61,46$, vì vậy chúng ta cần chọn một mẫu có 62 quan sát.

Ví dụ 3: Một lần nữa giả sử tổng thể tuân theo phân phối chuẩn, phương sai tổng thể được biết là $\sigma^2 = 16$, nhưng trung bình tổng thể cũ ng chưa xác định. Giả sử rằng chúng ta muốn xây dựng một khoảng tin cậy 95% đối với giá trị trung bình chưa biết μ với cỡ mẫu là 100.

Trong trường hợp này, chúng ta viết trình bày xác suất như sau:

$$P\left(Z_{0,025} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{0,975} \right) = P\left(-1,96 \leq \frac{\bar{X} - \mu}{0,4} \leq 1,96 \right) = 0,95$$

Nếu chúng ta biến đổi trình bày xác suất nêu trên nhằm đặt giá trị trung bình chưa biết ở giữa, với cách làm như vậy chúng ta có:

$$P\left(\bar{X} - 1,96 \times 0,4 \leq \mu \leq \bar{X} + 1,96 \times 0,4 \right) = 0,95$$

Khoảng ngẫu nhiên $\bar{X} \pm 1,96 \times 0,4$ là khoảng tin cậy 95% đối với giá trị trung bình chưa biết.

Lưu ý về giải thích: khi một mẫu cụ thể được rút ra và một khoảng tin cậy được tính toán, chúng ta có thể nói rằng có xác suất 95% để khoảng này chứa giá trị trung bình chưa biết. Khoảng này hoặc có chứa hoặc không chứa và chúng ta không biết liệu nó có chứa giá trị trung bình chưa biết hay không.

Nếu chúng ta muốn viết một trình bày xác suất, thì chúng ta có thể đưa ra một trình bày xác suất cho quá trình tính toán khoảng tin cậy này. Nếu chúng ta lặp lại qui trình này nhiều lần, thì 95% những khoảng như vậy, dựa trên các mẫu ngẫu nhiên có n quan sát được rút ra từ tổng thể cụ thể, sẽ chứa giá trị trung bình chưa biết này.

Lưu ý về ký hiệu: nhìn chung, chúng ta viết mức độ tin cậy là $(1 - \alpha)$, trong đó α là mức ý nghĩa (chúng ta định nghĩa mức ý nghĩa ngay sau đây). Với định nghĩa này trong đầu, chúng ta định ra khoảng tin cậy $(1 - \alpha)\%$ là:

$$\bar{X} \pm Z_{(1 - \alpha/2)} \times \frac{\sigma_x}{\sqrt{n}}$$

Tiếp theo, chúng ta thảo luận điều gì chúng ta có thể làm nếu không biết tổng thể có tuân theo phân phối chuẩn hay không, nhưng phương sai của tổng thể đã xác định.

Tổng thể $X \sim$ không có phân phối chuẩn, phương sai σ^2 đã cho, và cỡ mẫu lớn

Trong trường hợp này, chúng ta có thể dựa vào Thuyết Giới hạn Trung tâm (CLT) nếu cỡ mẫu đủ lớn. Tuy nhiên, chúng ta cần luôn nhận thức rằng các kết quả chỉ là gần đúng, và chúng ta có thể không biết sự gần đúng này tốt đến mức độ nào.

Giá trị trung bình và phương sai của phân phối chọn mẫu không phụ thuộc vào dạng của phân phối mẹ. Ngoài ra, nếu n đủ lớn thì phân phối chọn mẫu của \bar{X} được coi là xấp xỉ tuân theo phân phối chuẩn với giá trị kỳ vọng μ và phương sai $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$.

Ví dụ 4: Bây giờ giả sử rằng $X \sim (100, 16)$, nhưng không có phân phối xác suất chuẩn. Nếu chúng ta rút ra một mẫu có $n = 400$ quan sát, xác suất để trung bình mẫu lớn hơn hoặc bằng 100,2 là bao nhiêu?

$$\text{Đầu tiên hãy tính } \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{16}{400} = 0,04$$

Tiếp theo, viết trình bày xác suất:

$$P(100,2 \leq \bar{X}) = P\left(\frac{100,2 - \mu}{0,2} \leq \frac{\bar{X} - \mu}{0,2}\right) = P\left(\frac{100,2 - 100}{0,2} \leq Z\right) = P(1 \leq Z)$$

Đánh giá biểu thức trên cho ta xác suất gần đúng.

Ví dụ 5: Giả sử rằng ta không biết tổng thể tuân theo phân phối chuẩn, và trung bình tổng thể chưa biết. Tuy nhiên, ta giả sử rằng phương sai tổng thể là $\sigma^2 = 16$.

Cần có cỡ mẫu là bao nhiêu để bảo đảm rằng chênh lệch giữa trung bình mẫu và trung bình tổng thể nằm trong phạm vi 1 đơn vị với xác suất 95%?

Trong trường hợp này, chúng ta viết trình bày xác suất chính xác như trước đây:

$$P(-1 \leq \bar{X} - \mu \leq 1) = P\left(\frac{-1}{4/\sqrt{n}} \leq \frac{\bar{X} - \mu}{4/\sqrt{n}} \leq \frac{1}{4/\sqrt{n}}\right) = P\left(\frac{-1}{4/\sqrt{n}} \leq Z \leq \frac{1}{4/\sqrt{n}}\right) = 0,95$$

Nếu chúng ta giả định rằng phân phối chọn mẫu của \bar{X} là xấp xỉ chuẩn, thì trình bày xác suất này là gần đúng nếu:

$$Z_{0,025} = -1,96 = \frac{-1}{4/\sqrt{n}} \quad \text{và} \quad Z_{0,975} = 1,96 = \frac{1}{4/\sqrt{n}}$$

Kết quả cỡ mẫu tính ra là $n = 61,46$, từ đó chúng ta cần chọn một mẫu có 62 quan sát.

Liệu có phải $n = 62$ là một mẫu đủ lớn phụ thuộc vào độ lệch của phân phối mẹ so với phân phối chuẩn. Nếu phân phối mẹ này lệch ra khỏi phân phối chuẩn càng nhiều, thì $n = 62$ có thể chưa đủ lớn. Tuy nhiên, nếu phân phối này là đối xứng, thì $n = 62$ có lẽ là đủ lớn để cung cấp một xấp xỉ tốt cho kết quả mong muốn.

Nếu các kết quả tính toán của chúng ta cho ra một cỡ mẫu nhỏ, thì có lẽ chúng ta nên thận trọng về độ tin cậy của những kết quả có được.

Ví dụ 6: Một lần nữa giả định rằng không biết tổng thể có tuân theo phân phối chuẩn hay không, phương sai của tổng thể là $\sigma^2 = 16$, nhưng trung bình của tổng thể lại chưa biết. Giả sử rằng chúng ta muốn xây dựng một khoảng tin cậy 95% cho giá trị trung bình chưa biết μ , dựa trên một mẫu có 100 quan sát. Liệu cỡ mẫu này có đủ lớn và có phụ thuộc vào độ lệch của phân phối tổng thể so với phân phối chuẩn hay không, cụ thể là mức độ mất cân đối của phân phối tổng thể. Nếu mức độ mất cân đối không nhiều, thì khoảng tin cậy 95% xấp xỉ được tính toán bằng những tính toán y như đã cho trong Ví dụ 3.

Tổng thể $X \sim$ không có phân phối chuẩn, phương sai σ^2 đã biết, và cỡ mẫu nhỏ

Trong trường hợp này chúng ta KHÔNG thể dựa vào Thuyết Giới hạn Trung tâm (CLT). Nếu phân phối mẹ đã biết, thì có thể tìm được phân phối chọn mẫu của trung bình mẫu này. Việc này đòi hỏi các kỹ thuật ngoài phạm vi khóa học ngắn hạn này.

Phân phối chọn mẫu của s^2

Trong mỗi ví dụ vừa nêu ở trên, chúng ta đều đã giả định rằng giá trị của phương sai tổng thể σ^2 đã biết. Giả định này không thực tế, nên chúng ta phải ước tính phương sai tổng thể dựa trên dữ liệu mẫu.

Ước lượng mà chúng ta sử dụng là

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

Ước lượng này là ước lượng không chệch đối với σ^2 : $E[s^2] = \sigma^2$

Để chứng minh điều này, hãy sử dụng chiến lược sau:

$$E\left[\sum (X_i - \bar{X})^2\right] = E\left[(X_i - \mu) - (\bar{X} - \mu)\right]^2$$

Ngoài ra, chúng ta có thể chứng minh rằng:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Để chứng minh điều này, hãy bắt đầu với chiến lược y như đã cho trên đây, sau đó sử dụng định nghĩa chuẩn hoá và định nghĩa của phân phối Khi bình phương.

Để đạt tới tính nhất quán (consistency) của s^2 chúng ta lưu ý rằng phương sai của phân phối Khi bình phương với $(n-1)$ bậc tự do là $2(n-1)$.

Lấy biểu thức này: $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ và nhân toàn bộ với $\frac{\sigma^2}{(n-1)}$. Bây giờ về phải là một biến ngẫu nhiên có phương sai $\left(\frac{\sigma^2}{(n-1)}\right)^2 (n-1) = \frac{\sigma^4}{n-1}$.

Do đó, phương sai của s^2 nhỏ dần khi cỡ mẫu tiến tới vô cùng. Kết hợp điều này với thực tế s^2 là ước lượng không chệch, chúng ta kết luận rằng s^2 là một ước lượng nhất quán của σ^2 . Kết luận này sẽ có ích cho chúng ta về sau.

Một ứng dụng phổ biến của phân phối chọn mẫu cho phương sai mẫu là xây dựng các khoảng tin cậy cho phương sai tổng thể.

Hãy bắt đầu bằng cách viết trình bày xác suất:

$$P\left(\chi^2_{(n-1, \alpha/2)} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{(n-1, 1-\alpha/2)}\right) = 1-\alpha$$

Anh/Chị sẽ có khả năng làm một số ví dụ bằng số vào cuối tuần với Bài tập 5. Hãy ghi nhớ rằng phân phối Khi bình phương là không đối xứng!

Thống kê t-Student

Chúng ta có thể kết hợp các phân phối chọn mẫu của trung bình mẫu và phương sai mẫu để giải quyết vấn đề viết trình bày xác suất cho trung bình mẫu này khi phương sai tổng thể chưa biết.

Các giả định: $X \sim N(\mu, \sigma^2)$ và cả μ lẫn σ^2 đều chưa biết

Thực tế : trong các giả định đã cho, \bar{X} và s^2 là độc lập thống kê.

Hãy nhớ lại định nghĩa của phân phối t- student :

$$\frac{Z}{\sqrt{\chi_v^2/v}} \sim t_{(v)}$$

trong đó các biến ngẫu nhiên có phân phối chuẩn chuẩn hoá và phân phối Khi bình phương là độc lập thống kê.

Bây giờ chú ý hai điều sau đây:

$$\frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} = Z \sim N(0, 1)$$

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

Đưa những biểu thức này và định nghĩa của trị thống kê t và thực hiện các thủ tục rút gọn đơn giản, và chúng ta có kết quả sau :

$$\frac{(\bar{X} - \mu)}{s/\sqrt{n}} = T\text{-stat} \sim t_{(n-1)}$$

Bây giờ chúng ta có thể tính các khoảng tin cậy đối với giá trị trung bình chưa biết khi cả giá trị trung bình lẫn phương sai đều chưa biết.

Bắt đầu bằng cách viết trình bày xác suất :

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

Sau đó biến đổi để đưa μ chưa biết và giữ α biểu thức như sau:

$$P\left(\bar{X} + t_{(v, \alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right) \leq \mu \leq \bar{X} + t_{(v, 1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right) \right) = 1 - \alpha$$

Khoảng ngẫu nhiên có được là

$$\bar{X} \pm t_{(v, 1-\alpha/2)} \left(\frac{\sigma}{\sqrt{n}} \right)$$

là khoảng tin cậy $(1-\alpha)\%$ đối với giá trị trung bình chưa biết .

Nhớ rằng, đối với một mức tin cậy cho trước, những khoảng dựa trên phân phối t lớn hơn những khoảng dựa trên phân phối chuẩn chuẩn hoá. Điều này phản ánh thực tế rằng chúng ta đang làm việc với ít thông tin hơn: chúng ta không biết giá trị của phương sai tổng thể. Khi đó chúng ta phải ước lượng phương sai này, điều này là làm tăng thêm tính ngẫu nhiên trong tính toán của chúng ta, và nó được phản ánh bằng khoảng tin cậy dài hơn .

Nhớ rằng trong phần giới thiệu của bài giảng này, chúng ta đã giả định rằng phân phối mẹ là chuẩn: X tuân theo phân phối chuẩn. Nếu giả định này không đáng tin cậy, thì chúng ta không thể sử dụng phân phối t -Student; chúng ta lại phải nhờ sự trợ giúp của Thuyết Giới hạn Trung tâm .

Tổng thể $X \sim$ không có phân phối chuẩn , phương sai σ^2 chưa biết, và cỡ mẫu lớn

Một lần nữa, hãy lưu ý mẫu số trong định nghĩa của biến ngẫu nhiên t : nó chứa phương sai mẫu. Do ước lượng này là nhất quán (consistent), khi cỡ mẫu tiến tới vô hạn, mẫu số hội tụ tới phương sai tổng thể và chúng ta còn lại biến ngẫu nhiên chuẩn chuẩn hoá này trong tử số .

Theo lô gic này, khi phân phối tổng thể và phương sai của tổng thể chưa biết, và cỡ mẫu lớn, thì trị thống kê

$$\frac{\bar{X} - \mu}{s / \sqrt{n}}$$

xấp xỉ tuân theo phân phối chuẩn chuẩn hoá.

