14.1 Consistency and asymptotic normality

We showed last lecture that given data $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$, the maximum likelihood estimator for $\lambda$ is simply $\hat{\lambda} = \bar{X}$. How accurate is $\hat{\lambda}$ for $\lambda$? Recall from Lecture 12 the following computations:

$$
E_\lambda[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \lambda,
$$

$$
\text{Var}_\lambda[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}[X_i] = \frac{\lambda}{n}.
$$

So $\hat{\lambda}$ is unbiased, with variance $\lambda/n$.

When $n$ is large, asymptotic theory provides us with a more complete picture of the “accuracy” of $\hat{\lambda}$: By the Law of Large Numbers, $\bar{X}$ converges to $\lambda$ in probability as $n \to \infty$. Furthermore, by the Central Limit Theorem,

$$
\sqrt{n}(\bar{X} - \lambda) \to \mathcal{N}(0, \text{Var}[X_i]) = \mathcal{N}(0, \lambda)
$$

in distribution as $n \to \infty$. So for large $n$, we expect $\hat{\lambda}$ to be close to $\lambda$, and the sampling distribution of $\hat{\lambda}$ is approximately $\mathcal{N}(\lambda, \frac{\lambda}{n})$. This normal approximation is useful for many reasons—for example, it allows us to understand other measures of error (such as $E[|\hat{\lambda} - \lambda|]$ or $\mathbb{P}(|\hat{\lambda} - \lambda| > 0.01])$, and (later in the course) will allow us to obtain a confidence interval for $\hat{\lambda}$.

In a parametric model, we say that an estimator $\hat{\theta}$ based on $X_1, \ldots, X_n$ is consistent if $\hat{\theta} \to \theta$ in probability as $n \to \infty$. We say that it is asymptotically normal if $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to a normal distribution (or a multivariate normal distribution, if $\theta$ has more than 1 parameter). So $\hat{\lambda}$ above is consistent and asymptotically normal.

The goal of this lecture is to explain why, rather than being a curiosity of this Poisson example, consistency and asymptotic normality of the MLE hold quite generally for many “typical” parametric models, and there is a general formula for its asymptotic variance. The following is one statement of such a result:

**Theorem 14.1.** Let $\{f(x|\theta) : \theta \in \Omega\}$ be a parametric model, where $\theta \in \mathbb{R}$ is a single parameter. Let $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta_0)$ for $\theta_0 \in \Omega$, and let $\hat{\theta}$ be the MLE based on $X_1, \ldots, X_n$. Suppose certain regularity conditions hold, including:1

---

1Some technical conditions in addition to the ones stated are required to make this theorem rigorously true; these additional conditions will hold for the examples we discuss, and we won’t worry about them in this class.
• All PDFs/PMFs $f(x|\theta)$ in the model have the same support,

• $\theta_0$ is an interior point (i.e., not on the boundary) of $\Omega$,

• The log-likelihood $l(\theta)$ is differentiable in $\theta$, and

• $\hat{\theta}$ is the unique value of $\theta \in \Omega$ that solves the equation $0 = l'(\theta)$.

Then $\hat{\theta}$ is consistent and asymptotically normal, with $\sqrt{n}(\hat{\theta} - \theta_0) \to N(0, \frac{1}{I(\theta_0)})$ in distribution. Here, $I(\theta)$ is defined by the two equivalent expressions

$$I(\theta) := \text{Var}_\theta[z(X, \theta)] = -\mathbb{E}_\theta[z'(X, \theta)],$$

where $\text{Var}_\theta$ and $\mathbb{E}_\theta$ denote variance and expectation with respect to $X \sim f(x|\theta)$, and

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x|\theta), \quad z'(x, \theta) = \frac{\partial^2}{\partial \theta^2} \log f(x|\theta).$$

$z(x, \theta)$ is called the **score function**, and $I(\theta)$ is called the **Fisher information**. Heuristically for large $n$, the above theorem tells us the following about the MLE $\hat{\theta}$:

• $\hat{\theta}$ is *asymptotically unbiased*. More precisely, the bias of $\hat{\theta}$ is less than order $1/\sqrt{n}$. (Otherwise $\sqrt{n}(\hat{\theta} - \theta_0)$ should not converge to a distribution with mean 0.)

• The variance of $\hat{\theta}$ is approximately $\frac{1}{nI(\theta_0)}$. In particular, the standard error is of order $1/\sqrt{n}$, and the variance (rather than the squared bias) is the main contributing factor to the mean-squared-error of $\hat{\theta}$.

• If the true parameter is $\theta_0$, the sampling distribution of $\hat{\theta}$ is approximately $N(\theta_0, \frac{1}{nI(\theta_0)})$.

**Example 14.2.** Let’s verify that this theorem is correct for the above Poisson example. There,

$$\log f(x|\lambda) = \log \frac{\lambda^x e^{-\lambda}}{x!} = x \log \lambda - \lambda - \log(x!),$$

so the score function and its derivative are given by

$$z(x, \lambda) = \frac{\partial}{\partial \lambda} \log f(x|\lambda) = \frac{x}{\lambda} - 1, \quad z'(x, \lambda) = \frac{\partial^2}{\partial \lambda^2} \log f(x|\lambda) = -\frac{x}{\lambda^2}.$$

We may compute the Fisher information as

$$I(\lambda) = -\mathbb{E}_\lambda[z'(X, \lambda)] = \mathbb{E}_\lambda \left[ \frac{X}{\lambda^2} \right] = \frac{1}{\lambda},$$

so $\sqrt{n}(\hat{\lambda} - \lambda) \to N(0, \lambda)$ in distribution. This is the same result as what we obtained using a direct application of the CLT.
14.2 Proof sketch

We’ll sketch heuristically the proof of Theorem 14.1, assuming \( f(x|\theta) \) is the PDF of a continuous distribution. (The discrete case is analogous with integrals replaced by sums.) To see why the MLE \( \hat{\theta} \) is consistent, note that \( \hat{\theta} \) is the value of \( \theta \) which maximizes

\[
\frac{1}{n} l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta).
\]

Suppose the true parameter is \( \theta_0 \), i.e. \( X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta_0) \). Then for any \( \theta \in \Omega \) (not necessarily \( \theta_0 \)), the Law of Large Numbers implies the convergence in probability

\[
\frac{1}{n} \sum_{i=1}^{n} \log f(X_i|\theta) \to E_{\theta_0}[\log f(X|\theta)]. \tag{14.1}
\]

Under suitable regularity conditions, this implies that the value of \( \theta \) maximizing the left side, which is \( \hat{\theta} \), converges in probability to the value of \( \theta \) maximizing the right side, which we claim is \( \theta_0 \). Indeed, for any \( \theta \in \Omega \),

\[
E_{\theta_0}[\log f(X|\theta)] - E_{\theta_0}[\log f(X|\theta_0)] = E_{\theta_0} \left[ \log \frac{f(X|\theta)}{f(X|\theta_0)} \right].
\]

Noting that \( x \mapsto \log x \) is concave, Jensen’s inequality implies \( E[\log X] \leq \log E[X] \) for any positive random variable \( X \), so

\[
E_{\theta_0} \left[ \log \frac{f(X|\theta)}{f(X|\theta_0)} \right] \leq \log E_{\theta_0} \left[ \frac{f(X|\theta)}{f(X|\theta_0)} \right] = \log \int \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0) dx = \log \int f(x|\theta) dx = 0.
\]

So \( \theta \mapsto E_{\theta_0}[\log f(X|\theta)] \) is maximized at \( \theta = \theta_0 \), which establishes consistency of \( \hat{\theta} \).

To show asymptotic normality, we first compute the mean and variance of the score:

Lemma 14.1 (Properties of the score). For \( \theta \in \Omega \),

\[
E_{\theta}[z(X, \theta)] = 0, \quad \text{Var}_{\theta}[z(X, \theta)] = -E[z'(X, \theta)].
\]

Proof. By the chain rule of differentiation,

\[
z(x, \theta)f(x|\theta) = \left( \frac{\partial}{\partial \theta} \log f(x|\theta) \right) f(x|\theta) = \frac{\partial}{\partial \theta} f(x|\theta) = \frac{\partial}{\partial \theta} f(x|\theta). \tag{14.2}
\]

Then, since \( \int f(x|\theta) dx = 1 \),

\[
E_{\theta}[z(X, \theta)] = \int z(x, \theta)f(x|\theta) dx = \int \frac{\partial}{\partial \theta} f(x|\theta) dx = \frac{\partial}{\partial \theta} \int f(x|\theta) dx = 0.
\]
Next, we differentiate this identity with respect to $\theta$:

\[
0 = \frac{\partial}{\partial \theta} \mathbb{E}_\theta[z(X, \theta)]
\]

\[
= \frac{\partial}{\partial \theta} \int z(x, \theta) f(x | \theta) dx
\]

\[
= \int \left( z'(x, \theta) f(x | \theta) + z(x, \theta) \left( \frac{\partial}{\partial \theta} f(x | \theta) \right) \right) dx
\]

\[
= \int \left( z'(x, \theta) f(x | \theta) + z(x, \theta)^2 f(x | \theta) \right) dx
\]

\[
= \mathbb{E}_\theta[z'(X, \theta)] + \mathbb{E}_\theta[z(X, \theta)^2]
\]

\[
= \mathbb{E}_\theta[z'(X, \theta)] + \text{Var}_\theta[z(X, \theta)],
\]

where the fourth line above applies (14.2) and the last line uses $\mathbb{E}_\theta[z(X, \theta)] = 0$.

Since $\hat{\theta}$ maximizes $l(\theta)$, we must have $0 = l'(\hat{\theta})$. Consistency of $\hat{\theta}$ ensures that (when $n$ is large) $\hat{\theta}$ is close to $\theta_0$ with high probability. This allows us to apply a first-order Taylor expansion to the equation $0 = l'(\hat{\theta})$ around $\hat{\theta} = \theta_0$:

\[
0 \approx l'(\theta_0) + (\hat{\theta} - \theta_0) l''(\theta_0),
\]

so

\[
\sqrt{n}(\hat{\theta} - \theta_0) \approx -\sqrt{n} \frac{l'(\theta_0)}{l''(\theta_0)} = - \frac{1}{\sqrt{n}} \frac{l'(\theta_0)}{l''(\theta_0)}. \tag{14.3}
\]

For the denominator, by the Law of Large Numbers,

\[
\frac{1}{n} l''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \left( \log f(X_i | \theta) \right)_{\theta=\theta_0} = \frac{1}{n} \sum_{i=1}^{n} z'(X_i, \theta_0) \to \mathbb{E}_{\theta_0}[z'(X, \theta_0)] = -I(\theta_0)
\]

in probability. For the numerator, recall by Lemma 14.1 that $z(X, \theta_0)$ has mean 0 and variance $I(\theta_0)$ when $X \sim f(x | \theta_0)$. Then by the Central Limit Theorem,

\[
\frac{1}{\sqrt{n}} l'(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \left[ \log f(X_i | \theta) \right]_{\theta=\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z(X_i, \theta_0) \to N(0, I(\theta_0))
\]

in distribution. Applying these conclusions, the Continuous Mapping Theorem, and Slutsky’s Lemma\(^2\) to (14.3),

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \to \frac{1}{I(\theta_0)} N(0, I(\theta_0)) = N(0, I(\theta_0)^{-1}),
\]

as desired.

\(^2\)Slutsky’s Lemma says: If $X_n \to c$ in probability and $Y_n \to Y$ in distribution, then $X_nY_n \to cY$ in distribution.