Tóm tắt chương I: Giới thiệu tổng quan

- **population**
- **Sample**
- **descriptive statistic: (**Tables, graphs, or numerical summary tool, identification of patterns in the data, the population or sample of interest**)**
- **Statistical inference: (**draw conclusion about population parameters**)**
- **Collecting data** (có 3 cách: Retrospective; observation; experiment)
- **Categorise data** (có hai loại : qualitative and quantitative).

Tóm tắt chương II: **Các công thức tính xác suất**

- **Sample space = S** (Không gian mẫu là tập hợp các kết cục (outcome) có thể có khi thực hiện một phép thử).
- **Event:** là một tập con của sample space.
- $P(event) = \dfrac{|event|}{|S|}$ (nếu các outcome là đồng khả năng (**equally likely**))
- $P(event) = \sum P(outcome \in event)$.
- **Addition rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- Nếu $A \cap B = \phi$ thì A và B được gọi là xung khắc (**mutually exclusive**).
- Conditional probability: $P(A/B) = \dfrac{P(A \cap B)}{P(B)}$
- **Multiple rule**: $P(A \cap B) = P(A)P(B/A)$
- Nếu $P(A \cap B) = P(A)P(B)$ thì A và B được gọi là độc lập (**independence**).
- **Total probability**: Nếu $E_1, ..., E_n$ là hệ biến cố đầy đủ và xung khắc từng đôi thì $P(A) = \sum_i P(E_i)P(A/E_i)$.
- **Bayes theorem**: $P(E_k / A) = \dfrac{P(E_k)P(A/E_k)}{\sum_i P(E_i)P(A/E_i)}$

Tóm tắt chương III: **Biến ngẫu nhiên rời rạc (Discrete random variable)**

I) Kiến thức chung:

1) Hàm xác suất (**probabiliy mass function**): $f(x) = P(X = x)$.

   Nếu biết hàm xác suất f(x) thì có thể tính $P(a < X < b) = \sum_{a < x_i < b} f(x_i)$

2) Hàm phân phối tích lũy: (**Cumulative dist function**):

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$$

Nếu biết hàm phân phối tích lũy thì có thể tính được $P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$ hoặc làm theo cách khác là tìm hàm xác suất f(x) rồi dựa vào đó để tính.

3) Kì vọng (**mean or expected value**) và phương sai (**Variance**)

$$\mu = E(X) = \sum_i x_i f(x_i); V(X) = \sum_i x_i^2 f(x_i) - \mu^2; \sigma = \sqrt{V(X)}$$

Chú ý: Nếu cần tính $E(h(X)) = \sum_i h(x_i) f(x_i)$

II) Các biến ngẫu nhiên rời rạc thường gặp.

1) Biến ngẫu nhiên đều (**Uniform random variable**).

Nếu X nhận n giá trị $x_1, x_2, ..., x_n$ thì $P(X = x_i) = \frac{1}{n}$.

2) Biến ngẫu nhiên nhị thức (**Binomial**)

ĐN: Một thí nghiệm gồm n phép thử (**trial**)Bernoulli thỏa mãn *3* điều kiện:

+ Các PT độc lập (**independent**)

+ Mỗi PT chỉ có 2 kết quả ký hiệu là "thành công" (**success**) và "thất bại" (**failure**)

+ Xác suất của kết quả "thành công" trong một PT luôn không đổi và bằng *p*

Bnn X nhận giá trị <u>bằng số phép thử có kết quả là thành công</u> trong dãy n PT trên được gọi là bnn nhị thức với tham số n, p.

- **Hàm xác suất: $P(X = x) = f(x) = C^x_n\, p^x\, (1-p)^{n-x}$, $x = 0, 1,\ldots,n$**
- **Trung bình: $\mu = np$**
- **Phương sai: $\sigma^2 = np(1-p)$**

3) Phân phối **geometric** và **Negative Binomial**

ĐN: Gọi bnn X : **số phép thử cần thiết cho đến khi có 1 phép thử cho kết quả "thành công"** trong dãy n phép thử Bernoulli , X là bnn **Geometric**

Hàm xác suất: $P(X=x) = f(x) = (1-p)^{x-1}p$

Trung bình: $\mu = 1/p$

Phương sai: $\sigma^2 = (1-p)/p^2$

ĐN: Gọi bnn X : **số phép thử cần thiết cho đến khi có r phép thử cho kết quả "thành công"** trong dãy n phép thử Bernoulli , X là bnn **Negative Geometric**

Hàm xác suất: $P(X=x) = f(x) = C^{x-1}_{r-1}(1-p)^{x-r}\, p^r$

Trung bình: $\mu = r/p$

Phương sai: $\sigma^2 = r(1-p)/p^2$

4) Phân phối **Poisson**

ĐN: Bnn X chỉ **số 'event' xảy ra trong một 'interval'** được gọi là bnn Poisson.

Hàm xác suất: P(X=x) = f(x) = $e^{-\lambda}\lambda^{x}$/x! (λ là số trung bình các event trong interval đó**)**

Trung bình: $\mu = \lambda$

Phương sai: $\sigma^2 = \lambda$

Tóm tắt chương IV: **Biến ngẫu nhiên liên tục (Continuous random variable)**

I) Kiến thức chung:

1) Hàm mật độ (**probabiliy densty function**):

$$f(x): \int_{-\infty}^{+\infty} f(x)dx = 1; P(a < X < b) = \int_{a}^{b} f(x)dx.$$

Nếu biết hàm mật độ f(x) thì có thể tính $P(a < X < b) = \int_{a}^{b} f(x)dx$

2) Hàm phân phối tích lũy: (**Cumulative dist function**):

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$$

Nếu biết hàm phân phối tích lũy thì có thể tính được
$P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$

Nếu biết hàm phân phối tích lũy thì có thể tìm hàm mật độ f(x) = F'(x) .

3) Kì vọng (**mean or expected value**) và phương sai (**Variance**)

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx; V(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - \mu^2; \sigma = \sqrt{V(X)}$$

Chú ý: Nếu cần tính $E(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x)dx$

II) Các biến ngẫu nhiên liên tục thường gặp.

1) Biến ngẫu nhiên liên tục đều trên đoạn [a, b](**Uniform continuous random variable**).

   Là biến ngẫu nhiên *Z* có **hàm mật độ f(x) =1/(b-a) với a ≤ x ≤ b.**
   - **Trung bình μ = E(X) = (a+b)/2,**
   - **phương sai: $\sigma^2 = (b - a)^2 / 12$.**

2) Biến ngẫu nhiên tiêu chuẩn (**Standard normal**)

- **Hàm mật độ:** $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$;
- **Hàm phân phối tích lũy** $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ (giá trị của hàm này được tra tại table III-A6- textbook)
- **Trung bình:** $\mu = 0$

- **Phương sai:** $\sigma^2 = 1$
- **Tính xác suất** $P(Z < a) = \Phi(a)$; $P(a < Z < b) = \Phi(b) - \Phi(a)$ (tra bảng table III-A6- textbook).
- **Tìm a để** $P(Z < a) = \alpha$ thì a sẽ là số thỏa mãn $\Phi(a) = \alpha$

3) Biến ngẫu nhiên chuẩn (**Normal**) với trung bình μ và phương sai σ^2

- **Tính xác suất**

$$P(X < a) = P(Z < \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma});$$

$$P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

- **Tìm a để** $P(X < a) = \alpha \Leftrightarrow P(Z < \frac{a-\mu}{\sigma}) = \alpha$ thì a sẽ là số thỏa mãn

$\Phi(\frac{a-\mu}{\sigma}) = \alpha$

4) Phân phối mũ **(Exponential)**

Là biến ngẫu nhiên **chỉ khoảng thời gian giữa hai 'event'** trong quy trình Poisson.

Nếu λ là số trung bình các 'event' trên một interval có độ dài là 1 đơn vị thì

Hàm mật độ xác suất: $f(x) = \lambda e^{-\lambda x}$

Hàm phân phối tích lũy: $F(x) = 1 - e^{-\lambda x}$

Trung bình: $\mu = 1/\lambda$

Phương sai: $\sigma^2 = 1/\lambda^2$

5) Xấp xỉ chuẩn của phân phối nhị thức và phân phối Poisson.

a) Phân phối nhị thức

$$P(X \le x) \approx P\left(Z \le \frac{x + 0.5 - np}{\sqrt{np(1-p)}} \right)$$

$$P(X \ge x) \approx P\left(Z \ge \frac{x - 0.5 - np}{\sqrt{np(1-p)}} \right)$$

b) Phân phối Poisson

$$P(X \le x) = P\left(Z \le \frac{x-\lambda}{\sqrt{\lambda}} \right) \approx \Phi\left(\frac{x-\lambda}{\sqrt{\lambda}} \right)$$

Tóm tắt chương VI: **Thống kê mô tả (Statistical Descriptive)**

Kiến thức chung:

- **Trung bình mẫu** (sample mean)
- **Phương sai mẫu** (sample variance)
- **Độ lệch tiêu chuẩn** (sample standard deviation)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

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

$$s = \sqrt{s^2}$$

- **Sample range(**Khoảng biến thiên mẫu**):**

$$r = \max(x_i) - \min(x_i)$$

- **Median (**trung vị mẫu**):** Là số chia mẫu ra làm hai

  + Nếu kích thước mẫu là lẻ thì median chính là số đứng ở vị trí chính giữa của mẫu
  + Nếu kích thước mẫu là chẵn thì median là trung bình của hai số đứng giữa mẫu.
- **Mode:** Là số xuất hiện nhiều nhất trong mẫu.
- **Quartiles (**Tứ phân vị**):** là ba số mà chúng chia mẫu thành 4 phần bằng nhau, kí hiệu là q_1, q_2, q_3
- Khoảng tứ phân vị**: IQR = $q_3 - q_1$**

- Phân phối tần số **(Frequency distribution)**
- Phân phối tần suất **(Relative frequency distribution)**
- Phân phối tích lũy **(Cumulative distribution)**
- **Stem and leaf.**
- **Box plot**

Tóm tắt chương VII: **Point Estimate and The Central Limit Theorem (Ước lượng điểm và định lí giới hạn trung tâm)**

1) Point estimate

- For μ, the point estimate is the $\bar{x}$ - sample mean.
- For σ^2, the point estimate is s^2 - the sample variance.
- For p, the point estimate is $\hat{p}$ - the sample proportion.
- For $\mu_1 - \mu_2$, the estimate is $\bar{x}_1 - \bar{x}_2$.

- For $p_1 - p_2$, the estimate is $\hat{p}_1 - \hat{p}_2$

2) The Central Limit Theorem

- $\overline{X} \approx N(\mu; \frac{\sigma^2}{n})$ (khi n đủ lớn)

  (với μ là trung bình tổng thể ; σ^2 là phương sai tổng thể, n là kích thước mẫu)

  Ứng dụng để tính $P(a < \overline{X} < b) \approx P(\frac{a-\mu}{\sigma/\sqrt{n}} < Z < \frac{b-\mu}{\sigma/\sqrt{n}}) = \Phi(\frac{b-\mu}{\sigma/\sqrt{n}}) - \Phi(\frac{a-\mu}{\sigma/\sqrt{n}})$

- $\overline{X}_1 - \overline{X}_2 \approx N(\mu_1 - \mu_2; \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2})$ (Khi n_1, n_2 đủ lớn)

  (với μ_1, μ_2 là trung bình tổng thể 1, 2 ; ; $\sigma^2{}_1$; $\sigma^2{}_2$ là phương sai tổng thể 1, 2, n_1; n_2 là kích thước mẫu 1 và 2).

$$P(a < \overline{X}_1 - \overline{X}_2 < b) \approx P(\frac{a-\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < Z < \frac{b-\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}) = \Phi(\frac{b-\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}) - \Phi(\frac{a-\mu_1-\mu_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}})$$

Tóm tắt chương VIII: **Confidence Interval (viết tắt là CI) for population parameter (Khoảng tin cậy cho tham số của tổng thể)**

- Bài toán tổng quát của chương này:

Dựa vào mẫu, tìm ***ước lượng khoảng (confidence interval = CI)*** của một *tham số* (μ hoặc σ hoặc p) của biến ngẫu nhiên.

- Định nghĩa chung: Ước lượng khoảng của tham số θ với độ tin cậy (**confidence level**) 1- α là khoảng [L, U] sao cho

$$P(L \le \theta \le U) = 1 - \alpha,$$

  trong đó L, U là các hàm của mẫu ngẫu nhiên $(X_1,\ldots,X_n)$.

- **CI on μ: Với độ tin cậy (confidence level) 1- α thì khoảng tin cậy cho μ là:**

  TH1 (tổng thể có phân phối chuẩn, σ đã biết).

  - Hai phía: $\mu \in [L, U] = [\overline{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}]$

    với $z_{\alpha/2}$ là số mà $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$.

  - Một phía $\begin{matrix} upper: \mu \le \overline{X} + z_\alpha \sigma/\sqrt{n} \\ lower: \mu \ge \overline{X} - z_\alpha \sigma/\sqrt{n} \end{matrix}$

  - Nếu dùng $\overline{x}$ để xấp xỉ cho μ thì muốn sai số không vượt quá E thì với độ tin cậy 1-α kích thước mẫu cần tìm là:

$$n = \left(\frac{\sigma . z_{\alpha/2}}{E}\right)^2$$

TH2: (kích thước mẫu lớn) (n > = 30)

- Hai phía: $\mu \in [L,U] = [\overline{X} - \frac{z_{\alpha/2}S}{\sqrt{n}}, \overline{X} + \frac{z_{\alpha/2}S}{\sqrt{n}}]$

  với $z_{\alpha/2}$ là số mà $\Phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$.

- Một phía $\begin{aligned} &upper: \mu \le \overline{X} + z_\alpha S/\sqrt{n} \\ &lower: \mu \ge \overline{X} - z_\alpha S/\sqrt{n} \end{aligned}$

TH3: (tổng thể có phân phối chuẩn**,** σ chưa biết)

- Hai phía: $\mu \in [L,U] = [\overline{X} - \frac{t_{\alpha/2;n-1}S}{\sqrt{n}}, \overline{X} + \frac{t_{\alpha/2;n-1}S}{\sqrt{n}}]$

với $t_{\alpha/2}; n-1$ là số được tra bởi table V-A9-textbook bằng cách lấy giao của cột α/2 và dòng n-1.

- Một phía $\begin{aligned} &upper: \mu \le \overline{X} + t_{\alpha;n-1}S/\sqrt{n} \\ &lower: \mu \ge \overline{X} - t_{\alpha;n-1}S/\sqrt{n} \end{aligned}$

- **CI on σ^2: Với độ tin cậy (confidence level) 1- α thì khoảng tin cậy cho σ^2 là:**

- Hai phía $\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \right]$

  Với $\chi^2_{\alpha/2;n-1}$ là giá trị được tra từ Table IV-A8-textbook.

- Một phía $\begin{aligned} &upper: \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}} \\ &lower: \sigma^2 \ge \frac{(n-1)s^2}{\chi^2_{\alpha,n-1}} \end{aligned}$

- **CI on p: Với độ tin cậy (confidence level) 1- α thì khoảng tin cậy cho p là:**

- Hai phía $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- Một phía : $\begin{aligned} &upper: p \le \hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &lower: p \ge \hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \end{aligned}$

- **Với độ tin cậy 1 – α, muốn sai số khi xấp xỉ *p* bởi $\hat{p}$ không vượt quá E thì kích thước mẫu cần thiết là:**

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p})$$

- **Với độ tin cậy ít nhất bằng 1 – α, muốn sai số khi xấpxỉ *p* bởi $\hat{p}$ không vượt quá E thì kích thước mẫu cần thiết là:**

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 0.25$$

Tóm tắt chương XI: **Hồi quy tuyến tính (Linear Regression) và hệ số tương quan (Correlation).**

Với n cặp quan sát $(x_1; y_1);...(x_n; y_n)$ lấy từ tổng thể của biến ngẫu nhiên (X, Y) thì

- **The estimated (fitted) linear regression line :**

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x \quad \textbf{với}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{Sxx} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}};$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 : slope$$

$$\hat{\beta}_0 : \text{intercept}$$

- **Sample Correlation -** Hệ số tương quan mẫu

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Ý nghĩa của R: Đo mức độ tương quan tuyến tính mẫu của X và Y. Giá trị của R càng xấp xỉ lớn thì mức độ tương quan tuyến tính càng mạnh.

- **Tổng bình phương các sai số (SS_E).**

$$SS_E = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}$$

- **Test on β_1:**
  1) H_0: $\beta_1 = \beta_{1,0}$ H_1: $\beta_1 \neq \beta_{1,0}$
  2) Test statistic: $T_0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\dfrac{SS_E}{(n-2)S_{xx}}}}$
  3) Critical values $\pm t_{\alpha/2;n-2}$
  4) Reject H0 nếu $T_0 > t_{\alpha/2;n-2} or T_0 < -t_{\alpha/2;n-2}$

- **Test on β_0**

  1) $H_0: \beta_0 = \beta_{0,0}$ $H_1: \beta_0 \neq \beta_{0,0}$

  2) Test statistic $T_0 = \dfrac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\dfrac{SS_E}{n-2}\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]}}$

  3) Critical values $\pm t_{\alpha/2;n-2}$

  4) Reject H0 nếu $T_0 > t_{\alpha/2;n-2} or T_0 < -t_{\alpha/2;n-2}$

- **Test on ρ**: = population correlation

  1) $H_0: \rho = 0$ $H_1: \rho \neq 0$

  2) Test statistic $T_0 = \dfrac{R\sqrt{n-2}}{\sqrt{1-R^2}}$

  3) Critical values $\pm t_{\alpha/2;n-2}$

  4) Reject H0 nếu $T_0 > t_{\alpha/2;n-2} or T_0 < -t_{\alpha/2;n-2}$