MAS (1)

- Population
  - parameter : characteristic of population
  - sample
  - statistics : characteristic of sample
  - variable : characteristic of elements
  - data : value of variable

- Phương pháp collect data
  - retrospective study : các data có từ quá khứ
  - observational study : data từ quan sát, đo đạc
  - experment study : data từ thực nghiệm
  - simulation study : using models -> data
  - survey : [sample / population : CENSUS

- Type of data
  - <u>qualitative</u> (định tính) : gender, color, major, place, size (~những thứ đã đc phân loại)
  - <u>quatitative</u> (định lượng) — discrete (~~rời~~ rời rạc) / continuous (liên tục)

- Sampling method
  - representative : lấy sao cho <u>đại diện</u> đc population
  - replacement / without replacement : chọn xg bỏ ra (k° lấy nx) / chọn xg bỏ lại (có thể lấy tiếp)
  - non random (not representative)
  - random sampling :
    - + simple random
    - + stratified : bốc đều, từ các nhóm (class), mỗi class simple random

I. Basic probability formulas

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \mid B) = \dfrac{P(B \mid A) \,.\, P(A)}{P(B)}$
- If A, B independent: $P(A \cap B) = P(A) \,.\, P(B)$

II. Discrete random variables

- $\mathcal{M} = E(x) = \sum x_i \,.\, P(x=x_i)$
- $\sigma^2 = V(x) = \sum (x_i - \mathcal{M})^2 \,.\, P(x=x_i)$

$= \sum x_i^2 . P(x=x_i) - \mathcal{M}^2$

- $E(ax + by) = a.E(x) + b.E(y)$
- $V(ax + by) = a^2 . V(x) + b^2 . V(y)$
- Probability mass function: $f(x_i) = P(x=x_i)$
- Cumulative distribution function: $F(x_i) = P(x \leq x_i)$
- Some special distribution:
  1. Discrete uniform distribution
     - $P(x=X_i) = \frac{1}{n}$
     - $\mathcal{M} = \frac{a+b}{2}$
     - $\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$
  2. Binomial distribution
     - $P(x=k) = nCk . p^k . (1-p)^{n-k}$
     - $\mathcal{M} = n.p$
     - $\sigma^2 = n.p . (1-p)$
  3. Poisson distribution
     - $P(x=k) = \frac{e^{-\lambda.T}}{k!} (\lambda.T)^k$
     - $\mathcal{M} = \lambda.T$
     - $\sigma^2 = \lambda.T$
  4. Hypergeometric distribution
     - $P(x=k) = \frac{KCk . (N-K)C(n-k)}{NCn}$
     - $\mathcal{M} = n.p$
     - $\sigma^2 = n.p.(1-p). \frac{N-n}{N-1}$
  5. Geometric distribution
     - $P(x=k) = (1-p)^{k-1} . p$
     - $\mathcal{M} = \frac{1}{p}$
     - $\sigma^2 = \frac{1-p}{p^2}$

6. Negative binomial distribution
    - $P(x=k) = (k-1)C(r-1) \cdot p^r \cdot (1-p)^{k-r}$
    - $\mathcal{M} = \frac{r}{p}$
    - $\sigma^2 = \frac{r \cdot (1-p)}{p^2}$

III. Continuous random variable

- Probability density function f(x): $P(a<x<b) = \int_a^b f(x)\, d_x$
- Cumulative distribution function F(x):
    - $F(x_i) = P(x \le x_i)$
    - $F(x_i)' = f(x_i)$
- $\mathcal{M} = E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)\, d_x$
- $E(x^n) = \int_{-\infty}^{+\infty} x^n \cdot f(x)\, d_x$
- $\sigma^2 = V(x) = \int_{-\infty}^{+\infty} x^2 \cdot f(x)\, d_x - \mathcal{M}^2$
- Some special distribution:
    1. Continuous uniform distribution
        - $f(x) = \frac{1}{b-a}$, $a \le x \le b$

          $= 0$, elsewhere
        - $\mathcal{M} = \frac{a+b}{2}$
        - $\sigma^2 = \frac{(b-a)^2}{12}$
    2. Normal distribution $N(\mathcal{M}, \sigma^2)$
        - $z = \frac{x - \mathcal{M}}{\sigma}$
        - $f(z) = \frac{1}{\sqrt{2\Pi}} \cdot e^{\frac{x^2}{2}}$
        - $\phi(x) = p(z<x_i)$

- $\phi(-x) = 1 - \phi(x)$

3. Normal distribution approximate binomial and poisson distribution
   a. Binomial ($np > 5$ and $n(1-p) > 5$)
      - $z = \frac{x - n.p}{\sqrt{n.p.(1-p)}}$
      - $P(X_{BINORM} \leq a) = P(X_{NORMAL} \leq a+0.5)$
      - $P(X_{BINORM} \geqq a) = P(X_{NORMAL} \geqq a-0.5)$

   b. Poisson
      - $z = \frac{x - \lambda}{\sqrt{\lambda}}$
      - $P(X_{POISSON} \leq a) = P(X_{NORMAL} \leq a+0.5)$
      - $P(X_{POISSON} \geqq a) = P(X_{NORMAL} \geqq a-0.5)$

4. Exponential distribution
   - $f(x) = \lambda \,.\, e^{-\lambda.T}$, $x > 0$
   - $= 0$, elsewhere
   - $P(x \geqq a) = e^{-\lambda.a}$,($a > 0$)
   - $\mathcal{M} = \frac{1}{\lambda}$
   - $\sigma^2 = \frac{1}{\lambda^2}$

IV. Descriptive statistic (Take a sample of size n from population N)

- Sample mean: $\overline{x} = \frac{\sum x_i}{n}$
- Sample median: $L = \frac{n+1}{2}$ so Median $= \frac{x_{ceil(L)} + x_{floor(L)}}{2}$
- Mode: Số phần tử xuất hiện nhiều nhất
- Range: max - min
- Sample variance: $s^2 = \frac{\sum (\overline{x} - x_i)^2}{n-1}$
- Quatiles:

- $L_1 = \frac{n+1}{4}$ so $Q_1 = \frac{x_{ceil(L_1)} + x_{floor(L_1)}}{2}$
- $L_2 = \frac{n+1}{2}$ so $Q_2 = \frac{x_{ceil(L_2)} + x_{floor(L_2)}}{2}$
- $L_3 = \frac{3.(n+1)}{4}$ so $Q_3 = \frac{x_{ceil(L_3)} + x_{floor(L_3)}}{2}$

V. Sampling distribution

- Population mean $\mathcal{M}$, variance σ^2. Sample size n. *(Normal distribution or n > 30)*:
  - Phân phối của $\overline{X}$ có dạng: N($\mathcal{M}$, $\frac{\sigma^2}{n}$)
  - Phân phối của $\overline{X_1}$ - $\overline{X_2}$ có dạng: N($\mathcal{M}_1$ - $\mathcal{M}_2$, $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$)
- For proportion of population p, sample size n. *(np ≧ 5 or n.(1-p) ≧ 5)*:
  - Phân phối của $\widehat{P}$ có dạng: N(P , $\frac{P.(1-P)}{n}$)
  - Phân phối của $\widehat{P_1}$ - $\widehat{P_2}$ có dạng: N(P_1 - P_2 , $\frac{P_1.(1-P_1)}{n_1} + \frac{P_2.(1-P_2)}{n_2}$)

VI. Statistical intervals - Test claims for one sample

- (l, u) = ($\overline{X}$ - E, $\overline{X}$ + E)
- width = 2E
- P-value = 2 . P(Z > $|Z_0|$)

1. Population variance known
   - $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
   - $z_0 = \frac{\overline{X} - \mathcal{M}}{\sigma / \sqrt{n}}$
2. Population variance unknown
   - n > 30:
     - $E = z_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$

        - $z_0 = \dfrac{\overline{X} - \mathcal{M}}{S / \sqrt{n}}$
    - n ≤ 30:
        - $E = t_{\alpha/2, n-1} \cdot \dfrac{S}{\sqrt{n}}$
        - $t_0 = \dfrac{\overline{X} - \mathcal{M}}{S / \sqrt{n}}$
- For propotion:
    - $(l, u) = (\widehat{P} - E, \widehat{P} + E)$
    - $E = z_{\alpha/2} \cdot \sqrt{\dfrac{P.(1-P)}{n}}$
    - $z_0 = \dfrac{\widehat{P} - P}{\sqrt{\frac{P.(1-P)}{n}}}$
    - Nếu đề không cho $\widehat{P}$, mặc định $\widehat{P} = 0.5$
- Nếu là one-side thì tương tự nhưng thay $\alpha/2$ thành α

VII. Test claims for 2 samples (2 population independent, normal distribution or both n_1, $n_2 > 30$)

- $(l, u) = (\overline{X}_1 - \overline{X}_2 - E, \overline{X}_1 - \overline{X}_2 + E)$

1. Population variance known
    - $E = z_{\alpha/2} \cdot \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
    - $z_0 = \dfrac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
2. Population variance unknown
    - Assume $\sigma_1^2 = \sigma_2^2$
        - Degree of freedom: $df = n_1 + n_1 + 2$
        - $S_p^2 = \dfrac{(n_1 - 1) . S_1^2 + (n_2 - 1) . S_2^2}{n_1 + n_2 - 2}$

        - $E = t_{\alpha/2,\, df} \cdot \sqrt{\frac{S_p^{\ 2}}{n_1} + \frac{S_p^{\ 2}}{n_2}}$
        - $t_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_p^{\ 2}}{n_1} + \frac{S_p^{\ 2}}{n_2}}}$
    - Not assume $\sigma_1^{\ 2} = \sigma_2^{\ 2}$
        - Degree of freedom: $df = \frac{\left(\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}\right)^2}{\frac{S_1^{\ 4}}{n_1^{\ 2}.(n_1 - 1)} + \frac{S_2^{\ 4}}{n_2^{\ 2}.(n_2 - 1)}}$
        - $E = t_{\alpha/2,\, df} \cdot \sqrt{\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}}$
        - $t_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^{\ 2}}{n_1} + \frac{S_2^{\ 2}}{n_2}}}$
- For propotion:
    - $(l, u) = (\widehat{P}_1 - \widehat{P}_2 - E\, ,\ \widehat{P}_1 - \widehat{P}_2 + E)$
    - $E = z_{\alpha/2} \cdot \sqrt{\frac{\widehat{P}_1 . (1 - \widehat{P}_1)}{n_1} + \frac{\widehat{P}_2 . (1 - \widehat{P}_2)}{n_2}}$
    - $\widehat{P} = \frac{x_1 + x_2}{n_1 + n_2}$ (trong đó $x_i = n . \widehat{P}_i$)
    - $z_0 = \frac{\widehat{P}_1 - \widehat{P}_2 - \Delta_0}{\sqrt{\widehat{P} . (1 - \widehat{P}) . \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

VIII. Linear Regression

- $S_{XY} = \sum\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \sum x_i y_i \ - \ n . \overline{x} . \overline{y}$
- $S_{XX} = \sum\left(x_i - \overline{x}\right)^2 = \sum x_i^{\ 2} - n . \overline{x}^2$

- $S_{YY} = \sum\left(y_i - \bar{y}\right)^2 = \sum y_i^2 - n \cdot \bar{y}^2$
- Slope: $\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}} = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - n \cdot \bar{x}^2}$
- Intercept: $\widehat{\beta_0} = \bar{y}$ - $\widehat{\beta_1} \cdot \bar{x}$
- Error sum of square: $SS_E = \sum\left(y_i - \widehat{y_i}\right)^2$
- Regression sum of square: $SS_R = \sum\left(\widehat{y_i} - \bar{y}\right)^2$
- Total sum of square: $SS_T = \sum\left(y_i - \bar{y}\right)^2$
- $SS_E + SS_R = SS_T$
- Standard error: $\widehat{\sigma} = \sqrt{\frac{SS_E}{n-2}}$
- Coefficient of correlation: $R = \sqrt{\frac{SS_R}{SS_T}} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$
- Test claims about the slope *(df = n-2)*:
  - $se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma}^2}{S_{XX}}}$
  - $t_0 = \frac{\widehat{\beta_1} - \beta_{1,0}}{se(\widehat{\beta_1})}$
- Test claims about the intercept *(df = n-2)*:
  - $se(\widehat{\beta_0}) = \sqrt{\widehat{\sigma}^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right)}$
  - $t_0 = \frac{\widehat{\beta_0} - \beta_{0,0}}{se(\widehat{\beta_0})}$
- Test claims about the coefficient of correlation *(df = n-2)*: $t_0 = \frac{R - 0}{\sqrt{\frac{1-R^2}{n-2}}}$