

Lecture 0: Review on Probability and Statistics

Instructor: Yen-Chi Chen

0.1 Random Variables

Here we will ignore the formal mathematical definition of a random variable and directly talk about its property. For a random variable X , the *cumulative distribution function (CDF)* of X is

$$P_X(x) = F(x) = P(X \leq x).$$

Actually, the distribution of X is completely determined by the CDF $F(x)$, regardless of X being a discrete random variable or a continuous random variable (or a mix of them).

If X is discrete, its probability mass function (PMF) is

$$p(x) = P(X = x).$$

If X is continuous, its probability density function (PDF) is

$$p(x) = F'(x) = \frac{d}{dx}F(x).$$

Moreover, the CDF can be written as

$$F(x) = P(X \leq x) = \int_{-\infty}^x p(x')dx'.$$

Generally, we write $X \sim F$ or $X \sim p$ indicating that the random variable X has a CDF F or a PMF/PDF p .

For two random variables X, Y , their joint CDF is

$$P_{XY}(x, y) = F(x, y) = P(X \leq x, Y \leq y).$$

The corresponding joint PDF is

$$p(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

The *conditional PDF* of Y given $X = x$ is

$$p(y|x) = \frac{p(x, y)}{p(x)},$$

where $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$ is sometimes called the marginal density function. Note that you can define the joint PMF and conditional PMF using a similar way.

0.2 Expected Value

For a function $g(x)$, the quantity $g(X)$ will also be a random variable and its expected value is

$$\mathbb{E}(g(X)) = \int g(x)dF(x) = \begin{cases} \int_{-\infty}^{\infty} g(x)p(x)dx, & \text{if } X \text{ is continuous} \\ \sum_x g(x)p(x), & \text{if } X \text{ is discrete} \end{cases}.$$

When $f(x) = x$, this reduces to the usual definition of expected value.

Here are some useful properties and quantities related to the expected value:

- $\mathbb{E}(\sum_{j=1}^k c_j g_j(X)) = \sum_{j=1}^k c_j \cdot \mathbb{E}(g_j(X_i))$.
- We often write $\mu = \mathbb{E}(X)$ as the mean (expectation) of X .
- $\text{Var}(X) = \mathbb{E}((X - \mu)^2)$ is the variance of X .
- If X_1, \dots, X_n are independent, then

$$\mathbb{E}(X_1 \cdot X_2 \cdots X_n) = \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \cdots \mathbb{E}(X_n).$$

- If X_1, \dots, X_n are independent, then

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \cdot \text{Var}(X_i).$$

- For two random variables X and Y with their mean being μ_X and μ_Y and variance being σ_X^2 and σ_Y^2 .
The covariance

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_x)(Y - \mu_y)) = \mathbb{E}(XY) - \mu_x \mu_y$$

and the (Pearson's) correlation

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

The **conditional expectation** of Y given X is the random variable $\mathbb{E}(Y|X) = g(X)$ such that when $X = x$, its value is

$$\mathbb{E}(Y|X = x) = \int yp(y|x)dy,$$

where $p(y|x) = p(x, y)/p(x)$.

0.3 Common Distributions

0.3.1 Discrete Random Variables

Bernoulli. If X is a Bernoulli random variable with parameter p , then $X = 0$ or, 1 such that

$$P(X = 1) = p, \quad P(X = 0) = 1 - p.$$

In this case, we write $X \sim \text{Ber}(p)$.

Binomial. If X is a binomial random variable with parameter (n, p) , then $X = 0, 1, \dots, n$ such that

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

In this case, we write $X \sim \text{Bin}(n, p)$. Note that if $X_1, \dots, X_n \sim \text{Ber}(p)$, then the sum $S_n = X_1 + X_2 + \dots + X_n$ is a binomial random variable with parameter (n, p) .

Poisson. If X is a Poisson random variable with parameter λ , then $X = 0, 1, 2, 3, \dots$ and

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

In this case, we write $X \sim \text{Poi}(\lambda)$.

0.3.2 Continuous Random Variables

Uniform. If X is a uniform random variable over the interval $[a, b]$, then

$$p(x) = \frac{1}{b-a} I(a \leq x \leq b),$$

where $I(\text{statement})$ is the indicator function such that if the **statement** is true, then it outputs 1 otherwise 0. Namely, $p(x)$ takes value $\frac{1}{b-a}$ when $x \in [a, b]$ and $p(x) = 0$ in other regions. In this case, we write $X \sim \text{Uni}[a, b]$.

Normal. If X is a normal random variable with parameter (μ, σ^2) , then

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

In this case, we write $X \sim N(\mu, \sigma^2)$.

Exponential. If X is an exponential random variable with parameter λ , then X takes values in $[0, \infty)$ and

$$p(x) = \lambda e^{-\lambda x}.$$

In this case, we write $X \sim \text{Exp}(\lambda)$. Note that we can also write

$$p(x) = \lambda e^{-\lambda x} I(x \geq 0).$$

0.4 Useful Theorems

We write $X_1, \dots, X_n \sim F$ when X_1, \dots, X_n are IID (independently, identically distributed) from a CDF F . In this case, X_1, \dots, X_n is called a *random sample*.

For a sequence of random variables Z_1, \dots, Z_n, \dots , we say Z_n **converges in probability** to a fixed number μ if for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|Z_n - \mu| > \epsilon) = 0$$

and we will write

$$Z_n \xrightarrow{P} \mu.$$

In other words, Z_n converges in probability implies that the distribution is concentrating at the targeting point.

Let F_1, \dots, F_n, \dots be the corresponding CDFs of Z_1, \dots, Z_n, \dots . For a random variable Z with CDF F , we say Z_n **converges in distribution** to Z if for every x ,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x).$$

In this case, we write

$$Z_n \xrightarrow{D} Z.$$

Namely, the CDF's of the sequence of random variables converge to a the CDF of a fixed random variable.

Theorem 0.1 (Weak) Law of Large Number. Let $X_1, \dots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$. If $\mathbb{E}|X_1| < \infty$, then the sample average

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

converges in probability to μ . i.e.,

$$\bar{X}_n \xrightarrow{P} \mu.$$

Theorem 0.2 Central Limit Theorem. Let $X_1, \dots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \text{Var}(X_1) < \infty$. Let \bar{X}_n be the sample average. Then

$$\sqrt{n} \left(\frac{\bar{X}_n - \mu}{\sigma} \right) \xrightarrow{D} N(0, 1).$$

Note that $N(0, 1)$ is also called standard normal random variable.

0.5 Estimators and Estimation Theory

Let $X_1, \dots, X_n \sim F$ be a random sample. Here we can interpret F as the population distribution we are sampling from (that's why we are generating data from this distribution). Any numerical quantity (or even non-numerical quantity) of F that we are interested in is called the **parameter of interest**. For instance, the parameter of interest can be the mean of F , the median of F , standard deviation of F , first quartile of F , ... etc. The parameter of interest can even be $P(X \geq t) = 1 - F(t) = S(t)$. The function $S(t)$ is called the *survival function*, which is a central topic in biostatistics and medical research.

When we know (or assume) that F is a certain distribution with some parameters, then the parameter of interest can be the parameter describing that distribution. For instance, if we assume F is an exponential distribution with an unknown parameter λ . Then this unknown parameter λ might be the parameter of interest.

Most of the statistical analysis is concerned with the following question:

“given the parameter of interest, how can I use the random sample to infer it?”

Let $\theta = \theta(F)$ be the parameter of interest and let $\hat{\theta}_n$ be a statistic (a function of the random sample X_1, \dots, X_n) that we use to estimate θ . In this case, $\hat{\theta}_n$ is called an *estimator*. For an estimator, there are two important quantities measuring its quality. The first quantity is the **bias**:

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta,$$

which captures the systematic deviation of the estimator from its target. The other quantity is the **variance** $\text{Var}(\hat{\theta}_n)$, which measures the size of stochastic fluctuation.

Example. Let $X_1, \dots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \text{Var}(X)$. Assume the parameter of interest is the population mean μ . Then a natural estimator is the sample average $\hat{\mu}_n = \bar{X}_n$. Using this estimator, then

$$\text{bias}(\hat{\mu}_n) = \mu - \mu = 0, \quad \text{Var}(\hat{\mu}_n) = \frac{\sigma^2}{n}.$$

Therefore, when $n \rightarrow \infty$, both bias and variance converge to 0. Thus, we say $\hat{\mu}_n$ is a **consistent** estimator of μ . Formally, an estimator $\hat{\theta}_n$ is called a consistent estimator of θ if $\hat{\theta}_n \xrightarrow{P} \theta$.

The following lemma is a common approach to prove consistency:

Lemma 0.3 Let $\hat{\theta}_n$ be an estimator of θ . If $\text{bias}(\hat{\theta}_n) \rightarrow 0$ and $\text{Var}(\hat{\theta}_n) \rightarrow 0$, then $\hat{\theta}_n \xrightarrow{P} \theta$. i.e., $\hat{\theta}_n$ is a consistent estimator of θ .

In many statistical analysis, a common measure of the quality of the estimator is the *mean square error* (MSE), which is defined as

$$\text{MSE}(\hat{\theta}_n) = \text{MSE}(\hat{\theta}_n, \theta) = \mathbb{E} \left((\hat{\theta}_n - \theta)^2 \right).$$

By simple algebra, the MSE of $\hat{\theta}_n$ equals

$$\begin{aligned} \text{MSE}(\hat{\theta}_n, \theta) &= \mathbb{E} \left((\hat{\theta}_n - \theta)^2 \right) \\ &= \mathbb{E} \left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + \mathbb{E}(\hat{\theta}_n) - \theta)^2 \right) \\ &= \underbrace{\mathbb{E} \left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2 \right)}_{=\text{Var}(\hat{\theta}_n)} + 2 \underbrace{\mathbb{E} \left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) \right)}_{=0} \cdot (\mathbb{E}(\hat{\theta}_n) - \theta) + \underbrace{\left(\mathbb{E}(\hat{\theta}_n) - \theta \right)^2}_{=\text{bias}^2(\hat{\theta}_n)} \\ &= \text{Var}(\hat{\theta}_n) + \text{bias}^2(\hat{\theta}_n). \end{aligned}$$

Namely, the MSE of an estimator is the variance plus the square of bias. This decomposition is also known as the *bias-variance tradeoff* (or bias-variance decomposition). By the Markov inequality,

$$\text{MSE}(\hat{\theta}_n, \theta) \rightarrow 0 \implies \hat{\theta}_n \xrightarrow{P} \theta.$$

i.e., if an estimator has MSE converging to 0, then it is a consistent estimator. The convergence of MSE is related to the L_2 convergence in probability theory.

Note that we write $\theta = \theta(F)$ for the parameter of interest because θ is a quantity derived from the population distribution F . Thus, we may say that the parameter of interest θ is a ‘functional’ (function of function; the input is a function, and the output is a real number).

◆ : There are two common methods of finding an estimator: the first one is called the MLE (maximum likelihood estimator), the other one is called the MOM (method of moments)¹. You can google these two terms and you will find lots of references about them.

Question to think: if the parameter of interest is $F(x) = P(X \leq x)$, what will be the estimator of it?

¹[https://en.wikipedia.org/wiki/Method_of_moments_\(statistics\)](https://en.wikipedia.org/wiki/Method_of_moments_(statistics)) and MIT open course

0.6 O_P and o_P Notations

For a sequence of numbers a_n (indexed by n), we write $a_n = o(1)$ if $a_n \rightarrow 0$ when $n \rightarrow \infty$. For another sequence b_n indexed by n , we write $a_n = o(b_n)$ if $a_n/b_n = o(1)$.

For a sequence of numbers a_n , we write $a_n = O(1)$ if for all large n , there exists a constant C such that $|a_n| \leq C$. For another sequence b_n , we write $a_n = O(b_n)$ if $a_n/b_n = O(1)$.

Examples.

- Let $a_n = \frac{2}{n}$. Then $a_n = o(1)$ and $a_n = O(\frac{1}{n})$.
- Let $b_n = n + 5 + \log n$. Then $b_n = O(n)$ and $b_n = o(n^2)$ and $b_n = o(n^3)$.
- Let $c_n = 1000n + 10^{-10}n^2$. Then $c_n = O(n^2)$ and $c_n = o(n^2 \cdot \log n)$.

Essentially, the big O and small o notation give us a way to compare the leading convergence/divergence rate of a sequence of (non-random) numbers.

The O_P and o_P are similar notations to O and o but are designed for random numbers. For a sequence of random variables X_n , we write $X_n = o_P(1)$ if for any $\epsilon > 0$,

$$P(|X_n| > \epsilon) \rightarrow 0$$

when $n \rightarrow \infty$. Namely, $P(|X_n| > \epsilon) = o(1)$ for any $\epsilon > 0$. Let a_n be a nonrandom sequence, we write $X_n = o_P(a_n)$ if $X_n/a_n = o_P(1)$.

In the case of O_P , we write $X_n = O_P(1)$ if for every $\epsilon > 0$, there exists a constant C such that

$$P(|X_n| > C) \leq \epsilon.$$

We write $X_n = O_P(a_n)$ if $X_n/a_n = O_P(1)$.

Examples.

- Let X_n be an R.V. (random variable) from a Exponential distribution with $\lambda = n$. Then $X_n = O_P(\frac{1}{n})$
- Let Y_n be an R.V from a normal distribution with mean 0 and variance n^2 . Then $Y_n = O_P(n)$ and $Y_n = o_P(n^2)$.
- Let A_n be an R.V. from a normal distribution with mean 0 and variance $10^{100} \cdot n^2$ and B_n be an R.V. from a normal distribution with mean 0 and variance $0.1 \cdot n^4$. Then $A_n + B_n = O_P(n^2)$.

If we have a sequence of random variables $X_n = Y_n + a_n$, where Y_n is random and a_n is non-random such that $Y_n = O_P(b_n)$ and $a_n = O(c_n)$. Then we write

$$X_n = O_P(b_n) + O(c_n).$$

Examples.

- Let A_n be an R.V. from a uniform distribution over the interval $[n^2 - 2n, n^2 + 2n]$. Then $A_n = O(n^2) + O_P(n)$.
- Let X_n be an R.V from a normal distribution with mean $\log n$ and variance 10^{100} , then $X_n = O(\log n) + O_P(1)$.

The following lemma is an important property for a sequence of random variables X_n .

Lemma 0.4 *Let X_n be a sequence of random variables. If there exists a sequence of numbers a_n, b_n such that*

$$|\mathbb{E}(X_n)| \leq a_n, \quad \text{Var}(X_n) \leq b_n^2.$$

Then

$$X_n = O(a_n) + O_P(b_n).$$

Examples.

- Let X_1, \dots, X_n be IID from $\text{Exp}(5)$. Then the sample average

$$\bar{X}_n = O(1) + O_P(1/\sqrt{n}).$$

- Let Y_1, \dots, Y_n be IID from $N(5 \log n, 1)$. Then the sample average

$$\bar{Y}_n = O(\log n) + O_P(1/\sqrt{n}).$$

The following is a useful method for obtaining bounds on O_P :

Lemma 0.5 *Let X be a non-negative random variable. Then for any positive number t ,*

$$P(X \geq t) \leq \frac{\mathbb{E}(X)}{t}.$$

Application.

- Let X_n be a sequence of random variables that are uniformly distributed over $[-n^2, n^2]$. It is easy to see that $|X_n| \leq n^2$ so $\mathbb{E}(|X_n|) \leq n^2$. Then by Markov's inequality,

$$P(|X_n| \geq t) \leq \frac{\mathbb{E}(|X_n|)}{t} \leq \frac{n^2}{t}.$$

Let $Y_n = \frac{1}{n^2} X_n$. Then

$$P(|Y_n| \geq t) = P\left(\frac{1}{n^2}|X_n| \geq t\right) = P(|X_n| \geq n^2 \cdot t) \leq \frac{n^2}{n^2 \cdot t} = \frac{1}{t}$$

for any positive t . This implies $Y_n = O_P(1)$ so $X_n = O_P(n^2)$.