TỔNG HỢP KIẾN THỨC
LÝ THUYẾT XÁC SUẤT THỐNG KÊ

THỐNG KÊ MÔ TẢ

	Tổng thể (Population)	Mẫu (Sample)
Kích thước (size)	N	n
Liệt kê giá trị	$(x_1, x_2, ..., x_N)$	$(x_1, x_2, ..., x_n)$
Trung bình (mean)	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Phương sai (variance)	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Độ lệch chuẩn (standard deviation)	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$
Hệ số biến thiên (Coef. of variation)	$CV = \frac{\sigma}{\mu} \cdot 100\%$	$CV = \frac{s}{\bar{x}} \cdot 100\%$
Tứ phân vị (Quartile)		Q_1, Q_2, Q_3
Khoảng tứ phân vị (Interquartile Range)		$IQR = Q_3 - Q_1$
Giá trị chuẩn hóa (Z-score)	$z_i = \frac{x_i - \mu}{\sigma}$	$z_i = \frac{x_i - \bar{x}}{s}$
Hệ số bất đối xứng (Skewness)		$a_3 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3 / n}{s^3}$
Hệ số nhọn (Kurtorsis)		$a_4 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{s^4}$ $Kurt = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4 / n}{s^4} - 3$
Hiệp phương sai (Covariance)	$Cov(X,Y) = \frac{\sum_{i=1}^{N} (x_i - \mu_X)(y$	$cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$
Hệ số tương quan (Correlation coef.)	$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$	$r(X,Y) = \frac{cov(X,Y)}{s_X s_Y}$

CÁC CÔNG THỨC XÁC SUẤT

Xác suất theo định nghĩa cổ điển (Classical definition)	$P(A) = \frac{N_A}{N}$
Xác suất theo định nghĩa thống kê (Statistical definition)	$P(A) \approx \frac{f_A}{n}$ khi $n \to \infty$
Xác suất hai biến cố đối lập (Prob. of complement events)	$P(\bar{A}) + P(A) = 1$
Xác suất tích hai biến cố (Prob. of intersection)	$P(A \cdot B) = P(A) \cdot P(B\|A) = P(B) \cdot P(A\|B)$
Xác suất có điều kiện (Conditional probability)	$P(A\|B) = \frac{P(A \cdot B)}{P(B)}$
Hai biến cố độc lập (Independent events)	$P(A\|B) = P(B)$ và $P(B\|A) = P(B)$ $P(A \cdot B) = P(A) \cdot P(B)$
Nhiều biến cố độc lập toàn phần (Totally independent events)	$P\left(\prod_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i)$
Xác suất tổng hai biến cố (Prob. of union)	$P(A + B) = P(A) + P(B) - P(A \cdot B)$
Hai biến cố xung khắc (Mutually exclusive events)	$P(A + B) = P(A) + P(B)$
Nhiều biến cố xung khắc (Mutually exclusive events)	$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$
Công thức xác suất đầy đủ (Total probability)	$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B\|A_i)$
Công thức Bayes (Bayes's theorem)	$P(A_i\|B) = \frac{P(B \cdot A_i)}{P(B)} = \frac{P(A_i) \cdot P(B\|A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(B\|A_i)}$

Bảng phân phối xác suất của BNN rời rạc	X | x_1 | x_2 | ... | x_n $P(X)$ | p_1 | p_2 | ... | p_n $\sum_{i=1}^{n} p_i = 1$
Hàm phân phối xác suất	$F(x) = P(X < x)$ $P(a \le X < b) = F(b) - F(a)$
Hàm mật độ xác suất của BNN liên tục	$f(x) = F'(x)$ $\int_{-\infty}^{+\infty} f(x)dx = 1$ $P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a)$
Kỳ vọng	
Phương sai	$V(X) = E\left[X - E(X)\right]^2 = E(X^2) - \left[E(X)\right]^2$
Độ lệch chuẩn	$\sigma = \sqrt{V(X)}$
Mốt	

BIẾN NGẪU NHIÊN

Biến ngẫu nhiên hai chiều rời rạc

$X \setminus Y$	y_1	y_2	...	y_m	$\sum = P(X)$
x_1	p_{11}	p_{12}	...	p_{1m}	$P(x_1)$
x_2	p_{21}	p_{22}	...	p_{2m}	$P(x_2)$
⋮	⋮	⋮	⋱	⋮	⋮
x_n	p_{n1}	p_{n2}	...	p_{nm}	$P(x_n)$
$\sum = P(Y)$	$P(y_1)$	$P(y_2)$	...	$P(y_m)$	1

<table>
<tr><td>Hiệp phương sai</td><td colspan="2">$$Cov(X,Y) = E\Big[\big[X - E(X)\big]\big[Y - E(Y)\big]\Big]$$
$$= E(X \cdot Y) - E(X) \cdot E(Y) = \sum_i \sum_j x_i y_j p_{ij} - E(X) \cdot E(Y)$$</td></tr>
<tr><td>Hệ số tương quan</td><td colspan="2">$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$</td></tr>
<tr><td>Nếu X, Y độc lập</td><td colspan="2">$$Cov(X,Y) = \rho(X,Y) = 0$$</td></tr>
<tr><td rowspan="5">Tính chất của kì vọng, phương sai
Với c là hằng số</td><td>Kì vọng</td><td>Phương sai</td></tr>
<tr><td>$E(c) = c$</td><td>$V(c) = 0$</td></tr>
<tr><td>$E(X + c) = E(X) + c$</td><td>$V(X + c) = V(X)$</td></tr>
<tr><td>$E(c \cdot X) = c \cdot E(X)$</td><td>$V(c \cdot X) = c^2 \cdot V(X)$</td></tr>
<tr><td>$E(X \pm Y) = E(X) \pm E(Y)$</td><td>$V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X,Y$</td></tr>
<tr><td></td><td>$E(\sum X_i) = \sum E(X_i)$</td><td>$V(\sum X_i) = \sum V(X_i)$ nếu các X_i độc lập</td></tr>
</table>

PHÂN PHỐI XÁC SUẤT THÔNG DỤNG

Phân phối Không-một Bernoulli: $A(p)$	Công thức tính xác suất	$P(X=x)=p^x(1-p)^{1-x} \quad ; \quad x=0,1$
	Tham số	$E(X)=p \quad ; \quad V(X)=p(1-p)$
Phân phối Nhị thức Binomial: $B(n,p)$	Công thức tính xác suất	$P(X=x)=C_n^x p^x(1-p)^{n-x} \quad ; \quad x=0,1,2,...,n$
	Tham số	$E(X)=np \quad ; \quad V(X)=np(1-p)$
Phân phối Poisson $P(\lambda)$	Công thức tính xác suất	$P(X=x)=\dfrac{\lambda^x e^{-\lambda}}{x!} \quad ; \quad x=0,1,2,...$
	Tham số	$E(X)=\lambda \quad ; \quad V(X)=\lambda$
Phân phối Đều Uniform: $U(a,b)$	Hàm mật độ	$f(x)=\begin{cases}\dfrac{1}{b-a} & : x\in(a,b)\\ 0 & : x\notin(a,b)\end{cases}$
	Tham số	$E(X)=\dfrac{a+b}{2} \quad ; \quad V(X)=\dfrac{(b-a)^2}{12}$
Phân phối Chuẩn Normal: $N(\mu,\sigma^2)$	Hàm mật độ	$f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad ; \quad x\in\mathbb{R}$
	Tham số	$E(X)=\mu \quad ; \quad V(X)=\sigma^2$
	Chuẩn hóa	$f(z)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \quad ; \quad z\in\mathbb{R}$
	Công thức xác suất	$P(a<X<b)=P\left(\dfrac{a-\mu}{\sigma}<Z<\dfrac{b-\mu}{\sigma}\right)$ $P(\lvert X-\mu\rvert<\varepsilon)=2\cdot P\left(Z<\dfrac{\varepsilon}{\sigma}\right)$
	Quy tắc	$P(\mu-\sigma<X<\mu+\sigma)=0.6826$ $P(\mu-2\sigma<X<\mu+2\sigma)=0.9544$ $P(\mu-3\sigma<X<\mu+3\sigma)=0.9974$
	Giá trị tới hạn	$z_\alpha : P(Z>z_\alpha)=\alpha$
Phân phối Khi-bình phương Chi-squared: $\chi^2(n)$	Giá trị tới hạn	$\chi_\alpha^{2(n)} : P\left[\chi^2(n)>\chi_\alpha^{2(n)}\right]=\alpha$
Phân phối Student $T(n)$	Giá trị tới hạn	$t_\alpha^{(n)} : P\left[T(n)>t_\alpha^{(n)}\right]=\alpha$
Phân phối Fisher	Giá trị tới hạn	$f_\alpha^{(n_1,n_2)} : P\left[F(n_1,n_2)>f_\alpha^{(n_1,n_2)}\right]=\alpha$

$F(n_1, n_2)$		

MẪU NGẪU NHIÊN

Mẫu kích thước n	$W_n = (X_1, X_2, ..., X_n)$	
Trung bình mẫu (sample mean)	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ $E(\bar{X}) = \mu$; $V(\bar{X}) = \frac{\sigma^2}{n}$	$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$; $\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T^{(n-1)}$ khi $X \sim N(\mu, \sigma^2)$ hoặc khi n đủ lớn
Phương sai mẫu (sample variance)	$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ $E(S^2) = \sigma^2$	$\frac{(n-1)S^2}{\sigma^2} \sim \chi^{2(n-1)}$ khi $X \sim N(\mu, \sigma^2)$
Tần suất mẫu (sample proportion)	$\hat{p} = \frac{X_A}{n}$ $E(\hat{p}) = p$; $V(\hat{p}) = \frac{p(1-p)}{n}$	$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ khi n đủ lớn
Hiệp phương sai mẫu (sample covariance)	$Cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$	
Hệ số tương quan mẫu (sample correlation)	$R_{X,Y} = \frac{Cov(X, Y)}{S_X S_Y}$	

ƯỚC LƯỢNG ĐIỂM

Tính chất ước lượng điểm	Không chệch (unbiasness)	$E(\hat{\theta}) = \theta$
	Hiệu quả (efficient)	không chệch và $V(\hat{\theta})$ nhỏ nhất
Ước lượng hợp lý tối đa (maximum likelihood estimator)	Hàm hợp lý	$L(\theta) = \begin{cases} \prod_i P(x_i) & : \text{discrete} \\ \prod_i f(x_i) & : \text{continous} \end{cases}$
	Tối đa hóa hàm hợp lý hoặc logarit hàm hợp lý	$L(\theta) \to max$ hoặc $\ln L(\theta) \to max$

KHOẢNG TIN CẬY (Confidence Interval)

Trung bình tổng thể khi không biết σ	Hai phía	$\bar{X} - t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}^{(n-1)}\frac{S}{\sqrt{n}}$ hay $\bar{X} \pm \varepsilon$	$n = \left(t_{\alpha/2}^{(n-1)}\frac{S}{\varepsilon}\right)^2$
	Tối đa	$\mu < \bar{X} + t_{\alpha}^{(n-1)}\frac{S}{\sqrt{n}}$	
	Tối thiểu	$\bar{X} - t_{\alpha}^{(n-1)}\frac{S}{\sqrt{n}} < \mu$	
TB tổng thể khi biết σ	Hai phía	$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$n = \left(z_{\alpha/2}\frac{\sigma}{\varepsilon}\right)^2$
Phương sai tổng thể	Hai phía	$\frac{(n-1)S^2}{\chi^{2(n-1)}_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^{2(n-1)}_{1-\alpha/2}}$	
Tần suất tổng thể	Hai phía	$\hat{p} - z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < p < \hat{p} + z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ hay $\hat{p} \pm \varepsilon$	$n = z_{\alpha/2}^2\frac{\hat{p}(1-\hat{p})}{\varepsilon^2}$

KIỂM ĐỊNH GIẢ THUYẾT VỀ THAM SỐ (Parametric Hypothesis Testing)

Kiểm định một tham số, một tổng thể, một mẫu

Kiểm đinh	Giả thuyết gốc Thống kê	Giả thuyết đối	Miền bác bỏ
Trung bình tổng thể phân phối chuẩn, biết phương sai tổng thể	$H_0 : \mu_1 = \mu_2$ $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	$H_1 : \mu \neq \mu_0$	$\|Z\| > z_{\alpha/2}$
		$H_1 : \mu > \mu_0$	$Z > z_\alpha$
		$H_1 : \mu < \mu_0$	$Z < -z_\alpha$
Trung bình tổng thể phân phối chuẩn, không biết phương sai tổng thể	$H_0 : \mu_1 = \mu_2$ $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$	$H_1 : \mu \neq \mu_0$	$\|T\| > t_{\alpha/2}^{(n-1)}$
		$H_1 : \mu > \mu_0$	$T > t_\alpha^{(n-1)}$
		$H_1 : \mu < \mu_0$	$T < -t_\alpha^{(n-1)}$
Phương sai tổng thể phân phối chuẩn	$H_0 : \sigma^2 = \sigma_0^2$ $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$H_1 : \sigma^2 \neq \sigma_0^2$	$\chi^2 > \chi_{\alpha/2}^{2(n-1)}$ hoặc $\chi^2 < \chi_{1-\alpha/2}^{2(n-1)}$
		$H_1 : \sigma^2 > \sigma_0^2$	$\chi^2 > \chi_\alpha^{2(n-1)}$
		$H_1 : \sigma^2 < \sigma_0^2$	$\chi^2 < \chi_{1-\alpha}^{2(n-1)}$
Tần suất tổng thể	$H_0 : p = p_0$ $Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$	$H_1 : p \neq p_0$	$\|Z\| > z_{\alpha/2}$
		$H_1 : p > p_0$	$Z > z_\alpha$
		$H_1 : p < p_0$	$Z < -z_\alpha$

Kiểm định hai tham số, hai tổng thể, hai mẫu

Kiểm định	Giả thuyết gốc Thống kê	Giả thuyết đối	Miền bác bỏ
Hai trung bình tổng thể phân phối chuẩn, giả sử phương sai bằng nhau	$H_0 : \mu_1 = \mu_2$ $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$	$H_1 : \mu_1 \neq \mu_2$	$\lvert T\rvert > t_{\alpha/2}^{(n_1+n_2-2)}$
		$H_1 : \mu_1 > \mu_2$	$T > t_{\alpha}^{(n_1+n_2-2)}$
		$H_1 : \mu_1 < \mu_2$	$T < -t_{\alpha}^{(n_1+n_2-2)}$
Hai trung bình tổng thể phân phối chuẩn, giả sử phương sai khác nhau	$H_0 : \mu_1 = \mu_2$ $T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ $n_1 > 30, n_2 > 30$	$H_1 : \mu_1 \neq \mu_2$	$\lvert T\rvert > z_{\alpha/2}$
		$H_1 : \mu_1 > \mu_2$	$T > z_{\alpha}$
		$H_1 : \mu_1 < \mu_2$	$T < -z_{\alpha}$
Hai phương sai tổng thể phân phối chuẩn	$H_0 : \sigma_1^2 = \sigma_2^2$ $F = \dfrac{S_1^2}{S_2^2}$	$H_1 : \sigma_1^2 \neq \sigma_2^2$	$F > f_{\alpha/2}^{(n_1-1,n_2-1)}$ hoặc $F < f_{1-\alpha/2}^{(n_1-1,n_2-1)}$
		$H_1 : \sigma_1^2 > \sigma_2^2$	$F > f_{\alpha}^{(n_1-1,n_2-1)}$
		$H_1 : \sigma_1^2 < \sigma_2^2$	$F < f_{1-\alpha}^{(n_1-1,n_2-1)}$
Hai tần suất tổng thể	$H_0 : p_1 = p_2$ $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $\bar{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2}$	$H_1 : p_1 \neq p_2$	$\lvert Z\rvert > z_{\alpha/2}$
		$H_1 : p_1 > p_2$	$Z > z_{\alpha}$
		$H_1 : p_1 < p_2$	$Z < -z_{\alpha}$

KIỂM ĐỊNH PHI THAM SỐ (Non-parametric Testing)

	Thống kê	Cặp giả thuyết	Miền bác bỏ
Kiểm định tính độc lập của hai dấu hiệu định tính	$\chi^2 = n\left[\sum_i \sum_j \frac{n_i m_j}{n_{ij}^2} - 1\right]$	H_0 : hai dấu hiệu độc lập H_1 : hai dấu hiệu không độc lập	$\chi^2 > \chi_\alpha^{2((h-1)(k-1))}$
Jacque-Berra Kiểm định tính phân phối chuẩn	$\chi^2 = n\left[\frac{Skew^2}{6} + \frac{K^2}{24}\right]$	H_0 : biến phân phối chuẩn H_1 : biến không phân phối chuẩn	$\chi^2 > \chi_\alpha^{2(2)}$