1 Events

An **experiment**'s outcomes are defined by its **sample space** S. An event $E \subseteq S$ is a collection of possible outcomes. *Extreme* events are $\emptyset$ and S; *elementary* events are singleton subsets of S. For an **outcome** $s^* \in S$, an event E has occurred iff $s^* \in E \subseteq S$. $\emptyset$ will *never occur* and S will *always occur*. The event $\bigcup_i E_i$ will occur if any event E_x occurs, and $\bigcap_i E_i$ will occur if all events E_x occur. Events are **mutually exclusive** if $\forall i, j. E_i \cap E_j = \emptyset$. *An event occurs if any of its elements occur.*

To define a p.f. on S we agree on a collection of subsets of S to assign probability to, a σ-algebra $\mathcal{F}$. This means $\forall E, E_1, \cdots$:

- **Non-Empty**: $S \in \mathcal{F}$.
- **Closed complements**: $E \in \mathcal{F} \Rightarrow \overline{E} \in \mathcal{F}$.
- **Closed countable unions**: $\bigcup_i E_i \in \mathcal{F}$.

A **probability measure** on $\langle S, \mathcal{F} \rangle$ is a mapping $P : \mathcal{F} \to [0,1]$, satisfying the following axioms $\forall E$ on which it is defined:

- $\forall E \in \mathcal{F}. 0 \le P(E) \le 1$.
- $P(S) = 1$.
- **Countably additive**: for *mut. excl.* $E_1, \cdots \in \mathcal{F}$, we have $P(\bigcup_i E_i) = \sum_i P(E_i)$.

It is easy to derive $P(\emptyset) = 0$, $P(\overline{E}) = 1 - P(E)$ and *for any* E_1, E_2: $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cup E_2)$. Also, a **joint event** $E \cap F$ is **independent** iff $P(E \cap F) = P(E)P(F)$. More generally, $\{E_1, \cdots\}$ are independent if for any finite subset $\{E_{i_1}, E_{i_2}, \cdots\}$ where $\{i_j \mid 1 \le j \le n\}$, we have $P(\bigcap_{j=1}^n E_{i_j}) = \prod_{j=1}^n P(E_{i_j})$.

The **conditional prob** of E occuring given F with $P(F) \neq 0$:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

If E and F are independent, $P(E|F) = P(E)$. Also, $P(\cdot|F)$ defines a valid probability measure. E_1 and E_2 are **condtionally independent** given F iff $P(E_1 \cap E_2|F) = P(E_1|F)P(E_2|F)$.

The **law of total probability** states $\forall$ partitions of S: $\{F_1, \cdots\}$, and events $E \subseteq S$:

$$P(E) = \sum_i P(E|F_i)P(F_i)$$

Bayes' Theorem states for any $E, F \subseteq S$:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

2 Combinatorics

- **Multiplication Rule:** For independent events: $P(A \cap B) = P(A) \cdot P(B)$
- **Addition Rule:** For mutually exclusive events: $P(A \cup B) = P(A) + P(B)$
- **Combinations (unordered):** $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- **Permutations (ordered):** $P(n,k) = \frac{n!}{(n-k)!}$
- **Multinomial Coefficient:** Number of ways to divide n objects into r groups of sizes $k_1, k_2, \ldots, k_r$: $\frac{n!}{k_1!k_2!\ldots k_r!}$
- **Multinomial Probability:** For n independent trials with r outcomes: $P = \frac{n!}{k_1!k_2!\ldots k_r!} \cdot p_1^{k_1} p_2^{k_2} \ldots p_r^{k_r}$
- **Binomial Probability (2 outcomes):** $P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$
- **Complement Rule:** $P(A) = 1 - P(A^c)$
- **Conditional Probability:** $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
- **Expected Value (Discrete):** $\mathbb{E}[X] = \sum x \cdot P(X = x)$

3 Random Variables

A **probability space** is $\langle S, \mathcal{F}, P \rangle$. A **random variable** is a mapping $X : S \to \mathbb{R}$. Finite set of outcomes means *simple*, countable means *discrete*, otherwise *continuous*.

Induced prob.: P_X is a new PF on RV X with $\forall x \in \mathbb{R}$ let $S_X \subseteq S$ be $S_X = \{s \in S \mid X(s) \le x\}$, then $P_X(X \le x) \equiv P(S_X)$. The **image** of S under X is the **support** of X: $\text{supp}(X) = X(S) = \{x \in \mathbb{R} \mid \exists s \in S. X(s) = x\}$. $P_X(X \le x)$ is defined $\forall x \in \text{supp}(X)$.

The **CDF** of RV X is $F_X(x) = P_X(X \le x)$. F_X is *right-continuous*, meaning for decreasing seq. $x_1, \cdots \to x_\infty$, then $F_X(x_1), \cdots \to F_X(x_\infty)$. A valid CDF:

- **Monotonic**: $\forall x_1, x_2 \in \mathbb{R}. x_1 < x_2 \Rightarrow F_X(x_1) \le F_X(x_2)$.
- $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- F_X is **right continuous**.

The first two imply $\forall x \in \mathbb{R}. F_X(x) \in [0,1]$. For finite intervals $(a,b] \subseteq \mathbb{R}$, we can check $P_X(a < X \le b) = F_X(b) - F_X(a)$ by noting $E = \{X \le b\}$ may be rewritten as $E = (-\infty, a] \cup (a, b]$.

4 Discrete Random Variables

An RV X is **discrete** iff $\text{supp}(X) = \{x_1, \cdots\}$ is *countable*. If $\text{supp}(X)$ is ordered s.t. $x_1 < x_2 < \cdots$; then $S_X = \{s \in S \mid X(s) \le X\}$ is constant as we increase x in interval $[x_{i-1}, x_i)$. Once $x = x_i$, S_X grows larger to include outcomes that map to x_i. Thus, F_X will be a monotonic increaasing step function with vertical jumps at points in $\text{supp}(X)$. $P_X(X = x_i) = F_X(x_i) - F_X(x_{i-1})$.

For DRV X we define **PMF** $p(x) = P_X(X = x)$. If X can take values in $\text{supp}(X)$ then $\forall x \in \mathbb{R}. 0 \le p(x) \le 1$ and $\sum_i p(x_i) = 1$.

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

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

The **expectation** of X, $E[X] = \sum_x x p(x)$ is the weighted avg of possible values of X, or the **mean** of the distribution:

- $E[g(X)] = \sum_x g(x) p(x)$
- $\forall a, b \in \mathbb{R}. E[aX + b] = aE[X] + b$
- $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$

$E[X^n]$ is the *n*-**th moment** of X. The **central moment** is recentered to characterize deviation from the mean. The **variance** of X is the *second central moment* of X:

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

The **standard deviation** is the sqrt of the variance. $\forall a, b \in \mathbb{R}. \text{Var}(aX + b) = a^2 \text{Var}(X)$. The **skewness** of X is a measure of its *assymetry*, $\gamma_1 = \frac{E[(X - E[X])^3]}{\text{sd}(X)^3}$.

Let $S_n = \sum_{i=1}^n X_i$ be the *sum* of n non independent RVs of unkown distributions, and $\overline{X} = \frac{S_n}{n}$ be their average:

$$E[S_n] = \sum_{i=1}^n E[X_i]$$

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

If the vars are *independent*:

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

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

If the vars are *also identically distributed*:

$$E[\overline{X}] = \mu_X$$

$$\text{Var}(\overline{X}) = \frac{1}{n} \sigma_X^2$$

4.1 Bernoulli Distribution

An experiment with two possible outcomes $X \sim \text{Bernoulli}(p)$ with $p(x) = p^x(1-p)^{1-x}$ *for* $x \in \{0,1\}$. It follows that $\mu = p$ and $\sigma^2 = p(1-p)$.

4.2 Binomial Distribution

An experiment with n identical Bernoulli trials $X \sim \texttt{Binomial}(n, p)$ with $p(x) = \binom{n}{x} p^x (1-p)^{(n-x)}$, remembering $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. Also, $\mu = np$, $\sigma^2 = np(1-p)$ and $\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}}$.

4.3 Geometric Distribution

Consider a potentially infinite sequence of independent Bernoulli(p) RVs. Let X be the first successful trial, then $X \in \mathbb{N}^+$ and $X \sim \text{Geometric}(p)$ with $p(x) = p(1-p)^{x-1}$. Also, $\mu = \frac{1}{p}$, $\sigma^2 = \frac{1-p}{p^2}$ and $\gamma_1 = \frac{2-p}{\sqrt{1-p}}$.

4.4 Poisson Distribution

Poisson is concerned with number of random events happening per *unit* space. For $\lambda > 0$, $X \sim \text{Poisson}(\lambda)$ with $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$. Also $\mu = \sigma^2 = \lambda$ and $\gamma_1 = \frac{1}{\sqrt{\lambda}}$. For non-unit intervals, λt replaces λ, where λ is the rates at which events occur, and t is a time period.

4.5 Discrete Uniform Distribution

If $X \in \{1, \cdots, n\}$ then $X \sim U(\{1, \cdots, n\})$ with $p(x) = \frac{1}{n}$. Also, $\mu = \frac{n+1}{2}$ and $\sigma^2 = \frac{n^2-1}{12}$.

5 Continuous Random Variables

An RV X is **continuous** if $\exists f_X : \mathbb{R} \to \mathbb{R}$ such that $F_X(x) = \int_{-\infty}^x f_X(u)\, du$. Then f_X is the **pdf** of X, and $P_X(a < X \le b) = \int_a^b f_X(x)\, \text{d}x$. Hence, $\forall x \in \mathbb{R}. P_X(X = x) = 0$, hence the *support of a CRV must be uncountable* to sum to 1. $f_X(x) = \frac{\text{d}}{\text{d}x} F_X(x)$. The pdf is *non-negative*, and $\int_{-\infty}^{\infty} f_X(x)\, \text{d}x = 1$.

For CRV X, $E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$ and $Var(X) = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)\, dx$. The **$\alpha$-quartile** $Q_X(\alpha)$ for $0 \le \alpha \le 1$ is the least number satisfying $P(X \le Q_X(\alpha)) = \alpha$: $Q_X(\alpha) = F_X^{-1}(\alpha)$. e.g. the median of X solves $F_X(x) = 0.5$.

5.1 Continuous Uniform Distribution

If $X \in (a,b)$ is uniformly distributed, $X \sim U(a,b)$ with $f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & o.w. \end{cases}$ and $F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \ge b \end{cases}$. Also $\mu = \frac{a+b}{2}$ and $\sigma^2 = \frac{(b-a)^2}{12}$.

5.2 Exponential Distribution

If CRV X is exponentially distributed with rate $\lambda > 0$, $X \sim \exp(\lambda)$ with $f(x) = \lambda e^{-\lambda x}$ *for* $x \ge 0$ and $F(x) = 1 - e^{-\lambda x}$ *for* $x \ge 0$. Also $\mu = \frac{1}{\lambda}$ and $\sigma^2 = \frac{1}{\lambda^2}$.

The **memoryless property** states $\forall s, t \ge 0. P(X > s + t \mid X > s) = P(X > t)$. e.g. if we have waited s time for a random event, this doesn't affect how long we have left to wait.

If random events occur with Poisson(λ), the time between them $\sim \exp(\lambda)$.

5.3 Normal Distribution

A normal RV $X \sim N(\mu, \sigma^2)$ with $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ and $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt$.

When $\mu = 0$ and $\sigma = 1$ we get **standard normal** $Z \sim N(0,1)$ with $f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ and $F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{t^2}{2}}\, dt$. We can **standardize** with $X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$. Hence, $F_X(x) = \Phi(\frac{x-\mu}{\sigma})$, and $P(Z > z) = 1 - \Phi(z) = \Phi(-z)$.

5.4 Lognormal Distribution

If $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ then Y has a **longnormal dist.** with $f_Y(y) = \frac{1}{\sigma y \sqrt{2\pi}} \exp\left\{-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right\}$.

6 Moment Generating Functions

The **MGF** of CRV X is $M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$, or for DRV Y is $M_Y(t) = E[e^{tY}] = \sum_{y_i \in supp(Y)} e^{ty_i} p(y_i)$. This provides an alternative way to obtain $E[X^n] = \frac{\text{d}^n}{\text{d}t^n} M_X(t)|_{t=0}$.

The **characteristic func** modifies the mgf and is defined $\forall$ RVs: $\phi_X(t) = M_X(it) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx$ and $E[X^n] = i^{-n} \frac{\text{d}^n}{\text{d}t^n} \phi_X(t)|_{t=0}$.

> Since $E[\prod_{i=1}^n Z_i] = \prod_{i=1}^n E[Z_i]$, we have $M_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n M_{X_j}(t)$.

7 Random Variable Inequalities

The **markov inequality** states for any RV $X \ge 0$: $\forall a > 0. \left[P(X \ge a) \le \frac{E[X]}{a}\right]$.

The **chebyshev inequality** states for RV X: $\forall k > 0. \left[P(|X - \mu| \ge k) \le \frac{\sigma^2}{k^2}\right]$. This can be proven by applying the markov inequality to $Y = (X - \mu)^2$ and $a = k^2$.

8 Joint Random Variables

$\forall \langle x, y \rangle \in \mathbb{R}^2$ let $S \supseteq S_{xy} = \{s \in S \mid X(s) \le x \wedge Y(s) \le y\}$. Then when $Z = \langle X, Y \rangle$, $F(x,y) = P_Z(X \le x, Y \le y) = P(S_{xy})$. The **marginal CDF** $F_X(x) = F(x, \infty)$ and $F_Y(y) = F(\infty, y)$.

- $\forall x, y \in \mathbb{R}. 0 \le F(x,y) \le 1$
- **Monotonic**: $\forall x_1, x_2, y_1, y_2 \in \mathbb{R}. [(x_1 < x_2 \Rightarrow F(x_1, y_1) \le F(x_2, y_1) \wedge (y_1 < y_2 \Rightarrow F(x_1, y_1) \le F(x_1, y_2))]$.
- $\forall x, y \in \mathbb{R}. [F(x, -\infty) = F(-\infty, y) = 0 \wedge F(\infty, \infty) = 1]$.

$$P_Z(x_1 < X \le X_2, y_1 < Y \le y_2) =$$
$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$$

We can define **joint PMF** as $p(x,y) = P_Z(X = x, Y = y)$, and **marginal PMF** as $p_X(x) = \sum_y p(x,y)$ and $p_Y(y) = \sum_x p(x,y)$. $\forall x, y \in \mathbb{R}. 0 \le p(x,y) \le 1$ and $\sum_y \sum_x p(x,y) = 1$.

We can define **joint PDF** as $f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$ s.t. $F(x,y) = \int_{-\infty}^y \int_{-\infty}^x f(u,v)\, \text{d}u\, \text{d}v$ and **marginal PDFs** as $f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, \text{d}y$ and $f_Y(y) \int_{-\infty}^{\infty} f(x,y)\, \text{d}x$.

8.1 Joint Definition On Subsets

Let X, Y be random variables on sample space S with probability measure P. For subsets $B_X, B_Y \subseteq \mathbb{R}$, the joint probability is: $P_{XY}(B_X, B_Y) = P(\{\omega \in S : X(\omega) \in B_X, Y(\omega) \in B_Y\})$ That is, $P_{XY}(B_X, B_Y) = P(X \in B_X, Y \in B_Y)$.

8.2 More Joint Stuff

1. Joint PDF / PMF: - $f_{X,Y}(x,y)$: probability density (or mass) of (X,Y) - Must satisfy: $\iint f_{X,Y}(x,y)\, dx\, dy = 1$

2. Marginals: - $f_X(x) = \int f_{X,Y}(x,y)\, dy$ - $f_Y(y) = \int f_{X,Y}(x,y)\, dx$

3. Independence: - $X \perp Y$ iff $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

4. Conditional Density: - $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ (if $f_Y(y) > 0$)

5. Expectation: - $\mathbb{E}[g(X,Y)] = \iint g(x,y) f_{X,Y}(x,y)\, dx\, dy$

6. Covariance: - $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

7. Correlation: - $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$

8. Law of Total Expectation: - $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$

9. Sum of Independent RVs: - $f_Z(z) = \int f_X(x) f_Y(z-x)\, dx$ (convolution)

10. Transformation: - For $Z = g(X,Y)$:

$$P(Z \in B) = \iint_{(x,y) \in g^{-1}(B)} f_{X,Y}(x,y)\, dx\, dy$$

8.3 Convolution Theorem

Let X, Y be independent continuous random variables with PDFs $f_X(x), f_Y(y)$. Then the PDF of $Z = X + Y$ is the **convolution** of f_X and f_Y: $f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$.

- Valid iff X and Y are **independent**.
- $P(Z = z) = \sum_k P(X = k) P(Y = z - k)$.

- Only - Same idea for discrete case: $P(Z = z) = \sum_k P(X = k) P(Y = z - k)$ - Convolution mixes the distributions to give the distribution of the sum.

9 Independence & Expectation

X and Y are **independent** iff $\forall x, y. [F(x,y) = F_X(x) F_Y(y)]$, implying $\forall x, y. [p(x,y) = p_X(x) p_Y(y)]$ and $\forall x, y. [f(x,y) = f_X(x) f_Y(y)]$. Hence:

- If $g(X,Y) = g_1(X) + g_2(Y)$ then $E[g(X,Y)] = E[g_1(X)] + E[g_2(Y)]$.
- If $g(X,Y) = g_1(X) g_2(Y)$ and X, Y are *independent* then $E[g(X,Y)] = E[g_1(X)] E[g_2(Y)]$.
- Hence, $E[XY] = E[X]E[Y]$ if X, Y are independent.

For an RV X, $\sigma_X^2 = E[(X - \mu_X)^2]$. The bivariate ext of this is the **covariance** $\sigma_{XY} = Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$. When X, Y independent, $\sigma_{XY} = 0$.

Covariance measures how RVs change in relation to one another. The **correlation coeff.** $\rho_{XY} = Cor(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$. When X, Y are independent, $\rho_{XY} = 0$.

9.1 Multivariate Normal Distribution

A random vec $X = \langle X_1, \cdots X_n \rangle$ with $\mu = \langle \mu_1, \cdots \mu_n \rangle$ is **multivariate normal** with $f_X = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$ where Σ is the *positive definite* **covariance matrix** of X:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

$X_1, \cdots, X_n$ need not be independent.

10 Conditional Distributions

$$f(x|X>Y)=\frac{f(x)}{P(X>Y)}$$

A **conditional PMF** $p_{X|Y}(x \mid y)=\frac{p(x,y)}{p_Y(y)}$ is valid $\forall p_Y(y)>0$. **Bayes' Theorem** states:

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

A **conditional PDF** is $f_{X|Y}(x \mid y)=\frac{f(x,y)}{f_Y(y)}$. Now, X,Y are independent iff $\forall x,y \in \mathbb{R}.[f_{Y|X}(y \mid x)=f_Y(y)]$. **Bayes' theorem**:

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

A **conditional CDF** is $F_{X|Y}(x \mid y)=P(X\le x \mid Y=y)=\sum_{u=-\infty}^{x}p_{X|Y}(u \mid y)$ or $\int_{-\infty}^{x}f_{X|Y}(u \mid y)\,du$. From this, $P(a<X\le b \mid Y=y)=F_{X|Y}(b \mid y)-F_{X|Y}(a \mid y)$.

The **law of total probability** states:

1. $p_X(x)=\sum_y p_{X|Y}(x \mid y)p_Y(y)$.
2. $f_X(x)=\int_{-\infty}^{\infty}f_{X|Y}(x \mid y)f_Y(y)\,\mathrm{d}y$.
3. $F_X(x)=\int_{-\infty}^{\infty}F_{X|Y}(x \mid y)F_Y(y)\,\mathrm{d}y$.

The **conditional expectation** of *DRV* Y is $E_{Y|X}[Y \mid X=x]=\sum_y y\,p_{Y|X}(y \mid x)$.
The **conditional expectation** of *CRV* Y is $E_{Y|X}[Y \mid X=x]=\int_{-\infty}^{\infty}y\,f_{Y|X}(y \mid x)\,dy$.
In either case, expectation is a func of x but not Y.
The **law of total expectation** states $E_{Y|X}[Y \mid X]$ is an RV s.t. $E_Y[Y]=E_X[E_{Y|X}[Y \mid X]]$ for both discrete and cts.

11 Markov Chains

Discrete Time Markov Chains (DTMC) support arbitrary and dependent RVs:

- J is the **state space** of possible states.
- $X_{n\ge 0}\in J$, models the state at time n.
- Realization $X_0,X_1,\cdots$ is **sample path**.
- Goal: calculate $P(X_n=j)$.

We assume the **markov property** (next state depends only on current state): $P(X_{n+1}=j_{n+1} \mid X_n=j_n,\cdots,X_0=j_0)=P(X_{n+1}=j_{n+1} \mid X_n=j_n)$. We require an **initial prob vector** $\pi_0=[\pi_{0i}]^T$ *where* $P(X_0=i)=\pi_{0i}$ and **translation prob matrix** $R=[r_{ij}]$ *where* $r_{ij}=P(X_{n+1}=j \mid X_n=i)$. This gives rise to the following props:

- Each r_{ij} is independent of time n.
- Stuck states allowed (e.g. $r_{ii}=1$).
- R is a non-negative **stochastic** matrix (rows sum to 1).

In general, *transient analysis* shows that:

$$P(X_{n+1}=j \mid X_n=i)=r_{ij}$$
$$P(X_n=j \mid X_0=i)=(R^n)_{ij}$$
$$P(X_n=j)=(\pi_0R^n)_j$$
$$P(X_n=i)=\pi_{\infty i}$$

DTMC stabilize as a **limiting distribution**: $\pi_\infty=\lim_{n\to\infty}\pi_0R^n$ or **steady state distribution**: π^*_∞ that is invariant under R (i.e. $\forall n\ge 0\forall j\in J.\left[P(X_n=j)=1\pi^*_{\infty j}\right]$). These may *not* be unique. All limiting dists are steady state dists.
A DTMC is **irreducable** if the directed graph associated to R is **strongly connected**: $\forall\langle i,j\rangle\ \exists$ sample path from i to j. A DTMC is **periodic** if its states can only be visited at integer multiples of a fixed period. If it is *irreducable and aperiodic*:

- There exists unique $\pi_\infty=\pi^*_\infty$.
- The elements of π_∞ are >0.
- π_∞ solves $\pi_\infty R=\pi_\infty$ subject to $\sum_i\pi_{\infty_i}=1$. *Don't worry about the last case. Simply subssite st first few are valid.*

Without aperiodicity, an irreducable DTMC has no valid limiting distribution, however $\exists\pi^*_\infty$ s.t. π^*_∞ solves $\pi^*_\infty=R\pi^*_\infty$ subject to $\sum_i\pi^*_{\infty_i}=1$.

12 Estimation Theory

A **sample** of a **population**, $x=\langle x_1,\cdots,x_n\rangle$ is a realisation of RVs $X=\{X_1,\cdots,X_n\}$. A single draw follows $P(\cdot \mid \theta)$ where $\theta=\langle\theta_1,\cdots,\theta_n\rangle$ are the **params** to estimate, assuming X_i are *independent & identically distributed (iid)*. A **statistic** $T(X)$ is an RV:

- If approxes θ, T is an **estimator** of θ.
- Realisation $t(x)$ is an **estimate** of θ.
- We study $P(T \mid \theta)$ and its moments.

The **bias** of T is $\text{bias}(T)=E[T \mid \theta]-\theta$. For any X, the sample mean $\overline{X}$ is an unbiased estimate for μ: $E[\overline{X}]=E\left[\frac{\sum_{i=1}^n X_i}{n}\right]=\mu$.
For variance, we use **Bessel's Correction**: $E[S^2]=E[\frac{1}{n-1}\sum_{i=1}^n(X_i-\overline{X})^2]=\sigma^2$.
T is **more efficient** than H if $\forall\theta.[\text{Var}(T \mid \theta)\le\text{Var}(H \mid \theta)]$ and $\exists\theta.[\text{Var}(T \mid \theta)<\text{Var}(H \mid \theta)]$. If $\forall H$ T is more eff. than H, then T is **efficient**.
T is **consistent** if $\forall\epsilon>0.[P(|T(X)-\theta|>\epsilon)\to 0 \text{ as } n\to\infty]$, or if it is unbiased and $\lim_{n\to\infty}\text{Var}(T(X))=0$.
Sample Variance as a Biased Estimator: Let $X_1,X_2,\ldots,X_n$ be a sample from a population with mean μ and variance σ^2. The sample variance is defined as:

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

where $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean.

12.1 Bias in Sample Variance

We want to show that $\mathbb{E}[S^2]\ne\sigma^2$. Start by expanding S^2:

$$S^2=\tfrac{1}{n-1}\sum_{i=1}^n(X_i-\bar{X})^2$$
$$=\tfrac{1}{n-1}\left(\sum_{i=1}^n(X_i-\mu)^2-n(\bar{X}-\mu)^2\right)$$

Taking the expectation:

$$\mathbb{E}[S^2]=\tfrac{1}{n-1}\left(\mathbb{E}\left[\sum_{i=1}^n(X_i-\mu)^2\right]-n\mathbb{E}[(\bar{X}-\mu)^2]\right)$$

We know that:

$$\mathbb{E}[(X_i-\mu)^2]=\sigma^2\quad\text{for each } i$$

Thus, $\mathbb{E}[\sum_{i=1}^n(X_i-\mu)^2]=n\sigma^2$. Also, $\mathbb{E}[(\bar{X}-\mu)^2]=\frac{\sigma^2}{n}$, so:

$$\mathbb{E}[S^2]=\tfrac{1}{n-1}\left(n\sigma^2-n\cdot\tfrac{\sigma^2}{n}\right)=\tfrac{n-1}{n}\sigma^2$$

Therefore, $\mathbb{E}[S^2]=\frac{n-1}{n}\sigma^2$ shows that the sample variance is a biased estimator of the population variance.
Correction Factor: The bias can be corrected by using the factor $\frac{n}{n-1}$, resulting in the unbiased estimator:

$$\hat{\sigma}^2=\tfrac{n}{n-1}S^2$$

12.2 Extra Bias Notes

1. Bias of an Estimator: The bias of an estimator $\hat{\theta}$ for a parameter θ is defined as:

$$\text{Bias}(\hat{\theta})=\mathbb{E}[\hat{\theta}]-\theta$$

If $\text{Bias}(\hat{\theta})=0$, $\hat{\theta}$ is an **unbiased estimator**. If $\text{Bias}(\hat{\theta})\ne 0$, $\hat{\theta}$ is **biased**.
2. Unbiased Estimators: For an estimator $\hat{\theta}$ to be unbiased:

$$\mathbb{E}[\hat{\theta}]=\theta$$

Common unbiased estimators:

- Sample mean: $\mathbb{E}[\bar{X}]=\mu$
- Sample variance (corrected): $\mathbb{E}[\hat{\sigma}^2]=\sigma^2$
- Sample proportion: $\mathbb{E}[\hat{p}]=p$

3. Mean Squared Error (MSE): MSE is used to measure the quality of an estimator and combines both bias and variance:

$$\text{MSE}(\hat{\theta})=\mathbb{E}[(\hat{\theta}-\theta)^2]=\text{Var}(\hat{\theta})+\left(\text{Bias}(\hat{\theta})\right)^2$$

MSE is minimized when the estimator is both unbiased and has minimal variance.
4. Consistency of Estimators: An estimator $\hat{\theta}_n$ is **consistent** for θ if:

$$\hat{\theta}_n\xrightarrow{P}\theta\quad\text{as}\quad n\to\infty$$

This means that as the sample size increases, $\hat{\theta}_n$ converges in probability to θ.
5. Efficient Estimators: An estimator is **efficient** if it has the smallest variance among all unbiased estimators. The Cramer-Rao lower bound gives the theoretical minimum variance for an unbiased estimator:

$$\text{Var}(\hat{\theta})\ge\frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information.
6. Bias-Variance Tradeoff: For many estimators (e.g., in regression), theres a tradeoff between bias and variance. Reducing bias often increases variance, and vice versa. The optimal estimator minimizes the MSE.

13 Maximum Likelihood

The **likelihood func** $L(\theta)=\prod_{i=1}^n f(x_i \mid \theta)$ is the product of n pdfs viewed as a func of θ. We can find $\hat{\theta}$ that solves $\frac{\mathrm{d}}{\mathrm{d}\theta}\log(L(\hat{\theta}))=0$. If $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log(L(\hat{\theta}))<0$, $\hat{\theta}$ is a **maximum likelihood estimator** of θ - the best estimate for the parameter is the one that maximizes the likelihood of the observed data.

14 Central Limit Theorem

Let $X_1,\cdots,X_n$ be iid RVs with mean μ and var σ^2. We know $E[S_n]=n\mu$ and $\text{Var}(S_n)=n\sigma^2$. Thus, $E[S_n-\mu]=0$ and $\text{Var}(S_n)=n\sigma^2$. Finally, $E[\frac{S_n-n\mu}{\sigma\sqrt{n}}]=0$ and $\text{Var}(\frac{S_n-n\mu}{\sigma\sqrt{n}})=1$: $\lim_{n\to\infty}\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$. If $X_i\sim N(\mu,\sigma^2)$, the result is exact.

15 Hypothesis Testing

Consider hyp H_0 that takes param θ and value θ_0. We can test with a *two sided* $H_1:\theta\ne\theta_0$ or *one sided* $H_1:\theta>\theta_0$. For **test statistic** T, we find a distribution under H_0. We define a **rejection region** $R\subseteq\mathbb{R}$ such that $P(T\in R \mid H_0)=\alpha$, the **significance level**. If $t\in R$, we reject H_0.
To test the mean, we define R as the tails of N: $R=(-\infty,-z_{\alpha/2})\cup(z_{\alpha/2},\infty)$. σ^2 may be unkown, but S^2 is known. In this case, use **t-distribution** with $\nu=n-1$ *degrees of freedom* s.t. $T=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim t_{n-1}$. Now, $R=(-\infty,-t_{\langle n-1,1-\alpha/2\rangle})\cup(t_{\langle n-1,1-\alpha/2\rangle},\infty)$.
The **p-value** is the probability that a test statistic is at least as extreme as observed. Thus for fixed α, we reject H_0 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	$p=2(1-\Phi(\lvert z\rvert))$
2	-	S^2	$p=2(1-F(\lvert t\rvert))$ †

† F is the CDF of the t-distribution.

16 Discrete Event Simulation

A **DES** generates a random **sample path** through a state transition system with time delays at each state. Times between events are RVs - getting a sample path involves sampling these. To design a DES:

1. Identify the **entities** to be modelled.
2. Identify the **model states**.
3. Identify the **event types**.
4. For each **event** specify *how it changes curr state*, *what new events need to be cancelled/scheduled when it fires*.
5. Add code to calc **measurements** when the sim is running.
6. Add code to **output results**.

17 Output Analysis

A **non-terminating sim** seeks to model a system at **equilibrium** ($\forall s\in\text{States}.[\text{as } t\to\infty, p_s(t)\to p_s]$). A **terminating sim** models a system over a period with no notion of equilibrium. **Initial state** is fixed, and distribution changes after $t\gg 0$, which takes some time to converge. To avoid **initialization bias**, we discard *initialization transient* by resetting measures after some warm up time, or render long-enough to make bias insignificant.
DES are **stochastic**, so outputs are RVs and observations of a measure θ. If RVs $X_1,\cdots,X_n$ are steady-state observations from a sim, then an estimator for θ is the mean $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$.
By *CLT*, $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$. As n is large and σ^2 is known:

1. $P\left(-1.96\le\frac{\overline{X}-\theta}{\sigma/\sqrt{n}}\le 1.96\right)\approx 0.95$.
2. μ_0 is unknown, but by generating many intervals $[\overline{X}-1.96\frac{\sigma}{\sqrt{n}},\overline{X}+1.96\frac{\sigma}{\sqrt{n}}]$ using different simulations, we conclude with 95% confidence the true μ lies within one of the intervals.
3. $>$ **95% Confidence Interval** for μ.

To find a $100(1-\alpha)$% confidence interval for μ: $\overline{X}\pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. When the variance is unknown, we use S^2: $\overline{X}\pm t_{\langle n-1,1-\alpha/2\rangle}\frac{S}{\sqrt{n}}$.
Applying to DES: We could run the sim many times, once we reach a narrow confidence interval, we stop.
Another approach is to run the sim once until approx equilibrium is reached. Then, divide measurement into *batches*. If each X_i is sample mean of batch i, this is called the **sample means** method. X_i may be dependent, so we need *covariance* to construct the confidence interval:

$$\text{Var}(\overline{X})=\frac{\sigma^2}{n}+\frac{1}{n^2}\left[2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\text{Cov}(X_i,X_j)\right]$$

If covs are >0, then $\frac{S^2}{n}$ is an **under-estimate** of the var of $\overline{X}$ and the confidence interval is too narrow.

18 Distribution Sampling

Sims depend on the ability to **sample** cts random distributions. For RV X, we want a **sampling func** $U(0,1)\to\text{supp}(X)$.

18.1 Inverse Transform Method

Suppose X is a cts RV with CDF $F(x)=P(X\le x)$. Then by setting an RV $U\sim U(0,1)$ as $U=F(X)$, and solving for X (*invert*), we get a transformation from U to X. This also works for discrete RVs.

18.2 Acceptance Rejection Method

If $F(X)$ cannot be inverted, we choose a density function $g(x)$ that is easy to sample from. Now, we try to find a constant c s.t. $cg(x)=h(x)$ and $\forall x.h(x)\ge f(x)$. By construction, $c=c\int_X g(x)\,dx=\int_X h(x)\,dx$. $c=\max_{x\in\mathbf{supp}(X)}\frac{f(x)}{g(x)}$.

1. Let X be a sample from RV whose density function is $g(x)$.
2. Generate a $U(0,1)$ sample, U.
3. Let $Y=Uh(X)$.
4. If $Y\le f(x)$ (i.e. $U\le\frac{f(X)}{h(X)}$), then *accept* X, otherwise *reject* it and start again.

The probability of accepting X is $p=\frac{1}{c}$. Number of required iterations before accepting is *geometrically* distributed, so expected iterations $E[I]=c-1$.

18.3 Convolution Method

To sample a *sum of independent RVs*, sample them individually and then add the results.

18.4 Composition Method

Consider a discrete RV Y with $\text{supp}(Y)=\{1,\cdots,n\}$ and ctx RV X with $f_i(x)=f(x \mid Y=i)$. Now, we pick an i with probability $P(Y=i)$, then sample from density $f_i(x)$.

19 Common Formulae

- $e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!}$
- $\int x^{-1}\,\mathrm{d}x=\ln x+c$
- $\frac{d}{dx}(f(x)g(x))=f(x)\frac{d}{dx}g(x)+\frac{d}{dx}f(x)g(x)$
- $\int u\,dv=uv-\int v\,du$
- $\frac{d}{dx}e^{nx}=ne^{nx}$
- $\int e^{nx}\,dx=\frac{1}{n}e^{nx}+C\quad(\text{for } n\ne 0)$
- $\lim_{x\to c}\frac{f(x)}{g(x)}=\lim_{x\to c}\frac{f'(x)}{g'(x)}$
- $\frac{d}{dx}\frac{f(x)}{g(x)}=\frac{g(x)f'(x)-f(x)g'(x)}{[g(x)]^2}$
- $\frac{d}{dx}u=u\frac{dv}{dx}+v\frac{du}{dx}$
- $\frac{dy}{dx}=\frac{dy}{du}\frac{du}{dx}$