Probability and Statistics

Luyang Chen

September 20, 2016

1 Basic Probability Theory

1.1 Probability Spaces

- A **probability space** is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set of "outcomes", $\mathcal{F}$ is a set of "events", and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to events.
- A **σ-algebra** (or **σ-field**) $\mathcal{F}$ is a collection of subsets of Ω that satisfy
  1. $\emptyset, \Omega \in \mathcal{F}$.
  2. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
  3. if $A_i \in \mathcal{F}$ is a countable sequence of sets, then $\cup_i A_i \in \mathcal{F}$.
- A **measurable space** $(\Omega, \mathcal{F})$ is a space on which we can put a measure.
- A **measure** $\mu : \mathcal{F} \to \mathbb{R}$ is a nonnegative countably additive set function that satisfies
  1. $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathcal{F}$.
  2. if $A_i \in \mathcal{F}$ is a countable sequence of disjoint sets, then
  $$\mu(\cup_i A_i) = \sum_i \mu(A_i)$$

  If $\mu(\Omega) = 1$, we call μ a **probability measure**.
- Let μ be a measure on $(\Omega, \mathcal{F})$.
  1. **Monotonicity**. If $A \subseteq B$, then $\mu(A) \leq \mu(B)$.
  2. **Subadditivity**. If $A \subseteq \cup_{m=1}^{\infty} A_m$, then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.
  3. **Continuity from below**. If $A_i \uparrow A$ (i.e., $A_1 \subseteq A_2 \subseteq \ldots$ and $\cup_i A_i = A$), then $\mu(A_i) \uparrow \mu(A)$.
  4. **Continuity from above**. If $A_i \downarrow A$ (i.e., $A_i \supseteq A_2 \supseteq \ldots$ and $\cap_i A_i = A$), with $\mu(A_1) < \infty$, then $\mu(A_i) \downarrow \mu(A)$.

1.2 Distributions

- A **random variable** X is a real-valued function defined on Ω, such that for every Borel set $B \subseteq \mathbb{R}$, we have $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$.
- A random variable X is **discrete** if its possible values are finite or countably infinite.
- A random variable X is **continuous** if its possible values form an uncountable set and the probability that X equals any such value exactly is zero.
- A trivial, but useful, type of example of a random variable is the **indicator function** of a set $A \in \mathcal{F}$:
$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

- If X is a random variable, then X induces a probability measure on $\mathbb{R}$ called its **distribution**, by setting $\mu(A) = \mathbb{P}(X \in A)$ for Borel sets A.
- The distribution of a random variable X is described by giving its **distribution function** $F(x) = \mathbb{P}(X \leq x)$.
- Any distribution function F has the following properties:
  1. F is nondecreasing.
  2. $\lim_{x\to\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.
  3. F is right continuous, that is, $\lim_{y\downarrow x} F(y) = F(x)$.
  4. $\lim_{y\uparrow x} F(y) = F(x-) = \mathbb{P}(X < x)$.
- Any function F satisfying $1-3$ above is the distribution function of some random variable.
- When the distribution function $F(x)$ has the form

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

  we say that X has **density function** f.

1.3 Integration & Expected Value

- Suppose f and g are integrable functions on $(\Omega, \mathcal{F}, \mu)$.
  1. If $f \geq 0$ a.e., then $\int f d\mu \geq 0$.
  2. For all $a \in \mathbb{R}$, $\int af d\mu = a \int f d\mu$.
  3. $\int f + g d\mu = \int f d\mu + \int g d\mu$.
  4. If $g \leq f$ a.e., then $\int g d\mu \leq \int f d\mu$.
  5. If $g = f$ a.e., then $\int g d\mu = \int f d\mu$.
  6. $|\int f d\mu| \leq \int |f| d\mu$.
- If X is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, then we define its **expected value** to be $\mathbb{E}[X] = \int X d\mathbb{P}$. $\mathbb{E}[X]$ does not always exist.
- **Jensen's inequality**. Suppose ϕ is convex, and X and $\phi(X)$ are both integrable, then $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$.
- **Hölder's inequality**. If $p, q \in (1, \infty)$ with $1/p + 1/q = 1$, then $\mathbb{E}[|XY|] \leq (\mathbb{E}[|X|^p])^{\frac{1}{p}} (\mathbb{E}[|Y|^q])^{\frac{1}{q}}$.
- The special case $p = q = 2$ is called the **Cauchy-Schwarz inequality**.
- **Markov's inequality**. $\mathbb{P}(|X| \geq a) \leq a^{-1}\mathbb{E}[|X|]$.
- **Chebyshev's inequality**. $\mathbb{P}(|X| \geq a) \leq a^{-2}\mathbb{E}[|X|^2]$.
- If k is a positive integer, then $\mathbb{E}[X^k]$ is called the **kth moment** of X. The first moment $\mathbb{E}[X]$ is usually called the **mean** and denoted by μ. If $\mathbb{E}[X^2] < \infty$, then the **variance** of X is defined to be $\text{var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2$.
- The **covariance** of two random variables X and Y is defined as $\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mu_X\mu_Y$.

1.4 Integration to the Limit

- **Dominated Convergence Theorem**. If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n and $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.
- **Monotone Convergence Theorem**. If $0 \leq X_n \uparrow X$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$.
- **Fatou's Lemma**. If $X_n \geq 0$, then

$$\liminf_{n\to\infty} \mathbb{E}[X_n] \geq \mathbb{E}[\liminf_{n\to\infty} X_n]$$

1.5 Fubini's Theorem

- **Fubini's theorem**. If $f \geq 0$ or $\int |f| d\mu < \infty$, then

$$\int_X \int_Y f(x,y)\mu_2(dy)\mu_1(dx) = \int_{X\times Y} f d\mu = \int_Y \int_X f(x,y)\mu_1(dx)\mu_2(dy)$$

- **Exercise**. Let X be a nonnegative random variable. Show that

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)dt$$

2 Convergence

2.1 Convergence Concepts

- **Converge in probability**. We say that $X_n \to X$ **in probability**, if for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$.
- **Converge in L^p**. We say that $X_n \to X$ **in L^p**, if $\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0$.
- **Converge almost surely**. We say that $X_n \to X$ **a.s.**, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.
- **Converge in distribution**. We say that $X_n \to X$ **in distribution**, their CDFs converge, i.e. $F_n(x) \to F(x)$ for any continuous point x of F.
- **Note**. The following three statements are equivalent:
  1. $\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$ for all bounded and continuous $g(x)$.
  2. $\lim_{n\to\infty} \mathbb{E}[e^{i\alpha X_n}] = \mathbb{E}[e^{i\alpha X}]$ pointwise for all $\alpha \in \mathbb{R}$.
  3. $\lim_{n\to\infty} F_n(x) = F(x)$ for any continuous point x of F.

2.2 Relationship between Different Convergeces

- If $X_n \overset{a.s.}{\to} X$, then $X_n \overset{\mathbb{P}}{\to} X$.
  Proof.

$$\begin{aligned}
& \mathbb{P}(\cap_{\varepsilon>0} \cup_{N>0} \cap_{n\geq N} \{|X_n - X| < \varepsilon\}) = 1 \\
\Longrightarrow\ & \mathbb{P}(\cup_{\varepsilon>0} \cap_{N>0} \cup_{n\geq N} \{|X_n - X| \geq \varepsilon\}) = 0 \\
\Longrightarrow\ & \mathbb{P}(\cap_{N>0} \cup_{n\geq N} \{|X_n - N| \geq \varepsilon\}) = 0 \quad \forall \varepsilon > 0 \\
\Longrightarrow\ & \lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0
\end{aligned}$$

- $X_n \xrightarrow{\mathbb{P}} X$ doesn't imply $X_n \xrightarrow{a.s.} X$.
  Counterexample.

$$f_{2^k+i}(t) = \begin{cases} 1 & \frac{i}{2^k} \le t < \frac{i+1}{2^k} \\ 0 & \text{otherwise} \end{cases} \qquad i = 0, 1, \ldots, 2^k - 1, k = 0, 1, \ldots$$

$$X_n = f_n(U)$$

  where U is uniformly distributed on $[0, 1]$. X_n converges to 0 in probability, but not a.s.

- If $X_n \xrightarrow{L^p} X$, then $X_n \xrightarrow{\mathbb{P}} X$.
  Proof.

$$\mathbb{P}(|X_n - X| \ge \varepsilon) \le \frac{\mathbb{E}[|X_n - X|^p]}{\varepsilon^p} \to 0$$

- $X_n \xrightarrow{\mathbb{P}} X$ doesn't imply $X_n \xrightarrow{L^p} X$.
  Counterexample.

$$f_n(t) = \begin{cases} n^{1/p} & 0 \le t < \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

$$X_n = f_n(U)$$

  where U is uniformly distributed on $[0, 1]$. X_n converges to 0 in probability, but not in L^p.

- If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{\mathcal{D}} X$.

- If $X_n \xrightarrow{\mathcal{D}} a$ (constant), then $X_n \xrightarrow{\mathbb{P}} a$.

2.3 Continuous Mapping Theorem and Slutsky's Theorem

- **Continuous Mapping Theorem**. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a continuous function.
  1. If $X_n \xrightarrow{\mathcal{D}} X$, then $g(X_n) \xrightarrow{\mathcal{D}} g(X)$.
  2. If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.
  3. If $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.

- **Slutsky's Theorem** If $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{\mathbb{P}} a$ (constant), then $X_n + Y_n \xrightarrow{\mathcal{D}} X + a$ and $X_n Y_n \xrightarrow{\mathcal{D}} aX$.

2.4 Delta Method

- **Theorem**. Let $X_1, X_2, \ldots$ be a sequence of random variables such that $\sqrt{n}(X_n - a) \xrightarrow{\mathcal{D}} Z$ for some random variable Z and constant a. Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at a. Then $\sqrt{n}(g(X_n) - g(a)) \xrightarrow{\mathcal{D}} g'(a)Z$.
  Proof.

$$\sqrt{n}(g(X_n) - g(a)) = g'(\tilde{X}_n)\sqrt{n}(X_n - a)$$

  where $|\tilde{X}_n - a| \le |X_n - a|$.

$$\sqrt{n}(X_n - a) \xrightarrow{\mathcal{D}} Z \quad \Rightarrow \quad X_n \xrightarrow{\mathbb{P}} a \quad \Rightarrow \quad \tilde{X}_n \xrightarrow{\mathbb{P}} a \quad \Rightarrow \quad g'(\tilde{X}_n) \xrightarrow{\mathbb{P}} g'(a)$$

  Then use Slutsky's Theorem.

2.5 Weak Laws of Large Numbers (WLLN)

- **Theorem**. Let $X_1, X_2, \dots$ be uncorrelated random variables with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) \le C < \infty$. If $S_n = X_1 + \cdots + X_n$ then as $n \to \infty$, $S_n/n \to \mu$ in L^2 and also in probability.
Proof.

$$\mathbb{E}[S_n/n] = \mu$$

$$\mathbb{E}[|S_n/n - \mu|^2] = \text{var}(S_n/n) = \frac{1}{n^2}\text{var}(S_n) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) \le \frac{C}{n} \to 0$$

- **Theorem**. Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[|X_i|] < \infty$. If $S_n = X_1 + \cdots + X_n$ then as $n \to \infty$, $S_n/n \to \mu$ in probability.
Proof.

$$\begin{aligned}
S_n/n - \mu &= \frac{1}{n}\sum_{i=1}^{n}(X_i 1_{\{|X_i|\le n\}} + X_i 1_{\{|X_i|>n\}}) - \mathbb{E}[X_1 1_{\{|X_1|\le n\}}] + \mathbb{E}[X_1 1_{\{|X_1|\le n\}}] - \mathbb{E}[X_1] \\
&= \Big(\frac{1}{n}\sum_{i=1}^{n}(X_i 1_{\{|X_i|\le n\}} - \mathbb{E}[X_1 1_{\{|X_1|\le n\}}])\Big) + \frac{1}{n}\sum_{i=1}^{n} X_i 1_{\{|X_i|>n\}} + \Big(\mathbb{E}[X_1 1_{\{|X_1|\le n\}}] - \mathbb{E}[X_1]\Big) \\
&= I + II + III
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[|I|^2] =& \frac{1}{n}\mathbb{E}[|X_1 1_{\{|X_1|\le n\}} - \mathbb{E}[X_1 1_{\{|X_1|\le n\}}]|^2] \le \frac{1}{n}\mathbb{E}[|X_1|^2 1_{\{|X_1|\le n\}}] \\
=& \frac{1}{n}\mathbb{E}[|X_1|^2 1_{\{|X_1|\le \varepsilon\sqrt{n}\}}] + \frac{1}{n}\mathbb{E}[|X_1|^2 1_{\{\varepsilon\sqrt{n}<|X_1|\le n\}}] \\
\le& \varepsilon^2 + \mathbb{E}[|X_1| 1_{\{|X_1|>\varepsilon\sqrt{n}\}}]
\end{aligned}$$

$$\mathbb{E}[|II|] = \mathbb{E}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} X_i 1_{\{|X_i|>n\}}\Big|\Big] \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_i| 1_{\{|X_i|>n\}}] = \mathbb{E}[|X_1| 1_{\{|X_1|>n\}}] \to 0$$

$$|III| = |\mathbb{E}[X_1 1_{\{|X_1|\le n\}}] - \mathbb{E}[X_1]| \le \mathbb{E}[|X_1| 1_{\{|X_1|>n\}}] \to 0$$

- **Note**. Neither independence of the X_i nor their finite variance are needed for the validity of WLLN.

2.6 Strong Laws of Large Numbers (SLLN)

- **Theorem**. Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathbb{E}[|X_i|] < \infty$. If $S_n = X_1 + \cdots + X_n$ then as $n \to \infty$, $S_n/n \to \mu$ a.s..
- If the i.i.d. random variables $\{X_i\}$ have finite forth order moments, $\mathbb{E}[|X_i|^4] < \infty$ or $\mathbb{E}[|X_i - \mu|^4] < \infty$, then an application of the Chebyshev inequality with $p = 4$ gives the needed estimate and we have the SLLN in this case. Of course, this is only a sufficient condition for its validity. As with the WLLN, it is enough that $\mathbb{E}[|X_i|] < \infty$.

2.7 Central Limit Theorem

- **Theorem**. Let $X_1, x_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. If $S_n = X_1 + \cdots + X_n$ then $\sqrt{n}(S_n/n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$.
Proof.

$$\mathbb{E}[e^{i\alpha\sqrt{n}(S_n/n-\mu)}] = \mathbb{E}[e^{i\frac{\alpha}{\sqrt{n}}\sum_{j=1}^{n}(X_j-\mu)}] = \phi^n(\frac{\alpha}{\sqrt{n}})$$

where $\phi(\alpha) = \mathbb{E}[e^{i\alpha(X_1-\mu)}]$. Then $\phi(0) = 1$, $\phi'(0) = 0$, $\phi''(0) = -\sigma^2$.
By Taylor's theorem, we have

$$\phi(\frac{\alpha}{\sqrt{n}}) = 1 - \phi''(\alpha_n)\frac{\alpha^2}{2n}$$

where $0 < \alpha_n < \frac{\alpha}{\sqrt{n}}$.

$$\phi^n(\frac{\alpha}{\sqrt{n}}) \to e^{-\frac{\alpha^2\sigma^2}{2}}$$

3 Statistics

3.1 Probability and Statistics

- The basic problem of probability is: Given the distribution of the data, what are the properties (e.g. expectation, variance, etc.) of the outcomes?
- The basic problem of statistics is: Given the outcomes, what can we say about the distribution of the data? (Given $X_1, \ldots, X_n \sim F$, what can we say about F?)

3.2 Fundamental Concepts

- **Point estimation** involves the use of sample data to calculate a single value (known as a statistic) which is to serve as a "best guess" or "best estimate" of an unknown (fixed or random) population parameter.
  Let $X_1, \ldots, X_n$ be i.i.d. data points from some distribution $F(x; \theta^*)$. A point estimator $\hat{\theta}_n$ of parameter θ is some function of $X_1, \ldots, X_n$:
  $$\hat{\theta}_n = g(X_1, \ldots, X_n)$$
  We introduce the following two methods: **Method of Moments** and **Maximum Likelihood**.
- In statistics, the **bias** of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator with zero bias is called **unbiased**. Otherwise the estimator is said to be **biased**.
- Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be **consistent** in probability if $\hat{\theta}_n$ converges in probability to θ as n approaches infinity.
- A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that $\mathbb{P}(\theta \in C_n) \geq 1 - \alpha$.

3.3 The Methods of Moments

- The kth moment of a probability law is defined as $\mu_k = \mathbb{E}[X^k]$, where X is a random variable following that probability law.
- If $X_1, \ldots, X_n$ are i.i.d. random variables from that distribution, the kth **sample moment** is defined as $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^n X_i^k$. We can view $\hat{\mu}_k$ as an estimate of μ_k.
- **The method of moments** estimates parameters by finding expressions for them in term of the lowest possible order moments and then substituting sample moments into the expressions.
- **Example**. The first and second moments for the normal distribution $\mathcal{N}(\mu, \sigma^2)$ are
  $$\mu_1 = \mathbb{E}[X] = \mu$$
  $$\mu_2 = \mathbb{E}[X^2] = \mu^2 + \sigma^2$$
  Therefore, $\mu = \mu_1$ and $\sigma^2 = \mu_2 - \mu_1^2$.
  The corresponding estimates of μ and σ^2 from the sample moments are
  $$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$$
  $$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \Big(\frac{1}{n}\sum_{i=1}^n X_i\Big)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

- **Question**. Are the two estimators above unbiased? Are the two estimators above consistent? What are the confidence intervals?

$$\mathbb{E}[\hat{\mu}] = \mu$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\bar{X} - \mu)^2$$

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2$$

$\hat{\mu}$ is unbiased. $\hat{\sigma}^2$ is biased. Both $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent estimators. A $1-\alpha$ confidence interval of $\hat{\mu}$ is $[\mu - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha/2), \mu + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha/2)]$. $(n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1))$.

3.4 The Method of Maximum Likelihood

- Suppose that random variables $X_1, \ldots, X_n$ have a joint density $f(x_1, \ldots, x_n|\theta)$. Given observed values $X_i = x_i$, $i = 1, \ldots, n$, the **likelihood** of θ as a function of $x_1, \ldots, x_n$ is defined as $L(\theta) = f(x_1, \ldots, x_n|\theta)$.
- If X_i are assumed to be i.i.d., the likelihood is $L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$. The **log likelihood** is $l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$.
- The **maximum likelihood estimate (MLE)** of θ is that value of θ that maximizes the likelihood, that is, makes the observed data "most probable" or "most likely".
- The estimates obtained by the method of maximum likelihood are not always the same as those obtained by the method of moments.
- **Example**. If $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, their joint density is the product of their marginal densities:

$$f(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\Big[\frac{x_i - \mu}{\sigma}\Big]^2\Big)$$

The log likelihood is thus

$$l(\mu, \sigma) = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

The partials with respect to μ and σ are

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\hat{\mu}_{MLE} = \bar{X}$$

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- The following are the good properties of the MLE:
  1. Under appropriate smoothness conditions on f, the MLE from an i.i.d. sample is consistent.
  2. Under appropriate smoothness conditions on f, $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1/I(\theta^*))$.
  3. The MLE achieves the Cramer-Rao lower bound.
- **Fisher Information**.

$$I(\theta) = \mathbb{E}\Big[\frac{\partial}{\partial\theta}\log f(X|\theta)\Big]^2 = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\Big]$$

3.5 Hypothesis Testing

- H_0: the **null hypotheses**. H_1 (or H_A): the **alternative hypothesis**.
- Rejecting H_0 when it is true is called a **type I error**.
- The probability of a type I error is called the **significance level** of the test and is usually denoted by α.
- Accepting the null hypothesis when it is false is called a **type II error**. Its probability is usually denoted by β.
- The set of values of the **test statistic** that leads to rejection of the null hypothesis is called the **rejection region**, and the set of values that leads to acceptance is called the **acceptance region**.
- The probability distribution of the test statistic when the null hypothesis is true is called the **null distribution**.
- The **p-value** is the probability of a result as or more extreme than that actually observed if the null hypothesis were true.
- Some familiar hypothesis tests: **z-test**, **Student's t-test**, ...
- **Generalized Likelihood Ratio Test**. Suppose that the observations $X = (X_1, \ldots, X_n)$ have a joint density function $f(x_1, \ldots, x_n|\theta)$. H_0 specifies that $\theta \in \omega_0$ and H_1 specifies that $\theta \in \omega_1$, where $\omega_0 \cap \omega_1 = \emptyset$ and $\Omega = \omega_0 \cup \omega_1$. The test statistic

$$\Lambda = \frac{\max\limits_{\theta \in \omega_0}[L(\theta)]}{\max\limits_{\theta \in \Omega}[L(\theta)]}$$

  Under smoothness conditions on the probability density, the null distribution of $-2\log\Lambda$ tends to a chi-square distribution with degrees of freedom equal to $\dim\Omega - \dim\omega_0$ as the sample size tends to infinity.

3.6 Linear Regression

- Consider the following regression model:

$$Y = X\beta + \varepsilon$$

  where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

  The least square estimator

$$\hat{\beta}_{LS} = \arg\min ||Y - X\beta||_2^2$$

- Consider the model above and we have the following assumptions:
  1. X is non-random matrix with full column rank.
  2. $\mathbb{E}[\varepsilon] = 0$.
  3. $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2\delta_{ij}$.
  4. $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.
- $\hat{\beta}_{LS} = (X^TX)^{-1}X^TY$.
- Under assumption 1-2, $\hat{\beta}_{LS}$ is an unbiased estimator.

- Under assumption 1-3, $\text{Cov}(\hat{\beta}_{LS}) = \sigma^2(X^TX)^{-1}$. An unbiased estimator of σ^2 is

$$s^2 = \frac{1}{n-p}\text{RSS} = \frac{1}{n-p}(Y - X\hat{\beta}_{LS})^T(Y - X\hat{\beta}_{LS})$$

- Under assumption 1 and 4,

$$\hat{\beta}_{LS} \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$$

$$\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}$$

$$\frac{\hat{\beta}_{LS,j} - \beta_j}{s\sqrt{c_{jj}}} \sim t_{n-p}$$

  where c_{jj} is the jth element on the diagonal of $(X^TX)^{-1}$.