Topic 2: Scalar random variables

- Discrete and continuous random variables
- Probability distribution and densities (cdf, pmf, pdf)
- Important random variables
- Expectation, mean, variance, moments
- Markov and Chebyshev inequalities
- Testing the fit of a distribution to data

Definition of random variables

- A *random variable* is a function that assigns a real number, $X(s)$, to each outcome s in a sample space Ω.
  - Ω is the *domain* of the random variable
  - The set R_X of all values of X is its *range* $\Rightarrow R_X \subset \mathcal{R}$.
- The notation $\{X \leq x\}$ denotes a subset of Ω consisting of all outcomes s such that $X(s) \leq x$. Similarly for $\geq$, $=$ and $\in$.
- The function as a random variable must satisfy two conditions:
  - The set $\{X \leq x\}$ is an event for every x.
  - The probability of the events $\{X = \infty\}$ and $\{X = -\infty\}$ is zero:

$$P\{X = \infty\} = P\{X = -\infty\} = 0$$

Random variables

A random variable can be either discrete, continuous, or of mixed type.

$$X(s) : \Omega \to R_X$$

- Discrete variable: The range R_X is discrete, it can be either finite or countably infinite

  $$R_X = \{x_1, x_2, \ldots\}$$

  The sample space Ω can be discrete, continuous, or a mixture of both. $X(s)$ partitions Ω into the sets $\{S_i | X(s) = x_i \ \ \forall s \in S_i\}$.

- Continuous variable: The range is continuous. The sample space must also be continuous.

- Mixed type: The range is a combination of discrete values and continuous regions.

Distribution function

The distribution function of a random variable relates to the probability of an event described by the random variable. It is defined as

$$F_X(x) = P\{X \leq x\}$$

Properties of $F_X(x)$:

- $0 \leq F_X(x) \leq 1$
- $F(\infty) = 1$ and $F(-\infty) = 0$
- It is a non-decreasing function of x

  $$x_1 < x_2 \quad \to \quad F_X(x_1) \leq F_X(x_2)$$

- It is continuous from the right

  $$F_X(x^+) = \lim_{\epsilon \to 0} F_X(x + \epsilon) = F_X(x)$$

- $P\{X > x\} = 1 - F_X(x)$
- $P\{X = x\} = F_X(x) - F_X(x^-)$
- $P\{x_1 < X \leq x_2\} = F_X(x_2) - F_X(x_1)$

The distribution of different types of random variables

- Discrete: $F_X(x)$ is a stair-case function of x with jumps at a countable set of points $\{x_0, x_1, \ldots\}$

$$F_X(x) = \sum_k p_X(x_k) u(x - x_k)$$

where $p_X(x_k)$ is the probability of $\{X = x_k\}$.

- Continuous: $F_X(x)$ is continuous everywhere and can be written as an integral of a non-negative function

$$F_X(x) = \int_{-\infty}^{x} f_X(t) dt.$$

The continuity implies that at any point x,
$P\{X = x\} = F_X(x^+) - F_X(x) = 0$.

- Mixed: $F_X(x)$ has jumps on a countable set of points but is also continuous on at least one interval.

We will mostly study discrete *and* continuous *random variables.*

Discrete random variables – Pmf

A discrete random variable can be completely specified by its *probability mass function* $p_X(x)$

$$p_X(x) = P\{X = x\} \quad \text{for} \;\; x \in R_X$$

- $p_X(x) \geq 0$ for any $x \in R_X$
- $\sum_k p_X(x_k) = 1$ for all $x_k \in R_X$
- For any set A

$$P(X \in A) = \sum_k p_X(x_k) \quad \text{for all} \;\; x_k \in A \cap R_X$$

We use $X \sim p_X(x)$ or just simply $X \sim p(x)$ to denote discrete random variable X with pmf $p_X(x)$ or $p(x)$.

Some important discrete random variables

- *Bernoulli*: The success or failure of an experiment (Bernoulli trial).

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

  – Example: Flipping a bias coin.

- *Binomial*: The number of successes in a sequence of n independent Bernoulli trials.

$$p_X(k) = \binom{n}{k} p^k (1-p)^{N-k} \quad \text{for} \;\; k = 0, \ldots, n$$

  – Example: The number of heads in n independent coin flips.

- *Geometric*: The number of trials until the first success.

$$p_X(k) = (1-p)^{k-1} p \quad \text{for} \;\; k = 1, 2, \ldots$$

  The geometric probability is strictly decreasing with k.

  – Example: The number of coin flips until the first head shows up.

- *Poisson*: Number of occurrences of an event within a certain time period or region in space.

$$p_X(k) = \frac{\alpha^k}{k!} e^{-\alpha} \quad \text{for} \;\; k = 1, 2, \ldots$$

  where $\alpha \in \mathcal{R}^+$ is the average number of occurrences.

  – The Poisson probabilities can approximate the binomial probabilities. If n is large and p is small, then for $\alpha = np$

$$p_X(k) = \binom{n}{k} p^k (1-p)^{N-k} \approx \frac{\alpha^k}{k!} e^{-\alpha}$$

  The approximation becomes exact in the limit of $n \to \infty$, provided $\alpha = np$ is fixed.

Continuous random variables – Pdf

A continuous random variable can be completely specified by its *probability density function*, which is a nonnegative function such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt.$$

Properties of $f_X(x)$:

- $f_X(x) = \frac{dF_X(x)}{dx}$
- $f_X(x) \geq 0 \quad \text{for all } x \in \mathcal{R}$
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- $P\{X \in A\} = \int_A f_X(x) dx \quad \text{for any event } A \in \mathcal{R}$
- $P\{x_1 < X \leq x_2\} = \int_{x_1}^{x_2} f_X(x) dx$

However, $f_X(x)$ should not be interpreted as the probability at $X = x$. In fact, $f_X(x)$ is *not* a probability measure since it can be > 1.

Some important continuous random variables

- *Uniform* $U[a, b]$:

$$f_X(x) = \begin{cases} 0 & \text{for} \quad x < a \\ \frac{1}{b-a} & \text{for} \quad a \leq x \leq b \\ 0 & \text{for} \quad x > b \end{cases}$$

  - Example: A wireless signal $x(t) = A\cos(\omega t + \theta)$ has the phase $\theta \sim U[-\pi, \pi]$ because of random scattering.

- *Exponential*: $X \sim \exp(\lambda)$

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

  - Examples: The arrival time of packets at an internet router, cell-phone call durations can be modeled as exponential RVs.

- *Gaussian (normal)*: $X \sim \mathcal{N}(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\ , \quad \sigma > 0\ ,\ -\infty < x < \infty$$

  - When $\mu = 0$ and $\sigma = 1$, we call $f(x)$ the *standard Gaussian* density.

- The Gaussian distribution is very important and is often used in EE, for example, to model thermal noise in circuits, in communication and control systems.
- It also arises naturally from the sum of independent random variables. We will study more about this in a later lecture.
- The Q function

$$Q(\alpha) = \Pr[x \geq \alpha] = \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{+\infty} e^{-x^2/2} dx$$

  * Often used to calculate the error probability in communications.
  * Has no closed-form but good approximations exist.
  * A related function is the *complementary error function*

$$\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_{z}^{+\infty} e^{-x^2} dx = 2Q\left(\sqrt{2}z\right)$$

  Matlab has the command $\operatorname{erfc}(z)$.

- *Chi-square*: $X \sim \mathcal{X}_k^2$

$$f_X(x) = \frac{x^{k/2-1} e^{-x/2}}{\Gamma(k/2) 2^{k/2}}, \quad x \geq 0, \quad \text{where } \Gamma(p) := \int_0^{\infty} z^{p-1} e^{-z} dz$$

- Here k is called the *degree of freedom*. When k is an integer, $\Gamma(k) = (k-1)! = (k-1)(k-2)\ldots 2 \cdot 1$
- The chi-squared random variable X arises from the sum of k i.i.d. standard Gaussian RVs

$$X = \sum_{i=1}^{k} Z_i \ , \quad Z_i \sim \mathcal{N}(0,1) \ , \text{ independent}$$

- A $\mathcal{X}_2^2$ random variable ($k=2$) is the same as $\exp(\frac{1}{2})$.

- *Rayleigh*:

$$f_X(x) = \frac{x}{\lambda^2} e^{-(x/2)^2/2} \ , \quad x \geq 0$$

- Example: The magnitude of a wireless signal.

- *Cauchy*: $X \sim \text{Cauchy}(\lambda)$

$$f_X(x) = \frac{\lambda/\pi}{\lambda^2 + x^2} \ , \quad -\infty < x < \infty$$

- The Cauchy random variable arises as the tangent of a uniform RV.

Expectation

The *expected value* (also called *expectation* or *mean*) of a random variable X is defined

- for continuous X as:

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

- for discrete X as:

$$E[X] = \sum_k x_k p_X(x_k)$$

provided the integral or sum converges *absolutely* ($E[|X|] < \infty$).

- The mean can be thought of as the *average* value of X in a large number of independent repetitions of the experiment.
- $E[X]$ is the "center of gravity" of the pdf, considering $f_X(x)$ as the distribution of mass on the real line.

Questions: Find the mean of the following random variables: Binomial, Poison, uniform, exponential, Gaussian, Cauchy.

Variance and moments

- Expectation of a function of X

$$E[g(X)] = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x) & \text{for continuous } X\ dx \\ \sum_k g(x_k) p_X(x_k) & \text{for discrete } X \end{cases}$$

- The *variance* of a random variable X is defined as

$$\mathrm{var}(X) = E\left[(X - E[X])^2\right]$$

  - The variance provides a measure of the dispersion of X around its mean.
  - The variance is always non-negative.
  - The *standard deviation* $\sigma_X = \sqrt{\mathrm{var}(X)}$ has the same unit as X.

- The k^{th} *moment* of X is defined as

$$m_k = E\left[X^k\right]$$

  The mean and variance can be expressed in terms of the first two moments $E[X]$ and $E[X^2]$: $\quad \mathrm{var}(X) = E[X^2] - (E[X])^2$.

Properties of mean and variance

- Expectation is linear

$$E\left[\sum_{k=1}^{n} g_k(X)\right] = \sum_{k=1}^{n} E\left[g_k(X)\right]$$

- Let c be a constant scalar. Then

$$\begin{aligned} E[c] &= c & \mathrm{var}(c) &= 0 \\ E[X+c] &= E[X]+c & \mathrm{var}(X+c) &= \mathrm{var}(X) \\ E[cX] &= cE[X] & \mathrm{var}(cX) &= c^2\mathrm{var}(X) \end{aligned}$$

- Example: A random binary NRZ signal $x = \{1, 1, -1, -1, 1, -1, 1, \ldots\}$

$$x = \begin{cases} 1 & \text{with prob. } \frac{1}{2} \\ -1 & \text{with prob. } \frac{1}{2} \end{cases}$$

  - Mean $E[X] = 0$: the signal is unbiased.
  - Variance $\sigma_X^2 = 1$ is the average signal power.

  What happens to the mean and variance if you scale the signal to a different voltage V?

Markov and Chebyshev inequalities

For a Gaussian r.v., the mean and variance completely specify its pdf

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

In general, however, the mean and variance are insufficient in specifying a random variable (determining its pmf/pdf/cdf).

They can be used to bound the probabilities of the form $P[X \geq t]$.

- Markov inequality: For X nonnegative

$$P[X \geq a] \leq \frac{E[X]}{a}, \quad a > 0$$

  This bound is useful when the right-hand-side expression is < 1. It can be tight for certain distributions.

- Chebyshev inequality: For X with mean m and variance σ^2

$$P[|X - m| \geq a] \leq \frac{\sigma^2}{a^2}$$

  The Chebyshev inequality can be obtained by applying the Markov inequality to $Y = (X - m)^2$.

Testing the fit of a distribution to data

We have a set of observation data. How do we determine how well a model distribution fits the data?

The Chi-square test.

- Partition the sample space S_X into the union of K disjoint intervals.
- Based on the modeled distribution, calculate the expected number of outcomes that fall in the kth interval as m_k.
- Let N_k be the observed number of outcomes in the interval k.
- Form the weighted difference

$$D^2 = \sum_{k=1}^{K} \frac{(N_k - m_k)^2}{m_k}$$

If D^2 is small then the fit is good. If $D^2 > t_\alpha$ then reject.
Here t_α is a predetermined threshold based on the significant level of the test. It is calculated from $P[X \geq t_\alpha] = \alpha$, e.g. for $\alpha = 1\%$.