Lecture Notes 0 – Probability

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

March 2025

Random Variables

Expected Value

Common Distributions

Useful Theorems

Estimators and Estimation Theory

Random Variables

$X \sim F$ or $X \sim p$
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 $p(x) = \int_{-\infty}^{\infty} p(x, y) dy$

Expected Value

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

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

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

- If $X_1, \cdots, X_n$ are independent, then

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

- Covariance

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

- (Pearson's) correlation

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

Conditional Expectation

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

Common discrete distributions

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 $X = 0, 1, \cdots, n$ 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 $X_1, \cdots, X_n \sim \text{Ber}(p)$, then the sum $S_n = X_1 + X_2 + \cdots + X_n$ is a binomial random variable with parameter (n, p).
Poisson. If X is a Poisson random variable with parameter λ, then $X = 0, 1, 2, 3, \cdots$ and

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

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

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(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

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

Beta. If X is a beta random variable with parameter (α, β), then X takes values in $[0, 1]$ and

$$p(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad B(\alpha, \beta) = \int x^{\alpha-1}(1-x)^{\beta-1} dx.$$

Convergence for random variables

- **random sample** $X_1, \cdots, X_n \sim F$ are IID (independently, identically distributed) from a CDF F.
- For a sequence of random variables $Z_1, \cdots, Z_n, \cdots$, we say Z_n **converges in probability** to a fixed number μ iff for any $\epsilon > 0$,

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

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

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

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

  Notation $Z_n \xrightarrow{D} Z$.

Theorem ((Weak) Law of Large Numbers)

Let $X_1, \cdots, X_n \sim F$ *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 (**Central Limit Theorem**)

Let $X_1, \cdots, X_n \sim F$ *and* $\mu = \mathbb{E}(X_1)$ *and* $\sigma^2 = \mathrm{Var}(X_1) < \infty$. *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.*

Estimators and Estimation Theory

Let $X_1, \cdots, X_n \sim F$ be a random sample.

- **parameter of interest** $\theta = \theta(F)$
  - the mean of F, the median of F, standard deviation of F, first quartile of F, ...
  - $P(X \geq t) = 1 - F(t) = S(t)$ survival function
  - unknown parameter λ for $\exp(\lambda)$ distribution

- **statistic** $T_n \equiv T(X_1, \cdots, X_n)$ a function of the random sample

- **estimator** $\hat{\theta}_n =$ a statistics we use to estimate $\theta(F)$

Question *"given the parameter of interest, how can I use the random sample to infer it?"*

How good is an estimator?

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

Example.

- Let $X_1, \cdots, X_n \sim F$ and $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \text{Var}(X)$.
- parameter of interest is the population mean μ.
- a natural estimator is the sample average $\hat{\mu}_n = \bar{X}_n$.
- $\text{bias}(\hat{\mu}_n) = \mu - \mu = 0, \quad \text{Var}(\hat{\mu}_n) = \frac{\sigma^2}{n}$.
- Hence, when $n \to \infty$, both bias and variance converge to 0.
- we say $\hat{\mu}_n$ is a **consistent** estimator of μ.

Definition

An estimator $\hat{\theta}_n$ is called a **consistent** *estimator of θ if $\hat{\theta}_n \xrightarrow{P} \theta$.*

Lemma

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

Mean Square Error (MSE)

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

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

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

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).

Lemma

If an estimator has MSE converging to 0*, then it is a consistent estimator.*

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

MLE and MOM

There are two common methods of finding an estimator:

MLE Maximum Likelihood Estimator
MOM Method of Moments

Exercise If the parameter of interest is $F(x) = P(X \leq x)$, what will be the estimator of it?