Probability Review

Liming Feng

Dept. of Industrial & Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Random variables

- A **random variable (r.v.)** is a function whose value is random, depending on the outcome of a random experiment
- **Probability mass function** of a discrete r.v. X

$$p(x_i) = \mathbb{P}(X = x_i), \quad i = 1, 2, \cdots$$

- **Probability density function (pdf)** of a continuous r.v. X

$$\mathbb{P}(a < X < b) = \int_a^b p(x)dx$$

- **Cumulative distribution function (cdf)** of a r.v. X

$$F(x) = \mathbb{P}(X \leq x)$$

- **Mean** (expectation, expected value)

$$\text{discrete:} \quad \mathbb{E}[X] = \sum_i x_i p(x_i)$$

$$\text{continuous:} \quad \mathbb{E}[X] = \int_{\mathbb{R}} x p(x) dx$$

- **Variance**, standard deviation

$$\sigma_X^2 = \text{var}(X) = \mathbb{E}(X - \mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

- Expectation of $f(X)$

$$\text{discrete:} \quad \mathbb{E}[f(X)] = \sum_i f(x_i)p(x_i)$$

$$\text{continuous:} \quad \mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x)p(x)dx$$

- Generally,

$$\mathbb{E}[f(X)] \neq f(\mathbb{E}[X])$$

For a convex function $f(x)$ (e.g., $f(x) = x^2$, $f(x) = e^x$), we have **Jensen's inequality**:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

- X and Y are **independent** iff

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

- Suppose X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

The converse is not true

- **Covariance** and **correlation**

$$\sigma_{XY} = \text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y], \quad \rho = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where σ_X, σ_Y are the standard deviations of X and Y

- Covariance of X and itself: $\text{cov}(X,X) = \text{var}(X)$
- Variance of $X+Y$

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$$

- Variance of $X_1 + \cdots + X_n$

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i,j} \text{cov}(X_i, X_j) = \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{1 \le i < j \le n} \text{cov}(X_i, X_j)$$

Parameter estimation

- Sample mean:

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

- Sample variance:

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

- Sample covariance:

$$\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- Sample correlation:

$$\hat{\rho} = \hat{\sigma}_{XY}/(\hat{\sigma}_X\hat{\sigma}_Y)$$

Normal distribution

- The pdf of a normal r.v. $X \sim N(\mu, \sigma^2)$:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  In particular,

$$\mathbb{E}[X] = \mu, \quad \mathrm{var}(X) = \sigma^2$$

- Normal distributions are also called **Gaussian** distributions
- **Standard normal distribution**: $\mu = 0$, $\sigma = 1$
- **Standardization**: if $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

- cdf of a standard normal r.v. $X \sim N(0,1)$:

$$\Phi(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy$$

- No analytical expression for $\Phi(x)$: in **Excel**, $normsdist(x)$, in **Matlab**, $normcdf(x)$
- **Symmetry**:

$$\Phi(-x) = 1 - \Phi(x)$$

In fact, suppose $X \sim N(0,1)$

$$\Phi(-x) = \mathbb{P}(X \leq -x) = \mathbb{P}(X \geq x) = 1 - \mathbb{P}(X < x) = 1 - \Phi(x)$$

- Suppose X and Y are independent, $X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$, then

$$X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

- For arbitrary normal r.v.'s X and Y, is $X + Y$ still normal?
- Suppose $X \sim N(\mu, \sigma^2)$, then

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- **Example (normal distribution):** *the daily returns of a certain stock in a week are denoted by* $r_1, r_2, \cdots, r_5$*. The weekly return is given by*

$$r = r_1 + \cdots + r_5.$$

Suppose the daily returns are independent and normally distributed with mean $\mu = 0.001$ *and standard deviation* $\sigma = 0.02$*. Which of the following is the distribution of the weekly return* r*?*

1. $r \sim 5N(\mu, \sigma^2) = N(5\mu, 25\sigma^2)$?
2. $r \sim N(5\mu, 5\sigma^2)$?

- **Lognormal distribution**: suppose $X \sim N(\mu, \sigma^2)$. Then $Y = \exp(X)$ is lognormally distributed.
- Stock price is assumed to have a lognormal distribution in the Black-Scholes-Merton model
- The mean of Y: $\mathbb{E}[Y] = \mathbb{E}[\exp(X)] = \exp(\mu)???$

Monte carlo simulation

- In financial engineering, often need to compute

$$\mu = \mathbb{E}[X]$$

  where $\mathbb{E}[X]$ has no analytical expression
- Simulate i.i.d. $\{X_i, i \geq 1\}$ with the same distribution as X (**independent and identically distributed**). Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

- Attractive for high dimensions (where other numerical methods may fail)

LLN and CLT

- **Law of large numbers:** suppose $X_1, X_2, \cdots$ are i.i.d. with finite expectation μ. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$$

- **Central limit theorem**: suppose $X_1, X_2, \cdots$ are i.i.d. with finite expectation μ and variance σ^2. Then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \Rightarrow N(0,1), \quad n \to \infty$$

- Characteristic rate of convergence of monte carlo simulation: $O(1/\sqrt{n})$

$$\bar{X}_n - \mu \approx \frac{\sigma}{\sqrt{n}} N(0,1)$$

- Error measurement: **standard error** $\sigma/\sqrt{n}$
- Let z_α, $\alpha \in (0,1)$, be the **percentage point** of $Z \sim N(0,1)$

$$\mathbb{P}(Z > z_\alpha) = \mathbb{P}(Z < -z_\alpha) = \mathbb{P}(|Z| > z_{\alpha/2}) = \alpha, \quad z_{0.025} = 1.96$$

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \Rightarrow \mathbb{P}\left(|\bar{X}_n - \mu| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx \alpha$$

Estimated standard error

- σ^2 is usually not known and needs to be estimated

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

- Report **estimated standard error**

$$\frac{s_n}{\sqrt{n}}$$

- For fixed n, we obtain an estimate $\bar{X}_n$ and a measure of the error $s_n/\sqrt{n}$
- Main goal of efficient monte carlo simulation: reducing s_n (**variance reduction techniques**)

Uniform and exponential distributions

- The pdf of a uniformly distributed r.v. $X \sim U[0,1]$:

$$p(x) = 1, \quad x \in [0,1]$$

  The cdf of $X \sim U[0,1]$:

$$F(x) = \mathbb{P}(X \leq x) = x, \quad \forall x \in [0,1]$$

- The pdf of an exponentially distributed r.v. $X \sim \exp(\lambda)$:

$$p(x) = \lambda e^{-\lambda x}, \quad x > 0$$

Inverse transform method

- **Simulate a continuous r.v.**

  Let $U \sim U[0,1]$. For a continuous r.v. X with cdf $F(x)$,

$$\begin{aligned} \mathbb{P}(X \leq x) &= F(x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= \mathbb{P}(F^{-1}(U) \leq x) \end{aligned}$$

  So X can be simulated from $F^{-1}(U)$

- **Example (simulate exponential r.v.s)**: for $X \sim \exp(\lambda)$,

$$F(x) = 1 - e^{-\lambda x} \Rightarrow F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$$

Replace x by a uniformly distributed r.v. $U \sim U[0,1]$. $U[0,1]$ can be simulated on computers (in **Excel**, *rand()*, in **Matlab**, *rand(row,col)*, in C/C++, *ran1* or *ran2* from *"Numerical Recipes in C"*, 1992, Press et al.)

- Simulate a standard normal r.v.
  - **Inverse transform method**: $\Phi^{-1}(U)$ where $U \sim U[0,1]$, $\Phi(x)$ is the cdf of $N(0,1)$
  - **Box-Muller**: Suppose $X, Y \sim N(0,1)$ are i.i.d. Let

$$X = r\cos(\theta), \quad Y = r\sin(\theta)$$

Then $r^2 \sim \exp(1/2)$ and $\theta \sim U[0, 2\pi]$, r and θ are independent.

Box Muller algorithm:

1. *Simulate two independent uniform r.v.s:* $U_1, U_2 \sim U[0,1]$
2. $r = \sqrt{-2\ln(U_1)}, \quad \theta = 2\pi U_2$
3. $X = r\cos(\theta), \quad Y = r\sin(\theta)$

Binomial distribution

- Binomial distribution $B(n, p)$ models the number of successes out of n independent Bernoulli trials (p is the success rate)
- If $X \sim B(n, p)$, then

$$\mathbb{P}(X = k) = \binom{k}{n} p^k (1-p)^{n-k}, \quad k = 0, 1, \cdots, n$$

In particular,

$$\mathbb{E}[X] = np, \quad \mathrm{var}(X) = np(1-p)$$

- **Example (binomial distribution)**: *Suppose we are observing the daily stock price movement of IBM. The stock price moves up with probability* 54%, *and otherwise with probability* 46%. *Suppose day-to-day price movements are independent.*
- *What is the probability the IBM stock price moves up in 4 days of a certain week (5 days per week)?*

$$\left(\begin{array}{c} 4 \\ 5 \end{array} \right) 0.54^4 \cdot 0.46^1 = 19.6\%$$

- *What is the expected number of up moves in a year (252 business days per year)?*

$$\mathbb{E}[X] = np = 252 \times 0.54 = 136.08$$

- **Poisson distribution** with arrival rate (intensity) λ

$$\mathbb{P}(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n = 0, 1, \cdots$$

In particular,

$$\mathbb{E}[N] = \lambda$$

E.g., the number of market shocks (arrivals of good/bad news) in a year

- **Poisson process** N_t

$$\mathbb{P}(N_t = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \quad n = 0, 1, \cdots$$