Lecture 4: Parameter Estimation (Part I)

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 14, 2024

Today

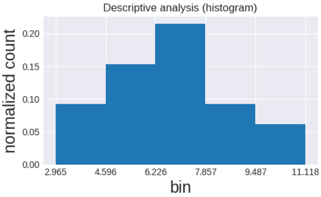
1. Mathematical modeling
2. Parameter estimation
   - Maximum likelihood estimation (MLE)
   - Likelihood and likelihood function
   - Joint probability distribution
   - Independence
3. Summary

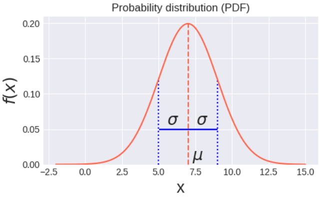
Learning outcome

- Be able to explain different components in a mathematical model $y = g(x; \theta \mid h)$
- Understand the purpose and general steps of parameter estimation
- Be able to explain these concepts: joint probability distribution, independent and identically distributed (i.i.d.) random variables, likelihood function, maximum likelihood

Recap: three questions from lecture 2

Jack suggested to use a Gaussian distribution to model your data.





- ☑ Question 1: Why should I use probability distributions instead of histograms?
- ☑ Question 2: How do you know if my data follows a Gaussian distribution?
- ? Question 3: How do I find the unknown parameters?

In today's lecture, we are going to address question 3.

Today

1. Mathematical modeling
2. Parameter estimation
3. Summary

What you will learn from this section

In the previous section, we have touched upon the topic of choosing a probabilistic model to describe a given data set.

Generally speaking, given a data set and a problem to be solved, you need to formulate the solution mathematically so that you can write a computer program to solve the problem. This is the main task for a data scientist.

This section aims to help you get started by providing explicit components and steps for formulating mathematical models.

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
    - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
    - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
  - Symbols:
    - Semicolon (";") is used to emphasize that θ is not known for free - it needs to be estimated
    - Bar ("|" pronounced "*given*") is used to indicate that h is known to you
  - Note: x, y, θ and h are not necessarily scalars; they can be multiple scalars, vectors or more complex data structures; g can be complex functions, for instance, a machine learning model or a deep neural network.

Five questions

Overwhelmed? Take it easy! Here is something that helps you get started! Answer these five questions in the language of mathematics step by step:

- 1) What do we want to predict, i.e. what is the target y?
- 2) What are the variables x?
- 3) What is the mathematical function g that relates variables x to the target y?
- 4) Are there any hyperparameters h in the function g? How do we choose them?
- 5) What are the unknown parameters θ in g? **How do we estimate them from data?**

Probabilistic modeling

Model a duck's weight using a probability distribution.
Example:

$$P(\text{A duck weighs between any two given kilograms}) = P(x_1 \leq X \leq x_2) \Leftrightarrow$$

$$y = g(x_1, x_2; \mu, \sigma) = \int_{x_1}^{x_2} f_X(t)dt = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

- 1) What do we want to predict, i.e. what is the target y? - The probability of the event $x_1 \leq weight \leq x_2$
- 2) What are the variables x? - x_1 and x_2 (we want to predict y for <u>any</u> x_1 and x_2)
- 3) What is the mathematical function g that relates variables x to the target y? - The integral of the Gaussian PDF
- 4) Are there any hyperparameters h in the function g? How do we choose them?- There is none in this case
- 5) What are the unknown parameters - μ and σ **How do we estimate them from data?**

General steps of parameter estimation for probabilistic models

- Note: the estimate of θ is denoted as $\hat{\theta}$.
- General steps for parameter estimation for a probabilistic model g:
  - a) Describe the **experiments**
  - b) Describe the **data** generated from the experiments
  - c) Describe the **random variables**
  - d) Identify **parameters of interest** θ
  - e) Choose an **estimation method**, e.g. MLE/MAP
  - f) **Compute** $\hat{\theta}$ typically by solving an optimization problem
    - Closed-form solution for simple cases
    - Iterative methods for general cases
  - g) Evaluation: estimate and report the uncertainty of $\hat{\theta}$ (later)
- **Underlying assumption**: the data used for parameter estimation is drawn from the same distribution as the data used for prediction.

Today

1 Mathematical modeling

2 Parameter estimation

- Maximum likelihood estimation (MLE)
- Likelihood and likelihood function
- Joint probability distribution
- Independence

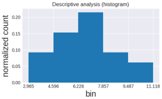
3 Summary

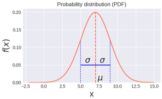
Overview

Given a dataset and a problem to be solved, now you know how to choose a probability distribution. However, the model has unknown parameters. In this section, you will learn how to estimate these parameters from data.

- There are two important parameter estimation methods: 1) the **maximum likelihood estimation (MLE)** and 2) the **maximum a posteriori estimation**.
- Concepts such as likelihood function, independent and identically distributed random variables, prior, posterior, Bayes' rule, etc are important building blocks for future machine learning models.

Overview (cont.)





- In a Gaussian distribution, what are the parameters to be estimated? mean μ and standard deviation σ
- The **maximum likelihood estimates** are the sample mean $\bar{x}$ and the sample standard deviation s for parameters μ and σ, respectively.

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\hat{\sigma} = s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- Straightforward for Gaussian distribution! Gaussian is great!
- However, it is not straightforward for all distributions - it is important to properly understand the MLE framework.

Maximum likelihood estimation (MLE)

Simplest case study: estimate one parameter given one observation

- **Model** g (cf. lecture 3):
  - **Assumption**: a duck's weight is drawn from a Gaussian distribution with standard deviation σ and mean μ

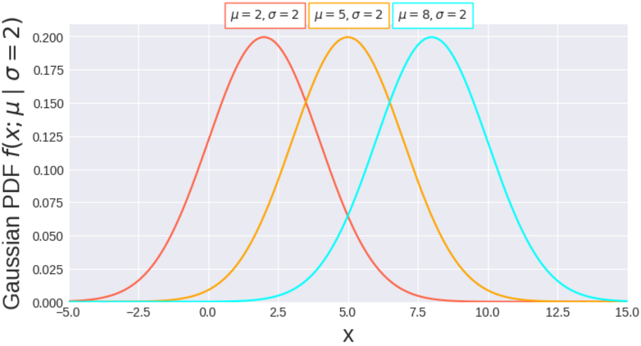
  To simplify the problem for illustration purposes, let's only look at one parameter for now:
  - We assume that σ is known to us: $\sigma = 2$
  - Unknown parameter: μ

  We want to estimate this unknown parameter by collecting some data from experiments.
- **Experiment**: we weigh a duck and observe its weight
- **Data**: the duck weighs 4 kg
- **Random variable**: $X = x$ if a duck weighs x kg
- **Parameter of interest**: μ
- **Estimation method**: the maximum likelihood estimation for μ
- **Compute** $\hat{\mu}_{MLE}$ by maximizing the **likelihood function**

Can you guess what result we are going to get? $\hat{\mu}_{MLE} = 4$

Intuition

Which Gaussian distribution is most “likely” to be the underlying model for the given data $x = 4$?

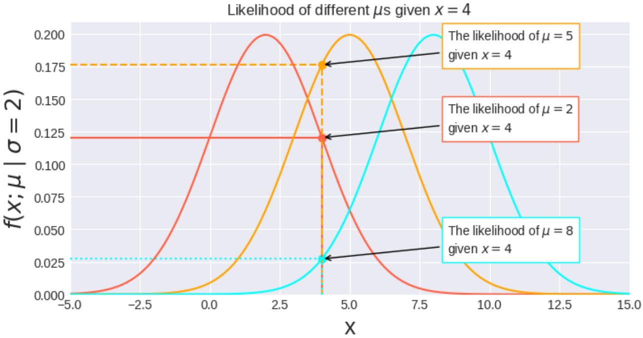


Likelihood and likelihood function

🚨 Terminology alert 🚨 - likelihood

Assumption (reminder): weights follow a Gaussian distribution with unknown parameter μ and known $\sigma = 2$

- **Likelihood of μ** given data $x = 4$ is $f(x = 4; \mu \mid \sigma = 2)$



A nonrigorous note on functions and variables

- Let g be a function that relates input variables x to a target y:

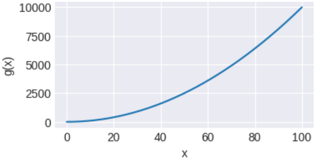
$$y = g(x)$$

- Typically, we care about the behavior of y for **all possible values for** x. This is called **generalization** in machine learning.
- Even if we add parameters θ and hyperparameters h to g, $g(x; \theta \mid h)$ is still a function of x.
- In a plot, we typically place the variable on the x-axis!
- If we are interested in the behavior of y in terms of θ, we can construct a different function L that takes θ as its input, $y = L(\theta)$, to relate θ to y.

A nonrigorous note on functions and variables (cont.)

- Example: $y = g(x) = x^2$
- In Python, all possible values for x means something like this:

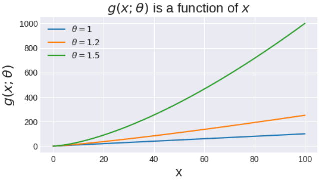
```
# Assume x can take any value between 0 and 100
xmin, xmax = 0, 100
N = 10000 # ideally, N should be infinity. But sadly, computers are discrete
          # so N has to be finite.
x = np.linspace(xmin, xmax, num=N) # all possible values for x
# Plot a function
def g(t):
    return np.power(t, 2)
y = g(x)
plt.plot(x, y)
```



A nonrigorous note on functions and variables (cont.)

- Now we add a parameter θ to g: $y = g(x;\theta) = x^{\theta}$

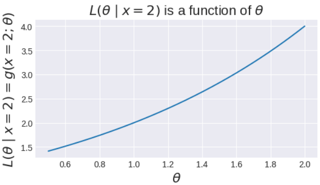
```
def g_theta(t, theta):
    return np.power(t, theta)
xmin, xmax = 0, 100 # assume x can take any value between 0 and 100
N = 10000
x = np.linspace(xmin, xmax, num=N) # all possible values for x
y = g_theta(x, 1)
plt.plot(x, y)
y = g_theta(x, 1.2)
plt.plot(x, y)
y = g_theta(x, 1.5)
plt.plot(x, y) # x is still on the x-axis
```



A nonrigorous note on functions and variables (cont.)

- Now we define a new function: $y = L(\theta \mid x = 2) = g(x = 2; \theta) = 2^{\theta}$

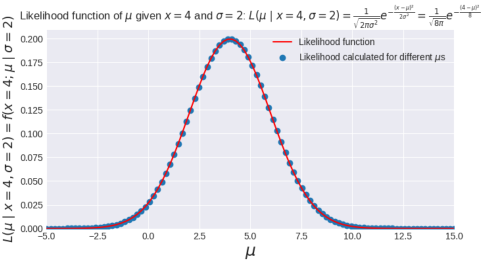
```
def L(t):
    return g_theta(2, t)
# Now theta is the variable! So we need to get all possible values for theta
# Assume theta can take any value between 0.5 and 2
theta_min, theta_max = 0.5, 2
N = 10000
thetas = np.linspace(theta_min, theta_max, num=N) # all possible values for theta
y = L(thetas)
plt.plot(thetas, y) # theta is on the x-axis now
```



🚨 Terminology alert 🚨 - likelihood function

- **Likelihood function of μ** given data $x = 4$ for $-\infty \leq \mu \leq \infty$:

$$\begin{aligned} L(\mu \mid x = 4, \sigma = 2) &= f(x = 4; \mu \mid \sigma = 2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}} \end{aligned}$$

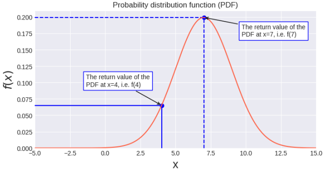


- A tiny note about symbol (most abstract), definition (less abstract) and computation (concrete - something you can implement it in Python)

Recall: probability density function and probability of events

Gaussian distribution with $\mu = 7$, $\sigma = 2$

- Probability density function $f(x \mid \mu = 7, \sigma = 2)$:

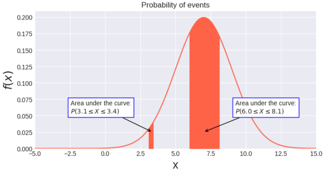


Probability distribution function (PDF)

The return value of the PDF at x=7, i.e. f(7)

The return value of the PDF at x=4, i.e. f(4)

- Probability of events $P(x_1 \leq X \leq x_2)$:



Probability of events

Area under the curve: $P(3.1 \leq X \leq 3.4)$

Area under the curve: $P(6.0 \leq X \leq 8.1)$

Probability of events vs likelihood function

- Probability of events given $\mu = 7$ and $\sigma = 2$:

$$\begin{aligned} g(x_1, x_2 \mid \mu = 7, \sigma = 2) &= P(x_1 \le X \le x_2) \\ &= \int_{x_1}^{x_2} f_X(t)dt = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= \int_{x_1}^{x_2} \frac{1}{\sqrt{8\pi}} e^{-\frac{(t-7)^2}{8}} dt \end{aligned}$$

  Here x_1 and x_2 are the **variables** - when we change x_1 and x_2, we get a different probability $g(x_1, x_2 \mid \mu = 7, \sigma = 2)$.
- Likelihood function for a given observation $x = 4$ (with known $\sigma = 2$):

$$L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

  Here μ is the **variable** - when we change μ, we get a different likelihood $L(\mu \mid x = 4, \sigma = 2)$.

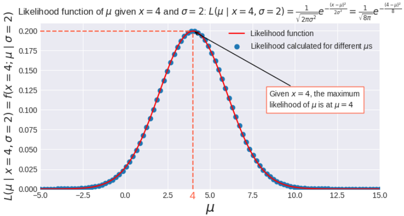
Maximum likelihood

From the likelihood function

$$L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

We can now define the **maximum likelihood** of μ given $x = 4$:

$$\text{the maximum likelihood of } \mu = \max(L(\mu \mid x = 4, \sigma = 2))$$



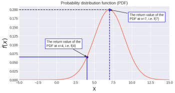
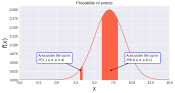
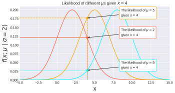
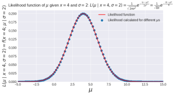
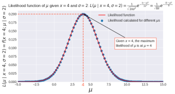
Likelihood function of μ given $x = 4$ and $\sigma = 2$: $L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$

Given $x = 4$, the maximum likelihood of μ is at $\mu = 4$

The value of μ that maximizes the likelihood function is called the **maximum likelihood estimation** (MLE) of μ. In this case, $\hat{\mu}_{\text{MLE}} = 4$.

Note: $\hat{}$ here means that $\hat{\mu}$ is an estimate instead of the true value μ.

Comparison

Probability density function	
Probability of events	
Likelihood of a parameter given data	
Likelihood function of a parameter given data	
Maximum likelihood estimation	

Summary: what have we done so far?

- We observe one data point $x = 4$.
- We assume that duck weights are drawn from a *Gaussian distribution* with known $\sigma = 2$ and unknown μ. We need to estimate μ.
- We write down the likelihood function: $L(\mu \mid x = 4, \sigma = 2) = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$.
- The maximum likelihood estimation of μ is defined as:

$$\hat{\mu}_{\text{MLE}} = \arg\max_{\mu} L(\mu \mid x = 4, \sigma = 2) = \arg\max_{\mu} \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}} \qquad (1)$$

Remaining questions

- We can't estimate the whole distribution from only one data point $x = 4$! What if we have more than one observation?
- How can we maximize the likelihood function and find the value of $\hat{\mu}_{MLE}$ analytically?
- What if σ is also unknown?
- What about discrete distributions?

Case study: parameter estimation given more observations

- **Model**:
  - Assumption: a duck's weight is drawn from a Gaussian distribution with known standard deviation $\sigma = 2$ and unknown mean μ
- **Experiment**: we observe 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
weight	6.98	5.43	2.97	7.07	$\cdots$	4.63	7.27

- **Parameter of interest**: μ
- **Estimation method**: maximum likelihood estimation
- **Compute $\hat{\mu}_{MLE}$** by maximizing the likelihood function

- Recall: when we only have one observation $x = 4$, the likelihood function looks like this

$$L(\mu \mid x = 4, \sigma = 2) = f(x = 4; \mu \mid \sigma = 2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{8\pi}} e^{-\frac{(4-\mu)^2}{8}}$$

- Educated guess 🤔 - now we have more observations, the likelihood function probably should look like this:

$$L(\mu \mid x_1 = 6.98, \cdots, x_{20} = 7.27, \sigma = 2) = \boxed{f(x_1 = 6.98, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2)}$$

Joint probability distribution

🚨 Terminology alert 🚨 - joint probability distribution 👯

Given two random variables X and Y, we use their **joint probability distribution** to characterize their behaviors:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) \ \text{ joint CDF}$$

- X, Y discrete: joint PMF $f_{X,Y}(x,y) = P(X = x, Y = y)$
- X, Y continuous: joint PDF $f_{X,Y}(x,y)$

- Bummer: these expressions are usually quite hard to obtain...
- Solution: we impose some assumptions to make the calculation easier.

Independence

Independence 😼

- Recall independent events: two events A and B are independent if and only if

$$P(A \text{ and } B) = P(A \cap B) = P(A)P(B)$$

New! **Independent random variables**: random variables X, Y are independent if and only if

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y)$$

  - X, Y discrete:

    $$f_{X,Y}(x,y) = P(X = x, Y = y) = P(X = x)P(Y = y) = f_X(x)f_Y(y)$$

    where $f_{X,Y}(x,y)$ is the joint PMF
  - X, Y continuous:

    $$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

    where $f_{X,Y}(x,y)$ is the joint PDF
- This idea generalizes to more than two random variables

Independence 😼

Any number of random variables:

- Given n random variables $X_1, X_2, \cdots, X_n$ with CDF $F_{X_i}(x_i)$,

$$F_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

  where $F_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint CDF

  - X_i discrete with PMF $f_{X_i}(x)$:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

    where $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint PMF

  - X_i continuous with PDF $f_{X_i}(x)$:

$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

    where $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n)$ is the joint PDF

- Now we have turned the joint probability distribution into multiplications of things we know how to compute. Yay!

Back to the case study

The likelihood function

$$L(\mu \mid x_1 = 6.98, \cdots, x_{20} = 7.27, \sigma = 2) = \boxed{f(x_1 = 6.98, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2)}$$

- **Model**:
  - Assumption: the weight is drawn from a Gaussian distribution with known standard deviation $\sigma = 2$ and unknown mean μ
- **Experiment**: we weigh 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
weight	6.98	5.43	2.97	7.07	$\cdots$	4.63	7.27

- **Random variable**: we define 20 random variables X_i : duck weight $\to \mathbb{R}$, where X_i are **independent and identically distributed (i.i.d)** Gaussian random variables
  - $X_1, \cdots, X_{20}$ are independent - [new!] assumption:

$$f_{X_1,\cdots,X_{20}}(x_1 = 6.98, \cdots, x_{20} = 7.27) = f_{X_1}(x = 6.98) \cdots f_{X_{20}}(x = 7.27)$$

  - $X_1, \cdots, X_{20}$ are identically distributed - they have the same PDF:

$$f_{X_1}(x; \mu \mid \sigma) = \cdots = f_{X_{20}}(x; \mu \mid \sigma) = f(x; \mu \mid \sigma)$$

  where $\sigma = \sigma_1 = \sigma_2 = \cdots = \sigma_{20} = 2$ and $\mu = \mu_1 = \mu_2 = \cdots = \mu_{20}$.
- **Parameter of interest**: μ
- **Estimation method**: maximum likelihood estimation
- **Compute** $\hat{\mu}_{MLE}$ by maximizing the likelihood function

Case study: the likelihood function given i.i.d. observations

- Put everything together, the likelihood function is expressed as

$$
\begin{aligned}
& L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\
= \; & f(x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27; \mu \mid \sigma = 2) \text{ - joint PDF} \\
= \; & f(6.98; \mu \mid \sigma = 2) f(5.43; \mu \mid \sigma = 2) \cdots f(7.27; \mu \mid \sigma = 2) \text{ - i.i.d. assumption} \\
= \; & \frac{1}{\sqrt{8\pi}} e^{-\frac{(6.98-\mu)^2}{8}} \frac{1}{\sqrt{8\pi}} e^{-\frac{(5.43-\mu)^2}{8}} \cdots \frac{1}{\sqrt{8\pi}} e^{-\frac{(7.27-\mu)^2}{8}} \text{ - plug in data} \\
= \; & (\frac{1}{\sqrt{8\pi}})^{20} e^{-\frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8}} \text{ - something concrete}
\end{aligned}
$$

- MLE can be computed as:

$$
\begin{aligned}
\hat{\mu}_{MLE} &= \arg\max_{\mu} L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\
&= \arg\max_{\mu} (\frac{1}{\sqrt{8\pi}})^{20} e^{-\frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8}}
\end{aligned}
$$

- From page 10, the only step left is f) compute $\hat{\mu}_{MLE}$.

Compute $\hat{\mu}_{MLE}$ from the likelihood function 🙈

The maximum likelihood estimate of μ is the value that maximizes the likelihood function

$$\begin{aligned}
\hat{\mu}_{MLE} \quad &= \quad \underset{\mu}{\arg\max}\ L(\mu \mid x_1 = 6.98, x_2 = 5.43, \cdots, x_{20} = 7.27, \sigma = 2) \\
&= \quad \underset{\mu}{\arg\max}\ L(\mu \mid \boldsymbol{x}) \text{ (for convenience)} \\
&= \quad \underset{\mu}{\arg\max}\ \log L(\mu \mid \boldsymbol{x}) \text{ (does not change the location of the maximum)}
\end{aligned}$$

Log function turns multiplications into sums, which makes everything much easier to compute

$$\begin{aligned}
&= \quad \underset{\mu}{\arg\max}\ l(\mu \mid \boldsymbol{x}) \text{ (l is called the } \textbf{log likelihood}\text{)} \\
&= \quad \underset{\mu}{\arg\min}\ -l(\mu \mid \boldsymbol{x}) \text{ (this is the } \textbf{negative log likelihood} \text{ - mainly for standardization purposes)} \\
&= \quad \underset{\mu}{\arg\min}\ -\log(\frac{1}{\sqrt{8\pi}})^{20} + \frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8} \text{ (plug in the expression of the PDFs)}
\end{aligned}$$

Set the derivative to zero to find the optimal solution

$$\begin{aligned}
&\Longleftrightarrow \quad \frac{\partial}{\partial \mu}\left(-\log(\frac{1}{\sqrt{8\pi}})^{20} + \frac{(6.98-\mu)^2 + \cdots + (7.27-\mu)^2}{8}\right) = 0 \\
&\Longleftrightarrow \quad \sum_{i=1}^{20} x_i - 20\mu = 0 \\
&\Longleftrightarrow \quad \hat{\mu}_{MLE} = \frac{1}{20}\sum_{i=1}^{20} x_i = 6.647
\end{aligned}$$

Compute $\hat{\mu}_{MLE}$ from the likelihood function (cont.)

Okay, here is the takeaway...

- The MLE of the mean value of a Gaussian distribution is the sample mean.
- It's an **optimization problem**. Read about the context here.
- Given the i.i.d. assumption, the MLE of parameters have closed-form solutions for many known distributions, e.g. Gaussian, uniform, Bernoulli, etc, e.g. google "mle Gaussian distribution".
- For general cases, the solution typically does not have closed-form and needs to be found using **iterative methods** and **approximation techniques**.
- The good news is 1) i.i.d. is usually a reasonable assumption; 2) many things in the world are naturally Gaussian!

Note: solving optimization problems is an important building block of data science and machine learning. For example, if you want to understand deep learning with backpropagation, you need to know at least why fiddling with the derivative leads to optimality.

What if σ is also unknown?

- Now we have two unknown parameters μ and σ, the likelihood function becomes:

$$L(\mu, \sigma \mid \boldsymbol{x}) = (\frac{1}{\sqrt{2\sigma^2\pi}})^{20} e^{-\frac{(6.98-\mu)^2+\cdots+(7.27-\mu)^2}{2\sigma^2}}$$

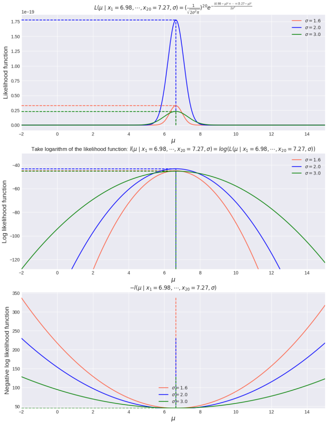
- Now we need to solve for both μ and σ:

$$\hat{\mu}_{MLE} = \arg\max_{\mu} L(\mu, \sigma \mid \boldsymbol{x}) \iff \frac{\partial}{\partial\mu} L(\mu, \sigma \mid \boldsymbol{x}) = 0$$
$$\hat{\sigma}_{MLE} = \arg\max_{\sigma} L(\mu, \sigma \mid \boldsymbol{x}) \iff \frac{\partial}{\partial\sigma} L(\mu, \sigma \mid \boldsymbol{x}) = 0$$

Note: left as an exercise.

- What happens to the partial derivative of σ with respect to μ and vise versa?

$\hat{\mu}_{MLE}$ given different σ



NOTE: the maximum likelihood estimation of μ does not depend on σ

What about discrete distributions?

Likelihood function for discrete distributions:

- Likelihood function given one discrete random variable X with PMF $f_X(x)$:
$$L(\theta \mid x) = P(X = x) = f_X(x; \theta)$$
- Likelihood function given N i.i.d. discrete random variables with PMF $f(x)$:
$$L(\theta \mid x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i; \theta)$$

Example: MLE for discrete distributions

- **Model**:
  - Assumption: the color of a duck is drawn from a Bernoulli distribution with PMF

$$f(k;p) = pk + (1-p)(1-k), k \in \{0,1\}$$

- **Experiment**: we observe the colors from 20 ducks
- **Data**:

duck id	1	2	3	4	$\cdots$	19	20
color	red	blue	red	red	$\cdots$	red	blue

- **Random variable**: let $X_i = \begin{cases} 0, & \text{a duck is red} \\ 1, & \text{a duck is blue} \end{cases}$ be independent and identically distributed (i.i.d.) Bernoulli random variables.
- **Parameter of interest**: p
- **Estimation method**: maximum likelihood estimation
- **Compute** $\hat{p}_{MLE}$ by maximizing the likelihood function

Example: MLE for discrete distributions 🙈 🙈

- Likelihood function:

$$
\begin{aligned}
L(p \mid \boldsymbol{x}) &= f(x_1; p) f(x_2; p) \cdots f(x_{20}; p) \text{ (using ";" instead of "|" because } p \text{ is unknown)} \\
&= (1-p) \times (1-0) \times p \times 1 \cdots p \times 1 \text{ (plug in Bernoulli PMF)} \\
&= p^{(\text{\# of blue ducks})}(1-p)^{(\text{\# of red ducks})} \\
&= p^6 (1-p)^{14}
\end{aligned}
$$

- Negative log likelihood function:

$$
-l(p \mid \boldsymbol{x}) = -\log(p^6(1-p)^{14}) = -(6\log(p) + 14\log(1-p))
$$

- Minimize the negative log likelihood function by setting the derivative to zero:

$$
\begin{aligned}
& \frac{\partial}{\partial p}(-l(p \mid \boldsymbol{x})) = 0 \\
\iff \quad & \frac{\partial}{\partial p}(6\log(p) + 14\log(1-p)) = 0 \text{ (we need to use a tiny chain rule here)} \\
\iff \quad & \frac{6}{p} - \frac{14}{1-p} = 0 \Rightarrow p = \frac{6}{20} \\
\Rightarrow \quad & \hat{p}_{MLE} = \arg\min_{p} \; -l(p \mid \boldsymbol{x}) = \frac{6}{20}
\end{aligned}
$$

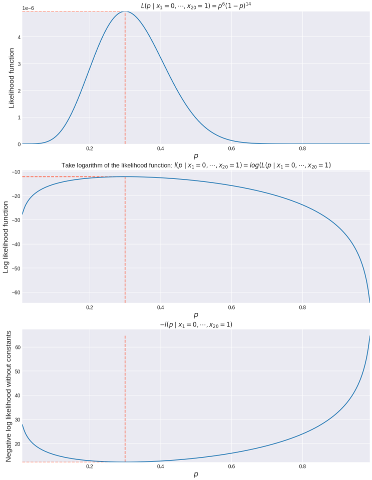
MLE for discrete distributions: two notes

Note 1: the MLE of the parameter p of the Bernoulli distribution is the sample mean of the observed data, i.e.,

$$\hat{p}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{\text{count}(x_i = 1)}{N}$$

MLE for discrete distributions: two notes

Note 2: the likelihood function of the parameter p in the Bernoulli distribution (discrete) is a *continuous* function.



Properties of MLE for interested readers

- The point of learning how to obtain the MLE instead of just "taking the average or something" is to have a theoretical framework to analyze the behaviors of these estimates.
- There are two example properties:
  - The maximum likelihood estimators are unbiased.
  - Under some regularity conditions (that are usually true in practice), MLE is asymptotically Gaussian. More precisely, for large N, $\hat{\theta}_{MLE}$ approximately follows a Gaussian distribution with mean θ and variance $\frac{1}{I_N(\theta)N}$, where $I_N(\theta)$ is called the **Fisher information**, which is defined as

$$I_N(\theta) = E_{\Theta}\left[-\frac{d^2}{d\theta^2} l(\theta \mid X_1, \cdots, X_N)\right]$$

    where $l(\theta \mid X_1, \cdots, X_N)$ is the log likelihood.

Note: in this context, "asymptotically" means "it is true when you have a lot of data"

Summary: steps to find the maximum likelihood estimation

Given a model $y = g(x; \mathcal{O} \mid h)$, where $\mathcal{O}$ is a set of parameters

- a) Describe the experiments
- b) Describe the data generated from the experiments
- c) Describe the random variables (typically with i.i.d. assumption)
- d) Identify parameters of interest $\theta \in \mathcal{O}$
- e) Choose the maximum likelihood estimation as the estimation method: Given data $x_1, \cdots, x_N$, find the likelihood function

$$L(\theta \mid x_1, \cdots, x_N) = f_{X_1 \cdots X_N}(x_1, \cdots, x_N; \theta)$$

If the random variables X_i are assumed to be i.i.d., then we have:

$$L(\theta \mid x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i; \theta)$$

where f is the PMF for discrete random variables and the PDF for continuous random variables.

Summary: steps to find the maximum likelihood estimation

- f) Compute $\hat{\theta}_{MLE}$ by maximizing the likelihood function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid x_1, \cdots, x_N) = \arg\max_{\theta} f_{X_1 \cdots X_N}(x_1, \cdots, x_N; \theta)$$

Given the i.i.d. assumption, we have

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta \mid x_1, \cdots, x_N) = \arg\max_{\theta} \prod_{i=1}^{N} f(x_i; \theta)$$

or equivalently, minimizing the negative log likelihood function (easier to compute/avoiding underflow):

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \; -\sum_{i=1}^{N} \log\left(f(x_i; \theta)\right)$$

  - Simple case, e.g. i.i.d. Gaussian, find the closed-form solution by:
    - Taking the partial derivative with respect to the parameter
    - Setting the derivative to zero
    - Solving for the parameter

    Note: the solutions for many distributions are already available in closed-form if you google them.
  - In general, the estimate needs to be found by iterative methods, e.g. gradient descent

Today

1. Mathematical modeling
2. Parameter estimation
3. Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation

Next (part II):

- Maximum a posteriori estimation

Before next lecture:

- Conditional probability, i.i.d. random variables

Until next time!