Statistical Theory and Modeling (ST2601)
Lecture 7 - Point estimation and Maximum likelihood

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Maximum likelihood**
- **Sampling distributions**
- **Bias-variance trade-off**
- **Consistency**
- **Sufficiency**

Probability vs Inference

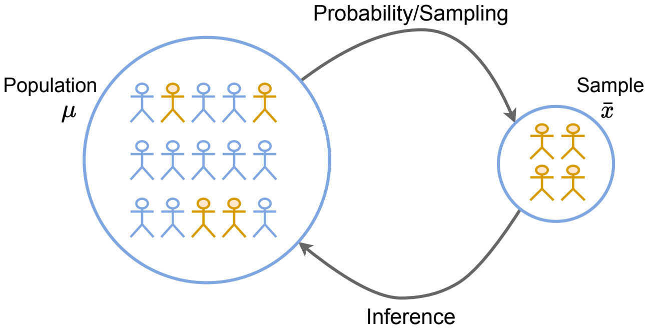
- **Probability theory**: **given a distribution with parameter θ** what are the properties of random variables (data)?
  - $X \sim \mathrm{Pois}(\lambda)$. Then: $\mathbb{E}(X) = \lambda$ and $\mathbb{V}(X) = \lambda$.
  - What is $\Pr(X > 4)$ for a given λ?
  - If $X_1, \ldots, X_n \sim \mathrm{Pois}(\lambda)$ for a given λ, what is $\mathbb{E}(\bar{X}_n)$?

- **Inference/Learning**: **given observed data $x_1, \ldots, x_n$**, which distribution and parameter value θ generated the data?
  - **Point estimation** $\hat{\lambda} = \bar{x}$
  - **Uncertainty quantification**:
    - standard errors $\mathbb{S}(\hat{\lambda})$
    - confidence intervals
    - Bayesian posterior distributions 😍

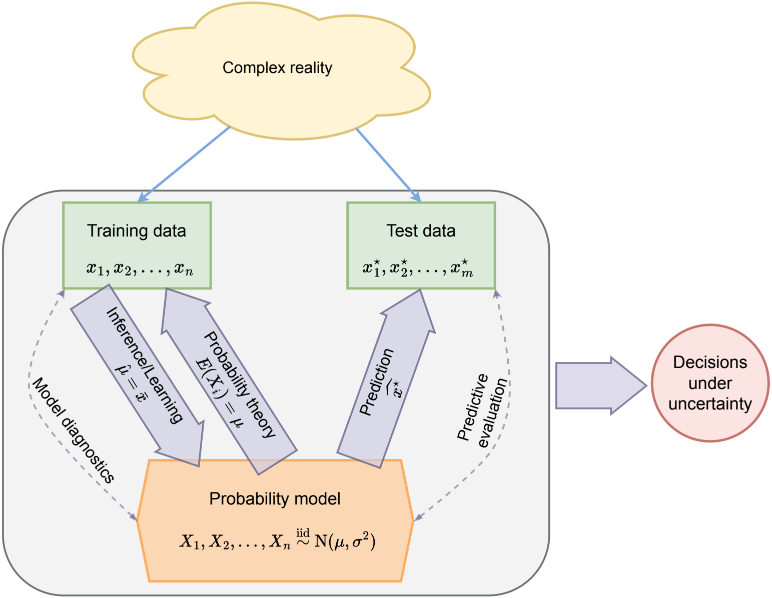
Probability vs Inference

- **Probability theory**: Models and Parameters $\Longrightarrow$ Data.
- **Inference**: Data $\Longrightarrow$ Models and Parameters $\rightsquigarrow$ Reality

- Often described as (particularly in finite populations):
- **Probability theory**: Population $\Longrightarrow$ Sample
- **Inference**: Sample $\Longrightarrow$ Population



The big picture of Statistics



The likelihood function

- **Probability distribution** for the dataset: $p(X_1, X_2, \ldots, X_n|\theta)$.
- **Probability for the observed data** $p(x_1, x_2, \ldots, x_n|\theta)$.
- **Inference**: given observed data $x_1, \ldots, x_n$, what is a "good" value for θ?
- Good values for $\theta \iff$ high probability for the observed data.
- Bad values for $\theta \iff$ low probability for the observed data.
- Find parameter value θ that **maximizes** the **likelihood function**

$$p(x_1, \ldots, x_n|\theta)$$

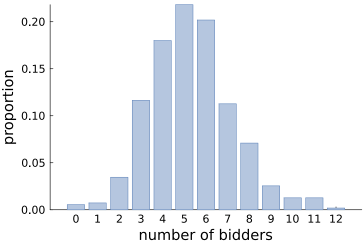
Different notations for the likelihood function

- $p(x_1, \ldots, x_n|\theta)$ [My Bayesian 😘 preference]
- $L(x_1, \ldots, x_n|\theta)$ [L instead of p is for Likelihood]
- $L(\theta)$ [Hiding the data. But convenient.]
- $L(x_1, \ldots, x_n; \theta)$ [Well, now we're just doing random symbols?]

Likelihood function - bit by bit

- eBay auction data with 1000 auctions for collectors' coins.
- We focus here on the **number of bidders** in the auctions.
- Use only the $n = 550$ auctions with smallest reservation prices.
- Count data: let's try a Poisson!

	BookVal	MinorBlem	MajorBlem	PowerSeller	IDSeller	Sealed	NegFeedback	ReservePriceFrac	NBidders	FinalPrice
1	18.95	0	0	0	0	0	0	0.368865435356201	2	15.5
2	43.5	0	0	1	0	0	0	0.229885057471264	6	41
3	24.5	0	0	1	0	0	0	1.02	1	24.99
4	34.5	1	0	0	0	0	0	0.721739130434783	1	24.9
5	99.5	0	0	0	0	0	1	0.167236180904523	4	72.65



Likelihood function for the first observation y_1

- First data point: $y_1 = 2$.
- Probability of observing $y_1 = 2$ in the Poisson model?
- Poisson probability function:

$$p(Y_1 = y_1|\lambda) = \frac{\lambda^{y_1} e^{-\lambda}}{y_1!} = \frac{\lambda^2 e^{-\lambda}}{2!}$$

- Let's try with $\lambda = 3$.
  - Mathematically:

$$p(Y_1 = 2|\lambda = 3) = \frac{3^2 e^{-3}}{2!} = 0.2240418$$

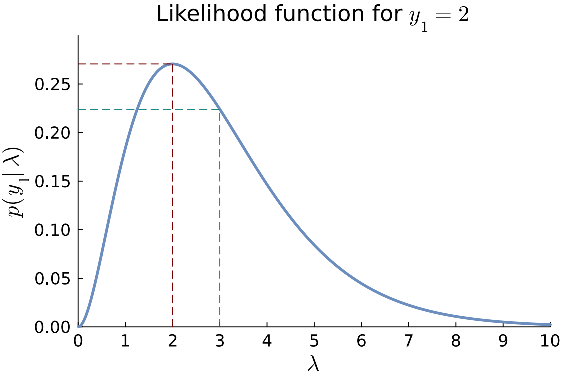
  - In R: `dpois(x = 2, lambda = 3)`
- For $\lambda = 2$:
  - Mathematically:

$$p(Y_1 = 2|\lambda = 2) = \frac{2^2 e^{-2}}{2!} = 0.2706706$$

  - In R: `dpois(x = 2, lambda = 2)`

Likelihood function for the first observation y_1

- So, $\lambda = 2$ gave a higher probability to the data $y_1 = 2$ compared to $\lambda = 3$.
- How about other λ values? Let's do them all!



Likelihood function for y_1 and y_2

- Data: $y_1 = 2$ and $y_2 = 6$.
- Likelihood function is the **joint probability**

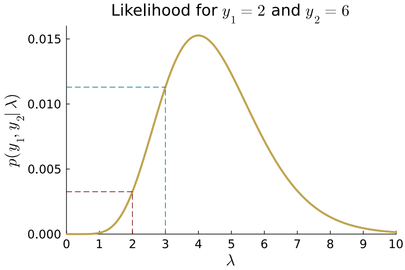
$$p(Y_1 = 2, Y_2 = 6|\lambda) \overset{\text{indep}}{=} p(Y_1 = 2|\lambda) \cdot p(Y_2 = 6|\lambda) = \frac{\lambda^{y_1} e^{-\lambda}}{y_1!} \cdot \frac{\lambda^{y_2} e^{-\lambda}}{y_2!}$$

- For $\lambda = 2$

$$p(Y_1 = 2, Y_2 = 6|\lambda = 2) = \frac{2^2 e^{-2}}{2!} \cdot \frac{2^6 e^{-2}}{6!}$$

- Let R do the work

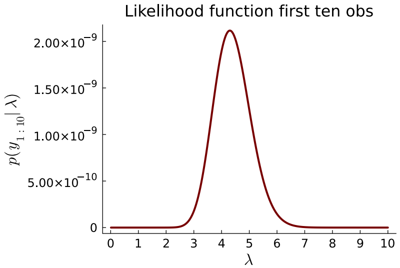
```
dpois(x = 2, lambda = 2)*dpois(x = 6, lambda = 2) = 0.003256114
```



Likelihood function for $y_1, \ldots, y_{10}$

- Likelihood function using first ten observations

$$p(Y_1 = y_1, \ldots, Y_{10} = y_{10}|\lambda) \overset{\text{indep}}{=} \prod_{i=1}^{10} p(y_i|\lambda)$$

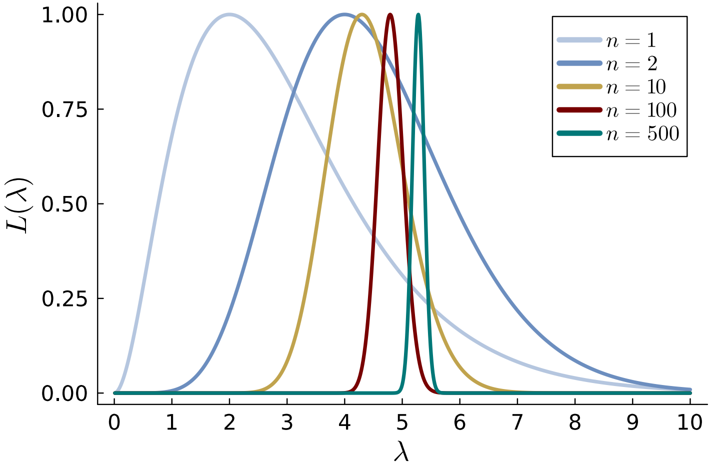


- **Likelihood function** for all $n = 550$ observations

$$p(Y_1 = y_1, \ldots, Y_n = y_n|\lambda) = \prod_{i=1}^{n} p(y_i|\lambda)$$

- Product of 1000 probabilities is a tiny number. Let's do logs.

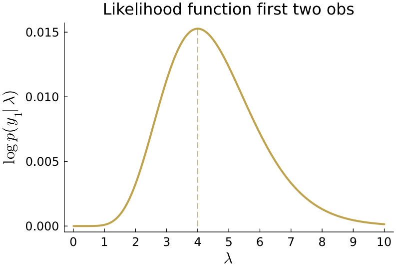
Likelihood concentrates with more data

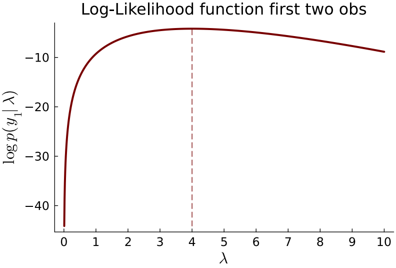


Log-likelihood function for two observations

- Log-Likelihood function using first two observations

$$\log p(Y_1 = 2, Y_2 = 6|\lambda) = \log p(Y_1 = 2|\lambda) + \log p(Y_2 = 6|\lambda)$$





- Since $\log(x)$ is monotonically increasing: the λ that maximizes the likelihood also maximizes the log-likelihood.
- **Logs simplifies the derivative** needed to find the maximum.
- **Maximum likelihood estimator** of λ: the value of λ that maximizes the (log-)likehood function.

Log-likelihood function for all observations

- **Log-likelihood** for all n data points

$$\ell(\lambda) = \log L(\lambda) = \sum_{i=1}^{n} \log p(y_i|\lambda)$$

- Poisson distribution

$$p(y_i|\lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} \quad \text{and} \quad \log p(y_i|\lambda) = y_i \log \lambda - \lambda - \log(y_i!)$$

- Log-likelihood for iid Poisson model

$$\ell(\lambda) = \sum_{i=1}^{n} \log p(y_i|\lambda) = \sum_{i=1}^{n} \left(y_i \log \lambda - \lambda - \log(y_i!)\right)$$

$$= \log \lambda \sum_{i=1}^{n} y_i - n\lambda - \sum_{i=1}^{n} \log(y_i!)$$

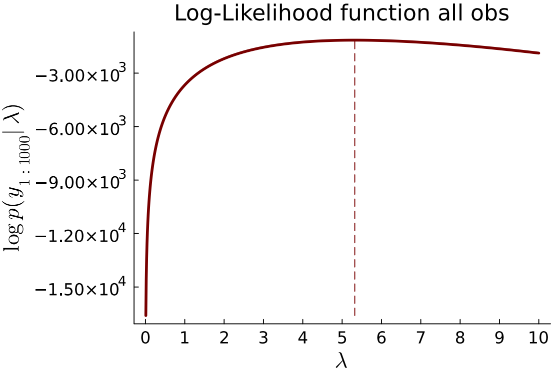
- Since $\sum_{i=1}^{n} y_i = n\bar{y}$ we can write

$$\ell(\lambda) = \log \lambda \cdot n\bar{y} - n\lambda - \sum_{i=1}^{n} \log(y_i!)$$

Log-likelihood function for all observations

- Log-likelihood for **iid Poisson model**

$$\ell(\lambda) = \log \lambda \cdot n\bar{y} - n\lambda - \sum_{i=1}^{n} \log(y_i!)$$



The MLE in the iid Poisson model

- **Maximum likelihood estimate** (**MLE**) of λ

$$\hat{\lambda}_{ML} = \underset{\lambda}{\operatorname{argmax}}\, \ell(\lambda)$$

- Finding a maximum of a function? Set first derivate to zero and solve for λ

$$\ell'(\lambda) = 0$$

- Check for (local) maximum by checking second derivative

$$\ell''(\hat{\lambda}_{ML}) < 0$$

- When $\ell'(\lambda) = 0$ cannot be solved mathematically. Use computer. More later!

The MLE in the iid Poisson model

- Log-likelihood

$$\ell(\lambda) = \log\lambda \cdot n\bar{y} - n\lambda - \sum_{i=1}^{n} \log(y_i!)$$

$$\ell'(\lambda) = \frac{n\bar{y}}{\lambda} - n = 0$$

has solution

$$\hat{\lambda}_{ML} = \bar{y}$$

- Second derivative shows that this indeed a (local) maximizer

$$\ell''(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\ell'(\lambda) = -\frac{n\bar{y}}{\lambda^2} < 0$$

for all λ and therefore also at $\hat{\lambda}_{ML}$.

- For a dataset with $\bar{y} = 0$ the second derivative test is inconclusive. Why?

The MLE in the iid Exponential model

- Model

$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Expon}(\beta)$$

- Likelihood (densities because of continuous random variables!)

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-y_i/\beta} = \frac{1}{\beta^n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} y_i} = \frac{1}{\beta^n} e^{-\frac{n\bar{y}}{\beta}}$$

- Log-likelihood

$$\ell(\beta) = \log L(\beta) = -n \log \beta - \frac{n\bar{y}}{\beta}$$

$$\ell'(\beta) = -\frac{n}{\beta} + \frac{n\bar{y}}{\beta^2} = 0$$

$$-n + \frac{n\bar{y}}{\beta} = 0$$

so

$$\hat{\beta}_{ML} = \bar{y}$$

The MLE in the iid Exponential model

- First derivative
$$\ell'(\beta) = -\frac{n}{\beta} + \frac{n\bar{y}}{\beta^2}$$
- Second derivative
$$\ell''(\beta) = \frac{n}{\beta^2} - \frac{2n\bar{y}}{\beta^3}$$
- Evaluate at $\hat{\beta}_{ML} = \bar{y}$
$$\ell''(\hat{\beta}_{ML}) = \frac{n}{\bar{y}^2} - \frac{2n\bar{y}}{\bar{y}^3} = \frac{n}{\bar{y}^2} - \frac{2n}{\bar{y}^2} = -\frac{n}{\bar{y}^2} < 0$$

since $n > 0$ and $\bar{y} > 0$ (exponential is used for positive data).

Sampling distribution of an estimator

- An estimator $\hat{\theta}$ depends on the sample

$$\hat{\theta}_n(X_1, \ldots, X_n)$$

- **Sampling distribution** of $\hat{\theta}$: how $\hat{\theta}$ **varies from sample to sample**.

- **Confidence intervals** are based on this.

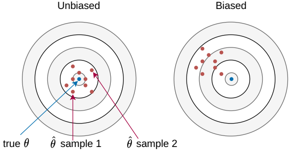
- **Asymptotic sampling distribution** for $\hat{\theta}_n$: the sampling distribution when n is large ($n \to \infty$).

- **Central limit theorem**: the asymptotic sampling distribution of the sample mean $\bar{X}_n$ is normal.

Bias-variance trade-off

Unbiased estimator

$$\mathbb{E}(\hat{\theta}) = \theta$$

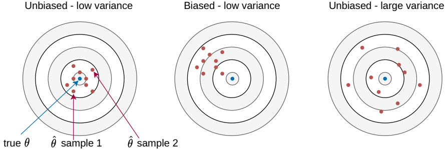


Bias

$$\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Mean square error (MSE)

$$\mathbb{E}(\hat{\theta} - \theta)^2 = \mathbb{V}(\hat{\theta}) + \left(\text{bias}(\hat{\theta})\right)^2$$



Consistent estimator

- Law of large numbers

$$\bar{X}_n \xrightarrow{p} \mu$$

- An estimator $\hat{\theta}$ is **consistent** for a population parameter θ if

$$\hat{\theta}_n \xrightarrow{p} \theta$$

which, by convergence in probability, means that for any $\epsilon > 0$

$$\Pr(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \quad \text{as } n \to \infty$$

- **Result**: An unbiased estimator $\hat{\theta}$ is consistent if

$$\mathbb{V}(\hat{\theta}_n) \to 0 \quad \text{as } n \to \infty$$

Properties of the MLE

- **Invariance of the MLE**: Let $\hat{\theta}$ be the MLE for θ and $g(\theta)$ a function of the parameter. Then, the MLE for $g(\theta)$ is $g(\hat{\theta}_{ML})$.
- Useful: obtain the MLE $\hat{\theta}$ and pop that into the function $g(\theta)$

$$\widehat{g(\theta)}_{ML} = g(\hat{\theta}_{ML})$$

- Example 1: $\mathrm{Pois}(\lambda)$ data. MLE of e^{λ} is $e^{\hat{\lambda}_{ML}} = e^{\bar{y}}$.
- Example 2:
  - Common model for income distribution:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{LogNormal}(\mu, \sigma^2)$$

  - The **Gini coefficient** $0 \leq G \leq 1$ is a measure of income inequality. For LogNormal data it can be shown that

$$G = 2\Phi(\sigma/\sqrt{2}) - 1$$

    where $\Phi(z)$ is the cdf for the standard normal distribution.
  - MLE for the Gini coefficient

$$\hat{G}_{ML} = 2\Phi(\hat{\sigma}_{ML}/\sqrt{2}) - 1$$

Sufficiency (not exam material)

- A **statistic** $T = t(X_1, \ldots, X_n)$ is a **compression of the data** into some lower-dimensional quantity.
- Examples: sample mean $\bar{X}_n$ or the sample variance s^2.
- A statistic $T = t(X_1, \ldots, X_n)$ is **sufficient** for a parameter θ if

$$\Pr(X_1, \ldots, X_n | T = t, \theta) = \Pr(X_1, \ldots, X_n | T = t)$$

- A **sufficient statistic** captures all the information in the data about the parameter θ.
- **Factorization criterion**. A statistic **T is sufficient for θ** if and only if the likelihood can be written

$$L(x_1, \ldots, x_n | \theta) = g(t, \theta) h(x_1, \ldots, x_n),$$

where $h(x_1, \ldots, x_n)$ is a function that does not involve θ.

Sufficiency and the MLE

- Assume that a data compression $T = t(X_1, \ldots, X_n)$ is sufficient for θ. We observe $T = t$.

- Since T is sufficient for θ, the log-likelihood can be written

$$\log L(\theta) = \log g(t, \theta) + \log h(x_1, \ldots, x_n)$$

- The maximum likelihood estimator $\hat{\theta}_{ML}$ is obtained by solving

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log L(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \log g(t, \theta)$$

- Enough to only keep the compressed data when finding $\hat{\theta}_{ML}$.
- Useful for online learning with streaming data.

Sufficiency in the iid Poisson model

- Likelihood when $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{n\bar{y}} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} = g(t, \theta) \cdot h(y_1, \ldots, y_n)$$

  where $t = \bar{y}$,

$$g(\bar{y}, \theta) = \lambda^{n\bar{y}} e^{-n\lambda} \qquad \text{and} \qquad h(y_1, \ldots, y_n) = \frac{1}{\prod_{i=1}^{n} y_i!}$$

  so $\bar{y}$ is a sufficient statistic for the parameter λ.

- The sample size n is a known constant, not a random variable.