Probability Spaces

Sample Space Ω => set of all outcomes (*mutually exclusive*) of random experiment

Event $E \subseteq \Omega$ => any subset of sample space

Extreme Events => null event $\emptyset$ & universal event S

Elementary Event => *singleton subsets* of S

If $s \in E$ is *experiment outcome*, then E occurred

Null event $\emptyset$ never & *universal event* S always -occurs

For events $E_1, E_2, \dots$:

E_1 or E_2 or ... => $\bigcup_i E_i$

E_1 and E_2 and ... => $\bigcap_i E_i$

$E_1, E_2, \dots$ are **Mutually Exclusive** => $\forall i,j.\ E_i \cap E_j = \emptyset$ i.e. they're *pairwise-disjoint*

$E_1, E_2, \dots$ are **Independent Events** => $P(\bigcap_{j=1}^n E_{i_j}) = \prod_{j=1}^n P(E_{i_j})$ for any finite subset $\{E_{i_1}, E_{i_2}, \dots, E_{i_n}\}$

If events A, B are *independent*, then $\bar{A}, \bar{B}$ are *also independent*

σ-algebra $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ => family of subsets of Ω s.t.

nonempty: $\Omega \in \mathcal{F}$

closed under *complements*: $E \in \mathcal{F} \Longrightarrow \bar{E} \in \mathcal{F}$

closed under *countable union*: $E_1, E_2, \dots \in \mathcal{F} \Longrightarrow \bigcup_i E_i \in \mathcal{F}$

Immediate Basic Results:

$\emptyset \in \mathcal{F}$

closed under *countable intersection*: $E_1, E_2, \dots \in \mathcal{F} \Longrightarrow \bigcap_i E_i \in \mathcal{F}$

$\{\emptyset, \Omega\}$ is smallest & $\mathcal{P}(\Omega)$ is largest σ-algebras

Generated σ-algebras:

$\sigma(\mathcal{G})$ => for family of subsets $\mathcal{G} \subseteq \mathcal{P}(\Omega)$, its *smallest* σ-algebra to contain $\mathcal{G}$ (exists & unique)

$\sigma(f)$ => for $f: \Omega \to E$ where $(E, \mathcal{E})$ is *measurable space*, $\sigma(f) = \{f^{-1}[F] \mid F \in \mathcal{E}\}$ i.e. all pre-images

trace σ-algebra of $B \in \mathcal{F}$ => $\mathcal{F}_B = \{B \cap A \mid A \in \mathcal{F}\}$

Probability Measure $P: \mathcal{F} \to [0,1]$ on $(\Omega, \mathcal{F})$

$\forall E \in \mathcal{F}, 0 \le P(E) \le 1$ i.e. between 0 and 1

$P(\Omega) = 1$ i.e. *universal event* S always occurs

σ-additive (*countably additive*) => $P(\bigcup_i E_i) = \sum_i P(E_i)$ for *pairwise-disjoint* events $E_1, E_2, \dots \in \mathcal{F}$

Immediate Basic Results:

$P(\bar{E}) = 1 - P(E)$

$P(\emptyset) = 0$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Measurable Space $(\Omega, \mathcal{F})$ => *sample space* Ω with *σ-algebra* $\mathcal{F}$ on it

Probability Space $(\Omega, \mathcal{F}, P)$ => *measurable space* $(\Omega, \mathcal{F})$ with *probability measure* P on it

Conditional Probability

Conditional Probability => $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ where $A, B \subseteq \Omega$ and $P(B) \neq 0$

The **conditional probability space** is $(B, \mathcal{F}_B, P(\cdot \mid B))$:

Sample space $B \subseteq \Omega$

Trace σ-algebra $\mathcal{F}_B = \{B \cap A \mid A \in \mathcal{F}\}$

Probability measure $P(\cdot \mid B)$

If A, B are *independent* then $P(A \mid B) = P(A)$

A_1, A_2 are **Conditionally Independent** given B **iff** $P(A_1 \cap A_2 \mid B) = P(A_1 \mid B) P(A_2 \mid B)$

Law Of Total Probability => for any *events* $\{B_1, B_2, \dots\}$ which *partition* Ω, $P(A) = \sum_i P(A \mid B_i) P(B_i) = \sum_i P(A \cap B_i)$

Special Case =>

$P(A) = P(A \cap B) + P(A \cap \bar{B}) = P(A \mid B)P(B) + P(A \mid \bar{B})P(\bar{B})$

Bayes Theorem => $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

General Random Variables

Random Variable => *measurable function* $X: \Omega \to E$

$(\Omega, \mathcal{F}, P)$ is a *probability space*, $(E, \mathcal{E})$ is a *measurable space*

For every $B \in \mathcal{E}$ the pre-image of B under X is in $\mathcal{F}$, i.e. $X^{-1}[B] = \{s \in \Omega \mid X(s) \in B\} \in \mathcal{F}$

i.e. $\sigma(X) \subseteq \mathcal{F}$ where $\sigma(X)$ is generated by *function* X

$g(X)(s) = (g \circ X)(s)$ is *also random variable*, for *measurable function* $g: E \to E$

Induced Probability $P_X(X \in B)$ => probability that X takes on value in $B \in \mathcal{E}$

$P_X(X \in B) = P(X^{-1}[B]) = P(\{s \in \Omega \mid X(s) \in B\})$

Also called **Pushforward Measure** of P onto $(E, \mathcal{E})$

induced by X => $(E, \mathcal{E}, P_X)$ is a *probability space*

Also called the **Probability Distribution** of X

Real Random Variables

Real Random Variable is *RV* who's *co-domain* is $E = \mathbb{R}$

Support $\mathrm{supp}(X)$ => is range of X i.e. $\mathrm{supp}(X) = X[\Omega]$

Simple RRV iff *finite* $\mathrm{supp}(X)$

Discrete RRV iff *countable* $\mathrm{supp}(X)$

Continuous RRV iff *uncountable* $\mathrm{supp}(X)$

Induced Probability => $P_X(X \le x) = P(\{s \in \Omega \mid X(s) \le x\})$

Cumulative Distribution Func. (CDF) $F_X(x) = P_X(X \le x)$

$F_X(x)$ is *right-continuous* => for any *decreasing* (x_n)

$\lim_{n\to\infty} x_n = x_L \Longrightarrow \lim_{n\to\infty} F_X(x_n) = F_X(x_L)$

To check that function is valid CDF, must obey:

Monotonicity $\forall x_1, x_2 \in \mathbb{R}, x_1 < x_2 \Longrightarrow F_X(x_1) \le F_X(x_2)$

$F_X(-\infty) = 0, F_X(\infty) = 1$

F_X is right-continuous

Simple Properties

$0 \le F_X(x) \le 1, \forall x \in \mathbb{R}$

$P_X(a < X \le b) = F_X(b) - F_X(a)$ for finite intervals $(a, b] \subseteq \mathbb{R}$

Moments of RRVs

Expectation $E[X]$ => the **mean** μ_X of *distribution* of X

Discrete => $E_X[g(X)] = \sum_x g(x) p_X(x)$

Continuous => $E_X[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

Linearity => $E[\alpha g(X) + h(X) + \beta] = \alpha E[g(X)] + E[h(X)] + \beta$

Sum => for any $X_1, \dots, X_n$

$E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i]$ $E[\bar{X}] = E[\frac{\sum_{i=1}^n X_i}{n}] = \frac{\sum_{i=1}^n E[X_i]}{n}$

Independent Product => $E[\prod_{i=1}^n X_i] = \prod_{i=1}^n E[X_i]$

Independent and Identically Distributed (i.i.d.) => $E(\bar{X}) = \mu_X$

n-th Raw Moment μ'_n $E[X^n]$ => i.e. about zero

n-th Central Moment $\mu_n = E[(X - E[X])^n]$

Variance $\sigma_X^2 = \mathrm{Var}(X) = E[(X - E(X))^2]$

$\mathrm{Var}(X) = E[X^2] - (E[X])^2$

$\mathrm{Var}(\alpha X + b) = \alpha^2 \mathrm{Var}(X)$

Sum => for *independent* $X_1, \dots, X_n$

$\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$

$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{\sum_{i=1}^n X_i}{n^2}) = \frac{\sum_{i=1}^n \mathrm{Var}(X_i)}{n}$

Independent and Identically Distributed (i.i.d.) => $\mathrm{Var}(\bar{X}) = \frac{\sigma_X^2}{n}$

Standard Deviation $\sigma_X = \mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$

n-th Standardized Moment $\tilde{\mu}_n = \frac{\mu_n}{\sigma^n} = \frac{E[(X - E[X])^n]}{\sqrt{\mathrm{Var}(X)}^n}$

Skewness $\gamma_1 = \tilde{\mu}_3 = \frac{E[(X-\mu)^3]}{\sigma^3}$ *measures asymmetry*

positive skew => distribution *leans left*

negative skew => distribution *leans right*

Moment Generating Function (MGF) $M_X(t) = E[e^{tX}]$

$E[X^n] = \frac{d^n M_X}{dt^n}\Big|_{t=0}$ if *open interval* around $t = 0$ exists

because $e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \dots + \frac{t^nX^n}{n!} + \dots$

so $M_X(t) = E[e^{tX}] = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \dots + \frac{t^nE[X^n]}{n!} + \dots$

Sum => for *independent* $X_1, \dots, X_n$

let $X = \sum_{i=1}^n X_i$ => $M_X(t) = \prod_{i=1}^n M_{X_i}(t)$

Discrete Random Variables

Discrete RRV iff *countable* $\mathrm{supp}(X)$

Let $\mathrm{supp}(X) = \{x_1, x_2, \dots\}$ be ordered s.t. $x_1 < x_2 < \dots$

CDF F_X will be *monotonic increasing step function*,

i.e. $F_X(x_i) = F_X(x_{i-1}) + P_X(X = x_i)$

i.e. $P_X(X = x_i) = F_X(x_i) - F_X(x_{i-1})$

Probability Mass Function (PMF) $p(x) = P_X(X = x)$

$0 \le p(x) \le 1, \forall x \in \mathbb{R}$

$\sum_{x \in \mathrm{supp}(X)} p(x) = 1$

$p(x_i) = F(x_i) - F(x_{i-1})$

$F(x_i) = \sum_{j=1}^i p(x_j)$

Bernoulli(P) => TODO: HEREEE!!!!!

Binomial(N,p) => TODO: HEREEE!!!!!

Poisson(P) => TODO: HEREEE!!!!!

Poisson(Λ) => TODO: HEREEE!!!!!

Uniform(U) => TODO: HEREEE!!!!!

Negative Binomial Distribution(U) => TODO: HEREEE!!!!!

Poisson Binomial Distribution => TODO: HEREEE!!!!!

Continuous Random Variables

X is **(Absolutely) Continuous RRV** if $\exists f_X: \mathbb{R} \to \mathbb{R}$ such that $F_X(x) = \int_{u=-\infty}^x f_X(u) du$

f_X called **Probability Density Function (PDF)** of X

$P_X(a < X \le b) = P_X(X \le b) - P_X(X \le a)$

$P_X(a < X \le b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$

$P_X(X = x) = P_X(\{x\}) = 0$ and

$P_X(X \in \{x_1, x_2, \dots\}) = P_X(X = x_1) + P_X(X = x_2) + \dots$ for *countable sets*

Properties of PDFs:

$F_X(x) = \int_{-\infty}^x f_X(t) dt$ & $f_X(x) = \frac{d}{dx} F_X(x)$

$f_X(x) \ge 0$ & $\int_{-\infty}^{\infty} f_X(x) dx = 1$

$f_X(x) \cdot h \approx P_X(X \in [x, x+h])$

Quantiles and Percentiles:

The **lower** and **upper quartiles** and **median** of *sample of data* are the points $(\frac{1}{4}, \frac{3}{4}, \frac{1}{2})$-way through the ordered dataset, respectively

α-quantile $Q_X(\alpha)$ => for continuous X and $0 \le \alpha \le 1$, the least number satisfying $P(X \le Q_X(\alpha)) = \alpha$

i.e. $Q_X(\alpha) = F_X^{-1}(\alpha)$

$Q(1/2)$ is *median* & *k*-th percentile is $Q(k/100)$

Uniform(A,b) => TODO: HEREEE!!!!!

Exp(Λ) => TODO: HEREEE!!!!!

Normal(M, Σ2) => TODO: HEREEE!!!!!

Lognormal => TODO: HEREEE!!!!!

Joint Random Variables

Product Probability Space $(\prod_i \Omega_i, \bigotimes_i \mathcal{F}_i, \times_i P_i)$

Let $(\Omega_1, \mathcal{F}_1, P_1), \dots, (\Omega_n, \mathcal{F}_n, P_n)$ be *probability spaces*

$\prod_i \Omega_i$ is *n-ary Cartesian product* of $\Omega_1, \dots, \Omega_n$

$\bigotimes_i \mathcal{F}_i$ is *product σ-algebra* of $\mathcal{F}_1, \dots, \mathcal{F}_n$

$\bigotimes_i \mathcal{F}_i \triangleq \sigma(\{\prod_{i=1}^n E_i \mid E_1 \in \mathcal{F}_1, \dots, E_n \in \mathcal{F}_n\})$

where $\sigma(\cdot)$ is the *generated σ-algebra*

$\times_i P_i$ is *unique measure* such that

$(\times_i P_i)(E_1 \times \dots \times E_n) = P_1(E_1) \cdots P_n(E_n)$

for every $E_1 \in \mathcal{F}_1, \dots, E_n \in \mathcal{F}_n$

Multivariate Random Variable => $X: \Omega \to \mathbb{R}^n$

Consider *RRVs* $X_1, \dots, X_n: \Omega \to \mathbb{R}$

NOTE: take *product space* if their *sample spaces differ*

$\mathbf{X} = (X_1, \dots, X_n): \Omega \to \mathbb{R}^n$ is a **Random Vector**

$[\mathbf{X}(s)]_i = X_i(s)$ => it's a *random variable*

from $(\Omega, \mathcal{F}, P)$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$

$\mathcal{B}(\mathbb{R}^n)$ is **Borel σ-algebra** on $\mathbb{R}^n$

i.e. requirement that each X_i is *also an RV*

Induced Probability =>

$P_{\mathbf{X}}(X_1 \le x_1, \dots, X_n \le x_n) = P(\{s \in \Omega \mid X_1(s) \le x_1 \wedge \dots \wedge X_n(s) \le x_n\})$

RECALL: this *pushforward measure* is also called *probability distribution* of $\mathbf{X}$ => this *measure* on *Borel σ-algebras* is also called the **Joint Probability Distribution**

Joint CDF $F_{\mathbf{X}}(x_1, \dots, x_n) = P_{\mathbf{X}}(X_1 \le x_1, \dots, X_n \le x_n)$

Recover **Marginal CDFs** for each X_i with $F_{X_i}(x_i) = \lim_{j \ne i, x_j \to \infty} F_{\mathbf{X}}(x_1, \dots, x_n)$

e.g. for *RVs* X, Y => $F_X(x) = F(x, \infty)$ and $F_Y(y) = F(\infty, y)$

To check that function is valid CDF, must obey:

Monotonicity:

For every i, *hold fixed* all-but-the i-th component

Then for every x_{i_1}, x_{i_2} we have

$x_{i_1} < x_{i_2} \Longrightarrow F_{\mathbf{X}}(\dots, x_{i_1}, \dots) \le F_{\mathbf{X}}(\dots, x_{i_2}, \dots)$

$0 \le F_{\mathbf{X}}(x_1, \dots, x_n) \le 1$

$F_{\mathbf{X}}(\dots, -\infty, \dots) = 0, F_{\mathbf{X}}(\infty, \dots, \infty) = 1$

Interval Probabilities for *bi-variate* case $Z = (X, Y)$

$P_Z(x_1 < X \le x_2, Y \le y) = F(x_2, y) - F(x_1, y)$, hence

$P_Z(x_1 < X \le x_2, y_1 < Y \le y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$

Joint PMF $p(x_1, \dots, x_n) = P_{\mathbf{X}}(X_1 = x_1, \dots, X_n = x_n)$

Recover **Marginal PMFs** for each X_i with

$p_{X_i}(x_i) = \sum_{x_n} \cdots \sum_{x_{i+1}} \sum_{x_{i-1}} \cdots \sum_{x_1} p_{\mathbf{X}}(x_1, \dots, x_n)$

e.g. for *RVs* X, Y => $p_X(x) = \sum_y p(x,y)$ & $p_Y(y) = \sum_x p(x,y)$

To check that function is valid PMF, must obey:

$0 \le p_{\mathbf{X}}(x_1, \dots, x_n) \le 1$ & $\sum_{x_1} \cdots \sum_{x_n} p_{\mathbf{X}}(x_n, \dots, x_1) = 1$

Multinomial Distribution => TODO: HEREEE!!!!!

$X_1, \dots, X_n$ are **Jointly Continuous** if $\exists f_{\mathbf{X}}: \mathbb{R}^n \to \mathbb{R}$ such

$F_{\mathbf{X}}(x_1, \dots, x_n) = \int_{t_n=-\infty}^{x_n} \cdots \int_{t_1=-\infty}^{x_1} f_{\mathbf{X}}(t_1, \dots, t_n)\, dt_1 \cdots dt_n$

$f_{\mathbf{X}}$ called **Joint PDF** of $\mathbf{X} = (X_1, \dots, X_n)$

$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\partial^n}{\partial x_1 \cdots \partial x_n} F_{\mathbf{X}}(x_1, \dots, x_n)$

To check that function is valid PMF, must obey:

$f_{\mathbf{X}}(x_1, \dots, x_n) \ge 0$ &

$\int_{t_n=-\infty}^{\infty} \cdots \int_{t_1=-\infty}^{\infty} f_{\mathbf{X}}(t_1, \dots, t_n) dt_1 \cdots dt_n = 1$

Recover **Marginal PDFs** for each X_i with

$f_{X_i}(x_i) = \int_{x_n=-\infty}^{\infty} \cdots \int_{x_{i+1}=-\infty}^{\infty} \int_{x_{i-1}=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} f_{\mathbf{X}}(x_1, \dots, x_n)\, dx_1 \cdots dx_n$

e.g. for *RVs* X, Y => $f_X(x) = \int_{y=-\infty}^{\infty} f(x,y) dy$ and

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

X, Y are **Independent RRVs if-and-only-if**

General => $F(x,y) = F_X(x) F_Y(y)$

Discrete => $p(x,y) = p_X(x) p_Y(y)$

Continuous => $f(x,y) = f_X(x) f_Y(y)$

Joint Expectation $E[\mathbf{X}] = E[X_1, \dots, X_n]$

Discrete =>

$E_{\mathbf{X}}[g(\mathbf{X})] = \sum_{x_n} \cdots \sum_{x_1} g(x_1, \dots, x_n) p_{\mathbf{X}}(x_1, \dots, x_n)$

Continuous =>

$E_{\mathbf{X}}[g(\mathbf{X})] = \int_{x_n=-\infty}^{\infty} \cdots \int_{x_1=-\infty}^{\infty} g(x_1, \dots, x_n) f_{\mathbf{X}}(x_1, \dots, x_n) dx_1 \cdots dx_n$

Sum => $E[\sum_i g_i(X_i)] = \sum_i E_{X_i}[g_i(X_i)]$

Independent Product => $E[\prod_i g_i(X_i)] = \prod_i E_{X_i}[g_i(X_i)]$

Covariance =>

$\sigma_{XY} = \mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$

NOTE: $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$

For *independent RVs* $E[XY] = E_X[X] E_Y[Y]$ so $\sigma_{XY} = 0$

Correlation => $\rho_{XY} = \mathrm{Cor}(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)}$

Correlation is *invariant to the scale* of the X, Y

NOTE: $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$

For *independent RVs* $\sigma_{XY} = \rho_{XY} = 0$

Multivariate Normal Distribution => TODO: HEREEE!!!!!

Conditional Distributions

Conditional Distribution $P_{Y|X}(B_Y \mid B_X) = \frac{P_{XY}(B_X, B_Y)}{P_X(B_X)}$

for any $B_X, B_Y \in \mathbb{R}$

Probability of Y falling inside B_Y given that we know X fell inside B_X

Conditional CDF $F_{Y|X}(y \mid x) = P(Y \le y \mid X = x)$

Discrete => $P(Y \le y \mid X = x) = \sum_{u=-\infty}^{y} p_{Y|X}(u \mid x)$

Continuous => $P(Y \le y \mid X = x) = \int_{u=-\infty}^{y} f_{Y|X}(u \mid x) du$

Interval Probabilities =>

$P(a < X \le b \mid Y = y) = F_{X|Y}(b \mid y) - F_{X|Y}(a \mid y)$

Law of Total Probability CDF =>

$F_X(x) = \int_{y=-\infty}^{\infty} F_{X|Y}(x \mid y) f_Y(y) dy$

Conditional PMF $p_{Y|X}(y \mid x) = \frac{p_{XY}(x,y)}{p_X(x)}$ if $p_X(x) > 0$

Bayes Theorem PMF => $p_{X|Y}(x \mid y) = \frac{p_{Y|X}(y|x) p_X(x)}{p_Y(y)}$

Law of Total Probability PMF =>

$p_X(x) = \sum_y p_{X|Y}(x \mid y) p_Y(y)$

Conditional PDF $f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}$ if $p_X(x) > 0$

Bayes Theorem PDF => $f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}$

Law of Total Probability PDF =>

$f_X(x) = \int_{y=-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy$

Conditional Expectation $E_{Y|X}[Y \mid X = x]$

Discrete => $E_{Y|X}[Y \mid X = x] = \sum_y y p_{Y|X}(y \mid x)$

Continuous => $E_{Y|X}[Y \mid X = x] = \int_{y=-\infty}^{\infty} y f_{Y|X}(y \mid x) dy$

NOTE: $E_{Y|X}[Y \mid X = \cdot]$ is a function of x

Law of Total Expectation:

Define $E_{Y|X}[Y \mid X](s) = E_{Y|X}[Y \mid X = X(s)]$

Now $E_{Y|X}[Y \mid X]: \Omega \to \mathbb{R}$ is *function of RRV* X

Then we have $E_Y[Y] = E_X[E_{Y|X}[Y \mid X]]$

e.g. in *continuous case* => $\int_x \int_y y f_{Y|X}(y \mid x) f_X(x) dy dx = \int_y \int_x y f(x,y) dx dy = \int_y y f_Y(y) dy$

Markov Chains

Discrete Time Markov Chains (DTMC) => generalization of *sequences of i.i.d. RVs* to support *arbitrary (and possibly dependent) RVs*

State Space J => elements $j \in J$ are *states*

$X_n, n \ge 0$ => takes values in J & models state at time n

Realization of $X_0, X_1, \dots$ is called **Sample Path**

GOAL => calculate $P(X_n = j)$

i.e. probability that at time n system reaches state j

Markov property =>

$P(X_{n+1} = j_{n+1} \mid X_n = j_n, \dots, X_1 = j_1, X_0 = j_0)$
$= P(X_{n+1} = j_{n+1} \mid X_n = j_n)$

i.e. choice of the next state depends on the current state only

How to *specify DTMCs*:

Initial Probability Vector => row-vector $\pi_0^T \in \mathbb{R}^n$

where $(\pi_0)_i = P(X_0 = i)$

Transition Probability Matrix => $R \in \mathbb{R}^{n \times n}$ where

$R_{ij} = P(X_{n+1} = j \mid X_n = i)$

Each transition probability R_{ij} is independent of the time n

Self-loops are allowed => e.g. $R_{ii} = 1$ means DTMC can *never leave* state i (*e.g. a permanent fault*)

R is a *non-nonegative matrix* who's rows sum to 1

Also called **stochastic matrix**

$P(X_n = j \mid X_0 = i) = (R^n)_{ij}$ => $P(X_n = j) = (\pi_0 R^n)_j$

Classification Of DTMCs:

Irreducible DTMC => *directed graph* associated to R is *strongly connected*:

For any pair (i,j) there *exist some sample path* where, starting in state i, the DTMC eventually reaches state j

Periodic DTMC => *time to return to a state* is an *integer multiple of a fixed period*; otherwise it's **aperiodic DTMC**

Common Long-Term Behaviours Of DTMC:

Steady-State/Stationary Distribution => row-vector π_∞^* thats invariant under R,

i.e. $P(X_n = j) = (\pi_\infty^*)_j$ for all $n \ge 0, j \in J$

Limiting Distribution => row-vector π_∞ such that $\pi_\infty = \lim_{n\to+\infty} \pi_0 R^n$

If exists, its *also a steady-state distribution* => converse isn't true

Uniqueness and existence:

The two above are **not unique in general**

If *irreducible and aperiodic DTMC*:

Limiting & steady-state distributions both exist => *unique* and identical to each other

Elements of π_∞ all *strictly positive*

π_∞ is solution of $\pi_\infty R = \pi_\infty$ subject to $\sum_j (\pi_\infty)_j = 1$

i.e. left-eigenvector with left-eigenvalue $\lambda = 1$

If *irreducible but* ***not*** *aperiodic DTMC*:

limiting distribution π_∞ no longer exists

Steady-state distribution π_∞^* still exists =>

its the *unique positive solution* of $\pi_\infty^* R = \pi_\infty^*$ subject to $\sum_j (\pi_\infty^*)_j = 1$

Estimator Theory

Sample of Data $x = (x_1, \dots, x_n)$ => may be seen as *realisation of RVs* $\mathbf{X} = (X_1, \dots, X_n)$ which are **i.i.d**

Assume that a single draw X_i follows a *distribution* $P(\cdot \mid \theta)$,

where $\theta = (\theta_1, \dots, \theta_r)$ are parameters we wish to estimate

I.i.d assumption implies => sampling is with replacement, i.e. any number of samples n may be collected

Statistic $T = T(\mathbf{X}) = T(X_1, \dots, X_n)$ => a function of *random sample vector* $\mathbf{X}$

Since T is a *random vector*, its a *random variable*

If $T(\mathbf{X})$ used to *approximate parameters* in θ then, T is an **Estimator** for θ

realisation $t(x)$ of *estimator* for *particular data sample* x is called **estimate** of θ

GOAL => study **sampling distribution**

$P(T \mid \theta) \equiv P(T(X) \mid \theta)$ for *statistic* T and *its moments*

Estimators for Mean => TODO: HEREEE!!!!!

Bias of Estimator $\mathrm{bias}(T) = E[T \mid \theta] - \theta$

Requires to determine the *expectation* of $T(\mathbf{X})$ based on *sampling distribution*

If $\mathrm{bias}(T) = 0$ we say T is **Unbiased**

Sample mean $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ is *unbiased estimate* for the *population mean* μ, i.e. $E[\bar{X}] = \mu$

Sample variance $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is known to be *biased estimate* of σ^2

Only if we know *population mean* μ then

$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ is *unbiased estimate* of σ^2

Bessel's correction => $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is an is *unbiased estimate* of σ^2, i.e. $E[S^2] = \sigma^2$

Efficiency of Estimators

Let $T \equiv T(\mathbf{X}), H \equiv H(\mathbf{X})$ be unbiased estimators for a parameter θ

Assume we know the *corresponding sampling distributions* $P(T \mid \theta), P(H \mid \theta)$ => we can calculate their variances

T is **More Efficient** than H if:

1) $\forall\theta \quad \mathrm{Var}(T \mid \theta) \le \mathrm{Var}(H \mid \theta)$

2) $\exists\theta \quad \mathrm{Var}(T \mid \theta) < \mathrm{Var}(H \mid \theta)$

T is **Efficient if** its *more efficient* than *any other possible estimator*

e.g. $T = \bar{X}$ is *more efficient* than $H = X_1$ as *estimator* for μ (*for* $n \ge 2$)

Consistency of Estimators

Consistency allows us to recognize bad patterns as n grows large

T is a **Consistent Estimator** for parameter θ **if** $\forall\varepsilon > 0, \quad P(|T(X) - \theta| > \varepsilon) \to 0$ as $n \to \infty$

i.e. all *probability mass* of the *estimator* (*seen as a random variable*) is *asymptotically* on the value θ

e.g. $T = \bar{X}$ is and $H = X_1$ isn't a *consistent estimator* for μ

Maximum Likelihood Estimation

1. The likelihood function, $L(\theta) = \prod_{i=1}^n f(x_i \mid \theta)$ is the product of the n pmf/pdf viewed as a function of a parameter θ.
2. Take the natural log of the likelihood to get the log-likelihood function $l(\hat{\theta}) = log(L(\hat{\theta}))$ and collect terms involving θ.
3. Find the value of θ for which log-likelihood is maximised. This is typically done by finding $\hat{\theta}$ that solves $l'(\hat{\theta}) = \frac{d}{d\theta} log(L(\hat{\theta})) = 0$.
4. If the estimate $\hat{\theta}$ obtained in step 3 corresponds to a maximum $\frac{d^2}{d\theta^2} l(\hat{\theta}) < 0$, then $\hat{\theta}$ is confirmed as the maximum likelihood estimator (MLE) of θ.

Remarks:

- In large sample sizes, the MLE progressively becomes unbiased, efficient and consistent. This can be proved under mild technical assumptions.
- In small sample sizes there is no such guarantee and the quality of a MLE can vary. Bayesian parameter estimation is an area of statistics that deals with this problem.
- MLE generalizes to multi-parameter distributions. Yet this requires multivariate calculus and the maximization may give more than one answer if l has several peaks (local maxima).

Central Limit Theorem (CLT)

The CLT is a general result for sums of random variables. In statistics, it helps to study the distribution of the sample mean. Let $X_1, X_2, \dots, X_n$ be now n independent and identically distributed (i.i.d.) random variables from any probability distribution with mean μ and variance σ^2 both finite.

$\lim_{n\to\infty} \frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$. Can also be written as:

$\lim_{n\to\infty} \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \bar{X} = \frac{S_n}{n} = \frac{\sum_{i=1}^n X_i}{n}$ is the sample mean.

Implications for the Sample Mean: The CLT thus implies that for large, but finite n, $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$

- A rule of thumb **for 'large n" is often n ≥ 30.**
- Amazingly, this result holds irrespective of the distribution of the $\{X_i\}$ (and including discrete random variables). CLT thus demonstrates that statistical regularity can arise even from the combination of highly-diverse random phenomena.
- If $X_i \sim N(\mu, \sigma^2), \forall i$, the result becomes exact even for finite n, since the sum of independent normal random variables is normally distributed.

Hypothesis Testing

The Null Hypothesis: there is no difference in user engagement between the two groups.

The Alternative Hypothesis: a difference exists, e.g., the new feature increases user engagement.

If the test group shows much higher engagement than the control group, then we will reject the null hypothesis and conclude that the evidence favours the alternative hypothesis.

Parametric tests: Suppose again we have a random i.i.d. sample $(X_1, \dots, X_n)$ of a random variable X from an unknown distribution P.

Parametric tests typically assume the sample comes from a parametric family $P(\cdot \mid \theta)$ and test whether we could reasonably assume $\theta = \theta_0$ for some particular value θ_0.

Hypotheses: Formally, let H_0 be the null hypothesis and H_1 be the alternative hypothesis.

Two-sided test: $H_0: \theta = \theta_0$ versus $H_0: \theta \neq \theta_0$.

One-sided test: $H_0: \theta = \theta_0$ versus $H_1: \theta < \theta_0$; $H_0: \theta = \theta_0$ versus $H_1: \theta > \theta_0$.

Usually, the null hypothesis is formulated with an equality sign (=), while the alternative hypothesis uses one of (≠, <, >).

Rejection Region for a Test Statistic: To test the validity of H_0, we choose a **test statistic** T(X) of the data for which we can find the distribution under H_0.

The "art" of hypothesis testing is to define the test by identifying a rejection region $R \subseteq \mathbb{R}$ of low probability values of T under the assumption that H_0 is true, so that $P(T \in R \mid H_0) = \alpha$ for some small probability α (say 5%). We call α the **significance level** of the test. α is the overall probability of all rejection region.

Confidence level = 1- α.

Rule: A well-chosen rejection region will have relatively high probability under H_1, whilst retaining low probability under H_0.

We calculate the observed test statistic t(x) for our sample x:

1. If $t \in R$ we "**reject** the null hypothesis at the 100α% level".
2. If $t \notin R$ we "**retain** the null hypothesis at the 100α% level".

Dealing with unknown variance: Student's t distribution:

$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$. $H_0: \mu = \mu_0$ **versus** $H_1: \mu \neq \mu_0$ **at the α level.**

Rejection region of observed test statistic $t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ is then give by $R = \left(-\infty, -t_{n-1,\frac{\alpha}{2}}\right) \cup \left(t_{n-1,\frac{\alpha}{2}}, \infty\right)$.

Two-sample t-test: $T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{S_P^2(\frac{1}{n_x} + \frac{1}{n_y})}} \sim t_{n_x + n_y - 2}$.

$$(\bar{x} - \bar{y}) \pm t_{\frac{\alpha}{2}, n_x + n_y - 2} \times s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

$$s_P^2 = \frac{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2}{n_x + n_y - 2} = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

p-Values: The **p-value** of the data is the probability of obtaining a test statistic at least as extreme as the one actually observed, assuming H_0 is correct. In other words, **the p-value is the maximum significance level at which we still reject the null hypothesis H_0 for that sample.** Thus, if we are given a fixed α, the null hypothesis H_0 is rejected if the p-value is less than or equal to α.

Rule: Smaller p-values suggest stronger evidence against H_0.

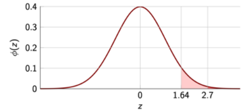
p-value in a One-sided Lower-Tailed Test ($H_1: \theta < \theta_0$):

- With known variance, the *p*-value is $\Phi(z)$.
- With unknown variance, the *p*-value is $F(t)$, where $F(\cdot)$ is the cdf of the Student's t distribution.

If $z = -2.7$ and $H_1: \theta < \theta_0$ then $p = 0.003467$. The lighter shading is R for $\alpha = 0.05$.

p-value in a One-sided Upper-Tailed Test ($H_1: \theta > \theta_0$):

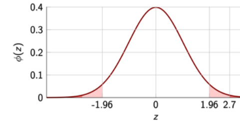
- With known variance, the *p*-value is $1 - \Phi(z)$.
- With unknown variance, the *p*-value is $1 - F(t)$.



If $z = 2.7$ and $H_1: \theta > \theta_0$ then $p = 0.003467$. The lighter shading is R for $\alpha = 0.05$.

p-value in a Two-sided Test:

- With known variance, the *p*-value is $2 \times (1 - \Phi(|z|))$.
- With unknown variance, the *p*-value is $2 \times (1 - F(|t|))$.



For example, if $z = 2.7$ then the **two-sided p-value** is $p = 0.006934$. The lighter shading is R for $\alpha = 0.05$.

p-value is the probability that a test statistic is at least as extreme as observed. Thus for fixed α, we reject $H0$ if $p \le \alpha$.

Side	Tail	Var	P-Value
1	Low	σ^2	$p = \Phi(z)$
1	Low	S^2	$p = F(t)$ †
1	Up	σ^2	$p = 1 - \Phi(z)$
1	Up	S^2	$p = 1 - F(t)$ †
2	-	σ^2	$p = 2(1 - \Phi(\|z\|))$
2	-	S^2	$p = 2(1 - F(\|t\|))$†

† F is the CDF of the t-distribution.

Simulation

Discrete-Event Simulation (DES): A (discrete-event) simulation is a program that generates a random **sample path** through a state transition system, where time delays are associated with each state.

There is a single global "clock" – a virtual time. Not to be confused with elapsed real time.

State transitions are triggered by **events** which are ordered in time on a virtual timeline, an event "diary".

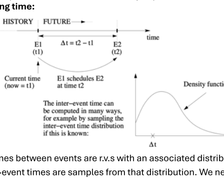
DES involves invoking events in time order; if an event is invoked ("occurs", "fires", "is triggered", ...) at virtual time t the clock is updated to t and the code for the event:

- Updates the model state.
- Schedules zero or more new future events on the time line.

Note that the state is unchanged between events.

- In practice, DES is based on a few core design principles:
- The virtual time is a floating-point number (call it *now*)
- The state is defined by a set of program variables, which are typically discrete (e.g. booleans, integers, ...)
- The timeline is a priority queue of *(Event, time)* pairs, ordered by virtual time – essentially an event *diary*.
- Events are implemented as objects, functions, procedures, methods, etc.
- A **scheduler** adds new *(Event, time)* pairs to the diary
- A **descheduler** similarly removes them from the diary.
- Additional measurement variables and code need to be added in order to accumulate performance measures (otherwise there will be no output!)

Evolving time:



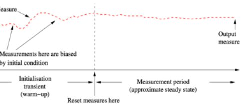
The times between events are r.v.s with an associated distribution – inter-event times are samples from that distribution. We need to be able to sample these distributions.

Designing a simulation model:

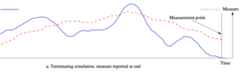
1. Identify the *entities* in the system that have to be modelled
2. Identify the *model states* (program state variables) – these specify where each entity is and what it is doing
3. Identify the *event types*, recalling that each state transition is triggered by an event (note that some events may be parameterisable, e.g. "arrival at location a")
4. For each *event*, specify i. how it changes the current state, ii. what new events need to be scheduled and what old events need to be cancelled (descheduled) when it fires
5. Add code to accumulate *measurements* whilst the simulation executes
6. Add code to *output results* when the program terminates, e.g. after T simulated time units, N occurrences of a specified event, etc.

Output Analysis

A **non-terminating simulation**, seeks to model a system at "equilibrium", a.k.a. "in steady state", where $p_s(t) \to p_s$ as $\to \infty$.



A **terminating simulation** models a system over a specified period during which there is no notion of equilibrium, e.g.



Equilibrium output analysis: We focus on non-terminating simulations.

Assume we are using simulation to estimate some "long-run" performance measure, e.g. mean population, mean response time, resource utilisation, etc.

The **initial state** is typically fixed (e.g. all queues empty), so the initial state probability distribution is different to the distribution after some time t >> 0, say, and measures take time to settle.

To avoid initialisation bias we must either:

1. Discard the initialisation transient by **resetting** the measured after some "warm-up" time has elapsed.
2. Measure for "long enough" to render any bias insignificant.

Confidence intervals: Discrete-event simulations are stochastic 随机的, so all outputs are random variables and each an **observation** of some measure, θ, of interest

If $X_i, 1 \le i \le n, n \ge 1$ are steady-state observations from a simulation then a **point estimate** for θ is the sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

If we draw a large number of samples, $100(1-\alpha)\%$ of them will have the mean value that lies under this interval.

For any desired coverage probability level $1 - \alpha$ we can define the **$100(1-\alpha)\%$ confidence interval** for μ by $[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$

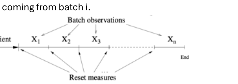
where z_α is the α-quantile of the standard normal.

For 95% confidence level: $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

If n>1, a **$100\alpha\%$ confidence interval** estimate for θ is $\bar{X} \pm t_{n-1,(1-\alpha)/2}\frac{S}{\sqrt{n}}$ where $t_{\nu,\alpha}$ is the α-quantile of the Student's t distribution with ν degrees of freedom.

Confidence intervals in simulation: How do we apply confidence intervals to simulation outputs?

- Running n independent replications, possibly in parallel, guarantees independence of the X_i.
- Another approach is to run the model once, wait for it to warm up and reach (approximate) equilibrium, then divide the measurement time into batches, with X_i coming from batch i.



- If each X_i is itself a mean then this is called the **batch means** method.
- Yet the X_i may not be independent because the state at the end of one batch is the same as that at the start of the next!

If the X_i are **dependent**, then we have to take **covariances** into account to build an exact confidence interval.

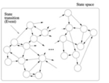
The above confidence interval makes the assumption that $Var(\bar{X}) = \sigma^2/n$, and then uses S^2 as an estimate of σ^2.

However, if the X_i are dependent, then it can be shown that $Var(\bar{X}) = \frac{\sigma^2}{n} + \frac{1}{n^2}[2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j)$.

If covariances are positive S^2/n becomes an **under-estimate** of $Var(\bar{X})$ and the computed confidence intervals are narrower than they should be.

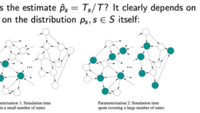
State-space coverage:

- Consider the underlying state transition system



s,

- Suppose the simulation is run for time T at (approximate) equilibrium and spends time T_s in state s
- T_s/T is then an estimate, $\hat{p}_s$, of (the unknown) state probability p_s
- Remark: Measures such as W could be computed directly, or in terms of $\hat{p}_s$ – the answer will be the same
- How good is the estimate $\hat{p}_s = T_s/T$? It clearly depends on T, but also on the distribution $p_s, s \in S$ itself:



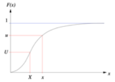
Filled states: largest subset $S' \subseteq S$ s.t. $\inf\{p_i \mid i \in S'\} \ge \sup\{p_j \mid j \in S - S'\}$ and $\sum_{i \in S'} p_i \le p_{max}$ for some p_{max}

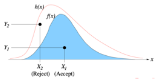
- For the same simulation time the C.I. for the left model (light load) should be narrower than the right (heavy load)

Distribution sampling

For r.v. X with support $supp(X)$, the objective is to define a sampling function: $U(0,1) \rightarrow supp(X)$ in terms of X's density/cdf (or pmf/cdf) function.

1. The Inverse Transform method: Suppose X is a continuous r.v. with cdf $F(x) = P(X \leq x)$ and that we are trying to sample X, Let U ~ U(0,1). Because F(x) increases monotonically, we have: $P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$, so **set $U = F(X)$ and invert to give $X = F^{-1}(U)$.**





(For Method 1) (For Method 2)

We can use the inverse transform method to sample a discrete r.v., X by inverting its cumulative distribution function, $F_X(x)$ (a"step function"). If $U \sim U(0,1)$, then the inverse transform methods returns $X = min\{x : F(x) \geq U\}$.

2. The Acceptance-Rejection (AR) Method: If F(x) cannot be explicitly inverted (e.g., normal cdf) we can sometimes work with the corresponding density function f(x). We choose a density function g(x) easy to sample from. Now we try to find a constant, c, so that **c g(x) = h(x) dominates f (x) for all x (i.e. c g(x)>f(x))**:

By construction, c is the area under h(x): $c = c\int_x g(x)dx = \int_x h(x)dx$.

To find c: We need to maximize $\frac{f(x)}{g(x)}$: Differentiate $\frac{f(x)}{g(x)}$, let $\frac{d}{dx}\frac{f(x)}{g(x)} = 0$ to find the maximum value of x, calculate c using the value of x ($c = max\frac{f(x)}{g(x)}$).

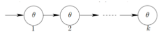
AR Algorithm:

1. Let X be a sample from the r.v. whose density function is g(x)
2. Generate a U(0,1) sample, U, and let $\boldsymbol{Y = Uh(X)}$
3. If $Y \leq f(X)$,i.e.if $U \leq \frac{f(X)}{h(X)} = \frac{f(X)}{cg(X)}$, where $U \sim (0,1)$, then **accept** X; otherwise **reject** it and start again.

It's a "dart throwing" exercise (Monte Carlo simulation).
By construction, the samples X and Y define a point that lies under h(X); if (X,Y) lies under f(X) as well we accept X.

3. The Convolution Method: Some random variables are defined as the sum of two or more independent random variables.
We can **sample the individual distributions and sum the results.**

- Example: An Erlang(k,θ) random variable, X say, is defined as the sum of k independent exponentially-distributed random variables X_i, each with rate parameter θ



- Notice that

$$E[X] = E[X_1 + \ldots + X_k] = \frac{1}{\theta} + \frac{1}{\theta} + \cdots + \frac{1}{\theta} = \frac{k}{\theta}$$

- We can generate Erlang(k,θ) samples using the sampler for the exponential distribution: if $X_i \sim \exp(\theta)$ then

$$X = \sum_{i=1}^{k} X_i \sim \text{Erlang}(k,\theta)$$

- If $U_i \sim U(0,1)$ then X_i is sampled using $-\log U_i/\theta$
- We can save the more expensive log calculations in the summation by turning the sum into a product:

$$X = \sum_{i=1}^{k} -\frac{\log U_i}{\theta} = -\frac{1}{\theta}\log\prod_{i=1}^{k} U_i$$

4. The Compositio Method:

- The *Composition Method* applies to density functions that have the following structure (called a *mixture*):

$$f(x) = w_1 f_1(x) + w_2 f_2(x) + \ldots + w_n f_n(x)$$

where the $f_i(x)$ are density (or mass) functions and where $0 < w_i < 1, 1 \leq i \leq n, n > 0$ and $\sum_{i=1}^{n} w_i = 1$
- The sampling algorithm is straightforward:
  - Pick i w.p. w_i (discrete distribution sampling)
  - Sample the distribution whose density is f_i, e.g. using one of the above methods

This applies also to cdfs and mass functions with a similar algorithm.

Misc

Combination: order doesn't matter. Select item without regard to how they are arranged. $nCr = \frac{n!}{(n-r)!r!}$.

Permutation: order is important. Count number of ways you can arrange a set of items where the sequence or arrangement matters. $nPr = \frac{n!}{(n-r)!}$.

$$\left(\sum_{i=0}^{n} X_i\right)^2 = \sum_{i=0}^{n}\sum_{j=0}^{n} X_i X_j$$

$$\frac{\sum(x-\overline{x})^2}{n} = \frac{\sum x^2}{n} - \overline{x}^2$$

Taylor Series: $e^x = \sum_{i=0}^{\infty}\frac{x^n}{n!}$

Summations: $\sum_{r=1}^{n} r = \frac{1}{2}n(n+1)$. $\sum_{r=1}^{n} r^2 = \frac{1}{6}n(n+1)(2n+1)$. $\sum_{r=1}^{n} r^3 = \frac{1}{4}n^2(n+1)^2$.

$\boldsymbol{f(x)}$	$\boldsymbol{\int f(x)\,dx}$
$\frac{1}{x}$	$ln\lvert x\rvert$
$\frac{1}{x^2-a^2}$	$\frac{1}{2a}ln\left\lvert\frac{x-a}{x+a}\right\rvert$ $(x > a)$
$\frac{1}{a^2-x^2}$	$\frac{1}{2a}ln\left\lvert\frac{a+x}{a-x}\right\rvert$ $(\lvert x\rvert < a)$

$\int uv'dx = uv - \int u'vdx$.

L(log)**A**(algebra)**T**(trigonometry)**E**(exponent)

$$\int\frac{f'(x)}{f(x)}dx = ln|f(x)|$$

$$\int f^n(x)f'(x)dx = \frac{1}{n+1}f^{n+1}(x) + C$$

$f(x)$	$f'(x)$
$ln\,x$	$\frac{1}{x}$
uv	$u'v + uv'$
$\frac{u}{v}$	$\frac{u'v - uv'}{v^2}$

$$x = rcos\theta, y = rsin\theta$$