B7. Inferential Statistics

Bổ sung

2019

Nội dung bổ sung

1. **Inferential Statistics**
2. **Standard Error**
3. **Confidence Intervals**
4. **Correlation**
5. **Hypothesis testing**

1. Inferential Statistics

- **Thống kê suy luận**
  - suy luận về tổng thể/quần thể (*population*) dựa trên mẫu (*sample*) chứa các quan sát (*observations*)
  - lấy mẫu (*sampling*) → *sampling error*: không thể hiện đặc trưng của tổng thể

> *"You don't have to eat the whole ox to know that the meat is tough."*
> [Samuel Johnson]

1. Inferential Statistics (tt.)

- **Lấy mẫu ngẫu nhiên đơn giản** (*simple random sampling*)

  A) Từ một tổng thể hữu hạn:
  - ở mỗi bước, các phần tử có cùng xs được chọn
  - hoàn lại hay không hoàn lại (sampling w./without replacement)
  - số lượng mẫu $\frac{N!}{n!(N-n)!}$: mỗi mẫu có cùng xs được chọn

  B) Từ một tổng thể vô hạn (vô cùng lớn): các phần tử được chọn một cách độc lập

1. Inferential Statistics (tt.)

- **Lấy mẫu ngẫu nhiên phân tầng** *(stratified random sampling)*
  - lấy mẫu theo xác suất (phương sai tương đối nhỏ)
  - tổng thể được phân chia thành nhiều tầng (cấu trúc phân cấp, tập hợp những phần tử "tương đồng")
  - một mẫu ngẫu nhiên đơn giản được lấy theo từng tầng

1. Inferential Statistics (tt.)

- **Lấy mẫu theo cụm** *(cluster sampling)*
  - lấy mẫu theo xác suất
  - tổng thể được chia thành nhiều cụm, mỗi cụm là một đại diện thu nhỏ của tổng thể (VD: khu vực địa lý)
  - một mẫu ngẫu nhiên đơn giản được lấy từ theo từng cụm

1. Inferential Statistics (tt.)

- **Lấy mẫu hệ thống** (*systematic sampling*)
  - lấy mẫu theo xác suất
  - phân tầng theo tỷ lệ
  - chọn ngẫu nhiên 1 trong k phần tử

  VD: lấy cỡ mẫu 50 phần tử từ tổng thể 5000 phần tử
  → lần lượt chọn 1 trong số 100 phần tử của hệ thống

1. Inferential Statistics (tt.)

- **Lấy mẫu thuận tiện** (*convenience sampling*)
  - lấy mẫu PHI xác suất
  - lấy mẫu dựa trên sự thuận tiện

  VD: lấy mẫu từ các sinh viên, những người tình nguyện, …

- **Lấy mẫu phán đoán** (*judgment sampling*)
  - lấy mẫu PHI xác suất
  - lấy mẫu dựa trên ý kiến phán đoán, đánh giá của chuyên gia

1. Inferential Statistics (tt.)

- **Sự thiên lệch (*bias*): mẫu không đại diện (đúng) cho tổng thể**
  - *convenience bias*: thiên lệch do chú trọng tính thuận lợi
  - *judgement bias*: thiên lệch do ý kiến phán đoán, đánh giá
  - *size bias*: cỡ mẫu quá nhỏ không chứa đủ các phần tử đại diện

1. Inferential Statistics (tt.)

- **Tham số và đặc trưng**
  - ước lượng giá trị tham số của tổng thể: μ, σ, …
  - tính toán đặc trưng của mẫu quan sát (*thống kê mẫu*): $\bar{x}$, s, …
- **Lấy mẫu N lần, mỗi lần n đối tượng (quan sát)**
  - các biến ngẫu nhiên X_1, X_2, …, X_n

1. Inferential Statistics (tt.)

- **Nội suy** (*interpolation*)
  - ước lượng các điểm dữ liệu mới TRONG phạm vi tập dữ liệu đã quan sát được

    $\{ (x_i, y_i) \} \Rightarrow (x, y)$: $x \in (\min(x_i), \max(x_i))$
  - suy luận dựa vào bản chất của hiện tượng
  - đường cong (thường là đa thức) nội suy các điểm đã quan sát

1. Inferential Statistics (tt.)

- **Ngoại suy** (*extrapolation*)
  - ước lượng các điểm dữ liệu mới NGOÀI phạm vi tập dữ liệu đã quan sát được → dựa vào mối quan hệ với các biến độc lập

    $\{ (x_i, y_i) \} \Rightarrow (x, y)$: $x \notin (\min(x_i), \max(x_i))$
  - quan sát sự biến động của hiện tượng → rút ra những quy luật → dự đoán sự phát triển của hiện tượng

Nội dung bổ sung

1. Inferential Statistics
2. **Standard Error**
3. Confidence Intervals
4. Correlation
5. Hypothesis testing

2. Standard Error

- **Bài toán xác suất**
  - tổng thể có r phần tử
  - thí nghiệm: chọn n (n << r) phần tử của tổng thể (lấy mẫu)
- **Trung bình mẫu, độ lệch chuẩn của mẫu: các biến ngẫu nhiên**

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad SEM = \frac{s}{\sqrt{n}}$$

  - có phân phối xác suất $\rightarrow \mu_{\bar{X}}, \sigma_{\bar{X}}$
  - $\bar{X}, s$: các *đại lượng ước lượng điểm* của μ, σ

2. Standard Error (tt.)

- **Phân phối mẫu** (*sampling distribution*): **phân phối xác suất của thống kê mẫu (các giá trị kết quả tính toán từ N lần lấy mẫu)**
  - phân phối của *trung bình mẫu* → ước lượng μ (của tổng thể)
  - phân phối của *phương sai mẫu* → ước lượng σ^2 (của tổng thể)

$$
\begin{array}{lcccc}
S_1 = \{x_{11}, \ldots, x_{1j}, \ldots, x_{1n}\} & \rightarrow & \bar{x}_1 & s_1^2 \\
\vdots & & \vdots & \vdots \\
S_i = \{x_{i1}, \ldots, x_{ij}, \ldots, x_{in}\} & \rightarrow & \bar{x}_i & s_i^2 \\
\vdots & & \vdots & \vdots \\
S_N = \{x_{N1}, \ldots, x_{Nj}, \ldots, x_{Nn}\} & \rightarrow & \bar{x}_N & s_N^2 \\
 & & \downarrow & \downarrow \\
 & & \mu & \sigma^2
\end{array}
$$

2. Standard Error (tt.)

- **Phân phối mẫu** (*sampling distribution*)
  - $X_1, X_2, \ldots, X_n$: *independent and identically distributed* (*i.i.d.*)
  - $X_1, X_2, \ldots, X_n$: cùng kỳ vọng và phương sai

$$\mu_{\bar{X}} = E\left[\bar{X}\right] = E[X] = \mu_X$$

$$\sigma_{\bar{X}}^2 = Var\left(\bar{X}\right) = \frac{Var(X)}{n} = \frac{\sigma_X^2}{n}$$

2. Standard Error (tt.)

- **Sai số chuẩn** (*Standard Error – SE*)
  - *standard deviation of the means*: thể hiện sự thay đổi của mean trong các lần lấy mẫu

$$s_{\bar{X}} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad SEM = \frac{s_{\bar{X}}}{\sqrt{n}}$$

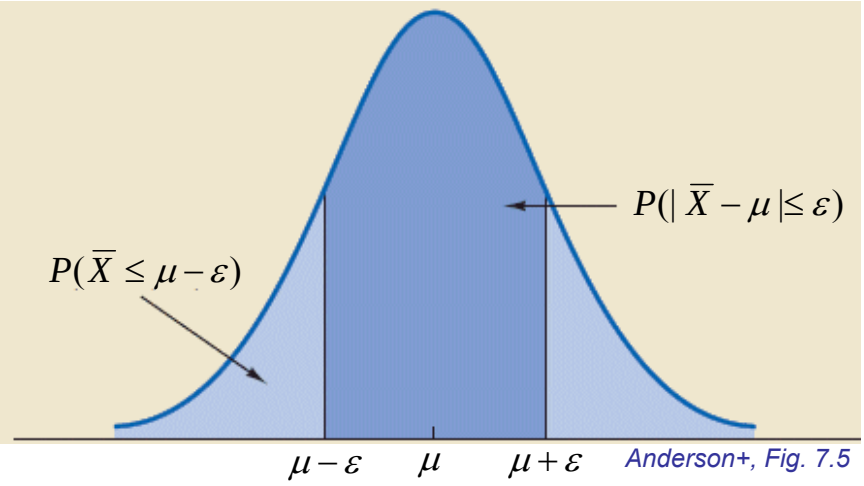
  - mức độ trung bình mẫu cách xa trung bình tổng thể σ
  - *đại lượng ước lượng điểm* của độ lệch chuẩn tổng thể σ
  - được dùng để ước lượng khoảng tin cậy (*Confidence Interval*)

2. Standard Error (tt.)

- **Tổng thể có phân phối chuẩn** $\Rightarrow \bar{X} \sim N(\mu, \sigma) \quad \forall n$
- **Tổng thể KHÔNG có phân phối chuẩn → áp dụng CLT**
  - nếu kích thước mẫu n đủ lớn thì trung bình mẫu gần xấp xỉ với phân phối chuẩn (phân bố 'xung quanh' μ)
  - mean của trung bình mẫu → μ

2. Standard Error (tt.)

- Sau một lần lấy mẫu, xác suất để trung bình mẫu sai lệch so với μ không vượt quá ε là bao nhiêu ?



Anderson+, Fig. 7.5

2. Standard Error (tt.)

- Các tham số chưa biết của 1 tổng thể
  - điểm (tốt nghiệp) trung bình (*mean* μ) của sinh viên trường A ?
  - tỉ lệ (*proportion* p) sinh viên trường A hút thuốc lá ?

2. Standard Error (tt.)

- **Ước lượng điểm** (*point estimate*)

  Tổng thể có tham số θ cần ước lượng (tìm giá trị $\Theta \approx \theta$)

  Không gian tham số (*parameter space*) Ω: các giá trị có thể của θ

  Các biến ngẫu nhiên: $X_1, X_2, \ldots, X_n$

  Mẫu $\{x_1, x_2, \ldots, x_n\}$: các g.trị quan sát tương ứng với $X_1, X_2, \ldots, X_n$

  Hàm ước lượng (*estimator*): $h(X_1, X_2, \ldots, X_n)$

  Ước lượng điểm là giá trị kết quả từ mẫu (thống kê mẫu):

  $$\Theta = h(x_1, x_2, \ldots, x_n) \in \Omega$$

2. Standard Error (tt.)

- **Ước lượng điểm** (*point estimator*)

  VD: $\Omega_{GPA} = \{ \mu \mid 0 \le \mu \le 10 \}$

  μ estimator: $h(X_1, X_2, \cdots, X_n) = \overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$

  μ point estimate (dựa trên mẫu): $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$

  σ^2 estimator: $h(X_1, X_2, \cdots, X_n) = S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

  σ^2 point estimate (dựa trên mẫu): $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  Sai số chuẩn = độ lệch chuẩn của ước lượng điểm

2. Standard Error (tt.)

- **Ước lượng điểm** (*point estimator*)
  - Tồn tại vô số khả năng chọn lựa estimator h
  - h tốt nhất → cho giá trị kết quả Θ xấp xỉ giá trị thật của θ
  - so sánh h_1 và h_2 ?

2. Standard Error (tt.)

- **Maximum Likelihood Estimation** (MLE)

Các biến ngẫu nhiên: $X_1, X_2, \ldots, X_n$ từ 1 phân phối:

Bộ tham số: $(\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$

Hàm phân phối PDF: $f(x_i;\theta_1,\theta_2,\cdots,\theta_n)$

Mẫu quan sát: $x=(x_1,x_2,\cdots,x_n)$

Likelihood function: $L(\theta_1,\theta_2,\cdots,\theta_n)=\prod_{i=1}^{n} f(x_i;\theta_1,\theta_2,\cdots,\theta_n)$

Maximum Likelihood Estimate: $\hat{\theta}=(h_1(x),h_2(x),\cdots,h_m(x))$

là điểm cực đại của hàm $L(\theta_1,\theta_2,\cdots,\theta_n)$

2. Standard Error (tt.)

- **Maximum Likelihood Estimation** (MLE)

  VD: Cân nặng của phụ nữ Mỹ ~ N(μ, σ)

  Lấy mẫu 10 phụ nữ có cân nặng (lbs):

  { 115, 122, 130, 127, 149, 160, 152, 138, 149, 180 }

$$f(x_i;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$L(\mu,\sigma)=\sigma^{-n}(2\pi)^{-n/2}e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}}$$

$$\hat{\mu}=\sum_{i=1}^{n}x_i=142.2$$

2. Standard Error (tt.)

- **Ước lượng điểm** (*point estimator*)
  - *không chệch* (*unbiased*): kỳ vọng của các mẫu = θ

    $E[\, h(X_1, X_2, \ldots, X_n)\,] = \theta$
  - *vững chắc* (*consistent*): cỡ mẫu n càng lớn thì ước lượng Θ càng chính xác
  - Hiệu quả (*most efficient*): unbiased, consistent, phương sai thấp nhấp (ít thay đổi theo các mẫu khác nhau)

2. Standard Error (tt.)

- **Ước lượng điểm** (*point estimator*)
  - có thật sự $\Theta \approx \theta$?
  - mức độ xấp xỉ giữa Θ và θ ?
  - 'khoảng' (đoạn) [L, U] chứa giá trị của tham số θ ?
  - mức độ tin cậy của khoảng [L, U] ?

Nội dung bổ sung

1. Inferential Statistics
2. Standard Error
3. **Confidence Intervals**
4. Correlation
5. Hypothesis testing

3. Confidence Intervals

- **'Khoảng' (đoạn) tin cậy** (*Confidence Interval for One Mean – CI*)

$\theta \in [L, U]$

CI = ước lượng điểm $\pm$ sai số biên (*margin of error*)

Mỗi CI có 1 *hệ số tin cậy* (*confidence coefficient* hay *proportion*), ký hiệu: $(1 - \alpha)$

hay mức độ tin cậy (*confidence level*), ký hiệu: $(1 - \alpha)100\%$

Các giá trị thông dụng của hệ số tin cậy (mức độ tin cậy):

0.90 (90%), 0.95 (95%), 0.99 (99%)

VD: Với mức tin cậy 95%, chiều cao trung bình của SV trường A nằm trong khoảng từ 158cm đến 165cm

3. Confidence Intervals (tt.)

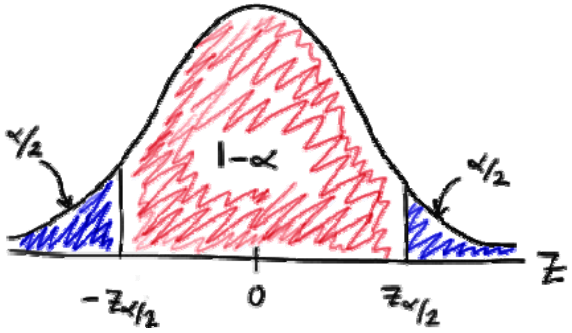
- **Z-interval của μ: một CI đặc biệt**

$z_{\alpha/2}$: Z-value tạo vùng bên *phải*, dưới đường phân phối có diện tích = $\alpha/2$

$$P(z_{\alpha/2} \leq Z) = \alpha/2$$

$-z_{\alpha/2}$: Z-value tạo vùng bên *trái*, dưới đường phân phối có diện tích = $\alpha/2$

$$P(Z \leq -z_{\alpha/2}) = \alpha/2$$



3. Confidence Intervals (tt.)

- **Z-interval của μ: một CI đặc biệt**

Giả sử: $X_1, X_2, \ldots, X_n$ ~ phân phối chuẩn

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{(\bar{X} - \mu)}{\sigma / \sqrt{n}} \sim N(0,1)$$

Nếu biết trước σ của tổng thể thì với mức độ tin cậy $(1 - \alpha)100\%$, khoảng tin cậy Z-interval của μ:

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Nếu $X_1, X_2, \ldots, X_n$ không phải phân phối chuẩn → n đủ lớn + CLT

3. Confidence Intervals (tt.)

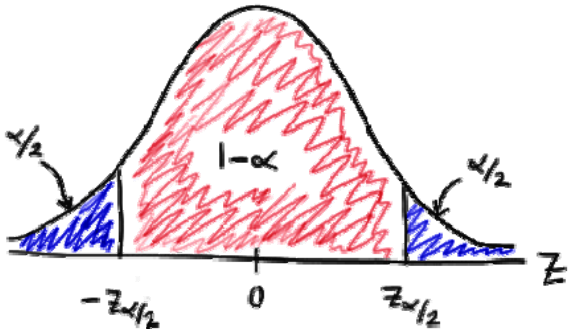
- **Z-interval của μ: một CI đặc biệt**

VD: Mức độ tin cậy 90% (0.90) ⇒ $\alpha = 0.10 \Rightarrow \alpha / 2 = 0.05$

Giá trị trong bảng Z: $P(Z \le \alpha)$

Giá trị cần có: $P(-\alpha/2 \le Z \le \alpha/2)$

Tra bảng Z với giá trị: $(0.90 + 0.05) = 0.95 \Rightarrow z_{\alpha/2} \approx 1.645$



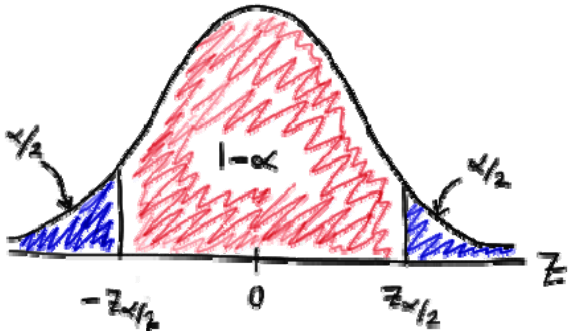
3. Confidence Intervals (tt.)

- **Z-interval của μ: một CI đặc biệt**

VD: Mức độ tin cậy 95% (0.95) $\Rightarrow \alpha / 2 = 0.025 \Rightarrow z_{\alpha/2} = 1.96$

Mức độ tin cậy 99% (0.99) $\Rightarrow \alpha / 2 = 0.005 \Rightarrow z_{\alpha/2} \approx 2.576$

Mức độ tin cậy 99.5% (0.995) $\Rightarrow \alpha / 2 = 0.0025 \Rightarrow z_{\alpha/2} = 2.81$



3. Confidence Intervals (tt.)

- **Z-interval của μ: một CI đặc biệt**

Độ dài Z-interval của μ: $d = 2z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

- σ đã cố định → không thể hiệu chỉnh để tăng, giảm d
- n tăng → d giảm
- mức độ tin cậy $z_{\alpha/2}$ giảm → d giảm

3. Confidence Intervals (tt.)

- **t-interval của μ**

Nếu không biết trước σ: ước lượng σ dựa trên phương sai mẫu s theo phân phối t

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

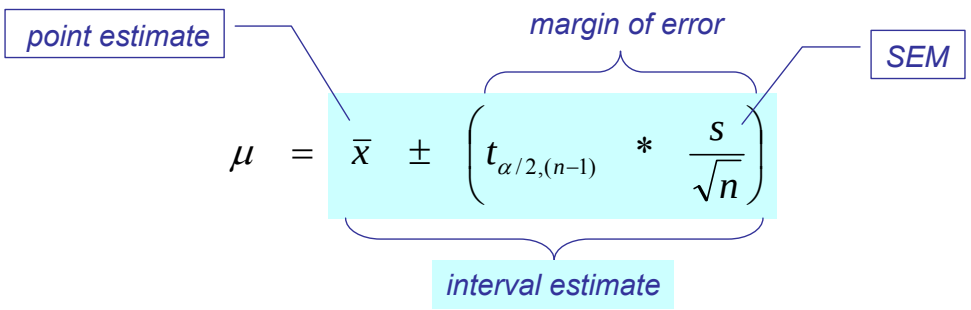
Ta có: $T = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$ có phân phối t với **(n – 1)** bậc tự do

Với mức độ tin cậy (1 – α)100%, khoảng tin cậy t-interval của μ:

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

3. Confidence Intervals (tt.)

- **t-interval của μ**



3. Confidence Intervals (tt.)

❑ **Trường hợp dữ liệu ban đầu không phải phân phối chuẩn**

$$T = \frac{(\bar{X} - \mu)}{s/\sqrt{n}}$$

- Khi n tăng: T ~ phân phối chuẩn bất chấp phân phối ban đầu
- Khi n đủ lớn: Z-interval và t-interval cho kết quả tương tự nhau

3. Confidence Intervals (tt.)

❑ **Xác định cỡ mẫu n khi biết phương sai tổng thể σ**

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Gọi ε là sai số biên mong muốn (chấp nhận được):

$$\varepsilon = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad n = \left(\frac{z_{\alpha/2}\sigma}{\varepsilon}\right)^2$$