

## I. Basic probability formulas

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A | B) = \frac{P(A \cap B)}{P(B)}$
- $P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$
- If A, B independent:  $P(A \cap B) = P(A) \cdot P(B)$

## II. Discrete random variables

- $\mathcal{M} = E(x) = \sum x_i \cdot P(x=x_i)$
- $\sigma^2 = V(x) = \sum (x_i - \mathcal{M})^2 \cdot P(x=x_i)$   
 $= \sum x_i^2 \cdot P(x=x_i) - \mathcal{M}^2$
- $E(ax + by) = a \cdot E(x) + b \cdot E(y)$
- $V(ax + by) = a^2 \cdot V(x) + b^2 \cdot V(y)$
- 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

- $P(x=k) = nCk \cdot p^k \cdot (1-p)^{n-k}$
- $\mathcal{M} = n \cdot p$
- $\sigma^2 = n \cdot p \cdot (1-p)$

### 3. Poisson distribution

- $P(x=k) = \frac{e^{-\lambda T}}{k!} (\lambda \cdot T)^k$
- $\mathcal{M} = \lambda \cdot T$
- $\sigma^2 = \lambda \cdot T$

### 4. Hypergeometric distribution

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

## 5. Geometric distribution

- $P(x=k) = (1-p)^{k-1} \cdot p$
- $\mathcal{M} = \frac{1}{p}$
- $\sigma^2 = \frac{1-p}{p^2}$

## 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  $f(x)$ :  $P(a < x < b) = \int_a^b f(x) dx$
- Cumulative distribution function  $F(x)$ :
  - $F(x_i) = P(x \leq x_i)$
  - $F(x_i)' = f(x_i)$
- $\mathcal{M} = E(x) = \int_{-\infty}^{+\infty} x \cdot f(x) dx$
- $E(x^n) = \int_{-\infty}^{+\infty} x^n \cdot f(x) dx$
- $\sigma^2 = V(x) = \int_{-\infty}^{+\infty} x^2 \cdot f(x) dx - \mathcal{M}^2$
- Some special distribution:

### 1. Continuous uniform distribution

- $f(x) = \frac{1}{b-a}$ ,  $a \leq x \leq b$   
 $= 0$ , elsewhere
- $\mathcal{M} = \frac{a+b}{2}$
- $\sigma^2 = \frac{(b-a)^2}{12}$

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

- $z = \frac{x - \mathcal{M}}{\sigma}$
- $f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$
- $\phi(x) = p(z < x_i)$

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

### 3. Normal distribution approximate binomial and poisson distribution

#### a. Binomial ( $np > 5$ and $n(1-p) > 5$ )

- $Z = \frac{x - np}{\sqrt{np.(1-p)}}$
- $P(X_{\text{BINORM}} \leq a) = P(X_{\text{NORMAL}} \leq a+0.5)$
- $P(X_{\text{BINORM}} \geq a) = P(X_{\text{NORMAL}} \geq a-0.5)$

#### b. Poisson

- $Z = \frac{x - \lambda}{\sqrt{\lambda}}$
- $P(X_{\text{POISSON}} \leq a) = P(X_{\text{NORMAL}} \leq a+0.5)$
- $P(X_{\text{POISSON}} \geq a) = P(X_{\text{NORMAL}} \geq a-0.5)$

### 4. Exponential distribution

- $f(x) = \lambda \cdot e^{-\lambda x}, x > 0$
- $= 0, \text{ elsewhere}$
- $P(x \geq a) = e^{-\lambda a}, (a > 0)$
- $M = \frac{1}{\lambda}$
- $\sigma^2 = \frac{1}{\lambda^2}$

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

- Sample mean:  $\bar{x} = \frac{\sum x_i}{n}$
- 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:  $s^2 = \frac{\sum (\bar{x} - x_i)^2}{n-1}$
- Quartiles:
  - $L_1 = \frac{n+1}{4}$  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  $\sigma^2$ . Sample size n. (*Normal distribution or n > 30*):
  - Phân phối của  $\bar{X}$  có dạng:  $N(\mathcal{M}, \frac{\sigma^2}{n})$
  - Phân phối của  $\bar{X}_1 - \bar{X}_2$  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. ( $np \geq 5$  or  $n.(1-p) \geq 5$ ):
  - Phân phối của  $\hat{P}$  có dạng:  $N(P, \frac{P.(1-P)}{n})$
  - Phân phối của  $\hat{P}_1 - \hat{P}_2$  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

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

### 1. Population variance known

- $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- $z_0 = \frac{\bar{X} - \mathcal{M}}{\sigma / \sqrt{n}}$

### 2. Population variance unknown

- $n > 30$ :
  - $E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
  - $z_0 = \frac{\bar{X} - \mathcal{M}}{s / \sqrt{n}}$
- $n \leq 30$ :
  - $E = t_{\alpha/2, n-1} \cdot \frac{s}{\sqrt{n}}$
  - $t_0 = \frac{\bar{X} - \mathcal{M}}{s / \sqrt{n}}$

### • For proportion:

- $(l, u) = (\hat{P} - E, \hat{P} + E)$
- $E = z_{\alpha/2} \cdot \sqrt{\frac{P.(1-P)}{n}}$
- $z_0 = \frac{\hat{P} - P}{\sqrt{\frac{P.(1-P)}{n}}}$
- 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 $n_1, n_2 > 30$ )

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

### 1. Population variance known

$$\circ \quad E = z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\circ \quad z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

### 2. Population variance unknown

$$\circ \quad \text{Assume } \sigma_1^2 = \sigma_2^2$$

■ Degree of freedom:  $df = n_1 + n_2 - 2$

$$\blacksquare \quad S_p^2 = \frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{n_1 + n_2 - 2}$$

$$\blacksquare \quad E = t_{\alpha/2, df} \cdot \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

$$\blacksquare \quad t_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

$$\circ \quad \text{Not assume } \sigma_1^2 = \sigma_2^2$$

$$\blacksquare \quad \text{Degree of freedom: } df = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{S_1^4}{n_1^2(n_1 - 1)} + \frac{S_2^4}{n_2^2(n_2 - 1)}}$$

$$\blacksquare \quad E = t_{\alpha/2, df} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

$$\blacksquare \quad t_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

### • For proportion:

$$\circ \quad (l, u) = (\hat{P}_1 - \hat{P}_2 - E, \hat{P}_1 - \hat{P}_2 + E)$$

$$\circ \quad E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{P}_1 \cdot (1 - \hat{P}_1)}{n_1} + \frac{\hat{P}_2 \cdot (1 - \hat{P}_2)}{n_2}}$$

- $\hat{P} = \frac{x_1 + x_2}{n_1 + n_2}$  (trong đó  $x_i = n \cdot \hat{P}_i$ )

$$\circ \quad z_0 = \frac{\hat{P}_1 - \hat{P}_2 - \Delta_0}{\sqrt{\hat{P} \cdot (1 - \hat{P}) \cdot \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

### VIII. Linear Regression

- $S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}$

- $S_{XX} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n \cdot \bar{x}^2$

- $S_{YY} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n \cdot \bar{y}^2$

- Slope:  $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - n \cdot \bar{x}^2}$

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

- Error sum of square:  $SS_E = \sum (y_i - \hat{y}_i)^2$

- Regression sum of square:  $SS_R = \sum (\hat{y}_i - \bar{y})^2$

- Total sum of square:  $SS_T = \sum (y_i - \bar{y})^2$

- $SS_E + SS_R = SS_T$

- Standard error:  $\hat{\sigma} = \sqrt{\frac{SS_E}{n-2}}$

- Coefficient of correlation:  $R = \sqrt{\frac{SS_R}{SS_T}} = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$

- Test claims about the slope ( $df = n-2$ ):

- $se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}$

- $t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$

- Test claims about the intercept ( $df = n-2$ ):

- $se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \cdot \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right)}$
- $t_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$
- Test claims about the coefficient of correlation ( $df = n-2$ ):  $t_0 = \frac{R - 0}{\sqrt{\frac{1 - R^2}{n - 2}}}$

Thứ	ngày
s — MAS (1)	
• - population	
- parameter : characteristic of population	
- sample	
- statistics : characteristic of sample	
- variable : characteristic of elements	
- data : value of variable	
• Phác họa cách collect data	
- retrospective study : các data có từ quá khứ	
- observational study : data từ quan sát, đúc kết	
- 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 xác định loại)	
- quantitative (định lượng) : discrete (tách rời rạc) continuous (liền tục)	
• Sampling method	
- representative : lấy sao cho đại diện cho population	
- replacement / without replacement : chọn xg bỏ ra (kèm lại) / chọn xg bỏ lui (có thể lấy tiếp)	
- non random (not representative)	
- random sampling :	
+ simple random	
+ stratified : bốc đều, dù có nhầm (class), mỗi class simple random	HONGHA

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)	