Lecture 2 Probability Distributions in Statistics

Tiejun Li

1 Discrete Examples

We will concentrate on the elementary and intuitive aspects of probability here. In the discrete case, the function \( P(X) \) is called the probability mass function (pmf).

- **Bernoulli distribution:**
  \[
P(X) = \begin{cases} p, & X = 1 \\ q, & X = 0 \end{cases}
\]
  where \( p > 0, q > 0, p + q = 1 \). The mean and variance are
  \[
  \mathbb{E}X = p, \quad \text{Var}(X) = pq.
  \]
  If \( p = q = \frac{1}{2} \), it is the well-known fair-coin tossing game.

- **Binomial distribution** \( B(n, p) \):
  \( n \) independent experiments of Bernoulli distribution \( X_k, X := X_1 + \ldots + X_n \), then
  \[
P(X = k) = \binom{n}{k} p^k q^{n-k}.
\]
  The mean and variance are
  \[
  \mathbb{E}X = np, \quad \text{Var}(X) = npq.
  \]

- **Negative binomial distribution** \( B^*(r, p) \):
  Suppose we repeat Bernoulli trials, we are interested in the number of the trial \( X \) on which the \( r \)-th 1 occurs. We may understand \( X \) as the binomial waiting time when the \( r \)-th 1 occurs. It is obvious that \( X \geq r \). Simple derivation gives
  \[
P(X = k) = \binom{r-1}{k-r-1} p^{r-1} q^{k-r} \cdot p = \binom{r-1}{k-r} p^{r-k} q^k, \quad k = r, r + 1, \ldots
\]
  Now we define \( k = m + r \) (\( m \in \mathbb{N} \)), then we have
  \[
  \binom{r-1}{k-r} q^{k-r} = \binom{m}{r} (-q)^m.
  \]
  Here take the convention that
  \[
  \binom{m}{r} = \frac{\alpha(\alpha - 1) \cdots (\alpha - m + 1)}{m!}, \quad \text{for } m \in \mathbb{N}, \quad \alpha \in \mathbb{C}.
  \]
From the Newton’s binomial theorem
\[(1 + x)^\alpha = \sum_{m=0}^{\infty} C_m^\alpha x^m, \quad |x| < 1, \ \alpha \in \mathbb{C}\]
we have
\[\sum_{m=0}^{\infty} C_r^{-p} (-q)^m = p^r (1 - q)^{-r} = 1.\]

Negative binomial distribution is also called Pascal distribution. When \(r = 1\), \(B^*(k; 1, p) = pq^{k-1}\) is called geometric distribution.

The mean and variance are
\[\mathbb{E}X = \frac{r}{p}, \quad \text{Var}(X) = \frac{r(1 - p)}{p^2}. \quad (1)\]

- **Multinomial distribution** \(M(p_1, \ldots, p_r)\):

  Multinomial distribution is a simple generalization of binomial distribution, in which each trial results in exactly one of some fixed number \(r\) possible outcomes with probability \(p_1, p_2, \ldots, p_r\), where

  \[\sum_{i=1}^{r} p_i = 1, \quad 0 \leq p_i \leq 1, \quad i = 1, \ldots, r,\]

  and we have \(n\) independent trials. Let the random variables \(X_i\) indicate the number of times the \(i\)-th outcome was observed over the \(n\) trials. \(X = (X_1, \ldots, X_r)\) follows a multinomial distribution with parameters \(n\) and \(p\), where \(p = (p_1, \ldots, p_r)\).

  The pmf of the multinomial distribution is:
  \[P(X_1 = x_1, \ldots, X_r = x_r) = \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \quad n = x_1 + \cdots + x_r.\]

  The mean, variance and covariance are

  \[\mathbb{E}(X_i) = np_i, \quad \text{Var}(X_i) = np_i(1 - p_i), \quad \text{Cov}(X_i, X_j) = -np_ip_j \ (i \neq j).\]

- **Hypergeometric distribution:**

  To explain the hypergeometric distribution, we take an example that there is a shipment of \(N\) objects in which \(M\) are defective. The hypergeometric distribution describes the probability that exactly \(X = k\) objects are defective in a sample of \(n\) distinctive objects drawn from the shipment. We have

  \[P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, \quad k \leq M, \quad n - k \leq N - M.\]

  The mean and variance:

  \[\mathbb{E}X = \frac{nM}{N}, \quad \text{Var}(X) = \frac{nM(N - M)(N - n)}{N^2(N - 1)}.\]

- **Poisson distribution:**
The number $X$ of radiated particles in a fixed time $\tau$ (or the number of received calls in a fixed time $\tau$, etc.) obeys

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

where $\lambda$ is the average number of radiated particles each time. The mean and variance are

$$\mathbb{E}X = \lambda, \text{Var}(X) = \lambda.$$  

Poisson distribution may be viewed as the limit of binomial distribution (the law of rare events)

$$C_n^k p^k q^{n-k} \rightarrow \frac{\lambda^k}{k!} e^{-\lambda} \quad (n \rightarrow \infty, np = \lambda).$$

Poisson distribution can also describe the spatial distribution of randomly scattered points. For example, Let $A$ be a set in $\mathbb{R}^2$. $X_A(\omega)$ be the number of points in $A$. If the points are uniformly distributed on the plane, then

$$\lambda = \text{area of } A \times \text{number of points/area}.$$

$X_A$ has Poisson distribution

$$P(X_A = n) = \frac{\lambda^n}{n!} e^{-\lambda}.$$

- Geometric probability.
  - Probability = Ratio of areas
  - Special case of continuous examples — uniform distribution.

**Example 1.** Maxwell-Boltzmann, Bose-Einstein, Fermi-Dirac statistics.

Suppose there are $n$ particles and $N$ bins, where $N > n$.

1. Given $n$ bins, what is the probability that each bin has one particle? (Boson)
2. What is the probability that there exist $n$ bins such that each bin has exactly one particle? (Fermion, Pauli exclusion principle)

The answer to these questions depends on whether the particles are distinguishable or not (classical or quantum). For answers see [1].

## 2 Continuous Examples

In continuous case, the function $p(x)$ is called the probability density function (pdf).

- Uniform distribution $U[0, 1]$:

$$p(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

The mean and variance are

$$\mathbb{E}X = \frac{1}{2}, \text{Var}(X) = \frac{1}{12}.$$
• Exponential distribution: \((\lambda > 0)\)

\[
p(x) = \begin{cases} 
0 & \text{if } x < 0 \\
\lambda e^{-\lambda x} & \text{if } x \geq 0 
\end{cases}
\]

The mean and variance are

\[
E(X) = \frac{1}{\lambda}, \ Var(X) = \frac{1}{\lambda^2}.
\]

Special cases include the Boltzmann distribution

\[
f(E) = \begin{cases} 
\frac{1}{Z} e^{-\beta E} & E \geq 0 \\
0 & E < 0
\end{cases}
\]

\[
\beta = \frac{1}{k_B T}, k_B - \text{Boltzmann constant, } T - \text{temperature.}
\]

Waiting time for continuous time Markov process also has exponential distribution, where \(\lambda\) is the rate of the process.

The characteristic function of exponential distribution is \(g(t) = E(\exp(i t X)) = (1 - it/\lambda)^{-1}\).

• Normal distribution (Gaussian distribution) \((N(0,1))\):

\[
p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}
\]

or more generally \(N(\mu, \sigma)\)

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

where \(\mu\) is the mean (expectation), \(\sigma^2\) is the variance.

High dimensional case \((N(\mu, \Sigma))\)

\[
p(x) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
\]

where \(\mu\) is the mean, \(\Sigma\) is a symmetric positive definite matrix, which is the covariance matrix of \(X\), \(\det \Sigma\) is the determinant of \(\Sigma\). More general high dimensional normal distribution is defined with characteristic functions \(g(t) = \exp\left(i \mu \cdot t - \frac{1}{2} t^T \Sigma t\right)\).

**Remark 1.** In 1D case, the normal distribution \(N(np, npq)\) may be viewed as the limit of the Binomial distribution \(B(n, p)\) when \(n\) is large. This is the famous De Moivre-Laplace limit theorem. It is a special case of the central limit theorem (CLT). Notice that

\[
\frac{B(n, p) - np}{\sqrt{npq}} \xrightarrow{d} N(0,1) \text{ as } n \to \infty.
\]

**Remark 2.** In 1D case, the normal distribution \(N(\lambda, \lambda)\) may be viewed as the limit of the Poisson distribution \(Poisson(\lambda)\) when \(\lambda\) is large. Notice the simple fact that the sum of two independent \(Poisson(\lambda)\) and \(Poisson(\mu)\) is \(Poisson(\lambda + \mu)\) (why?), we can decompose \(Poisson(\lambda)\) into the sum of \(n\) i.i.d. \(Poisson(\lambda/n)\), we have

\[
\frac{\text{Poisson}(\lambda) - \lambda}{\sqrt{\lambda}} \xrightarrow{d} N(0,1) \text{ when } \lambda \text{ is large.}
\]

Here the condition that \(\lambda\) is large ensures the discrete points of PMF form a continuous profile.
• Beta distribution: Beta(\(\alpha, \beta\))

The beta distribution is a family of continuous probability distributions defined on the interval \([0, 1]\) parameterized by two positive shape parameters, typically denoted by \(\alpha\) and \(\beta\).

The pdf of Beta distribution is \((x \in [0, 1])\)
\[
p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}
\]
where \(B(\alpha, \beta)\) is the Beta function
\[
B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} \, dt.
\]
We have
\[
\mathbb{E}X = \frac{\alpha}{\alpha + \beta}, \quad \text{Var}X = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.
\]
Actually it is a special case of the Dirichlet distribution with only two components. If \(\alpha = \beta = 1\), it is \(U[0, 1]\); if \(\alpha, \beta < 1\), it is a \(U\)-shape curve; if \(\alpha, \beta > 1\), it is unimodal.

• Gamma distribution: \(\Gamma(k, \theta)\) \((k, \theta > 0)\)

The gamma distribution is a two-parameter family of continuous probability distributions. It has a scale parameter \(\theta\) and a shape parameter \(k\). If \(k\) is an integer then the distribution represents the sum of \(k\) exponentially distributed random variables, each of which has a mean of \(\theta\). But in general, \(k\) may be positive reals.

We have
\[
p(x) = x^{k-1} e^{-x/\theta} \frac{1}{\theta^k \Gamma(k)}, \quad x > 0,
\]
where \(\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} \, dt\) is the Gamma function \((\Gamma(n+1) = n!)\). We have the mean and variance
\[
\mathbb{E}X = k\theta, \quad \text{Var}X = k\theta^2.
\]
The Gamma distribution has the property that if \(X \sim \Gamma(k, 1)\), then \(\theta X \sim \Gamma(k, \theta)\).

The characteristic function of Gamma distribution is \(g(t) = (1 - it\theta)^{-k}\). Thus if \(X_i\) has a \(\Gamma(k_i, \theta)\) distribution for \(i = 1, 2, \ldots, N\), then
\[
\sum_{i=1}^N X_i \sim \Gamma \left( \sum_{i=1}^N k_i, \theta \right)
\]
provided all \(X_i\) are independent. Thus Gamma distribution exhibits the infinite divisibility.

When \(\theta = 2\), Gamma distribution becomes \(\chi^2(2k)\) distribution.

When \(k = 1\), Gamma distribution becomes exponential distribution with parameter \(1/\theta\).

• Dirichlet distribution: \(\text{Dir}(\alpha)\) \((K \geq 2)\)

\[
p((x_1, x_2, \ldots, x_{K-1})) = \frac{1}{B(\alpha)} \prod_{i=1}^K x_i^{\alpha_i-1}
\]
where \( x := (x_1, \ldots, x_K) \geq 0 \) and \( \sum_{i=1}^{K} x_i = 1 \). Here
\[
B(\alpha) = \int_{x_i \geq 0, \sum_{i=1}^{K} x_i < 1} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \, dx_1 \cdots dx_{K-1} = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)}
\]
is called the multinomial Beta function. Dirichlet distribution is a multinomial generalization of Beta distribution. We have \( X_i \sim Beta(\alpha_i, \alpha_0 - \alpha_i) \).

Define \( X_K = 1 - (X_1 + \cdots + X_{K-1}) \), we have
\[
\mathbb{E} X_i = \frac{\alpha_i}{\alpha_0}, \quad \text{Var} X_i = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad i = 1, 2, \ldots, K
\]
where \( \alpha_0 = \sum_{i=1}^{K} \alpha_i \). We have the aggregation property
\[
(X_1, \ldots, X_i + X_j, \ldots, X_K) \sim Dir(\alpha_1, \ldots, \alpha_i + \alpha_j, \ldots, \alpha_K).
\]

One intuitive explanation of Dirichlet distribution may be as follows. If one wants to cut strings (each of initial length 1) into \( K \) pieces with average length \( \alpha_i/\alpha_0 \) for different pieces, and we allow some variation in the relative sizes of the pieces. One choice is \( \text{Dir}(\alpha) \).

- Inverse Gaussian distribution: \( IG(\mu, \lambda) \) (\( \lambda, \mu > 0 \))

It is also called the Wald distribution in probability. We have
\[
p(x) = \left[ \frac{\lambda}{2\pi \mu^3} \right]^{1/2} \exp \left\{ -\frac{\lambda(x - \mu)^2}{2\mu^2} \right\}, \quad x > 0
\]
and \( \mathbb{E} X = \mu, \text{Var} X = \frac{\mu^3}{\lambda} \). As \( \lambda \) tends to infinity, the inverse Gaussian distribution becomes more like a Gaussian distribution.

If \( X_i \) has distribution \( IG(\mu_0 w_i, \lambda_0 w_i^2) \) for \( i = 1, 2, \ldots, n \) and all \( X_i \) are independent, then
\[
S = \sum_{i=1}^{n} X_i \sim IG(\mu_0 \bar{w}, \lambda_0 \bar{w}^2),
\]
where \( \bar{w} = \sum_{i=1}^{n} w_i \).

For process \( X_t = \nu t + \sigma W_t \) (\( \nu > 0 \)), where \( W_t \) is a standard Brownian motion. The first passage time for a fixed level \( \alpha > 0 \) by \( X_t \) is defined as
\[
T_\alpha = \inf \{0 < t < \infty \mid X_t = \alpha\}.
\]
Then the IG parameters become \( \mu = \frac{\alpha}{\nu} \lambda = \frac{\alpha^2}{\sigma^2} \), and
\[
T_\alpha \sim IG\left(\frac{\alpha}{\nu}, \frac{\alpha^2}{\sigma^2}\right).
\]

- Chi-square distribution \( \chi^2(k) \):

If \( X_i \sim N(0, 1), \text{i.i.d.} \), then we say \( Y = \sum_{i=1}^{n} X_i^2 \) satisfies the \( \chi^2 \)-distribution with \( n \) degrees of freedom. In general if the pdf \( (k > 0) \)
\[
p(x) = \begin{cases} \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2 - 1} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{for } x \leq 0, \end{cases}
\]
we say $X$ has the chi-square distribution with $k$ degrees of freedom.

The characteristic function of chi-square distribution is

$$g(t) = (1 - 2it)^{-k/2}.$$ 

Thus we have if $X_1, \ldots, X_n$ are independent R.V. having chi-square distributions with $k_1, \ldots, k_n$ degrees of freedom, then $Y = \sum_{i=1}^n X_i$ has the chi-square distribution with $k_1 + \cdots + k_n$ degrees of freedom.

The mean and variance of $\chi^2(k)$ are

$$E(X) = k, \ Var(X) = 2k.$$ 

Define $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, we also have

**Lemma 1.** If $\bar{X}$ and $S^2$ are the mean and the variance of a random sample of size $n$ from a normal population with the mean $\mu$ and variance $\sigma^2$, then $\bar{X}$ and $S^2$ are independent and $\frac{(n-1)S^2}{\sigma^2}$ has $\chi^2(n-1)$ distribution.

- **Student’s t distribution $t(k)$:**

  The “student’s” distribution was actually published in 1908 by W. S. Gosset. Gosset, however, was employed at a brewery that forbade the publication of research by its staff members. To circumvent this restriction, Gosset used the name “Student”, and consequently the distribution was named “Student t-distribution”.

  Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, i.i.d.. We have

  $$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$ 

  Gosset studied a related pivotal quantity,

  $$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}},$$ 

  which differs from $Z$ in that the exact standard deviation $\sigma$ is replaced by the random variable $S_n$.

  Gosset’s work showed that $T$ has the probability density function

  $$p(t) = \frac{\Gamma\left(k+\frac{1}{2}\right)}{\sqrt{k\pi} \Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}, \quad t \in \mathbb{R}.$$ 

  with $k$ equals to $n-1$.

  The distribution of $T$ is now called the $t$-distribution. The parameter $k$ is called the number of degrees of freedom. The distribution depends on $k$, but not $\mu$ or $\sigma$; the lack of dependence on $\mu$ and $\sigma$ is what makes the $t$-distribution important in both theory and practice.

  When $k > 1$, $E(T) = 0$, otherwise it is not defined; when $k > 2$, $Var(T) = \frac{k}{k-2}$, otherwise it is not defined! When $k = 1$, we have

  $$p(t) = \frac{1}{\pi(1+t^2)}.$$ 

  It is called Cauchy’s distribution, which has neither mean and variance.
• \textit{F} distribution $F(k_1, k_2)$:

A random variate of the $F$-distribution arises as the ratio of two chi-squared variates:

$$\frac{U_1/k_1}{U_2/k_2},$$

where $U_1 \sim \chi^2(k_1)$ and $U_2 \sim \chi^2(k_2)$, and $U_1$ and $U_2$ are independent.

The probability density function of an $F(k_1, k_2)$ distributed random variable is given by

$$p(x) = \frac{\Gamma\left(\frac{k_1 + k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \cdot x^{\frac{k_1}{2} - 1} \left(1 + \frac{k_1}{k_2}x\right)^{-\frac{1}{2}(k_1 + k_2)} , \ x \geq 0$$

usually $k_1$ and $k_2$ are positive integers.

\textbf{Lemma 2.} If $S^2_1$ and $S^2_2$ are the variances of independent random samples of size $n_1$ and $n_2$ from normal populations with the variance $\sigma_1^2$ and $\sigma_2^2$, then

$$F = \frac{S^2_1/\sigma_1^2}{S^2_2/\sigma_2^2}$$

is a random variable having an $F$ distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

In this sense, the $F$ distribution is also known as the variance-ratio distribution.

\section{3 Limit Theorems}

Let $\{X_n\}$ be a sequence of i.i.d. random variables, and let $\eta = \mathbb{E}X_1$,

$$S_n = X_1 + X_2 + \ldots + X_n.$$

We have the following limit theorems concerning the average of $S_n$.

\textbf{Theorem 1.} (Weak Law of Large Numbers, WLLN) If $\mathbb{E}|X_i| < +\infty$, then

$$\frac{S_n}{n} \to \eta$$

in probability.

\textbf{Theorem 2.} (Strong Law of Large Numbers, SLLN)

$$\frac{S_n}{n} \to \eta \quad \text{a.s.}$$

if and only if $\mathbb{E}|X_i| < +\infty$

\textbf{Theorem 3.} (Central Limit Theorem, CLT) Assume that $\mathbb{E}X^2_i < +\infty$ and let $\sigma^2 = \text{Var}(X_i)$. Then

$$\frac{S_n - n\eta}{\sqrt{n\sigma^2}} \to N(0, 1)$$

in the sense of distribution.
Proof. Assume without loss of generality $\eta = 0$, let $f$ be the characteristic function of $X_i$ and let $g_n$ be the characteristic function of $S_n \sqrt{n\sigma^2}$. Then
\[
g_n(\xi) = \mathbb{E}e^{i\xi \frac{S_n}{\sqrt{n\sigma^2}}} = \prod_{j=1}^{n} \mathbb{E}e^{i\xi \frac{X_i}{\sqrt{n\sigma^2}}} = \prod_{j=1}^{n} f\left(\frac{\xi}{\sqrt{n\sigma^2}}\right) = f\left(\frac{\xi}{\sqrt{n\sigma^2}}\right)^n.
\]
Using Taylor expansion
\[
f\left(\frac{\xi}{\sqrt{n\sigma^2}}\right) = f(0) + \frac{\xi}{\sqrt{n\sigma^2}} f'(0) + \frac{1}{2} \left(\frac{\xi}{\sqrt{n\sigma^2}}\right)^2 f''(0) + \ldots
\]
\[
= 1 - \frac{1}{2} \left(\frac{\xi^2}{n}\right) + o\left(\frac{1}{n}\right).
\]
Hence
\[
g_n(\xi) = f\left(\frac{\xi}{\sqrt{n\sigma^2}}\right)^n = \left(1 - \frac{\xi^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{1}{2}\xi^2}
\]
for every $\xi \in \mathbb{R}^1$. The continuity of $e^{-\frac{1}{2}\xi^2}$ at 0 is trivial. This completes the proof.

4 Homeworks

- HW. Numerically investigate the limit process
  Binomial $\rightarrow$ Poisson $\rightarrow$ Normal distribution
  with MATLAB. Find the suitable parameter regime that the limit holds.

- Prove the formula (1).

- Prove that in the limit $n \to \infty, p \to 0, np = \lambda, \delta t = p/\lambda = 1/n, k = t/\delta t$, the geometric distribution approaches to the exponential distribution.

- Prove that for large $n$, the sampling distribution of the median for random samples of size $2n + 1$ is approximately normal with the mean $\tilde{\mu}$ and the variance $\frac{1}{8p^2(\tilde{\mu})n}$. Here we ask that the probability density $p(x)$ is continuous and nonzero at the median $\tilde{\mu}$, and the median means $\int_{-\infty}^{\tilde{\mu}} p(x)dx = \frac{1}{2}$

References
