Chapter 6: Likelihood Inference

1 The Likelihood Function

Let θ be an (unknown) parameter or vector. The likelihood function $L(\theta)$ is the joint PMF of PDF of data. The loglikelihood function is $\ell(\theta) = \log L(\theta)$. The book uses notations $L(\theta|\mathbf{x})$ and $\ell(\theta|\mathbf{x})$, respectively, where $\mathbf{x}$ represents data.

In statistics, we only have the data. Statistical models or assumptions are proposed. They may not be correct. Therefore, it is important to justify the assumptions. The task is to get information from the data for our interest. The interest is often treated as a parameter, which may be a value, a vector, or a space. In this course, we only focus on the case when it is a value or a two-dimensional vector.

(a): Some important concepts:

- Parameters: a parameter is an unknown constant which affects the distribution of random variables.
- Statistic: a function of data, which becomes a real value or a real vector if data are available. We can only use statistics (plural of statistic) in our method.
- Parameters can only be estimated (learned, CS term) from data; comparing it with the term *machine learning.*
- Estimation, estimator, and estimate. Estimation means a method. Estimator means a formula. Estimate means a value, computed from the formula.
- Data are treated as random before they are available and as values after they are available. A statistical method is often evaluated by repeating data collection and then its implementation.
- Statistics (a subject name) provides methods for data.
- Statistical models or assumptions: artificial or subjective, may be changed, but useful. Only data are real.
- Upper cases mean random variables (e.g. X, Y, and etc). Lower cases mean ovservations (e.g. x, y and etc.

Some examples for the concepts:

- Example 6.1.2. Flip a coin 10 times. Let X be the total of heads with the observed value $x = 4$.

  Solution: The data set is composed of 10 of $\{0, 1\}$, where 0 represents a tail and 1 represents a head. It has 4 one and 6 zero. An example is

  $$(0, 1, 1, 0, 0, 1, 0, 0, 0, 1).$$

  Note that the above is an array. Order matters. $X \sim Bin(10, \theta)$ is a random variable, where $\theta = P(1)$ is the probability of seeing a head. $x = 4$ is the result. If the coin is balanced, then $\theta = 1/2$; otherwise $\theta \neq 1/2$. In statistics, we assume θ is unknown. We need to estimate (learn, CS term) it. An estimator of θ is $X/10$. An estimate of θ is 0.4.

- Example 6.1.3. Flip a coin until four heads are obtained. Let X be the total times of flipping with the observed value $x = 9$.

  Solution: The length of the sequence (denoted by X) of 0 or 1 may vary. The sequence contains only 4 ones. The last one must be one. For examples, $(1,1,1,1)$, $(0,1,1,0,1,1)$, $(0,0,1,1,0,0,0,1,0,1)$. At this time, the observed value of X equals 9, i.e., $x = 9$. Many cases could be contained. You may write a distribution. Note that only data are known. The distribution may not **fit** the data. This is a statistical problem.

- Example 6.1.4. Let $\mathbf{x} = (X_1, \cdots, X_n)^\top$, where $X_i \sim^{iid} N(\mu, \sigma_0^2)$, where σ_0^2 is known.

  Solution: We only have $\mathbf{x}$, which is an n-dimensional vector. We assume $X_i \sim^{iid} N(\mu, \sigma_0^2)$, but this may not hold. However, we have experience that the conclusion based on this assumption is reliable even if the assumption is incorrect. The reason is that we have the CLT (central limit theorem). The assumption that σ_0^2 is known is usually unrealistic, but it can help us understand the methodology. Then, you can use σ_0^2 in your statistical answer.

(b): Sufficient Statistics (SS):

- Definition: $T = T(\mathbf{x})$, a function of data, is called sufficient statistic, if $L[\theta|T(\mathbf{x}_1)]/L[\theta|T(\mathbf{x}_2)]$ does not depend on θ.

- Factorization theorem: If and only if $L(\theta|\mathbf{x}) = h(\mathbf{x})g_\theta(T)$, then T is a sufficient statistic.

- Minimal sufficient statistics (MSS): if the size (i.e. the dimension) of T is mimimized.

- The dimension of MSS cannot be less than the dimension of $\boldsymbol{\theta}$. If the dimension of an SS is equal to the dimension of $\boldsymbol{\theta}$, then it is MSS.

- SS is an important concept in statistical inference. It concludes that one only needs to know SS in statistical inference.

Remark: Keep in mind that we want to get some knowledge for θ from data. We need to remove redundancy. After that, we get the SS. This will be interpreted in a few examples later.

Remark: We cannot do statistical inference without assumptions. The assumptions are often derived based on the distribution of the data.

Remark: Descriptive statistics does not need assumptions. It only describes the performance of the data, which may help us to formulate our models (i.e., assumptions). Inferential statistics wants to draw conclusion not only for the data. Then, we need models.

2 Maximum Likelihood Estimation (MLE)

The MLE, which attempts to maximize $L(\theta)$ to estimate θ, is the most important approach in statistics. A nice property is that the MLE is invariant under transformations: if $\hat{\theta}$ is the MLE of θ, then $u(\hat{\theta})$ is the MLE of $u(\theta)$ for any function $u(\cdot)$.

Let $L(\boldsymbol{\theta})$ (or $L(\boldsymbol{\theta}|\mathbf{x})$) is the likelihood function. The maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}$ is the maximum of $L(\boldsymbol{\theta})$, i.e,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

Some properties:

- The MLE $\hat{\boldsymbol{\theta}}$ is a function of data, which is random.

- The MLE of $g(\boldsymbol{\theta})$ is $g(\hat{\boldsymbol{\theta}})$, which means it is transformation invariant.

- The choice of distributions is important in maximum likelihood estimation.

Unbiased estimator: As estimator $\tilde{\theta}$ of θ is unbiased if $\mathrm{E}\tilde{\theta} = \theta$. An unbiased estimator is not invariant under transformations.

2.1 Computation of the MLE

Let $\ell(\boldsymbol{\theta}) = \log L(\theta)$ is the loglikelihood function. Then, $\boldsymbol{\theta}$ is one of the solutions of

$$\nabla \ell(\boldsymbol{\theta}) = (\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_k})^\top = \mathbf{0}.$$

Mostly, the solution is unique; otherwise, we need to make sure the solution is a global maximum (this is a hard topic in research).

Examples for SS, MSS and MLE as well as the unbiasedness:

- Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$.

  Solution: The PMF of the Bernoulli distribution is $P(X = 1) = \theta$ and $P(X = 1) = 1 - \theta$, which can be expressed as

  $$f(x) = \theta^x (1-\theta)^{1-x}.$$

  The likelihood function is the joint PMF, which is

  $$\begin{aligned} L(\theta) =& \prod_{i=1}^n p(X_i) \\ =& \prod_{i=1}^n \theta^{X_i} (1-\theta)^{1-X_i} \\ =& \theta^{\sum_{i=1}^n X_i} (1-\theta)^{n - \sum_{i=1}^n X_i}. \end{aligned}$$

  Thus, $SS = \{\sum_{i=1}^n X_i\}$ or $SS = \{\bar{X}\}$. It is the MSS since its size is equal to the size of θ (from a theorem). The MLE is derived by maximizing $L(\theta)$. We want to maximize the logarithm of the likelihood function. Let

  $$\ell(\theta) = \log L(\theta) = \sum_{i=1}^n X_i \log \theta + (n - \sum_{i=1}^n X_i) \log(1-\theta) = n\bar{X} \log \theta + n(1-\bar{X}) \log(1-\theta).$$

  We want to solve

  $$\ell'(\theta) = \frac{n\bar{X}}{\theta} - \frac{n(1-\bar{X})}{1-\theta} = 0.$$

  for θ. The solution is

  $$\hat{\theta} = \bar{X}$$

  which is the MLE (maximum likelihood estimator) of θ.

- Let $X_1, \cdots, X_n$ be iid $Poisson(\theta)$.

  Solution: The PMF of the Poisson distribution if

  $$f(x) = \frac{\theta^x}{x!} e^{-\theta}.$$

The likelihood function is the joint PMF, which is

$$
\begin{aligned}
L(\theta) = &\prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!} e^{-\theta} \\
=&(\prod_{i=1}^{n} \frac{1}{X_i!})(\theta^{\sum_{i=1}^{n} X_i})(e^{-n\theta}) \\
=&(\prod_{i=1}^{n} \frac{1}{X_i!})(\theta^{n\bar{X}})(e^{-n\theta}).
\end{aligned}
$$

Thus, $SS = \{\sum_{i=1}^{n} X_i\}$ or $SS = \{\bar{X}\}$, which is also the MSS. We still study the log-likelihood function (i.e., the logarithm of the likelihood function), which is

$$\ell(\theta) = \log L(\theta) = -\log(\prod_{i=1}^{n} \frac{1}{X_i!}) + n\bar{X}\log\theta - n\theta.$$

By

$$\ell'(\theta) = \frac{n\bar{X}}{\theta} - n = 0$$

we obtain the MLE of θ as

$$\hat{\theta} = \bar{X}.$$

Remark: Supose we have applied $\hat{\theta} = \bar{X}$ to a real data and obtain $\bar{x} = 4.7$. Then, the maximum likelihood estimate of θ is 4.7, which means that the maximum likelihood estimate of the PMF is

$$f(x) = \frac{4.7^x}{x!} e^{-4.7}.$$

Then, we can study the probability problems. For instance, if we want to compute $P(X \leq 1)$, then we have

$$P(X \leq 1) = e^{-4.7} + 4.7e^{-4.7} = 0.05184,$$

which is the maximum likelihood estimate of the probability. The point is the solution $\hat{\theta} = \bar{X}$ is an answer for θ. It is used to replace θ in the computation of probabilities once we have the data. I am going to provide a remark for the normal model later.

- Example 6.2.3: Let $X_1, \cdots, X_n$ be iid $Exp(\theta)$.

  Solution: The PDF of $Exp(\theta)$ is

  $$f(\theta) = \theta e^{-\theta x}.$$

  The likelihood function is the joint PDF, which is

  $$
  \begin{aligned}
  L(\theta) = \log L(\theta) = &\prod_{i=1}^{n} \theta e^{-\theta X_i} \\
  =&\theta^n e^{-\theta \sum_{i=1}^{n} X_i} \\
  =&\theta^n e^{-\theta n \bar{X}_i}.
  \end{aligned}
  $$

  Thus, $SS = \{\sum_{i=1}^{n} X_i\}$ or $SS = \{\bar{X}\}$, which is the MSS. The log-likelihood function is

  $$\ell(\theta) = n\log\theta - \theta n\bar{X}.$$

  By

  $$\ell'(\theta) = \frac{n}{\theta} - n\bar{X},$$

  the MLE of θ is

  $$\hat{\theta} = \frac{1}{\bar{X}}.$$

- Example 6.2.4: Let $X_1, \cdots, X_n$ be iid from PMF $p_1 = P(X = 1) = \theta$, $p_2 = P(X_2) = \theta^2$ and $p_3 = P(X = 3) = 1 - \theta - \theta^2$. Find SS and MSS.

  Solution: This problem is a little bit tricky. We need to use one equation to express the PMF, which is

$$
\begin{aligned}
f(x) =& \theta^{I(x=1)} (\theta^2)^{I(x=2)} (1 - \theta - \theta^2)^{I(x=3)} \\
=& \theta^{I(x=1)} \theta^{2I(x=2)} (1 - \theta - \theta^2)^{I(x=3)} \\
=& \theta^{I(x=1)+2I(x=2)} (1 - \theta - \theta^2)^{I(x=3)},
\end{aligned}
$$

  where $I(\cdot)$ is the indicator function. The likelihood function is

$$
\begin{aligned}
L(\theta) =& \prod_{i=1}^{n} f(X_i) \\
=& \prod_{i=1}^{n} \{\theta^{I(X_i=1)+2I(X_i=2)} (1 - \theta - \theta^2)^{I(X_i=3)} \\
=& \theta^{\sum_{i=1}^{n} \{I(X_i=1)+2I(X_i=2)} (1 - \theta - \theta^2) \sum_{i=1}^{n} I(X_i) \\
=& \theta^{n_1+2n_2} (1 - \theta - \theta^2)^{n_3} \\
=& \theta^{n_1+2n_2} (1 - \theta - \theta^2)^{n-n_1-n_2}
\end{aligned}
$$

  where $n_1 = \sum_{i=1}^{n} I(X_i = 1)$ is the total number of ones, $n_2 = \sum_{i=1}^{n} I(X_i = 2)$ is the total number of twos, $n_3 = \sum_{i=1}^{n} I(X_i = 3)$ is the total number of threes. Then, you may have $SS = \{n_1, n_2, n_3\}$ or $SS = \{n_1 + 2n_2, n_3\}$, where the MSS is the second choice since its size is lower. We need to show that the size of MSS cannot be one, which is hard.

  Remark: For the maximum likelihood estimator, we need

$$
\ell(\theta) = \log L(\theta) = (n_1 + 2n_2) \log \theta + n_3 \log(1 - \theta - \theta^2).
$$

  Its derivative is

$$
\ell'(\theta) = \frac{n_1 + 2n_2}{\theta} - \frac{n_3(1 + 2\theta)}{1 - \theta - \theta^2}.
$$

  Let $\ell'(\theta) = 0$. We have

$$
\begin{aligned}
& \frac{n_1 + 2n_2}{\theta} = \frac{n_3(1 + 2\theta)}{1 - \theta - \theta^2} \\
\Rightarrow& (n_1 + 2n_2)(1 - \theta - \theta^2) = n_3(1 + 2\theta)\theta \\
\Rightarrow& (n_1 + 2n_2 + 2n_3)\theta^2 + (n_1 + 2n_2 + n_3)\theta - (n_1 + 2n_2) = 0 \\
\Rightarrow& \theta = \frac{-(n_1 + 2n_2 + n_3) \pm \sqrt{(n_1 + 2n_2 + n_3) + 4(n_1 + 2n_2)(n_1 + 2n_2 + 2n_3)}}{2(n_1 + 2n_2 + 2n_3)}.
\end{aligned}
$$

  Note that we need $0 < \theta < 1$. The solution can only be

$$
\hat{\theta} = \frac{-(n_1 + 2n_2 + n_3) + \sqrt{(n_1 + 2n_2 + n_3) + 4(n_1 + 2n_2)(n_1 + 2n_2 + 2n_3)}}{2(n_1 + 2n_2 + 2n_3)},
$$

  which could still be a concern. We need a discussion. Thus, this problem is not easy.

- Example 6.2.5: Let $X_1, \cdots, X_n$ be iid $Uniform(\theta)$.

  Solution: The PDF is

$$
f(x) = \frac{1}{\theta} I(0 \le x \le \theta) = \begin{cases} 1/\theta, & 0 \le x \le \theta, \\ 0, & \text{otherwise.} \end{cases}
$$

The likelihood function is

$$\begin{aligned}
L(\theta) =& \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \le X_i \le \theta) \\
=& \frac{1}{\theta^n} I(0 \le \min(X_i) \le \max(X_i) \le \theta) \\
=& \frac{1}{\theta^n} I(0 \le X_{(1)} \le X_{(n)} \le \theta) \\
=& \frac{1}{\theta^n} I(0 \le X_{(1)}) I(X_{(n)} \le \theta).
\end{aligned}$$

where $X_{(1)} = \min(X_i)$ and $X_{(n)} = \max(X_i)$. Thus, $SS = \{X_{(n)}\} = \{\max(X_i)\}$.

Now, we look at the MLE. To make $L(\theta)$ large, we need to make θ small, but θ cannot be lower than $X_{(n)}$. Therefore,

$$\hat{\theta} = X_{(n)} = \max(X_i).$$

Remark: We need to compute the PDF of $X_{(n)}$. The conclusion is used very often. The idea is to compute the CDF of $X_{(n)}$ and then take the derivative. By the definition, the CDF of $X_{(n)}$ is given by $P(X_{(n)} \le x)$ for any $x \in [0, \theta]$. Then,

$$\begin{aligned}
P(X_{(n)} \le x) =& P(X_1 \le x, X_2 \le x, \cdots, X_n \le x) \\
=& P(X_1 \le x) P(X_2 \le x) \cdots P(X_n \le x) \\
=& [P(X_1 \le x)]^n \\
=& (\frac{x}{\theta})^n \\
=& \frac{x^n}{\theta^n}.
\end{aligned}$$

for any $x \in [0, \theta]$. Thus, the CDF of $X_{(n)}$ is $F(x) = x^n/\theta^n$ when $0 \le x \le \theta$. The PDF is $f(x) = F'(x) = nx^{n-1}/\theta^n$. Sometimes, we need to compute $\mathrm{E}(X_{(n)})$ and $V(X_{(n)})$. The expected value is

$$\mathrm{E}(X_{(n)}) = \int_0^\theta x f(x) dx = \int_0^\theta x \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{\theta^n} \times \frac{\theta^{n+1}}{n+1} = \frac{n\theta}{n+1}.$$

For the variance, we also need

$$\mathrm{E}(X_{(n)}^2) = \int_0^\theta x^2 f(x) dx = \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx = \frac{n}{\theta^n} \times \frac{\theta^{n+2}}{n+2} = \frac{n\theta^2}{n+2}.$$

Thus,

$$\mathrm{V}(X_{(n)}) = \mathrm{E}(X_{(n)}^2) - \mathrm{E}^2(X_{(n)}) = \frac{n\theta^2}{n+2} - (\frac{n\theta}{n+1})^2 = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

- Example 6.2.2: Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma_0^2)$ with known σ_0^2.

  Solution: Let $\theta = \mu$. The PDF if X_i is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(x-\mu)^2}.$$

The likelihood function is

$$
\begin{aligned}
L(\theta) =& \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(X_i-\mu)^2} \right\} \\
=& (2\pi)^{-\frac{n}{2}} \sigma_0^{-n} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\} \\
=& (2\pi)^{-\frac{n}{2}} \sigma_0^{-n} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (X_i^2 - 2X_i\mu + \mu^2) \right\} \\
=& (2\pi)^{-\frac{n}{2}} \sigma_0^{-n} \exp\left\{ -\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^{n} X_i^2 - 2\mu \sum_{i=1}^{n} X_i + n\mu^2 \right] \right\} \\
=& (2\pi)^{-\frac{n}{2}} \sigma_0^{-n} \exp\left\{ -\frac{1}{2\sigma_0^2} \left[\sum_{i=1}^{n} X_i^2 - 2\mu \sum_{i=1}^{n} X_i + n\mu^2 \right] \right\} \\
=& (2\pi)^{-\frac{n}{2}} \sigma_0^{-n} \exp\left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} X_i^2 \right\} \exp\left\{ \frac{n\mu\bar{X}}{\sigma_0^2} \right\} \left\{ -\frac{n\mu^2}{2\sigma_0^2} \right\}.
\end{aligned}
$$

Thus, $SS = \{\bar{X}\}$. For the MLE, we need

$$
\ell(\mu) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma_0^2) - \frac{1}{2\sigma_0^2} \left[\sum_{i=1}^{n} X_i^2 - 2\mu \sum_{i=1}^{n} X_i + n\mu^2 \right].
$$

Then,

$$
\ell'(\mu) = \frac{\partial \ell(\mu)}{\partial \mu} = -\frac{1}{2\sigma_0^2} (-2 \sum_{i=1}^{n} X_i + 2n\mu) = 0 \Rightarrow \hat{\mu} = \bar{X}.
$$

Thus, the MLE of μ is $\hat{\mu} = \bar{X}$.

- Example 6.2.6: Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$.

  Solution: Let $\theta = (\mu, \sigma^2)$. Simply repeat the above, we have

$$
L(\theta) = L(\mu, \sigma^2) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2 \right\} \exp\left\{ \frac{n\mu\bar{X}}{\sigma^2} \right\} \left\{ -\frac{n\mu^2}{2\sigma^2} \right\}.
$$

  Thus, $SS = \{\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\}$. This is not easy to use. We usually change $L(\theta)$ as

$$
L(\theta) = L(\mu, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^{n} (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 \right] \right\}.
$$

  To derive this, we need

$$
\begin{aligned}
\sum_{i=1}^{n} (X_i - \mu)^2 =& \sum_{i=1}^{n} (X_i - \bar{X} + \bar{X} - \mu)^2 \\
=& \sum_{i=1}^{n} \left\{ (X_i - \bar{X})^2 + 2(X_i - \bar{X})(\bar{X} - \mu) + (\bar{X} - \mu)^2 \right\} \\
=& \sum_{i=1}^{n} (X_i - \bar{X})^2 + 2 \sum_{i=1}^{n} (X_i - \bar{X})(\bar{X} - \mu) + n(\bar{X} - \mu)^2 \\
=& \sum_{i=1}^{n} (X_i - \bar{X})^2 + 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \bar{X}) + n(\bar{X} - \mu)^2 \\
=& \sum_{i=1}^{n} (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.
\end{aligned}
$$

Thus, we also have $SS = \{\bar{X}, \sum_{i=1}^{n}(X_i - \bar{X})^2\}$. The log-likelihood function is

$$\ell(\theta) = \ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\right].$$

Then,

$$\frac{\partial \ell(\theta)}{\partial \mu} = \frac{1}{2\sigma^2}[2n(\bar{X} - \mu)] \Rightarrow \hat{\mu} = \bar{X},$$

and

$$\frac{\partial \ell(\theta)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\right] \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Thus, the MLE of θ is

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

Remark: In many statistical software packages, the estimator of σ^2 is given by the

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

This is called the UMVUE (uniform minimum variance unbaised estimator). To understand this concept, you need to learn **STAT 517**. However, the book has used the formula already.

Remark: This is a typical statement in probability problems for normal distribution.

The time that it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. The article "Fast-Rise Brake Lamp as a Collision-Prevention Device" (*Ergonomics*, 1993, 391-395) suggests that reaction time for an in-traffic response to a brake signal from standard brake lights can be modeled with a normal distribution having mean value 1.25 seconds and standard deviation of 0.46 seconds. What is the probability that react time is higher than 2 seconds.

- The study has a data set.
- The maximum likelihood estimate of μ is $\hat{\mu} = 1.25$ seconds.
- The maximum likelihood estimate of σ is $\hat{\sigma} = 0.46$ seconds. Then, that of σ^2 is $\hat{\sigma}^2 = 0.46^2$.
- The next interest is the computation of probability. We have

  $$P(X \geq 2) = 1 - \Phi(\frac{2 - 1.25}{0.46}) = 1 - \Phi(1.63) = 0.0516,$$

  which is the maximum likelihood estimate of the probability (i.e., for an accident). Q: how to interpret it? A: two seconds rule in the drive license test.
- We can also compute

  $$P(X \geq 3) - 1 - \Phi(\frac{3 - 1.25}{0.46}) = 1 - \Phi(3.81) = 7.1 \times 10^{-5}.$$

  This is the estimate of the probability if you follow the three seconds rule.
- If you use one second rule, then

  $$P(X \geq 1) = 1 - \Phi(\frac{1 - 1.25}{0.46}) = 1 - \Phi(-0.54) = 0.71.$$

  The accident probability is 0.71.

3 Inferences Based on the MLE

Assume one considers an estimator $\tilde{\theta}$ for an unknown parameter $\theta \in \mathbb{R}$. Let $\hat{\theta}$ be the MLE of θ.

3.1 Standard Errors, Bias, and Consistency

- (MSE). The mean-squared error (MSE) of $\tilde{\theta}$ is $\mathrm{MSE}_\theta(\tilde{\theta}) = \mathrm{E}_\theta(\tilde{\theta} - \theta)^2$, which is a function of θ.
- There is $\mathrm{MSE}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta(\tilde{\theta}) + [\mathrm{E}_\theta(\tilde{\theta}) - \theta]^2$, where $\mathrm{E}_\theta(\tilde{\theta}) - \theta$ is called the bais.
- If $\mathrm{E}(\tilde{\theta}) = \theta$, then $\tilde{\theta}$ is an unbiased estimator. There is $\mathrm{MSE}_\theta(\tilde{\theta}) = \mathrm{V}_\theta(\tilde{\theta})$.
- The standard error of $\tilde{\theta}$ is the estimator of the variance of $\tilde{\theta}$, given by $\{\mathrm{Var}_{\tilde{\theta}}(\tilde{\theta})\}^{1/2}$. Comparing the standard deviation, which is $\{\mathrm{Var}_\theta(\tilde{\theta})\}^{1/2}$, the standard error is an estimator of the standard deviation.
- Assume there are two estimators $\tilde{\theta}_1$ and $\tilde{\theta}_2$. We say $\tilde{\theta}_1$ is not worse than $\tilde{\theta}_2$ if $\mathrm{MSE}_\theta(\tilde{\theta}_1) \le \mathrm{MSE}_\theta(\tilde{\theta}_2)$. We say $\tilde{\theta}_1$ is better than $\tilde{\theta}_2$ if $\mathrm{MSE}_\theta(\tilde{\theta}_1) \le \mathrm{MSE}_\theta(\tilde{\theta}_2)$ for all θ and $\mathrm{MSE}_\theta(\tilde{\theta}_1) < \mathrm{MSE}_\theta(\tilde{\theta}_2)$ for some θ.

Examples:

- Example 6.3.1. Let $X_1, \cdots, X_n \sim^{iid} N(\mu, \sigma_0^2)$, where σ_0^2 is known but μ is not. Find the MLE of θ as well as its bais, standard deviation, the standard error.

  Solution: We have already derived the MLE as $\hat{\mu} = \bar{X}$. Note that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{V}(\bar{X}) = \sigma_0^2/n$. We obtain the bias as

  $$\mathrm{Bias} = \mathrm{E}(\hat{\mu}) - \mu = \mu - \mu = 0.$$

  Thus, $\hat{\mu}$ is unbiased. The standard deviation is

  $$\sqrt{V(\hat{\mu})} = \frac{\sigma_0}{\sqrt{n}},$$

  which is also the standard error (since σ_0 is known).

- Examples 6.3.4 and 6.3.5. Let $X_1, \cdots, X_n \sim^{iid} N(\mu, \sigma^2)$, where both μ and σ^2 are known. Find the MLE of θ as well as its bias, standard deviation, and standard error. Find the MLE of σ^2 as well as its bias. If we observe $\bar{x} = 64.517$ and $s = 2.379$ with $n = 30$, what are those answers.

  Solution: The MLEs of μ and σ^2 are $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. For $\hat{\mu}$, we have $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{V}(\bar{X}) = \sigma^2/n$. Thus,

  $$\mathrm{Bias}(\bar{X}) = \mathrm{E}(\bar{X}) - \mu = 0$$

  and its standard deviation is

  $$\sqrt{\mathrm{V}(\bar{X})} = \frac{\sigma}{\sqrt{n}}.$$

  Since σ is unknown, we need to replace it by its estimator in the standard error computation. If we use $\hat{\sigma}^2$ as an estimator of σ^2, then the standard error is

  $$\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

  If we use S^2 as an estimator of σ^2, then the standard error is

  $$\frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Based on the data, we have $\hat{\sigma}^2 = 5.4710$ and $\hat{\sigma} = 2.339$. The two answers are $2.339/\sqrt{30} = 0.4270$ and $2.379/\sqrt{30} = 0.4343$, respectively.

For σ^2, we have

$$\begin{aligned}
\text{Bias}(\hat{\sigma}^2) =& \text{E}(\hat{\sigma}^2) - \sigma^2 = \text{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right] - \sigma^2 = \frac{1}{n}\text{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] - \sigma^2 \\
=& \frac{\sigma^2}{n}\text{E}\left[\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2\right] - \sigma^2 = \frac{\sigma^2}{n}\text{E}(\chi^2_{n-1}) - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 \\
=& \frac{\sigma^2}{n}.
\end{aligned}$$

Thus, it is not unbiased. Further, we have

$$\begin{aligned}
\text{V}(\hat{\sigma}^2) =& \text{V}\left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n^2}\text{V}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{\sigma^4}{n^2}\text{V}\left[\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{\sigma^4}{n^2}\text{V}(\chi^2_{n-1}) \\
=& \frac{2(n-1)\sigma^4}{n^2}.
\end{aligned}$$

The standard deviation of $\hat{\sigma}^2$ is

$$\sqrt{V(\hat{\sigma}^2)} = \frac{\sigma^2}{n}\sqrt{2(n-1)}.$$

To compute the standard error, we need to replace σ^2 by $\hat{\sigma}^2$ or s^2. The answers are $(2.339^2/30)\sqrt{58} = 1.3888$ and $(2.379/30)\sqrt{58} = 1.368$, respectively.

Remark: Here we use a conclusion that $\sum_{i=1}^n (X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$, which is also equivalent to $(1/\sigma^2)\sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$. You cannot show the conclusion in this course. We also use $\text{E}(\chi^2_{n-1}) = n-1$ and $\text{V}(\chi^2_{n-1}) = 2(n-1)$. This is given in my review.

Remark: S^2 is unbiased. It can be shown by

$$\begin{aligned}
\text{Bias}(S^2) =& \text{E}\left[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2\right] - \sigma^2 = \frac{1}{n-1}\text{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] - \sigma^2 \\
=& \frac{\sigma^2}{n-1}\text{E}(\chi^2_{n-1}) - \sigma^2 = 0.
\end{aligned}$$

- Examples 6.3.2 and 6.3.3. Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$. Find the MLE of θ as well as its bias, standard deviation, and standard error. If $n = 1000$ and $\sum_{i=1}^n x_i = 790$, then what are those answers.

  Solution: The MLE of θ is $\bar{X}$. Note that $\mu = \text{E}(X_i) = \theta$ and $\sigma^2 = \text{V}(X_i) = \theta(1-\theta)$. The bais is $\text{E}(\bar{X}) - \mu = 0$ and the standard deviation is $\text{V}^{1/2}(\hat{X}) = (\sigma^2/n)^{1/2} = [\theta(1-\theta)/n]^{1/2}$. By the data, we have $\hat{\theta} = 0.790$. The standard error is $\sqrt{\hat{\theta}(1-\hat{\theta})/n} = \sqrt{0.790(1-0.790)/1000} = 0.01288$.

- Example: Assume $X_1, \cdots, X_n \sim N(\theta, 1)$. Let $\tilde{\theta} = \sum_{i=1}^m X_i/m$, where $m < n$. Justify why $\tilde{\theta}$ is worse than $\hat{\theta} = \sum_{i=1}^n X_i/n$.

  Solution: Note that $\text{E}(\tilde{\theta}) = \theta$, $\text{V}(\tilde{\theta}) = \sigma^2/m = 1/m$, $\text{E}(\hat{\theta}) = \theta$, and $\text{V}(\hat{\theta}) = \sigma^2/n = 1/n$. We have

  $$\text{MSE}(\tilde{\theta}) = \text{Bias}^2(\tilde{\theta}) + \text{V}(\tilde{\theta}) = \text{V}(\tilde{\theta}) = \frac{1}{m}$$

  and

  $$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\tilde{\theta}) + \text{V}(\tilde{\theta}) = \text{V}(\tilde{\theta}) = \frac{1}{n}.$$

  Since $\text{MSE}(\hat{\theta}) < \text{MSE}(\tilde{\theta})$, $\hat{\theta}$ is better than $\tilde{\theta}$ (as estimators of θ).

Consistency of Estimators
An estimator $\tilde{\theta}$ of θ is consistent if $\tilde{\theta} \xrightarrow{P} \theta$ for every θ.

- Consistency is the minimum requirement of estimators: if an estimator is not consistent, then it cannot be used.
- Q: Why $\tilde{\theta}$ equal to a constant cannot be used.
- Show consistency of the MLE in the previous examples.

3.2 Confidence Intervals

Let $l = l(data)$ and $u = u(data)$ be two statistics, always satisfying $l < u$. We say $[l, u]$ is a γ-level confidence interval of γ-confidence interval for θ if $P_\theta(l \leq \theta \leq u) \geq \gamma$ for every θ. We refer to γ as the confidence level of the interval.

Understanding the confidence interval and the confidence level:

(a) Collecting data many times (n times);

(b) calculate the interval $[l, u]$;

(c) the proportion of $\theta \in [l, u]$ is approximately greater than or equal to $1 - \alpha$;

(d) it becomes exactly greater than or equal to $1 - \alpha$ if $n \to \infty$.

Examples:

- Example 6.3.6. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma_0^2)$, where σ_0^2 is known. We use

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1)$$

  to find the confidence interval. The result is

$$[\bar{x} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}],$$

  where $P(N(0, 1) > z_{\alpha/2}) = \alpha/2$ is the upper $\alpha/2$ quantile (inverse CDF) of $N(0, 1)$. For example, if $\alpha = 0.05$, we have

$$[\bar{x} - 1.96\frac{\sigma_0}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma_0}{\sqrt{n}}].$$

  To under the concept, we need to do a simulated experiment.

  - Assume $n = 30$ and $\sigma_0 = 1$.
  - Collect 30 data points and compute the confidence interval.
  - Check whether $\theta \in [l, u]$.
  - Repeat the entire procedure and look at the proportion for the correct confidence intevrals.
  - Plot it for θ.

- Example 6.3.7. Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$. Then, a $1 - \alpha$-level confidence interval for θ is

$$[\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}].$$

This is the Wald method, which uses properties of the $N(0,1)$ distribution. In particular, let $\tilde{\theta}$ be an estimator of θ and $\sigma_{\tilde{\theta}}$ be its standard deviation. Then, the $1-\alpha$ level confidence interval is formulated by

$$\tilde{\theta} \pm z_{\alpha/2}\sigma_{\tilde{\theta}}.$$

If $\sigma_{\tilde{\theta}}$ is unknown, then we can simply replace it by the standard error. In this problem, we use $\hat{\theta} = \bar{X}$ as the estimator of θ. Note that $\mathrm{V}(\bar{X}) = \theta(1-\theta)/n$. The $1-\alpha$ confidendence interval for θ has the form of

$$\bar{X} \pm z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}.$$

Since θ is unknown, we replace θ by $\bar{X}$ and obtain the formula.

- Use $\sum_{i=1}^{1000} x_i = 790$ with $n = 1000$ (Example 6.3.3), we can calculate the confidence interval.
- We also need to do a simulated experiment to understand it.

t-Confidence Intervals

Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Then,

- $\bar{X}$ is the MLE of μ;
- $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$ is the MLE of σ^2
- $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is the UMVUE (uniform minimum unbiased estimator, not to be taught) of σ^2;
- $\bar{X}$ and S^2 are independent;
- $\bar{X} \sim N(\mu, \sigma^2/n)$;
- $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$.

Therefore

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

which provides a $1-\alpha$-level t-confidence interval for μ as

$$[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}],$$

where $t_{\alpha/2,n-1}$ is the quantile values of t_{n-1} distribution. If $\gamma = 0.05$, then the 95%-confidence interval for μ is

$$[\bar{X} - t_{0.025,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{0.025,n-1}\frac{S}{\sqrt{n}}].$$

In Example 6.3.5, we have $n = 30$, $\bar{x} = 64.517$, and $s/\sqrt{30} = 0.43434$, then $t_{0.025,29} = 2.0452$, implying that

$$64.517 \pm 2.0452(0.043434) = [63.629, 65.405].$$

To understand the t-confidence interval, we also need a simulated example.

- Assume $n = 30$.
- Collect 30 data. Compute the t-confidence interval.
- Look at the proportion of the interval which contains the true value of μ.

Remark: The book uses lower probability instead of upper probability. It defines z_γ as $P(N(0,1) \leq z_\gamma) = \gamma$. The classical definition is z_α but not z_γ. Similar issues can also be found in other distributions.

3.3 Testing Hypotheses and P-values

Testing hypotheses is an important statistical problem. It concerns whether a statement is correct or not. It contains the following items.

- A statement.
- Null hypothesis H_0: the statement is true; and the alternative hypothesis H_1 (H_a, or H_A): the statement is false.
- A test statistic T.
- Rejection region C: if $T \in C$, then conclude H_1.
- Q: how to define T and C.

Type I error, Type II error, significance level, power function, and P-value: I decide to move Sections 8.2.1 and 8.2.2 here.

Since the decision can only be made based on data, one cannot guarantee that the decision is alway consistent with the truth. Therefore, we propose two types of errors based on the following table.

	Truth	
Conclusion	True	False
True	Correct	Type II error
False	Type I error	Correct

Type I error probability is

$$P(\text{Conclude } H_1 | H_0).$$

Type II error probability is

$$P(\text{Conclude } H_0 | H_1).$$

The significance level is

$$\alpha = \max\{\text{Type I error probabilities}\}.$$

The power function is

$$P(\text{Conclude } H_1).$$

The p-value is the largest α which can reject H_0. Therefore, if the p-value is larger than α, we conclude H_0; otherwise, we conclude H_1.

Examples:

- Examples 6.3.9 and 6.3.10. Let $X_1, \cdots, X_n \sim N(\mu, \sigma_0^2)$, where σ_0^2 is unknown. Assume we want to know whether $\mu = \mu_0$, where μ_0 is a preselected number. Then,
  - Statement: $\mu = \mu_0$.
  - Null hypothesis $H_0 : \mu = \mu_0$; alternative hypothesis: $H_1 : \mu \neq \mu_0$.
  - Test statistic: $\bar{X}$.
  - Rejection region (either $\bar{X}$ is too large or $\bar{X}$ is too small):

$$C = \{|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}| > a\} = \{\bar{X} > \mu_0 + \frac{a\sigma_0}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - \frac{a\sigma_0}{\sqrt{n}}\},$$

  where a is a value to be determined.

- Type I error probabilities:

$$\begin{aligned}P(\text{Conclude } H_1|H_0) =& P(\bar{X} > \mu_0 + \frac{a\sigma_0}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - \frac{a\sigma_0}{\sqrt{n}}|\mu = \mu_0)\\ =& P(\bar{X} > \mu_0 + \frac{a\sigma_0}{\sqrt{n}}|\mu = \mu_0) + P(\bar{X} < \mu_0 - \frac{a\sigma_0}{\sqrt{n}}|\mu = \mu_0)\\ =& [1 - \Phi(a)] + \Phi(-a)\\ =& 2\Phi(-a),\end{aligned}$$

  where $a > 0$.
- Type II error probabilities:

$$\begin{aligned}P(\text{Conclude } H_0|H_1) =& P(\mu_0 - \frac{a\sigma_0}{\sqrt{n}} \le \bar{X} \le \mu_0 + \frac{a\sigma_0}{\sqrt{n}}|\mu \ne \mu_0)\\ =& \Phi(a + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}) - \Phi(a - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}).\end{aligned}$$

- The significance level is $2\Phi(-a)$.
- The power function is

$$\begin{aligned}P(\text{Conclude } H_1) =& P(\bar{X} > \mu_0 + \frac{a\sigma_0}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - \frac{a\sigma_0}{\sqrt{n}})\\ =& P(\bar{X} > \mu_0 + \frac{a\sigma_0}{\sqrt{n}}) + P(\bar{X} < \mu_0 - \frac{a\sigma_0}{\sqrt{n}})\\ =& \left[1 - \Phi(a + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0})\right] + \Phi\left(a - \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}\right).\end{aligned}$$

- To understand the P-value, we need to change a such that we can just conclude H_1. Then, we have

$$a_0 = |\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}|.$$

  implying that the P-value is

$$p = 2\Phi\left(-|\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}}|\right).$$

- Using data of Example 6.3.10, we have $\sigma_0^2 = 4$ and $\mu = 26$. Suppose we want to know whether $\mu = 25$. Then, we choose $\mu_0 = 25$. From the data, we have $\bar{x} = 26.6808$ and $n = 10$. Therefore the P-value is

$$2\Phi\left(-|\frac{26.6808 - 25}{2\sqrt{10}}|\right) = 2\Phi(-2.6576) = 0.0078.$$

  If we choose $\alpha < 0.0078$, then we conclude H_0; otherwise, we conclude H_1. Therefore, 0.0078 is the largest significance value for us to conclude H_1.

- Example 6.3.11. Let $X_1, \cdots, X_n$ be an iid sample from $Bernoulli(\theta)$. Suppose we want to test $h_0 : \theta = \theta_0$. Let $T = \sum_{i=1}^n X_i$. Then, $T \sim Bin(n, \theta)$. Then, we reject H_0 if $T \le a$ or $T \ge b$ for some $a < b$. Therefore, the rejection region is

$$\{T \le a \text{ or } T \ge b\}.$$

  Then,

  - $H_0 : \theta = \theta_0$ against $H_1 : \theta \ne \theta_0$.

- Type I error probability

$$\begin{aligned}P(T \le a \text{ or } T \ge b|\theta = \theta_0) =& P(Bin(n,\theta_0) \le a) + [1 - P(Bin(n,\theta_0) \ge b)] \\ \approx& [1 - \Phi(\frac{b - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}})] + \Phi(\frac{a - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}).\end{aligned}$$

This is also the significance level since it has just one value.

- Type II error probability

$$\begin{aligned}P(a < T < b|\theta \ne \theta_0) =& P(a < Bin(n,\theta) < b) \\ \approx& \Phi(\frac{b - n\theta}{\sqrt{n\theta(1-\theta)}}) - \Phi(\frac{a - n\theta}{\sqrt{n\theta(1-\theta)}}),\end{aligned}$$

where $\theta \ne \theta_0$.

- We often choose a and b symmetric about θ_0. This is based on the approximation

$$\frac{T - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \sim^{approx} N(0,1).$$

Then, the p-value is about

$$2\Phi(-|\frac{T - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}|).$$

- Suppose we want to test $H_0 : \theta = 1/2$ with $n = 100$. If we observe $T = 54$, then the P-value is

$$2\Phi(-|\frac{54 - 50}{\sqrt{100(0.5)(0.5)}}|) = 2\Phi(-0.8) = 0.4238.$$

Consistency of a test: what is statistical significance practically significant?

We want both the Type I error probabilities and the Type II error probabilities small. Usually there is

$$\max\{\text{Type I error probabilities}\} + \max\{\text{Type II error probabilities}\} = 1.$$

Since the first term is controlled by α (the significance level), we cannot control the second term. Therefore, type II error probabilities are often considered at individual points. Consider the normal case, where at an individual $\mu = \mu_1 \ne \mu_0$ the Type II error probability is

$$\Phi(a + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma_0}) - \Phi(a - \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma_0}), a > 0.$$

If $\mu_1 > \mu_0$, then $\sqrt{n}(\mu_0 - \mu_1) \to -\infty$ implying that the about limit is 0 as $n \to 0$. If $\mu_1 < \mu_0$, then $\sqrt{n}(\mu_0 - \mu_1) \to \infty$ also implying that the above limit is 0. Therefore, the Type II probability goes to 0 at the individual level as $n \to \infty$.

Hypothesis assessment via confidence intervals. Theoretically, the confidence interval problem is equivalent to the (two-sided) testing problem. If want to test $H_0 : \theta = \theta_0$ at 0.05 significance level, then we can compute the 95% confidence interval for θ. We conclude $H_0 : \theta = \theta_0$ if and only if the confidence interval contains θ_0. We can use **Example 6.3.12** to understand such an issue.

t-Tests. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$. Then, $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2 \sim \sigma^2\chi^2_{n-1}$ independently. Then,

$$\frac{\sqrt{n}}{\sigma}(\bar{X} - \mu) \sim N(0,1)$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Then,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Therefore, the t-test at α significance level rejects $H_0 : \mu = \mu_0$ if

$$|\frac{\bar{X}-\mu_0}{s/\sqrt{n}}| \geq t_{\alpha/2}.$$

In Example 6.3.10, we obtain $n = 10$, $\bar{x} = 26.6808$ and $s = 2.2050$. For $H_0 : \mu = 25$ against $H_1 : \mu \neq 25$, we obtain the t-statistic value as

$$|t| = |\frac{26.6808 - 25}{2.2050/\sqrt{10}}| = 2.4105 > t_{0.025,9} = 2.2622.$$

Thus, we reject H_0.

One-Side Tests. We want to understand the concepts of type I error probabilities, type II error probabilities, power functions, significance levels, and p-values. We also evaluate consistency. If a test is not consistent, then it cannot be used.

Example 6.3.12: Normal distribution with known variances. Let $X_1, \cdots, X_n \sim^{iid} N(\mu, \sigma_0^2)$. Suppose we want to test

$$H_0 : \mu \leq \mu_0 \leftrightarrow \mu > \mu_0.$$

Then, the rejection region is

$$C = \{\bar{X} \geq a\}$$

for some $a > 0$. The type I error probability is

$$P(\bar{X} \geq a | \mu \leq \mu_0) = 1 - \Phi(\frac{a-\mu}{\sigma_0/\sqrt{n}}), \mu \leq \mu_0,$$

which is increasing in μ. The type II error probability is

$$P(\bar{X} < a | \mu \geq \mu_0) = \Phi(\frac{a-\mu}{\sigma_0/\sqrt{n}}), \mu > \mu_0,$$

which is decreasing in μ. The power function is

$$P(\bar{X} \geq a) = 1 - \Phi(\frac{a-\mu}{\sigma_0/\sqrt{n}}), \mu \in \mathbb{R}.$$

It is the type I error probability if $\mu \leq \mu_0$ (i.e., H_0 holds) or one minus the type II error probability if $\mu > \mu_0$ (i.e., H_1 holds). The significance level is

$$\alpha = \max_{\mu \leq \mu_0}[1 - \Phi(\frac{a-\mu}{\sigma_0/\sqrt{n}})] = [1 - \Phi(\frac{a-\mu_0}{\sigma_0/\sqrt{n}})].$$

If we want to control it by not over 0.05, then we need to select a such that

$$[1 - \Phi(\frac{a-\mu_0}{\sigma_0/\sqrt{n}})] = 0.05 \Rightarrow a = \mu_0 + 1.645\sigma_0/\sqrt{n}.$$

If the data set of Example 6.3.10 is used, then the p-value is

$$1 - \Phi(\frac{26.6808 - 25}{2/\sqrt{10}}) = 1 - \Phi(2.6576) = 0.0039.$$

Example: Binomial or Bernoulli distribution. Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$. Then $T = \sum_{i=1}^n X_i \sim Bin(n, \theta)$. Suppose we want to test $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$. Then, we reject H_0 if $T \ge a$. Thus, the rejection region should be

$$C = \{T \ge a\}.$$

The type I error probability is

$$P(T \ge a | \theta \le \theta_0) = P(Bin(n, \theta) \ge a) = 1 - P(Bin(n, \theta) \le a - 1), \theta \le \theta_0.$$

The type II error probability is

$$P(T < a | \theta > \theta_0) = P(Bin(n, \theta) \le a - 1), \theta \ge \theta_0.$$

The power function is

$$P(T \ge a) = 1 - P(Bin(n, \theta) \le a - 1), \theta \in (0, 1).$$

The significance level is

$$\alpha = \max_{\theta \le \theta_0} [1 - P(Bin(n, \theta) \le a - 1)] = 1 - P(Bin(n, \theta_0) \le a - 1).$$

If $\alpha = 0.05$ and $\theta_0 = 0.05$ are chosen, then we can find a by choosing the minimum a satisfying $\alpha \le 0.05$. We have the following table.

	10	15	20	25	30	40	50	100
a	9	12	15	18	20	26	32	59

If $n = 100$ and $T = 60$, the p-value is

$$P(Bin(100, 0.5) \ge 60) = 0.02844.$$

3.4 Inferences for the Variance

Let $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. Then,

$$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Then, we can find a and b such that

$$P(\chi^2_{n-1} \ge a) = P(\chi^2_{n-1} \le b) = \alpha/2.$$

This provides the $1 - \alpha$-level confidence interval as

$$[\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}].$$

For example 6.3.10, we have $s^2 = 4.8620$ and $n = 10$. Then, the 95% confidence interval for σ^2 is

$$[\frac{9(4.8620)}{19.023}, \frac{9(4.8620)}{2.700}] = [2.3002, 16.207].$$

3.5 Sample-Size Calculations: Confidence Intervals

Example 6.3.16. Note that the length of confidence interval is $2z_{\alpha/2}\sigma_0/\sqrt{n}$. We can obtain the minimum n such that the length is less than 2δ, a preselected value. Then, we have

$$n \geq \sigma_0^2(\frac{z_{\alpha/2}}{\delta})^2.$$

If $\sigma_0^2 = 10$, $\gamma = 1-\alpha = 0.95$, $\delta = 0.5$, then we want the 95% confidence interval not over 1, leading

$$n \geq 10(1.96/0.5)^2 = 153.6 \Rightarrow n = 154.$$

However, if σ^2 is unknown, then it becomes

$$n \geq s^2(\frac{t_{\alpha/2,n-1}}{\delta})^2,$$

which depends on the sample. Therefore, the method cannot be used. This is still a research problem today.

Example 6.3.17. Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$. Then $T = \sum_{i=1}^n \sim Bin(n,\theta)$. The approximate confidence interval is

$$\bar{x} \pm z_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}.$$

If the length not over 2δ, then we need

$$n \geq \bar{x}(1-\bar{x})(\frac{z_{\alpha/2}}{\delta})^2.$$

Note that $\bar{x}(1-\bar{x}) \leq 1/4$, then we can choose

$$n \geq \frac{1}{4}(\frac{z_{\alpha/2}}{\delta})^2.$$

This choice guarantees the length of the confidence interval not over 2δ. If $\gamma = 0.95$ and $\delta = 0.1$ (the length not over 0.2), then we need

$$n \geq \frac{1}{4}(\frac{1.96}{0.1})^2 = 96.04$$

and we choose $n = 97$. If $\delta = 0.01$ (the length not over 0.02), then

$$n \geq \frac{1}{4}(\frac{1.96}{0.01})^2 = 9604.$$

3.6 Sample-Size Calculations: Power

We can only control type II error probabilities at the individual level. Note that the type II error probability equals one minus the power function. We need to make the power function as large as possible. This is called the *more powerful* way.

Example 6.3.18. Let $X_1, \cdots, X_n \sim N(\mu, \sigma_0^2)$ with a known σ_0^2. If we test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at α significance level, then the rule is that we reject H_0 if

$$\bar{X} \leq \mu_0 - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \text{ or } \bar{X} \geq \mu_0 + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}.$$

Then, the acceptence is

$$\mu_0 - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}.$$

The type II error probability at $\mu \neq \mu_0$

$$
\begin{aligned}
P(\text{Accept } H_0|\mu) =& P(\mu_0 - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}|\mu) \\
=& \Phi(\frac{\mu_0 + z_{\frac{\alpha}{2}}\sigma_0/\sqrt{n} - \mu}{\sigma_0/\sqrt{n}}) - \Phi(\frac{\mu_0 - z_{\frac{\alpha}{2}}\sigma_0/\sqrt{n} - \mu}{\sigma_0/\sqrt{n}}) \\
=& \Phi(\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} + z_{\frac{\alpha}{2}}) - \Phi(\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} - z_{\frac{\alpha}{2}}) \\
=& \Phi(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} + z_{\frac{\alpha}{2}}) - \Phi(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} - z_{\frac{\alpha}{2}}).
\end{aligned}
$$

Rather than the method introduce by the book, I decide to introduce another method.

For any $\mu \neq \mu_0$, if $\mu < \mu_0$, then

$$\lim_{n\to\infty} \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} = \infty,$$

implying that

$$\lim_{n\to\infty} P(\text{Accept } H_0|\mu) = 1 - 1 = 0.$$

If $\mu > \mu_0$, then

$$\lim_{n\to\infty} \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} = -\infty,$$

$$\lim_{n\to\infty} P(\text{Accept } H_0|\mu) = 0 - 0 = 0.$$

If $\mu > \mu_0$, to ensure

$$P(\text{Accept } H_0|\mu) \leq \beta_0$$

for a given β_0, it is enough to have

$$\Phi(\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} + z_{\frac{\alpha}{2}}) \leq \beta_0,$$

which means the first term is less than β_0. Then, we need

$$\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} + z_{\frac{\alpha}{2}} \leq -z_{\beta_0},$$

which provides

$$\frac{\mu - \mu_0}{\sigma_0/\sqrt{n}} \geq z_{\beta_0} + z_{\frac{\alpha}{2}} \Rightarrow n \geq \sigma_0^2 \left(\frac{z_{\beta_0} + z_{\alpha/2}}{\mu_0 - \mu}\right)^2.$$

We can similar derive the same conclusion in the case when $\mu < \mu_0$. This guarantees that $\beta(\mu) \leq \beta_0$ at μ.

Example. Let $X_1, \cdots, X_n \sim N(\mu, \sigma_0^2)$ with a known σ_0^2. Consider the test for

$$H_0 : \mu \leq \mu_0 \leftrightarrow \mu > \mu_0.$$

Assume we reject H_0 if $\bar{X} \geq \mu_0 + z_\alpha \sigma_0/\sqrt{n}$. For any $\mu > \mu_0$, the type II error probability is

$$
\begin{aligned}
\beta(\mu) =& P(\bar{X} \leq \mu_0 + z_\alpha \frac{\sigma_0}{\sqrt{n}}|\mu > \mu_0) \\
=& \Phi(\frac{\mu_0 + z_\alpha\sigma_0/\sqrt{n} - \mu}{\sigma_0/\sqrt{n}}) \\
=& \Phi(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}).
\end{aligned}
$$

If we want $\beta(\mu) \leq \beta_0$, then we need

$$\Phi(z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0}) \leq \beta_0 \Rightarrow z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma_0} \leq -z_{\beta_0} \Rightarrow n \geq \sigma_0^2 \left(\frac{z_{\beta_0} + z_\alpha}{\mu_0 - \mu}\right)^2.$$

For instance, assume $\mu_0 = 0$, $\sigma_0 = 2$, and $\mu = 0.1$. If $\alpha = 0.05$, then $z_\alpha = z_{0.05} = 1.645$. If we want $\beta(0.1) \le 0.01$, then $z_{\beta_0} = z_{0.01} = 2.33$. Thus,

$$n \ge \sigma_0^2 \left(\frac{1.645 + 2.33}{0.1} \right)^2 = 6320.25 \Rightarrow n = 6321.$$

Example 6.3.19. Binomial case, not analytically solvable, but we can numerically derive the result.
Example 6.3.20. If σ^2 is unknown, there is not a clear way to find n. This is a research problem.

4 Distribution-Free Methods

I am going to introduce Section 4.1 only.

4.1 Method of Moments

The moment estimation (method of moments) attempts to estimate θ using moment conditions. Look at a few examples for the comparison of the ME (Moment estimator) and the MLE.

- Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$.

  Solution: By $\mathrm{E}(X_i) = \mu$ and $\mathrm{E}(X_i^2) = \mu^2 + \sigma^2$. We have equations

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

  and

  $$\hat{\mu}_{ME}^2 + \hat{\sigma}_{ME}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2,$$

  which give the ME of μ as $\hat{\mu}_{ME} = \bar{X}$ and $\hat{\sigma}_{ME}^2 = \sum_{i=1}^n (X_i - \bar{X})^2/n$. They are identical to the MLEs.

- Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$.

  Solution: By $\mathrm{E}(X_i) = \theta$, we have $\hat{\theta}_{ME} = \bar{X}$, which is identical to the MLE.

- Let $X_1, \cdots, X_n$ be iid $Poisson(\theta)$.

  Solution: By $\mathrm{E}(X_i) = \theta$, we have $\hat{\theta}_{ME} = \bar{X}$, which is still the MLE.

- Let $X_1, \cdots, X_n$ be iid $Uniform(\theta)$.

  Solution: By $\mathrm{E}(X_i) = \theta/2$, we have $\hat{\theta}_{ME}/2 = \bar{X}$, leading to $\hat{\theta}_{ME} = 2\bar{X}$, which is different from the MLE.

- Let $X_1, \cdots, X_n$ be $Gamma(\alpha, \beta)$.

  Solution: By $\mathrm{E}(X_i) = \alpha/\beta$ and $\mathrm{V}(X_i) = \alpha/\beta^2$. We have equations

  $$\frac{\hat{\alpha}_{ME}}{\hat{\beta}_{ME}} = \bar{X}$$

  and

  $$\frac{\hat{\alpha}_{ME}}{\hat{\beta}_{ME}^2} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2,$$

  leading to

  $$\hat{\alpha}_{ME} = \frac{\bar{X}^2}{(1/n) \sum_{i=1}^n (X_i - \bar{X})^2}.$$

and

$$\hat{\beta}_{ME} = \frac{\bar{X}}{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Remark: Moment estimation is not as important as maximum likelihood estimation. You need to know the method, but you do not need to pay too much attention to it.

5 Asymptotic Properties

Note that we want to have results for confidence interval and hypotheses testing for the MLE. We need the variance of the estimator. In particular, let $\hat{\theta}$ be the MLE of θ. If we have $\mathrm{V}(\hat{\theta})$, then the $1-\alpha$ level confidence interval for θ is

$$\hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\hat{\mathrm{V}}(\hat{\theta})},$$

which is derived based on the Wald method. In theory, confidence interval and hypothesis testing problems are equivalent. Suppose you want to test

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0.$$

Then, we reject H_0 at α significance level if θ_0 is not inside the $1-\alpha$ level confidence interval. Thus, we should have a way to derive the variance of the MLE. This is given by the Fisher Information. In the following, I am going to provide the detail for the method.

Let $f_\theta(x)$ be the PDF/PMF. Then, the loglikelihood function is

$$\ell(\theta) = \sum_{i=1}^n \log f_\theta(X_i).$$

Then

$$\ell'(\theta) = \sum_{i=1}^n \frac{\partial \log f_\theta(X_i)}{\partial \theta} = \sum_{i=1}^n \frac{1}{f_\theta(X_i)} \frac{\partial f_\theta(X_i)}{\partial \theta}.$$

Let θ_0 be the true value. Then

$$\begin{aligned} \mathrm{E}[\ell'(\theta)] =& n\mathrm{E}\left[\frac{1}{f_\theta(X_i)} \frac{\partial f_\theta(X_i)}{\partial \theta}\right] \\ =& n\int_{-\infty}^{\infty} \frac{1}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta} f_{\theta_0}(x)dx. \end{aligned}$$

If $\theta = \theta_0$, then

$$\begin{aligned} \mathrm{E}[\ell'(\theta_0)] =& n\int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial \theta}|_{\theta=\theta_0} dx \\ =& n\frac{\partial}{\partial \theta}\left[\int_{-\infty}^{\infty} f_\theta(x)dx\right] |\theta = \theta_0 \\ =& 0. \end{aligned}$$

Then (detailed proofs are not included), $\hat{\theta}$ is the root of $\ell'(\theta)$ and θ_0, the true value of θ, is the root of $\mathrm{E}[\ell'(\theta)]$, indicating that $\hat{\theta} \xrightarrow{P} \theta_0$ (the exact proof is very hard).

Consider the Taylor expansion

$$\ell'(\hat{\theta}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0).$$

We obtain

$$\hat{\theta} - \theta_0 \approx -\ell'(\theta_0)/\ell''(\theta_0).$$

Note that

$$\frac{1}{n}\ell'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f_\theta(X_i)}{\partial\theta}|_{\theta=\theta_0}$$

and

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f_\theta(X_i)}{\partial\theta^2}|\theta=\theta_0$$

and average of iid samples. By SLLN, we have

$$\frac{1}{n}\ell'(\theta_0) \xrightarrow{P} \mathrm{E}\left[\frac{\partial \log f_\theta(X)}{\partial\theta}|_{\theta=\theta_0}\right] = 0$$

and

$$\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} \mathrm{E}\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}|_{\theta=\theta_0}\right].$$

By the CLT, we have

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) \xrightarrow{D} N(0,\tau^2),$$

where

$$\tau^2 = \mathrm{V}\left[\frac{\partial \log f_\theta(X)}{\partial\theta}|_{\theta=\theta_0}\right] = \mathrm{E}\left[\left(\frac{\partial \log f_\theta(X)}{\partial\theta}|_{\theta=\theta_0}\right)^2\right].$$

Then,

$$\mathrm{E}\left[\left(\frac{\partial \log f_\theta(X)}{\partial\theta}|_{\theta=\theta_0}\right)^2\right] = -\mathrm{E}\left[\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}|_{\theta=\theta_0}\right].$$

Using

$$\int_{-\infty}^{\infty}\frac{\partial \log f_\theta(x)}{\partial\theta}f_\theta(x)dx = 0,$$

we obtain

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\frac{\partial \log f_\theta(x)}{\partial\theta}f_\theta(x)dx = 0.$$

Then,

$$\int_{-\infty}^{\infty}\left[\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2}\right]f_\theta(x)dx + \int_{-\infty}^{\infty}\left[\frac{\partial \log f_\theta(x)}{\partial\theta}\right]^2 f_\theta(x)dx = 0.$$

We draw the conclusion. Therefore,

$$\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} \tau^2.$$

Thus,

$$\begin{aligned}\sqrt{n}(\hat\theta-\theta_0) &= \left(\frac{1}{n}\ell''(\theta_0)\right)^{-1}\left[\frac{1}{\sqrt{n}}\ell'(\theta_0)\right]\\ &\xrightarrow{D}\tau^{-2}N(0,\tau^2) = N(0,1/\tau^2).\end{aligned}$$

Therefore, we have the following

Theorem 1 *Let $X_1,\cdots,X_n$ be iid $f_\theta(x)$. Let*

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial \log f_\theta(X)}{\partial\theta}\right)^2\right] = -\mathrm{E}\left[\left(\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2}\right)\right]$$

be the Fisher Information. Then, the MLE $\hat\theta$ satisfies

$$\sqrt{n}(\hat\theta-\theta_0) \xrightarrow{D} N(0,I^{-1}(\theta_0)).$$

Theorem 2 *(The Delta Theorem). Let g be a smooth function. If θ is univariate, then*

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \xrightarrow{D} N(0, I^{-1}(\theta)[g'(\theta)]^2).$$

If θ is multivariate, then

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \xrightarrow{D} N(0, \nabla^\top g(\theta) I^{-1}(\theta) \nabla g(\theta)).$$

Proof: I just write the proof for the univariate case. Using the Taylor expansion, we have

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta) \Rightarrow g(\hat{\theta}) - g(\theta) \approx g'(\theta)(\hat{\theta} - \theta).$$

Then,

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \approx g'(\theta)\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, I^{-1}(\theta)[g'(\theta)]^2).$$

Similarly, we can provde the second conclusion. $\diamondsuit$

Based on the two theorems, we can provide the 95% confidence interval for θ as

$$\hat{\theta} \pm 1.96\sqrt{\frac{1}{nI(\hat{\theta})}}$$

or the 95% confidence interval for $g(\theta)$ as

$$g(\hat{\theta}) \pm 1.96 g'(\hat{\theta})\sqrt{\frac{1}{nI(\hat{\theta})}}.$$

We can also derive formulae for hypotheses testing.

Compute the Fisher Information and provide the asymptotic distribution of the MLE.

- (Examples 6.5.2 and 6.5.3). Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$.

  Solution: The PDF of $N(\mu, \sigma^2)$ is

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

  Its logarithm is

$$\log f_\theta(x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}.$$

  The first-order partial derivatives are

$$\begin{aligned}
\frac{\partial \log f_\theta(x)}{\partial \mu} &= \frac{x-\mu}{\sigma^2}, \\
\frac{\partial \log f_\theta(x)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}.
\end{aligned}$$

  The second-order partial derivatives are

$$\begin{aligned}
\frac{\partial^2 \log f_\theta(x)}{\partial \mu^2} &= -\frac{1}{\sigma^2}, \\
\frac{\partial^2 \log f_\theta(x)}{\partial (\sigma^2)^2} &= \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}, \\
\frac{\partial^2 \log f_\theta(x)}{\partial \mu \partial \sigma^2} &= -\frac{x-\mu}{\sigma^4},
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{E}\frac{\partial^2 \log f_\theta(x)}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\
\mathrm{E}\frac{\partial^2 \log f_\theta(x)}{\partial (\sigma^2)^2} &= -\frac{1}{2\sigma^4} \\
\mathrm{E}\frac{\partial^2 \log f_\theta(x)}{\partial \mu \partial \sigma^2} &= 0.
\end{aligned}$$

The Fisher information matrix is

$$I(\theta) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Using

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

we obtain the asymptotic distribution of the MLE as

$$\sqrt{n}\left[\begin{pmatrix} \bar{X} \\ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right] \xrightarrow{D} N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\right].$$

Next, we want to use the Delta Theorem to find the asymptotic distribution of $\eta = \mu/\sigma$. Clearly, there is

$$\hat{\eta} = \frac{\bar{X}}{[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2]^{1/2}}.$$

Let $g(z_1, z_2) = z_1/\sqrt{z_2}$. Then

$$g(\hat{\mu}, \hat{\sigma}^2) = \hat{\eta}$$

and

$$g(\mu, \sigma^2) = \eta.$$

Then,

$$\begin{aligned}
\frac{\partial g(z_1, z_2)}{\partial z_1} &= 1/\sqrt{z_2}, \\
\frac{\partial g(z_1, z_2)}{\partial z_2} &= -\frac{z_1}{2z_2^{3/2}}.
\end{aligned}$$

Thus,

$$\nabla g(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{pmatrix}.$$

We obtain

$$\nabla^\top g(\mu, \sigma^2) I^{-1}(\theta) \nabla^\top g(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma} & -\frac{\mu}{2\sigma^3} \end{pmatrix}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}\begin{pmatrix} \frac{1}{\sigma} \\ -\frac{\mu}{2\sigma^3} \end{pmatrix} = 1 + \frac{\mu^2}{2\sigma^2}.$$

Thus,

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N(0, 1 + \frac{\mu^2}{2\sigma^2}).$$

Then, the 95% confidence interval for μ/σ is

$$\begin{aligned}
&\frac{\hat{\mu}}{\hat{\sigma}} \pm 1.96\sqrt{\frac{1}{n}(1 + \frac{\hat{\mu}^2}{2\hat{\sigma}^2})} \\
=&\frac{\hat{X}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/n}} \pm 1.96\sqrt{\frac{1}{n}(1 + \frac{\bar{X}^2}{2\sum_{i=1}^n (X_i - \bar{X})^2/n})}.
\end{aligned}$$

In addition, we can compute the asymptotic distributions of $\sqrt{n}(\hat{\mu}^2 - \mu^2)$, $\sqrt{n}(\hat{\sigma} - \sigma)$, and many others.

- (Example 6.5.4). Let $X_1, \cdots, X_n$ be iid $Bernoulli(\theta)$.

  Solution: The PMF is
  $$f_\theta(x) = \theta^x (1-\theta)^{1-x}.$$
  Its logarithm is
  $$\log f_\theta(x) = x \log\theta + (1-x)\log(1-\theta).$$
  Its partial derivative is
  $$\frac{\partial \log f_\theta(x)}{\partial\theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}.$$
  The second-order partial derivative is
  $$\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}.$$
  The Fisher information is
  $$I(\theta) = -\mathrm{E}\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$
  Thus, the asymptotic distribution of the MLE is
  $$\sqrt{n}(\bar{X} - \theta) \xrightarrow{D} N(0, \theta(1-\theta)).$$
  Let $\eta = \log[\theta/(1-\theta)]$. Then, we can define $g(z) = \log[z/(1-z)]$. We obtain $g'(z) = 1/[z(1-z)]$. Therefore,
  $$\sqrt{n}(\log\frac{\bar{X}}{1-\bar{X}} - \log\frac{\theta}{1-\theta}) \xrightarrow{D} N(0, \frac{1}{\theta(1-\theta)}) = N(0, \frac{1}{\theta} + \frac{1}{1-\theta}).$$
  This is also a popular formula.

- (Example 6.5.5). Let $X_1, \cdots, X_n$ be iid $Poisson(\theta)$.

  Solution: The logarithm of the PMF is
  $$\log f_\theta(x) = -\log x! + x\log\theta - \theta.$$
  Its partial derivative is
  $$\frac{\partial \log f_\theta(x)}{\partial\theta} = \frac{x}{\theta} - 1$$
  Its second-order partial derivative is
  $$\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = -\frac{x}{\theta^2}.$$
  The Fisher information is
  $$I(\theta) = -\mathrm{E}\frac{\partial^2 \log f_\theta(X)}{\partial\theta^2} = \frac{1}{\theta}.$$
  Thus, the asymptotic distribution of the MLE is
  $$\sqrt{n}(\bar{X} - \theta) \xrightarrow{D} N(0, \theta).$$

- Let $X_1, \cdots, X_n$ be iid $Exp(\theta)$.

  Solution: The logarithm of the PDF is
$$\log f_\theta(x) = \log\theta - \theta x.$$
  Its first-order partial derivative is
$$\frac{\partial \log f_\theta(x)}{\partial\theta} = \frac{1}{\theta} - x.$$
  Its second-order partial derivative is
$$\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2} = -\frac{1}{\theta^2}.$$
  Thus, the Fisher information is
$$I(\theta) = \frac{1}{\theta^2}.$$
  The asymptotic distribution of the MLE is
$$\sqrt{n}(\bar{X}^{-1} - \theta) \xrightarrow{D} N(0, \theta^2).$$

- Let $X_1, \cdots, X_n$ be iid with common PDF $f(x) = (\theta+1)x^\theta$ for $x \in (0,1)$ and $\theta > -1$.

  Solution: The logarithm of the PDF is
$$\log f_\theta(x) = \log(1+\theta) + \theta\log x.$$
  Its first-order partial derivative is
$$\frac{\partial f_\theta(x)}{\partial\theta} = \frac{1}{1+\theta} + \log x.$$
  Its second-order partial derivative is
$$\frac{\partial^2 f_\theta(x)}{\partial\theta^2} = -\frac{1}{(1+\theta)^2}.$$
  Thus, the Fisher information is
$$I(\theta) = \frac{1}{(1+\theta)^2}.$$
  The asymptotic distribution of the MLE is
$$\sqrt{n}(-1 - \frac{n}{\sum_{i=1}^n \log X_i} - \theta) \xrightarrow{D} N(0, (1+\theta)^2).$$

- Q: how about $X_1, \cdots, X_n$ is iid $Uniform(\theta)$. This is an irregular case.

Remark:

- The method based on the Fisher information has been extended to not iid cases.
- It is very basic in all statistical inferences.
- it is important since it can provide testing and confidence interval methods.
- There is another way to defined the Fisher information. It is based on the the inverse of a matrix constructed from
$$-\frac{1}{n}\mathrm{E}\frac{\partial^2 \log f_\theta(\mathbf{x})}{\partial\theta_i\partial\theta_j} = -\frac{1}{n}\mathrm{E}\frac{\partial^2 \log \prod_{i=1}^n f_\theta(X_i)}{\partial\theta_i\partial\theta_j}.$$
  This kind of definitions can be easily extended to more general cases.