Probability **Note**

我不知道

1 Variance

Property 1.1. For two independent RVs X, Y, $\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$.

Proof.

$$\begin{aligned}
& \operatorname{Var}(X+Y) \\
= & E[X+Y]^2 - E[(X+Y)^2] \\
= & E[X]^2 + 2E[X]E[Y] + E[Y]^2 - E[X^2] - 2E[XY] - E[Y]^2 \\
= & \operatorname{Var}(X) + \operatorname{Var}(Y)
\end{aligned}$$

■

Property 1.2. For RV X, $Y = cX$ has variance $\operatorname{Var}(Y) = c^2 \operatorname{Var}(x)$.

Proof.

$$\begin{aligned}
& \operatorname{Var}(cX) \\
= & E[cX]^2 - E[(cX)^2] \\
= & c^2 \operatorname{Var}(X)
\end{aligned}$$

■

Lemma 1.1. Let $x_1, x_2, \ldots, x_n$ be n samples from a probability distribution. Let $\bar{x}$ denote the mean of the samples and μ be the *real mean*. If the *real mean* is known, the sample variance is

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Otherwise,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Proof. Call $f(x)$ an *unbiased estimator* of θ if

$$\mathbb{E}[f(u(x_1, x_2, \ldots, x_n))] = \theta$$

where u means uniformly chosen from all samples x_i.

Let us show that the given formula is indeed an unbiased estimator of $\mathrm{Var}(x)$.

$$\begin{aligned}
&E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\\
=&\ \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\\
=&\ \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2\right]\\
=&\ \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i-\mu)^2+2\sum_{i=1}^{n}(x_i-\mu)(\mu-\bar{x})+\sum_{i=1}^{n}(\mu-\bar{x})^2\right]\\
=&\ \frac{n}{n-1}\left(\mathrm{Var}(x)-2\,\mathrm{Var}(\bar{x})+\mathrm{Var}(\bar{x})\right)\\
=&\ \frac{n}{n-1}\left(\mathrm{Var}(x)-\mathrm{Var}(\bar{x})\right)\\
=&\ \frac{n}{n-1}\left(\mathrm{Var}(x)-\mathrm{Var}\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)\right)\\
=&\ \frac{n}{n-1}\left(\mathrm{Var}(x)-\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n}x_i\right)\right)\\
=&\ \frac{n}{n-1}\left(\mathrm{Var}(x)-\frac{n}{n^2}\mathrm{Var}(x)\right)\\
=&\ \frac{n}{n-1}\cdot\frac{n-1}{n}\mathrm{Var}(x)\\
=&\ \mathrm{Var}(x)
\end{aligned}$$

■

2 Discrete Distributions

2.1 Binomial Distribution

Definition 2.1. A *Bernoulli Trial* is a RE that only has two outcomes, success and fail, each with probability of p and $1-p$.

Definition 2.2. A sequence of RVs is *independent and identically distributed* if each of them has the same probability distribution and they are mutally independent.

Definition 2.3. A *Binomial Distribution*, denoted by $B(n,p)$, is the distribution the number of success of n i.i.d. Bernoulli Trial with success probability p. The PMF of $X \sim B(n,p)$ is

$$f(X)=\binom{n}{X}p^X(1-p)^X$$

Property 2.1. Let $X \sim B(n,p)$.

- $\mu = np$.

- $\sigma^2 = np(1-p)$.

Proof. We will prove the variance. The MGF of X is

$$M(t) = \sum_{X=0}^{n} e^{tX}\binom{n}{X}p^X(1-p)^X = \left(pe^t + (1-p)\right)^n$$

Therefore

$$M'(t) = n\left(pe^t + (1-p)\right)^{n-1} pe^t \Rightarrow \mu = E[X] = M'(0) = np$$

$$M''(t) = n(n-1)\left(pe^t + (1-p)\right)^{n-2} p^2e^{2t} + n\left(pe^t + (1-p)\right)^{n-1} pe^t \Rightarrow E[X^2] = M''(0) = n(n-1)p^2 + np$$

$$\sigma^2 = E[X^2] - E[X]^2 = np - np^2 = np(1-p)$$

■

Definition 2.4. A *Negative Binomial Distribution*, denoted by $NB(r,p)$, is the distribution of the number of i.i.d. Bernoulli Trial with success probability p needs to performed until r success. The PMF of $X \sim NB(r,p)$ is

$$f(X) = \binom{r+X-1}{r-1}p^r(1-p)^X$$

Definition 2.5. A *Geometric Distribution* with probability p, denoted by $G(p)$, is defined as $NB(1,p)$. Its PMF is

$$f(X) = p(1-p)^X$$

Definition 2.6 (Poisson Distribution)**.** Observe the number of arrivals during a given period of time. Under the assumption of

1. Each arrival is independent.
2. The probability of exactly one arrival in a sufficient short period of time Δt is $\lambda\Delta t$.
3. The probability of two or more arrival in a sufficient short period of time is 0.

Then the distribution is a *Poisson Distribution*, denoted by $\text{Pois}(\lambda)$.

Property 2.2. Let $X \sim \text{Pois}(\lambda)$

- PMF $f(X) = \dfrac{\lambda^X e^{-\lambda}}{X!}$
- $\mu = \lambda$.
- $\sigma^2 = \lambda$.

Proof. We will prove the all of them. Assume the time period is splited into n segments, then

$$X \sim B(n, \frac{\lambda}{n})$$

When $n \to \infty$,

$$\begin{aligned} P(X=x) &= \lim_{n\to\infty}\binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \lim_{n\to\infty}\frac{\prod_{n-x+1}^{i=n} i}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \\ &= 1\cdot\frac{\lambda^x}{x!}\cdot e^{-\lambda}\cdot 1 \end{aligned}$$

The MGF of the distribution is

$$M(t) = \sum_{x=0}^{\infty} \frac{\lambda^X e^{-\lambda}}{X!} e^{tX} = e^{-\lambda} e^{\lambda e^t}$$

Therefore

$$M'(t) = \lambda e^t e^{\lambda e^t - \lambda} \Rightarrow \mu = E[x] = M'(0) = \lambda$$

$$M''(t) = \lambda e^t e^{\lambda e^t - \lambda} + \lambda^2 e^{2t} e^{\lambda e^t - \lambda} \Rightarrow E[x^2] = \lambda + \lambda^2 \Rightarrow \sigma^2 = \lambda$$

■

3 Continuous Distributions

Definition 3.1. For a RV X, the *cumulative distribution function* is defined

$$F_X(t) = P(X \leq t)$$

which the following must hold:

- $\lim_{t\to\infty} F_X(t) = 1$.
- $\lim_{t\to-\infty} F_X(t) = 0$.
- $a < b \Rightarrow F_X(a) \leq F_X(b)$.

Definition 3.2. For a continuous RV X, the *probability density function* is defined

$$f_X(t) = \frac{d}{dt} F_X(t)$$

or equivlantly,

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

Property 3.1. A valid PDF f_X with space S must have

- $f(x) \geq 0 \quad (x \in S)$
- $\int_S f(x)dx = 1$
- $P(a < X < b) = \int_a^b f(x)dx \quad ((a, b) \subseteq S)$.

Remark. Notice the following consequences:

- It is possible that $f(x) > 1$.
- $P(X = x) = 0$.
- It is possible f_X is non-continuous
- $P(X \in [a, b]) = P(X \in [a, b)) = P(X \in (a, b]) = P(X \in (a, b))$

Definition 3.3. For a continuous RV X with PDF f_X,

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

- Variance $\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f_X(x)dx$
- MGF $M(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$, if there exists a open interval $(-\delta, \delta)$ which the integral converges.

Theorem 3.1 (Moment Generating Function Uniqueness Theorem)**.** If two MGF are the same, the PDF will also be the same.

Proof will not be present because it requires graduate courses.

Example. $X \sim U(a,b)$ means X is uniformly random selected from the interval $[a,b]$, which has the following property.

- Mean $\mu = \dfrac{a+b}{2}$
- Variance $\sigma^2 = \dfrac{(b-a)^2}{12}$
- MGF $M(t) = \begin{cases} \dfrac{e^{bt}-e^{at}}{t(b-a)} & (t \neq 0) \\ 1 & (t = 0) \end{cases}$

From discrete possion distribution, the waiting time between two successive arrival is also an RV. We denote this distribution as *Exponential Distribution*.

Definition 3.4. $X \sim \mathrm{Exp}(\lambda)$ denotes the distribution of two successive arrival under the setting of possion distribution. The CDF can be written as

$$P(X \leq \omega) = 1 - \frac{(\omega\lambda)^0 e^{-\omega\lambda}}{0!} = 1 - e^{-\omega\lambda}$$

Therefore

$$f_X(\omega) = \lambda e^{-\lambda\omega} \quad (\omega \geq 0)$$

Property 3.2. Let $X \sim \mathrm{Exp}(\lambda)$, then the MGF is

$$M(t) = \frac{1}{1-\lambda^{-1}t}$$

Consequently, $\mu = \lambda^{-1}, \sigma^2 = \lambda^{-2}$.

Proof. When $t = \lambda$ the integral diverges, otherwise

$$M(t) = \int_0^{\infty} \lambda e^{-\lambda x} e^{tx} dx = \frac{\lambda}{t-\lambda} e^{(t-\lambda)x}\Big|_0^{\infty}$$

When $t > \lambda$ the integral also diverges. Therefore when $t < \lambda$,

$$M(t) = -\frac{\lambda}{t-\lambda} = \frac{1}{1-\lambda^{-1}t}$$

∎

Property 3.3. Exponential distribution is *memoryless*. That is, for $a < b$, $P(X > b \mid X > a) = P(X > (b-a))$.

It is not hard to think of, because intuitively, in Possion distribution, every arrival is independent. In mathematical form,

$$P(X > b \mid X > a) = \frac{P(X > b \cap X > a)}{P(X > a)} = \frac{P(X > b)}{P(X > a)} = \frac{e^{-b\lambda}}{e^{-a\lambda}} = e^{-(b-a)\lambda}$$

It is possible to generalize the exponential distribution,

Definition 3.5. $X \sim \Gamma(\alpha, \lambda)$ denotes the distribution of waiting time before the α-th under the setting of possion distribution. The CDF can be written as

$$P(X \leq \omega) = 1 - \sum_{k=0}^{\alpha-1} \frac{(\omega\lambda)^k e^{-\omega\lambda}}{k!}$$

Therefore

$$f_X(\omega) = \lambda e^{-\omega\lambda} - \sum_{k=1}^{\alpha-1} \left(\frac{-\lambda(\omega\lambda)^k e^{-\omega\lambda}}{k!} + \frac{k\lambda(\omega\lambda)^{k-1} e^{-\omega\lambda}}{k!} \right) = \lambda e^{-\omega\lambda} - \lambda e^{-\omega\lambda} \left(1 - \frac{(\lambda\omega)^{\alpha-1}}{(\alpha-1)!} \right) = \frac{\omega^{\alpha-1}\lambda^\alpha e^{-\omega\lambda}}{(\alpha-1)!}$$

Property 3.4. Let $X \sim \Gamma(\alpha, \lambda)$, then the MGF is

$$M(t) = (1 - \lambda^{-1}t)^{-\alpha} \quad (t < \lambda)$$

Consequently, $\mu = \alpha\lambda^{-1}, \sigma^2 = \alpha\lambda^{-2}$.

Proof. Let us prove by mathematical induction. Observe when $\alpha = 1$, the result is the same as exponential distribution, which we have already shown. Assume the property is correct when $\alpha = k - 1$. When $\alpha = k$:

$$\begin{aligned}
\int_0^\infty \frac{x^{\alpha-1}\lambda^\alpha e^{(t-\lambda)x}}{(\alpha-1)!} dx &= \int_0^\infty \frac{x^{\alpha-1}\lambda^\alpha}{(\alpha-1)!} \cdot e^{(t-\lambda)x} dx \\
&= \frac{x^{\alpha-1}\lambda^\alpha}{(\alpha-1)!} \cdot \frac{e^{(t-\lambda)x}}{t-\lambda}\Bigg|_0^\infty - \int_0^\infty \frac{x^{\alpha-2}\lambda^\alpha}{(\alpha-2)!} \cdot \frac{e^{(t-\lambda)x}}{t-\lambda} dx \\
&= -\int_0^\infty \frac{x^{\alpha-2}\lambda^\alpha}{(\alpha-2)!} \cdot \frac{e^{(t-\lambda)x}}{t-\lambda} dx \\
&= \frac{\lambda}{\lambda - t} \int_0^\infty \frac{x^{\alpha-2}\lambda^{\alpha-1} e^{(t-\lambda)x}}{(\alpha-2)!} dx \\
&= (1 - \lambda^{-1}t)^{-\alpha}
\end{aligned}$$

By induction, the property holds. ■

Definition 3.6. The *Chi-square Distribution* with r degrees of freedom, denoted as $\chi^2(r)$, is equivlant to $\Gamma(\frac{r}{2}, \frac{1}{2})$.

Definition 3.7. The *Beta Distribution* with shape parameters α, β, denoted as $\mathrm{Beta}(\alpha, \beta)$, is defined on interval $[0, 1]$, and has PDF

$$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

where $\mathrm{B}(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is called beta function, which is a normalization constant.

Beta distribution is the beta distribution is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial, and geometric distributions. Recall the Baye's Law:

$$\underbrace{\boxed{P(A \mid B)}}^{\text{Posterior}} = \frac{\overbrace{\boxed{P(B \mid A)}}^{\text{Likelihood}} \times \overbrace{\boxed{P(A)}}^{\text{Prior}}}{\underbrace{\boxed{P(B)}}_{\text{Evidence}}}$$

Suppose we suspect the target distribution is binomial distribution, but with probability p unknown, then we can choose beta distribution as the *prior* probability, with parameters α, β reflecting existing belief or information. For example if we know the p is selected uniformly at random, we might choose $\alpha = \beta = 1$.
We know that if we sampled s success and f faliure, then

$$P(s, f \mid p = x) = \binom{s+f}{s} x^s (1-x)^f$$

$$P(p = x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$$

To compute the probability of p being x given the result of s, f, we use Baye's Law:

$$\begin{aligned}
P(p = x \mid s, f) &= \frac{P(s, f \mid p = x) P(p = x)}{P(s, f)} \\
&= \frac{\binom{s+f}{s} x^s (1-x)^f \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}}{\int P(s, f \mid p = y) P(p = y) dy} \\
&= \frac{\binom{s+f}{s} x^s (1-x)^f \cdot \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}}{\int_0^1 \binom{s+f}{s} y^s (1-y)^f \cdot \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} dy} \\
&= \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{\mathrm{B}(s+\alpha, f+\beta)} \\
&\sim \mathrm{Beta}(s+\alpha, f+\beta)
\end{aligned}$$

We call this *conjugate prior probability*, because the beta distribution appraring as prior results a beta distribution as posterior probability distribution.

4 Transformation of Random Variable

Basic transformation:

Property 4.1. For a random variable X of PDF f_X. The CDF of $Y = g(X)$ is

$$F_Y(y) = P(Y \le y) = P(g(X) \le Y) = \int_{g(x) \le y} f_X(x) dx$$

Property 4.2. Assume $Y = u(X)$ is **monotonic increasing** function. If the support (c_1, c_2) maps to the $(d_1, d_2) = (u(c_1), u(c_2))$. Then the distribution of Y is

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(u^{-1}(y)) = f_X(u^{-1}(y))(u^{-1})'(y)$$

If u is **monotonic decreasing** function, then the result is

$$f_Y(y) = -\frac{d}{dy}F_Y(y) = -\frac{d}{dy}F_X(u^{-1}(y)) = -f_X(u^{-1}(y))(u^{-1})'(y)$$

Applying the whole domain might not be possible, but partition the domain into several monotonic intervals is a good idea.

Example. $X \sim \Gamma(\alpha, \lambda)$, find Y where $Y = e^X$.
$X = \ln Y$, therefore

$$f_Y = -f_X(\ln Y) \cdot \frac{1}{Y} = \frac{(\ln Y)^{\alpha-1} Y^{-\lambda} \lambda^\alpha}{\Gamma(\alpha) Y}$$

Example. Assume a RV X with $Y = aX + b$, find PDF of Y.

$$f_Y = f_X\left(\frac{Y-b}{a}\right)\left|\frac{1}{a}\right|$$

Example. Assume a RV X with $Y = X^n$ (where $n \in \mathbb{N}$), find PDF of Y.
When n is odd, the transformation property is applicable:

$$f_Y = f_X\left(Y^{\frac{1}{n}}\right)\frac{Y^{\frac{1}{n}-1}}{n}$$

When n is even, consider the both side:

$$f_Y = \left(f_X\left(Y^{\frac{1}{n}}\right) + f_X\left(-Y^{\frac{1}{n}}\right)\right)\frac{Y^{\frac{1}{n}-1}}{n} \quad (Y \geq 0)$$

5 Multivariable Distribution

Definition 5.1. A *joint PMF* of k discrete RVs $X_1, X_2, \ldots, X_k$ is a function that $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = f(x_1, x_2, \ldots, x_k)$. Such function satisfies

- $0 \leq f(x_1, x_2, \ldots, x_k) \leq 1$
- $\sum_{(x_1, x_2, \ldots, x_k) \in S} f(x_1, x_2, \ldots, x_k) = 1$

The *marginal PMF* of X_i is

$$f_{X_i}(x) = P(X_i = x) = \sum f(x_1, x_2, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_k)$$

Definition 5.2. A *joint PDF* of k continuous RVs $X_1, X_2, \ldots, X_k$ is a function that $P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = f(x_1, x_2, \ldots, x_k)$. Such function satisfies

- $0 \leq f(x_1, x_2, \ldots, x_k)$
- $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k$

The *marginal PDF* of X_i is

$$\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_k) dx_i$$