95% khoảng tin cậy & giá trị p

Nguyễn Quang Vinh – Nguyễn Thị Từ Vân

Giới thiệu

Thống kê

- Mô tả:
  - Độ tập trung
  - Độ phân tán
- Suy lý / Suy luận / Suy rộng:
  - Ước lượng
  - Kiểm định giả thuyết

SỐ ĐO XU HƯỚNG TẬP TRUNG

Trung bình (trung bình đại số)

Trung bình mẫu : $\overline{x} = \frac{\sum x}{n}$

Trung bình tổng thể : $\mu = \frac{\sum x}{N}$

- Duy nhất
- Đơn giản
- Giá trị ngoại lai & trung bình (!)

Trung điểm

$$Mr = \frac{L + H}{2}$$

- Ít dùng hơn trung bình và trung vị
- Dễ tính toán
- Đơn giản
- Giá trị ngoại lai & *Mr* (!)

Trung vị (Md)

- Duy nhất
- Đơn giản
- Giá trị ngoại lai & trung vị (!)

Mode

- dùng để mô tả dữ liệu định tính

SỐ ĐO ĐỘ PHÂN TÁN

(*dispersion, variation, spread, scatter*)

1. Khoảng giá trị
2. Phương sai
3. Độ lệch chuẩn
4. Hệ số biến thiên

SỐ ĐO ĐỘ PHÂN TÁN

3. Độ lệch chuẩn

Độ lệch chuẩn từ mẫu, s2

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]}$$

Độ lệch chuẩn tổng thể, σ

$$\sigma = \sqrt{\frac{\sum (x - \mu)}{N}}$$

4. Hệ số biến thiên*

$$C.V. = \frac{s}{\bar{x}}.100$$

* for data sets with extreme variation, it is possible to obtain a $C.V. > 100\%$

SỐ ĐO ĐỘ PHÂN TÁN
(dispersion, variation, spread, scatter)

1. Khoảng giá trị: H - L

2. Phương sai

Phương sai mẫu, s2

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]$$

Phương sai tổng thể, σ2

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

PHÂN PHỐI MẪU

Phân phối xác suất của *trị số thống kê có từ mẫu nghiên cứu* được gọi là phân phối mẫu.

Sai số chuẩn

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

còn gọi là:

- Sai số chuẩn của trị số trung bình, hoặc
- Sai số chuẩn, hoặc
- Độ lệch chuẩn của các trị số trung bình từ các mẫu nghiên cứu

Ước lượng

Số ước lượng → Tham số

Các tham số:

- Trung bình tổng thể
- Tỷ lệ của tổng thể
- Phương sai của tổng thể
- Khác biệt giữa 2 trung bình
- Khác biệt giữa 2 tỷ lệ
- Tỷ số giữa 2 phương sai

Số ước lượng → Tham số

- Mỗi tham số:

  Ước lượng điểm

  Ước lượng khoảng

KHOẢNG TIN CẬY CỦA TRUNG BÌNH TỔNG THỂ

Ước lượng khoảng tin cậy có công thức chung:

estimator ± (reliability coefficient) x (standard error)

Thực tế, khi mẫu được chọn từ tổng thể có phân phối ***bình thường*** với ***phương sai biết trước***, ước lượng khoảng cho trung bình μ sẽ là:

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$

Cách diễn giải kết quả khoảng ước lượng theo công thức này

- *Nếu lấy mẫu lặp đi lặp lại càng nhiều lần, từ tổng thể có phân phối bình thường, 100(1 - α)% của tất cả các khoảng ước lượng tính theo công thức trên sẽ chứa trung bình của tổng thể μ*

- Con số 1 - α, gọi là *hệ số tin cậy*, &

  Khoảng $\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$, gọi là *khoảng tin cậy của μ*

Khi cỡ mẫu *lớn* $\rightarrow$ dùng *z*, và s là xấp xỉ của σ. Các giá trị sử dụng thường dùng cho

Cách diễn giải thực tế

- *Chúng ta tin cậy ở mức 100(1 - α)% là khoảng ước lượng tính được này*

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$

sẽ chứa trung bình của tổng thể, μ

- *Gọi E = biên sai số = sai số lớn nhất = sai số có thể chấp nhận được trong thực hành / lâm sàng:*

$$E = z_{\alpha/2}\sigma_{\bar{x}} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

PHÂN PHỐI CỦA TRUNG BÌNH MẪU $\overline{X}$

- Khi mẫu được chọn từ tổng thể có phân phối không bình thường:

Định lý giới hạn trung tâm:

*Một tổng thể cho trước với dạng phân phối bất kỳ có trung bình μ và phương sai hữu hạn σ^2; phân phối của trung bình mẫu $\overline{X}$, tính từ các mẫu có cũng cỡ mẫu n được rút ra từ tổng thể này, sẽ có phân phối xấp xỉ normal với trung bình μ, và phương sai σ^2/n khi cỡ mẫu **đủ lớn**.*

Cỡ mẫu đủ lớn bao nhiêu để có thể áp dụng định lý giới hạn trung tâm?

- Không có một câu trả lời, bởi vì cỡ mẫu cần lấy phụ thuộc vào mức độ phân phối không bình thường hiện hữu trong tổng thể.
- Quy tắc chung: trong thực tế ở hầu hết các tình huống, cỡ mẫu từ 30 trở lên là đủ lớn.
- Nói chung, việc xấp xỉ phân phối bình thường sẽ càng tốt hơn khi tăng cỡ mẫu lên.

CHỌN MẪU TỪ TỔNG THỂ CÓ PHÂN PHỐI NONNORMAL

→ Việc chọn mẫu từ:

- tổng thể có phân phối nonnormal
- tổng thể có hình dạng không biết trước

→ Lấy cỡ mẫu đủ lớn→ *áp dụng định lý giới hạn trung tâm*

KHOẢNG TIN CẬY CHO KHÁC BiỆT GiỮA TRUNG BÌNH 2 MẪU

Khi **biết phương sai của hai tổng thể**, 100(1 - α)% khoảng tin cậy của μ_1 - μ_2 là:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sigma_{\bar{x}_1 - \bar{x}_2}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Nếu chọn mẫu từ tổng thể có phân phối nonnormal: lấy cở mẫu n_1, n_2 *đủ lớn*→ *áp dụng định lý giới hạn trung tâm*

KHOẢNG TIN CẬY CHO KHÁC BiỆT GiỮA TRUNG BÌNH 2 MẪU

Khi không biết phương sai của hai tổng thể, cần phân biệt hai tình huống:

<u>(1) Phương sai của hai tổng thể không khác nhau</u>

- Nếu giả định này thỏa mãn, công thức của phương sai gộp (*pooled estimate)* là:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- 100(1 - α)% khoảng tin cậy của μ_1 - μ_2 là:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

KHOẢNG TIN CẬY CHO KHÁC BiỆT GiỮA TRUNG BÌNH 2 MẪU

(2) Phương sai của hai tổng thể khác nhau

- Khi thỏa điều kiện này, 100(1 - α)% khoảng tin cậy của $\mu_1 - \mu_2$ là

$$(\bar{x}_1 - \bar{x}_2) \pm t'_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t'_{\alpha/2} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

$$w_1 = \frac{s_1^2}{n_1}$$

$$w_2 = \frac{s_2^2}{n_2}$$

$$t_1 = t_{\alpha/2, n_1 - 1}$$

$$t_2 = t_{\alpha/2, n_2 - 1}$$

$t'_{\alpha/2}$ gọi là hệ số tin cậy Cochran

KHOẢNG TIN CẬY CỦA TỶ LỆ TỔNG THỂ

- Tỷ lệ của mẫu, ký hiệu $\hat{p}$ được dùng như là số ước lượng điểm của tỷ lệ của tổng thể, ký hiệu *p*, khi đó khoảng tin cậy theo công thức chung:

  estimator ± (reliability coefficient) x (standard error)

- Khi *np & n(1-p)* đều lớn hơn 5, sampling distribution của tỷ lệ mẫu $\hat{p}$ có phân phối bình thường.

  vì thế, hệ số tin cậy là giá trị z tính từ phân phối bình thường chuẩn.

- Sai số chuẩn là: $$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

  Vì không biết *p* , ta phải dùng $\hat{p}$ để ước lượng. Vì thế σ ước lượng bởi

$$\sigma_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$$

KHOẢNG TIN CẬY CỦA TỶ LỆ TỔNG THỂ

- 100 (1 - α)% khoảng tin cậy của p:

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}}$$

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- Vì thế, 95% khoảng tin cậy của *p* là

$$\hat{p} \pm 1.96\sigma_{\hat{p}}$$

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

KHOẢNG TIN CẬY CỦA KHÁC BIỆT TỶ LỆ CỦA HAI TỔNG THỂ

$$(\hat{p}_1 - \hat{p}_2) \rightarrow (p_1 - p_2)$$

Khi: cỡ mẫu n_1 & n_2 lớn & các tỷ lệ tổng thể, p_1 - p_2, không gần 0 hoặc 1

→ áp dụng định lý giới hạn trung tâm & lý thuyết phân phối normal để xác định khoảng tin cậy

$$S.E. = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

100 (1 - α)% khoảng tin cậy của p_1 - p_2

Ghi chú

* Thông thường không biết phương sai $\sigma^2 \rightarrow$ cần phải ước lượng σ^2

* Việc ước lượng σ^2 từ các nguồn sau đây:

1. Mẫu nghiên cứu thử
2. Kết quả nghiên cứu trước hoặc tương tự
3. $\sigma \approx$ R/4 (hoặc R/6) (phân phối xấp xỉ normal & biết giá trị nhỏ nhất và giá trị lớn nhất của biến số trong tổng thể)
4. $s \approx$ IQR/1.35

KHOẢNG TIN CẬY CỦA PHƯƠNG SAI TỔNG THỂ CÓ PHÂN PHỐI NORMAL

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

100 - α)% KTC của *(n-1)s²/σ²*

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

(100 - α)% KTC của σ^2

$$\Leftrightarrow \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$$

(100 - α)% KTC của σ

KHOẢNG TIN CẬY CỦA PHƯƠNG SAI TỔNG THỂ CÓ PHÂN PHỐI NORMAL

Nhược điểm

Mặc dù phương pháp này để xác định khoảng tin cậy của σ^2 được dùng rộng rãi, nhưng không phải không có một số nhược điểm:

- Giả định phân phối của tổng thể mà mẫu được rút ra là *normal*, đây là điều quan trọng.
- Số ước lượng không nằm giữa khoảng tin cậy bởi vì phân phối χ^2 không đối xứng như phân phối normal.

KHOẢNG TIN CẬY CỦA PHƯƠNG SAI TỔNG THỂ CÓ PHÂN PHỐI NORMAL

Nếu cỡ mẫu đủ lớn:

$$s - z_{1-\alpha/2}\frac{s}{\sqrt{2n}} < \sigma < s + z_{1-\alpha/2}\frac{s}{\sqrt{2n}}$$

(100 - α)% khoảng tin cậy của σ

KHOẢNG TIN CẬY CỦA TỶ SỐ HAI PHƯƠNG SAI CỦA HAI TỔNG THỂ CÓ PHÂN PHỐI NORMAL

$$\frac{s_1^2}{\sigma_1^2} \Big/ \frac{s_2^2}{\sigma_2^2}$$

tuân theo phân phối ***F*** *.*

Với giả định: $s_1^2 \,\&\, s_2^2$

được tính từ hai mẫu độc lập n_1 & n_2, lần lượt, được rút ra từ hai tổng thể có phân phối bình thường

KHOẢNG TIN CẬY CỦA TỶ SỐ HAI PHƯƠNG SAI CỦA HAI TỔNG THỂ CÓ PHÂN PHỐI NORMAL

$$F_{\alpha/2} < \frac{s_1^2}{\sigma_1^2} \Big/ \frac{s_2^2}{\sigma_2^2} < F_{1-\alpha/2}$$

(100 - α)% KTC của

$$\frac{s_1^2}{\sigma_1^2} \Big/ \frac{s_2^2}{\sigma_2^2}$$

$$\Leftrightarrow \frac{s_1^2 / s_2^2}{F_{1-\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2 / s_2^2}{F_{\alpha/2}}$$

(100 - α)% KTC của

$$\sigma_1^2 / \sigma_2^2$$

Ghi chú

$$F_{(1-\alpha),\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}$$

Trong đó:

$\nu_1 = n_1 - 1$ *(độ tự do của tử số)*

$\nu_2 = n_2 - 1$ *(độ tự do của mẫu số)*

Kiểm định giả thuyết thống kê

Giới thiệu

- *Đưa ra một quyết định* về một tổng thể bằng cách khảo sát một mẫu lấy ra từ tổng thể đó.
- Hai loại giả thuyết:

(1) *Giả thuyết nghiên cứu:*
- một phán đoán hoặc một khả năng
- có thể là kết quả của nhiều năm quan sát
- dẫn trực tiếp đến giả thuyết thống kê.

(2) *Giả thuyết thống kê:*
giả thuyết được phát biểu theo một cách thức mà có thể đánh giá bằng các kỹ thuật kiểm định thống kê phù hợp

<table>
<tr><td colspan="2" rowspan="2">Các trường hợp mắc sai lầm loại I & sai lầm loại II (4 khả năng)</td><td colspan="2">Sự thật trong tổng thể</td></tr>
<tr><td>Có mối liên hệ giữa tiếp xúc & kết cục
(H_o sai)</td><td>Không có mối liên hệ giữa tiếp xúc & kết cục
(H_o đúng)</td></tr>
<tr><td rowspan="2">Kết quả của nghiên cứu từ mẫu → Kết luận:</td><td>Bác bỏ H_o</td><td>Quyết định đúng</td><td>Sai lầm loại I</td></tr>
<tr><td>Không bác bỏ H_o</td><td>Sai lầm loại II</td><td>Quyết định đúng</td></tr>
</table>

Lưu ý

- H_o, H_A & α *phải được xác lập trước khi nhìn vào dữ liệu.* Nói cách khác, *không để dữ liệu dẫn đường giả thuyết*
- *Giá trị* α càng nhỏ, *giá trị* β càng lớn → *nếu* muốn β nhỏ, cần chọn giá trị α lớn
- *Trong hầu hết mọi trường hợp* khoảng giá trị chấp nhận của α thay đổi từ 0.01 đến 0.1
- *Nếu* không có khác biệt đáng kể giữa hiệu lực của sai lầm loại I so với sai lầm loại II, nên chọn α= .05

Năm bước trong kiểm định giả thuyết thống kê

Bước 1: Đặc tính của dữ liệu - K*iểm tra các giả định - Xác lập* H$_o$ - H$_A$

Bước 2: *xác định test thống kê, và phân phối*

Bước 3: *xác định vùng bác bỏ*: đưa ra giá tri α

Bước 4: *tính giá trị của test thống kê và kiểm định với phân phối tương ứng* → ***giá trị p***

Phát biểu quyết định thống kê: bác bỏ H$_o$ *hoặc không bác bỏ* H$_o$

Bước 5: *đưa ra kết luận – không có* thuật ngữ thống kê.

- Nếu H_0 bị bác bỏ, chúng ta kết luận là H_A đúng.
- Nếu không bác bỏ được H_0, chúng ta kết luận có lẽ H_0 đúng.

- Tránh dùng "chấp nhận" trong trường hợp không bác bỏ được H_o, khi đó ta nên phát biểu là H_o "không được bác bỏ"

Tóm tắt

(1) Nói chung, kiểm định giả thuyết không đi chứng minh một giả thuyết - Chỉ đơn giản là dữ liệu hiện trong tay có ủng hộ hay không ủng hộ cho giả thuyết không.

(2) Điều mà chúng ta mong chờ có thể để kết luận là kết quả của kiểm định thường nên để trong giả thuyết H_A

(3) Giả thuyết H_O là giả thuyết được kiểm định

(4) Hai giả thuyết H_O & H_A là bổ sung của nhau