X	Discrete random variable	Y	Continuous random variable	n	Sample size
$p; \hat{p}$	Probability; Sample probability	$q; \hat{q}$	Complementary probability: $1-p$; $1-\hat{p}$	V_p	p^{th} percentile
$E[X]/\mu$	Population mean/ Expected value	$\bar{x}$	Sample mean/ Arithmetic mean	λ	$\frac{d}{dt}$ (expected number of events)
SD/σ	Population standard deviation	s	Sample standard deviation	Cov	Covariance of dependent variables X, Y
Var/σ^2	Population variance $(\frac{1}{n})$	s^2	Sample variance $(\frac{1}{n-1})$	$Corr/\rho$	Correlation between X, Y: $\rho \in [-1,1]$
PMF	Probability mass function: $P(X=k)$	CDF/F	Cumulative distribution function: $P(X \le x)$	PDF	Probability density function: $f(y)$
H_0	Null hypothesis: hypothesis to be tested	α	Significance level: $P(\text{reject } H_0 \mid H_0 \text{ is true})$	Type I error	$P(\text{reject } H_0 \mid H_0 \text{ is true})$
H_1	Alternative hypothesis: contradicts H_0	$1-\beta$	Power: $P(\text{reject } H_0 \mid H_0 \text{ is false})$	Type II error	$P(\text{accept } H_0 \mid H_0 \text{ is false}) = \beta$

Standard probability stuff

$P(B\|A) = \dfrac{P(A\|B) \times P(B)}{P(A\|B) \times P(B) + P(A\|\bar{B}) \times P(\bar{B})}$	$\sigma^2 = E[X^2] - \mu^2$	$Bin(n,p); n \ge 100; p \le 0.01 \to Po(np)$
	$Po(\lambda); \lambda \ge 10 \to N(\lambda, \lambda)$	$Bin(n,p); npq \ge 5 \to N(np, npq)$
Outlying values: $x < V_{25} - 1.5(V_{75} - V_{25}) \mid V_{75} + 1.5(V_{75} - V_{25}) < x$		
$n = \frac{\sigma^2}{(\mu_0 - \mu)^2}\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2 *$	power $= 1 - \beta = \Phi\left(\frac{\|\mu-\mu_0\|\sqrt{n}}{\sigma} - z_{1-\alpha}\right) = \Phi\left(\frac{(\mu_0-\mu)\sqrt{n}}{\sigma} - z_{1-\frac{\alpha}{2}}\right) + \Phi\left(\frac{(\mu-\mu_0)\sqrt{n}}{\sigma} - z_{1-\frac{\alpha}{2}}\right) *$	

Binomial distribution stuff - $Bin(n,p)$

$PMF = P(X=k) = C_k^n p^k q^{n-k}$	$P(a \le X \le b) = P(a - 0.5 \le Z \le b + 0.5)$	$\mu = np$	$Var[X] = npq$

Continuous distribution stuff

$P(a < b) = P(a \le b)$	$P(a \le Y < b) = \int_a^b f(y)\,dy$	$\mu = \int_{-\infty}^{\infty} y f(y)\,dy$	$Var[Y] = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy$
$Cov(X,Y) = E\left[(X - \mu_x)(Y - \mu_y)\right] = E[XY] - \mu_x \mu_y$			$\rho = \frac{1}{\sigma_x \sigma_y} Cov(X,Y)$

Gaussian/Normal distribution stuff - $N(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}; E[Y] = \mu; Var[Y] = \sigma^2$$

Standard normal distribution stuff - $N(0,1)$: $\mu = 0; \sigma^2 = 1$

$z_u = V_{100u}; u \in [0,1]$	$P(-z_u < Z < z_u) = u$
$CDF_{N(0,1)}(x) = P(Z \le z_u) = \frac{1 + P(-z_u \le Z \le z_u)}{2} = \frac{1+u}{2} = \Phi(z_u) = 1 - \Phi(-z_u) *$	
Standardisation: $N(0,1) = \frac{N(\mu,\sigma^2) - \mu}{\sigma}$	$P(a < Y < b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) *$

Confidence interval stuff

${\sigma^2}_{100(1-\alpha)\%\text{ CI}} = \left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}*}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}*}}\right)$	① Check $n\hat{p}\hat{q} \ge 5$ ② $p_{100(1-\alpha)\%\text{ CI}} = \hat{p} \pm \sqrt{\frac{\hat{p}\hat{q}}{n}}\, z_{1-\frac{\alpha}{2}}$	$\boldsymbol{\mu}_{\mathbf{100(1-\alpha)\%\ CI}} \in \text{CI}_{tz}$; $\boldsymbol{m} = \bar{x}$; $\boldsymbol{d} = n - 1$ $\bar{\boldsymbol{\Delta}}_{\mathbf{100(1-\alpha)\%\ CI}} \in \text{CI}_{tz}$; $\boldsymbol{m} = \bar{x}$; $\boldsymbol{d} = n - 1$
$\mathbf{CI}_{tz} = \begin{cases} \boldsymbol{m} \pm \frac{s}{\sqrt{n}} t_{\boldsymbol{d},1-\frac{\alpha}{2}} *, & n \le 200 \\ \boldsymbol{m} \pm \frac{s}{\sqrt{n}} z_{1-\frac{\alpha}{2}} *, & n > 200 \end{cases}$	$\boldsymbol{\rho}_{\mathbf{100(1-\alpha)\%\ CI}} \in \left(\frac{e^{2z_1}-1}{e^{2z_1}+1}, \frac{e^{2z_2}-1}{e^{2z_2}+1}\right)$; $z_{1,2} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \mp \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}$; $r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$	
$\|\boldsymbol{\mu_1} - \boldsymbol{\mu_2}\|_{\mathbf{100(1-\alpha)\%\ CI}} \in \text{CI}_t$; $\boldsymbol{m} = \|\overline{x_1} - \overline{x_2}\|$; $\boldsymbol{d} = n_1 + n_2 - 2$; $\frac{\mathbf{1}}{\boldsymbol{n'}} = \frac{1}{n_1} + \frac{1}{n_2}$; $\boldsymbol{s'} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$		

Hypothesis stuff

μ hypothesis (H_0: $\mu = \mu_0$; H_1: $\mu <\neq> \mu_0$) ① Unknown σ ② Find α_p' from $t_{obs} = \frac{\sqrt{n}}{s}(\bar{x} - \mu_0) = t_{n-1,\alpha_p'} *$	**μ hypothesis (H_0: $\mu = \mu_0$; H_1: $\mu <\neq> \mu_0$)** ① Known σ ② Find α_p' from $z_{obs} = \frac{\sqrt{n}}{\sigma}(\bar{x} - \mu_0) = z_{\alpha_p'} *$
σ hypothesis (H_0: $\sigma = \sigma_0$; H_1: $\sigma <\neq> \sigma_0$) ① ${\chi^2}_{obs} = \frac{s^2}{\sigma_0^2}(n-1)$ ② Find α_p' from ${\chi^2}_{obs} = {\chi^2}_{n-1,\alpha_p'} *$ ③ <u>Accept</u> H_0 if ${\chi^2}_{obs} \in \left[{\chi^2}_{n-1,\alpha_c\|\frac{\alpha_c}{2}}, {\chi^2}_{n-1,1-\alpha_c\|1-\frac{\alpha_c}{2}}\right]$	**p hypothesis (H_0: $p = p_0$; H_1: $p <\neq> p_0$)** ① Check normal approximation: $np_0q_0 \ge 5$ ② $z_{obs} = \sqrt{\frac{n}{p_0q_0}}(\hat{p} - p_0)$ ③ Find α_p' from $z_{obs} = z_{\alpha_p'} *$
Δ hypothesis (H_0: $\Delta = \Delta_0$; H_1: $\Delta <\neq> \Delta_0$) ① Set $\Delta_0 = 0$ (usually) ② $t_{obs} = \frac{\sqrt{n}}{s_d}(\bar{d} - \Delta_0)$ ③ Find α_p' from $t_{obs} = t_{n-1,\alpha_p'} *$	**α_p' interpretation** ($\alpha_{\text{cutoff}} = 0.05$) ① $\alpha_p = \begin{cases} 1 - \alpha_p', & \text{one tailed} \\ 2(1 - \alpha_p'), & \text{two tailed} \end{cases}$ ② $\begin{cases} \text{Accept } H_0, & \alpha_p > 0.05, \quad \text{(not significant)} \\ \text{Reject } H_0, & \alpha_p < 0.05, \quad \text{(significant)} \end{cases}$

σ_1 σ_2 hypothesis ($H_0: \sigma_1 = \sigma_2; H_1: \sigma_1 \neq \sigma_2$)

① Assume normal distribution & independent

② $F_{obs} = \frac{s_1^2}{s_2^2}$

③ Find α_p from $F_{obs} = F_{n_1-1,n_2-1,\alpha_p}$*

④ Accept H_0 if $F_{obs} \in \left[F_{n_1-1,n_2-1,\frac{\alpha_c}{2}}, \frac{1}{F_{n_2-1,n_1-1,\frac{\alpha_c}{2}}}\right]$

μ_1 μ_2 hypothesis ($H_0: \mu_1 = \mu_2; H_1: \mu_1 <\neq> \mu_2$)

① Test/Assume $\sigma_1 = \sigma_2$

② $\frac{1}{n'} = \frac{1}{n_1} + \frac{1}{n_2}; s' = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

③ $t_{obs} = \frac{\sqrt{n'}}{s'}(\overline{x_1} - \overline{x_2})$

④ Find α_p' from $t_{obs} = t_{n_1+n_2-2,\alpha_p'}$*

μ_1 μ_2 hypothesis ($H_0: \mu_1 = \mu_2; H_1: \mu_1 <\neq> \mu_2$)

① Test/Assume $\sigma_1 \neq \sigma_2$

② $\begin{cases} z_{obs} = \frac{1}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}(\overline{x_1} - \overline{x_2}), & \sigma \text{ known} \\ t_{obs} = \frac{1}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}(\overline{x_1} - \overline{x_2}), & \sigma \text{ unknown} \end{cases}$

③ Find α_p' from $\begin{cases} z_{obs} = z_{\alpha_p'} \text{*}, & \sigma \text{ known} \\ t_{obs} = t_{d'',\alpha_p'} \text{*}, & \sigma \text{ unknown} \end{cases}$;

$$d'' = \left\lfloor \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right)} \right\rfloor$$

Regression hypothesis ($H_0: \beta = 0; H_1: \beta \neq 0$)

① Reject H_0 if $F_{obs} = \frac{Reg}{Res_{MS}} > F_{1,n-2,1-\alpha}$*

Regression hypothesis ($H_0: \beta = 0; H_1: \beta \neq 0$)

① Find α_p' from $t_{obs} = b\sqrt{\frac{L_{xx}}{Res_{MS}}} = t_{n-2,\alpha_p'}$*

Correlation hypothesis ($H_0: \rho = 0; H_1: \rho \neq 0$)

① Find α_p' from $t_{obs} = r\sqrt{\frac{n-2}{1-r^2}} = t_{n-2,\alpha_p'}$*; $r = \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$

Correlation hypothesis ($H_0: \rho = \rho_0; H_1: \rho \neq \rho_0$)

① Find α_p' from $\lambda = \frac{\sqrt{n-3}}{2}\left(\ln\frac{1+r}{1-r} - \ln\frac{1+\rho_0}{1-\rho_0}\right) = a_{\alpha_p'}$*

Non-parametric stuff

Sign test ($H_0: \Delta = 0; H_1: \Delta \neq 0$)

① Let C = number of data points with $d_i > 0$

② $z_{obs} = \frac{\sqrt{n}}{2}\left(C - \frac{n}{2}\right)$

③ Find α_p' from $z_{obs} = z_{\alpha_p'}$*

Signed-rank test ($H_0: \Delta = 0; H_1: \Delta \neq 0$)

① Rank d_i by ascending $|d_i|$ from 1 to n; ignore $d_i = 0$

② Let R = sum of ranks for data points $d_i > 0$

③ Find α_p' from $T = \frac{2\left|R-\frac{1}{4}n(n+1)\right|}{\sqrt{\sum_{i=1}^{n} rank_i^2}} \approx \frac{2\sqrt{6}\left|R-\frac{1}{4}n(n+1)\right|}{\sqrt{n(n+1)(2n+1)}} = z_{\alpha_p'}$*

Rank-sum test ($H_0: F_1(x) = F_2(x); H_1: F_1(x) = F_2(x - \Delta)$)

① Rank d_i by ascending $|d_i|$ of both samples from 1 to $n_1 + n_2$; ignore $d_i = 0$

② Let R_1 = sum of ranks for data points $d_i > 0$ for one sample

③ Find α_p' from $T = \sqrt{\frac{(n_1+n_2)(n_1+n_2+1)}{n_1 n_2}} \frac{\left|R_1-\frac{n_1(n_1+n_2+1)}{2}\right|}{\sqrt{\sum_{i=1}^{n_1+n_2}\left(rank_i-\frac{n_1+n_2+1}{2}\right)^2}} \approx \frac{2\sqrt{3}\left|R_1-\frac{n_1(n_1+n_2+1)}{2}\right|}{\sqrt{n_1 n_2(n_1+n_2+1)}} = z_{\alpha_p'}$*

Contingency table stuff

Contingency table test (H_0: factors are independent; H_1: factors are associated)

①
$$\begin{array}{c|ccccc} n_{i,j} & \boldsymbol{f_{2,1}} & & \boldsymbol{f_{2,j}} & & \boldsymbol{\Sigma_{i+}} \\ \hline \boldsymbol{f_{1,1}} & n_{1,1} & \cdots & n_{1,j} & \rightarrow & n_{1+} \\ \vdots & \vdots & \ddots & \vdots & \rightarrow & \vdots \\ \boldsymbol{f_{1,i}} & n_{i,1} & \cdots & n_{i,j} & \rightarrow & n_{i+} \\ & \downarrow & \downarrow & \downarrow & \searrow & \\ \boldsymbol{\Sigma_{+j}} & n_{+1} & \cdots & n_{+j} & & \boldsymbol{n} \end{array}$$

② Compute $E_{i,j} = \frac{1}{n}\left(n_{i+}\, n_{+j}\right)$ table

③ Compute $\chi_{i,j}^2 = \frac{1}{E_{i,j}}\left(E_{i,j} - n_{i,j}\right)^2$ table

④ Reject H_0 if $\sum\chi^2 > \chi^2_{(R-1)(C-1),1-\alpha}$*

Fisher's exact test (H_0: factors are independent; H_1: factors are associated)

① Follow contingency table test, check that ≥ 20% of $E_{i,j} < 5$ or any $E_{i,j} < 1$

② Rearrange table such that $\Sigma_{1+} < \Sigma_{2+}$ and $\Sigma_{+1} < \Sigma_{+2}$

③ Enumerate tables $\begin{bmatrix} 0 & b+a \\ c+a & d-a \end{bmatrix} \cdots \begin{bmatrix} a-1 & b+1 \\ c+1 & d-1 \end{bmatrix}, \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \begin{bmatrix} a+1 & b-1 \\ c-1 & d+1 \end{bmatrix} \cdots \begin{bmatrix} a+c & b-c \\ 0 & d+c \end{bmatrix}$

④ Find $P(\text{Table } n) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$, compare p_n with hypothesis

Linear regression stuff

$E[Y\|X = x] = f(x) = \alpha + \beta x$	$a = \bar{y} - b\bar{x}$	$b = \frac{L_{xy}}{L_{xx}}$	$L_{xx} = -n\bar{x}^2 + \sum_{i=1}^{n} x_i^2$	$L_{yy} = -n\bar{y}^2 + \sum_{i=1}^{n} y_i^2$
$Reg = \frac{L_{xy}^2}{L_{xx}}$	$Res_{MS} = \frac{1}{n-2}\left(L_{yy} - Reg\right)$	$\sigma^2 = s_{y.x}^2 = Res_{MS}$	$L_{xy} = -n\bar{x}\bar{y} + \sum_{i=1}^{n} x_i y_i$	
$R^2 = \frac{Reg_{MS}}{L_{yy}} = \frac{L_{xy}^2}{L_{xx}L_{yy}}$	$Adjusted\ R^2 = 1 - \frac{s_{y.x}^2}{s_y^2}$	$s_x^2 = \frac{1}{n-1}L_{xx}$	$s_y^2 = \frac{1}{n-1}L_{yy}$	