

# Maximum Likelihood Estimation

Julia Kempe & David S. Rosenberg

CDS, NYU

March 5, 2019

- 1 Likelihood of an Estimated Probability Distribution
- 2 Parametric Families of Distributions
- 3 Maximum Likelihood Estimation

## Likelihood of an Estimated Probability Distribution

# Estimating a Probability Distribution: Setting

- Let  $p(y)$  represent a probability distribution on  $\mathcal{Y}$ .
- $p(y)$  is **unknown** and we want to **estimate** it.
- Assume that  $p(y)$  is either a
  - probability density function on a continuous space  $\mathcal{Y}$ , or a
  - probability mass function on a discrete space  $\mathcal{Y}$ .
- Typical  $\mathcal{Y}$ 's:
  - $\mathcal{Y} = \mathbf{R}$ ;  $\mathcal{Y} = \mathbf{R}^d$  [typical continuous distributions]
  - $\mathcal{Y} = \{-1, 1\}$  [e.g. binary classification]
  - $\mathcal{Y} = \{0, 1, 2, \dots, K\}$  [e.g. multiclass problem]
  - $\mathcal{Y} = \{0, 1, 2, 3, 4 \dots\}$  [unbounded counts]

# Evaluating a Probability Distribution Estimate

- Before we talk about estimation, let's talk about evaluation.
- Somebody gives us an estimate of the probability distribution

$$\hat{p}(y).$$

- How can we evaluate how good it is?
- We want  $\hat{p}(y)$  to be descriptive of **future** data.

# Likelihood of a Predicted Distribution

- Suppose we have

$\mathcal{D} = (y_1, \dots, y_n)$  sampled i.i.d. from true distribution  $p(y)$ .

- Then the **likelihood** of  $\hat{p}$  for the data  $\mathcal{D}$  is defined to be

$$\hat{p}(\mathcal{D}) = \prod_{i=1}^n \hat{p}(y_i).$$

- If  $\hat{p}$  is a probability mass function, then likelihood is probability.

# Parametric Families of Distributions

# Parametric Models

## Definition

A **parametric model** is a set of probability distributions indexed by a parameter  $\theta \in \Theta$ . We denote this as

$$\{p(y; \theta) \mid \theta \in \Theta\},$$

where  $\theta$  is the **parameter** and  $\Theta$  is the **parameter space**.

- Below we'll give some examples of common parametric models.
  - But it's worth doing research to find a parametric model most appropriate for your data.
- We'll sometimes say **family of distributions** for a probability model.



# Poisson Family

- Support  $\mathcal{Y} = \{0, 1, 2, 3, \dots\}$ .
- Parameter space:  $\{\lambda \in \mathbf{R} \mid \lambda > 0\}$
- Probability mass function on  $k \in \mathcal{Y}$ :

$$p(k; \lambda) = \lambda^k e^{-\lambda} / (k!)$$

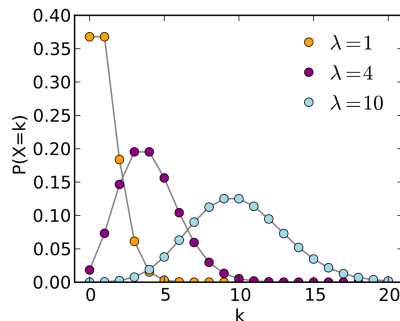


Figure is "Poisson pmf" by Skbkakas - Own work. Licensed under CC BY 3.0 via Wikimedia Commons - [http://commons.wikimedia.org/wiki/File:Poisson\\_pmf.svg#/media/File:Poisson\\_pmf.svg](http://commons.wikimedia.org/wiki/File:Poisson_pmf.svg#/media/File:Poisson_pmf.svg).

# Beta Family

- Support  $\mathcal{Y} = (0, 1)$ . [The unit interval.]
- Parameter space:  $\{\theta = (\alpha, \beta) \mid \alpha, \beta > 0\}$
- Probability density function on  $y \in \mathcal{Y}$ :

$$p(y; a, b) = \frac{y^{\alpha-1} (1-y)^{\beta-1}}{B(\alpha, \beta)}$$

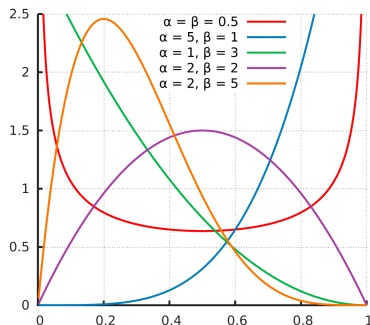
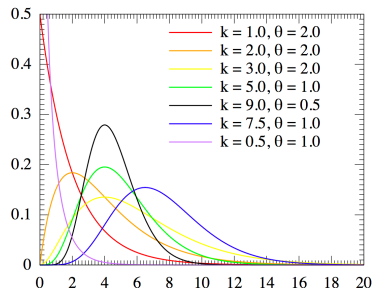


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via [Wikimedia Commons](#).

# Gamma Family

- Support  $\mathcal{Y} = (0, \infty)$ . [Positive real numbers]
- Parameter space:  $\{\theta = (k, \theta) \mid k > 0, \theta > 0\}$
- Probability density function on  $y \in \mathcal{Y}$ :

$$p(y; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-y/\theta}.$$



- Special cases: exponential distribution, chi-squared distribution, Erlang distribution

Figure from Wikipedia [https://commons.wikimedia.org/wiki/File:Gamma\\_distribution\\_pdf.svg](https://commons.wikimedia.org/wiki/File:Gamma_distribution_pdf.svg).

# Maximum Likelihood Estimation

# Likelihood in a Parametric Model

Suppose we have a parametric model  $\{p(y; \theta) \mid \theta \in \Theta\}$  and a sample  $\mathcal{D} = (y_1, \dots, y_n)$ .

- The **likelihood** of parameter estimate  $\hat{\theta} \in \Theta$  for sample  $\mathcal{D}$  is

$$p(\mathcal{D}; \hat{\theta}) = \prod_{i=1}^n p(y_i; \hat{\theta}).$$

- In practice, we prefer to work with the **log-likelihood**. Same maximizer, but

$$\log p(\mathcal{D}; \hat{\theta}) = \sum_{i=1}^n \log p(y_i; \hat{\theta}),$$

and sums are easier to work with than products.

# Maximum Likelihood Estimation

- Suppose  $\mathcal{D} = (y_1, \dots, y_n)$  is an i.i.d. sample from some distribution.

## Definition

A **maximum likelihood estimator (MLE)** for  $\theta$  in the model  $\{p(y; \theta) \mid \theta \in \Theta\}$  is

$$\begin{aligned}\hat{\theta} &\in \arg \max_{\theta \in \Theta} \log p(\mathcal{D}, \theta) \\ &= \arg \max_{\theta \in \Theta} \sum_{i=1}^n \log p(y_i; \theta).\end{aligned}$$

# Maximum Likelihood Estimation

- Finding the MLE is an **optimization problem**.
- For some model families, calculus gives a closed form for the MLE.
- Can also use numerical methods we know (e.g. SGD).

# MLE Existence

- In certain situations, the MLE may not exist.
- But there is usually a good reason for this.
- e.g. Gaussian family  $\{\mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbf{R}, \sigma^2 > 0\}$
- We have a single observation  $y$ .
- Is there an MLE?
- Taking  $\mu = y$  and  $\sigma^2 \rightarrow 0$  drives likelihood to infinity.
- MLE doesn't exist.



## Example: MLE for Poisson

- Observed counts  $\mathcal{D} = (k_1, \dots, k_n)$  for taxi cab pickups over  $n$  weeks.
  - $k_i$  is number of pickups at Penn Station Mon, 7-8pm, for week  $i$ .
- We want to fit a Poisson distribution to this data.
- The Poisson log-likelihood for a single count is

$$\begin{aligned}\log [p(k; \lambda)] &= \log \left[ \frac{\lambda^k e^{-\lambda}}{k!} \right] \\ &= k \log \lambda - \lambda - \log(k!)\end{aligned}$$

- The full log-likelihood is

$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^n [k_i \log \lambda - \lambda - \log(k_i!)] .$$

## Example: MLE for Poisson

- The full log-likelihood is

$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^n [k_i \log \lambda - \lambda - \log(k_i!)]$$

- First order condition gives

$$\begin{aligned} 0 = \frac{\partial}{\partial \lambda} [\log p(\mathcal{D}, \lambda)] &= \sum_{i=1}^n \left[ \frac{k_i}{\lambda} - 1 \right] \\ \implies \lambda &= \frac{1}{n} \sum_{i=1}^n k_i \end{aligned}$$

- So MLE  $\hat{\lambda}$  is just the mean of the counts.

## Test Set Log Likelihood for Penn Station, Mon-Fri 7-8pm

Method	Test Log-Likelihood
Poisson	-392.16
<b>Negative Binomial</b>	-188.67
Histogram (Bin width = 7)	$-\infty$
.95 Histogram + .05 NegBin	-203.89

# Estimating Distributions, Overfitting, and Hypothesis Spaces

- Just as in classification and regression, MLE can overfit!
- Example Probability Models:
  - $\mathcal{F} = \{\text{Poisson distributions}\}$ .
  - $\mathcal{F} = \{\text{Negative binomial distributions}\}$ .
  - $\mathcal{F} = \{\text{Histogram with 10 bins}\}$
  - $\mathcal{F} = \{\text{Histogram with bin for every } y \in \mathcal{Y}\}$  [will likely overfit for continuous data]
- How to judge which model works the best?
- Choose the model with the **highest likelihood on validation set**.