Core Concepts

General Distributions: Discrete and Continuous

Discrete random variable
X: takes countable values.
PDF $p_X(x) = \mathbb{P}(X = x)$, CDF $F_X(x) = \mathbb{P}(X \leqslant x)$.
$\mathbb{E}[X] = \sum_x x\, p_X(x)$.
$\mathbb{V}ar[X] = \sum_x (x^2\, p_X(x)) - [\sum_x x\, p_X(x)]^2$.

Continuous random variable
X: has a probability density function $f_X(x)$ with $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$.
$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$.
$\mathbb{V}\mathrm{ar}[X] = \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx - \left(\int_{-\infty}^{\infty} x\, f_X(x)\, dx\right)^2$

Joint distributions:
CDF:
$F_{XY}(a, b) = \mathbb{P}(X \leqslant a, Y \leqslant b)$
PDF:
$\mathbb{P}(a < X \leqslant b, c < Y \leqslant d) = \int_a^b \int_c^d f_{XY}(x, y) dx dy, \forall a \leqslant b,\ c \leqslant d.$

Conditional distribution
PDF:
$f_{X|Y}(x, y) = f_{X|Y}(x, y) = \frac{f_{XY}(x,y)}{f_Y(y)}$

Conditional CDF given a quantile We know that $X > q_\alpha$
Let q_α be the α-quantile, i.e. $F(q_\alpha) = \alpha$.
$F_\alpha(x) = \mathbb{P}(X < x \mid X > q_\alpha) = \frac{F(x) - F(q_\alpha)}{1 - F(q_\alpha)} \cdot \mathbf{1}_{\{x \geqslant q_\alpha\}}$
$f_\alpha(x) = \frac{f(x)}{1 - F(q_\alpha)} \cdot \mathbf{1}_{\{x \geqslant q_\alpha\}}$

Arrangement and Combinations:

Arrangement (Permutation): Number of ways to choose and order k elements from n distinct objects:
$P_n^k = \frac{n!}{(n-k)!}$ (also written $A(n, k)$ or ${}^nP_k$)
Permutation (Full): Special case when $k = n$:
$n!$ total ways to order n distinct elements.
Combination: Number of ways to choose k elements from n without regard to order:
$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
Key identities:
- $\binom{n}{k} = \binom{n}{n-k}$
- Total number of subsets of size k: $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

Probability rules:

- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
- If A and B are disjoint: $\mathbb{P}(A \cap B) = 0$
- **Conditional probability:** $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, if $\mathbb{P}(B) > 0$
- **Law of Total Probability (Discrete):** $\mathbb{P}(B) = \sum_i \mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i)$ (where $\{A_i\}$ is a partition of the sample space)
- **Bayes' Rule (Discrete):** $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B|A)\, \mathbb{P}(A)}{\mathbb{P}(B)}$
- **Joint Probability Decomposition:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B) \cdot \mathbb{P}(B)$
- **Independence:** $A \perp B \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\, \mathbb{P}(B)$
  $\Rightarrow \mathbb{P}(A \mid B) = \mathbb{P}(A)$
- **Chain rule for multiple events:**
  $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A) \cdot \mathbb{P}(C \mid A, B)$
- **Continuous version (densities):**
  - $f_{U|W}(u \mid w) = \frac{f_{UW}(u,w)}{f_W(w)}$, if $f_W(w) > 0$
  - $f_{UW}(u, w) = f_{U|W}(u \mid w) \cdot f_W(w)$
- *Note: Replace $\mathbb{P}$ with f for densities in the continuous case.*

Expectation:

- $\mathbb{E}[aX + b] = a\, \mathbb{E}[X] + b,$
- $\mathbb{E}\left(\sum_i X_i\right) = \sum_i \mathbb{E}[X_i]$. (Holds regardless of whether X_i are independent)
- **If X and Y are independent:** $\mathbb{E}[XY] = \mathbb{E}[X]\, \mathbb{E}[Y]$.
  More generally, for any functions g, h: $\mathbb{E}(g(X)h(Y)) = \mathbb{E}[g(X)]\, \mathbb{E}[h(Y)]$.
- **Law of Iterated Expectations:** $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]$ $\mathbb{E}(X|Y)$ is a function of Y!, $\mathbb{V}ar(X) < \infty$
- **Jensen's Inequality (for convex/concave g):**
  $g(\mathbb{E}[X]) \leqslant \mathbb{E}[g(X)]$ if g is convex,
  $g(\mathbb{E}[X]) \geqslant \mathbb{E}[g(X)]$ if g is concave.
- **Handling Transformations:** For $Y = g(X)$,
  $\mathbb{E}[Y] = \int g(x)\, f_X(x)\, dx$ (continuous),
  $\mathbb{E}[Y] = \sum_x g(x)\, p_X(x)$ (discrete).

Variance:

- $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
- $\mathrm{Var}(aX + b) = a^2\, \mathrm{Var}(X)$.
- $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j)$.
- **If X and Y are independent:** $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
  More generally, if $\mathrm{Cov}(X, Y) = 0$, then this still holds.
- **Law of Total Variance:** $\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$. $\mathbb{V}ar(X) < \infty$
- **Variance of sample mean:** For $X_1, \ldots, X_n$ i.i.d. with variance σ^2,
  $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathrm{Var}(X_1 + \cdots + X_n) = n\, \sigma^2$.
- **Conditional scaling (e.g. Gaussian case):**
  If $X \mid Y \sim \mathcal{N}(\mu(Y), \sigma^2(Y))$, then
  $\mathrm{Var}(X) = \mathbb{E}[\sigma^2(Y)] + \mathrm{Var}(\mu(Y))$.
- **Population variance:** $\sigma^2 = \mathbb{E}[(X - \mu)^2]$
- **Sample variance (unbiased):** $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
  Unbiased estimator of σ^2 when X_i i.i.d. with finite variance.

Covariance:

- $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- **Properties:**
  - $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
  - $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
  - $\mathrm{Cov}(aX + b, Y) = a\, \mathrm{Cov}(X, Y)$
  - $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$
- **Cauchy-Schwarz Inequality:** $|\mathrm{Cov}(X, Y)| \leqslant \sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}$

Correlation:

- **Correlation coefficient:** $\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\, \mathrm{Var}(Y)}}$
  Measures the linear association between X and Y. $\rho \in [-1, 1]$.
- **Properties:**
  - $\rho_{XY} = \rho_{YX}$
  - $\rho_{X,Y} = 0$ does not imply independence
  - $\rho_{X,Y} = \pm 1$ perfect linear relationship
  - $|\rho_{X,Y}| \leqslant 1$ (from Cauchy-Schwarz inequality)
- **Sample correlation:** $r_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$
- **Partial correlation:** The correlation between residuals of y and z after removing the effect of X:
  $$\mathbb{C}orr(Y, Z|X) = r_{yz \cdot X} = \frac{z^{*\prime} y^*}{\sqrt{z^{*\prime} z^*} \cdot \sqrt{y^{*\prime} y^*}}$$
  where y^* and z^* are residuals from regressing y and z on X, respectively.

Independance:

Two random variables are independent iff the joint c.d.f. of X and Y is given by: $f_{XY}(x, y) = f_X(x) \times f_Y(y)$
or equivalently
$F_{XY}(x, y) = F_X(x) \times F_Y(y)$
The following observations are not sufficient to conclude independance Observations If X and Y are independent :

1. $f_{X|Y}(x, y) = f_X(x)$.
2. $\mathbb{E}(X|Y) = \mathbb{E}(X)$
3. $\mathbb{E}(g(X)|Y) = \mathbb{E}(g(X))$, where g is any function.
4. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$
5. $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$
6. $\mathbb{C}ov(X, Y) = 0$.
7. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Common Discrete Laws

Uniform(a, b):
A random variable X is said to follow a discrete uniform law on $\{a, a+1, \ldots, b\}$ if:
$\mathbb{P}(X = k) = \frac{1}{b-a+1}$ for $k = a, a+1, \ldots, b$.
$\mathbb{E}[X] = \frac{a+b}{2}, \mathrm{Var}(X) = \frac{\left((b-a+1)^2 - 1\right)}{12}$.

Bernoulli(p): 1 experience with 2 possibles outcomes
$\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$
$\mathbb{E}[X] = p$
$\mathrm{Var}(X) = p(1 - p)$.

Binomial(n, p): number of successes in n independent Bernoulli(p) trials.

$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$,
$\mathbb{E}[X] = np$,
$\text{Var}(X) = np(1-p)$.

Geometric(p): number of trials (Bernoulli(p)) needed until first success.
$\mathbb{P}(X = k) = (1-p)^{k-1} p$,
$\mathbb{E}[X] = \frac{1}{p}$,
$\text{Var}(X) = \frac{1-p}{p^2}$.

Hypergeometric(N, K, n): N items total, K successes in population. Draw n items *without replacement*, let X be # of successes drawn.
$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$,
$\mathbb{E}[X] = n\frac{K}{N}$,
$\text{Var}(X) = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$.

Poisson(λ): counts the number of events in fixed time/space if events happen at constant rate λ independently.
$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$,
$\mathbb{E}[X] = \lambda$, $\text{Var}(X) = \lambda$.
If 2 RV's follow a poisson law and are independant, the sum of these 2 RV's will follow a poisson law with $\lambda = \lambda_1 + \lambda_2$

Common Continuous Laws

Uniform(a, b): $X \sim \text{Unif}(a, b)$
$f_X(x) = \frac{1}{b-a} \mathbf{1}_{\{a \leqslant x \leqslant b\}}$,
$\mathbb{E}[X] = \frac{a+b}{2}$,
$\text{Var}(X) = \frac{(b-a)^2}{12}$.

Exponential(λ): $X \sim \text{Exp}(\lambda)$
$f_X(x) = \lambda e^{-\lambda x}$, $x \geqslant 0$,
$\mathbb{E}[X] = \frac{1}{\lambda}$,
$\text{Var}(X) = \frac{1}{\lambda^2}$.
Memoryless property: $\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$.

Normal(μ, σ^2): $X \sim \mathcal{N}(\mu, \sigma^2)$
$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$,
$\mathbb{E}[X] = \mu$,
$\text{Var}(X) = \sigma^2$.
$\psi_3 =$ Skewness $= 0$
$\psi_4 =$ Kurtosis $= 3$

Chi-squared(χ^2_ν): If $Z_i \sim \mathcal{N}(0, 1)$ i.i.d., then $\chi^2_\nu = \sum_{i=1}^{\nu} Z_i^2 \sim \chi^2(\nu)$
$\mathbb{E}[X] = \nu$, $\text{Var}(X) = 2\nu$.

Student(t_ν): $T = \frac{Z}{\sqrt{U/\nu}}$ with $Z \sim \mathcal{N}(0, 1)$ and $U \sim \chi^2_\nu$ independent.
$\mathbb{E}[T] = 0$ (if $\nu > 1$, undefined otherwise),
$\text{Var}(T) = \frac{\nu}{\nu-2}$ (if $\nu > 2$), undefined if $\nu = 1$ and infinite if $\nu = 2$).

Fisher(F_{d_1,d_2}): ratio of scaled chi-squared variables. Often used in ANOVA or regression tests.
If $U_1 \sim \chi^2_{d_1}$, $U_2 \sim \chi^2_{d_2}$ (independent), then

$$F = \frac{\frac{U_1}{d_1}}{\frac{U_2}{d_2}} \sim F_{d_1,d_2}.$$

$$\mathbb{E}[F] = \begin{cases} \dfrac{d_2}{d_2 - 2}, & \text{if } d_2 > 2, \\ \text{undefined}, & \text{if } d_2 \leqslant 2, \end{cases}$$

$$\text{Var}(F) = \begin{cases} \dfrac{2\,d_2^2\,(d_1 + d_2 - 2)}{d_1\,(d_2-2)^2\,(d_2-4)}, & \text{if } d_2 > 4, \\ \text{undefined}, & \text{if } d_2 \leqslant 4. \end{cases}$$

Moments

Central moment of order k: $\mu_k = \mathbb{E}[(Y - \mu)^k]$, where $\mu = \mathbb{E}[Y]$
Standardized moment of order k:
$\psi_k = \frac{\mu_k}{(\mathbb{V}\text{ar}(Y))^{k/2}} = \frac{\mathbb{E}[(Y-\mu)^k]}{(\mathbb{V}\text{ar}(Y))^{k/2}}$

- $\psi_1 = 0$ for any distribution (centered)
- $\psi_2 = 1$ by definition (variance standardized)
- $\psi_3 =$ **Skewness** $= 0$ for symmetric distributions (e.g. Gaussian)
- $\psi_4 =$ **Kurtosis** $= 3$ for Gaussian

Excess kurtosis: $\psi_4 - 3 \rightarrow$ Measures heaviness of tails vs. normal distribution.
Sample central moment of order k: $m_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k$
Standardized sample moment:
$g_k = \frac{m_k}{(m_2)^{k/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{k/2}}$

Law of Large Numbers and Central Limit Theorem

LLN (Law of Large Numbers):
If $X_1, \dots, X_n$ are i.i.d. with mean μ, then: $\bar{X}_n \xrightarrow{p} \mu$
Key Points:

- **Consistency:** $\bar{X}_n$ is a consistent estimator of μ
- **Unbiasedness:** $\mathbb{E}[\bar{X}_n] = \mu$
- **Variance:** $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$

Basic CLT (sample mean):
If $X_i \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$, then:
$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, Var(X_i)) \quad \Rightarrow \quad \bar{X}_n \approx \mathcal{N}\left(\mu, \frac{Var(X_i)}{n}\right)$ for large n
Estimation error shrinks at rate $\sqrt{n}$ (convergence in distribution).

CLT for sample sum: $S_n = \sum_{i=1}^{n} X_i = n\bar{X} \approx \mathcal{N}(n\mu, n\sigma^2)$

CLT for linear combinations: If $a_i \in \mathbb{R}$, and $X_i \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$, then: $\sum_{i=1}^{n} a_i X_i \xrightarrow{d} \mathcal{N}\left(\sum a_i \mu, \sum a_i^2 \sigma^2\right)$

CLT for difference of sample means: If $\bar{X}$ and $\bar{Y}$ are independent: $\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Multivariate CLT:
If $X_i \in \mathbb{R}^m$ i.i.d., with mean $\boldsymbol{\mu}$ and covariance matrix Σ (positive definite), then: $\sqrt{n}(\bar{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma)$

Types of Convergence

1. Convergence in Probability: $X_n \xrightarrow{p} X$

- $\forall \varepsilon > 0$: $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$ as $n \rightarrow \infty$
- Used for consistency (e.g., $\hat{\theta}_n \xrightarrow{p} \theta$)

2. Convergence in Distribution: $X_n \xrightarrow{d} X$

- CDF of X_n converges to CDF of X
- Used in asymptotic approximations (e.g., CLT: $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$)

3. Convergence in Mean Square (L2): $X_n \xrightarrow{L^2} X$

- $\mathbb{E}[(X_n - X)^2] \rightarrow 0$
- Implies convergence in probability

4. Almost Sure Convergence: $X_n \xrightarrow{a.s.} X$

- $\mathbb{P}(\lim_{n\rightarrow\infty} X_n = X) = 1$
- Strongest form of convergence

Difference with Expected Value:

- $\mathbb{E}[X_n] \rightarrow \mathbb{E}[X]$ (not a type of convergence of RVs)
- Convergence in distribution does NOT imply convergence of expectations

Two-Sample Mean Test (CLT-based)

Compare two i.i.d. samples (with known variances:

- $m_1, \dots, m_{n_m} \sim (\mu_m, \sigma_m^2)$
- $w_1, \dots, w_{n_w} \sim (\mu_w, \sigma_w^2)$

Then by CLT: $\bar{m} - \bar{w} \sim \mathcal{N}\left(\mu_m - \mu_w, \sigma^2\right)$, with $\sigma^2 = \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}$
Use this for constructing confidence intervals or performing a two-sided $H_0 : \mu_m = \mu_w$ test.

Two-Sample t-Test (Unequal Variances)
If sample size ¡ 30 : use this, otherwise can be approximated with normal comparison of mean
Test $H_0 : \mu_m = \mu_w$ based on: $\frac{\bar{m} - \bar{w}}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}} \sim t_\nu$ (approx)
with Welch's degrees of freedom: $\nu = \frac{\left(\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}\right)^2}{\frac{(s_m^2/n_m)^2}{n_m - 1} + \frac{(s_w^2/n_w)^2}{n_w - 1}}$

Chebyshev's Inequality

Let X be a random variable and $c \in \mathbb{R}$ (typically the mean or median). If $\mathbb{E}(|X - c|^r) < \infty$ for some $r > 0$, then for all $\varepsilon > 0$:
$\mathbb{P}(|X - c| > \varepsilon) \leqslant \frac{\mathbb{E}(|X-c|^r)}{\varepsilon^r}$
If $r = 2$ and X has finite variance, then:
$\mathbb{P}(|X - \mu| > \varepsilon) \leqslant \frac{\text{Var}(X)}{\varepsilon^2}$
Use: Bounds tail probabilities; proves convergence in probability (LLN).

Markov's Inequality

Let X be a non-negative random variable with $\mathbb{E}[X] < \infty$. Then for any $a > 0$:

$$\boxed{\mathbb{P}(X \geqslant a) \leqslant \frac{\mathbb{E}[X]}{a}}$$

Interpretation: Upper bound on the probability that X exceeds a threshold, using only the mean.

Bias, Constitency:

Bias vs Consistency:

- An estimator $\hat{\theta}$ is **unbiased** for θ if $\mathbb{E}[\hat{\theta}] = \theta$.
- An estimator $\hat{\theta}_n$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.
- Bias is a finite-sample property, while consistency is an asymptotic property.

Bias: $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$
Unbiasedness: $\mathbb{E}[\hat{\theta}] = \theta$
Expected squared error: $\text{MSE}(\hat{\theta}) = (\mathbb{E}[\hat{\theta} - \theta]^2) = \text{Var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$
Consistency: $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$
Good estimator: low variance and low expected bias

Linear Regression

Linear Regression Model (Matrix Form)

$y_i = \beta' x_i + \varepsilon_i \quad$ (for $i = 1, \ldots, n$)
If $x_{i1} = 1$ for all i, then β_1 is the intercept.

Matrix form: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

Key Matrices in Regression

Symbol	Description	Form / Dimensions
y	Outcome vector	$n \times 1$
X	Design matrix	$n \times k$
β	Coefficient vector	$k \times 1$, unknown
b	OLS estimator	$k \times 1$, $b = (X'X)^{-1}X'y$
$\hat{y}$	Predicted values	$n \times 1$, $\hat{y} = Xb$
e	Residuals	$n \times 1$, $e = y - \hat{y}$
ε	Errors	$n \times 1$, $y = X\beta + \varepsilon$
H	Hat matrix	$n \times n$, $H = X(X'X)^{-1}X'$
M	Residual maker	$n \times n$, $M = I_n - H$
M^0	Mean-centering matrix	$M^0 = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$
P	Projection matrix	$P = X(X'X)^{-1}X' = H$
s^2	Estimator of σ^2	$\frac{e'e}{n-k}$
Z	Instrumental variable matrix	$n \times l$, $l \geqslant k$
P_Z	Projection on Z	$P_Z = Z(Z'Z)^{-1}Z'$

OLS Estimation

Objective: minimize residual sum of squares $f(b) = (y - Xb)'(y - Xb) = e'e$

First-order condition:
$\frac{\partial f}{\partial b} = -2X'y + 2X'Xb = 0 \Rightarrow X'Xb = X'y \Rightarrow \boxed{b = (X'X)^{-1}X'y}$

Also: $b = \beta + (X'X)^{-1}X'\varepsilon \quad$ (under assumption 1)

Consistency (as $n \to \infty$):
If assumptions 1–2 hold and $\frac{1}{n}X'X \xrightarrow{p} Q > 0$, then:
$(X'X)^{-1}X'\varepsilon \xrightarrow{p} 0 \quad \Rightarrow \quad b \xrightarrow{p} \beta$ (we can say b=β)

Assumptions:
1. Full rank: $\text{rank}(X) = k$ (no perfect collinearity) x 2. Exogeneity: $\mathbb{E}[\varepsilon_i | X] = 0$
3. Homoskedasticity: $\text{Var}(\varepsilon_i | X) = \sigma^2$ constant
4. No autocorrelation: $\text{Cov}(\varepsilon_i, \varepsilon_j | X) = 0$ for $i \neq j$
5. Normality (optional): $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Consequences:

- $\mathbb{C}orr(x_{ij}\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i) = 0$ under 2,
- $\text{Var}(\varepsilon | X) = \sigma^2 I_n$ under 3-4,
- $\mathbb{E}[b|X] = \beta$ b is unbiased, under 1-2,
- $\sqrt{n}(b - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1})$ under 1-4 and $\frac{1}{n}X'X \xrightarrow{p} Q$
- $\text{Var}(b|X) = \sigma^2(X'X)^{-1}$, which is the **variance-covariance matrix**, under 1-4.
  - Diagonal elements: variances of the estimated coefficients.
  - Off-diagonal elements: covariances between coefficients.
- $s^2 = \hat{\sigma}^2 = \frac{e'e}{n-k} \xrightarrow{p} \sigma^2$, under 1-4,
  - s^2 is NOT normally distributed, even for large n.
  - with k = df = number of regressors (incl. intercept)
  - $\mathbb{E}(s^2 \mid \mathbf{X}) = \sigma^2$ (unbiased estimator).
- $(n - K)\frac{s^2}{\sigma^2}|\text{X} \sim \chi^2_{n-K} \quad \Rightarrow \quad s^2$ is a scaled chi-squared variable.
- $X'e = 0$ is a mechanical result of the OLS first-order conditions and holds by construction, even if the exogeneity (assumption 2) fails.

Gauss-Markov Theorem (BLUE): If assumptions 1–4 hold, b is the Best Linear Unbiased Estimator of β.
The "Linear" in BLUE explicitly means the estimator b is linear in y

Warning: This does not refer to the model being linear in variables. The model $y = X\beta + e$ is assumed linear in parameters, but "linear" in BLUE is about the estimator's form.

Frisch–Waugh–Lovell Theorem

Goal: Estimate $\mathbf{b}_2$ from the model $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon$ after accounting for $\mathbf{X}_1$

Theorem: The coefficient $\mathbf{b}_2$ from the full regression is the same as the coefficient from the regression: $\boxed{\mathbf{b}_2 = \left(\mathbf{X}_2'\mathbf{M}^{\mathbf{X}_1}\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}^{\mathbf{X}_1}\mathbf{y}}$

Interpretation:

- Remove the part of $\mathbf{y}$ explained by $\mathbf{X}_1$ to get residuals: $\mathbf{y}^* = \mathbf{M}^{\mathbf{X}_1}\mathbf{y}$
- Remove the part of $\mathbf{X}_2$ explained by $\mathbf{X}_1$: $\mathbf{X}_2^* = \mathbf{M}^{\mathbf{X}_1}\mathbf{X}_2$
- Regress $\mathbf{y}^*$ on $\mathbf{X}_2^*$:

$$\boxed{\mathbf{b}_2 = \frac{(\mathbf{X}_2^*)'\mathbf{y}^*}{(\mathbf{X}_2^*)'\mathbf{X}_2^*}}$$

Matrix Definitions:

- $\mathbf{X}_1 \in \mathbb{R}^{n \times k_1}$: Matrix of control regressors (e.g., dummies)
- $\mathbf{X}_2 \in \mathbb{R}^{n \times k_2}$: Regressor(s) of interest
- $\mathbf{M}^{\mathbf{X}_1} = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$: Residual-maker matrix projecting orthogonally to $\mathbf{X}_1$
- $\mathbf{y}^* = \mathbf{M}^{\mathbf{X}_1}\mathbf{y}$: Residuals from regressing $\mathbf{y}$ on $\mathbf{X}_1$
- $\mathbf{X}_2^* = \mathbf{M}^{\mathbf{X}_1}\mathbf{X}_2$: Residuals from regressing $\mathbf{X}_2$ on $\mathbf{X}_1$

Projection & Residual Matrices

Hat matrix: $\hat{y} = Py = Xb$
We can also note that $y = \hat{y} + e$

Projection Matrix: $P = X(X'X)^{-1}X'$
We can also note that $\hat{y} = Py$

Residual Maker Matrix: $e = My, \quad M = I_n - P$
We can also note that $My = e = y - Xb$

Properties of P and M:

- P, M are symmetric $(A = A')$, idempotent $(A = A^k, \quad \forall k > 0)$
- $PX = X$, $MX = 0$
- $PM = MP = 0$
- $My = Me$
- $y = Py + My$, decomposition of y in two orthogonal parts

Key Property: Orthogonality of residuals (OLS projection result)

The residuals $e = y - Xb$ are **orthogonal to all regressors in** X:
$X'e = 0 \quad \Leftrightarrow \quad \sum_{i=1}^n x_{ij}e_i = 0$ for each regressor j
That implies:

- $\sum e_i = 0$ (orthogonal to the intercept, (CONDITIONAL ON HAVING AN INTERCEPT))
- $\sum z_i e_i = 0$, $\sum w_i e_i = 0$, etc.

This comes from the first-order condition of the OLS minimization problem.

$\boxed{X'e = 0}$ (residuals orthogonal to regressors)

Regression Specifications (How to interpret β_j)

Continuous:
$y = \beta_0 + \beta_j x_j + \varepsilon \Rightarrow$ 1 unit increase in $x_j \to \beta_j$ change in y

Dummy (binary):
$y = \beta_0 + \beta_j D_j + \varepsilon \Rightarrow D_j = 1$ vs $D_j = 0 \to \beta_j$ change in y

Log-Linear:
$\log(y) = \beta_0 + \beta_j x_j + \varepsilon \Rightarrow$ 1 unit increase in $x_j \to \beta_j \cdot 100\%$ change in y

Linear-Log:
$y = \beta_0 + \beta_j \log(x_j) + \varepsilon \Rightarrow$ 1% increase in $x_j \to \frac{\beta_j}{100}$ change in y

Log-Log:
$\log(y) = \beta_0 + \beta_j \log(x_j) + \varepsilon \Rightarrow$ 1% increase in $x_j \to \beta_j\%$ change in y

Goodness of Fit

Total Sum of Squares:
$TSS = \sum(y_i - \bar{y})^2 = y'M^0 y$
The TSS can be imagined as a sum of squared residuals on a regression with only a constant equal to $\bar{y}$.

Explained Sum of Squares (ESS):
$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = b'X'M^0 Xb$

Residual Sum of Squares (SSR):
$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = e'e = (y - Xb)'(y - Xb)$

Decomposition: $TSS = ESS + SSR$

Coefficient of Determination:
$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{e'e}{y'M^0 y}$
Note: R^2 does not penalize irrelevant regressors — prefer $\bar{R}^2$, AIC, or BIC for model selection.

Asymptotic Limit of R^2: Even as $n \to \infty$, $R^2 < 1$ if $\sigma^2 > 0$ (irreducible noise in y). $\Rightarrow \hat{\beta} \xrightarrow{p} \beta$ but y still noisy.

Adjusted R^2:
$\bar{R}^2 = 1 - \frac{e'e/(n-k)}{y'M^0 y/(n-1)} = 1 - \frac{n-1}{n-k}(1 - R^2)$

Change in R^2 from adding variable z:
$R^2_{X,z} = R^2_X + (1 - R^2_X)(r^X_{yz})^2$ One instrument per endogenous regressor system is just identified. Use standard IV formula.

Common Pitfalls in Linear Regression

1. Multicollinearity

- Occurs when some regressors are nearly linear combinations of others.
- Leads to large variances of OLS estimators $\to$ wide confidence intervals.
- Reduces power of t-tests $\to$ harder to reject $H_0 : \beta_j = 0$.
- Example: if $\mathrm{Corr}(x_1, x_2)$ close to 1, variance of b_1 inflates.
- Variance of b_j: $s^2 \cdot [(X'X)^{-1}]_{jj}$ increases when X has high collinearity.

— **2. Omitted Variable Bias**

- Suppose the **true model:** $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, but you estimate: $b_1 = (X_1'X_1)^{-1}X_1'y \Rightarrow \mathbb{E}[b_1|X] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$
- If $X_1'X_2 \neq 0$ (no orthogonality, linear dependence) and $\beta_2 \neq 0$, b_1 is biased.
- Intuition: X_1 "captures" the effect of omitted X_2
- Remedy: include control variables (i.e., add X_2 to the regression)

—

3. Irrelevant Variables (Overfitting)

- Suppose the true model is: $y = X_1\beta_1 + \varepsilon$ but you estimate: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ with $\beta_2 = 0$
- Estimates are still unbiased.
- But adding irrelevant X_2 increases the variance of b_1
- Leads to inefficiency and reduced test power (higher risk of Type II errors)
- Adjusted R^2 and AIC/BIC help guard against overfitting

Instrumental Variables (IV)

If $\mathbb{E}[\varepsilon_i|x_i] \neq 0$, then x_i is endogenous, and OLS is no longer consistent.

Why? $b = \beta + (X'X)^{-1}X'\varepsilon \quad \Rightarrow \quad b \xrightarrow{p} \beta + Q_{xx}^{-1}\gamma \neq \beta$ with $\gamma = \mathbb{E}[x_i\varepsilon_i]$
Here, $Q_{xx} = \frac{X'X}{n}$ is the probability limit of the scaled Gram matrix — it represents the asymptotic second moment matrix of the regressors. Its inverse, Q_{xx}^{-1}, appears in the asymptotic bias term and plays a role similar to $(X'X)^{-1}$ in finite samples.

Remedy: Instrumental Variables (IV) Use valid instruments z_i such that: $\mathbb{E}[\varepsilon_i|z_i] = 0$ and $\mathrm{Cov}(z_i, x_i) \neq 0$

—

Valid instruments z_i satisfy:

- **Relevance:** $\mathrm{Cov}(z_i, x_i) \neq 0$
- **Exogeneity:** $\mathbb{E}[\varepsilon_i|z_i] = 0$
- No multicollinearity in projected X on Z *Why?* To ensure that $(Z'X)$ is invertible and that all parameters in β are identified. $\Rightarrow$ Without this, b_{IV} is undefined (underidentified model).

—

Just-identified IV estimator ($L = K$):
One instrument per endogenous regressor system is just identified. Use standard IV formula.
$\boxed{b_{IV} = (Z'X)^{-1}Z'y}$
Wald estimator = IV estimator in the special case with:
- One endogenous regressor x_i
- One binary instrument z_i
Then: $\hat{\beta}_{\text{Wald}} = \frac{\mathrm{Cov}(z_i, y_i)}{\mathrm{Cov}(z_i, x_i)}$
More general cases use full IV or 2SLS formula.

Asymptotic distribution (if $L = K$): $b_{IV} \xrightarrow{d} \mathcal{N}\left(\beta, \frac{\sigma^2}{n}[Q_{xz}Q_{zz}^{-1}Q_{zx}]^{-1}\right)$

Where: $Q_{xz} = \frac{X'Z}{n}, \quad Q_{zx} = \frac{Z'X}{n}, \quad Q_{zz} = \frac{Z'Z}{n}$

Estimate variance in practice: $\widehat{\mathrm{Var}}(b_{IV}) = s^2_{IV} \cdot Q_{zx}^{-1}Q_{zz}Q_{xz}^{-1}$ with $s^2_{IV} = \frac{1}{n}\sum(y_i - x_i'b_{IV})^2$

—

Overidentified case ($L > K$): 2SLS
More instruments than regressors overidentified system. Project X on Z, then regress y on $\hat{X}$.
Steps:
1. Regress X on Z: $\hat{X} = P_Z X$ with $P_Z = Z(Z'Z)^{-1}Z'$
2. Regress y on $\hat{X}$

2SLS estimator: $\boxed{b_{2SLS} = (X'P_Z X)^{-1}X'P_Z y}$

—

Weak instruments:
- Instruments only weakly correlated with x_i (relevance not resepcted)
- Check: **F-statistic from first stage** regression. Low F weak instruments

Inference and Confidence Intervals

Under assumptions 4.1–4.5 (including normality):
$\mathbf{b}|X \sim \mathcal{N}\left(\beta, \sigma^2(X'X)^{-1}\right)$
$s^2 = \hat{\sigma}^2 = \frac{e'e}{n-k} \xrightarrow{p} \sigma^2)$ (under 1-4)
with k = df = number of regressors (incl. intercept)

Distribution of b_k (component of $\mathbf{b}, \sigma^2$ known):
$b_k \mid X \sim \mathcal{N}\left(\beta_k, \sigma^2 v_k\right)$
$\sqrt{n}\frac{b_k - \beta_k}{\sqrt{\sigma^2 v_k}} \mid X \sim \mathcal{N}(0, 1)$
Distribution of b_k (component of $\mathbf{b}, \sigma^2$ unknown):
$\sqrt{n}\frac{b_k - \beta_k}{\sqrt{s^2 v_k}} \mid X \sim t(n-k)$

with $v_k = [(X'X)^{-1}]_{kk}$, which means: the k-th diagonal element of the matrix $(X'X)^{-1}$, and is the variance weight associated with the k-th coefficient b_k

t-statistic:
$$t_k = \frac{\frac{b_k - \beta_k}{\sqrt{\sigma^2 v_k}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2(n-K)}}} = \frac{b_k - \beta_k}{\sqrt{s^2 v_k}} \sim t(n-K)$$

t-Test Requirements:

- Gauss-Markov (Assumptions 1–4) **are insufficient** for valid t-tests in small samples.
- **Normality** (Assumption 5: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$) is strictly required for exact t-distributions in finite samples.
- Without normality, t-tests rely on asymptotic approximations (CLT).

Confidence interval for b_k at $1 - \alpha\%$:
$[b_k \pm t_{1-\frac{\alpha}{2}, n-k} \cdot \sqrt{s^2 v_k}]$
We use $\mathcal{N}(0, 1)$ quantiles as $n \to \infty$ (CLT)

Inference on linear combinations: Let $\alpha'b$ estimate $\alpha'\beta$:
$\alpha'b \mid X \sim \mathcal{N}(\alpha'\beta, \sigma^2\alpha'(X'X)^{-1}\alpha)$
$\Rightarrow \frac{\alpha'b - \alpha'\beta}{\sqrt{s^2\alpha'(X'X)^{-1}\alpha}} \sim t(n-k)$

Hypothesis Testing:

We test hypotheses about parameters θ (imperfectly observed) using data $\mathbf{x}$ whose distribution depends on θ.

Null and alternative hypotheses:
$H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1 = \Theta_0^c$
(or equivalently, $H_0 : h(\theta) = 0$)

A statistical test requires:

- a parameter vector θ (partially or fully unknown)
- a test statistic $S(\mathbf{x})$ (function of the sample)
- a critical region Ω (set of implausible values under H_0)

Decision rule:

- Reject H_0 if $S(\mathbf{x}) \in \Omega$
- Fail to reject H_0 if $S(\mathbf{x}) \notin \Omega$

Errors and test performance:

- **Type I error (false positive):** reject H_0 when true (α)
- **Type II error (false negative):** fail to reject H_0 when false (β)
- **Power:** $\gamma = 1-\beta \to 1$ as $n \to \infty$ (probability of correctly rejecting H_0)

Error probabilities:
$\alpha = \mathbb{P}(S \in \Omega \mid H_0)$ $\beta = \mathbb{P}(S \notin \Omega \mid H_1)$
Note: There is often a tradeoff between α and power $(1-\beta)$

Decision / Truth	H_0 True	H_0 False
Not rejected	**Correct decision** $(1-\alpha)$	**Type II error** (β)
Rejected	**Type I error** (α)	**Correct decision** $(1-\beta)$

Common Tests:

Generic Setup: Let S_n be a test statistic that (under H_0) follows a known distribution D, or compare S_n to quantiles of D at level α.

p-value: Reject H_0 at level α if $p < \alpha$

1. Student's t-distribution: Arises when: $\hat{\theta}$ is normally distributed with variance estimated from sample
Assumptions: i.i.d. observations, normality (or large n), unknown variance
Statistic: $T = \frac{\hat{\theta}-\theta_0}{\widehat{\mathrm{SE}}(\hat{\theta})} \sim t_{df}$
Reject H_0 if: $|T| > t_{\alpha/2,df}$ (two-sided) or $T > t_{\alpha,df}$ (one-sided)

2. Chi-squared (χ^2) distribution:
Arises when: testing variance, goodness-of-fit, or quadratic forms in normals
Statistic: $\chi^2 = \sum_{i=1}^{k} \left(\frac{O_i - E_i}{\sqrt{E_i}}\right)^2$ or $\hat{\varepsilon}' A \hat{\varepsilon}$
Degrees of freedom = number of independent components
Reject H_0 if: $\chi^2_{\text{obs}} > \chi^2_{\alpha,df}$

3. F-Test
When used: To test joint linear restrictions (e.g. $H_0 : R\beta = q$), compare models, or test variance equality.

General formula (Wald-based):
$F = \frac{(Rb-q)'[R(X'X)^{-1}R']^{-1}(Rb-q)}{Js^2} \sim \mathcal{F}(J, n-k)$

Alternative formula (SSR-based):
$F = \frac{(SSR_{restr} - SSR_{unrestr})/J}{SSR_{unrestr}/(n-k)} = \frac{(R^2 - R^2_*)/J}{(1-R^2)/(n-k)}$
Where SSR_{restr} is the model under H_0

Degrees of freedom:

- J = number of restrictions (numerator df)
- $n-k$ = number of residual df (denominator)

Reject H_0 if $F_{\text{obs}} > F_{\alpha,J,n-k}$
Asymptotic approx (large $n-k$): $F(J,\infty) \approx \chi^2(J)/J \to$ Wald and F converge.

Caution (CLT misconception): Even as $n \to \infty$, F-distributions do **not** converge to Normal. F is a ratio of χ^2 variables the CLT does not apply.

4. Standard Normal ($\mathcal{N}(0,1)$):
Arises when: variance is known, or from large sample CLTs
Statistic: $Z = \frac{\hat{\theta}-\theta_0}{\mathrm{SE}(\hat{\theta})} \sim \mathcal{N}(0,1)$
Reject H_0 if: $|Z| > z_{\alpha/2}$ (two-sided) or $Z > z_\alpha$ (one-sided)
Critical values for $\mathcal{N}(0,1)$:

Significance Level α	z_α (one-sided)	$z_{\alpha/2}$ (two-sided)
0.10	1.28	1.64
0.05	1.645	1.96
0.01	2.33	2.58
0.001	3.09	3.29

5. Jarque-Bera Test (Normality):
Tests whether a sample's skewness and kurtosis match those of a normal distribution
Under H_0: $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, so: $g_3 \to 0$, $g_4 \to 3$ Asymptotic properties:

- $\sqrt{n} g_3 \xrightarrow{d} \mathcal{N}(0,6)$
- $\sqrt{n}(g_4 - 3) \xrightarrow{d} \mathcal{N}(0,24)$

JB Statistic: $JB = \frac{n}{6}\left(g_3^2 + \frac{(g_4-3)^2}{4}\right)$ where g_3, g_4 are standardized sample moments
Under H_0, $JB \xrightarrow{d} \chi^2(2)$

6. Durbin–Wu–Hausman Test (for endogeneity):
Test whether OLS is inconsistent and IV is necessary.
$H = (b_{IV} - b_{OLS})' \left[\widehat{\mathrm{Var}}(b_{IV}) - \widehat{\mathrm{Var}}(b_{OLS})\right]^+ (b_{IV} - b_{OLS})$
Where:
- $+$ = Moore-Penrose pseudo-inverse (in case matrix isn't full rank)
- Under H_0: both estimators are consistent (OLS preferred for efficiency)
- Under H_0: $H \sim \chi^2(q)$, where q is the rank of $\mathrm{Var}(\mathbf{b}_1) - \mathrm{Var}(\mathbf{b}_0)$ - Under H_1: only IV is consistent (OLS is biased)

Rejecting H_0 OLS inconsistent prefer IV.

Testing Linear Restrictions (F-test)

Test joint restrictions: $H_0 : R\beta = q$ vs. $H_1 : R\beta \neq q$

Discrepancy vector: $m = Rb - q$
Under H_0: $\mathbb{E}[m|X] = 0$
Variance: (Under 4.1–4.4) $\mathrm{Var}(m|X) = \sigma^2 R(X'X)^{-1}R'$

Wald statistic (σ^2 known:)
$W = m' [\mathrm{Var}(m|X)]^{-1} m = \frac{m'[R(X'X)^{-1}R']^{-1}m}{\sigma^2} \sim \chi^2(J)$

F statistic (when σ^2 is unknown, use s^2):
$F = \frac{1}{J} \cdot \frac{m'[R(X'X)^{-1}R']^{-1}m}{s^2} \sim F(J, n-K)$

We can use the SSR (alternative) formula (see F-test). The restricted SSR is the one of the test, with less parameters, i.e. the one under H_0.