# Chapter 8.3. Maximum Likelihood Estimation

Prof. Tesler

Math 283
Fall 2016

# Estimating parameters

- Let $Y$ be a random variable with a distribution of known type but unknown parameter value $\theta$.
  - **Bernoulli or geometric** with unknown $p$.
  - **Poisson** with unknown mean $\mu$.
- Denote the pdf of $Y$ by
$$P_Y(y;\theta)$$
to emphasize that there is a parameter $\theta$.
- Do $n$ independent trials to get data $y_1, y_2, y_3, \ldots, y_n$.
The joint pdf is
$$P_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta) = P_Y(y_1;\theta)\cdots P_Y(y_n;\theta)$$
- **Goal:** Use the data to estimate $\theta$.

# Likelihood function

- Previously, we knew the parameter $\theta$ and regarded the $y$'s as unknowns (occurring with certain probabilities).
- Define the *likelihood* of $\theta$ given data $y_1, \ldots, y_n$ to be

$$L(\theta; y_1, \ldots, y_n) = P_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \theta) = P_Y(y_1; \theta) \cdots P_Y(y_n; \theta)$$

- It's the exact same formula as the joint pdf; the difference is the interpretation. Now we consider the data $y_1, \ldots, y_n$ to be given and $\theta$ to be an unknown.

## Definition (Maximum Likelihood Estimate, or MLE)

The value $\theta = \widehat{\theta}$ that maximizes the likelihood is the *Maximum Likelihood Estimate*. Often, it is found using Calculus by locating a critical point:

$$\frac{dL}{d\theta} = 0 \qquad \frac{d^2L}{d\theta^2} < 0$$

However, be sure to check for complications such as discontinuities and boundary values of $\theta$.

# MLE for the Poisson distribution

- $Y$ has a Poisson distribution with unknown parameter $\mu \geqslant 0$.
- Collect data from independent trials:

$$Y_1 = y_1,\ Y_2 = y_2,\ \cdots,\ Y_n = y_n$$

- **Likelihood:**

$$L(\mu; y_1, \ldots, y_n) = \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{y_i}}{y_i!} = \frac{e^{-n\mu} \mu^{y_1 + \cdots + y_n}}{y_1! \cdots y_n!}$$

- **Log likelihood** is maximized at the same $\mu$ and is easier to use:

$$\ln L(\mu; y_1, \ldots, y_n) = -n\mu + (y_1 + \cdots + y_n) \ln \mu - \ln(y_1! \cdots y_n!)$$

- **Critical point:** Solve $d(\ln L)/d\mu = 0$:

$$\frac{d(\ln L)}{d\mu} = -n + \frac{y_1 + \cdots + y_n}{\mu} = 0 \qquad \text{so} \qquad \mu = \frac{y_1 + \cdots + y_n}{n}$$

# MLE for the Poisson distribution

- **Log likelihood** is maximized at the same $\mu$ and is easier to use:
  $$\ln L(\mu; y_1, \ldots, y_n) = -n\mu + (y_1 + \cdots + y_n)\ln\mu - \ln(y_1! \cdots y_n!)$$
- **Critical point:** Solve $d(\ln L)/d\mu = 0$:
  $$\frac{d(\ln L)}{d\mu} = -n + \frac{y_1 + \cdots + y_n}{\mu} = 0 \qquad \text{so} \qquad \mu = \frac{y_1 + \cdots + y_n}{n}$$
- **Check second derivative is negative:**
  $$\frac{d^2(\ln L)}{d\mu^2} = -\frac{y_1 + \cdots + y_n}{\mu^2} = -\frac{n^2}{y_1 + \cdots + y_n} < 0$$
  provided $y_1 + \cdots + y_n > 0$. So it's a max unless $y_1 + \cdots + y_n = 0$.
- **Boundaries for range $\mu \geq 0$:** Must check $\mu \to 0^+$ and $\mu \to \infty$. Both send $\ln L \to -\infty$, so the $\mu$ identified above gives the max.

**The Maximum Likelihood Estimate for the Poisson distribution**

$$\hat{\mu} = \frac{y_1 + \cdots + y_n}{n} = \frac{0(\text{# of 0's}) + 1(\text{# of 1's}) + 2(\text{# of 2's}) + \cdots}{n}$$

# MLE for the Poisson distribution

- The exceptional case on the previous slide was $y_1 + \cdots + y_n = 0$, giving $y_1 = \cdots = y_n = 0$ (since all $y_i \geqslant 0$).
- In this case,

$$\begin{aligned}
\ln L(\mu; y_1, \ldots, y_n) &= -n\mu + (y_1 + \cdots + y_n) \ln \mu - \ln(y_1! \cdots y_n!) \\
&= -n\mu + 0 \ln \mu - \ln(0! \cdots 0!) \\
&= -n\mu
\end{aligned}$$

- On the range $\mu \geqslant 0$, this is maximized at $\hat{\mu} = 0$, which agrees with

$$\hat{\mu} = \frac{y_1 + \cdots + y_n}{n} = \frac{0 + \cdots + 0}{n} = 0$$

# Repeating the estimation gives different results

- $A$ does $n$ trials $y_{A1}, y_{A2}, \ldots, y_{An}$, leading to MLE $\widehat{\theta}_A$.
  $B$ does $n$ trials $y_{B1}, y_{B2}, \ldots, y_{Bn}$, leading to MLE $\widehat{\theta}_B$.
  $\ldots$
  How do $\widehat{\theta}_A, \widehat{\theta}_B, \ldots$ compare?
- Treat the $n$ trials in each experiment as random variables $Y_1, \ldots, Y_n$ and the MLE as a random variable $\widehat{\Theta}$.

### Estimate Poisson parameter with $n = 10$ trials (secret: $\mu = 1.23$)

| Experiment | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ | $Y_6$ | $Y_7$ | $Y_8$ | $Y_9$ | $Y_{10}$ | $\widehat{\Theta}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 | 3 | 0 | 2 | 2 | 0 | 2 | 1.0 |
| B | 1 | 2 | 0 | 1 | 1 | 3 | 0 | 0 | 0 | 1 | 0.9 |
| C | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1.5 |
| D | 1 | 2 | 1 | 2 | 1 | 4 | 2 | 3 | 2 | 1 | 1.9 |
| E | 0 | 3 | 0 | 1 | 1 | 0 | 0 | 1 | 2 | 2 | 1.0 |
| Mean | 1.2 | 1.8 | 0.6 | 1 | 1.4 | 1.6 | 1 | 1.6 | 1 | 1.4 | 1.26 |

## Desireable properties of an estimator $\widehat{\Theta}$

- $\widehat{\Theta}$ should be narrowly distributed around the correct value of $\theta$.
- Increasing $n$ should improve the estimate.
- The distribution of $\widehat{\Theta}$ should be known.

The MLE often does this.

# Bias

- Suppose $Y$ is Poisson with secret parameter $\mu$.
- Poisson MLE from data is
$$\hat{\mu} = \frac{Y_1 + \cdots + Y_n}{n}$$
- If many MLEs are computed from independent data sets, the average tends to
$$E(\hat{\mu}) = E\left(\frac{Y_1 + \cdots + Y_n}{n}\right) = \frac{E(Y_1) + \cdots + E(Y_n)}{n}$$
$$= \frac{\mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu$$
- Since $E(\hat{\mu}) = \mu$, we say $\hat{\mu}$ is an *unbiased estimator* of $\mu$.
- If the formula were different such that we had $E(\hat{\mu}) \neq \mu$, we would say $\hat{\mu}$ is a *biased estimator* of $\mu$.
E.g.: $\hat{\mu}' = 2Y_1$ has $E(\hat{\mu}') = 2\mu$, so it's biased (unless $\mu = 0$).

# Efficiency (want estimates to have small spread)

- Continue with Poisson MLE $\hat{\mu} = \frac{Y_1+\cdots+Y_n}{n}$ and secret mean $\mu$.
- The variance is

$$\operatorname{Var}(\hat{\mu}) = \operatorname{Var}\left(\frac{Y_1+\cdots+Y_n}{n}\right) = \frac{\operatorname{Var}(Y_1)+\cdots+\operatorname{Var}(Y_n)}{n^2}$$

$$= \frac{n\operatorname{Var}(Y_1)}{n^2} = \frac{\operatorname{Var}(Y_1)}{n} = \frac{\mu}{n}$$

  Increasing $n$ makes the variance smaller ($\hat{\mu}$ is more efficient).
- **Here's a second estimator:** Use $Y_1, Y_2$ and discard $Y_3, \ldots, Y_n$.

$$\hat{\mu}' = \frac{Y_1 + 2Y_2}{3}$$

$$E(\hat{\mu}') = \frac{\mu + 2\mu}{3} = \mu \qquad \text{so unbiased}$$

$$\operatorname{Var}(\hat{\mu}') = \frac{\operatorname{Var}(Y_1) + 4\operatorname{Var}(Y_2)}{9} = \frac{\mu + 4\mu}{9} = \frac{5\mu}{9}$$

  so it has higher variance (less efficient) than the MLE.