sample variance $S_X^2 = \frac{\sum_{\text{all } x}(x-\bar{x})^2}{n-1}$

Joint pmf
$p(x,y) = P(X=x, Y=y)$ for all x,y: $0 \le p(x,y) \le 1$, $\sum_{\text{all } x}\sum_{\text{all } y} p(x,y) = 1$, $P((X,Y)\in A) = \sum\sum_{(x,y)\in A} p(x,y)$

$p_X(x) = \sum_{\text{all } y} p(x,y)$ $p_Y(y) = \sum_{\text{all } x} p(x,y)$ $E(h(X,Y)) = \sum_{\text{all } x}\sum_{\text{all } y} h(x,y)p(x,y)$ $E(X+Y) = E(X)+E(Y)$

Two discrete rvs X and Y are independent
$\Leftrightarrow \forall x,y,\ P(X=x, Y=y) = P(X=x)P(Y=y)$
$\Leftrightarrow \forall x,y,\ p_{X,Y}(x,y) = p_X(x)p_Y(y)$

conditional pmf
$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$ for $p_Y(y) > 0$ $P(X=x|Y=y) = \frac{P(X=x \cap Y=y)}{P(Y=y)}$ for $P(Y=y)>0$

Covariance def $Cov(X,Y) = E[(X-E(X))(Y-E(Y))] = \sigma_{XY}$
properties $Cov(X,Y) = E(XY) - E(X)E(Y)$ $Cov(X,X) = Var(X)$
$Cov(aX+b, cY+d) = ac\,Cov(X,Y)$
$Var(X+Y) = Var(X)+Var(Y)+Cov(X,Y)$

$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x)$

Correlation def $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} = Corr(X,Y) = \rho_{XY}$
Weak $|\rho| \le 0.5$
Moderate $0.5 < |\rho| \le 0.8$
Strong $|\rho| > 0.8$

properties $-1 \le \rho \le 1$
$Corr(aX+b, cY+d) = Corr(X,Y)$

distribution of sample mean $X_1 \dots X_n \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
$\bar{x}$ — sample mean S^2 — sample variance
μ — population mean σ^2 — population variance
$E(\bar{X}) = \mu$ $Var(\bar{X}) = \frac{\sigma^2}{n}$ $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
central limit theorem
Let $X_1, \dots, X_n$ be iid from a distribution with mean μ and variance σ^2.
For $n \ge 30$, $\bar{X}$ is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

Confidence Intervals
Use tree $\alpha = 1 - CL$
$n = \left(\frac{2 z^*_{\alpha/2}\sigma}{w}\right)^2$ round up

CL	α	$z^*_{\alpha/2}$	
90%	0.10	1.645	$= z^*_{0.05}$
95%	0.05	1.96	$= z^*_{0.025}$
99%	0.01	2.58	$= z^*_{0.005}$

$CL\uparrow \Rightarrow w\uparrow$ $n\uparrow \Rightarrow w\downarrow$
$\alpha\downarrow \Rightarrow w\uparrow$ $\sigma\uparrow \Rightarrow w\uparrow$

Proportions
let X be # successes in sample of size n $X \sim Binomial(n, p)$
$E(X) = np$ $Var(X) = np(1-p)$
$\hat{p} = \frac{X}{n}$

score CI for p
$\tilde{p} = \frac{\hat{p} + \frac{(z^*_{\alpha/2})^2}{2n}}{1 + \frac{(z^*_{\alpha/2})^2}{n}}$ the CI is $\tilde{p} \pm z^*_{\alpha/2}\frac{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{(z^*_{\alpha/2})^2}{4n^2}}}{1 + \frac{(z^*_{\alpha/2})^2}{n}}$

traditional CI is bad when $n \le 100$ or p is close to 0 or 1

traditional CI for p $\hat{p} \pm z^*_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

T distribution $T = \frac{\bar{X}-\mu}{\left(\frac{S}{\sqrt{n}}\right)}$

CI for μ
- Normal Dist
  - σ known → $\bar{x} \pm z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
  - σ unknown
    - $n < 40$ → $\bar{x} \pm t_{\frac{\alpha}{2}, n-1}\left(\frac{S}{\sqrt{n}}\right)$
    - $n \ge 40$ → $\bar{x} \pm z^*_{\alpha/2}\frac{S}{\sqrt{n}}$
- Unknown Dist
  - $n < 40$ → ✗
  - $n \ge 40$ → $\bar{x} \pm z^*_{\alpha/2}\left(\frac{\sigma \text{ or } S}{\sqrt{n}}\right)$

$X \sim N$, σ unknown
$\Rightarrow$ use $\bar{x} \pm t_{\frac{\alpha}{2}, n-1}\left(\frac{S}{\sqrt{n}}\right)$
as $1-\alpha$ CI for μ

prediction interval CI for X_{n+1}
$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} S\sqrt{1 + \frac{1}{n}}$

Hypothesis Testing
Normal, σ known → Z test
Any distribution, $n \ge 40$ → Z test
σ unknown, or unknown distribution with small n → T test

1. Define H_0 and H_a
2. Select significance level α
3. Collect data
4. Calculate test statistic and find rejection region
5. Make a decision
6. Write conclusion

errors
Type I — reject H_0 when H_0 is true $P(\text{Type I error}) = \alpha$
Type II — fail to reject H_0 when H_0 is False $P(\text{Type II error}) = \beta$

Z test
H_a:
- $\mu > \mu_0 \to Z \ge z^*_\alpha$
- $\mu < \mu_0 \to Z \le -z^*_\alpha = z_\alpha$
- $\mu \ne \mu_0 \to Z \le -z^*_{\alpha/2} = z_{\frac{\alpha}{2}}$ or $Z \ge z^*_{\alpha/2}$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

T test
H_a:
- $\mu > \mu_0 \to T \ge t_{\alpha, n-1}$
- $\mu < \mu_0 \to T \le -t_{\alpha, n-1}$
- $\mu \ne \mu_0 \to T \le -t_{\frac{\alpha}{2}, n-1}$ or $T \ge t_{\frac{\alpha}{2}, n-1}$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Hypothesis Test for Population Proportions
point estimate $\hat{p} = \frac{X}{n}$ $H_0: p = p_0$
When $np_0 \ge 10$ and $n(1-p_0) \ge 10$ $\Rightarrow \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$
$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ rejection region H_a: $p > p_0 \to Z \ge z^*_\alpha$; $p < p_0 \to Z \le -z^*_\alpha$; H_a: $p \ne p_0 \to Z \le -z^*_{\alpha/2}$ or $Z \ge z^*_{\alpha/2}$

probability density function — $f(x)$

$f(x) \geq 0$ for all $x \in R$

$\int_{-\infty}^{\infty} f(x)\,dx = 1$

$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$

$P(X = a) = 0$

cumulative distribution function — $F(x)$

$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$ for all $x \in R$

$P(X > a) = 1 - F(a)$

$P(a \leq x \leq b) = F(b) - F(a)$ $\quad \int_{-\infty}^{\tilde{\mu}} f(x)\,dx = 0.5$

$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx \quad E(h(x)) = \int_{-\infty}^{\infty} h(x) f(x)\,dx$

$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$

$Var(X) = E(X^2) - [E(x)]^2$

uniform $\quad X \sim U(a,b) \Rightarrow \begin{cases} f(x) = \frac{1}{b-a} \text{ for } a<x<b \\ F(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a<x<b \\ 1 & \text{if } x \geq b \end{cases} \end{cases}$

$E(X) = \frac{a+b}{2} \quad Var(X) = \frac{(b-a)^2}{12}$

normal $\quad X \sim N(\mu, \sigma^2) \Rightarrow$ pdf $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

σ 68.26%
2σ 95.44%
3σ 99.74%

standard normal $\quad Z \sim N(0,1) \quad \Phi(z) = P(Z \leq z) \quad Z = \frac{X-\mu}{\sigma}$ $\quad$ pdf $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ $\quad$ cdf $\Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt$

percentiles $\quad x_p = \mu + \sigma z_p \quad x_p^* = \mu + \sigma z_p^*$

gamma $\quad \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad \Gamma(n) = (n-1)! \quad \Gamma(a) = (a-1)\Gamma(a-1) \quad \Gamma(\tfrac{1}{2}) = \sqrt{\pi}$

$X \sim Gam(\alpha, \beta) \Rightarrow f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$

$\wedge\ E(X) = \alpha\beta \ \wedge\ Var(X) = \alpha\beta^2$

chi-square $\quad X \sim \chi_r^2 \Leftrightarrow X \sim Gam(\alpha = \frac{r}{2}, \beta = 2) \quad r \in Z^+$

exponential $\quad X \sim Exp(\lambda) \Leftrightarrow X \sim Gam(\alpha = 1, \beta = \frac{1}{\lambda}) \Rightarrow f(x; \lambda) = \lambda e^{-\lambda x} \wedge F(x; \lambda) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$

$\wedge\ E(X) = \frac{1}{\lambda} \ \wedge\ Var(X) = \frac{1}{\lambda^2} \Rightarrow \sigma_X = \frac{1}{\lambda}$

P-values with z-test $\quad p = \begin{cases} P(Z \geq z) = 1 - \Phi(z) & \text{upper tailed } (\mu > \mu_0) \\ P(Z \leq z) = \Phi(z) & \text{lower tailed } (\mu < \mu_0) \\ P(Z \geq |z| \text{ or } Z \leq -|z|) = 2(1 - \Phi(|z|)) & \text{two tailed } (\mu \neq \mu_0) \end{cases}$

reject $H_0 \Leftrightarrow p \leq \alpha$