**I. Basic probability formulas**

- P(A ∪ B) = P(A) + P(B) - P(A ∩ B)
- P(A | B) = $\frac{P(A \cap B)}{P(B)}$
- P(A | B) = $\frac{P(B \mid A)\,.P(A)}{P(B)}$
- If A, B independent: P(A ∩ 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≤$x_i$)
- 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 . $p^k$ . $(1\text{-}p)^{n-k}$
     - $\mathcal{M}$ = 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

- $P(x=k) = \dfrac{KCk\,.(N-K)C(n-k)}{NCn}$
- $\mathcal{M}$ = n.p
- $\sigma^2$ = n.p.(1-p). $\dfrac{N-n}{N-1}$

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

### III. Continuous random variable

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

       = 0 , elsewhere
     - $\mathcal{M} = \dfrac{a+b}{2}$
     - $\sigma^2 = \dfrac{(b-a)^2}{12}$

2. Normal distribution N($\mathcal{M}$, $\sigma^2$)

  - z = $\dfrac{x - \mathcal{M}}{\sigma}$
  - f(z) = $\dfrac{1}{\sqrt{2\Pi}} . e^{\frac{x^2}{2}}$
  - $\phi$(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)

   $\mu = n.p$

   $\sigma^2 = n.p(1-p$

      - z = $\dfrac{x - n.p}{\sqrt{n.p.(1-p)}}$
      - 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 = $\dfrac{x - \lambda}{\sqrt{\lambda}}$
      - 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

   $\lambda$

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

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

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

- Sample variance: $s^2 = \dfrac{\sum(\underline{x} - x_i)^2}{n-1}$
- Quatiles:
  - $L_1 = \dfrac{n+1}{4}$ so $Q_1 = \dfrac{x_{ceil(L_1)} + x_{floor(L_1)}}{2}$
  - $L_2 = \dfrac{n+1}{2}$ so $Q_2 = \dfrac{x_{ceil(L_2)} + x_{floor(L_2)}}{2}$
  - $L_3 = \dfrac{3.(n+1)}{4}$ so $Q_3 = \dfrac{x_{ceil(L_3)} + x_{floor(L_3)}}{2}$

**V. Sampling distribution**

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

**VI. Statistical intervals - Test claims for one sample**

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

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

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

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

- (l, u) = ($\underline{X_1}$ - $\underline{X_2}$ - E , $\underline{X_1}$ - $\underline{X_2}$ + E)

1. Population variance known
   - E = $z_{\alpha/2} \cdot \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$
   - $z_0 = \frac{\underline{\underline{X_1}} - \underline{\underline{X_2}} - \Delta_0}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$
2. Population variance unknown
   - Assume ${\sigma_1}^2 = {\sigma_2}^2$
     - Degree of freedom: df = $n_1$ + $n_1$ + 2
     - ${S_p}^2 = \frac{(n_1 - 1) \,.\, {S_1}^2 + (n_2 - 1) \,.\, {S_2}^2}{n_1 + n_2 - 2}$
     - E = $t_{\alpha/2,df} \cdot \sqrt{\frac{{S_p}^2}{n_1} + \frac{{S_p}^2}{n_2}}$
     - $t_0 = \frac{\underline{\underline{X_1}} - \underline{\underline{X_2}} - \Delta_0}{\sqrt{\frac{{S_p}^2}{n_1} + \frac{{S_p}^2}{n_2}}}$
   - Not assume ${\sigma_1}^2 = {\sigma_2}^2$

        - Degree of freedom: df = $\dfrac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2.(n_1-1)}+\frac{S_2^4}{n_2^2.(n_2-1)}}$
        - E = $t_{\alpha/2,df}$ . $\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}$
        - $t_0 = \dfrac{\underline{X_1}-\underline{X_2}-\Delta_0}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$
- For propotion:
    - (l, u) = ($\widehat{P_1}$ - $\widehat{P_2}$ - E , $\widehat{P_1}$ - $\widehat{P_2}$ + E)
    - E = $z_{\alpha/2}$ . $\sqrt{\dfrac{\widehat{P_1}.(1-\widehat{P_1})}{n_1}+\dfrac{\widehat{P_2}.(1-\widehat{P_2})}{n_2}}$
    - $\widehat{P} = \dfrac{x_1+x_2}{n_1+n_2}$ (trong đó $x_i$ = n . $\widehat{P_i}$ )
    - $z_0 = \dfrac{\widehat{P_1}-\widehat{P_2}-\Delta_0}{\sqrt{\widehat{P}.(1-\widehat{P}).\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$

**VIII. Linear Regression**

- $S_{XY} = \sum\left(x_i-\underline{x}\right)\left(y_i-\underline{y}\right) = \sum x_iy_i - n.\underline{x}.\underline{y}$
- $S_{XX} = \sum\left(x_i-\underline{x}\right)^2 = \sum x_i^2 - n.\underline{x}^2$
- $S_{YY} = \sum\left(y_i-\underline{y}\right)^2 = \sum y_i^2 - n.\underline{y}^2$
- Slope: $\widehat{\beta_1} = \frac{S_{XY}}{S_{XX}} = \dfrac{\sum x_iy_i - n.\underline{x}.\underline{y}}{\sum x_i^2 - n.\underline{x}^2}$
- Intercept: $\widehat{\beta_0} = \underline{y} - \widehat{\beta_1}.\underline{x}$
- Error sum of square: $SS_E = \sum\left(y_i-\widehat{y_i}\right)^2$
- Regression sum of square: $SS_R = \sum\left(\widehat{y_i}-\underline{y}\right)^2$
- Total sum of square: $SS_T = \sum\left(y_i-\underline{y}\right)^2$
- $SS_E + SS_R = SS_T$
- Standard error: $\hat{\sigma} = \sqrt{\dfrac{SS_E}{n-2}}$

- Coefficient of correlation: R = $\sqrt{\frac{SS_R}{SS_T}} = \frac{S_{XY}}{\sqrt{S_{XX}.S_{YY}}}$
- Test claims about the slope *(df = n-2)*:
    - se($\widehat{\beta_1}$) = $\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$
    - $t_0$ = $\frac{\widehat{\beta_1} - \beta_{1,0}}{se(\widehat{\beta_1})}$
- Test claims about the intercept *(df = n-2)*:
    - se($\widehat{\beta_0}$) = $\sqrt{\hat{\sigma}^2 . \left(\frac{1}{n} + \frac{\underline{x}^2}{S_{XX}}\right)}$
    - $t_0$ = $\frac{\widehat{\beta_0} - \beta_{0,0}}{se(\widehat{\beta_0})}$
- Test claims about the coefficient of correlation *(df = n-2)*: $t_0$ = $\frac{R - 0}{\sqrt{\frac{1 - R^2}{n - 2}}}$

Thứ ngày

# 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
  - experiment study : data từ thực nghiệm
  - simulation study : using models → data
  - survey : [ sample / population : CENSUS

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

- Sampling method
  - representative : lấy sao cho đại diện 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

Thứ ngày

    + cluster random : chỉ bốc 1 số cluster (class) nhất định
    + system random : random 1 cách có hệ thống (VD chỉ chọn những số chẵn)