Chapter 6: Likelihood Inference

1 The Likelihood Function

Let \( \theta \) 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 \( L(\theta|x) \) and \( \ell(\theta|x) \), respectively, where \( x \) represents data.

(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.
- Parameters can only be estimated (learned, CS term) from data; comparing it with the term machine learning.
- Estimation, estimator, and estimate.
- Data are treated as random before they are available and as values after they are available.
- Statistics is to provide methods to address problems of distributions from 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 observations (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 \( x = 4 \).
- Example 6.1.3. Flip a coin until four heads are obtained. Let \( X \) be the total times of flipping with the observed value \( x = 9 \).
- Example 6.1.4. Let \( x = (X_1, \cdots, X_n) \), where \( X_i \sim iid N(\mu, \sigma^2_0) \), where \( \sigma^2_0 \) is known.

(b) Sufficient Statistics (SS):

- Definition: \( T = T(x) \), a function of data, is called sufficient statistic, if \( L[\theta|T(x_1)] / L[\theta|T(x_2)] \) does not depend on \( \theta \).
- Factorization theorem: If and only if \( L(\theta|x) = h(x)g_\theta(T) \), then \( T \) is a sufficient statistic.
- Minimal sufficient statistics (MSS): if the size (i.e. the dimension) of \( T \) is minimized.
- The dimension of MSS cannot be less than the dimension of \( \theta \). If the dimension of an SS is equal to the dimension of \( \theta \), then it is MSS.
- SS is an important concept in statistical inference. It is concluded that one only needs to known SS in statistical inference.
2 Maximum Likelihood Estimation (MLE)

The MLE, which attempts to maximize $L(\theta)$ to estimate $\theta$, 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 $\theta$, then $u(\hat{\theta})$ is the MLE of $u(\theta)$ for any function $u(\cdot)$.

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

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

Some properties:

- The MLE $\hat{\theta}$ is a function of data, which is random.
- The MLE of $g(\theta)$ is $g(\hat{\theta})$, which means it is transformation invariant.
- The choice of distributions is important in maximum likelihood estimation.

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

2.1 Computation of the MLE

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

$$\nabla \ell(\theta) = \left( \frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_k} \right)^\top = 0.$$

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 $X_1, \ldots, X_n$ be iid $Bernoulli(\theta)$.
- Let $X_1, \ldots, X_n$ be iid $Poisson(\theta)$.
- Example 6.2.3: Let $X_1, \ldots, X_n$ be iid $Exp(\theta)$.
- Example 6.2.4: Let $X_1, \ldots, X_n$ be iid from PMF $p_1 = P(X = 1) = \theta$, $p_2 = P(X = 2) = \theta^2$ and $p_3 = P(X = 3) = 1 - \theta - \theta^2$.
- Example 6.2.5: Let $X_1, \ldots, X_n$ be iid $Uniform(\theta)$.
- Example 6.2.2: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2_0)$ with known $\sigma^2_0$.
- Example 6.2.6: Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$.

3 Inferences Based on the MLE

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

3.1 Standard Errors, Bias, and Consistency

- (MSE). The mean-squared error (MSE) of \( \hat{\theta} \) is \( \text{MSE}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta)^2 \), which is a function of \( \theta \).

- There is \( \text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + [\text{E}(\hat{\theta}) - \theta]^2 \), where \( \text{E}(\hat{\theta}) - \theta \) is called the bias.

- If \( \text{E}(\hat{\theta}) = \theta \), then \( \hat{\theta} \) is an unbiased estimator. There is \( \text{MSE}(\hat{\theta}) = \text{V}(\hat{\theta}) \).

- The standard error of \( \hat{\theta} \) is the estimator of the variance of \( \hat{\theta} \), given by \( \{\text{Var}(\hat{\theta})\}^{1/2} \). Comparing the standard deviation, which is \( \{\text{Var}(\hat{\theta})\}^{1/2} \), the standard error is an estimator of the standard deviation.

- Assume there are two estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \). We say \( \hat{\theta}_1 \) is not worse than \( \hat{\theta}_2 \) if \( \text{MSE}(\hat{\theta}_1) \leq \text{MSE}(\hat{\theta}_2) \).

We say \( \hat{\theta}_1 \) is better than \( \hat{\theta}_2 \) if \( \text{MSE}(\hat{\theta}_1) \leq \text{MSE}(\hat{\theta}_2) \) for all \( \theta \) and \( \text{MSE}(\hat{\theta}_1) < \text{MSE}(\hat{\theta}_2) \) for some \( \theta \).

Examples:

- Example 6.3.1. Let \( X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma_0^2) \), where \( \sigma_0^2 \) is known but \( \mu \) is not. Find the MLE of \( \theta \) as well as its bias, standard deviation, and standard error.

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

- Examples 6.3.2 and 6.3.3. Let \( X_1, \ldots, X_n \) be iid \( \text{Bernoulli}(\theta) \). Find the MLE of \( \theta \) 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.

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

Consistency of Estimators

An estimator \( \hat{\theta} \) of \( \theta \) is consistent if \( \hat{\theta} \xrightarrow{P} \theta \) for every \( \theta \).

- Consistency is the minimum requirement of estimators: if an estimator is not consistent, then it cannot be used.

- Q: Why \( \hat{\theta} \) 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 \( \gamma \)-level confidence interval of \( \gamma \)-confidence interval for \( \theta \) if \( P_{\theta}(l \leq \theta \leq u) \geq \gamma \) for every \( \theta \). We refer to \( \gamma \) 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 \( \gamma \);
it becomes exactly greater than or equal to $\gamma$ if $n \to \infty$.

**Examples:**

- **Example 6.3.6.** Let $X_1, \cdots, X_n$ be iid $N(0, \sigma_0^2)$, where $\sigma_0^2$ is known. We use
  $$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$$
  to find the confidence interval. The result is
  $$[\bar{x} - z(1+\gamma)/2 \frac{\sigma_0}{\sqrt{n}}, \bar{x} + z(1+\gamma)/2 \frac{\sigma_0}{\sqrt{n}}],$$
  where $z(1+\gamma)/2$ is the quantile (inverse CDF) of $N(0,1)$. For example, if $\gamma = 0.95$, we have
  $$[\bar{x} - 1.96 \frac{\sigma_0}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma_0}{\sqrt{n}}].$$
  To understand 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 intervals.
  - Plot it for $\theta$.

- **Example 6.3.7.** Let $X_1, \cdots, X_n$ be iid $\text{Bernoulli}(\theta)$. Then, a $\gamma$-level confidence interval for $\theta$ is
  $$[\bar{X} - z(1+\gamma)/2 \sqrt{\frac{X(1-X)}{n}}, \bar{X} + z(1+\gamma)/2 \sqrt{\frac{X(1-X)}{n}}].$$
  - 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 $\mu$ and $\sigma^2$ are unknown. Then,

- $\bar{X}$ is the MLE of $\mu$;
- $\sigma^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / n$ is the MLE of $\sigma^2$
- $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n - 1)$ is the UMVUE (uniform minimum unbiased estimator, not to be taught) of $\sigma^2$;
- $\bar{X}$ 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 $\gamma$-level $t$-confidence interval for $\mu$ as

$$[\bar{X} - t_{(1+\gamma)/2, n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{(1+\gamma)/2, n-1} \frac{S}{\sqrt{n}}].$$
where $t_{\gamma/n-1}$ is the quantile values of $t_{n-1}$ distribution. If $\gamma = 0.05$, then the 95%-confidence interval for $\mu$ is

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

In Example 6.3.5, we have $n = 30$, $\bar{x} = 64.517$, and $s/\sqrt{30} = 0.43434$, then $t_{0.975,29} = 2.0452$, implying that

$$64.517 \pm 2.0452(0.43434) = [63.629, 65.405].$$

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 $\mu$.

### 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 always consistent with the truth. Therefore, we propose two types of errors based on the following table.

<table>
<thead>
<tr>
<th>Conclusion</th>
<th>Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
</tr>
<tr>
<td>False</td>
<td>Type I error</td>
</tr>
</tbody>
</table>

Type I error probabilities is

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

Type II error probabilities 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 $\alpha$ which can reject $H_0$. Therefore, if the $p$-value is larger than $\alpha$, we conclude $H_0$; otherwise, we conclude $H_1$.

**Examples:**
Examples 6.3.9 and 6.3.10. Let $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$, where $\sigma_0^2$ is unknown. Assume we want to know whether $\mu = \mu_0$, where $\mu_0$ 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):

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

where $a$ is a value to be determined.
- Type I error probabilities:

$$P(\text{Conclude } H_1 | H_0) = P(\bar{X} > \mu_0 + \frac{a \sigma_0}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - \frac{a \sigma_0}{\sqrt{n}} | \mu = \mu_0)$$

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

$$= \Phi(\frac{a}{\sqrt{n} (\mu_0 - \mu)}) + \Phi(-\frac{a}{\sqrt{n} (\mu_0 - \mu)}).$$

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

$$P(\text{Conclude } H_0 | H_1) = P(\mu_0 - \frac{a \sigma_0}{\sqrt{n}} \leq \bar{X} \leq \mu_0 + \frac{a \sigma_0}{\sqrt{n}} | \mu \neq \mu_0)$$

$$= \Phi(\frac{a}{\sqrt{n} (\mu_0 - \mu)}) - \Phi(\frac{a}{\sqrt{n} (\mu_0 - \mu)}).$$

- The significance level is $2 \Phi(\frac{a}{\sqrt{n}})$.
- The power function is

$$P(\text{Conclude } H_1) = P(\bar{X} > \mu_0 + \frac{a \sigma_0}{\sqrt{n}} \text{ or } \bar{X} < \mu_0 - \frac{a \sigma_0}{\sqrt{n}})$$

$$= P(\bar{X} > \mu_0 + \frac{a \sigma_0}{\sqrt{n}}) + P(\bar{X} < \mu_0 - \frac{a \sigma_0}{\sqrt{n}})$$

$$= \left[ 1 - \Phi(\frac{a}{\sqrt{n} (\mu_0 - \mu)} \right) + \Phi(\frac{a}{\sqrt{n} (\mu_0 - \mu)}).$$

- 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

$$2 \Phi \left( -\frac{26.6808 - 25}{2\sqrt{10}} \right) = 2 \Phi(-2.6576) = 0.0078.$$
Example 6.3.11. Let \( X_1, \ldots, X_n \) be an iid sample from \( \text{Bernoulli}(\theta) \). Suppose we want to test \( H_0 : \theta = \theta_0 \). Let \( T = \sum_{i=1}^{n} X_i \). Then, \( T \sim \text{Bin}(n, \theta) \). Then, we reject \( H_0 \) if \( T \leq a \) or \( T \geq b \) for some \( a < b \). Therefore, the rejection region is \( \{ T \leq a \text{ or } T \geq b \} \).

Then,
- \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \).
- Type I error probability
\[
P(T \leq a \text{ or } T \geq b | \theta = \theta_0) = P(\text{Bin}(n, \theta_0) \leq a) + [1 - P(\text{Bin}(n, \theta_0) \geq b)]
\]
\[\approx [1 - \Phi\left(\frac{b - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}\right)] + \Phi\left(\frac{a - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}\right).
\]
This is also the significance level since it has just one value.
- Type II error probability
\[
P(a < T < b | \theta \neq \theta_0) = P(a < \text{Bin}(n, \theta) < b)
\]
\[\approx \Phi\left(\frac{b - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}\right) - \Phi\left(\frac{a - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}\right),
\]
where \( \theta \neq \theta_0 \).
- We often choose \( a \) and \( b \) symmetric about \( \theta_0 \). This is based on the approximation
\[
\frac{T - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}} \sim N(0, 1).
\]
Then, the \( p \)-value is about
\[
2\Phi\left(-\frac{T - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}\right).
\]
- Suppose we want to test \( H_0 : \theta = 1/2 \) with \( n = 100 \). If we observe \( T = 54 \), then the \( P \)-value is
\[
2\Phi\left(-\frac{54 - 50}{\sqrt{100(0.5)(0.5)}}\right)2\Phi(-0.8) = 0.4238.
\]

**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 \( \alpha \) (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 \( \mu = \mu_1 \neq \mu_0 \) 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 we want to test \( H_0 : \theta = \theta_0 \) at 0.05 significance level, then
we can compute the 95% confidence interval for $\theta$. We conclude $H_0 : \theta = \theta_0$ if and only if the confidence interval contains $\theta_0$. 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 $(n-1)S^2 \sim \chi^2_{n-1}$ 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,

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

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

$$\frac{|\bar{X} - \mu_0|}{s/\sqrt{n}} \geq t_{1-\alpha/2}.$$ 

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

$$|t| = \left| \frac{26.6808 - 25}{2.2050/\sqrt{10}} \right| = 2.4105 > t_{0.975,9} = 2.2622.$$ 

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.

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

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

which is increasing in $\mu$. The type II error probability is

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

which is decreasing in $\mu$. The power function is

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

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

$$\alpha = \max_{\mu \leq \mu_0} [1 - \Phi\left( \frac{a - \mu}{\sigma_0/\sqrt{n}} \right)] = [1 - \Phi\left( \frac{a - \mu_0}{\sigma_0/\sqrt{n}} \right)].$$
If we want to control it by not over 0.05, then we need to select $a$ such that

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

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

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

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 $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. Then, we reject $H_0$ if $T \geq a$. Thus, the rejection region should be

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

The type I error probability is

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

The type II error probability is

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

The power function is

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

The significance level is

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

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

<table>
<thead>
<tr>
<th>$n$</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>26</td>
<td>32</td>
<td>59</td>
</tr>
</tbody>
</table>

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

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

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

$$P(\chi^2_{n-1} \leq a) = P(\chi^2_{n-1} \leq b) = (1 - \gamma)/2.$$ 

This provides the $\gamma$-level confidence interval as

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

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

$$\left[ \frac{9(4.8620)}{19.023}, \frac{9(4.8620)}{2.700} \right] = [2.3002, 16.207].$$
3.5 Sample-Size Calculations: Confidence Intervals

Example 6.3.16. Note that the length of confidence interval is $2 z_{(1+\gamma)/2} \sigma_0 / \sqrt{n}$. We can obtain the minimum $n$ such that the length is less than $2 \delta$, a preselected value. Then, we have

$$ n \geq \sigma_0^2 \left( \frac{z_{(1+\gamma)/2}}{\delta} \right)^2. $$

If $\sigma_0^2 = 10$, $\gamma = 0.96$, $\delta = 0.5$, then we want the 95% confidence interval not over 1, leading

$$ n \geq 10(1.96/0.5)^2 = 153.6 \Rightarrow n = 164. $$

However, if $\sigma^2$ is unknown, then it becomes

$$ n \geq s^2 \left( \frac{t_{(1+\gamma)/2, n-1}}{\delta} \right)^2, $$

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

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

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

If the length not over $2 \delta$, then we need

$$ n \geq \bar{x}(1-\bar{x}) \left( \frac{z_{1+\gamma/2}}{\delta} \right)^2. $$

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

$$ n \geq \frac{1}{4} \left( \frac{z_{1+\gamma/2}}{\delta} \right)^2. $$

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

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

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

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

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 $X_1, \ldots, X_n \sim N(\mu, \sigma_0^2)$ with a known $\sigma_0^2$. If we test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at $\alpha$ significance level, then the power function

$$ \beta(\mu) = 1 - \Phi\left( \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} + z_{1-\alpha/2} \right) + \Phi\left( \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} \right) $$

$$ = \Phi\left( -\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} \right) + \Phi\left( \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} \right) $$

Rather than the method introduce by the book, I decide to introduce another method. If $\mu > \mu_0$ (same for the case when $\mu < \mu_0$), then the first term is larger. There is

$$ \beta(\mu) \leq \Phi\left( -\frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} \right). $$
If a preselected $\beta$ is chosen and we solve $n$ by
\[
\beta_0 = \Phi\left(-\frac{\mu_0 - \mu}{\sigma_0 / \sqrt{n}} - z_{1-\alpha/2}\right) \Rightarrow n \geq \sigma_0^2 \left(\frac{z_{1-\beta_0} + z_{1-\alpha/2}}{\mu_0 - \mu}\right)^2.
\]
This guarantees that $\beta(\mu) \leq \beta_0$ at $\mu$.

Example 6.3.19. Binomial case, not analytically solvable, but we can numerically derive the result.

Example 6.3.20. If $\sigma^2$ 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 $\theta$ using moment conditions. Look at a few examples for the comparison of the ME (Moment estimator) and the MLE.

- Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$.
- Let $X_1, \ldots, X_n$ be iid $Bernoulli(\theta)$.
- Let $X_1, \ldots, X_n$ be iid $Uniform(\theta)$.
- Let $X_1, \ldots, X_n$ be $Gamma(\alpha, \beta)$.

5 Asymptotic Properties

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 $\theta_0$ be the true value. Then
\[
E[\ell'(\theta)] = nE\left[\frac{1}{f_\theta(X_i)} \frac{\partial f_\theta(X_i)}{\partial \theta}\right] = n \int_{-\infty}^{\infty} \frac{1}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial \theta} f_\theta_0(x) dx.
\]
If $\theta = \theta_0$, then
\[
E[\ell'(\theta_0)] = n \int_{-\infty}^{\infty} \frac{\partial f_\theta(x)}{\partial \theta} \bigg|_{\theta = \theta_0} dx = n \frac{\partial}{\partial \theta} \left[\int_{-\infty}^{\infty} f_\theta(x) dx \right]|_{\theta = \theta_0} = 0.
\]

Then (detailed proofs are not included), $\hat{\theta}$ maximizes $\ell'(\theta)$ and $\theta_0$, the true value of $\theta$, maximizes $E[\ell'(\theta)]$, indicating that $\hat{\theta} \xrightarrow{D} \theta_0$. 

11
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

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

and

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

By the CLT, we have

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

where

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

Then,

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

**Proof:** 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,

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

Therefore, we have the following
Theorem 1 Let $X_1, \cdots, X_n$ be iid $f_\theta(x)$. Let

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

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

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

- Let $X_1, \cdots, X_n$ be iid Bernoulli($\theta$).
- Let $X_1, \cdots, X_n$ be iid $N(0, \sigma^2)$.
- Let $X_1, \cdots, X_n$ be iid $Exp(\theta)$.
- Let $X_1, \cdots, X_n$ be iid Poisson($\theta$).
- Q: how about $X_1, \cdots, X_n$ is iid $Uniform(\theta)$. 

13