STATISTICAL COMPUTATIONAL METHODS

Review of Random Variables and Common Distributions

Random Variables

$(S, \mathcal{K}, P)$ a probability space.
Random variable: $X : S \to \mathbb{R}$ s. t. $\forall x \in \mathbb{R}$, the inverse image

$$X^{-1}\big((-\infty, x]\big) = \{e \in S |\ X(e) \le x\} = \big(X \le x\big) \in \mathcal{K}.$$

- $X(S) \subset \mathbb{R}$ a discrete subset, then **discrete random variable**;
- $X(S) \subseteq \mathbb{R}$ a continuous subset (interval), then **continuous random variable**.

Cumulative distribution function (cdf): $F : \mathbb{R} \to \mathbb{R},\ F(x) = P(X \le x)$.

Probability distribution (density) function (pdf):

- X d. r. v., $X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$, $p_i = P(X = x_i)$, $F(x) = \sum\limits_{x_i \le x} p_i$;
- X c. r. v., $f : \mathbb{R} \to \mathbb{R}$, $F(x) = \int\limits_{-\infty}^{x} f(t)dt$.

Expected value:

- X d. r. v., $E(X) = \sum\limits_{i \in I} x_i p_i$;
- X c. r. v., $E(X) = \int\limits_{\mathbb{R}} x f(x) dx$.

Variance: $V(X) = E\left((X - E(X))^2\right) = E\left(X^2\right) - (E(X))^2$.

Standard deviation: $\sigma(X) = \sqrt{V(X)}$.

Discrete Distributions

Bernoulli Distribution, *Bern*(p), with parameter $p \in (0,1)$:

pdf $X\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, $E(X) = p$, $V(X) = pq$.

Binomial Distribution, *Bino*(n, p), with parameters $n \in \mathbb{N}, p \in (0,1)$:

pdf $X\begin{pmatrix} k \\ C_n^k p^k q^{n-k} \end{pmatrix}_{k=\overline{0,n}}$, $E(X) = np$, $V(X) = npq$.

X is the number of successes in n Bernoulli trials, with probability of success p.

Discrete Uniform Distribution, *Unid*(m), with parameter $m \in \mathbb{N}$:

pdf $X\begin{pmatrix} k \\ \dfrac{1}{m} \end{pmatrix}_{k=\overline{1,m}}$, $E(X) = \dfrac{m+1}{2}$, $V(X) = \dfrac{m^2-1}{12}$.

Poisson Distribution, *Poiss*(λ), with parameter $\lambda > 0$:

pdf $X\begin{pmatrix} k \\ \dfrac{\lambda^k}{k!}e^{-\lambda} \end{pmatrix}_{k\in\mathbb{N}}$, $E(X) = V(X) = \lambda$.

X is the number of "rare events" that occur in a fixed period of time; λ is the average number of events occurring per time unit.

Geometric Distribution, *Geo*(p), with parameter $p \in (0,1)$:

pdf $X\begin{pmatrix} k \\ pq^k \end{pmatrix}_{k\in\mathbb{N}}$, cdf $F(x) = 1 - q^{x-1}$, for $x = 1, 2, \ldots$, $E(X) = \dfrac{q}{p}$, $V(X) = \dfrac{q}{p^2}$. **X is the number of <u>failures</u> that occur before the first success, in an infinite sequence of Bernoulli trials, with probability of success p.**

"Almost" Geometric Distribution, *AGeo*(p), with parameter $p \in (0,1)$:

pdf $X\begin{pmatrix} l \\ pq^{l-1} \end{pmatrix}_{l=1,2,\ldots}$, cdf $F(x) = 1 - q^x$, for $x = 1, 2, \ldots$, $E(X) = \dfrac{1}{p}$, $V(X) = \dfrac{q}{p^2}$.

X is the number of <u>trials</u> needed to get the first success, in an infinite sequence of Bernoulli trials, with probability of success p.

Negative Binomial (Pascal) Distribution, *Nbin*(n, p), with parameters $n \in \mathbb{N}, p \in (0,1)$:

pdf $X\begin{pmatrix} k \\ C_{n+k-1}^k p^n q^k \end{pmatrix}_{k\in\mathbb{N}}$, $E(X) = \dfrac{nq}{p}$, $V(X) = \dfrac{nq}{p^2}$.

X is the number of failures that occur before the n^{th} success, in an infinite sequence of Bernoulli trials, with probability of success p.

Continuous Distributions

Normal Distribution, *Norm*(μ,σ), with parameters $\mu \in \mathbb{R},\ \sigma > 0$:

pdf $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$, cdf $F(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-\dfrac{(t-\mu)^2}{2\sigma^2}} dt = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$,
$E(X) = \mu,\ V(X) = \sigma^2$.

Standard (Reduced) Normal Distribution, *Norm*$(0,1)$:

pdf $f(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\dfrac{x^2}{2}}, x \in \mathbb{R}$, cdf $F(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-\dfrac{t^2}{2}} dt = \Phi(x)$, $E(X) = 0,\ V(X) = 1$.

Uniform Distribution, *Unif*(a, b), with parameters $a, b \in \mathbb{R},\ a < b$:

pdf $f(x) = \dfrac{1}{b-a},\ x \in [a,b]$, cdf $F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x < b \\ 1, & x \ge b \end{cases}$, $E(X) = \dfrac{a+b}{2},\ V(X) = \dfrac{(b-a)^2}{12}$.

Standard Uniform Distribution, *Unif*(0, 1):

pdf $f(x) = 1,\ x \in [0,1]$, cdf $F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$, $E(X) = \dfrac{1}{2},\ V(X) = \dfrac{1}{12}$.

Exponential Distribution, *Exp*(λ) = *Gam*$(1, 1/\lambda)$, with parameter $\lambda > 0$:

pdf $f(x) = \lambda e^{-\lambda x}$, $x > 0$ (**Caution!** in Matlab, $f(x) = \dfrac{1}{\lambda} e^{-\dfrac{1}{\lambda}x}$, $x > 0$), cdf $F(x) = 1 - e^{-\lambda x}$,
$E(X) = \dfrac{1}{\lambda},\ V(X) = \dfrac{1}{\lambda^2}$.
$X \in$ Exp(λ) models time: waiting time, interarrival time, failure time, time between rare events, etc. The parameter λ represents the frequency of rare events, measured in time^{-1}.

Gamma Distribution, *Gam*(α,λ), with parameters $\alpha, \lambda > 0$:

pdf $f(x) = \dfrac{1}{\lambda^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\dfrac{x}{\lambda}}$, $x > 0$, $E(X) = \alpha\lambda,\ V(X) = \alpha\lambda^2$.
$X \in$ Gam(α,λ) models the total time of a multistage scheme, where each (independent) step takes Exp$(1/\lambda)$ amount of time.

<u>Properties</u>

1. $Bino(n, p)$ is the sum of n independent $Bern(p)$ variables; $Bern(p) = Bino(1, p)$.

2. $Nbin(n, p)$ is the sum of n independent $Geo(p)$ variables; $Geo(p) = Nbin(1, p)$.

3. For $\alpha \in \mathbb{N}$, $Gam(\alpha, \lambda)$ is the sum of α independent $Exp(1/\lambda)$ variables; $Exp(\lambda) = Gam(1, 1/\lambda)$.

4. In a Poisson process, the time between rare events is Exponentially distributed and the time of the α-th event is *Gamma*- distributed.

5. If T represents the time of the occurrence of the α-th rare event and X is the number of rare events that occur in time $t > 0$, then $T \in Gam(\alpha, 1/\lambda)$ and $X \in Poiss(\lambda t)$.

6. Memoryless Property: Exponential $Exp(\lambda), \lambda > 0$ and "Almost" Geometric $AGeo(p), p \in (0, 1)$ variables "lose memory"; in predicting the future, the past gets "forgotten", only the present matters,

$$\begin{aligned} X \in Exp(\lambda), &\quad P(X > x + y \mid X > y) = P(X > x), \quad \forall x, y \geq 0, \\ X \in AGeo(p), &\quad P(X > x + y \mid X > y) = P(X > x), \quad \forall x, y \in \mathbb{N}. \end{aligned}$$

7. In a sense, Exponential distribution is a continuous version of a AGeometric distribution: An Exponential variable describes the time (measured continuously) until the next "rare event" occurs, a AGeometric variable is the time ("measured" discreetly, as the number of Bernoulli trials) until the next success. Also, they both have the memoryless property, which *no other* (discrete or continuous) distribution has.

8. Gamma-Poisson Formula For $T \in Gam(\alpha, \lambda)$ and $X \in Poiss\left(\frac{1}{\lambda}t\right)$, $\alpha \in \mathbb{N}, \lambda, t > 0$, the following formulas hold:

$$\begin{aligned} P(T > t) &= P(X < \alpha), \\ P(T \leq t) &= P(X \geq \alpha). \end{aligned}$$