

MH3500 Summary for Chapter 1 and 2

MH3500

April 8, 2024

1 Discrete Random Variables

- (a) $X \sim \text{Bernoulli}(p)$.
 - $P(X = x) = (1 - p)^{1-x}p^x$, for $x = 0, 1$.
 - $E(X) = p$, $\text{Var}(X) = p(1 - p)$, $M_X(t) = pe^t + 1 - p$
- (b) $X \sim \text{Binomial}(n, p)$.
 - $P(X = x) = \binom{n}{x}(1 - p)^{n-x}p^x$, for $x = 0, 1, \dots, n$.
 - $E(X) = np$, $\text{Var}(X) = np(1 - p)$, $M_X(t) = (pe^t + 1 - p)^n$
- (c) $X \sim \text{Geometric}(p)$.
 - $P(X = x) = p(1 - p)^{x-1}$, for $x = 1, 2, \dots$
 - $E(X) = \frac{1}{p}$, $\text{Var}(X) = \frac{1-p}{p^2}$, $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$, for $t < -\ln(1 - p)$
- (d) $X \sim \text{Poisson}(\lambda)$.
 - $P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$, for $x = 0, 1, \dots$
 - $E(X) = \lambda$, $\text{Var}(X) = \lambda$, $M_X(t) = e^{\lambda(e^t - 1)}$

2 Continuous Random Variables

- (a) $X \sim \text{Uniform}(a, b)$.
 - $f_X(x) = \frac{1}{b-a}$, for $x \in [a, b]$.
 - $E(X) = \frac{a+b}{2}$, $\text{Var}(X) = \frac{(b-a)^2}{12}$, $M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$, for $t \neq 0$
- (b) $X \sim N(\mu, \sigma^2)$.
 - $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$, for $x \in \mathbb{R}$.
 - $E(X) = \mu$, $\text{Var}(X) = \sigma^2$, $M_X(t) = \exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right)$
- (c) $X \sim \text{Exp}(\lambda)$.
 - $f_X(x) = \lambda e^{-\lambda x}$, for $x > 0$.
 - $E(X) = \frac{1}{\lambda}$, $\text{Var}(X) = \frac{1}{\lambda^2}$, $M_X(t) = \frac{\lambda}{\lambda - t}$, for $t < \lambda$
- (d) $X \sim \text{Gamma}(\alpha, \beta)$.
 - $f_X(x) = \frac{x^{\alpha-1}}{\beta^\alpha \Gamma(\alpha)} e^{-\frac{x}{\beta}}$, for $x > 0$.
 - $E(X) = \alpha\beta$, $\text{Var}(X) = \alpha\beta^2$, $M_X(t) = (1 - \beta t)^{-\alpha}$, for $t < \frac{1}{\beta}$
- (e) Chi-squared distribution: If $X_1, \dots, X_n \sim N(0, 1)$, then $X_1^2 + \dots + X_n^2 \sim \chi_n^2 \sim \text{Gamma}\left(\frac{n}{2}, 2\right)$.
- (f) t-distribution: If $X \sim N(0, 1)$ and $Y \sim \chi_k^2$, then $\frac{X}{\sqrt{Y/k}} \sim t_k$.
- (g) F-distribution: If $U \sim \chi_m^2$ and $V \sim \chi_n^2$, then $\frac{U/m}{V/n} \sim F(m, n)$.

3 Moment Generating Functions

Let X be a random variable. The moment generating function (MGF) of X is $M_X(t) = E(e^{tX})$.

- If $M_X(t) = M_Y(t)$ for all $t \in [-a, a]$, then X and Y has the same distribution.
- If $Y = aX + b$ for $a, b \in \mathbb{R}$, $M_Y(t) = e^{tb}M_X(at)$.
- If X and Y are independent random variables, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

If $M_X(t)$ exists in an open interval containing zero, then we can generate the moments by differentiating the MGF. The n -th moment of X is

$$E(X^n) = \frac{d^n}{dt^n} M_X(t)|_{t=0}.$$

The n -th central moment is defined as $E((X - E(X))^n)$.

4 Sample Mean and Sample Variance

Let X_1, \dots, X_n be i.i.d. random variables with population mean $E(X_1) = \mu$ and population variance $Var(X_1) = \sigma^2$. We define the following statistics:

$$\text{Sample mean, } \bar{X} := \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \text{Sample variance, } s^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

In general, $E(\bar{X}) = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n}$ and $E(s^2) = \sigma^2$.

Next, we restrict $X_i \sim N(\mu, \sigma^2)$, for all $i = 1, \dots, n$. Then,

$$\left[\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \implies \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1) \right] \quad \text{and} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Suppose we want to approximate μ with a given set of n samples.

- If σ^2 is known, use $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$.
- If σ^2 is unknown, use s^2 to approximate σ^2 and for $n > 30$, use $\frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim N(0, 1)$.
- If σ^2 is unknown, use s^2 to approximate σ^2 and for $n \leq 30$, use $\frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}$.

5 Important Definitions and Theorems

Definition 1 (Convergence in Probability). A sequence of random variables $\{X_n\}$ is said to converge in probability towards a random variable X if for all $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0.$$

Definition 2 (Convergence in Distribution). A sequence of random variables $\{X_n\}$ is said to converge in distribution towards a random variable X if

$$\lim_{n \rightarrow \infty} |P(X_n \leq x) - P(X \leq x)| = 0,$$

for all $x \in \mathbb{R}$ at which $P(X \leq x)$ is continuous.

Theorem 1 (Lévy's Continuity Theorem). Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables with corresponding MGF $M_n(t)$ and CDF F_n . Suppose there is a random variable X with MGF $M(t)$ and CDF F .

If for all $t \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} M_n(t) = M(t),$$

then

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

for $x \in \mathbb{R}$, where F is continuous. This means that $\{X_n\}$ converges in distribution to X .

Theorem 2 (Central Limit Theorem). Let X_1, \dots, X_n be i.i.d. r.v. with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2 < \infty$. Let $Z \sim N(0, 1)$ and write

$$Z_n = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma},$$

then Z_n converges in distribution to Z . Furthermore, the convergence is uniform, i.e.,

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |P(Z_n \leq x) - P(Z \leq x)| = 0.$$

Theorem 3 (Weak Law of Large Numbers). Let X_1, \dots, X_n be i.i.d. r.v. with $E(X_1) = \mu$, then \bar{X} converges to μ in probability, i.e., for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| > \varepsilon) = 0.$$

Theorem 4 (Chebyshev's Inequality). Let X be a random variable with mean μ and finite non-zero variance σ^2 . For any real number $k > 0$,

$$P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2}.$$

This is a special case of Markov's Inequality involving the second central moment of X .

6 Method of Moments Estimators

Let X_1, \dots, X_n be i.i.d. random variables of distribution \mathcal{D}_θ , where θ is unknown. The m -th sample moment is denoted as

$$s_m = \frac{1}{n} \sum_{i=1}^n X_i^m.$$

The m -th moment is denoted as $\mu_m = E(X_1^m)$. It can be shown using the Weak LLN, that S_m converges in probability to μ_m . We aim to write $\theta = h(\mu_1, \dots, \mu_m)$ for some function h . Then, the MME for θ is $\hat{\theta}_{MME} = h(s_1, \dots, s_m)$.

7 Maximum Likelihood Estimators

Let X_1, \dots, X_n be i.i.d. random variables of distribution \mathcal{D}_θ , where θ is unknown. Let $\mathbb{S} = (a, b)$ be the range of parameters of θ . Let $f(X_i|\theta)$ be the mass/density function for X_i . The likelihood function for X_1, \dots, X_n is

$$L(X_1, \dots, X_n|\theta) = \prod_{i=1}^n f(X_i|\theta).$$

The MLE for θ is defined as

$$\hat{\theta}_{MLE} = \arg \max_{\theta \in \mathbb{S}} L(X_1, \dots, X_n|\theta).$$

To ensure $\hat{\theta}$ exists, we need to verify the standard conditions for $L(\theta)$, for $\theta \in (a, b)$, i.e.,

- $L(\theta) > 0$, for all $\theta \in (a, b)$.
- $L'(\theta)$ exists for all $\theta \in (a, b)$.
- $\lim_{\theta \rightarrow a^+} L(\theta) = \lim_{\theta \rightarrow b^-} L(\theta) = 0$.

If the maximizer exists, we solve $\frac{d}{d\theta} L(\theta) = 0$ for θ . However, if this is complicated to solve, we can resolve to the log-likelihood function, that is to solve $\frac{d}{d\theta} \ln L(\theta) = 0$ for θ to obtain $\hat{\theta}_{MLE}$.

8 Bias, Variance, Consistency

Let X_1, \dots, X_n be i.i.d. random variables of distribution \mathcal{D}_θ , where θ is unknown. Let $\hat{\theta}$ be an estimator for θ based on the n samples. We have the following quantities:

- Bias($\hat{\theta}$) = $E(\hat{\theta}) - \theta$.
- Var($\hat{\theta}$) = $E((\hat{\theta} - E(\hat{\theta}))^2)$.
- Standard Error: SE($\hat{\theta}$) = $\sqrt{Var(\hat{\theta})}$.

- Mean Squared Error: $MSE(\hat{\theta}) = E((\hat{\theta} - \theta)^2) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$.

$\hat{\theta}$ can be computed explicitly based on a given set of observations. For this set of observations, the performance of $\hat{\theta}$ can be measured via the estimated standard error, $\widehat{SE}(\hat{\theta})$, the standard error evaluated based on observations.

Definition 3 (Consistency). $\hat{\theta}$ is said to be a consistent estimator for θ if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0.$$

Theorem 5 (Sufficient Condition for Consistency). Let $\hat{\theta}$ be an estimator for θ . If

$$\lim_{n \rightarrow \infty} \text{Bias}(\hat{\theta}) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}) = 0,$$

then $\hat{\theta}$ is a consistent estimator.

Theorem 6 (MME Consistency). Let $\hat{\theta} = h(s_1, \dots, s_m)$ be the method of moments estimator for θ . If h is continuous in all s_1, \dots, s_m , then $\hat{\theta}$ is a consistent estimator.

Theorem 7 (MLE Consistency). Let $\hat{\theta}$ be the maximum likelihood estimator for θ . Under appropriate regularity assumptions on $L(\theta)$, $\hat{\theta}$ is a consistent estimator.

9 Fisher Information

Let X_1, \dots, X_n be i.i.d. random variables of distribution \mathcal{D}_θ . The likelihood function for X_1, \dots, X_n is

$$L(X_1, \dots, X_n | \theta) = \prod_{i=1}^n f(X_i | \theta).$$

The Fisher information of X_1, \dots, X_n at θ_0 is

$$I_{X_1, \dots, X_n}(\theta_0) = E \left[\left(\frac{d}{d\theta} \ln L(X_1, \dots, X_n | \theta) \Big|_{\theta=\theta_0} \right)^2 \right].$$

If condition (*) is satisfied, i.e.,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x | \theta) dx = 0 \quad (\text{for PDF}) \quad \text{or} \quad \sum_x \frac{\partial}{\partial \theta} f(x | \theta) = 0 \quad (\text{for PMF}),$$

then $I_{X_1, \dots, X_n}(\theta) = n I_{X_1}(\theta)$.

Furthermore, if we have

$$\int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} f(x | \theta) dx = 0 \quad (\text{for PDF}) \quad \text{or} \quad \sum_x \frac{\partial^2}{\partial \theta^2} f(x | \theta) = 0 \quad (\text{for PMF}),$$

then we have the alternative formula for $I_{X_1}(\theta)$,

$$I_{X_1}(\theta) = E \left[\left(\frac{d}{d\theta} \ln L(X_1 | \theta) \right)^2 \right] = -E \left[\frac{d^2}{d\theta^2} \ln L(X_1 | \theta) \right]$$

10 Cramer-Rao Lower Bound (CRLB)

Theorem 8. Let (X_1, \dots, X_n) be any random sample with a joint PDF $f(X_1, \dots, X_n | \theta)$. Let $T = T(X_1, \dots, X_n)$ be an estimator of θ and $E(T)$ is differentiable with respect to θ . Suppose $f(x_1, \dots, x_n | \theta)$ satisfies

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} h(x_1, \dots, x_n) f(x_1, \dots, x_n | \theta) dx_1 \dots dx_n = \int_{\mathbb{R}^n} h(x_1, \dots, x_n) \frac{\partial}{\partial \theta} f(x_1, \dots, x_n | \theta) dx_1 \dots dx_n,$$

for any function h with $E(|h(x)|) < \infty$. Then

$$\text{Var}(T) \geq \frac{\left(\frac{\partial}{\partial \theta} E(T) \right)^2}{E \left[\left(\frac{d}{d\theta} \ln L(X_1, \dots, X_n | \theta) \right)^2 \right]} =: \text{CRLB}.$$

Corollary 1. Suppose X_1, \dots, X_n are i.i.d. random variables of distribution \mathcal{D}_θ , with PDF $f(x|\theta)$. Suppose T is an unbiased estimator of θ , then condition $(*)$ holds by choosing $h(x_1, \dots, x_n) = 1$ in the theorem above and we have

$$\text{Var}(T) \geq \frac{1}{nI_{X_1}(\theta)}.$$

Definition 4. An estimator T for θ is efficient if T is unbiased and $\text{Var}(T) = (nI_{X_1}(\theta))^{-1}$.

Theorem 9 (Asymptotic Normality of MLE). Suppose X_1, \dots, X_n are i.i.d. random variables of distribution \mathcal{D}_θ , with PDF or PMF $f(x|\theta)$, where the true parameter θ is unknown. Let $\hat{\theta}$ be the MLE for θ , under some regularity conditions, $\sqrt{nI_{X_1}(\theta)}(\hat{\theta} - \theta)$ converges in distribution to $N(0, 1)$.

The theorem above describes the asymptotic behaviour of MLE and $\frac{1}{nI_{X_1}(\theta)}$ is the asymptotic variance of $\hat{\theta}$.

To conclude, under some regularity conditions, MLE is

consistent, asymptotically unbiased, asymptotically normal and asymptotically efficient.

The last point is justified since the asymptotic variance is the CRLB.

11 Confidence Interval

Definition 5 (Confidence Interval). Suppose X_1, \dots, X_n are i.i.d. random variables of distribution \mathcal{D}_θ , where the true parameter θ is unknown. Suppose $L(X_1, \dots, X_n)$ and $U(X_1, \dots, X_n)$ are statistics such that

$$P(L \leq \theta \leq U) = 1 - \alpha,$$

then $(L(X_1, \dots, X_n), U(X_1, \dots, X_n))$ is an exact $100(1 - \alpha)\%$ confidence interval for θ .

Definition 6 (Pivotal Quantity). A random variable $Q(X_1, \dots, X_n, \theta)$ is a pivotal quantity if the distribution of Q does not depend on the unknown parameter θ .

To verify that a random variable $Q(X_1, \dots, X_n, \theta)$ is a pivotal quantity, check that the PDF/CDF/MGF of Q does not involve any unknown parameter θ .

Example 1. Suppose $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ are i.i.d. Let \bar{X} and s^2 denote the sample mean and sample variance respectively.

- $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$ is a pivotal quantity to construct the C.I. for σ .
- $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$ is a pivotal quantity to construct the C.I. for μ , if σ is known.
- $\frac{\sqrt{n}(\bar{X} - \mu)}{s} \sim t_{n-1}$ is a pivotal quantity to construct the C.I. for μ , if σ is unknown.

Due to the asymptotical normality of MLE, i.e. $\sqrt{nI_{X_1}(\theta)}(\hat{\theta} - \theta)$ converges in distribution to $N(0, 1)$, this means that $\sqrt{nI_{X_1}(\theta)}(\hat{\theta} - \theta)$ serves as an asymptotical pivotal quantity to construct a confidence interval for θ . The behaviour is only significant for large n . If we define z_α as the number such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0, 1)$, then the approximate $100(1 - \alpha)\%$ C.I. for θ is

$$\left[\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{nI_{X_1}(\hat{\theta})}}, \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{nI_{X_1}(\hat{\theta})}} \right].$$

12 Other Useful Properties

- Suppose X_1, \dots, X_n are i.i.d. random variables with a CDF F . Let $Y_1 = \min(X_1, \dots, X_n)$ and $Y_2 = \max(X_1, \dots, X_n)$, then

$$P(Y_1 \leq y) = 1 - (1 - F(y))^n \quad P(Y_2 \leq y) = (F(y))^n.$$

- If X and Y are independent random variables, then $E(XY) = E(X)E(Y)$. The converse is **TRUE** if X, Y are normally distributed. Generally, the converse is **FALSE**.
- Bernoulli distribution is a special case of binomial distribution with $n = 1$.

- Suppose $X_i \sim \text{Binomial}(n_i, p)$, for $i = 1, \dots, k$ are independent, then

$$X_1 + X_2 \dots + X_k \sim \text{Binomial}(n_1 + \dots + n_k, p).$$

- Suppose $X_i \sim \text{Poisson}(\lambda_i)$, for $i = 1, \dots, k$ are independent, then

$$X_1 + X_2 \dots + X_k \sim \text{Poisson}(\lambda_1 + \dots + \lambda_k).$$

- Suppose $X_i \sim \text{Gamma}(\alpha_i, \beta)$, for $i = 1, \dots, k$ are independent, then

$$X_1 + X_2 \dots + X_k \sim \text{Gamma}(\alpha_1 + \dots + \alpha_k, \beta).$$

- Suppose $X_i \sim N(\mu_i, \sigma_i^2)$, for $i = 1, \dots, k$ are independent, then

$$\sum_{i=1}^k a_i X_i \sim N\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right).$$

- Suppose $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ are independent random variables, then

$$X + Y \sim \chi_{m+n}^2$$

- Gamma function:

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

$$\Gamma(z+1) = z\Gamma(z) = z!, \text{ for } z \in \mathbb{Z}^+.$$

- Beta function:

$$B(m, n) = \int_0^1 x^{m-1} (1-x)^{n-1} dx = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

- Relationship between Exp, Gamma and χ^2 :

$$\text{Exp}(\lambda) \sim \text{Gamma}\left(1, \frac{1}{\lambda}\right), \quad \chi_n^2 \sim \text{Gamma}\left(\frac{n}{2}, 2\right).$$