Introduction to Statistics: Lecture Notes

Motonobu Kanagawa
Data Science Department, EURECOM

March 24, 2023

Contents

Chapter 1

Overview and Introduction

1.1 Introduction

Probability: Language for Describing Uncertainty

Uncertainty: Degree of being not certain about a certain statement.

Uncertainty arises from various sources.

- The statement may be about the future

— e.g., It will rain tonight.

- Lack of information

— e.g., Does this patient have a cancer?

— If we only have limited information about the patient, we cannot be certain about this statement (unless we conduct a more detailed investigation).

Why Uncertainty/Probability Matters?

Question. Why do we need to quantify uncertainty?

One answer. Because we need to make "decisions".

Example 1: Whether to Bring an Umbrella

- You need to decide whether to bring an umbrella for going out.

- This involves the uncertainty regarding whether it will rain today.

  - If you believe that it will rain, you would bring the umbrella.
  - If the weather forecast says probability 0% of rain, you would not bring the umbrella.

- It also involves costs:

  - The cost of bringing the umbrella.
  - The cost of being wet without bringing the umbrella.

- If you don't care about being wet, you would not bring the umbrella anyway.

- If your clothes are fragile, you would bring the umbrella even when the probability of rain is low.

Example 2: Which Drug Should a Doctor Give to a Patient?

- Assume that you are a doctor and seeing a patient.

- You need to decide which drug you should give to the patient.

- The decision should take various uncertainties into account.

  - Which disease does the patient have (disease A, B, C, ...)?
  - Whether the drug α (or β, γ ...) is effective to the disease A (or B, C, ...).

- The decision should also take costs into account.

  - How expensive / risky is the drug?
  - If the drug is risky, and if the disease is not serious, you would choose not to give the drug.

Probability, Statistics, and Decision Theory

The course involves the following three disciplines.

Probability Theory:

- provides a way of quantitatively modeling uncertainties.

Statistics:

- provides a way of estimating probabilities/uncertainties from data.

Statistical Decision Theory:

- provides a way of determining the optimal decisions/policies under uncertainty.

1.2 Schedule

Schedule

Week 1: Introduction and Probability Theory.

Week 2: Mean Estimation and Introduction to Estimation Theory.

Week 3: Parametric Models and Maximum Likelihood Estimation.

Week 4: Statistical Hypothesis Testing

Week 5: Statistical Hypothesis Testing (Contd)

Week 6: Bayesian Inference I

Week 7: Bayesian Inference II: Hypothesis Testing

1.3 Recommended reading

Recommended Books

- B. Efron and T. Hastie, "Computer Age Statistical Inference: Algorithms, Evidence and Data Science", Cambridge University Press, 2016.

- J. O. Berger, "Statistical Decision Theory and Bayesian Analysis", Springer, 1985.

Chapter 2

Probability Theory

2.1 Subjective and Objective Probabilities

What is the Meaning of "Probability"?

Subjective Probability (used in Bayesian statistics):

- A probability represents one's degree of belief or knowledge about a certain statement, expressed as a number between 0 and 1.

e.g.,

- This drug cures this disease with probability 0.6.

- It will rain today with probability 0.6.

What is the Meaning of "Probability"?

Objective Probability (used in Frequentist statistics):

- A probability represents the degree of randomness.

- Given as the frequency of a statement to be true over infinitely many repeated experiments.

- e.g., think about a biased coin, with the "probability of head 0.6".

- Assume that you toss the coin n times: then the probability statement can be understood as

$$\lim_{n\to\infty} \frac{\text{the number of heads out of } n \text{ trials}}{n} = 0.6.$$

What is the Meaning of "Probability"?

- This lecture introduces mathematical definition of probabilities, which may be used for both

(subjective and objective) interpretations.

- Note that this lecture may be the most mathematical among the other lectures in this STATS course.

- Being mathematically rigorous is similar to being rigorous about grammar in a language course

- But please don't be scared: the other lectures are less mathematical.

- Only very basic questions may appear in the exam.
- Don't quite the course because of the lecture today!

2.2 Probability Spaces

Probability Space: Definition

A triplet $(\Omega, \mathcal{F}, P)$ is called a **probability space**, if

i) Ω is a set (e.g., $\Omega = \mathbb{R}$):

— this is called a sample space, and each $\omega \in \Omega$ is called a sample or an elementary event.

ii) $\mathcal{F}$ is a σ-algebra, i.e., a set of subsets of Ω satisfying

1. $\phi \in \mathcal{F}$ and $\Omega \in \mathcal{F}$; (ϕ is an empty set)
2. If $A \in \mathcal{F}$, then $\Omega \backslash A \in \mathcal{F}$; ($\Omega \backslash A := \{\omega \in \Omega \mid \omega \notin A\}$.)
3. If $A_1, A_2, \ldots, \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

- Each $A \in \mathcal{F}$ is called a measurable set
- Each $A \in \mathcal{F}$ can be understood as a certain logical statement (we'll see later).
- Thus, the σ-algebra is a set of statements for which probabilities are defined.

Probability Space: Definition

iii) P: a probability measure (distribution), i.e.,

1. P is a function from $\mathcal{F}$ to $[0, 1]$:
    - The probability $P(A)$ of any statement $A \in \mathcal{F}$ being true between 0 and 1.

2. $P(\phi) = 0$ and $P(\Omega) = 1$.
3. For $A_1, A_2, \cdots \in \mathcal{F}$ such that $A_i \cap A_j = \phi$ with $i \neq j$, we have

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Probability Space: Interpretation

For any two statements $A, B \in \mathcal{F}$,

- The intersection

$$A \cap B := \{\omega \mid \omega \in A \text{ and } \omega \in B\}$$

may be understood as the statement that "both A and B are true."

- The union

$$A \cup B := \{\omega \mid \omega \in A \text{ or } \omega \in B\}$$

may be understood as the statement that "either A or B is true."

- $A \cap B = \phi$ means that A and B cannot be true simultaneously.

Probability Space: Interpretation

Thus, for statements $A, B \in \mathcal{F}$ with $A \cap B = \phi$,

- $P(A \cap B) = P(\phi) = 0$:

  - The probability of both A and B being true is 0.

- $P(A \cup B) = P(A) + P(B)$:

  - The probability that either A or B is true is the sum of the probability of A being true and the probability of B being true.

Probability Space: Interpretation

Note that for any $A \in \mathcal{F}$, the complement

$$\Omega \backslash A := \{\omega \in \Omega \mid \omega \notin A\}$$

maybe understood as the negation of A.

Thus,

$$1 = P(\Omega) = P(A \cup (\Omega \backslash A)) = P(A) + P(\Omega \backslash A).$$

and

$$P(\Omega \backslash A) = 1 - P(A).$$

- The probability of A being not true is 1 minus the probability of A being true.

Examples of Probability Spaces: Finite Discrete Sample Space

Consider fair coin tossing

i.e., the probabilities of "head" and "tail" are the same: $1/2$.

i) Sample space: $\Omega := \{H, T\}$.

i.e., the sample space consists of two elements:

"H" (Head) and "T" (Tail).

Examples of Probability Spaces: Finite Discrete Sample Space

ii) σ-algebra: $\mathcal{F} := \{\phi, \{H\}, \{T\}, \{H, T\}\}$

– i.e., $\mathcal{F} = 2^{\Omega}$ is the power set (= the set of all subsets of Ω).

- $\{H\}$: the statement that "the head appears"
- $\{T\}$: the statement that "the tail appears"
- $\{H, T\} = \{H\} \cup \{T\} = \Omega$: "the head or the tail appears"
- ϕ: the statement that "nothing appears".

iii) Probability measure:

$P(\phi) = 0$, $P(\{H\}) = 1/2$, $P(\{T\}) = 1/2$, $P(\{H, T\}) = 1$.

Examples of Probability Spaces: Infinite Discrete Sample Space

i) Sample space: $\Omega := \mathbb{N} := \{1, 2, \dots\}$ (i.e., all natural numbers).

ii) σ-algebra: $\mathcal{F} := 2^{\mathbb{N}}$ (i.e., all the subsets of $\mathbb{N}$).

iii) Probability measure: $P(\{n\}) := 2^{-n}$ for all $n \in \mathbb{N}$.

Exercise:

- Verify that this example satisfies the definition of a probability space.

Examples of Probability Spaces: Uncountable Sample Space

i) Sample space: $\Omega := [0, 1] \subset \mathbb{R}$.

ii) σ-algebra: $\mathcal{F} =$ the Borel σ-algebra (i.e., the smallest σ-algebra that contains all the open subsets of $[0, 1]$).

iii) Probability measure: $P((a, b)) := b - a$ for all $0 \leq a < b \leq 1$. (i.e., the uniform measure on $[0, 1]$).

- You can think about throwing a needle onto the interval [0,1] uniformly at random.

- Then $(a, b) \in \mathcal{F}$ is the statement that the needle lies between a and b.

Exercises:

- Verify that this example satisfies the definition of a probability space.
- Verify that $\mathcal{F}$ contains all the closed subsets of $[0, 1]$.

2.3 Random Variables

Random Variables

A random variable = a variable that is random.

- e.g., consider rolling a dice:

- Then the number (1, 2, ..., or 6) that appears in the top is a random variable.

How can we define a random variable mathematically?

Random Variables: Definition

Let $(P, \Omega, \mathcal{F})$ be a probability space.

Let $(\Omega_X, \mathcal{F}_X)$ be a measurable space, i.e.,

- Ω_X is a nonempty set;
- $\mathcal{F}_X$ is a σ-algebra of subsets of Ω_X.

Definition. A function $X : \Omega \to \Omega_X$ is called a random variable, if it is a measurable function, i.e.,

$$\text{For all } S \in \mathcal{F}_X, \text{ we have } X^{-1}(S) \in \mathcal{F};$$

where $X^{-1}(S) := \{\omega \in \Omega \mid X(\omega) \in S\}$ is the inverse image of S.

- In words,

$$\text{If } S \subset \Omega_X \text{ is measurable, then } X^{-1}(S) \subset \Omega \text{ is also measurable.}$$

Random Variables: The Distribution of a Random Variable

Random variable $X : \Omega \to \Omega_X$ induces a probability measure $P_X : \mathcal{F}_X \to [0, 1]$ by

$$P_X(S) := P(X^{-1}(S)), \quad S \in \mathcal{F}_X.$$

- $(P_X, \Omega_X, \mathcal{F}_X)$ is a probability space.

- P_X is called the distribution (or the law) of X.

— We write $X \sim P_X$.

- It is said that X takes values in Ω_X.

Exercise:

- Verify that $(\Omega_X, \mathcal{F}_X, P_X)$ is a probability space.

Random Variables: Discrete and Continuous

A random variable $X : \Omega \to \Omega_X$ is said to be

- discrete if X takes countably many (finite or countably infinite) values.

- continuous if X takes uncountably infinitely many values.

Random Variables: A Discrete Example

Consider a biased dice:

– Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

– σ-algebra: $\mathcal{F} := 2^{\Omega}$ (= the set of all subsets of Ω)

– Probabilities: $P(\{1\}) = 7/12$, $P(\{2\}) = \cdots P(\{6\}) = 1/12$.

Define a random variable X as follows:

– Sample space: $\Omega_X := \{0, 1\}$.

– Random variable: $X : \Omega \to \Omega_X$ defined by

$$X(\omega) := \begin{cases} 0 & \text{if } \omega = 1, 3, \text{ or } 5 \\ 1 & \text{if } \omega = 2, 4, \text{ or } 6. \end{cases}$$

i.e., $X(\omega)$ takes the value 0 if ω is odd, and takes 1 if ω is even.

Random Variables: A Discrete Example

– σ-algebra: $\mathcal{F}_X := 2^{\Omega_X} := \{\phi, \{0\}, \{1\}, \{0, 1\}\}$.

The measurability of $X : \Omega \to \Omega_X$ can be checked easily:

- $X^{-1}(\phi) = \phi \in \mathcal{F}$.
- $X^{-1}(\{0\}) = \{1, 3, 5\} \in \mathcal{F}$.
- $X^{-1}(\{1\}) = \{2, 4, 6\} \in \mathcal{F}$.
- $X^{-1}(\{0, 1\}) = \{1, 2, 3, 4, 5, 6\} \in \mathcal{F}$.

The distribution of X is thus given by

- $P_X(\phi) = P(\phi) = 0$.
- $P_X(\{0\}) = P(\{1, 3, 5\}) = 7/12 + 1/12 + 1/12 = 9/12$.
- $P_X(\{1\}) = P(\{2, 4, 6\}) = 1/12 + 1/12 + 1/12 = 3/12$.
- $P_X(\{0, 1\}) = P(\{1, 2, 3, 4, 5, 6\}) = 1$.

Random Variables: A Continuous Example

Consider a random variable following the uniform measure on the unit interval $[0, 1]$.

- The sample space: $\Omega = \Omega_X = [0, 1]$.

- The σ-algebra: $\mathcal{F} = \mathcal{F}_X =$ the Borel-σ algebra (i.e., the smallest σ-algebra containing all open subsets of $[0, 1]$).

- Random variable $X : \Omega \to \Omega_X$ by the identity, i.e., $X(\omega) = \omega$ for all $\omega \in [0, 1]$.

- Probability measure: $P = P_X$ is defined for any interval $(a, b) \subset [0, 1]$ as

$$P((a, b)) = P_X((a, b)) = b - a.$$

For instance, you can think of $X(\omega)$ as the needle that you throw on the interval $[0, 1]$ uniformly at random.

2.4 Expectation of a Random Variable

Expectation of a Random Variable

The expectation (or the mean, the average) of a random variable is an important concept in probability and statistics.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $f : \Omega_X \to \mathbb{R}$ be a measurable function (with respect to the Borel σ-algebra in $\mathbb{R}$).

- This implies, e.g., $f^{-1}(B) \in \mathcal{F}_X$ for any open (or closed) subset $B \subset \mathbb{R}$.

- Then $f(X)$ is a real-valued random variable.

- You can interpret $f(X)$ as a mapping $f(X) : \Omega \to \mathbb{R}$:

$$f(X)(\omega) := f(X(\omega)) \in \mathbb{R}.$$

Expectation of a Random Variable

- For a discrete random variable X, the expectation of $f(X)$ is defined by the sum of values $f(x)$ weighted by their probabilities $P_X(\{x\})$:

$$\mathbb{E}[f(X)] := \sum_{x \in \Omega_X} f(x) P_X(\{x\})$$

- For a continuous random variable X, the expectation is defined by the Lebesgue integral:

$$\mathbb{E}[f(X)] := \int_{\Omega_X} f(x) dP_X(x).$$

We will quickly look at the definition of the Lebesgue integral for completeness.

Lebesgue Integration: Simple Functions

First, consider a function $f : \Omega_X \to \mathbb{R}$ of the form (such a f is called a simple function)

$$f(x) = \sum_{i=1}^{n} a_i 1_{S_i}(x),$$

where

- $a_1, \ldots, a_n \in \mathbb{R}$ are constants;
- $S_i \in \mathcal{F}_X$ are disjoint to each other: $S_i \cap S_j = \phi$ for $i \neq j$;
- $1_{S_i} : \Omega_X \to \mathbb{R}$ are indicator functions:

$$1_{S_i}(x) = \begin{cases} 1 & \text{if } x \in S_i, \\ 0 & \text{if } x \notin S_i. \end{cases}$$

Lebesgue Integration: Simple Functions

The integral of the simple function $f = \sum_{i=1}^{n} a_i 1_{S_i}$ is defined by

$$\int_{\Omega_X} f(x) dP_X(x) := \sum_{i=1}^{n} a_i P_X(S_i)$$

- Note that $f(x) = a_i$ for all $x \in S_i$.
- Thus, this takes the same form as the discrete case: the sum of values $f(x)$ weighted by the probabilities $P_X(S_i)$.

In particular, for an indicator function $f(x) := 1_S(x)$ with $S \in \mathcal{F}_X$, the integral is the probability of S:

$$\int_{\Omega_X} 1_S(x) dP_X(x) = P_X(S), \quad S \in \mathcal{F}_X.$$

Lebesgue Integration: Non-negative Functions

Assume that f is measurable and non-negative, i.e., $f(x) \geq 0$ for any $x \in \Omega_X$. (See e.g. (Dudley, 2002, 4.1.5) for the details below.)

- For any $n \in \mathbb{N}$, consider the division of $[0, \infty]$ into disjoint $n \times 2^n + 1$ intervals:

$$\begin{aligned}
&[0, \infty] \\
&= \left[0, \frac{1}{2^n}\right] \cup \left(\frac{1}{2^n}, \frac{2}{2^n}\right] \cup \left(\frac{2}{2^n}, \frac{3}{2^n}\right] \cup \cdots \cup \left(\frac{n \times 2^n - 1}{2^n}, n\right] \cup (n, \infty] \\
&= \left[0, \frac{1}{2^n}\right] \cup \bigcup_{j=1}^{n \times 2^n - 1} \left(\frac{j}{2^n}, \frac{j+1}{2^n}\right] \cup (n, \infty].
\end{aligned}$$

- Define the corresponding subsets in Ω_X given by the inverse mapping f^{-1}:

$$S_{nj} := f^{-1}\left(\left(\frac{j}{2^n}, \frac{j+1}{2^n}\right]\right), \quad U_n := f^{-1}((n, \infty]).$$

Lebesgue Integration: Non-negative Functions

- These subsets S_{nj} (and U_n) are disjoint to each other.
- $S_{nj} \in \mathcal{F}_X$ and $U_n \in \mathcal{F}_X$ because f is Borel-measurable.
- Note that if $n > \max_{x \in \Omega_X} f(x)$, then $U_n = \phi$.

Then define a simple function $f_n : \Omega_X \to \mathbb{R}$ by

$$f_n(x) := \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} 1_{S_{nj}}(x) + n 1_{U_n}(x),$$

By construction,

- For $x \in S_{nj} = f^{-1}((\frac{j}{2^n}, \frac{j+1}{2^n}])$, we have $f_n(x) = \frac{j}{2^n} < f(x)$.

- For $x \in U_n = f^{-1}((n, \infty])$, we have $f_n(x) = n < f(x)$.
- For $x \in \Omega \backslash \left(\bigcup_j S_{nj} \cup U_n \right)$, we have $f_n(x) = 0 \leq f(x)$.

Therefore $f_n(x) \leq f(x)$ for all $x \in \Omega_X$.

Lebesgue Integration: Non-negative Functions

- We can also show that $f_n(x) \to f(x)$ for $n \to \infty$ for all $x \in \Omega_X$.

- Since f_n is a simple function, we can define the integral

$$\int_{\Omega_X} f_n(x) dP_X(x) := \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} P_X(S_{nj}) + n P_X(U_n)$$
$$= \sum_{j=1}^{n \times 2^n - 1} \frac{j}{2^n} P_X \left(f^{-1}(\left(\frac{j}{2^n}, \frac{j+1}{2^n} \right]) \right) + n P_X \left(f^{-1}((n, \infty]) \right)$$

- Then the integral of f can be defined as the limit of the integral of f_n as $n \to \infty$:

$$\int_{\Omega_X} f(x) dP_X(x) := \lim_{n \to \infty} \int_{\Omega_X} f_n(x) dP_X(x).$$

If $\int_{\Omega_X} f(x) dP_X(x)$ defined above is finite, f is called integrable.

Lebesgue Integration: General Functions

For a general measurable function $f : \Omega_X \to \mathbb{R}$, we can consider the following decomposition:

$$f(x) = f^+(x) - f^-(x),$$

where

$$f^+(x) := \max(f(x), 0), \quad f^-(x) := \max(-f(x), 0).$$

- These are both non-negative measurable functions:

$$f^+(x) \geq 0, \quad f^-(x) \geq 0.$$

- Thus we can define their integrals as in the previous slides:

$$\int_{\Omega_X} f^+(x) dP_X(x), \quad \int_{\Omega_X} f^-(x) dP_X(x).$$

- If both integrals are finite (i.e., f^+ and f^- are integrable), then f is called integrable, and the integral is given by

$$\int_{\Omega_X} f(x) dP_X(x) := \int_{\Omega_X} f^+(x) dP_X(x) - \int_{\Omega_X} f^-(x) dP(x).$$

Lebesgue Integration: Vector-valued Functions

- Consider a vector-valued function $\mathbf{f} : \Omega_X \to \mathbb{R}^d$ such that

$$\mathbf{f}(x) := (f_1(x), \ldots, f_d(x))^\top \in \mathbb{R}^d$$

where each $f_i : \Omega_X \to \mathbb{R}$ is measurable.

- Then the integral can be defined as

$$\int_{\Omega_X} \mathbf{f}(x) dP_X(x) := \left(\int_{\Omega_X} f_1(x) dP_X(x), \ldots, \int_{\Omega_X} f_d(x) dP_X(x) \right)^\top \in \mathbb{R}^d.$$

Important Examples: The Mean of a Random Variable

Consider the case $\Omega_X = \mathbb{R}^d$: i.e., X takes values in $\mathbb{R}^d$.

- Define $\mathbf{f} : \Omega_X \to \mathbb{R}^d$ as the identity: $\mathbf{f}(x) = x$.

- Then we can define the expected value (or the mean) of X as

$$\mu_X := \mathbb{E}[X] := \int \mathbf{f}(x) dP_X(x) = \int x \; dP_X(x).$$

- This is the average value that X takes.

Important Examples: The Variance of a Random Variable

- Let $g : \Omega_X \to \mathbb{R}$ be a measurable function.

  - Then $g(X)$ is a random variable.

- Let μ_g be its mean: $\mu_g := \mathbb{E}[g(X)] := \int g(x) \; dP_X(x)$.

- The variance of $g(X)$ can be defined as

$$\begin{aligned} \mathrm{Var}[g(X)] &:= \mathbb{E}[(g(X) - \mu_g)^2] \\ &= \int_{\Omega_X} (g(x) - \mu_g)^2 dP_X(x) = \int_{\Omega_X} f(x) dP_X(x). \end{aligned}$$

where we defined $f : \Omega_X \to \mathbb{R}$ by

$$f(x) := (g(x) - \mu_g)^2.$$

- The variance quantifies how much $g(X)$ may vary around the mean $\mu_g = \mathbb{E}[g(X)]$.

Important Examples: The Variance of a Random Variable

- In particular, for $\Omega_X = \mathbb{R}$, the variance of X is

$$\mathrm{Var}[X] := \mathbb{E}[(X - \mu_X)^2] = \int (x - \mu_X)^2 dP_X(x),$$

where

$$\mu_X := \mathbb{E}[X] := \int_{\Omega_X} x \; dP_X(x).$$

Notation

- I will often write the integral without writing the sample space Ω_X (which is obvious from the context):

$$\int_{\Omega_X} f(x) dP_X(x) =: \int f(x) dP_X(x).$$

- Some people also use the following notation

$$P_X f = \int f \; dP_X = \int_{\Omega_X} f(x) P_X(dx) = \int_{\Omega_X} f(x) dP_X(x)$$

- There are also variations in the notation of the expectation:

$$\mathbb{E}[f(X)] = \mathbb{E}_X[f(X)] = \mathbb{E}_{X \sim P_X}[f(X)] = \mathbb{E}_{P_X}[f(X)] = \int_{\Omega_X} f(x) dP_X(x)$$

- Anyway always pay attention to the definition!

2.5 Probability Density Functions and Dirac Distributions

Probability Density Functions

- Probability density functions are an important concept in probability and statistics.

- People often confuse probability density functions and probability distributions (measures)

- Distributions always exist, but density functions may not.

- An important example is Dirac distributions, another key concept in statistics

— This gives a representation of data.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

Base Measure

Density functions are defined with respect to another measure ν, which is usually called a base measure (or a reference measure).

A set function $\nu : \mathcal{F}_X \to \mathbb{R}$ is called a measure on the measurable space $(\Omega_X, \mathcal{F}_X)$, if it satisfies

- $\nu(A) \geq 0$ for all $A \in \mathcal{F}_X$.
- $\nu(\phi) = 0$.
- For any $A_1, A_2, \cdots \in \mathcal{F}_X$ with $A_i \cap A_j = \phi$ with $i \neq j$,

$$\nu\left(\bigcup_i A_i\right) = \sum_i \nu(A_i).$$

Note that ν is a probability measure if it in addition satisfies

$$\nu(\Omega_X) = 1.$$

But in general, we may have $\nu(\Omega_X) > 1$ or even $\nu(\Omega) = \infty$.

Base Measure

For $\Omega_X \subset \mathbb{R}^d$, a standard choice is ν being the Lebesgue measure:

- For any rectangle

$$A := [a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d, \quad -\infty < a_i < b_i < \infty,$$

the Lebesgue measure outputs its "volume":

$$\nu(A) = \prod_{i=1}^{n} (b_i - a_i).$$

For simplicity, we often write an integral in the following way, when ν is the Lebesgue measure:

$$\int f(x) d\nu(x) = \int f(x) dx$$

Probability Density Function

-We say that probability measure P_X (or random variable X) has a probability density function $p_X : \Omega_X \to [0, \infty)$ with respect to the base measure ν, if

$$P_X(A) = \int_A p_X(x) d\nu(x) := \int 1_A(x) p_X(x) d\nu(x), \quad \forall A \in \mathcal{F}_X,$$

where 1_A is the indicator function of A:

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

By taking $A := \Omega_X$, it follows that

$$\int_{\Omega_X} p_X(x) d\nu(x) = P_X(\Omega_X) = 1.$$

Example: Gaussian distributions and Gaussian densities

We consider a Gaussian random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$.

- The sample space: $\Omega = \Omega_X = \mathbb{R}$ (the real line).

- The σ-algebra: $\mathcal{F} = \mathcal{F}_X =$ the Borel-σ algebra (i.e., the smallest σ-algebra containing all open subsets of $\mathbb{R}$).

- Random variable $X : \Omega \to \Omega_X$ is the identity, i.e., $X(\omega) = \omega$.

- The Gaussian distribution $P = P_X$ is given by, for all $S \in \mathcal{F}_X$,

$$P(S) = P_X(S) = \int_S p_{\mu,\sigma^2}(x)dx, \quad S \in \mathcal{F}_X$$

where $p_{\mu,\sigma^2} : \mathbb{R} \to [0, \infty)$ is the Gaussian density

$$p_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Probability Distributions and Density Functions

Note that the probability distribution (measure) P_X and the probability density function p_X are **different**!

- Probability distribution (measure) P_X: a function that maps a measurable set to a value in $[0, 1]$:

$$P_X : S \in \mathcal{F}_X \to P_X(S) \in [0, 1].$$

- Probability density function p_X: a function that maps a sample point $x \in \Omega_X$ to a value in $[0, \infty)$.

$$p_X : x \in \Omega_X \to p_X(x) \in \mathbb{R}.$$

Not all probability distributions have density functions.

- Representative examples include Dirac distributions.

Probability Distributions and Density Functions

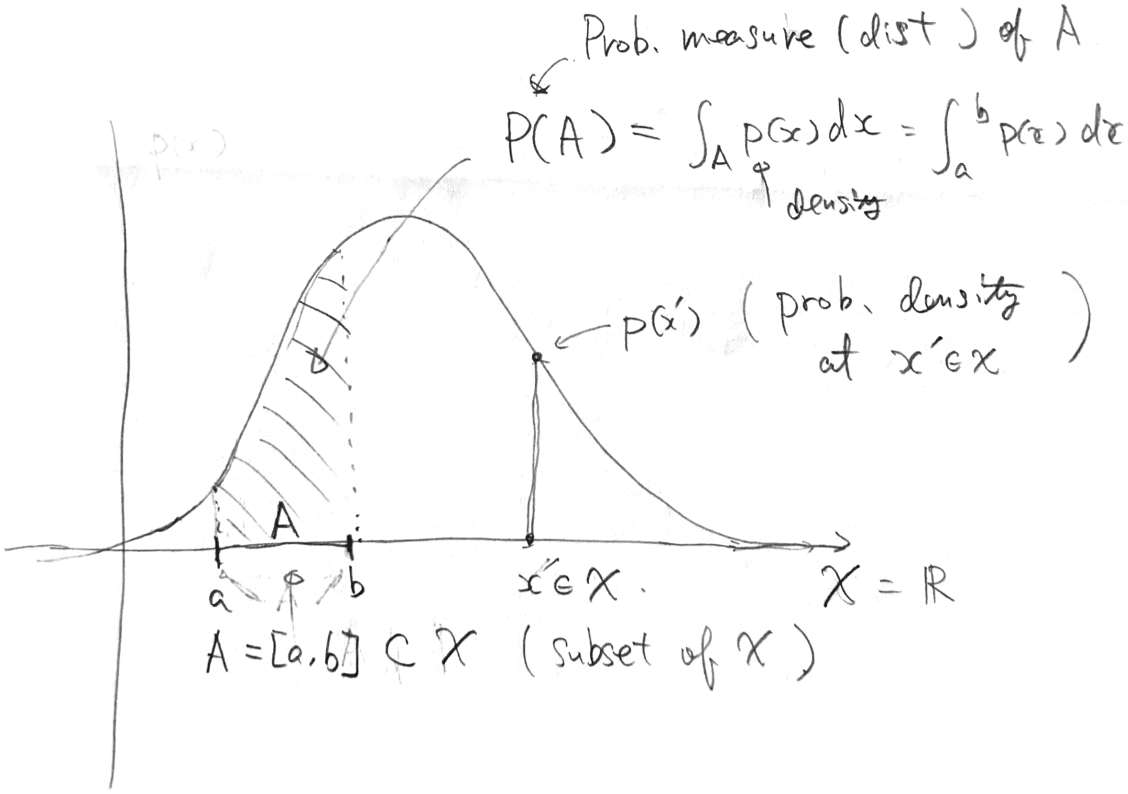
Dirac Distributions

- For any $z \in \Omega_X$, the Dirac distribution (or Dirac measure) at z, denoted by $\delta_z : \mathcal{F}_X \to \mathbb{R}$ is defined as

$$\delta_z(A) = \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{if } z \notin A. \end{cases}, \quad A \in \mathcal{F}_X.$$

- This is the distribution of a random variable $X : \Omega \to \Omega_X$ such that

$$X(\omega) := z, \quad \forall \omega \in \Omega.$$


Prob. measure (dist) of A
$P(A) = \int_A p(x)\,dx = \int_a^b p(x)\,dx$
density
$p(x')$ (prob. density at $x' \in X$)
A
a
b
$x' \in X$.
$X = \mathbb{R}$
$A = [a,b] \subset X$ (subset of X)

- i.e., X takes only the value z, whatever the sample ω is.

This can be seen as follows.

Dirac Distributions

Since $X(\omega) = z$ for all $\omega \in \Omega$, for any $A \in \mathcal{F}_X$ we have

$$\begin{aligned} X^{-1}(A) &:= \{\omega \in \Omega \mid X(\omega)(:= z) \in A\} \\ &= \begin{cases} \Omega & \text{if } z \in A \\ \phi & \text{if } z \notin A. \end{cases} \end{aligned}$$

Therefore,

$$\delta_z(A) := P_X(A) := P(X^{-1}(A)) = \begin{cases} P(\Omega) = 1 & \text{if } z \in A \\ P(\phi) = 0 & \text{if } z \notin A. \end{cases}$$

Dirac Distributions

- For any measurable function $f : \Omega_X \to \mathbb{R}$, the expected value of $f(X)$ with $X \sim \delta_z$ is given by

$$\mathbb{E}_{X \sim \delta_x}[f(X)] = \int_{\Omega_X} f(x) d\delta_z(x) = f(z).$$

- This is intuitively obvious, because $X(\omega) = z$ for all $\omega \in \Omega$.

For instance, assume that f is a simple function $f(x) = \sum_i a_i 1_{S_i}(x)$ with disjoint subsets $S_i \in \mathcal{F}_X$. Then:

$$\int f(x) d\delta_z(x) = \sum_i a_i \delta_z(S_i) = \begin{cases} a_j = f(z) & \text{if } z \in S_j \text{ for some } j \\ 0 & \text{otherwise .} \end{cases}$$

Exercise: Prove $\int f(x) d\delta_z(x) = f(z)$ for a general measurable function f. (Recall the definition of the Lebesgue integral)

Dirac Distributions do not Have Density Functions with respect to the Lebesgue Measure.

For instance, assume $\Omega_X = \mathbb{R}$ and let $z = 0$.

Assume that the Dirac distribution δ_z has a probability density function $p_z(x)$. (We'll show a contradiction)

Then for any $a > 0$, we have $z := 0 \in [-a, a]$, and thus

$$1 = \delta_z([-a, a]) = \int_{\Omega_X} 1_{[-a,a]}(x) p_z(x) d\nu(z) \leq 2a \max_{-a<x<a} p_z(x)$$

Therefore,

$$\frac{1}{2a} \leq \max_{-a<x<a} p_z(x).$$

This holds for all $a > 0$. Thus,

$$\infty = \lim_{a \to +0} \frac{1}{2a} \leq \lim_{a \to +0} \max_{-a < x < a} p_z(x).$$

Therefore, p_z is diverging at 0, which is a contradiction.

Another Example where Densities do not Exist

Define $\Omega := \Omega_X := [0,1]^2 \subset \mathbb{R}^2$.

Assume P is the uniform distribution on $[0,1]^2$.

Define $X : \Omega \to \Omega_X$ by

$$X(\omega) = (\omega_1, 1/2), \quad \omega := (\omega_1, \omega_2) \in \Omega.$$

Then the distribution P_X of X does not have a density function with respect to the Lebesgue measure.

$$P_X([a_1, b_1] \times [a_2, b_2]) = \begin{cases} b_1 - a_1 & \text{if } 1/2 \in [a_2, b_2] \\ 0 & \text{otherwise .} \end{cases}$$

2.6 Joint Random Variable and Joint Distribution

Dealing with Several Random Variables

- So far we have considered only one random variable $X : \Omega \to \Omega_X$.

- There might be another random variable, say $Y : \Omega \to \Omega_Y$ with sample space Ω_Y.

- By sharing the common probability space $(\Omega, \mathcal{F}, P)$, these random variables may be related to each other.

For instance:

- X may be whether or not it will rain tomorrow.

- Y may be whether or not your flight will be delayed tomorrow.

Dealing with Several Random Variables

Here we look at

- how to model several random variables and their joint probability distribution.

- how to quantify the degree of relatedness (independence, covariance etc.)

Joint Random Variable: The Sample Space

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with the associated probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $Y : \Omega \to \Omega_Y$ be a random variable, with the associated probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$.

Define a joint sample space as the product set

$$\Omega_X \times \Omega_Y := \{(\omega_1, \omega_2) \mid \omega_1 \in \Omega_X,\ \omega_2 \in \Omega_Y\}.$$

Joint Random Variable: The σ-algebra

Define $\mathcal{F}_X \otimes \mathcal{F}_Y \subset 2^{\Omega_X \times \Omega_Y}$ as the product σ-algebra, i.e., the smallest σ-algebra containing all subsets ("rectangles") of the form

$$\begin{aligned} S &= A \times B \\ &= \{(\omega_1, \omega_2) \in \Omega_X \times \Omega_Y \mid \omega_1 \in A,\ \omega_2 \in B\}, \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y. \end{aligned}$$

- Note that each $A \in \mathcal{F}_X$ is a certain statement for which a probability can be defined;

- Similarly, each $B \in \mathcal{F}_Y$ is a certain statement for which a probability can be defined.

- The measurable set $A \times B \in \mathcal{F}_X \otimes \mathcal{F}_Y$ is thus a combined statement that both A and B are true, for which we define a probability.

Joint Random Variable: The Definition

We define the joint random variable of X and Y as a mapping $(X, Y) : \Omega \to \Omega_X \times \Omega_Y$ by

$$(X, Y)(\omega) := (X(\omega), Y(\omega)) \in \Omega_X \times \Omega_Y, \quad \omega \in \Omega.$$

The inverse map $(X, Y)^{-1} : \mathcal{F}_{(X,Y)} \to \Omega$ is defined by

$$(X, Y)^{-1}(S) := \{\omega \in \Omega \mid (X(\omega), Y(\omega)) \in S\}, \quad S \in \mathcal{F}_{(X,Y)}.$$

In particular, for a set of the form $S = A \times B$,

$$\begin{aligned} &(X, Y)^{-1}(A \times B) := \{\omega \in \Omega \mid X(\omega) \in A \text{ and } Y(\omega) \in B\} \\ &= X^{-1}(A) \cap Y^{-1}(B), \quad A \in \mathcal{F}_X,\ B \in \mathcal{F}_Y \end{aligned}$$

- i.e., the inverse map $(X, Y)^{-1}(A \times B)$ is a subset in Ω for which both $X(\omega) \in A$ and $Y(\omega) \in B$ hold.

- Intuitively, $(X, Y)^{-1}(A \times B) \in \mathcal{F}$ is the statement that A and B are both true.

Joint Random Variable: The Joint Distribution

- $P_{(X,Y)}$: Joint distribution defined as

$$P_{(X,Y)}(S) := P((X, Y)^{-1}(S)), \quad S \in \mathcal{F}_X \otimes \mathcal{F}_Y$$

In particular, for $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$, we have

$$\begin{aligned} P_{(X,Y)}(A \times B) &:= P((X,Y)^{-1}(A \times B)) \\ &= P(\{\omega \in \Omega \mid X(\omega) \in A \text{ and } Y(\omega) \in B\}). \end{aligned}$$

Then the triplet

$$\big(\Omega_X \times \Omega_Y,\ \mathcal{F}_X \otimes \mathcal{F}_Y,\ P_{(X,Y)}\big)$$

is a probability space.

Joint Random Variable: Some Properties

If we take $A = \Omega_X$, then for any $B \in \mathcal{F}_Y$ we have

$$\begin{aligned} (X,Y)^{-1}(\Omega_X \times B) &= X^{-1}(\Omega_X) \cap Y^{-1}(B) \\ &= \Omega \cap Y^{-1}(B) \\ &= Y^{-1}(B). \end{aligned}$$

Therefore,

$$\begin{aligned} P_{(X,Y)}(\Omega_X \times B) &= P((X,Y)^{-1}(\Omega_X \times B))) \\ &= P(Y^{-1}(B)) \\ &= P_Y(B). \end{aligned}$$

Similarly, we have

$$P_{(X,Y)}(A \times \Omega_Y) = P_X(A), \quad \forall A \in \mathcal{F}_X$$

In this context, P_X and P_Y are called marginal distributions of $P_{(X,Y)}$.

Example: A Fair Dice

Let's consider a fair dice.

- **Sample space:** $\Omega = \{1,2,3,4,5,6\}$.

- **σ-algebra:** $\mathcal{F} = 2^\Omega$ (the power set, i.e., the set of all subsets of Ω).

- **Probability:** $P(\{1\}) = P(\{2\}) \cdots = P(\{6\}) = 1/6$.

Define random variables

$X : \Omega \to \Omega_X := \{a, b\}$:

$$X(\omega) := \begin{cases} a & \text{if } \omega \text{ is odd } (i.e., 1,3,5) \\ b & \text{if } \omega \text{ is even } (i.e., 2,4,6) \end{cases}$$

$Y : \Omega \to \Omega_Y := \{c, d\}$:

$$Y(\omega) := \begin{cases} c & \text{if } \omega = 1 \\ d & \text{if } \omega = 2,3,4,5,6 \end{cases}$$

Example: A Fair Dice

The probability distribution P_X of X:

$$P_X(\{a\}) = P(X^{-1}(a)) = P(\{1,3,5\}) = \frac{1}{2},$$
$$P_X(\{b\}) = P(X^{-1}(b)) = P(\{2,4,6\}) = \frac{1}{2}.$$

The probability distribution P_Y of Y:

$$P_Y(\{c\}) = P(Y^{-1}(\{c\})) = P(\{1\}) = \frac{1}{6},$$
$$P_Y(\{d\}) = P(Y^{-1}(\{d\})) = P(\{2,3,4,5,6\}) = \frac{5}{6}.$$

Example: A Fair Dice

The product σ-algebra is given by:

$$\mathcal{F}_X \otimes \mathcal{F}_Y = \{\phi, \{a\}, \{b\}, \{a,b\}\} \times \{\phi, \{c\}, \{d\}, \{c,d\}\}$$

For instance, consider $\{a\} \times \{c\} \in \mathcal{F}_X \otimes \mathcal{F}_Y$. Since

$$X^{-1}(\{a\}) = \{1,3,5\}, \quad Y^{-1}(\{c\}) = \{1\}$$

we have

$$(X,Y)^{-1}(\{a\} \times \{c\}) = X^{-1}(\{a\}) \cap Y^{-1}(\{c\}) = \{1\}.$$

Therefore the joint probability of $\{a\} \times \{c\}$ is

$$P_{(X,Y)}(\{a\} \times \{c\}) = P((X,Y)^{-1}(\{a\} \times \{c\})) = P(\{1\}) = 1/6.$$

Joint Probability Density Function

A related key concept is joint probability density functions.

- Let ν_X be a base measure on $(\Omega_X, \mathcal{F}_X)$.

- Let ν_Y be a base measure on $(\Omega_Y, \mathcal{F}_Y)$.

- Define $\nu_X \otimes \nu_Y$ as the product measure of $\nu_X \otimes \nu_Y$: i.e., a measure on $(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y)$ such that

$$\nu_X \otimes \nu_Y(A \times B) = \nu_X(A)\nu_Y(B), \quad A \in \mathcal{F}_X, \ B \in \mathcal{F}_Y.$$

For instance, assume that

- $\Omega_X = \mathbb{R}^p$ and ν_X is the Lebesgue measure on $\mathbb{R}^p$.

- $\Omega_Y = \mathbb{R}^q$ and ν_Y is the Lebesgue measure on $\mathbb{R}^q$.

Then, $\Omega_X \times \Omega_Y = \mathbb{R}^{p+q}$ and $\nu_X \otimes \nu_Y$ is the Lebesgue measure on $\mathbb{R}^{p+q}$.

Joint Probability Density Function

If the joint distribution $P_{(X,Y)}$ has a probability density function

$$p_{(X,Y)} : \Omega_X \times \Omega_Y \to [0, \infty)$$

with respect to the base measure $\nu_X \otimes \nu_Y$ such that

$$P_{(X,Y)}(S) = \int_S p_{(X,Y)}(x, y) d\nu_X \otimes \nu_Y(x, y), \quad \forall S \in \mathcal{F}_X \otimes \mathcal{F}_Y,$$

then we call $p_{(X,Y)}$ the joint probability density function of X and Y.

In particular, for $S = A \times B$ with $A \in \mathcal{F}_X$ and $B \in \mathcal{F}_Y$, the joint density function satisfies

$$P_{(X,Y)}(A \times B) = \int_B \int_A p_{(X,Y)}(x, y) d\nu_X(x) d\nu_Y(y),$$

Joint Probability Density Function

We look at some important properties of the joint density function.

- Assume that P_X has a density function $p_X : \Omega_X \to \mathbb{R}$ with respect to the base measure ν_X.

- Then we have

$$p_X(x) = \int_{\Omega_Y} p_{(X,Y)}(x, y) dP_Y(y), \quad x \in \Omega_X.$$

This operation is called the marginalization of Y, or the sum rule.

- This can be shown as follows.

Joint Probability Density Function

- Define $f(x) := \int_{\Omega_Y} p_{(X,Y)}(x, y) d\nu_Y(y)$.

- Then for any $A \in \mathcal{F}_X$, this function satisfies

$$\begin{aligned} &\int_A f(x) d\nu_X(x) = \int_A \int_{\Omega_Y} p_{(X,Y)}(x, y) d\nu_Y(y) d\nu_X(x) \\ &= P_{(X,Y)}(A \times \Omega_Y) = P_X(A). \end{aligned}$$

- Thus, f defined here satisfies the definition of a density function of P_X.

Similarly, we have

$$p_Y(y) = \int_{\Omega_X} p_{(X,Y)}(x, y) dP_X(x), \quad y \in \Omega_Y.$$

2.7 Conditional Probabilities and Conditional Distributions

Conditional Probabilities and Conditional Distributions

Another important concept is conditional probabilities and conditional distributions.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$

- Let $Y : \Omega \to \Omega_Y$ be a random variable with probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$

Conditional Probabilities

- Take a measurable set $A \in \mathcal{F}_X$, which is a certain statement for which a probability $P_X(A)$ is defined.

- Take a measurable set $B \in \mathcal{F}_Y$, which is another statement regarding the random variable Y.

We are interested in the probability of B being true, given that the statement A is true.

This is the conditional probability of B given A, which we write

$$P_{Y|X}(B|A), \quad A \in \mathcal{F}_X, \ B \in \mathcal{F}_Y.$$

Conditional Probabilities

Statement A can be expressed in the probability space $(\Omega, \mathcal{F}, P)$ as

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}.$$

- Thus, "A is true" can be interpreted as "$X^{-1}(A)$ is true" in the original probability space $(\Omega, \mathcal{F}, P)$.

- Therefore, the condition that "A is true" can be formulated as the restriction of the probability space $(\Omega, \mathcal{F}, P)$ onto $X^{-1}(A) \in \mathcal{F}$.

Conditional Probabilities

- i.e., we consider the restricted probability space

$$(\Omega_{|X^{-1}(A)}, \ \mathcal{F}_{|X^{-1}(A)}, \ P_{|X^{-1}(A)}),$$

defined with

- Restricted sample space: $\Omega_{|X^{-1}(A)} := X^{-1}(A)$;

- Restricted σ-algebra

$$\mathcal{F}_{|X^{-1}(A)} := \{S \cap X^{-1}(A) \mid S \in \mathcal{F}\} \subset \mathcal{F}.$$

- Restricted probability measure:

$$P_{|X^{-1}(A)}(C) := \frac{P(C)}{P(X^{-1}(A))}, \quad C \in \mathcal{F}_{|X^{-1}(A)}.$$

Conditional Probabilities

- Here, we assumed that $P_X(A) = P(X^{-1}(A)) > 0$, i.e., the statement A has a non-zero probability.

- The division by $P(X^{-1}(A))$ is needed to ensure

$$P_{|X^{-1}(A)}(X^{-1}(A)) = \frac{P(X^{-1}(A))}{P(X^{-1}(A))} = 1.$$

- We can check that $(\Omega_{|X^{-1}(A)}, \mathcal{F}_{|X^{-1}(A)}, P_{|X^{-1}(A)})$ satisfies the definition of a probability space (exercise).

Conditional Probabilities

- Take a statement $B \in \mathcal{F}_Y$, which is expressed as $Y^{-1}(B) \in \mathcal{F}$ in the original probability space $(\Omega, \mathcal{F}, P)$.

- In the restricted probability space $(\Omega_{|X^{-1}(A)}, \mathcal{F}_{|X^{-1}(A)}, P_{|X^{-1}(A)})$, $Y^{-1}(B)$ is expressed as

$$Y^{-1}(B) \cap X^{-1}(A) \in \mathcal{F}_{|X^{-1}(A)}.$$

- Thus, the conditional probability of B given A is defined by

$$\begin{aligned} P_{Y|X}(B|A) &:= P_{|X^{-1}(A)}(Y^{-1}(B) \cap X^{-1}(A)) \\ &= \frac{P(Y^{-1}(B) \cap X^{-1}(A))}{P(X^{-1}(A))} \\ &= \frac{P_{(X,Y)}(A \times B)}{P_X(A)}. \end{aligned}$$

Example: A Fair Dice

Let's consider a fair dice.

- **Sample space:** $\Omega = \{1, 2, 3, 4, 5, 6\}$.

- **σ-algebra:** $\mathcal{F} = 2^\Omega$ (the power set, i.e., the set of all subsets of Ω).

- **Probability:** $P(\{1\}) = P(\{2\}) \cdots = P(\{6\}) = 1/6$.

Define random variables

$X : \Omega \to \Omega_X := \{a, b\}$: $\quad (\mathcal{F}_X := 2^{\{a,b\}})$

$$X(\omega) := \begin{cases} a & \text{if } \omega \text{ is odd } (i.e., 1, 3, 5) \\ b & \text{if } \omega \text{ is even } (i.e., 2, 4, 6) \end{cases}$$

$Y : \Omega \to \Omega_Y := \{c, d\}$: $\quad (\mathcal{F}_Y := 2^{\{c,d\}})$

$$Y(\omega) := \begin{cases} c & \text{if } \omega = 1 \\ d & \text{if } \omega = 2, 3, 4, 5, 6 \end{cases}$$

Example: A Fair Dice

Let's consider conditioning with $A := \{a\} \in \mathcal{F}_X$.

Then

$$X^{-1}(\{a\}) = \{1, 3, 5\}, \quad P(X^{-1}(\{a\})) = 1/2.$$

The restricted sample space is

$$\Omega_{X^{-1}(\{a\})} = X^{-1}(\{a\}) = \{1, 3, 5\}.$$

The restricted σ-algebra is

$$\mathcal{F}_{X^{-1}(\{a\})} = \{S \cap \{1, 3, 5\} \mid S \in 2^{\{1,2,3,4,5,6\}}\} = 2^{\{1,3,5\}}.$$

The restricted probability measure is

$$P_{X^{-1}(\{a\})} := \frac{P(C)}{P(\{1, 3, 5\})}, \quad C \in \mathcal{F}_{X^{-1}(\{a\})}.$$

Example: A Fair Dice

Therefore, the conditional probability of $B \in \mathcal{F}_Y$ given $\{a\}$ is

$$P_{Y|X}(B \mid \{a\}) = \frac{P(Y^{-1}(B) \cap \{1, 3, 5\})}{P(\{1, 3, 5\})}.$$

For instance, since $Y^{-1}(\{c\}) = \{1\}$,

$$P_{Y|X}(\{c\} \mid \{a\}) = \frac{P(\{1\} \cap \{1, 3, 5\})}{P(\{1, 3, 5\})} = \frac{P(\{1\})}{P(\{1, 3, 5\})} = 1/3.$$

Similarly, since $Y^{-1}(\{d\}) = \{2, 3, 4, 5, 6\}$,

$$P_{Y|X}(\{d\} \mid \{a\}) = \frac{P(\{2, 3, 4, 5, 6\} \cap \{1, 3, 5\})}{P(\{1, 3, 5\})} = \frac{P(\{3, 5\})}{P(\{1, 3, 5\})} = 2/3.$$

Conditional Distributions

By construction,

$$P_{Y|X}(\cdot \mid A)$$

defines the probability distribution (measure) on $(\Omega_Y, \mathcal{F}_Y)$.

- This is the conditional distribution of Y given the statement A about X.

- In the case $A = \Omega_X$, $P_{Y|X}(\cdot \mid \Omega_X)$ is the marginal distribution of Y, i.e., P_Y:

In fact, since $X^{-1}(\Omega_X) = \Omega$

$$P_{Y|X}(B|\Omega_X) = \frac{P(Y^{-1}(B) \cap X^{-1}(\Omega_X))}{P(X^{-1}(\Omega_X))}$$
$$= \frac{P(Y^{-1}(B) \cap \Omega)}{P(\Omega)} = P(Y^{-1}(B)) = P_Y(B).$$

Conditional Distributions

- In the case where A is a singleton set, i.e.,

$$A = \{x\} \text{ for some } x \in \Omega_X,$$

then $P_{Y|X}(\cdot|\{x\})$ is usually called "the conditional distribution of Y given $X = x$."

- In this case we implicitly assume $P_X(\{x\}) > 0$, since otherwise the division

$$P_{Y|X}(B \mid \{x\}) = \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})}$$

is not well-defined.

- However, $P_X(\{x\}) = 0$ occurs in many important settings, in particular when X is a continuous random variable.

Conditional Distributions: Abstract Definition

- Thus, the conditional distribution of Y given $X = x$ is defined in a rather abstract way, as follows.

- For each $x \in \Omega_X$, let $P_{Y|X=x}$ be a probability distribution on $(\Omega_Y, \mathcal{F}_Y)$.

- Assume that for any $B \in \mathcal{F}_Y$, a function $f_B : \Omega_X \to \mathbb{R}$ defined by

$$f_B(x) := P_{Y|X=x}(B), \quad x \in \Omega_X$$

is a measurable function.

- Then $P_{Y|X=x}$ is called the conditional distribution of Y given $X = x$, if it satisfies

$$P_Y(B) = \int_{\Omega_X} P_{Y|X=x}(B) dP_X(x), \quad \forall B \in \mathcal{F}_Y.$$

Conditional Distributions: Abstract Definition

- For instance, assume that X is a discrete random variable and that $P_X(\{x\}) > 0$ for all $x \in \Omega_X$.

- In this case, the conditional probability of $B \in \mathcal{F}_Y$ given $\{x\} \in \mathcal{F}_X$ is given by

$$P_{Y|X}(B \mid \{x\}) = \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})}.$$

- This is consistent with the above abstract definition, since

$$\begin{aligned}
&\int_{\Omega_X} \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})} dP_X(x) \\
&= \sum_{x \in \Omega_X} \frac{P_{(X,Y)}(\{x\} \times B)}{P_X(\{x\})} P_X(\{x\}) \\
&= \sum_{x \in \Omega_X} P_{(X,Y)}(\{x\} \times B) = P_{(X,Y)}(\Omega_X, B) = P_Y(B).
\end{aligned}$$

Conditional Probability Density Functions

Assume that

- the joint distribution $P_{(X,Y)}$ has a density function

$$p_{(X,Y)} : \Omega_X \times \Omega_Y \to [0, \infty)$$

with respect to the base measure $\nu_X \otimes \nu_Y$:

$$\begin{aligned}
P_{(X,Y)}(A \times B) &= \int_{A \times B} p_{(X,Y)}(x, y) d\nu_X \otimes \nu_Y(x, y) \\
&= \int_B \int_A p_{(X,Y)}(x, y) d\nu_X(x) d\nu_Y(y)
\end{aligned}$$

- the marginal distribution P_X has a density function

$$p_X : \Omega_X \to [0, \infty)$$

such that

$$P_X(A) = \int_A p_X(x) d\nu_X(x)$$

Conditional Probability Density Functions

Then, the conditional probability density function

$$p_{Y|X} : \Omega_X \times \Omega_Y \to [0, \infty)$$

is defined by

$$p_{Y|X}(y \mid x) := \frac{p_{(X,Y)}(x, y)}{p_X(x)}, \quad x \in \Omega_X,\ y \in \Omega_Y,$$

assuming $p_X(x) > 0$.

- For each $x \in \Omega_X$, $p_{Y|X}(\cdot \mid x)$ is a probability density function on Ω_Y: In fact,

$$\int_{\Omega_Y} p_{Y|X}(y \mid x) d\nu_Y(y) = \int_{\Omega_Y} \frac{p_{(X,Y)}(x,y)}{p_X(x)} d\nu_Y(y)$$
$$= \frac{1}{p_X(x)} \int_{\Omega_Y} p_{(X,Y)}(x,y) d\nu_Y(y) = \frac{p_X(x)}{p_X(x)} = 1,$$

where we used the sum rule.

Conditional Probability Density Functions

Conditional density function $p_{Y|X}(\cdot \mid x)$ is consistent with the abstract definition of conditional distributions.

To see this, let

$$P_{Y|X=x}(B) := \int_B p_{Y|X}(y \mid x) d\nu_Y(y), \quad B \in \mathcal{F}_Y.$$

Then

$$\int_{\Omega_X} P_{Y|X=x}(B) dP_X(x) = \int_{\Omega_X} \int_B p_{Y|X}(y \mid x) d\nu_Y(y) dP_X(x)$$
$$= \int_{\Omega_X} \int_B \frac{p_{(X,Y)}(x,y)}{p_X(x)} d\nu_Y(y) dP_X(x)$$
$$= \int_B \int_{\Omega_X} \frac{p_{(X,Y)}(x,y)}{p_X(x)} p_X(x) d\nu_X(x) d\nu_Y(y)$$
$$\int_B \int_{\Omega_X} p_{(X,Y)}(x,y) d\nu_X(x) d\nu_Y(y) = \int_B p_Y(y) d\nu_Y(y) = P_Y(B),$$

where we used the sum rule.

2.8 Independence of Random Variables

Independence of Random Variables

The independence of random variables is another key concept in probability and statistics.

This characterizes "unrelatedness" of two random variables.

- Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- Let $X : \Omega \to \Omega_X$ be a random variable, with the associated probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

- Let $Y : \Omega \to \Omega_Y$ be a random variable, with the associated probability space $(\Omega_Y, \mathcal{F}_Y, P_Y)$.

- Let

$$\left(\Omega_X \times \Omega_Y, \mathcal{F}_X \otimes \mathcal{F}_Y, P_{(X,Y)}\right)$$

be the joint probability space.

Independence of Random Variables

Definition. Random variables X and Y are called independent if

$$P_{(X,Y)}(A \times B) = P_X(A)P_Y(B), \quad \forall A \in \mathcal{F}_X,\ \forall B \in \mathcal{F}_Y.$$

In other words, X and Y are independent, if the joint distribution $P_{(X,Y)}$ is equal to the product measure, i.e.,

$$P_{(X,Y)} = P_X \otimes P_Y,$$

where $P_X \otimes P_Y$ is the product measure such that

$$P_X \otimes P_Y(A \times B) = P_X(A)P_Y(B).$$

Independence of Random Variables

- If there exists a joint density function

$$p_{X,Y} : \Omega_X \times \Omega_Y \to [0, \infty),$$

then X and Y are independent if

$$p_{(X,Y)}(x, y) = p_X(x)p_Y(y), \quad \forall x \in \Omega_X,\ \forall y \in \Omega_Y,$$

where

- $p_X : \Omega_X \to [0, \infty)$ is the density function of X

- $p_Y : \Omega_Y \to [0, \infty)$ is the density function of Y

Exercise:

- Show that the above characterization leads to the definition of independence.

Independence of Random Variables: An Interpretation

The independence of X and Y implies that X does not have any information about Y (and vice versa).

- In fact, the independence implies that the conditional probability $P_{Y|X}(A|B)$ is equal to the marginal $P_Y(B)$:

$$P_{Y|X}(B|A) := \frac{P_{(X,Y)}(A \times B)}{P_X(A)} = \frac{P_X(A)P_Y(B)}{P_X(A)} = P_Y(B).$$

- Similarly, if there exist density functions, the conditional density function $p_{Y|X}(y|x)$ equals the marginal density function p_Y(y):

$$p_{Y|X}(y|x) = \frac{p_{(X,Y)}(x, y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(x)} = p_Y(y).$$

Intuitively, this means that the conditioning $X = x$ does not affect the distribution of Y.

Consequences of Independence

Let $f : \Omega_X \to \mathbb{R}$ and $g : \Omega_Y \to \mathbb{R}$ be any measurable functions.

Then, if X and Y are independent, we have

$$\mathbb{E}_{X,Y}[f(X)g(Y)] = \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)]$$

This is because, since $P_{(X,Y)} = P_X \otimes P_Y$ by the independence,

$$\begin{aligned}\mathbb{E}_{(X,Y)}[f(X)g(Y)] &:= \int_{\Omega_X \times \Omega_Y} f(x)g(y)dP_{(X,Y)}(x,y) \\ &= \int_{\Omega_X}\int_{\Omega_Y} f(x)g(y)dP_Y(y)dP_X(x) \\ &= \int_{\Omega_X} f(x)dP_X(x)\int_{\Omega_Y} g(y)dP_Y(x) \\ &= \mathbb{E}_X[f(X)]\mathbb{E}_Y[g(Y)].\end{aligned}$$

Independently and Identically Distributed (i.i.d.)

The i.i.d. is a very important concept, ubiquitous in statistics and machine learning.

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Let $X : \Omega_X \to \Omega$ be a random variable with probability space $(\Omega_X, \mathcal{F}_X, P_X)$.

Consider n random variables $X_1, X_2, \ldots, X_n$ such that

- $X_i : \Omega \to \Omega_{X_i}$ is a random variable with probability space $(\Omega_{X_i}, \mathcal{F}_{X_i}, P_{X_i})$ $(i = 1, \ldots, n)$.

Independently and Identically Distributed (i.i.d.)

Definition. Random variables $X_1, X_2, \ldots, X_n$ independently and identically distributed (i.i.d.) with X (or with P_X), if they satisfy the following:

– $(\Omega_{X_i}, \mathcal{F}_{X_i}, P_{X_i}) = (\Omega_X, \mathcal{F}_X, P_X)$ $(i = 1, \ldots, n)$.

- i.e., X_i is identically distributed with X.

– X_i and X are independent $(i = 1, \ldots, n)$.

- i.e., X_i is independently distributed with X.

– X_i and X_j are independent for $i \neq j$.

- i.e., X_i is independently (and identically) distributed with X_j.

We often write as $X_1, \ldots, X_n \sim P$ (i.i.d.).

2.9 Important Points to Remember

Some Important Points to Remember

When talking about random variables, people often omit mentioning the underlying probability space $(\Omega, \mathcal{F}, P)$.

- Often, we say something like "Let X be a random variable taking values in Ω_X".

- But remember that there is always such a probability space.

In this lecture, I will use a different notation like $\mathcal{X}$ in place of Ω_X.

- So, "Let X be a random variable taking values in $\mathcal{X}$"

should be understood as

"Let $(P, \mathcal{F}, \Omega)$ be a probability space, and let $X : \Omega \to \Omega_X$ be a random variable with probability space $(P_X, \mathcal{F}_X, \Omega_X)$ with $\mathcal{X} := \Omega_X$."

Further Reading

If you are interested in more detail of probability theory, you may look at the books in the references.

Specifically:

- (Dudley, 2002, Chapters 3, 4, 8, 10).

- (Rao, 1973, Chapter 2).

Chapter 3

Introduction to Estimation Theory

3.1 Mean Estimation Problem and Motivations

Estimation of the Mean. Let X be a random variable taking values in $\mathbb{R}$ with probability distribution P.[1] The mean (or the expected value) of X is defined by

$$\mu := \mathbb{E}_{X\sim P}[X] = \int x \; dP(x) \in \mathbb{R}.$$

Assume that we don't know P, and thus we don't know μ.

Assume instead that we are given some **data**:

$$X_1, \ldots, X_n \in \mathbb{R}$$

These are assumed to be random variables taking values in $\mathbb{R}$.

The task of mean estimation is estimating the unknown mean μ from the data $X_1, \ldots, X_n$. This is one of the most ubiquitous and fundamental problems in statistics. In this chapter, we look at this problem in detail.

Motivation 1: Relation to Many Problems. Many problems can be formulated as estimation of the mean. Examples include:

- Monte Carlo: Simulation-based mean estimation.
- Design of experiments: Average treatment (causal) effect.
- Regression: Estimation of the conditional mean.
- Supervised machine learning:

[1] In the previous chapter, P is used to denote the distribution of the underlying probability space, but here P denotes the distribution of X.

- Risk = the mean of a loss function.
- Stochastic gradient = approximation of the expected gradient.

Motivation 2: Different Statistical Approaches. Mean estimation can be used to illustrate different approaches.

- The "frequentist" approach - maximum likelihood estimation.
- The "Bayesian" approach - posterior inference.
- The "empirical Bayes" - the mixed approach.

Motivation 3: Key Notions We can learn key notions in statistics.

- Estimator and consistency.
- Bias-variance decomposition/trade-off
- Law of large numbers and the central limit theorem.

Most importantly, the key is how data are generated/obtained.

Is the Empirical Average a Good Approach? A standard approach is to take the empirical average of data points:

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In this chapter, we will address questions like:

- When is the empirical average a good estimate, and when is it not?
- When can we justify the use of the empirical average?
- What conditions do we need for the data $X_1, \ldots, X_n$?

3.2 The Data Generation Process Matters

Population and Data

In the mean estimation problem, we have two kinds of random variables:

[Population] Random variable X represents the hypothetical population of interest, with P being its probability distribution.

[Data] Random Variables $X_1, \ldots, X_n$ represent the given data.

- The data $X_1, \ldots, X_n$ are assumed to provide information about the population random variable X (or its distribution P).

- Otherwise, we cannot estimate the population mean $\mu = \mathbb{E}_{X \sim P}[X]$ from the data $X_1, \ldots, X_n$.

- Therefore, how the data are generated/obtained becomes very important.

Example: Estimating the Average Income in France

- Assume that $X \in \mathbb{R}$ represents the income of a randomly sampled French person, with P being its distribution.

- The population mean $\mu = \mathbb{E}_{X \sim P}[X]$ represents the average income of French people.

- The data $X_1, \ldots, X_n$ are the incomes of n French people randomly selected from the French population.

- Then, is the empirical average

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

a good estimate of the true average income $\mu = \mathbb{E}_{X \sim p}[X]$?

Example: Estimating the Average Income in France

- Assume that data $X_1, \ldots, X_n$ are the incomes of randomly sampled French persons in French Riviera.

- Then, the empirical average

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

would be higher than the average income of the French population.

Example: Estimating the Average Income in France

- Assume that the data $X_1, \ldots, X_n$ are the incomes of randomly sampled French people between age 20 and 30.

- Then, the empirical average

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

would give an estimate lower than the true average income.

The Data Generating Process Matters

These examples indicate that how the data are generated/obtained strongly affects the validity of the empirical average.

- We need to make sure that data $X_1, \ldots, X_n$ are sampled from the same population as that of the target random variable $X \sim P$.

- This requirement is mathematically formulated by assuming that random variables $X_1, \ldots, X_n$ are independently and identically distributed (i.i.d.) with $X \sim P$.

Independently and Identically Distributed (i.i.d.)

Recall that random variables $X_1, \ldots, X_n$ are i.i.d. with a random variable $X \sim P$ if they satisfy the following:

- Independence:

  - X_i and X_j are independent for all $i \neq j$.
  - X_i and X are independent for all $i = 1, \ldots, n$;
    - Recall that X represents the hypothetical population (e.g., randomly selected French person).

- Identity:

  - X_i follows the same probability distribution P of X (for all $i = 1, \ldots, n$).

We often write $X_1, \ldots, X_n \sim P$ *(i.i.d.)*.

See also the lecture slides on Probability Theory.

3.3 Preliminaries: Key Properties of Expectation and Variance

Preliminaries

Before going further, we collect here some key properties of Expectation and Variance of random variables.

Some Key Properties of Expectation

- For any real-valued random variable X and a constant $c \in \mathbb{R}$, we have

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

- For any real-valued random variables X and Y, we have

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

- If X and Y are independent,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$$

Variance of a Random Variable

In statistics, the variance of a random variable plays a key role.

- Let X be a real-valued random variable with probability distribution P.

Then the variance of X is defined by

$$\mathbb{V}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2] = \int (x - \mathbb{E}[X])^2 dP(x) \geq 0.$$

- Note that the mean $\mathbb{E}[X] \in \mathbb{R}$ is a constant.

Some Key Properties of Variance

Let X be a real-valued random variable.

Then we have

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof:

$$\begin{aligned}
&\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - 2(\mathbb{E}[X])^2 + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

Some Key Properties of Variance

Let X be a real-valued random variable. Then for any constant $c \in \mathbb{R}$, we have

$$\mathbb{V}[cX] = c^2 \mathbb{V}[X].$$

Proof:

$$\begin{aligned}
&\mathbb{V}[cX] := \mathbb{E}[(cX - \mathbb{E}[cX])^2] \\
&= c^2 \mathbb{E}[(X - \mathbb{E}[X])^2] = c^2 \mathbb{V}[X].
\end{aligned}$$

In particular, by setting $c = 1/n$, we have

$$\mathbb{V}\left[\frac{X}{n}\right] = \frac{1}{n^2}\mathbb{V}[X].$$

Some Key Properties of Variance

Let X and Y be real-valued random variables.

If X and Y are independent, then

$$\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y].$$

Proof:

$$\begin{aligned}
&\mathbb{V}[X+Y] := \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X]+Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2 + 2(X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) + (Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])^2] + \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{V}[X] + \mathbb{V}[Y],
\end{aligned}$$

where we used

$$\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[(X-\mathbb{E}[X])]\mathbb{E}[(Y-\mathbb{E}[Y])] = 0,$$

which follows from the independence of X and Y.

Some Key Properties of Variance

By recursive applications of the previous result, we have the following useful result:

Let $X_1, X_2, \ldots, X_n$ are independent real-valued random variables (note: they don't necessary identically distributed).

Then we have

$$\mathbb{V}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}[X_i]$$

Corollary:

- Let $X_1, \ldots, X_n$ be independent real-valued random variables.
- Let $c_1, \ldots, c_n \in \mathbb{R}$ be constants.

Then

$$\mathbb{V}[\sum_{i=1}^{n} c_i X_i] = \sum_{i=1}^{n} \mathbb{V}[c_i X_i] = \sum_{i=1}^{n} c_i^2 \mathbb{V}[X_i].$$

Some Key Properties of Variance

In particular, assuming that $X_1, \ldots, X_n$ are i.i.d. with a random variable X, and setting $c_i := 1/n$, we have

$$\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}[X_i] = \frac{1}{n}\mathbb{V}[X].$$

- Thus, the variance of the empirical average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is n times smaller than the variance of X.

- By taking the average over independent observations, the variance can be reduced.

3.4 Statistical Estimators

Estimators and Estimates

In statistics, the procedure of estimating a quantity of interest is formulated as a function of data.

- This function is called an estimator.

- The output from the estimator is called an estimate.

Estimators and Estimates

- Let $\theta^* \in \Theta$ be an unknown quantity of interest that we want to estimate (Θ is an appropriate set) (θ^* is also called an estimand).

- Assume that we are given some data D_n of size $n \in \mathbb{N}$ of the form

$$D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$$

where each $X_i \in \mathcal{X}$ is a random variable ($\mathcal{X}$ is a measurable space.).

Definition: a map

$$F_n : \mathcal{X}^n \to \Theta$$

is called an estimator (of θ^*).

- The estimator should be designed so that the estimate will be close to θ^*.

- $\hat{\theta}_n := F_n(D_n)$ is called an estimate (of θ^*).

Estimators and Estimates: Mean Estimation

Let's consider the mean estimation problem as an example.

The quantity of interest is the mean of the random variable $X \sim P$:

$$\theta^* := \mu := \mathbb{E}[X] \in \mathbb{R} =: \Theta.$$

Assume that n random variables $X_1, \ldots, X_n$ are given as data:

$$D_n = (X_1, \ldots, X_n) \in \mathcal{X}^n, \quad \mathcal{X} := \mathbb{R}.$$

Then one can define an estimator $F_n : \mathcal{X}^n \to \Theta$ of the mean θ^* by

$$F_n(D_n) := \frac{1}{n} \sum_{i=1}^{n} X_i = \hat{\mu} =: \hat{\theta}.$$

i.e., the empirical average of $X_1, \ldots, X_n$.

Which Estimator Should We Choose?

Note that the empirical average is not the only choice.

For instance, we can define various estimators for the mean estimation problem; e.g.,

1. $F_n(D_n) := (X_1 + \cdots + X_n)/n$.
2. $F_n(D_n) := X_1$ (i.e., discarding $X_2, \ldots, X_n$).
3. $F_n(D_n) := 0$ (i.e., always outputs constant 0, no matter what D_n is).
4. $F_n(D_n) := c_0 + c_1 X_1 + \cdots + c_n X_n$ for some $c_0, c_1, \ldots, c_n \geq 0$.

- Which estimator should we choose?

- When is the empirical average a good choice, and when is it not?

(Actually we'll see that the empirical average is not always a good choice).

Which Estimator Should We Choose?

To investigate these questions, we need to introduce criteria for comparing different estimators.

3.5 Mean Square Error and Bias-Variance Decomposition

Mean Square Error (MSE)

To discuss the quality of a statistical estimator, we need a certain error criterion.

Here we consider the mean square error (MSE), one of the most standard criteria.

- Let $\theta^* \in \Theta \subset \mathbb{R}$ be the unknown quantity of interest.

— We assume $\Theta \subset \mathbb{R}$ for simplicity, but the following argument also holds for more general situations.

- Consider an estimator $F_n : \mathcal{X}^n \to \Theta$ such that

$$\hat{\theta}_n := F_n(D_n) \in \Theta, \quad D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n.$$

- Note that the estimate $\hat{\theta}_n = F_n(D_n) = F_n((X_1, \ldots, X_n))$ is a random variable, since $X_1, \ldots, X_n$ are random variables.

Mean Square Error (MSE)

- Then we can consider the squared error between the target θ^* and estimate $\hat{\theta}_n$:

$$(\hat{\theta}_n - \theta^*)^2 = (F_n(D_n) - \theta^*)^2.$$

- This error is also a random variable, because the estimate $\hat{\theta}_n = F_n(D_n)$ is a random variable.

- Then the mean square error (MSE) of the estimator F_n is defined as the expectation of the squared error:

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(F_n(D_n) - \theta^*)^2]$$

where the expectation is with respect to the data $D_n = (X_1, \ldots, X_n)$.

- The MSE quantifies how the estimate $\hat{\theta}_n$ is close to (or far from) the target θ^* on average.

Mean Square Error (MSE)

Note that the MSE

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(F_n(D_n) - \theta^*)^2]$$

depends on

1. the target quantity θ^*
2. the estimator F_n
3. the distribution of the data $X_1, \ldots, X_n$

By theoretically studying the MSE, we can study

- which estimator F_n is good for estimating the target θ^*,
- when the data $X_1, \ldots, X_n$ are distributed in an assumed way.

Probabilistic Error Bound from MSE

- A general fact: For any non-negative real-valued random variable Z, Markov's inequality states that

$$\Pr(Z \geq c) \leq \frac{\mathbb{E}[Z]}{c}, \quad \forall c > 0.$$

- By setting $Z := (\hat{\theta}_n - \theta^*)^2$, we then have

$$\Pr((\hat{\theta}_n - \theta^*)^2 \geq c) \leq \frac{\mathbb{E}[(\hat{\theta}_n - \theta^*)^2]}{c}, \quad \forall c > 0.$$

- Thus, if the MSE $\mathbb{E}[(\hat{\theta} - \theta^*)^2]$ is small, then the probability of

$$(\hat{\theta}_n - \theta^*)^2 > c$$

becomes small for any $c > 0$.

Bias-Variance Decomposition

- The following is a very important result concerning the MSE.

Theorem: The MSE can be decomposed into the bias and the variance of the estimator, as follows:

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2]}_{Variance} + \underbrace{(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2}_{Bias}$$

This is called the bias-variance decomposition.

- The bias of the estimator $F_n : \mathcal{X}^n \to \Theta$ is defined as the difference between the expectation of the estimate $\mathbb{E}[\hat{\theta}_n]$ and the target θ^*:

$$\mathbb{E}[\hat{\theta}_n] - \theta^* = \mathbb{E}[F_n(D_n)] - \theta^*.$$

where the expectation is with respect to the data $D_n = (X_1, \ldots, X_n)$.

Bias-Variance Decomposition

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \underbrace{\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2]}_{Variance} + \underbrace{(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2}_{Bias}$$

- The variance of the estimator $F_n : \mathcal{X}^n \to \Theta$ is defined as

$$\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] = \mathbb{E}[(F_n(D_n) - \mathbb{E}[F_n(D_n)])^2].$$

- i.e., the average deviation of the estimate $\hat{\theta}_n := F_n(D_n)$ from its mean $\mathbb{E}[\hat{\theta}_n]$.
- Recall again that the estimate $\hat{\theta}_n$ is a random variable.

- To make the mean-square error small, both the bias and variance need to be small!

Proof of Bias-Variance Decomposition

- The mean square error can be expanded as

$$\begin{aligned}
&\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n] + \mathbb{E}[\hat{\theta}_n] - \theta^*)^2] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + \mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2] + 2\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta^*)] \\
=&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] + (\mathbb{E}[\hat{\theta}_n] - \theta^*)^2,
\end{aligned}$$

where the last line follows from $\mathbb{E}[\hat{\theta}_n]$ being a constant:

$$\begin{aligned}
&\mathbb{E}[(\mathbb{E}[\hat{\theta}_n] - \theta^*)^2] = (\mathbb{E}[\hat{\theta}_n] - \theta^*)^2, \\
&\mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])(\mathbb{E}[\hat{\theta}_n] - \theta^*)] = \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])\right](\mathbb{E}[\hat{\theta}_n] - \theta^*) = 0.
\end{aligned}$$

Remarks on the Bias-Variance Decomposition

- The bias-variance decomposition holds under a very generic situation.

- This is because the proof does not require any assumption about the joint distribution of the data $X_1, \dots, X_n$ (essentially).
- The only assumption is that the MSE is finite.

- Thus, for instance, we can consider cases like:

- where $X_1, \dots, X_n$ are not independently distributed
- where $X_1, \dots, X_n$ are not identically distributed.

- By considering a different setting for the distribution of the data $X_1, \dots, X_n$, we can study when a certain estimator is a good choice, when it is not.

- This is done by analyzing the bias and variance of the estimator.

Bias-Variance Decomposition: Multivariate Case

- Let $\theta^* \in \Theta \subset \mathbb{R}^d$ be the quantity of interest.

- Let $\hat{\theta}_n$ be any estimate of θ^* (you can just think of $\hat{\theta}_n$ as a random variable in $\mathbb{R}^d$).

- Define the mean square error by

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2],$$

where $\|\cdot\|$ is the norm of $\mathbb{R}^d$.

Theorem. - Assume that

$$\|\mathbb{E}[\hat{\theta}_n]\| < \infty, \quad \mathbb{E}[\|\hat{\theta}_n\|^2] < \infty.$$

Then the following bias-variance decomposition holds:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{Variance} + \| \underbrace{\theta^* - \mathbb{E}[\hat{\theta}_n]}_{Bias} \|^2$$

Bias-Variance Decomposition: Multivariate Case

Exercise: Prove the above bias-variance decomposition.

Hint: for any $a, b \in \mathbb{R}^d$,

$$\|a - b\|^2 = \langle a - b, a - b \rangle$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^d$.

3.6 Bias-Variance Decomposition in Mean Estimation

Mean Estimation Problem: Setup

- Now consider the mean estimation problem.

- Let $X \sim P$ be the random variable of interest, whose mean

$$\mu_P := \mathbb{E}[X] = \int x \, dP(x)$$

is the estimand.

- To deal with a generic situation, we assume that i.i.d. data $X_1, \ldots, X_n$ are generated from a probability distribution Q, which can be different from P:

$$X_1, \ldots, X_n \sim Q, i.i.d.$$

- Let $Y \sim Q$ be a random variable, with distribution Q;

- Then $X_1, \ldots, X_n$ are i.i.d. with Y.

Bias-Variance Decomposition in Mean Estimation

- Assume that the mean and the variance of $Y \sim Q$ are finite:

$$|\mu_Q| < \infty, \quad \mu_Q := \mathbb{E}_{Y \sim Q}[Y]$$
$$\sigma_Q^2 < \infty, \quad \sigma_Q^2 := \mathbb{V}_{Y \sim Q}[Y] := \mathbb{E}_{Y \sim Q}[(Y - \mu_Q)^2].$$

Theorem: The mean square error of the empirical average estimator

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

is given by

$$\begin{aligned} \mathbb{E}[(\hat{\mu} - \mu_P)^2] &= \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] + (\mathbb{E}[\hat{\mu}] - \mu_P)^2 \\ &= \frac{\sigma_Q^2}{n} + (\mu_Q - \mu_P)^2. \end{aligned}$$

Proof: Bias-Variance Decomposition in Mean Estimation

Proof:

- The first identity follows from the bias-variance decomposition.

- Thus, we show the second identity.

Variance term.

Because $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$, the variance term can be expressed as

$$\begin{aligned} \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] &= \mathbb{V}[\hat{\mu}] = \mathbb{V}[\frac{1}{n} \sum_{i=1}^{n} X_i] \\ &= \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n} X_i] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}[X_i] = \frac{1}{n} \mathbb{V}[Y] = \frac{\sigma_Q^2}{n}. \end{aligned}$$

Proof: Bias-Variance Decomposition in Mean Estimation

Bias term. On the other hand,

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[Y] = \mu_Q.$$

Therefore, the bias term is

$$(\mathbb{E}[\hat{\mu}] - \mu_P)^2 = (\mu_Q - \mu_P)^2.$$

□

Interpretation of the Bias-Variance Decomposition We proved the bias-variance decomposition:

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_Q^2}{n} + (\mu_Q - \mu_P)^2.$$

Let's study what this means.

- The bias of the estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is

$$\mathbb{E}[\hat{\mu}] - \mu_P = \mu_Q - \mu_P$$

i.e., the difference between

- the mean μ_Q of the data distribution Q, and
- the mean μ_P of the target distribution P.

Interpretation of the Bias-Variance Decomposition

Therefore,

- if the data $X_1, \ldots, X_n$ are independently generated from a distribution Q, and
- if the mean μ_Q of Q is different from the mean μ_P of the target random variable $X \sim P$,

then the use of the empirical average

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

causes a non-zero bias, $\mu_Q - \mu_P \neq 0$.

Note that in this case, since $(\mu_Q - \mu_P)^2 > 0$, the mean square error

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = (\mu_Q - \mu_P)^2 + \frac{\sigma_Q^2}{n} \geq (\mu_Q - \mu_P)^2 > 0$$

does not decrease to 0, even when $n \to \infty$.

Interpretation of the Bias-Variance Decomposition

This example shows the importance of the data distribution Q.

- If possible, we should collect data $X_1, \ldots, X_n$ generated from the same distribution P as the target random variable X, i.e., $Q = P$.

- In this case, the bias becomes 0: $(\mu_Q - \mu_P)^2 = 0$, and the MSE is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n},$$

where $\sigma_P^2 = \mathbb{V}[X]$ is the variance of $X \sim P$.

- Thus, the MSE decreases as the sample size n increases.

Interpretation of the Bias-Variance Decomposition

- On the other hand, the variance term

$$\mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \frac{\sigma_Q^2}{n}$$

depends only on the data $X_1, \ldots, X_n$, and not on the target μ_P.

- Therefore, whatever the data distribution Q is, the variance term converges to 0 as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \lim_{n\to\infty} \frac{\sigma_Q^2}{n} = 0.$$

Interpretation of the Bias-Variance Decomposition

- Note that in the derivation of the variance term, we used

$$\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n} X_i].$$

- This follows from the independence between $X_1, \ldots, X_n$. (see pp.21-22)

- Therefore, if the independence between $X_1, \ldots, X_n$ does not hold, the variance may not decrease to 0 (we'll see an example later).

Interpretation of the Bias-Variance Decomposition

- For example, recall the example where $X \sim P$ represents the income of a randomly picked-up French person.

- Assume that data $X_1, \ldots, X_n \sim Q$ (*i.i.d.*) are the incomes of randomly picked-up French persons in French Riviera.

- Then we would have

$$\mu_Q := \mathbb{E}_{Y\sim Q}[Y] > \mathbb{E}_{X\sim P}[X] =: \mu_P$$

i.e., the average income of French Riviera people μ_Q is higher than the average income of the whole population μ_P.

Interpretation of the Bias-Variance Decomposition

- Thus, the empirical average of the data

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

has a non-zero bias:

$$\mathbb{E}[\hat{\mu}] - \mu_P = \mu_Q - \mu_P \neq 0.$$

- Therefore, the MSE of the empirical average

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = (\mu_Q - \mu_P)^2 + \frac{\sigma_Q^2}{n}$$

does not decrease to 0, even when n is very large.

- Thus, we should make sure that data $X_1, \ldots, X_n$ are randomly picked-up from the whole French population. (i.e., $Q = P$).

Mean Estimation in the Multivariate Case

- Let $X \sim P$ be a random vector in $\mathbb{R}^d$. Define

$$\mu_P := \mathbb{E}_{X\sim P}[X] \in \mathbb{R}^d$$

- Let $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$ be random vectors in $\mathbb{R}^d$, and let $Y \sim Q$. Define

$$\mu_Q := \mathbb{E}_{Y\sim Q}[Y] \in \mathbb{R}^d, \quad \sigma_Q^2 := \mathbb{E}_{Y\sim Q}[\|Y - \mu_Q\|^2] \geq 0.$$

Theorem. Assume that

$$\|\mu_P\| < \infty, \quad \|\mu_Q\| < \infty, \quad \sigma_Q^2 < \infty.$$

Then, the empirical average estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies

$$\begin{aligned}\mathbb{E}[\|\hat{\mu} - \mu_P\|^2] &= \mathbb{E}[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2] + \|\mathbb{E}[\hat{\mu}] - \mu_P\|^2 \\ &= \frac{\sigma_Q^2}{n} + \|\mu_Q - \mu_P\|^2.\end{aligned}$$

Exercise. Prove this. (The first identity is the bias-variance decomposition)

How Large should the Sample Size be?

- In the mean estimation problem, when $X_1, \ldots, X_n \sim P$ i.i.d., the MSE is given by

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n}, \quad \sigma_P^2 := \mathbb{V}[X].$$

for the empirical average estimate $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$.

- Assume that one wants to make the MSE small in that sense that

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] \leq \varepsilon^2,$$

for some $\varepsilon > 0$. Then the sample size n should satisfy

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\sigma_P^2}{n} \leq \varepsilon^2$$

or equivalently

$$n \geq \frac{\sigma_P^2}{\varepsilon^2}.$$

How Large should the Sample Size be?

For instance, consider the example of estimating the average income.

- Assume $\mu_P = 2,000$ EUR/month (mean) and $\sigma_P = 500$ (standard deviation).

Then, the sample size n should satisfy

$$n \geq \frac{500^2}{\varepsilon^2}.$$

For instance,

- to achieve the precision of $\varepsilon = 10$, we need $n \geq 2500$.

- to achieve the precision of $\varepsilon = 1$, we need $n \geq 250,000$.

3.7 Consistency and Unbiasedness

Consistency

- Let $\theta^* \in \Theta \subset \mathbb{R}$ be an estimand (i.e., the quantity of interest).

- Let $X_1, \ldots, X_n$ be random variables such that $X_i \in \mathcal{X}$, and define the data as

$$D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$$

- Let $F_n : \mathcal{X}^n \to \mathbb{R}$ be an estimator, and let $\hat{\theta}_n := F_n(D_n)$ be an estimate.

Definition. We call F_n a consistent estimator of θ^*, if the estimate $\hat{\theta_n}$ converges to θ^* as $n \to \infty$ in an appropriate sense, e.g.,

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] \to 0 \quad \text{as } n \to \infty.$$

- The consistency means that, as we have more data $X_1, \ldots, X_n$, the estimate $\hat{\theta}_n$ becomes more accurate (in estimating θ^*).

- Consistency is one of the most important concepts in statistics.

Unbiasedness

Definition. We call F_n an unbiased estimator of θ^*, if the bias is zero for every $n \in \mathbb{N}$, i.e.,

$$\mathbb{E}[F_n(D_n)] - \theta^* = \mathbb{E}[\hat{\theta}_n] - \theta^* = 0, \quad \forall n \in \mathbb{N}.$$

- If this is not satisfied, we call F_n a biased estimator of θ^*.

Unbiasedness

For instance, consider the mean estimation problem.

- If the data $X_1, \ldots, X_n$ are i.i.d. with $X \sim P$, then the empirical average $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$ satisfies

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X] = \mathbb{E}[X] = \mu_P.$$

- So, in this case, the empirical average $\hat{\mu}$ is an unbiased estimator of the mean μ_P.

- If $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$, and if $\mu_Q \neq \mu_P$, then

$$\mathbb{E}[\hat{\mu}] = \mu_Q \neq \mu_P.$$

- So, in this case, the empirical average $\hat{\mu}$ is a biased estimator of the mean μ_P.

Unbiasedness

- If F_n is an unbiased estimator, then the MSE is given by

$$\mathbb{E}[(\hat{\theta}_n - \theta^*)^2] = \mathbb{E}[(\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n])^2] = \mathbb{V}[\hat{\theta}_n]$$

i.e., the MSE is equal to the variance of the estimate $\hat{\theta}_n$.

Some importance consequences of unbiasedness:

- If the variance $\mathbb{V}[\hat{\theta}_n]$ decreases to 0 as $n \to \infty$, then $\hat{\theta}_n$ converges to θ^*; thus F_n becomes a consistent estimator.

- If we can estimate the variance $\mathbb{V}[\hat{\theta}_n]$, then we can estimate the amount of error (MSE):

  - In other words, we can estimate how far the estimate $\hat{\theta}_n$ is from the target θ^*.

- Thus, an estimate of the variance $\mathbb{V}[\hat{\theta}_n]$ can be used for constructing a confidence interval for θ^* (not covered in the course).

Unbiasedness and Consistency

Note that

- the unbiasedness does not imply the consistency;

- An unbiased estimator can be inconsistent.

- the consistency does not require the unbiasedness;

- A biased estimator can be consistent (we'll see this later).

Example of an Unbiased Estimator that is not Consistent

Consider the mean estimation problem.

- Let $X \sim P$, and assume that $X_1, \ldots, X_n \sim P$ $(i.i.d.)$.

- Define an estimator F_n by

$$\hat{\mu} := F_n(X_1, \ldots, X_n) := X_1.$$

i.e., we only use X_1, and discard $X_2, \ldots, X_n$.

- Then, this estimator is unbiased: In fact,

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}[X_1] = \mathbb{E}[X] = \mu_P.$$

Example of an Unbiased Estimator that is not Consistent

- However, the variance of the estimate $\hat{\mu}$ is a constant:

$$\mathbb{V}[\hat{\mu}] = \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \mathbb{E}[(X_1 - \mu)^2] = \mathbb{E}[(X - \mu)^2] = \sigma_P^2.$$

- Thus, the MSE of this estimator is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \mathbb{E}[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2] = \sigma_P^2.$$

- Thus, the MSE does not decrease to 0, even if $n \to \infty$, i.e., the estimator is not consistent.

This example demonstrates that the unbiasedness does not imply consistency.

- For consistency, we need to make sure that the variance of the estimate decreases to 0 as $n \to \infty$.

Constructing an Unbiased Estimator by Weighting

Consider again the mean estimation problem.

- Let $X \sim P$, and assume $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$.

- Assume that the data distribution Q is different from the target P.

- We show here that we can still construct an unbiased estimator of the mean

$$\mu_P = \mathbb{E}_{X \sim P}[X]$$

from the data $X_1, \ldots, X_n \sim Q$ $(i.i.d.)$.

Constructing an Unbiased Estimator by Weighting

- To this end, assume that distributions P and Q have density functions p and q, respectively.

- Define a weight function by

$$w(x) := \frac{p(x)}{q(x)}, \quad x \in \mathbb{R}$$

- Assume that this weight function is well-defined and bounded:

$$\max_{x \in \mathbb{R}} w(x) =: C < \infty.$$

- Note that this requires $p(x)/q(x) < C$, and thus

$$p(x) < Cq(x) \quad \text{for all } x \in \mathbb{R}.$$

- Thus, if the target density has a positive value $p(x) > 0$, then the data density should also have a positive value $q(x) > 0$.

Constructing an Unbiased Estimator by Weighting

- We assume for simplicity that this weight function $w(x) = p(x)/q(x)$ is known.

  - Otherwise we need to estimate it from data.

- Define an estimator F_n of the mean μ_P as:

$$\hat{\mu} := F_n(X_1, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} w(X_i) X_i.$$

- This is an unbiased estimator of the mean μ_P of P: This can be shown as follows.

Constructing an Unbiased Estimator by Weighting

- Recall that $X_1, \ldots, X_n$ are i.i.d. with $Y \sim Q$. Therefore,

$$\begin{aligned}
\mathbb{E}[\hat{\mu}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[w(X_i)X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[w(Y)Y] \\
&= \mathbb{E}[w(Y)Y] = \int x \; w(x)dQ(x) = \int x \; \frac{p(x)}{q(x)}q(x)dx \\
&= \int x \; p(x)dx = \int x \; dP(x) = \mu_P.
\end{aligned}$$

- Thus, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i$ is an unbiased estimator of μ_P.

Constructing an Unbiased Estimator by Weighting

- On the other hand, the variance of the estimator is

$$\begin{aligned}
\mathbb{V}[\hat{\mu}] &= \mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i] = \sum_{i=1}^{n} \mathbb{V}[\frac{1}{n}w(X_i)X_i] \\
&= \sum_{i=1}^{n} \frac{1}{n^2}\mathbb{V}[w(X_i)X_i] = \sum_{i=1}^{n} \frac{1}{n^2}\mathbb{V}[w(Y)Y] \\
&= \frac{1}{n}\mathbb{V}[w(Y)Y].
\end{aligned}$$

- This can be upper-bounded as

$$\begin{aligned}
&\frac{1}{n}\mathbb{V}[w(Y)Y] = \frac{1}{n}\left(\mathbb{E}[(w(Y)Y)^2] - (\mathbb{E}[w(Y)Y]\right)^2 \\
&\leq \frac{1}{n}(\mathbb{E}[C^2Y^2] + \mu_P^2) = \frac{1}{n}(C^2\mathbb{E}[Y^2] - \mu_P^2) \\
&= \frac{1}{n}(C^2(\sigma_Q^2 + \mu_Q^2) - \mu_P^2).
\end{aligned}$$

Constructing an Unbiased Estimator by Weighting

- To summarize, the MSE of the estimator is

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{1}{n}\mathbb{V}[w(Y)Y] \leq \frac{1}{n}(C^2(\sigma_Q^2 + \mu_Q^2) - \mu_P^2).$$

- Therefore, the MSE decreases to 0 as $n \to \infty$:

  - i.e., the estimator $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)X_i$ is consistent in estimating μ_P.

- The weight function $w(x)$ is called the importance weight of a point x.

- The way of constructing an estimator by weighting each sample point X_i by $w(X_i)$ is called importance weighting.

Constructing an Unbiased Estimator by Weighting

- Importance weighting is a widely used technique, examples including:

  - Domain shift adaptation in machine learning.
  - Estimation of treatment effects in causal inference.
  - Monte Carlo for efficient simulations.

- If you are interested in the first, you can for instance look at Sugiyama and Kawanabe (2012).

3.8 Variance Reduction by Introducing a Bias

Variance of Unbiased Estimators may be Large

- We demonstrate here that sometimes biased estimators may be "better" than unbiased estimators.

- The key is an approach called shrinkage or regularization, which is ubiquitous in statistics and machine learning.

Variance of Unbiased Estimators may be Large

- We have seen the bias-variance decomposition of the MSE:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{Variance} + \| \underbrace{\mathbb{E}[\hat{\theta}_n] - \theta^*}_{Bias} \|^2$$

- The MSE decomposes into the bias and variance.

- For an unbiased estimator (i.e.,the bias is zero), the MSE is equal to the variance:

$$\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] = \underbrace{\mathbb{E}[\|\hat{\theta}_n - \mathbb{E}[\hat{\theta}_n]\|^2]}_{Variance}.$$

- This variance may be large if, e.g.,

  - the sample size n is small
  - the dimensionality of $\hat{\theta}_n$ is large (in multivariate cases).

- In such a situation, a biased estimator with a lower variance may have a smaller MSE than the unbiased estimator.

Variance Reduction in Mean Estimation

- To describe this, consider the mean estimation problem.

- Let $X \sim P$, and $X_1, \ldots, X_n \sim P$ (*i.i.d.*).

- We saw that the empirical average

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of the mean of

$$\mu_P := \mathbb{E}_{X\sim P}[X],$$

and the MSE is given by

$$\mathbb{E}[(\hat{\mu} - \mu_P)^2] = \frac{\mathbb{V}[X]}{n}.$$

- We'll show that there are biased estimators that have smaller MSE than the empirical average.

Empirical Average as a Least-Squares Solution

- We first show that the empirical average $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the solution to the following optimization problem

$$\hat{\mu} = \arg\min_{\alpha\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2.$$

- i.e., we consider a least-squares problem (fitting a constant α to the data $X_1, \ldots, X_n$).

- To solve this, set the the derivative of the objective function with respect to α to be zero:

$$\frac{d}{d\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2\right) = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha - X_i) = 2\alpha - \frac{2}{n}\sum_{i=1}^{n} X_i = 0.$$

- Thus, the α that minimizes the objective function is

$$\alpha = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

i.e., the empirical average.

Regularized Least Squares and Shrinkage Estimator

- We then consider a modified optimization problem, adding a regularization term:

$$\hat{\mu}_\lambda := \arg\min_{\alpha\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2 + \lambda\alpha^2,$$

where $\lambda \geq 0$ is a regularization constant.

- The solution is given by setting the derivative of the objective function to be 0:

$$\frac{d}{d\alpha}\left(\frac{1}{n}\sum_{i=1}^{n}(\alpha - X_i)^2 + \lambda\alpha^2\right) = \frac{1}{n}\sum_{i=1}^{n} 2(\alpha - X_i) + 2\lambda\alpha$$

$$= 2\alpha - \frac{2}{n}\sum_{i=1}^{n} X_i + 2\lambda\alpha = 2\alpha(1+\lambda) - \frac{2}{n}\sum_{i=1}^{n} X_i = 0.$$

- Thus, the solution is given by

$$\alpha = \frac{1}{(1+\lambda)}\frac{1}{n}\sum_{i=1}^{n} X_i =: \hat{\mu}_\lambda$$

Regularized Least Squares and Shrinkage Estimator

$$\hat{\mu}_\lambda = \frac{1}{(1+\lambda)}\frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Large λ shrinks the solution $\hat{\mu}_\lambda$ towards 0.

  - In this sense, this is called a shrinkage estimator.

- $\lambda = 0$ recovers the empirical average $\hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Mean Square Error of the Shrinkage Estimator

- The expectation of $\hat{\mu}_\lambda$ is

$$\mathbb{E}[\hat{\mu}_\lambda] = \mathbb{E}[\frac{1}{(1+\lambda)}\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{(1+\lambda)}\mu_P.$$

- Thus, the (squared) bias of $\hat{\mu}_\lambda$ is

$$(\mathbb{E}[\hat{\mu}_\lambda] - \mu_P)^2 = (\frac{1}{(1+\lambda)}\mu_P - \mu_P)^2 = \frac{\lambda^2\mu_P^2}{(1+\lambda)^2}.$$

- Thus, the bias increases as λ increases.

- On the other hand, the variance of $\hat{\mu}_\lambda$ is

$$\begin{aligned}\mathbb{V}[\hat{\mu}_\lambda] &= \mathbb{V}[\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{(1+\lambda)^2}\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{(1+\lambda)^2}\frac{\mathbb{V}[X]}{n}.\end{aligned}$$

- Thus, the variance decreases as λ increases.

Mean Square Error of the Shrinkage Estimator

- Thus, the MSE of $\hat{\mu}_\lambda$ is

$$\begin{aligned}\mathbb{E}[(\hat{\mu}_\lambda - \mu_P)^2] &= \mathbb{V}[\hat{\mu}_\lambda] + (\mathbb{E}[\hat{\mu}_\lambda] - \mu_P)^2 \\ &= \frac{1}{(1+\lambda)^2}\frac{\mathbb{V}[X]}{n} + \frac{\lambda^2\mu_P^2}{(1+\lambda)^2}\end{aligned}$$

- Let's draw some observations. Assume $\mu_P \neq 0$.

Mean Square Error of the Shrinkage Estimator

- By an easy calculation, the MSE of $\hat{\mu}_\lambda = \frac{1}{(1+\lambda)} \frac{1}{n} \sum_{i=1}^n X_i$ can be shown to be smaller than that of the empirical average $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$:

$$\mathbb{E}[(\hat{\mu}_\lambda - \mu_P)^2] < \mathbb{E}[(\hat{\mu} - \mu_P)^2]$$

if $\lambda > 0$ is chosen so that

$$\frac{\lambda}{2+\lambda} \leq \frac{\mathbb{V}[X]}{n\ \mu_P^2}.$$

Some interpretations:

- When $\mathbb{V}[X]/n$ is large (e.g., when n is small), a large λ can be taken (and more shrinkage).

- When the mean μ_P^2 is small, a large λ can be taken (and more shrinkage).

Exercise: Perform numerical experiments to confirm that the shrinkage estimator can have a smaller MSE.

Mean Square Error of the Shrinkage Estimator

- For a right choice of $\lambda > 0$, we need to know $\mathbb{V}[X]$ and μ_P.

  - Therefore this estimator is not practically useful.

- However, under some assumptions (e.g., P is a Gaussian), there is a way of choosing λ without the knowledge of $\mathbb{V}[X]$ and μ_P.

  - This resulting estimator is called the James-Stein estimator; see (Efron and Hastie, 2016, Section7) (Berger, 1985, Section 5.4).

Regularization for Variance Reduction

- Anyway, this example illustrates that artificially introducing a bias is often useful to reduce the variance.

- In this spirit, regularization has been widely used in many statistical methods: e.g.,

  - L_2 and L_1 regularization in regression and classification (supervised learning)
  - Early stopping in optimization algorithms for machine learning algorithms.

- In supervised learning problems, a good regularization constant can be chosen by, e.g., cross validation

  - See e.g. the MALIS and ASI courses.

Summary of the Lecture

- We introduced several important concepts in statistical estimation.

- When constructing statistical estimators, always pay attention to

  - what is your quantity of interest (in the population).
  - how your data were generated.
  - whether your estimator is biased or unbiased.
  - how much your estimate would have variance.

Chapter 4

Maximum Likelihood Estimation

4.1 Estimation in Parametric Models

Density Estimation Problem

- Let P be an unknown probability distribution on a measurable set $\mathcal{X} \subset \mathbb{R}^d$.

- Assume that P has a probability density function $p : \mathcal{X} \to \mathbb{R}$.

- Given i.i.d. data $X_1, \ldots, X_n$ from the unknown P, we are interested in estimating the density function p.

- This is the task of density estimation.

Notation

We may write $X_1, \ldots, X_n \sim p$ (i.i.d.) with the density function p.

Density Estimation Problem

- There are mainly two approaches to this problem: parametric and nonparametric.

Parametric approach

- Define a model of a finite degree of freedom for the unknown density p.
- This is called a parametric model, and indexed by a finite number of parameters.
- Assumptions of the model are often made on the shape of the unknown density p.
- Density estimation is done by estimating the parameters from the data $X_1, \ldots, X_n \sim p$.

Density Estimation Problem

Nonparametric approach

- Define a model with infinite degree of freedom.
- Increase the complexity of the model as more data become available.
- Assumptions of the model are often made on the smoothness of the unknown density p.
- e.g., kernel density estimation Silverman (1986).

- In this course we'll only focus on the parametric approach (while the nonparametric approach is also important).

Parameter Approach to Density Estimation

- In the parametric approach, we define a parametric model for the unknown density function p generating data $X_1, \ldots, X_n$.

Parametric Model

- Let Θ be a set of parameter vectors (e.g., $\Theta \subset \mathbb{R}^q$).
- For each $\theta \in \Theta$, define a probability density function $p_\theta : \mathcal{X} \to [0, \infty)$.
- A parametric model is defined as the set of such density functions:

$$\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}.$$

Parametric Approach to Density Estimation

Remarks on the Term "Parametric Models"

- The parametric model can be seen as a function $f : \mathcal{X} \times \Theta \to [0, \infty)$ such that

$$f(x, \theta) := p_\theta(x), \quad x \in \mathcal{X},\ \theta \in \Theta.$$

  We may say that f is a parametric model.
- Alternatively, regarding $\theta \in \Theta$ as a variable, we also say p_θ is a parametric model for simplicity.

Parametric Approach to Density Estimation

- The parametric model $\mathcal{P}_\Theta$ should be designed so that the unknown density p belongs to $\mathcal{P}_\Theta$, i.e., $p \in \mathcal{P}_\theta$;

- $p \in P_\theta = \{p_\theta \mid \theta \in \Theta\}$ is equivalent to the existence of some $\theta^* \in \Theta$ such that

$$p = p_{\theta^*} \in \mathcal{P}_\Theta.$$

- We may call such θ^* the true parameter (vector).

- Therefore the model $\mathcal{P}_\Theta$ should reflect our knowledge/belief about the unknown p.

Parametric Approach to Density Estimation

- If $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$, we say that the model $\mathcal{P}_\Theta$ is correctly specified.

- In this case, estimation of the unknown density $p = p_{\theta^*}$ can be done by estimating the true parameter θ^* from the data $X_1, \ldots, X_n$.

- If $p \notin \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$, we say that the model $\mathcal{P}_\Theta$ is misspecified.

Example: Gaussian Models

- Recall that the density function of a Gaussian distribution on $\mathcal{X} = \mathbb{R}$ is given by

$$p_{\text{gauss}}(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

where

- $\mu \in \mathbb{R}$ is the mean of p_{gauss}
- $\sigma^2 > 0$ is the variance of p_{gauss}.

Example: Gaussian Density Models

Example: Gaussian Models

There are several ways to define a probabilistic model.

1. Parametrizing the mean

- Assume that we know/believe that the variance of the unknown density p is σ^2.
- Then we can define a parametric model p_θ by treating the mean μ as a parameter θ:

$$p_\theta(x) := p_{\text{gauss}}(x; \theta, \sigma^2).$$

- In this case, the parameter set may be defined as $\Theta := [-a, a] \subset \mathbb{R}$ for some $a > 0$.

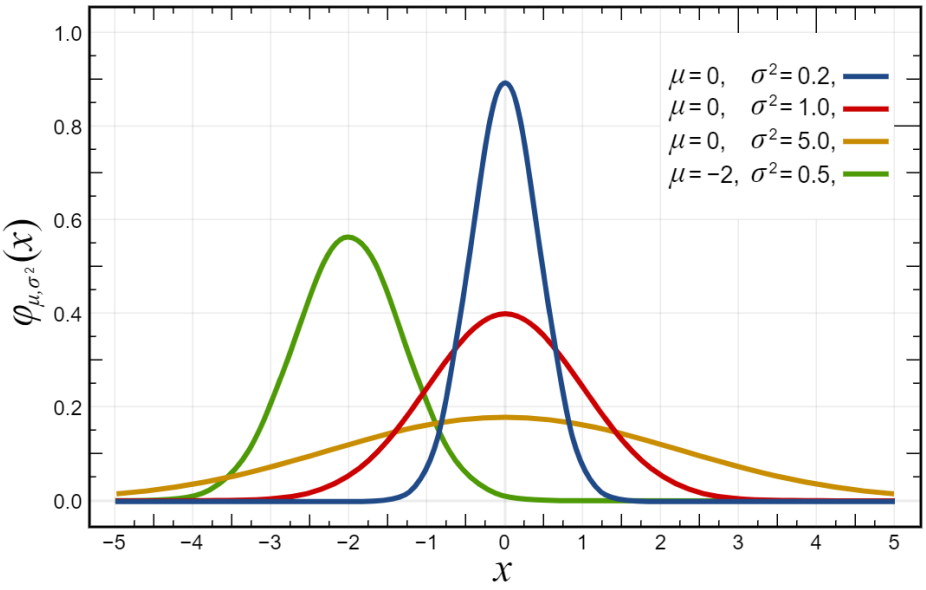


Figure 4.1: Gaussian density functions; From Wikipedia "Normal distribution."

Remarks

- Note that the definition of the parameter set Θ is a part of the model.
- e.g., the choice of the interval $[-a, a]$ implicitly represents our belief that the mean μ of p should satisfy $\mu \in [-a, a]$.

Example: Gaussian Models

2. Parametrizing both the mean and variance

- We can treat both mean μ and variance σ^2 as parameters.
- In this case, we can define a parametric model p_θ as

$$P_\theta(x) := p_{\text{gauss}}(x; \theta_1, \theta_2).$$

where

$$\theta := (\theta_1, \theta_2) \in \Theta \subset \mathbb{R} \times (0, \infty).$$

- The parameter set may be defined as

$$\Theta := [-a, a] \times [b, c] \subset \mathbb{R} \times (0, \infty)$$

for some $a, b, c > 0$.

Example: Gaussian Models

- By using the Gaussian model p_θ, we implicitly makes several assumptions about the unknown p:

Assumptions about the true p made in the Gaussian model

1. There is only one mode (or the "bump") in the density p .
2. $X \sim p$ may take an arbitrarily large value, but with an exponentially small probability.
3. $X \sim p$ takes both positive and negative values.
4. All the moments of p exist: $-\infty < \mathbb{E}_{X\sim p}[X^k] < \infty$ for all $k \in \mathbb{N}$.

- Gaussian models have been widely used in practice.

— This is because there are several mathematically and computationally convenient properties (we'll see this soon).

Example: Gaussian Mixture Models

- Assume instead that we know/believe that there two bumps in the true density p.

- Then the use of the above Gaussian model might be inappropriate.
- We can instead consider a two-component Gaussian mixture model:

$$p_\theta(x) := \frac{1}{2}p_{\text{gauss}}(x;\theta_1,\theta_2) + \frac{1}{2}p_{\text{gauss}}(x;\theta_3,\theta_4), \quad x \in \mathbb{R}$$

where

$$\theta := (\theta_1,\theta_2,\theta_3,\theta_4) \in \Theta \subset \mathbb{R}\times(0,\infty)\times\mathbb{R}\times(0,\infty).$$

4.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation

- **Maximum likelihood estimation (MLE)** is a classic but still widely used approach to estimating the parameter of a parametric model, advocated by Fisher (1922).

- The approach defines an estimator of the true parameter θ^* (in the correctly specified case) as a maximizer of the likelihood function.

Notation

In this lecture, we will use the notation

$$\arg\max_{\theta\in\Theta} A(\theta) = \left\{\theta^* \in \Theta \mid A(\theta^*) = \max_{\theta\in\Theta} A(\theta)\right\}$$

as a set of elements in Θ that maximize the objective function $A(\theta)$.

- Thus, if there are multiple maximizers of $A(\theta)$, then $\arg\max_{\theta\in\Theta} A(\theta)$ consists of multiple elements.

($\arg\min_{\theta\in\Theta} A(\theta)$ is defined in a similar way.)

Likelihood Function

- Let $X_1, \ldots, X_n \in \mathcal{X} \subset \mathbb{R}^d$ be i.i.d. data.

Likelihood Function

For a parametric model $\mathcal{P}_\Theta := \{p_\theta(x) \mid \theta \in \Theta\}$, the likelihood function $\ell_n : \Theta \to [0, \infty)$ for the data $X_1, \ldots, X_n$ is defined by:

$$\ell_n(\theta) := \prod_{i=1}^{n} p_\theta(X_i), \quad \theta \in \Theta.$$

Remarks

- $\ell_n(\theta)$ is a function of the parameter vector $\theta \in \Theta$ (with $X_1, \ldots, X_n$ being fixed).
- $\ell_n(\theta)$ is not a probability density function of $\theta \in \Theta$.
  In fact, its integral may not be 1: $\int \ell_n(\theta) d\theta = \int \left(\prod_{i=1}^{n} p_\theta(X_i)\right) d\theta \neq 1$.

Maximum Likelihood Estimation (MLE)

- Let $X_1, \ldots, X_n \sim p$ be i.i.d. data from the unknown density function p.

- Let $\ell_n(\theta) := \prod_{i=1}^{n} p_\theta(X_i)$ be the likelihood function.

Maximum Likelihood Estimation (MLE)

- Assume that there exists a true parameter $\theta^* \in \Theta$ such that $p = p_\theta^*$ (i.e., the correctly specified case $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$).
- MLE defines an estimate $\hat{\theta}_n$ of the true parameter $\theta^* \in \Theta$ as a solution to the following optimization problem:

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell_n(\theta) := \left\{ \theta' \in \Theta \mid \ell_n(\theta') = \max_{\theta\in\Theta} \ell_n(\theta) \right\}$$

- i.e., the estimate $\hat{\theta}_n$ is a maximizer of the likelihood function:

$$\ell_n(\hat{\theta}_n) = \max_{\theta\in\Theta} \ell_n(\theta).$$

Maximum Likelihood Estimation: Intuition

- We may interpret a parametric model p_θ as a conditional probability density function on $\mathcal{X}$ given $\theta \in \Theta$:

$$p(x \mid \theta) := p_\theta(x), \quad x \in \mathcal{X},\ \theta \in \Theta.$$

- Thus, the likelihood function may be interpreted as the conditional joint probability density of i.i.d. observations $X_1, \ldots, X_n$:

$$\ell_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \prod_{i=1}^{n} p(X_i \mid \theta).$$

— Note that the product form is due to the independence assumption of $X_1, \ldots, X_n$.

- Thus the MLE may be interpreted as searching for the parameter vector θ^* that maximizes the conditional probability (density) of the data $X_1, \ldots, X_n$.

- This interpretation of the likelihood function becomes important in Bayesian inference (we'll see this in a coming lecture)

MLE as Maximizing the Log Likelihood Function

- MLE can be equivalently defined as a maximizer of the log likelihood:

$$\hat{\theta} \in \arg\max_{\theta \in \Theta} \ell_n(\theta) = \arg\max_{\theta \in \Theta} \log\, \ell_n(\theta)$$

- This is because the logarithm is a monotonically increasing function:

$$\log(t) > \log(s) \iff t > s > 0.$$

- The log likelihood function is often easier to work with in practice, because the product becomes the sum.

$$\log \ell_n(\theta) = \log \prod_{i=1}^{n} p_\theta(X_i) = \sum_{i=1}^{n} \log p_\theta(X_i).$$

- We'll also see the use of log likelihood leads to a deeper understanding of MLE (Akaike, 1998).

Example: MLE with a Gaussian Density Model

- Consider a Gaussian density model on $\mathcal{X} = \mathbb{R}$, with a parametrized mean $\mu = \theta$ and a fixed variance $\sigma^2 > 0$:

$$p_\theta(x) := p_{\text{gauss}}(x; \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

- Assume that i.i.d. data $X_1, \ldots, X_n$ are given.

Example: MLE with a Gaussian Density Model

- Then the log likelihood function is given as

$$\begin{aligned}\log \ell_n(\theta) &:= \sum_{i=1}^{n} \log p_{\text{gauss}}(X_i; \theta, \sigma^2) \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \theta)^2}{2\sigma^2} \right) \right) \\ &= \sum_{i=1}^{n} \left(\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(X_i - \theta)^2}{2\sigma^2} \right) \\ &= n \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{2\sigma^2}.\end{aligned}$$

Example: MLE with a Gaussian Density Model

- To obtain the maximizer, compute the derivative w.r.t. θ and equate it to 0:

$$\frac{d \log \ell_n(\theta)}{d\theta} = \sum_{i=1}^{n} \frac{(X_i - \theta)}{\sigma^2} = 0.$$

- Solving this leads to the maximum likelihood estimator for the mean:

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- This is the empirical average of $X_1, \ldots, X_n$!

Exercise

- Think about why the empirical average is obtained as MLE for the Gaussian model. (Hint: recall that the empirical average can be given as a solution to the least-squares problem).

Example: MLE with a Gaussian Density Model

Exercise

- Consider the Gaussian model with both mean $\mu = \theta_1$ and variance $\sigma^2 = \theta_2$ parametrized:

$$p_\theta(x) := p_{\text{gauss}}(x; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{(x - \theta_1)^2}{2\theta_2^2} \right)$$

- Show that the MLE for $\theta = (\theta_1, \theta_2)$ with i.i.d. observations $X_1, \ldots, X_n$ is given by

$$\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i, \ \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\theta}_1)^2 \right).$$

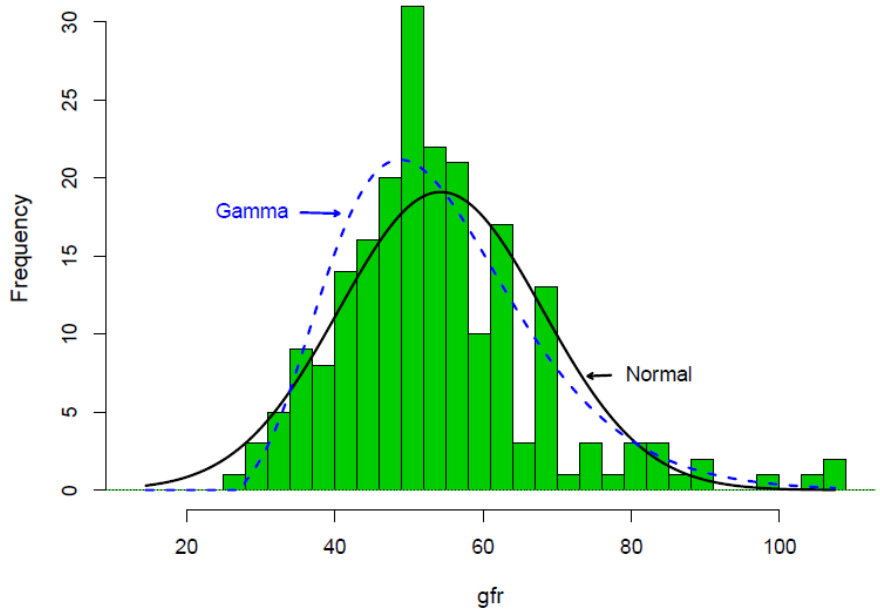


Figure 4.2: (Efron and Hastie, 2016, Fig 4.1)

Illustration

Maximum Likelihood Estimation (MLE)

- In general, this optimization problem of MLE has no analytical solution (e.g., consider Gaussian mixture models)

- In that case, one needs to use numerical optimization. e.g.,

- Gradient descent (see the Optim course for details.)
- Expectation-Maximization (EM) algorithm.

- In this lecture, we'll study statistical properties of MLE, assuming that we can obtain the maximizer $\hat{\theta}_n$.

4.3 MLE as Kullback-Leibler Divergence Minimization

MLE as Kullback-Leibler (KL) Divergence Minimization

- Here we'll see an interpretation of MLE as searching for the parameter $\theta \in \Theta$ that minimizes the KL divergence between the true density p and the model density p_θ Akaike (1998).

- This interpretation is very important, because it provides an understanding of the MLE in the misspecified case $p \notin \mathcal{P}_\Theta$:

- To describe this, we'll look at the definition and properties of the KL divergence.

Kullback-Leibler (KL) Divergence - KL divergence quantifies the discrepancy between two probability density functions.

Kullback-Leibler (KL) Divergence

- Let p and q be probability density functions on $\mathcal{X} \subset \mathbb{R}^d$ such that $p(x)/q(x) < \infty$ for all $x \in \mathcal{X}$.
- Then the KL divergence between p and q is defined as

$$\begin{aligned} KL(p||q) &:= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx. \end{aligned}$$

Intuition: KL Divergence as a Discrepancy Measure

- If $KL(p\|q)$ is large, then p and q are very different;
- If $KL(p\|q)$ is small, then p and q are similar.

Properties of the KL Divergence

Nonnegativity

- The KL divergence only take non-negative values: for any density functions p and q,

$$KL(p\|q) \geq 0.$$

- This can be seen as follows:

$$\begin{aligned} KL(p\|q) &= \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx = - \int p(x) \log \left(\frac{q(x)}{p(x)} \right) dx \\ &\geq - \log \left(\int p(x) \frac{q(x)}{p(x)} dx \right) \\ &= - \log \left(\int q(x) dx \right) = - \log(1) = 0 \end{aligned}$$

where the inequality follows from Jensen's inequality and $\log(t)$ being a convex function of $t > 0$ (see e.g.,(Berger, 1985, Sec 1.8)).

Properties of the KL Divergence

KL Divergence as a Discrepancy

- $KL(p\|q) = 0$ if and only if $p = q$ (almost everywhere).

-"if" part can be shown easily: If $p = q$, we have

$$KL(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx = \int p(x) \log(1) dx = 0.$$

Asymmetry of the KL Divergence

- That the KL divergence is not symmetric: in general,

$$KL(p\|q) \neq KL(q\|p)$$

- Therefore, KL divergence is not a distance (metric) between probability density functions. (A distance measure needs to be symmetric).

Properties of the KL Divergence

- KL divergence has its origin in Information Theory.

- Indeed, the KL divergence can be written as

$$\begin{aligned} KL(p\|\|q) &:= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \underbrace{- \int p(x) \log \frac{1}{p(x)} dx}_{\text{Entropy of } p} + \underbrace{\int p(x) \log \frac{1}{q(x)} dx}_{\text{Cross Entropy of } p \text{ and } q} \end{aligned}$$

- For details see e.g. Gray (2011) and the InfoTheo course.

Example: KL Divergence between Gaussians

- Consider the KL divergence between two Gaussian densities p and q on $\mathcal{X} := \mathbb{R}$.

KL Divergence between Univariate Gaussians

- Let $p(x) := p_{\text{gauss}}(x; \mu_p, \sigma_p^2)$ with mean $\mu_p \in \mathbb{R}$ and variance $\sigma_p^2 > 0$;
- Let $q(x) := p_{\text{gauss}}(x; \mu_q, \sigma_q^2)$ with mean $\mu_q \in \mathbb{R}$ and variance $\sigma_q^2 > 0$.
- Then the KL divergence between p and q is given by

$$KL(p\|q) = \frac{1}{2} \left(\frac{(\mu_p - \mu_q)^2}{\sigma_q^2} + \log \left(\frac{\sigma_q^2}{\sigma_p^2} \right) + \frac{\sigma_p^2}{\sigma_q^2} - 1 \right)$$

Exercise. Prove this.

Example: KL Divergence between Gaussians

- For instance, consider the equal variance case $\sigma_p^2 = \sigma_q^2 =: \sigma^2$.

- Then, the KL divergence simplifies to

$$\begin{aligned} KL(p\|q) &= \frac{1}{2}\left(\frac{(\mu_p - \mu_q)^2}{\sigma^2} + \log\left(\frac{\sigma^2}{\sigma^2}\right) + \frac{\sigma^2}{\sigma^2} - 1\right) \\ &= \frac{(\mu_p - \mu_q)^2}{2\sigma^2}. \end{aligned}$$

- We can make the following observations:

  - As difference between the means μ_p and μ_q approaches 0, the KL divergence converges to 0.
$$KL(p\|q) \to 0 \quad \text{as} \quad (\mu_p - \mu_q)^2 \to 0.$$
  - As the variance σ^2 increases, the KL divergence converges to 0
$$KL(p\|q) \to 0 \quad \text{as} \quad \sigma^2 \to \infty$$

MLE as KL Divergence Minimization

- We now look at a connection between MLE and KL divergence.

- The estimate $\hat{\theta}$ of MLE can be obtained as

$$\begin{aligned} \hat{\theta} \in \arg\max_{\theta\in\Theta} \log \ell_n(\theta) &= \arg\max_{\theta\in\Theta} \log \prod_{i=1}^n p_\theta(X_i) \\ &= \arg\max_{\theta\in\Theta} \sum_{i=1}^n \log p_\theta(X_i) = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i). \end{aligned}$$

- The objective function in the last expression is the empirical average of the log density $\log p_\theta(x)$ with the i.i.d. data $X_1, \ldots, X_n \sim p$:

$$\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i).$$

MLE as KL Divergence Minimization

- Thus, we can interpret the objective function of MLE as an empirical approximation to the expected log density:

$$\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i) \approx \mathbb{E}_{X\sim p}[\log p_\theta(X)] = \int (\log p_\theta(x))p(x)dx.$$

where the expectation is with respect to the true unknown density, $X \sim p$.

- Thus, under an appropriate identifiability condition (introduced later), we may expect that

$$\hat{\theta}_n \approx \theta^* \in \arg\max_{\theta \in \Theta} \int p(x) \log p_\theta(x) dx.$$

where $\theta^* \in \Theta$ is a maximizer of the expected log density.

- (We use the the notation θ^* as for the "true parameter" intentionally, for a reason that will be clear later).

MLE as KL Divergence Minimization

- We show that this maximizer θ^* is the minimizer of the KL divergence between the true density p and the model density p_θ:

$$\begin{aligned}
\theta^* &\in \arg\max_{\theta \in \Theta} \int p(x) \log p_\theta(x) dx \\
&= \arg\min_{\theta \in \Theta} - \int p(x) \log p_\theta(x) dx \\
&= \arg\min_{\theta \in \Theta} - \int p(x) \log p_\theta(x) dx + \int p(x) \log p(x) dx \\
&= \arg\min_{\theta \in \Theta} \int p(x) \left(- \log p_\theta(x) + \log p(x) \right) dx \\
&= \arg\min_{\theta \in \Theta} \int p(x) \log \frac{p(x)}{p_\theta(x)} dx \\
&= \arg\min_{\theta \in \Theta} KL(p \| p_\theta).
\end{aligned}$$

MLE as KL Divergence Minimization

- Thus, we have:

$$\theta^* \in \arg\max_{\theta \in \Theta} \int p(x) \log p_\theta(x) dx = \arg\min_{\theta \in \Theta} KL(p \| p_\theta).$$

- Therefore, the estimate $\hat{\theta}_n$ of MLE

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i)$$

can be seen as an approximation to the minimizer of the KL divergence:

$$\theta^* \in \arg\min_{\theta \in \Theta} KL(p \| p_\theta)$$

- We will look at closely the conditions required for this interpretation to be valid.

- These are conditions required for MLE to "succeed", thus providing a guideline for the use of MLE in practice.

4.4 Consistency of MLE

Consistency of MLE

- We saw that MLE may be interpreted as an estimator of the optimal parameter θ^* given by

$$\theta^* \in \arg\max_{\theta\in\Theta} \int p(x) \log p_\theta(x) dx = \arg\min_{\theta\in\Theta} KL(p\|p_\theta).$$

- We'll investigate the consistency of the estimate $\hat{\theta}_n$ in estimating such θ^* in a large sample limit $n \to \infty$.

- This is based on [White 82]; see this paper for details.

- The purpose is to clarify conditions under which MLE "works well."

- To this end, we'll introduce several assumptions (= conditions).

Assumptions on the Data Distribution

Assumption 1 (Data and the True Density)

The data $X_1, \ldots, X_n \in \mathcal{X} \subset \mathbb{R}^d$ are i.i.d. with a distribution P with a density function p.

Assumptions on the Parametric Model

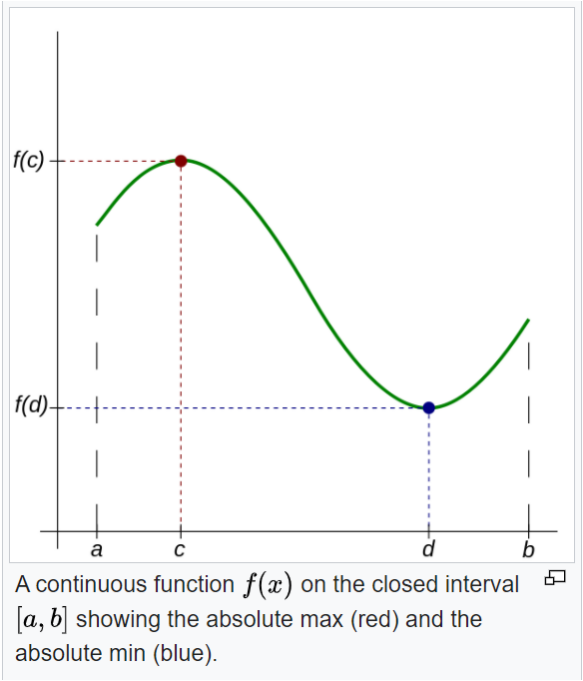
Assumption 2 (Model)

- The parameter set $\Theta \subset \mathbb{R}^q$ is compact.
  — i.e., Θ is a bounded and closed subset.
- For every $x \in \mathcal{X}$, the mapping
$$\theta \to p_\theta(x)$$
is a continuous function of $\theta \in \Theta$.

Consequence of the Continuity Assumption

- The likelihood function $\ell_n(\theta) = \prod_{i=1}^n p_\theta(X_i)$ is a continuous function of $\theta \in \Theta$, because the mapping
$$\theta \to p_\theta(X_i)$$
is continuous for all $i = 1, \ldots, n$.

Assumptions on the Parametric Model



A continuous function $f(x)$ on the closed interval $[a, b]$ showing the absolute max (red) and the absolute min (blue).

Figure 4.3: From Wikipedia "Extreme value theorem"

- Assumption 2 guarantees that the maximum of the likelihood function is bounded: i.e.,

$$\max_{\theta \in \Theta} \ell_n(\theta) = \max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i) < \infty.$$

This follows from

1. The likelihood function $\ell_n(\theta)$ is a continuous function of $\theta \in \Theta$;
2. Θ is compact;
3. **Extreme value theorem** (a general fact): a continuous function on a compact domain is bounded.

Assumptions on the Parametric Model

Assumptions on the Parametric Model

- Thus, Assumption 2 guarantees that MLE

$$\hat{\theta_n} \in \arg\max_{\theta \in \Theta} \ell_n(\theta)$$

is well-defined.

- If Assumption 2 is not satisfied, then we may have

$$\max_{\theta\in\Theta} \ell_n(x) = \infty$$

- In this case, MLE $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ell_n(x)$ is not well-defined.

Example where MLE is not Well-Defined

- Consider a 2-component Gaussian mixture model;

$$p_\theta(x) = \frac{1}{2} p_{\text{gauss}}(x; \theta_1, \theta_2) + \frac{1}{2} p_{\text{gauss}}(x; \theta_3, \theta_4),$$

with

$$\theta := (\theta_1, \theta_2, \theta_3, \theta_4) \in \Theta \subset \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

- Define the parameter set Θ as

$$\Theta := [-a, a] \times (0, c] \times [-a, a] \times (0, c]$$

for constants $a, c > 0$.

- In this case, Θ is not closed and thus not compact.

- Therefore Assumption 2 is not satisfied.

Example where MLE is not Well-Defined

- We'll show that in this case the maximum of the likelihood function is unbounded:

$$\max_{\theta\in\Theta} \ell_n(\theta) = \infty,$$

and thus MLE is not well-defined.

Example where MLE is not Well-Defined

- Define $\theta_1 := X_k$ for with $k \in \{1, \ldots, n\}$ arbitrary, and fix θ_3 and θ_4.

$$\begin{aligned} p_\theta(X_k) &= \frac{1}{2} p_{\text{gauss}}(X_k; \theta_1, \theta_2) + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} \exp\left(-\frac{(X_k - \theta_1)^2}{2\theta_2^2} \right) + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \\ &= \frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4). \end{aligned}$$

- Taking the limit $\theta_2 \to +0$ (i.e., the variance θ_2 going to 0), we have

$$\lim_{\theta_2 \to +0} p_\theta(X_k) = \lim_{\theta_2 \to +0} \left(\frac{1}{2} \frac{1}{\sqrt{2\pi\theta_2^2}} + \frac{1}{2} p_{\text{gauss}}(X_k; \theta_3, \theta_4) \right) = \infty.$$

- The limit $\theta_2 \to +0$ can be taken, because $\theta_2 \in (0, c]$.

Example where MLE is not Well-Defined

- On the other hand, for all $i \neq k$ we have

$$\begin{aligned} p_\theta(X_i) &= \frac{1}{2} p_{\text{gauss}}(X_i; \theta_1, \theta_2) + \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4) \\ &\geq \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4). \end{aligned}$$

- Therefore,

$$\begin{aligned} \lim_{\theta_2 \to +0} \ell_n(\theta) &= \lim_{\theta_2 \to +0} \prod_{i=1}^{n} p_\theta(X_i) = \lim_{\theta_2 \to +0} p_\theta(X_k) \prod_{i \neq k}^{n} p_\theta(X_i) \\ &\geq \left(\lim_{\theta_2 \to +0} p_\theta(X_k) \right) \prod_{i \neq k} \frac{1}{2} p_{\text{gauss}}(X_i; \theta_3, \theta_4) = \infty. \end{aligned}$$

This implies that

$$\max_{\theta \in \Theta} \ell_n(\theta) \geq \lim_{\theta_2 \to +0} \ell_n(\theta) = \infty.$$

Example where MLE is not Well-Defined

- This example shows that MLE is not always well-defined.

- We need to be careful about how the parameter set Θ is defined.

Exercise

Construct other examples where MLE is not well-defined.

Assumptions for the KL Divergence to be Well-Defined

Assumption 3 (The existence of the KL divergence)

- The true density $p(x)$ satisfies

$$-\infty < \int p(x) \log p(x) dx < \infty.$$

- For the model $p_\theta(x)$, there exists a function $g : \mathcal{X} \to [0, \infty)$ such that

$$|\log p_\theta(x)| \leq g(x) \quad \text{for all } x \in \mathcal{X} \text{ and } \theta \in \Theta$$

and

$$\int g(x) p(x) dx < \infty.$$

Assumptions for the KL Divergence to be Well-Defined

- The latter condition implies that

$$\left|\int p(x)\log p_\theta(x)dx\right| < \int p(x)|\log p_\theta(x)|dx \\ \leq \int p(x)g(x)dx < \infty.$$

- Therefore, the above conditions imply that the KL divergence

$$KL(p\|p_\theta) = \int p(x)\log\frac{p(x)}{p_\theta(x)}dx \\ = \int p(x)\log p(x)dx - \int p(x)\log p_\theta(x)dx$$

is finite and thus well-defined.

Exercise

- Construct examples of p and p_θ for which the KL divergence cannot be defined.

Assumption for the Identifiability

Assumption 4 (Identifiability)

- Expected log density $\int p(x)\log p_\theta(x)dx$ has a unique maximizer $\theta^* \in \Theta$: i.e.,

$$\int p(x)\log p_{\theta^*}(x)dx > \int p(x)\log p_\theta(x)dx \quad \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta^*.$$

- In other words, θ^* is the unique minimizer of the KL-divergence:

$$KL(p\|p_{\theta^*}) = \int p(x)\log p(x)dx - \int p(x)\log p_{\theta^*}(x)dx \\ < \int p(x)\log p(x)dx - \int p(x)\log p_\theta(x)dx = KL(p\|p_\theta) \\ \text{for all } \theta \in \Theta \text{ with } \theta \neq \theta^*.$$

- In this case, we call the model $P_\Theta = \{p_\theta \mid \theta \in \Theta\}$ is identifiable with respect to p.

Assumption for the Identifiability

- If Assumption 4 (identifiability) is true, the notation

$$\theta^* = \arg\max_{\theta\in\Theta} \int p(x)\log p_\theta(x)dx = \arg\min_{\theta\in\Theta} KL(p\|p_\theta)$$

is justified (because the "argmax" only consists of one element, θ^*).

- Assumption 4 enables us to define θ^* as the quantity of interest (or the estimand) in statistical estimation.

- Thus, we can discuss the "consistency" of the MLE $\hat{\theta}_n \to \theta^*$ as $n \to \infty$.

- This will be important in particular

- when we are interested in the optimal parameter θ^* itself; and
- when we want to perform hypothesis testing regarding θ^*.

Interpretation of the Optimal Parameter θ^*

- Let's consider what the optimal parameter θ^* is.

- Assume that the KL divergence between the true unknown density p and the optimal model density p_{θ^*} is zero:

$$KL(p\|p_{\theta^*}) = 0.$$

- In this case,

- We have $p = p_{\theta^*}$, because $KL(p\|p_\theta^*) = 0$ if and only if $p = p_{\theta^*}$.
- Therefore, $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$ i.e., the model $\mathcal{P}_\Theta$ is correctly specified.

- Thus, we can interpret θ^* as the true parameter in this case.

- The convergence of MLE $\hat{\theta}_n \to \theta^*$ implies that the MLE is consistent in estimating the true parameter θ^* .

Interpretation of the Optimal Parameter θ^*

Summary

- $KL(p\|p_{\theta^*}) = 0$ corresponds to the correctly specified case $p \in \mathcal{P}_\Theta$.
- Since $p = p_{\theta^*}$, the optimal parameter θ^* is interpreted as the true parameter.

Interpretation of the Optimal Parameter θ^*

Interpretation of the Optimal Parameter θ^*

- Assume the KL divergence between the true density p and the optimal model density p_{θ^*} is larger than zero:

$$KL(p\|p_{\theta^*}) = \min_{\theta\in\Theta} KL(p\|p_\theta) > 0,$$

- In this case,

- we have $p \neq p_{\theta^*}$, i.e., the optimal model density p_{θ^*} does not match the true density p;

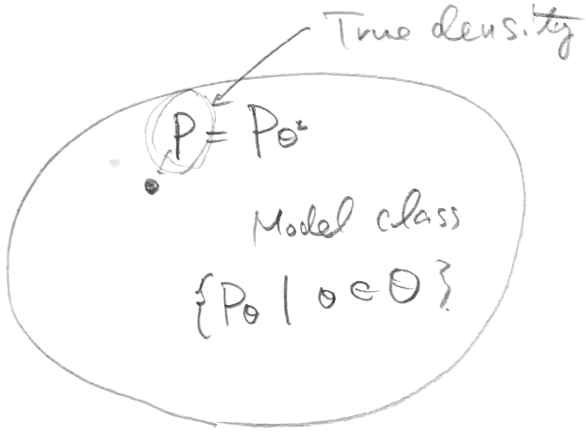


Figure 4.4: When $KL(p\|p_{\theta^*}) = 0$ (correctly specified case)

- thus $p \notin \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$, i.e., the model $\mathcal{P}_\Theta$ is misspecified.

- In this case, we can interpret p_{θ^*} as the best approximation to the true density p as measured by the KL divergence.

- Thus, we can interpret θ^* as the parameter that gives the best approximation of the model $\mathcal{P}_\Theta$ to the true p.

Interpretation of the Optimal Parameter θ^*

Summary

- $KL(p\|p_{\theta^*}) > 0$ corresponds to the misspecified case $p \notin \mathcal{P}_\Theta$.
- Since $KL(p\|p_{\theta^*}) = \min_{\theta\in\Theta} KL(p\|p_\theta)$, the optimal parameter θ^* is interpreted as the parameter that gives the best approximation p_{θ^*} to the true density p under the KL divergence.

Interpretation of the Optimal Parameter θ^*

Example where the Model is not Identifiable

- Consider a 2-component Gaussian mixture model;

$$p_\theta(x) = \frac{1}{2}p_{\text{gauss}}(x;\theta_1,\theta_2) + \frac{1}{2}p_{\text{gauss}}(x;\theta_3,\theta_4)$$

with

$$\theta = (\theta_1,\theta_2,\theta_3,\theta_4) \in \Theta \subset \mathbb{R}\times(0,\infty)\times\mathbb{R}\times(0,\infty).$$

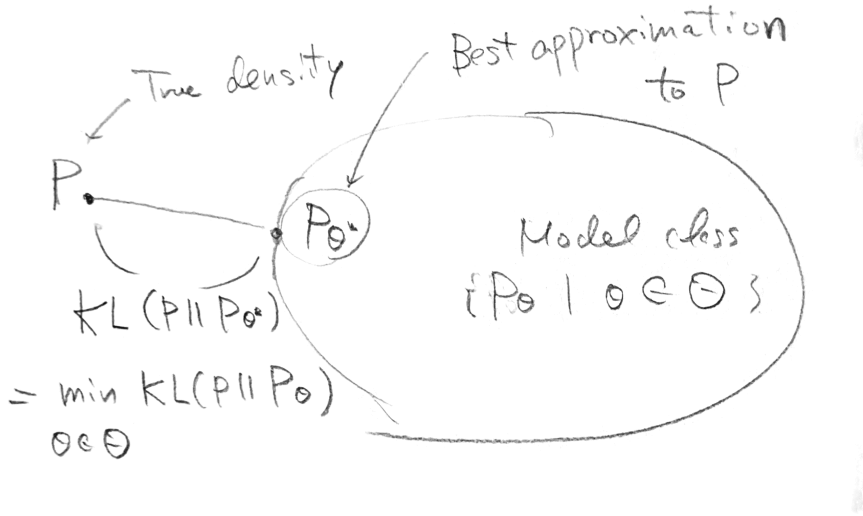


Figure 4.5: When $KL(p\|p_{\theta^*}) > 0$ (model misspecification).

- Define the parameter set Θ by

$$\Theta := [-a, a] \times [b, c] \times [-a, a] \times [b, c]$$

for constants $a, b, c > 0$.

- The model is not identifiable, because switching (θ_1, θ_2) and (θ_3, θ_4) produces the same density function.

Example where the Model is not Identifiable

- To show this, let

$$(\mu_1, \sigma_1^2) \in [-a, a] \times [b, c], \quad (\mu_2, \sigma_2^2) \in [-a, a] \times [b, c]$$

be arbitrary constants such that $\sigma_1^2 \neq \sigma_2^2$.

- Then, for $\theta^* := (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$, we have

$$p_{\theta^*}(x) = \frac{1}{2} p_{\text{gauss}}(x; \mu_1, \sigma_1^2) + \frac{1}{2} p_{\text{gauss}}(x; \mu_2, \sigma_2^2)$$

- For $\tilde{\theta}^* := (\mu_2, \sigma_2^2, \mu_1, \sigma_1^2)$

$$p_{\tilde{\theta}^*}(x) = \frac{1}{2} p_{\text{gauss}}(x; \mu_2, \sigma_2^2) + \frac{1}{2} p_{\text{gauss}}(x; \mu_1, \sigma_1^2)$$

- Thus, we have

$$p_{\theta^*} = p_{\tilde{\theta}^*} \quad \text{while} \quad \theta^* \neq \tilde{\theta}^*.$$

- Therefore the mixture model with this parameter set Θ is not identifiable.

Example where the Model is not Identifiable

- A simple trick to make this model identifiable is to restrict the parameter set Θ.

- For instance, if we define the parameter set as

$$\Theta := \{(\theta_1, \theta_2, \theta_3, \theta_4) \in [-a, a] \times [b, c] \times [-a, a] \times [b, c] \mid \theta_2 < \theta_4\}$$

then the mixture model becomes identifiable.

- This corresponds to assuming that one mixture component has a smaller variance than the other.

Exercise

Construct other examples where the model is not identifiable.

MLE Consistency Theorem

Theorem: Consistency of MLE (Theorem 2.2 of White (1982))

- Suppose that Assumptions 1, 2, 3 and 4 are satisfied.
- Let

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell_n(\theta) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p_\theta(X_i)$$

be the MLE with i.i.d. data $X_1, \ldots, X_n \sim p$.

- Let $\theta^* \in \Theta$ be the optimal parameter

$$\theta^* = \arg\max_{\theta \in \Theta} \int p(x) \log p_\theta(x) dx = \arg\min_{\theta \in \Theta} KL(p \| p_\theta)$$

.

- Then $\hat{\theta}_n$ converges to θ^* almost surely: i.e.,

$$\Pr(\lim_{n \to \infty} \hat{\theta}_n = \theta^*) = 1.$$

For the proof, see White (1982) and also (Van der Vaart, 1998, Sec 5.5).

MLE Consistency Theorem

The proof idea is that

1. First show that

$$\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) \to \int p(x) \log p_\theta(x) dx \quad \text{as} \quad n \to \infty$$

uniformly for all $\theta \in \Theta$.

2. Then conclude that

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log p_\theta(X_i) \to \theta^* = \arg\max_{\theta\in\Theta^*} \int p(x) \log p_\theta(x) dx.$$

as $n \to \infty$.

4.5 Conclusions and Further Readings

Conclusions

- MLE can be understood as searching for a model density that best approximates the true density in terms of the KL divergence.

- MLE makes sense also in the misspecified case where the true density does not belong to the model class.

- MLE is not always consistent; we need conditions = assumptions.

- These conditions provide a guideline for designing your parametric model.

Conclusions

More generic takeaways:

- A role of convergence analysis is to understand conditions under which the method of interest works well.

- Even the MLE - one of the simplest approaches - requires several conditions.

- So please always try to understand conditions under which your favorite statistical/ML method should work!

Further Readings

- (Fisher, 1922, Section 6).

- White (1982)

- (Efron and Hastie, 2016, Chapter 4)

Chapter 5

Hypothesis Testing

5.1 Introduction: The Lady Tasting Tea Experiment

The Lady Tasting Tea Experiment (Fisher, 1937, Chapter II)

- There was a lady who claimed that she can distinguish the tastes of tea with milk made in the following two different ways:

Way M: Milk is first poured into the cup, and tea later.

Way T: Tea is first poured into the cup, and milk later.

- Ronald Fisher Fisher (1937) came up with an idea of testing her claim by a randomized experiment.

The Lady Tasting Tea Experiment (Fisher, 1937, Chapter II)

1) Let's make 8 cups of tea, of which

- 4 cups are made in Way M.
- 4 cups are made in Way T.

2) Shuffle the order of the 8 cups randomly:

- For instance, assume that as a result, the cups are ordered as:

M-M-T-M-T-T-T-M

- This information was not shared to the lady.

- She only knew that 4 of them were made in M; and the other 4 cups in T.

The Lady Tasting Tea Experiment (Fisher, 1937, Chapter II)

3) Ask the lady

- to taste the 8 cups of tea in the given order; and
- to pick up 4 cups of M from the 8 cups.

- In the end, the lady correctly identified all the 4 cups of M from the 8 cups (i.e., did no mistake).

- Fisher concluded that it is likely that she can distinguish the two ways of making tea.

- What was Fisher's reasoning?

Fisher's Reasoning

- In total, there are 70 different ways of choosing the 4 cups for M from the 8 cups

$$70 = \frac{8!}{4!4!} = \frac{8 \times 7 \times 6 \times 5 \times 4}{4 \times 3 \times 2 \times 1}$$

- Assume that the lady

- was not able to distinguish the tastes (= null hypothesis); and
- just did a random guess, picking one of the 70 ways randomly.

- **Under this assumption**, the probability of correctly identifying 4 cups of M from the 8 cups is $1/70 \approx 0.014$:

- This probability is very small, so we can conclude that

- It is unlikely that the lady is doing a random guess.
- i.e., the null hypothesis is unlikely to be true.

Fisher's Reasoning

- Assume instead a situation where the lady

- correctly identified 3 M cups, but
- wrongly chose 1 cup.

- There are 16 different ways of choosing 3 M cups correctly and one cup wrongly (**Exercise**: confirm this).

- Thus, under the null hypothesis (= the lady is doing a random guess),

- the probability of correctly choosing 3 M cups and wrongly choosing 1 cup is $16/70 \approx 0.23$.

- This probability is "not very small," and therefore

- we cannot deny the null hypothesis that the lady was doing a random guess.

Fisher's Reasoning - This example illustrates the idea of statistical hypothesis testing and a randomized experiment.

- In this lecture, we'll learn basics of hypothesis testing.

- For reading, I recommend (Rao, 1973, Chapter 7).

5.2 Procedure of Statistical Hypothesis Testing

Hypothesis Testing: Statistical Proof by Contradiction

Hypothesis testing may be understood as a statistics version of Proof by Contradiction:

Proof by Contradiction (Mathematics)

1. To prove a statement A, assume that A is not true;
2. Starting from the assumption, derive a statement B that produces a contradiction.
3. Conclude that the statement A is true.

Procedure of Testing: Step 1. Defining Hypotheses

- Hypothesis testing starts from defining a null hypothesis H_0 and an alternative hypothesis H_1

Null Hypothesis H_0

The hypothesis that you try to reject in the end.

Alternative Hypothesis H_1

The hypothesis that you try to "prove" (statistically).

Example (The lady tasting tea experiment)

- The null hypothesis H_0:
  — The lady cannot distinguish the tastes of tea of different kinds.
- The alternative hypothesis H_1:
  — The lady can distinguish the tastes of tea of different kinds.

Procedure of Testing: Step 1. Defining Hypotheses

- Let $(\Omega, \mathcal{F})$ be a measurable space, where

- Ω is a sample space, consisting possible outcomes of the experiment.

- $\mathcal{F}$ is a σ-algebra, i.e., a set of subsets of Ω for which probabilities can be defined.

- For the null H_0 and alternative hypotheses H_1, define the associated probability distributions P_0 and P_1 on $(\Omega, \mathcal{F})$:

Distributions under the Null and Alternative Hypotheses

- P_0 is the probability distribution on Ω when the null H_0 is true.
- P_1 is the probability distribution on Ω when the alternative H_1 is true.
- We may write P_0 and P_1 in the form of conditional distribution:

$$P(S \mid H_0) := P_0(S), \quad P(S \mid H_1) := P_1(S), \quad S \in \mathcal{F}.$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- The sample space Ω consists of 70 different ways of choosing 4 cups of M from 8 cups:
$$\Omega := \{\omega_1, \omega_2, \ldots, \omega_{70}\},$$
where each $\omega_i \in \Omega$ represents one way of ordering, e.g.,
$$\begin{aligned}\omega_1 &:= \text{M-M-M-M-T-T-T-T}\\ \omega_2 &:= \text{M-M-M-T-M-T-T-T}\\ &\ldots\\ \omega_{69} &:= \text{M-T-T-T-M-M-M-T}\\ \omega_{70} &:= \text{T-T-T-T-M-M-M-M}\end{aligned}$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- Under the null hypothesis H_0, the lady gives a random guess; therefore the distribution P_0 under the null is
$$P_0(\{\omega_1\}) = P_0(\{\omega_2\}) = \cdots = P_0(\{\omega_{70}\}) = 1/70.$$
- Under the alternative hypothesis H_1, let's **assume** that the lady can identify the correct 4 cups of M with probability 1:
$$P_1(\{\omega_{32}\}) = 1, \quad P(\{\omega_i\}) = 0 \text{ for all } i \neq 32,$$
where $\omega_{32} \in \Omega$ is the correct ordering:
$$\omega_{32} := \text{M-M-T-M-T-T-T-M}.$$

Procedure of Testing: Step 1. Defining Hypotheses

Example (The lady tasting tea experiment)

- Note that the way of defining P_1 is not unique: we may define, e.g.,
$$P_1(\{\omega_{32}\}) = 0.9, \quad P_1(\{\omega_i\}) = 0.1/69 \text{ for all } i \neq 32,$$
- This may represent another alternative hypothesis H_1' that
  - the lady can distinguish tastes of tea of different kinds
  - but may loose her tasting ability with probability 1/10.

Step 2: Defining Significance Level and Critical Region

- The next step is to decide the level of significance and the critical region for the test.

Significance Level

- Define a small constant $\alpha > 0$, called the level of significance (e.g., $\alpha = 0.05$ or $\alpha = 0.01$).

Step 2: Defining Significance Level and Critical Region

Critical Region

- Given a significance level $\alpha > 0$, determine a subset $S_\alpha \subset \Omega$ (such that $S_\alpha \in \mathcal{F}$), called the critical region, such that

1. the probability of S_α under the null H_0 is less than or equal to α:

$$P_0(S_\alpha) \leq \alpha;$$

2. the probability of S_α under the alternative H_1

$$P_1(S_\alpha)$$

becomes as large as possible.

Remark

- The second requirement is equivalent to choosing S_α so that $P_1(\Omega \backslash S_\alpha) = 1 - P_1(S_\alpha)$ becomes as small as possible .

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Let's define $\alpha := 0.05$ as our significance level.

- We may define the critical region S_α as the singleton set of ω_{32}:

$$S_\alpha := \{\omega_{32}\},$$

where ω_{32} := M-M-T-M-T-T-T-M is the correct ordering of 8 cups.
- Then

1. The probability of S_α under the null H_0 (the lady cannot distinguish the tastes) is

$$P_0(S_\alpha) = 1/70 \approx 0.014 \leq 0.05 = \alpha.$$

2. The probability of S_α under the alternative H_1 (the lady can perfectly distinguish the tastes) is

$$P_1(S_\alpha) = 1.$$

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Note that $S_\alpha = \{\omega_{32}\}$ is not the only way of defining a critical region.

- For instance, we may define

$$S_\alpha := \{\omega_{31},\ \omega_{32},\ \omega_{33}\},$$

where ω_{31} and ω_{33} are two ways of wrongly identifying one M cup as T.

$$\omega_{31} := \text{M-M-T-M-T-T-M-T},$$
$$\omega_{33} := \text{M-T-M-M-T-T-T-M}$$

- In this case,

$$P_0(S_\alpha) = 3/70 \approx 0.043 \leq 0.05 = \alpha,$$
$$P_1(S_\alpha) = P_1(\{\omega_{32}\}) + P_1(\{\omega_{31}, \omega_{33}\}) = 1 + 0 = 1.$$

Step 2: Defining Significance Level and Critical Region

Example (The lady tasting tea experiment)

- Or even we may define the critical region S_α for arbitrary $i = 1, 2, \ldots, 70$ with $i \neq 32$ such that
$$S_\alpha := \{\omega_i\}$$
- In this case, we have
$$P_0(S_\alpha) = 1/70 \approx 0.014 \leq 0.05 = \alpha,$$
$$P_1(S_\alpha) = 0.$$
- Since $P_1(S_\alpha) = 0$, this critical region S_α should not be chosen for our alternative hypothesis H_1.

Step 3: Obtain a Sample, and Make a Decision

- After deciding a significance level $\alpha > 0$ and a critical region $S_\alpha \subset \Omega$, make a statistical decision in the following way:

Statistical decision of whether rejecting H_0 or not

- Obtain a sample $\omega_e \in \Omega$ by performing an experiment.
  - If $\omega_e \in S_\alpha$, we reject the null hypothesis H_0.
  - If $\omega_e \notin S_\alpha$, we don't reject the null hypothesis H_0.
- We may say that the test is significant with level α.

Step 3: Obtain a Sample, and Make a Decision

Example (The lady tasting tea experiment)

- Let $\alpha := 0.05$ and $S_\alpha := \{\omega_{32}\}$.

- As a result of the experiment, the lady correctly identified the 4 M cups out of 8 cups, i.e.,
$$\omega_e = \text{M-M-T-M-T-T-T-M} = \omega_{32}.$$
  - Thus we have $\omega_e \in S_\alpha$; and thus
  - We reject the null hypothesis H_0 that the lady cannot distinguish the tastes of tea of different kinds.
  - This test is significant with the level $\alpha = 0.05$.

Remarks on the Testing Procedure

- Ronald Fisher made the following remarks on the testing procedure.

(Fisher, 1937, Section 8)

- It should be noted that the null hypothesis is never proved or established, but is possibly **disproved**, in the course of experimentation.
- Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

- This means that $\omega_e \notin S_\alpha$ does not prove the null hypothesis H_0; we just don't reject the null H_0.

Remarks on the Testing Procedure

Example (The lady tasting tea experiment)

- Assume that the lady made one mistake: $\omega_e := \omega_{31} \neq \omega_{32}$.
- Then $\omega_e \notin S_\alpha = \{\omega_{32}\}$, and we don't reject the null H_0.
- But this **does not prove** the null hypothesis H_0 that the lady cannot distinguish the tastes of tea.

5.3 Type 1 Error, Type 2 Error and the Power of a Test

Type 1 Error and Type 2 Error

- In hypothesis testing, there are two kinds of errors: **Type 1** and **Type 2**.

Type 1 Error and Type 2 Error

- **Type 1 Error**:
  — Rejecting the null hypothesis H_0, when H_0 is true.
- **Type 2 Error**:
  — Not rejecting the null hypothesis H_0, when an alternative hypothesis H_1 is true.

Type 1 Error and Type 2 Error

Example (The lady tasting tea experiment)

- **Type 1 Error:**
  - Rejecting the null hypothesis H_0 that the lady is doing a random guess
  - when the lady is really doing a random guess (H_0 is true)
- **Type 2 Error:**
  - Not rejecting that the null hypothesis H_0 that the lady is doing a random guess
  - when the lady has the ability of distinguishing the tastes of tea (H_1 is true)

Type 1 Error and the Level of Significance

- Recall that

  - we reject the null H_0 when $\omega_e \in S_\alpha$;
  - we don't reject the null H_0 when $\omega_e \notin S_\alpha$ (i.e., when $\omega_e \in \Omega \backslash S_\alpha$).

- Thus, the probability of making the Type 1 error may be given by

$$P(S_\alpha \mid H_0) := P_0(S_\alpha) \leq \alpha,$$

where the inequality follows from the definition of critical region S_α.

- i.e., the level of the significance α is (the upper-bound of) the probability of making the Type 1 error.

Type 2 Error and Statistical Power

- On the other hand, the probability of making the Type 2 error is:

$$P_1(\Omega \backslash S_\alpha) = 1 - P_1(S_\alpha).$$

- Thus, the following ways of choosing a critical region S_α are equivalent:

1. $P_1(S_\alpha)$ is maximized.

Reality \ Test	Not Reject H_0	Reject H_0
H_0 is true	(prob. $1-\alpha$)	Type 1 Error (prob. α)
H_1 is true	Type 2 Error (prob. β)	(**Power** = prob. $1-\beta$)

2. $1 - P_1(S_\alpha)$ is minimized (probability of Type 2 error).

- This probability $P_1(S_\alpha)$ is called the power of the test.

Power of a Test, $P_1(S_\alpha)$

- The probability of rejecting the null hypothesis H_0, when the alternative hypothesis H_1 is true.

Recap: Critical Region

Critical Region

- Given a significance level $\alpha > 0$, determine a subset $S_\alpha \subset \Omega$ (such that $S_\alpha \in \mathcal{F}$), called the critical region, such that

1. the probability of S_α under the null H_0 is less than or equal to α:

$$P_0(S_\alpha) = \textbf{Probability of Type 1 Error} \leq \alpha;$$

2. the probability of S_α under the alternative H_1

$$P_1(S_\alpha) = \textbf{Power of the Test}$$

becomes as large as possible.

Remark

- The second requirement is equivalent to choosing S_α so that $P_1(\Omega \backslash S_\alpha) =$ **Prob. of Type 2 Error** $= 1 - P_1(S_\alpha)$ becomes as small as possible .

Type 1 Error, Type 2 Error, and Power of a Test

- Relations between the Type 1 error, Type 2 error and the power of a test can be summarized as follows:

5.4 Test Statistics

Test Statistics

- In practice, the determination of a critical region S_α is done by defining a test statistic.

Test Statistics

- Let Ω be a sample space.
- A test statistic T is a (measurable) function from Ω to $\mathbb{R}$:

$$T : \Omega \to \mathbb{R}.$$

Remark

- Depending on the problem, we may define a different range for a statistic T.
- e.g., $T : \Omega \to \mathbb{Z}$ (where where $\mathbb{Z}$ is the set of all integers).

- A test statistic $T : \Omega \to \mathbb{R}$ summarizes characteristics of an experiment outcome $\omega_e \in \Omega$ into one dimensional value $T(\omega_e) \in \mathbb{R}$.

Test Statistics

- For any (measurable) subset $A \subset \mathbb{R}$, we can define the corresponding subset in Ω by the inverse map of T as

$$T^{-1}(A) := \{\omega \in \Omega \mid T(\omega) \in A\} \subset \Omega$$

- Therefore, we can define a critical region $S_\alpha \subset \Omega$ by defining a corresponding subset $I_\alpha \subset \mathbb{R}$ for T:

$$S_\alpha := T^{-1}(I_\alpha) = \{\omega \in \Omega \mid T(\omega) \in I_\alpha\} \subset \Omega$$

- We thus call I_α a critical region with significance level $\alpha > 0$, if it satisfies

$$P_{0,T}(I_\alpha) := P_0(T^{-1}(I_\alpha)) = P_0(S_\alpha) \leq \alpha,$$

- Here, $P_{0,T}$ is the probability distribution on $\mathbb{R}$, induced from the test statistic $T : \Omega \to \mathbb{R}$ and the distribution P_0 on Ω under the null H_0.

Hypothesis Testing with a Test Statistic

- Hypothesis testing of significance level $\alpha > 0$ can be carried out, with the test statistic T and the critical region $I_\alpha \subset \mathbb{R}$ in the following way:

Hypothesis Testing with a Test Statistic

- Let $\omega_e \in \Omega$ be the outcome of an experiment.
  - Reject the null hypothesis H_0, if $T(\omega_e) \in I_\alpha$;
  - Not reject the null hypothesis H_0, if $T(\omega_e) \notin I_\alpha$.

- The question is how to choose the critical region $I_\alpha \subset \mathbb{R}$.

- To this end, we need to consider the probabilities of Type 1 and 2 errors, and the power of the test.

- This requires considering the distributions of the test statistic T under the null H_0 and alternative H_1, respectively.

Probability Distributions of a Test Statistic

Distribution of T under the Null Hypothesis H_0

- Let $(\Omega, \mathcal{F}, P_0)$ be the probability space associated with the null hypothesis H_0.
- Under the null H_0, the test statistic $T : \Omega \to \mathbb{R}$ can be interpreted as a random variable in $\mathbb{R}$ induced from $(\Omega, \mathcal{F}, P_0)$:

$$T(\omega), \quad \omega \sim P_0$$

- Then the probability distribution of T under the null hypothesis H_0, denoted by $P_{0,T}$, is given by

$$P_{0,T}(A) := P_0(T^{-1}(A)) \text{ for any measurable } A \subset \mathbb{R}$$

Probability Distributions of a Test Statistic

Distribution of T under the Alternative Hypothesis H_1

- Let $(\Omega, \mathcal{F}, P_1)$ be the probability space associated with the alternative hypothesis H_1.
- Under the alternative H_1, the test statistic $T : \Omega \to \mathbb{R}$ can be interpreted as a random variable in $\mathbb{R}$ induced from $(\Omega, \mathcal{F}, P_1)$:

$$T(\omega), \quad \omega \sim P_1$$

- Then the probability distribution of T under the alternative hypothesis H_1, denoted by $P_{1,T}$, is given by

$$P_{1,T}(A) := P_1(T^{-1}(A)) \text{ for any measurable } A \subset \mathbb{R}$$

Type 1 Error, Type 2 Error, and Power

- Recall that the Type 1 and Typer 2 errors of a test are defined as:

- **Type 1 Error:** rejecting the null H_0 when H_0 is true;
- **Type 2 Error:** not rejecting the null H_0 when an alternative H_1 is true.

- Since the test rejects H_0 when $T(\omega_e) \in I_\alpha$, the probability of making the Type 1 Error is thus given by

$$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha))$$

- Since the test does not reject H_0 when $T(\omega_e) \notin I_\alpha$, the probability of making the Type 2 Error is

$$P_{1,T}(\mathbb{R} \backslash I_\alpha) = 1 - P_{1,T}(I_\alpha)$$

- The **Test Power**, i.e., the probability of rejecting when H_1 is true, is thus

$$P_{1,T}(I_\alpha) = 1 - \text{Prob. Type 2 Error}$$

Test Statistics: How to Choose the Critical Region

- To summarize, the critical region $I_\alpha \subset \mathbb{R}$ should be chosen as follows:

Critical Region for a Test Statistic

- Let $T : \Omega \to \mathbb{R}$ be a test statistic.
- Given a significance level $\alpha > 0$, determine a subset $I_\alpha \subset \mathbb{R}$, called the critical region, such that

1. the probability of I_α under the null H_0 is less than or equal to α:

$$P_{0,T}(I_\alpha) := P_0(T^{-1}(I_\alpha)) = \textbf{Type 1 Error} \leq \alpha;$$

2. the probability of I_α under the alternative H_1

$$P_{1,T}(I_\alpha) := P_1(T^{-1}(I_\alpha)) = \textbf{Power of the Test}$$

becomes as large as possible.

Example: Testing the Location of a Gaussian Mean

- Let p^* be an unknown probability density function on $\mathbb{R}$.

- Assume that we know/believe that p^* is Gaussian, with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$:

$$p^*(x) = p_{\text{gauss}}(x; \mu, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

- Assume that we can perform an experiment to obtain an i.i.d. sample of size n from p^*:

$$x_1, \ldots, x_n \in \mathbb{R}$$

- Assume that we are interested in testing whether the unknown mean μ is equal to some specified value $\mu_0 \in \mathbb{R}$ or not.

- Thus, the null hypothesis H_0 and alternative hypothesis H_1 may be defined as

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

Example: Testing the Location of a Gaussian Mean

- For instance, assume that μ_0 is the average blood pressure of the whole French population.

- Assume that we are interested in the effect of a certain drug on the blood pressure.

- Let $\omega_e = (x_1, \ldots, x_n)$ be the blood pressures of n randomly selected French people, measured after each being treated the drug.

- By testing the null hypothesis $H_0 : \mu = \mu_0$, we could investigate whether the drug is effective in changing the blood pressure or not.

Example: Testing the Location of a Gaussian Mean

- We can define the sample space Ω as

$$\Omega := \mathbb{R}^n.$$

- Each $\omega := (x_1, \ldots, x_n) \in \Omega$ represents a possible experiment outcome of n i.i.d. observations.

- Thus, the distribution P_0 on Ω under the null hypothesis H_0 is given by the density function $p_0 : \Omega \to \mathbb{R}$:

$$\begin{aligned} p_0(\omega) &= \prod_{i=1}^{n} p_{\text{gauss}}(x_i; \mu_0, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i - \mu_0)^2}{2\sigma^2}) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}), \quad \omega := (x_1, \ldots, x_n) \in \Omega. \end{aligned}$$

Example: Testing the Location of a Gaussian Mean

- We can define a test statistic $T : \Omega \to \mathbb{R}$ as

$$\begin{aligned} T(\omega) := T((x_1, \ldots, x_n)) &:= \frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right), \\ \omega := (x_1, \ldots, x_n) &\in \Omega := \mathbb{R}^n. \end{aligned}$$

- Consider

$$\omega = (X_1, \ldots, X_n) \sim P_0 \quad (i.e., X_1, \ldots, X_n \sim p(x; \mu_0, \sigma^2), \quad i.i.d.)$$

as a random variable under the null hypothesis H_0.

- Then the distribution $P_{0,T}$ of the test statistic

$$T(\omega) = T((X_1, \ldots, X_n)) = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_0\right)$$

is Gaussian, with mean 0 and variance 1.

Example: Testing the Location of a Gaussian Mean

- In other words, the density function $p_{0,T}$ of the distribution $P_{0,T}$ of the test statistic T under the null hypothesis H_0 is

$$p_{0,T}(t) := p_{\text{gauss}}(t; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}), \quad t \in \mathbb{R}.$$

Exercise: Prove this.

Hint: First derive the probability distribution of $\frac{1}{n}\sum_{i=1}^{n} X_i$.

To this end, use the following facts (where $X \sim p_{\text{gauss}}(x; \mu_0, \sigma^2)$):

- The sum of Gaussian random variables is Gaussian.
- $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X] = \mu_0$
- $\mathbb{V}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\mathbb{V}[X] = \frac{\sigma^2}{n}$.

Example: Testing the Location of a Gaussian Mean

- Thus, we may define a critical region I_α with significance level $\alpha > 0$

$$I_\alpha := (-\infty, -c_\alpha] \cup [c_\alpha, \infty) \subset \mathbb{R}$$

where c_α is a constant satisfying

$$P_{0,T}(I_\alpha) = \int_{-\infty}^{-c_\alpha} p_{0,T}(t)dt + \int_{c_\alpha}^{\infty} p_{0,T}(t)dt = \alpha.$$

- For instance, if $\alpha := 0.05$, we can take $c_\alpha \approx 1.96$.

Example: Testing the Location of a Gaussian Mean

- The tail regions are the critical region I_α with $\alpha = 0.05$.

- We reject the null hypothesis $H_0 : \mu = \mu_0$ if

$$T(\omega_e) > 1.96 \quad \text{or} \quad T(\omega_e) < -1.96$$

for an experiment outcome $\omega_e = (x_1, \ldots, x_n)$.

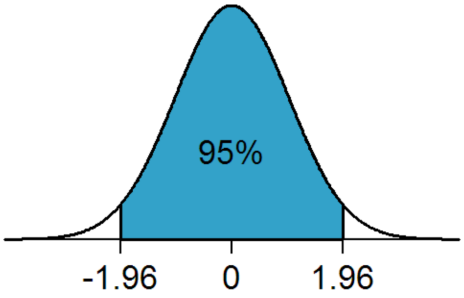


Figure 5.1: From Wikipedia "1.96"

Test Statistics: Important Points

- A test statistic $T : \Omega \to \mathbb{R}$ summarizes characteristics of an experiment outcome $\omega_e \in \Omega$ into one dimensional value $T(\omega_e) \in \mathbb{R}$.

- This summary $T(\omega_e)$ should capture important characteristics of ω_e for testing the null hypothesis H_0 against an alternative H_1.

- At the same time, $T : \Omega \to \mathbb{R}$ should be designed so that the distribution $P_{0,T}$ under the null hypothesis H_0 is easy to compute.

- This is needed to determine the critical region.

5.5 *P*-Value

P-Value

- Hypothesis testing outputs binary decisions ("Reject" or "Not reject") with a pre-specified significance level $\alpha > 0$.

- Recall that a lower value of α implies that the test is more significant, in the sense that the probability of Type 1 Error ($= \alpha$) is smaller.

- The *p*-value provides a continuous measure of statistical significance for an experimental outcome $\omega_e \in \Omega$ against the null hypothesis H_0.

- A lower *p*-value indicates more that the null hypothesis H_0 fails to explain the characteristics of the observed outcome ω_e.

P-Value

Definition of P-Value (Lehmann and Romano, 2005, Section 3.3)

- For each $\alpha > 0$, let $S_\alpha \subset \Omega$ be the critical region for the null hypothesis H_0 such that
$$P_0(S_\alpha) = \alpha.$$
- Assume that the critical regions are nested:
$$S_\alpha \subset S_{\alpha'} \subset \Omega \quad \text{for all } 0 < \alpha < \alpha' < 1$$
- Then the p-value for an experimental outcome ω_e is defined by
$$p\text{-value} := \mathbf{p}(\omega_e) := \min_{\alpha>0} \ \alpha \ \text{ such that } \ \omega_e \in S_\alpha$$
- i.e., the minimum significance level α such that the critical region S_α contains the outcome ω_e.

P-Value

- Note that the p-value depends on

- The definition of the probability distribution P_0 under the null hypothesis H_0;

- The definition of the critical regions S_α, $0 < \alpha < 1$ (i.e., the test).

P-Values for a Test Statistic

- In practice, p-values are defined for a given test statistic T and the distribution P_0 under the null hypothesis H_0.

P-Values for a Test Statistic

- Let $T : \Omega \to \mathbb{R}$ be a test statistic with probability distribution $P_{0,T}$ under the null hypothesis H_0.
- For each $\alpha > 0$, let $I_\alpha \subset \mathbb{R}$ be the critical region such that

$$P_{0,T}(I_\alpha) = \alpha \quad \text{for all } 0 < \alpha < 1.$$

- Assume that the critical regions are nested:

$$I_\alpha \subset I_{\alpha'} \subset \mathbb{R}, \quad 0 < \alpha < \alpha' < 0.$$

- Then the p-value of an observed outcome $\omega_e \in \Omega$ is given by

$$p\text{-value} := \mathbf{p}(\omega_e) = \min_{\alpha > 0} \alpha \ \text{ such that } \ T(\omega_e) \in I_\alpha.$$

P-Values for a Test Statistic

- Since $I_\alpha \subset I_{\alpha'}$ for $\alpha < \alpha'$, we have

$$S_\alpha = \{\omega \in \Omega \mid T(\omega) \in I_\alpha\} \subset \{\omega \in \Omega \mid T(\omega) \in I_{\alpha'}\} = S_{\alpha'}$$

- Thus, I_α being nested implies S_α being nested:

$$I_\alpha \subset I_{\alpha'} \Longrightarrow S_\alpha \subset S_{\alpha'}, \quad 0 < \alpha < \alpha' < 1.$$

- Therefore the definition of the p-value for a test statistic $T : \Omega \to \mathbb{R}$ is consistent with the definition of the p-value with significant regions S_α in the original sample space Ω.

P-Values for a Test Statistic

According to the **American Statistical Association**'s Statement on p-Values (Wasserstein and Lazar, 2016, Section 2):

- *Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data ... would be equal to or more extreme than its observed value.*

P-Values for a Test Statistic

- For instance, assume that the critical region I_α is given by

$$I_\alpha := [c_\alpha, \infty),$$

for constant c_α satisfying

$$c_{\alpha'} < c_\alpha \quad \text{for all } 0 < \alpha < \alpha' < 1$$

so that

$$I_\alpha = [c_\alpha, \infty) \subset [c_{\alpha'}, \infty) = I_{\alpha'}$$

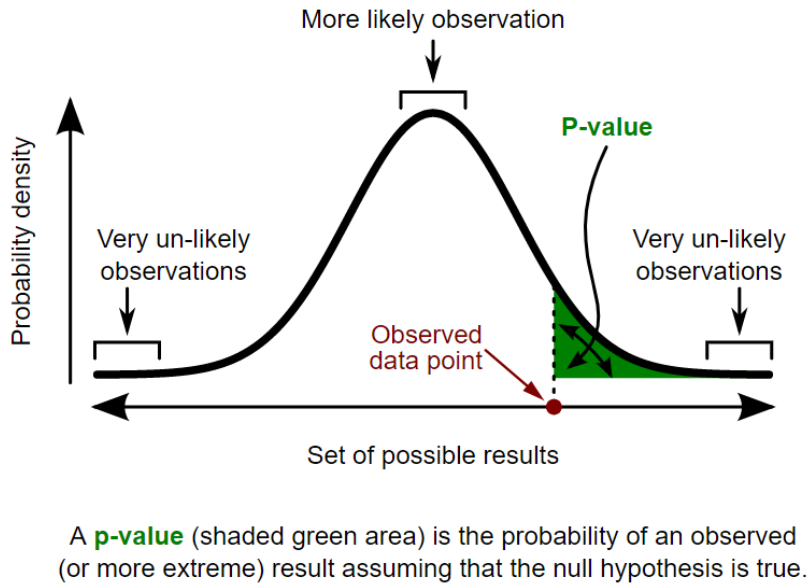


Figure 5.2: This figure illustrates the p-value for one-sided critical region of the form $[c_\alpha, \infty)$. From Wikipedia "p-value".

- Then the p-value is given by

$$\mathbf{p}(\omega_e) = \min_{\alpha>0} \alpha \quad \text{such that } T(\omega_e) \in [c_\alpha, \infty)$$

i.e., the minimum significance level α such that the critical region $[c_\alpha, \infty)$ contains the test statistic $T(\omega_e)$.

Illustration of P-Value

P-Value: Example of the Location Test of a Gaussian Mean

- Consider again the location test of a Gaussian mean.

- We constructed the two-sided critical regions I_α with a significance level $\alpha > 0$ as

$$I_\alpha := (-\infty, -c_\alpha] \cup [c_\alpha, \infty)$$

for a constant $c_\alpha > 0$ satisfying

$$P_{0,T}(I_\alpha) = \int_{-\infty}^{-c_\alpha} p_{0,T}(t)dt + \int_{c_\alpha}^{\infty} p_{0,T}(t)dt = \alpha.$$

- For instance, if $\alpha := 0.05$, we can take $c_\alpha \approx 1.96$.

P-Value: Example of the Location Test of a Gaussian Mean

- Assume that we obtained an experiment outcome $\omega_e := (x_1, \ldots, x_n) \in \Omega$ such that

$$T(\omega_e) = \frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^{n} x_i - \mu_0 \right) = 2.24$$

- In this case, the p-value is given by

$$\begin{aligned}\mathbf{p}(\omega_e) &= \min_{\alpha>0} \alpha \quad \text{such that } T(\omega_e) = 2.24 \in (-\infty, -c_\alpha] \cup [c_\alpha, \infty) \\ &\approx 0.025.\end{aligned}$$

- Thus, the null hypothesis $H_0 : \mu = \mu_0$ would have been rejected if the significance level was set to $\alpha = 0.05$ (since $c_\alpha \approx 1.96$ for $\alpha = 0.05$).

P-Value: Example of the Location Test of a Gaussian Mean

Exercise:

- Derive p-values for the cases where, e.g.,

$T(\omega_e) = 1.26$.

$T(\omega_e) = 3.42$.

- You can for instance use the table from

`https://en.wikipedia.org/wiki/Standard_normal_table`

Interpretation and Use of P-Value

- P-values have been widely used in scientific literature.

- However, the interpretation and use of p-values involve a lot of controversy.

- Ronald Fisher, the advocate of p-values, explains that (Fisher, 1934, Section 20):

  - If P is between 0.1 and 0.9 there is certainly no reason to suspect the hypothesis tested.
  - If it is below 0.02 it is strongly indicated that the hypothesis fails to account for the whole of the facts.

- Here "P" is the p-value, and

- "the hypothesis tested" is the null hypothesis H_0.

Interpretation and Use of P-Value

- The **American Statistical Association**'s Statement on p-Values Wasserstein and Lazar (2016) explains that

  1. P-values can indicate how incompatible the data are with a specified statistical model.
  2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based **only on** whether a p-value passes a specific threshold.

Interpretation and Use of P-Value

4. Proper inference requires full reporting and transparency.
5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

- The statement concludes that "*No single index should substitute for scientific reasoning.*"

- See also e.g. Berger and Sellke (1987); McShane et al. (2019) and references therein.

5.6 Neyman-Pearson Lemma and Likelihood Ratio Test

What is the Most Powerful Test?

- So far we have not discussed how to construct a test statistic.

- A test statistic $T : \Omega \to \mathbb{R}$ and a critical region $I_\alpha \subset \mathbb{R}$ should be constructed so that

- For a given $\alpha > 0$, the Type 1 Error probability is bounded by α

$$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha)) \leq \alpha.$$

- The test power

$$P_{1,T}(I_\alpha) = P_1(T^{-1}(I_\alpha))$$

  is as large as possible,

where P_0 and P_1 are the probability distributions on Ω under the null H_0 and alternative H_1 hypotheses, respectively.

- The question is how to construct a test statistic with a high test power.

What is the Most Powerful Test?

- One answer is provided by the Neyman-Pearson lemma Neyman and Pearson (1933).

- This lemma states that the likelihood ratio test statistic provides the most powerful test.

Likelihood Ratio Test

- Let P_0 and P_1 be the probability distributions on Ω under the null H_0 and alternative H_1 hypotheses, respectively.

- Assume P_0 and P_1 have density functions

$$p_0 : \Omega \to [0, \infty), \quad p_1 : \Omega \to [0, \infty)$$

with respect to a base measure ν (e.g., ν is the Lebesgue measure when $\Omega \subset \mathbb{R}^n$.)

- i.e., for any measurable subset $S \subset \Omega$, we have

$$P_0(S) = \int_S p_0(\omega) d\nu(\omega), \quad P_1(S) = \int_S p_1(\omega) d\nu(\omega).$$

Likelihood Ratio Test

- Define a test statistic $T : \Omega \to [0, \infty)$ by

$$T(\omega) := \frac{p_1(\omega)}{p_0(\omega)}, \quad \omega \in \Omega$$

- This is called the likelihood ratio test statistic.

- Define a test of the form

- Reject the null hypothesis H_0, if $T(\omega_e) \geq c_\alpha$;
- Not reject the null hypothesis H_0, if $T(\omega_e) < c_\alpha$,

where $c_\alpha \geq 0$ is defined so the Type 1 Error probability becomes $\alpha > 0$.

i.e., we define the critical region I_α for the test statistic T as

$$I_\alpha = [c_\alpha, \infty).$$

Neyman-Pearson Lemma

- The Neyman-Pearson Lemma states that

The likelihood ratio test is the most powerful test among all tests with the significance level α.

Neyman-Pearson Lemma

Neyman-Pearson Lemma Neyman and Pearson (1933)

- Define $\alpha > 0$ as the level of significance.
- Let $c_\alpha > 0$ be a constant such that the critical region defined by

$$S^*_\alpha := T^{-1}([c_\alpha, \infty)) = \left\{\omega \in \Omega \mid T(\omega) := \frac{p_1(\omega)}{p_0(\omega)} \geq c_\alpha\right\}$$

  satisfies

$$P_{0,T}([c_\alpha, \infty)) := P_0(S^*_\alpha) = \alpha.$$

- Then the test based on S^*_α has the highest power among all tests with the significance level α;
- i.e., for all $S_\alpha \subset \Omega$ such that $P_0(S_\alpha) = \alpha$, we have

$$P_1(S^*_\alpha) \geq P_1(S_\alpha).$$

Neyman-Pearson Lemma: Proof

- Since $S^*_\alpha \cap S_\alpha \subset S^*_\alpha$, we have

$$P_0(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) = P_0(S^*_\alpha) - P_0(S^*_\alpha \cap S_\alpha) = \alpha - P_0(S^*_\alpha \cap S_\alpha).$$

- Similarly, since $S^*_\alpha \cap S_\alpha \subset S_\alpha$, we have

$$P_0(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)) = P_0(S_\alpha) - P_0(S^*_\alpha \cap S_\alpha) = \alpha - P_0(S^*_\alpha \cap S_\alpha).$$

- Therefore

$$P_0(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) = P_0(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)).$$

Neyman-Pearson Lemma: Proof

- Recall that

$$\frac{p_1(\omega)}{p_0(\omega)} \geq c_\alpha, \quad \forall \omega \in S^*_\alpha, \qquad \frac{p_1(\omega)}{p_0(\omega)} < c_\alpha, \quad \forall \omega \in \Omega \backslash S^*_\alpha$$

- Therefore,

$$p_1(\omega) \geq c_\alpha p_0(\omega), \quad \forall \omega \in S^*_\alpha.$$

- Thus, for any subset $S \subset S^*_\alpha$, we have

$$P_1(S) = \int_S p_1(\omega) d\nu(\omega) \geq \int_S c_\alpha p_0(\omega) d\nu(\omega) = c_\alpha P_0(S).$$

- On the other hand,

$$p_1(\omega) < c_\alpha p_0(\omega), \quad \forall \omega \in \Omega \backslash S^*_\alpha.$$

- Thus, for all $S' \subset \Omega \backslash S^*_\alpha$,

$$P_1(S') = \int_{S'} p_1(\omega) d\nu(\omega) < \int_{S'} c_\alpha p_0(\omega) d\nu(\omega) = c_\alpha P_0(S').$$

Neyman-Pearson Lemma: Proof

- Since

$$S := S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha) \subset S^*_\alpha, \quad S' := S_\alpha \backslash (S^*_\alpha \cap S_\alpha) \subset \Omega \backslash S^*_\alpha,$$

and since

$$P_0(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) = P_0(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)),$$

we have

$$\begin{aligned} &P_1(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) \geq c_\alpha P_0(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) \\ &= c_\alpha P_0(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)) > P_1(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)). \end{aligned}$$

Therefore

$$\begin{aligned} P_1(S^*_\alpha) &= P_1(S^*_\alpha \backslash (S^*_\alpha \cap S_\alpha)) + P_1((S^*_\alpha \cap S_\alpha)) \\ &> P_1(S_\alpha \backslash (S^*_\alpha \cap S_\alpha)) + P_1((S^*_\alpha \cap S_\alpha)) = P_1(S_\alpha). \end{aligned}$$

Thus the proof completes. □

Example: Testing the Location of a Gaussian Mean

- Consider again testing the location of a Gaussian mean.

- Let p^* be an unknown probability density function on $\mathbb{R}$.

- Assume that we know/believe that p^* is Gaussian, with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$:

$$p^*(x) = p_{\text{gauss}}(x; \mu, \sigma_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

- Assume that we can perform an experiment to obtain an i.i.d. sample of size n from p^*:

$$x_1, \ldots, x_n \in \mathbb{R}$$

- Assume that we are interested in testing whether the unknown mean μ is equal to some specified value $\mu_0 \in \mathbb{R}$ or not.

Example: Testing the Location of a Gaussian Mean

- Thus, the null hypothesis H_0 is defined as

$$H_0 : \mu = \mu_0.$$

- For simplicity, we consider a simple alternative hypothesis H_1 where the unknown mean μ is another specified value $\mu_1 \neq \mu_0$:

$$H_1 : \mu = \mu_1.$$

Example: Testing the Location of a Gaussian Mean

- We can define the sample space Ω as

$$\Omega := \mathbb{R}^n.$$

- Each $\omega := (x_1, \ldots, x_n) \in \Omega$ represents a possible experiment outcome of n i.i.d. observations.

- Thus, the distribution P_0 on Ω under the null hypothesis H_0 is given by the density function $p_0 : \Omega \to \mathbb{R}$:

$$\begin{aligned} p_0(\omega) &= \prod_{i=1}^{n} p_{\text{gauss}}(x_i; \mu_0, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x_i - \mu_0)^2}{2\sigma^2}) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}), \quad \omega := (x_1, \ldots, x_n) \in \Omega. \end{aligned}$$

Example: Testing the Location of a Gaussian Mean

- Similarly, the density function p_1 of P_1 under the alternative is given by, for $\omega := (x_1, \ldots, x_n)$,

$$p_1(\omega) = \prod_{i=1}^{n} p_{\text{gauss}}(x_i; \mu_1, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2}{2\sigma^2})$$

- The likelihood ratio test statistic is thus given by, for $\omega := (x_1, \ldots, x_n)$,

$$\begin{aligned} T(\omega) &:= \frac{p_1(\omega)}{p_0(\omega)} = \exp(-\frac{\sum_{i=1}^{n}(x_i - \mu_1)^2 - \sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}) \\ &= \exp(-\frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\mu_1 + \mu_1^2) - \sum_{i=1}^{n}(x_i^2 - 2x_i\mu_0 + \mu_0^2)}{2\sigma^2}) \\ &= \exp(\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2}) \end{aligned}$$

Example: Testing the Location of a Gaussian Mean - Therefore, the test is given by the critical region determined by the threshold

$$\exp(\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2}) \geq c_\alpha$$

where $c_\alpha \geq 0$ is such that we have $P_0(S_\alpha) = \alpha$ for the critical region

$$S_\alpha := \{\omega := (x_1, \ldots, x_n) \in \mathbb{R} \mid T(\omega) \geq c_\alpha\}$$

- Taking the logarithm in the both sides, we have

$$\begin{aligned} &\frac{2(\mu_1 - \mu_0)\sum_{i=1}^{n} x_i - n(\mu_1^2 - \mu_0^2)}{2\sigma^2} \geq \log(c_\alpha) \\ \iff \quad &(\mu_1 - \mu_0)\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right) \end{aligned}$$

Example: Testing the Location of a Gaussian Mean

$$(\mu_1 - \mu_0)\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right)$$

-Thus, if $(\mu_1 - \mu_0) > 0$ (i.e., $\mu_1 > \mu_0$), the rejection threshold is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \frac{1}{2(\mu_1 - \mu_0)}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right) =: r_\alpha$$

- If $(\mu_1 - \mu_0) < 0$ (i.e., $\mu_1 < \mu_0$), the rejection threshold is given by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \frac{1}{2(\mu_1 - \mu_0)}\left(2\sigma^2 \log(c_\alpha) + (\mu_1^2 - \mu_0^2)\right) =: \ell_\alpha$$

Example: Testing the Location of a Gaussian Mean

- Note that, under the null H_0 where $x_1, \ldots, x_n \sim p_{\text{gauss}}(t; \mu_0, \sigma)$ (i.i.d.), we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i \sim p_{\text{gauss}}(t; \mu_0, \sigma^2/n).$$

- Thus, we can derive the rejection threshold r_α

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq r_\alpha$$

directly as r_α satisfying

$$\text{Type 1 Error Probability} = \int_{r_\alpha}^{\infty} p_{\text{gauss}}(t; \mu_0, \sigma^2/n)dt = \alpha.$$

- This shows that the rejection threshold r_α does not depends on the value of μ_1, as long as $\mu_1 > \mu_0$.

Example: Testing the Location of a Gaussian Mean

- This means that the likelihood ratio test is the uniformly most powerful for a composite alternative hypothesis

$$H_1 : \mu > \mu_0$$

- Similarly, if $\mu_1 < \mu_0$ we can derive the threshold ℓ_α as the one satisfying

$$\text{Type 1 Error Probability} = \int_{-\infty}^{\ell_\alpha} p_{\text{gauss}}(t; \mu_0, \sigma^2)dt = \alpha.$$

- This shows that the rejection threshold ℓ_α does not depends on the value of μ_1, as long as $\mu_1 < \mu_0$

- This means that the likelihood ratio test is the uniformly most powerful for a composite alternative hypothesis

$$H_1 : \mu < \mu_0$$

Example: Testing the Location of a Gaussian Mean

- However, this shows that there does not exist a uniformly most powerful test for a composite alternative hypothesis $H_1 : \mu \neq \mu_0$, i.e.,

$$H_1 : \mu < \mu_0 \quad \text{or} \quad \mu_0 < \mu$$

- This is because, when the true unknown mean μ satisfies $\mu > \mu_0$, then the test based on the right rejection threshold

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq r_\alpha$$

is the most powerful,

- while when the true unknown mean μ satisfies $\mu < \mu_0$, then the test based on the left rejection threshold

$$\frac{1}{n}\sum_{i=1}^{n} x_i \leq \ell_\alpha$$

becomes the most powerful.

Important Points to Remember

- The likelihood ratio test depends on how we define an alternative hypothesis.

  - This is true for any test, because the test power (or the Type 2 error) is defined for a given alternative hypothesis.

- For a composite alternative hypothesis (where the alternative contains a variable parameter), there might be no uniformly most powerful test.

- Anyway, the likelihood ratio test and the Neyman-Pearson lemma provides a guideline to design a powerful test.

5.7 Conclusions and Further Reading

Some Key Points to Remember

- To design a test, we need to specify the distribution P_0 on the space Ω of experiment outcomes (or data) under the null hypothesis H_0.

- We should be careful that P_0 may be misspecified.

- For instance, consider the example of testing the location of a Gaussian mean.

- We assumed that the data $\omega = (x_1, \ldots, x_n)$ are i.i.d. with a Gaussian distribution with known variance $\sigma^2 > 0$.

- The knowledge of the variance $\sigma^2 > 0$ is not available in practice, and we need to estimate it from data.
- This requires modifying the testing procedure, and results in the Student t-test.

Some Key Points to Remember

- More generally, the Gaussian assumption itself may be misspecified.

- Under such a misspecification, the Type 1 Error probability

$$P_{0,T}(I_\alpha) = P_0(T^{-1}I_\alpha)$$

may be deviated from a desired level α of significance.

- Thus, in general we should define a null hypothesis H_0 with a weaker assumption about the data distribution P_0.

Some Key Points to Remember

- To derive a critical region $I_\alpha \subset \mathbb{R}$, we need to be able to calculate the probability of I_α under the null H_0

$$P_{0,T}(I_\alpha) = P_0(T^{-1}(I_\alpha)).$$

- This may not be easy in general, in particular when we pose a less restrictive assumption about P_0.

- A modern approach to this purpose is the bootstrap method, developed by Bradley Efron (See (Efron and Hastie, 2016, Section 10)).

- This method uses Monte Carlo (or simulations) to approximate the distribution $P_{0,T}$ under the null.
- The approach can be used for a wide range of problems and easy to implement.

Further Reading

- Again, I recommend you to have a look at Rao (1973).

- The following are recommendations for further reading.

Introduction to Hypothesis Testing and Design of Experiments

Fisher (1934, 1937)

Introduction to the Neyman-Pearson Theory (or the Frequentist Theory)

Neyman and Pearson (1933)

About the Conflicts between the Fisher and Neyman-Pearson Theories

Lehmann (1993) (Efron and Hastie, 2016, Sections 2 and 4)

Further Reading

P-values and Statistical Significance

Berger and Sellke (1987) Wasserstein and Lazar (2016) McShane et al. (2019)

Connections between the Likelihood Ratio Test and the KL Divergence

(Rao, 1973, Section 7a. 3) Eguchi and Copas (2006)

Chapter 6

Introduction to Bayesian Inference

6.1 Introduction and Recap of Parametric Inference

Bayesian Inference: Introduction

- Bayesian inference is another major approach to statistical inference.

- It provides alternatives to the Frequentist and Fisherian approaches (e.g., MLE, hypothesis testing).

- Broad applications, including machine learning (see the "Advanced Statistical Inference" course by Prof Filipopne).

- Available knowledge or prior belief about the estimand (i.e., the quantity of interest in estimation) is expressed as a probability distribution - called the **prior distribution**.

- After observing data, the prior distribution is updated to the **posterior distribution** via **Bayes' rule**.

Bayesian Inference: Introduction

- The major strengths of the Bayesian approach is its coherent approach to quantifying uncertainties.

- The quantified uncertainties are themselves informative, indicating how much our estimates are reliable.

- Moreover, these uncertainties can be used in subsequent tasks, such as

  - Prediction and forecasting.
  - Decision making — including hypothesis testing.
  - Further estimation problems.

Recap: Density Estimation Problem

- Let P be an unknown probability distribution on a measurable set $\mathcal{X} \subset \mathbb{R}^d$.

- Assume that P has a probability density function $p : \mathcal{X} \to \mathbb{R}$.

- Given i.i.d. data $X_1, \ldots, X_n$ from the unknown P, we are interested in estimating the density function p.

- This is the task of density estimation.

Notation

We may write $X_1, \ldots, X_n \sim p$ (i.i.d.) with the density function p.

Recap: Parametric Approach to Density Estimation

- In the parametric approach, we define a parametric model for the unknown density function p generating data $X_1, \ldots, X_n$.

Parametric Model

- Let Θ be a set of parameter vectors (e.g., $\Theta \subset \mathbb{R}^q$).
- For each $\theta \in \Theta$, define a probability density function $p_\theta : \mathcal{X} \to [0, \infty)$.
- A parametric model is defined as the set of such density functions:

$$\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}.$$

Recap: Parametric Approach to Density Estimation

- The parametric model $\mathcal{P}_\Theta$ should be designed so that the unknown density p belongs to $\mathcal{P}_\Theta$, i.e., $p \in \mathcal{P}_\theta$;

  - $p \in P_\theta = \{p_\theta \mid \theta \in \Theta\}$ is equivalent to the existence of some $\theta^* \in \Theta$ such that

$$p = p_{\theta^*} \in \mathcal{P}_\Theta.$$

  - We may call such θ^* the true parameter (vector).

- Therefore the model $\mathcal{P}_\Theta$ should reflect our knowledge/belief about the unknown p.

Recap: Parametric Approach to Density Estimation

- If $p \in \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$, we say that the model $\mathcal{P}_\Theta$ is correctly specified.

  - In this case, estimation of the unknown density $p = p_{\theta^*}$ can be done by estimating the true parameter θ^* from the data $X_1, \ldots, X_n$.

- In the literature on Bayesian inference, this setting is sometimes caled M-closed ("M" stands for Model) Bissiri et al. (2016).

- If $p \notin \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$, we say that the model $\mathcal{P}_\Theta$ is misspecified.

- This setting is called M-open.

6.2 Prior and Posterior Distributions, and Bayes' Theorem

Parameter as a Random Variable

- For now, assume the correctly specified case where there is a "true parameter" $\theta^* \in \Theta$ such that

$$p = p_{\theta^*}.$$

- Given i.i.d. observations $X_1, \ldots, X_n$ from the unknown $p = p_{\theta^*}$, we are interested in making inference about θ^*.

- The core idea of Bayesian inference is to introduce a probability distribution on the parameter space Θ for expressing one's uncertainty about the true parameter θ^*.

Prior Distribution

- Assume that, before observing data, we have some prior belief or knowledge about the true parameter θ^* (e.g., its range).

- We express this as a probability distribution (or density function), called a prior distribution (or prior density function).

- In this lecture, we assume that the prior is given as a probability density function, denoted by

$$\pi(\theta), \quad \theta \in \Theta$$

Example:

- You may know that θ^* should lie in the interval $[-1, 1] \subset \Theta \subset \mathbb{R}$.
- Then we may set a prior density $\pi(\theta)$ as that of the uniform distribution on $[-1, 1]$:

$$\pi(\theta) = \frac{1}{2} 1_{[-1,1]}(\theta), \quad \theta \in \Theta$$

Prior Distribution

- In general, a prior density $\pi(\theta)$ can be interpreted as expressing the whole available knowledge obtained so far, such as data obtained from a previous experiment (we'll see this later).

Updating the Prior with the Likelihood Function

- Assume that i.i.d. data $D_n := (X_1, \ldots, X_n)$ from $p = p_{\theta^*}$ is provided.

- We use this data D_n to update our prior density function $\pi(\theta)$.

- This is done by transforming the data into the likelihood function:

$$\ell_n(\theta) := p(D_n|\theta) := \prod_{i=1}^{n} p_\theta(X_i), \quad \theta \in \Theta.$$

- and by updating the prior $\pi(\theta)$ by multiplying $\ell_n(\theta)$:

$$p_\pi(\theta|D_n) := C^{-1}\ell_n(\theta)\pi(\theta), \quad \theta \in \Theta,$$

where $C > 0$ is a normalization constant to make $p_\pi(\cdot|D_n)$ a density function:

$$\int p_\pi(\theta|D_n)d\theta = \int C^{-1}\ell_n(\theta)\pi(\theta)d\theta = 1.$$

- i.e.,

$$C = \int \ell_n(\theta)\pi(\theta)d\theta.$$

Updating the Prior with the Likelihood Function - The updated density function $p_\pi(\theta|D_n)$ is called the posterior density function.

- This update can be understood by recalling the definition of a conditional density function.

Recap: Joint and Conditional Densities

- Let $\mathcal{Y}$ and Θ measurable sets (e.g., $\mathcal{Y} = \mathcal{X}^n$).

- Assume that we have a joint probability density function $\mathcal{Y} \times \Theta$

$$p(y, \theta), \quad (y, \theta) \in \mathcal{Y} \times \Theta$$

- Then the conditional density function on Θ given $y \in \mathcal{Y}$ is given by

$$p(\theta|y) = \frac{p(y, \theta)}{p(y)}, \quad \theta \in \Theta$$

where $p(y)$ is the marginal density of $y \in \mathcal{Y}$

$$p(y) := \int p(y, \theta)d\theta.$$

- Here we assumed $p(y) > 0$.

Posterior as Conditional Density: Bayes' Theorem

- Recal that the likelihood function $\ell_n(\theta)$ is the conditional probability density function of the data $D_n = (X_1, \ldots, X_n)$ given a parameter $\theta \in \Theta$

$$\ell_n(\theta) = p(D_n|\theta) := \prod_{i=1}^{n} p(X_i|\theta) := \prod_{i=1}^{n} p_\theta(X_i),$$

where we used an interpretation of the model $p_\theta(x)$ as a conditional probability density of $x \in \mathcal{X}$ given $\theta \in \Theta$:

$$p(x|\theta) := p_\theta(x).$$

- Thus, the product of $\ell(\theta)$ and $\pi(\theta)$ can be understood as a joint probability density of the data D_n and the parameter θ:

$$\ell_n(\theta)\pi(\theta) = p(D_n|\theta)\pi(\theta) =: p_\pi(D_n, \theta).$$

Posterior as Conditional Density: Bayes' Theorem

- The posterior density $p_\pi(\theta|D_n)$ is nothing but the conditional probability density induced from this joint density:

$$p_\pi(\theta|D_n) = \frac{p_\pi(D_n, \theta)}{p_\pi(D_n)} = \frac{p(D_n|\theta)\pi(\theta)}{p_\pi(D_n)},$$

where $p_\pi(D_n)$ is the normalization constant

$$p_\pi(D_n) := \int p(D_n|\theta)\pi(\theta)d\theta = \int \ell_n(\theta)\pi(\theta)d\theta$$

- This formula for deriving the posterior $p_\pi(\theta|D_n)$ is called **Bayes' theorem** (or **Bayes' rule**, **Bayes' formula**, etc.)

Posterior as Conditional Density: Bayes' Theorem

- Note that the integral of the normalization constant $p_\pi(D_n)$ becomes the sum, if Θ is countable (i.e., finite or countably infinite), i.e.,

$$p_\pi(D_n) = \sum_{\theta \in \Theta} \ell_n(\theta)\pi(\theta).$$

Point Estimation with the Posterior

- There are mainly two ways of producing a point estimate of the true parameter θ^* from the posterior $p_\pi(\theta|D_n)$.

Maximum a Posteriori (MAP) Estimation

- Define an estimate $\hat{\theta}_{\mathrm{MAP}} \in \Theta$ as a maximizer of the posterior density:

$$\hat{\theta}_{\mathrm{MAP}} \in \arg\max_{\theta \in \Theta} p_\pi(\theta|D_n)$$

Posterior Mean

- Define an estimate $\hat{\theta}_{\text{mean}}$ as the mean of the posterior:

$$\hat{\theta}_{\text{mean}} := \int_{\Theta} \theta \; p_{\pi}(\theta|D_n) d\theta$$

MAP as Regularized Maximum Likelihood Estimation

- Taking the logarithm, and dividing by n, the MAP estimate can also be given as

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &\in \arg\max_{\theta\in\Theta} \frac{1}{n} \log p_{\pi}(\theta|D_n) \\ &= \arg\max_{\theta\in\Theta} \frac{1}{n} \log \ell_n(\theta) + \frac{1}{n}\pi(\theta) \\ &= \arg\max_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(X_i) + \frac{1}{n} \log \pi(\theta).\end{aligned}$$

- The first term is the objective function of MLE.

- Thus, MAP can be understood as a regularized version of MLE, where the regularization term is given by the log prior term, $\frac{1}{n} \log \pi(\theta)$.

MAP as Regularized Maximum Likelihood Estimation

- Since $X_1, \ldots, X_n \sim p$ (i.i.d.), the objective function may be approximated by the integral as

$$\frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(X_i) + \frac{1}{n} \log \pi(\theta) \approx \int p(x) \log p_{\theta}(x) dx + \frac{1}{n} \log \pi(\theta)$$

for a large enough n.

- This suggests that, as n increases, the effect of the prior $\pi(\theta)$ diminishes:

$$\frac{1}{n} \log \pi(\theta) \to 0 \quad \text{as} \quad n \to \infty, \quad \forall \theta \in \Theta$$

- Thus, as n increases, the MAP approaches MLE.

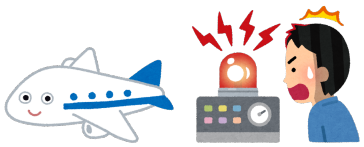
Uncertainty Quantification with Posterior Variance

- The posterior variance is useful in quantifying uncertainty about the true θ^*.

- If $\Theta \subset \mathbb{R}$, the posterior variance can be defined as

$$\hat{\sigma}^2_{\text{post}} := \int_{\Theta} (\theta - \hat{\theta}_{\text{mean}})^2 p_{\pi}(\theta|D_n) d\theta.$$

- This quantifies how much we are uncertain about the parameter, after observing the data (in a Bayesian way).



Example: Airplane Warning Light

- Let's consider a simple example where both $\mathcal{X}$ and Θ are finite sets (Berger, 1985, Example 4 in Section 4.2).

- Assume that you are a pilot of an airplane.

- Let $\Theta := \{\theta_a, \theta_b\}$, and assume that

  - θ_a represents the event that "the landing gear extends"
  - θ_b represents the event that "the landing gear fails to expand".

- Let $\mathcal{X} := \{x_a, x_b\}$, and assume that

  - x_a represents the event that "the warning light is on"
  - x_b represents the event that "the warning light is not on".

Example: Airplane Warning Light

- Assume that the conditional probability density $p(x|\theta)$ is given as

$$p(x_a|\theta_a) = 0.005, \quad p(x_b|\theta_a) = 0.995,$$
$$p(x_a|\theta_b) = 0.999, \quad p(x_b|\theta_b) = 0.001.$$

- From records, the prior density $\pi(\theta)$ is given as

$$\pi(\theta_a) = 0.997 \quad \pi(\theta_b) = 0.003$$

i.e., with probability 0.003, the landing gear fails to expand.

- Assume that we observed x_a, i.e., the warning light is on.

- What is the posterior probability of θ_b (the landing gear fails to expand)?

Example: Airplane Warning Light

- The likelihood function is given by

$$\ell_n(\theta) = p(x_a|\theta) = \begin{cases} 0.005 & \text{if } \theta = \theta_a, \\ 0.999 & \text{if } \theta = \theta_b. \end{cases}$$

- Here $n = 1$, as we only have one observation. — And thus $D_n := (x_a)$.

- Thus, the posterior probability density $p_\pi(\theta|D_n)$ is given by

$$p_\pi(\theta|D_n) = \frac{\ell_n(\theta)\pi(\theta)}{p_\pi(D_n)} = \frac{1}{p_\pi(D_n)} \times \begin{cases} 0.005 \times 0.997 & \text{if } \theta = \theta_a \\ 0.999 \times 0.003 & \text{if } \theta = \theta_b. \end{cases}$$

where the normalization constant is

$$\begin{aligned} p_\pi(D_n) &= \ell_n(\theta_a)\pi(\theta_a) + \ell_n(\theta_b)\pi(\theta_b) \\ &= 0.005 \times 0.997 + 0.999 \times 0.003. \end{aligned}$$

Example: Airplane Warning Light - Thus, the posterior probability is given by

$$p_\pi(\theta|D_n) \approx \begin{cases} 0.62 & \text{if } \theta = \theta_a \\ 0.38 & \text{if } \theta = \theta_b. \end{cases}$$

- In other words, the posterior probability of θ_b (the landing gear fails to extend) is 0.38, given that the warning light is on (i.e., x_a)

– Compare this with the prior probability.

$$\pi(\theta_a) = 0.997 \quad \pi(\theta_b) = 0.003$$

6.3 Normalization Constant, Conjugate Models and Gaussian Examples

Normalization Constant

- Many challenges in Bayesian inference arise from the existence of the normalization constant

$$p_\pi(D_n) = \int p(D_n|\theta)\pi(\theta)d\theta$$

in the denominator of Bayes' theorem

$$p_\pi(\theta|D_n) = \frac{p(D_n|\theta)\pi(\theta)}{p_\pi(D_n)},$$

- This integral is in general does not have an analytic solution.

- If we don't know the normalization constant, then we can only compute the posterior densities up to a constant.

Normalization Constant

- If the normalization constant is not analytically available, we need to approximately compute the posterior ("Bayesian computation").

- Two major approaches exist for this purpose:

1. Sampling from the Posterior with Monte Carlo methods (e.g, importance sampling, Markov Chain Monte Carlo).
2. Variational inference (Approximating the posterior by a parametric model, by optimizing the parameter).

- See Prof Filippone's course "Advanced Statistical Inference" for these methods.

- See also (Gelman et al., 2013, Chapter III).

Normalization Constant

- The normalization constant

$$p_\pi(D_n) := \int p(D_n|\theta)\pi(\theta)d\theta.$$

is also called marginal likelihood or model evidence.

- In fact, this is the likelihood of the data D_n averaged over all possible parameters $\theta \in \Theta$ under the prior.

- This is a very important quantity in Bayesian statistics, since it can be used for model evaluation and model selection.

- A higher value of $p_\pi(D_n)$ suggests that the model $p_\theta(x)$ and the prior $\pi(\theta)$ are supported more by the data (See (Berger, 1985, Chapter 4))

Conjugate Models and Priors

- There are pairs of a model $p_\theta(x)$ and a prior $\pi(\theta)$ for which the normalization constant $p_\pi(D_n)$ is analytically available, called conjugate models and priors.

Conjugate Priors

- $\pi(\theta)$ is called a conjugate prior for the model $p_\theta(x)$, if the posterior $p_\pi(\theta|D_n)$ takes the same functional form as $\pi(\theta)$.

Examples:

- Gaussian likelihood + Gaussian prior
- Binomial (Multinomial) likelihood + Beta (Dirichlet) prior for categorical data.

- For conjugate models, the posteriors are analytically available.

Example: Gaussian Model with a Gaussian Prior

- Consider a Gaussian model

$$p_\theta(x) := p_{\text{gauss}}(x; \theta, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right).$$

- We fix the variance σ^2 and only regard the mean θ as a parameter.

- Assuming that the data $D_n := (X_1, \ldots, X_n)$ are generated i.i.d. from p_{θ^*} for an unknown $\theta^* \in \Theta$, we are interested in making an inference about θ^*.

Example: Gaussian Model with a Gaussian Prior

- We set a prior probability density also as Gaussian:

$$\pi(\theta) := p_{\text{gauss}}(\theta; \mu_0, \sigma_0^2) := \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right).$$

for some $\mu_0 \in \mathbb{R}$ and $\sigma_0^2 > 0$.

- The prior mean μ_0 and variance σ_0^2 should be chosen to reflect your prior belief or knowledge about the true θ^*.

- Then we have

$$\begin{aligned} &p(D_n|\theta)\pi(\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_i - \theta)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right). \end{aligned}$$

Example: Gaussian Model with a Gaussian Prior

- Below C is some constant that does not depend on θ. Then,

$$p(D_n|\theta)\pi(\theta) = C \exp\left(-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma^2} - \frac{(\theta - \mu_0)^2}{2\sigma_0^2}\right).$$

- The exponent can be expanded as

$$\begin{aligned} &-\frac{\sum_{i=1}^n (X_i - \theta)^2}{2\sigma^2} - \frac{(\theta - \mu_0)^2}{2\sigma_0^2} \\ &= \frac{-\theta^2(\sigma_0^2 n + \sigma^2) + 2\theta(\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0) - \sigma_0^2 \sum_{i=1}^n X_i^2 - \sigma^2\mu_0}{2\sigma^2\sigma_0^2} \\ &= \frac{-\theta^2 + 2\theta(\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0)/(\sigma_0^2 n + \sigma^2)}{2\sigma^2\sigma_0^2/(\sigma_0^2 n + \sigma^2)} + C_1 \\ &= \frac{-\left[\theta - (\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0)/(\sigma_0^2 n + \sigma^2)\right]^2}{2\sigma^2\sigma_0^2/(\sigma_0^2 n + \sigma^2)} + C_2, \end{aligned}$$

where C_1 and C_2 are constants not depending on θ.

Example: Gaussian Model with a Gaussian Prior

- Therefore, recalling that the normalization constant $p_\pi(D_n)$ does not depend on θ, we have

$$\begin{aligned} p_\pi(\theta|D_n) &= \frac{p(D_n|\theta)\pi(\theta)}{p_\pi(D_n)} \\ &= C_3 \exp\left(\frac{-\left[\theta - (\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0)/(\sigma_0^2 n + \sigma^2)\right]^2}{2\sigma^2\sigma_0^2/(\sigma_0^2 n + \sigma^2)}\right) \\ &= C_3 \exp\left(-\frac{(\theta-\mu_n)^2}{2\sigma_n^2}\right), \end{aligned}$$

where μ_n and σ_n^2 are defined as

$$\mu_n := \frac{\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0}{\sigma_0^2 n + \sigma^2}, \quad \sigma_n^2 := \frac{\sigma^2\sigma_0^2}{\sigma_0^2 n + \sigma^2}.$$

This shows that $p_\pi(\theta \mid D_n)$ also takes the from of Gaussian.

Example: Gaussian Model with a Gaussian Prior

- Since $p_\pi(\theta \mid D_n)$ is a probability density function, we should have

$$C_3 = \frac{1}{\sqrt{2\pi\sigma_n^2}}$$

and the posterior density for the unknown mean θ is given by

$$p_\pi(\theta|D_n) = p_{\text{gauss}}(\theta; \mu_n, \sigma_n^2) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(\theta-\mu_n)^2}{2\sigma_n^2}\right)$$

with

$$\mu_n := \frac{\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0}{\sigma_0^2 n + \sigma^2}, \quad \sigma_n^2 := \frac{\sigma^2\sigma_0^2}{\sigma_0^2 n + \sigma^2}.$$

Interpreting the Gaussian Example (Large Sample Limit)

- The posterior mean μ_n can be written as

$$\mu_n := \frac{\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2\mu_0}{\sigma_0^2 n + \sigma^2} = \frac{\sum_{i=1}^n X_i + \mu_0\sigma^2/\sigma_0^2}{n + \sigma^2/\sigma_0^2},$$

- Therefore, if we take the large sample limit $n \to \infty$, we have

$$\mu_n \to \frac{1}{n}\sum_{i=1}^n X_i \quad \text{as} \quad n \to \infty$$

- In other words, as $n \to \infty$, the posterior mean μ_n approaches the maximum likelihood estimate

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i.$$

- i.e., the effects of the prior vanish as $n \to \infty$.

Interpreting the Gaussian Example (Large Sample Limit)

- Similarly, the posterior variance σ_n^2 can be written as

$$\sigma_n^2 := \frac{\sigma^2 \sigma_0^2}{\sigma_0^2 n + \sigma^2} = \frac{\sigma^2}{n + \sigma^2/\sigma_0^2}.$$

- And therefore,

$$\sigma_n^2 \to 0 \quad \text{as} \quad n \to \infty.$$

i.e., the posterior variance shrinks towards 0 as we observe more data.

- To summarize, informally we have

$$p_\pi(\theta|D_n) = p_{\text{gauss}}(\theta; \mu_n, \sigma_n^2) \to \delta(\theta - \hat{\mu}) \quad \text{as} \quad n \to \infty,$$

where $\delta(\theta - \hat{\mu})$ is the Dirac distribution at the MLE $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$.

- This suggests that the prior is effective when n is not very large.

Interpreting the Gaussian Example (Uninformative Prior)

- If we take the prior variance σ_0^2 very large, the prior density

$$\pi(\theta) = p_{\text{gauss}}(\theta; \mu_0, \sigma_0^2)$$

becomes uninformative (or noninformative (see (Berger, 1985, Chapter 3) for details of noninfomative priors).

- Let's consider this limit $\sigma_0^2 \to \infty$.

- For the posterior mean μ_n and variance σ_n^2, we have

$$\mu_n = \frac{\sum_{i=1}^n X_i + \mu_0 \sigma^2/\sigma_0^2}{n + \sigma^2/\sigma_0^2} \to \frac{1}{n}\sum_{i=1}^n X_i \quad \text{as} \quad \sigma_0^2 \to \infty,$$

$$\sigma_n^2 = \frac{\sigma^2}{n + \sigma^2/\sigma_0^2} \to \frac{\sigma^2}{n} \quad \text{as} \quad \sigma_0^2 \to \infty.$$

- The latter matches the variance of the empirical average estimator (recall the lecture "Mean Estimation").

Interpreting the Gaussian Example (Strong Prior)

- Let's consider next the limit $\sigma_0 \to 0$.

- i.e. the prior $\pi(\theta) = p_{\text{gauss}}(\theta; \mu_0, \sigma_0)$ is peaky at the prior mean μ_0.

- In this case, we have a strong prior knowledge/belief about the unknown mean θ^*, and new observations do not give more information.

- In fact,

$$\mu_n = \frac{\sigma_0^2 \sum_{i=1}^n X_i + \sigma^2 \mu_0}{\sigma_0^2 n + \sigma^2} \to \mu_0 \quad \text{as} \quad \sigma_0 \to 0,$$
$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma_0^2 n + \sigma^2} \to 0 \quad \text{as} \quad \sigma_0 \to 0.$$

Interpreting the Gaussian Example (Shrinkage Estimator)

- In the "Mean Estimation" lecture, we considered a shrinkage estimator for the mean μ of a distribution P

$$\hat{\mu}_\lambda = \frac{1}{(1+\lambda)} \frac{1}{n} \sum_{i=1}^n X_i.$$

obtained from a regularized least-squares problem

$$\hat{\mu}_\lambda := \arg\min_{\alpha \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (\alpha - X_i)^2 + \lambda \alpha^2,$$

where $\lambda \geq 0$ is a regularization constant.

- Large λ shrinks the solution $\hat{\mu}_\lambda$ towards 0.

- In this sense, this is called a shrinkage estimator.

- $\lambda = 0$ recovers the empirical average $\hat{\mu}_0 = \frac{1}{n} \sum_{i=1}^n X_i$.

Interpreting the Gaussian Example (Shrinkage Estimator)

- There is a Bayesian interpretation of this shrinkage estimator.

- Set the prior mean μ_0 to 0, i.e., $\pi(\theta) = p_{\text{gauss}}(\theta; 0, \sigma_0^2)$.

- Then the posterior density of the mean parameter θ is given by

$$\mu_n = \frac{\sigma_0^2 \sum_{i=1}^n X_i}{\sigma_0^2 n + \sigma^2} = \frac{1}{1 + \sigma^2/(\sigma_0^2 n)} \frac{1}{n} \sum_{i=1}^n X_i$$

- Thus, the shrinkage estimator is equal to the posterior mean, if we set the regularization constant as

$$\lambda = \frac{\sigma^2}{\sigma_0^2 n}$$

- For instance, if σ_0^2 is small, we have a prior belief $\pi(\theta) = p_{\text{gauss}}(0, \sigma_0^2)$ that the true mean θ^* should be close to 0;

- This results in a larger regularization constant λ, leading to more shrinkage towards 0.

6.4 Prior as a Previous Experience

Prior as a Previous Experience

- We'll make here an interpretation of a prior density $\pi(\theta)$ as representing all available knowledge available for problem so far.

- We show an example to describe this interpretation.

Notation

- We often use the following notation

$$p_\pi(\theta \mid D_n) \propto p(D_n \mid \theta)\pi(\theta)$$

to indicate

$$p_\pi(\theta \mid D_n) = C \times p(D_n \mid \theta)\pi(\theta)$$

for some constant C that does not depend on θ.

Prior as a Previous Experience

- As before, assume that i.i.d. observations $X_1, \ldots, X_n$ are available from $p = p_{\theta^*}$, and let

$$D_n := (X_1, \ldots, X_n).$$

- Let $\pi_0(\theta)$ be a prior density function on the parameter space Θ (e.g., consider it as an uninformative prior).

- Assume that another set of i.i.d. observations $Y_1, \ldots, Y_m \sim p$ were independently given previously, and let

$$E_m := (Y_1, \ldots, Y_m).$$

- Then the conditional joint density of D_n and E_m given a parameter θ

$$p(D_n, E_m|\theta) = p(D_n|\theta)p(E_m|\theta)$$

where the decomposition follows from the independence between D_n and E_m.

Prior as a Previous Experience

- Then the posterior density of θ after observing both D_n and E_m is given by

$$\begin{aligned} p_{\pi_0}(\theta|D_n, E_m) &\propto p(D_n, E_m|\theta)\pi_0(\theta) \\ &\propto p(D_n|\theta)p(E_m|\theta)\pi_0(\theta) \\ &\propto p(D_n|\theta)p_{\pi_0}(\theta|E_m), \end{aligned}$$

where $p_{\pi_0}(\theta|E_m)$ is the posterior density of θ after observing only E_m:

$$p_{\pi_0}(\theta|E_m) \propto p(E_m|\theta)\pi_0(\theta).$$

Prior as a Previous Experience

- Consider a situation where we observed E_m previously, and obtained the posterior

$$p_\pi(\theta|E_m) \propto p(E_m|\theta)\pi_0(\theta).$$

with the initial prior π_0.

- We then made the posterior $p_\pi(\theta|E_m)$ as our new updated prior density that incorporates the information of the past data E_m

$$\pi_m(\theta) := p_{\pi_0}(\theta|E_m).$$

- Now we have observed new data $D_n = (X_1, \ldots, X_n)$, and then update the prior $\pi_m(\theta)$ by multiplying the likelihood $p(D_n|\theta)$, obtaining a new posterior:

$$p_{\pi_m}(\theta|D_n) \propto p(D_n|\theta)\pi_m(\theta)$$

Prior as a Previous Experience

- From the above argument, this is equal to the posterior given both D_n and E_m with the initial prior π_0:

$$p_{\pi_0}(\theta|D_n, E_m) = p_{\pi_m}(\theta|D_n), \quad \pi_m(\theta) := p_{\pi_0}(\theta|E_m)$$

- The point here is that, a prior density $\pi(\theta)$ in general may be interpreted as representing all knowledge obtained previously, such as the one

$$\pi(\theta) = \pi_m(\theta) = p_{\pi_0}(\theta|E_m)$$

- For more discussions on priors, see (Berger, 1985, Chapter 3).

6.5 Posterior Distribution via KL-Regularized Loss Minimization

What is the Posterior under Model Misspecification?

- We have so far assumed the correctly specified case where there exists a true parameter $\theta^* \in \Theta$ such that

$$X_1, \ldots, X_n \sim p = p_{\theta^*} \quad (i.i.d.)$$

- In this case, we can interpret the posterior density $p_\pi(\theta|D_n)$ as representing the uncertainty regarding the true parameter θ^*.

- But in practice, the model p_θ is almost always misspecified, i.e.,

$$p \notin \mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}$$

and there may not exist such "true parameter" θ^* that gives rise the true unknown density p

- *How can we interpret the posterior density $p_\pi(\theta|D_n)$ in such a misspecified setting?*

What is the Posterior under Model Misspecification?

- To answer this question, we show here that the posterior density $p_\pi(\theta|D_n)$ can be obtained as a solution to a certain optimization problem involving the KL divergence (see e.g. (Bissiri et al., 2016, Eqs. 7 and 8)).

- Below $p_\pi(\cdot|D_n)$ denotes the posterior density function, where "$\cdot$" denotes the input argument, i.e., the mapping

$$\theta \in \Theta \to p_\pi(\theta|D_n)$$

Posterior via a Density Optimization Problem

Theorem: Posterior as a Loss Minimizer

- Let $D_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$ and $\ell_n(\theta) = \prod_{i=1}^n p(X_i|\theta)$ be the likelihood function.
- Let $\mathcal{P}$ be the set of all probability density functions on Θ.
- Let $\pi(\theta)$ be a prior density, and let $p_\pi(\theta|D_n)$ be the posterior:

$$p_\pi(\theta|D_n) \propto \ell_n(\theta)\pi(\theta), \quad \theta \in \Theta.$$

- Then we have

$$p_\pi(\cdot|D_n) = \arg\min_{q\in\mathcal{P}} - \int q(\theta) \log \ell_n(\theta) d\theta + KL(q\|\pi),$$

where $KL(q\|\pi)$ is the KL divergence between q an π:

$$KL(q\|\pi) := \int q(\theta) \log \frac{q(\theta)}{\pi(\theta)} dx$$

Posterior via a Density Optimization Problem

- Note that the notation

$$p_\pi(\cdot|D_n) = \arg\min_{q\in\mathcal{P}} - \int q(\theta) \log \ell_n(\theta) d\theta + KL(q\|\pi)$$

indicates that the solution of the optimization problem is unique (on the support of $\ell_n(\theta)\pi(\theta)$).

Posterior via a Density Optimization Problem: Proof

- The objective function can be expanded as

$$\begin{aligned}
&\int -\log \ell_n(\theta) q(\theta) dx + KL(q\|\pi) \\
&= \int -\log \ell_n(\theta) q(\theta) dx + \int q(\theta) \log \frac{q(\theta)}{\pi(\theta)} d\theta \\
&= \int q(\theta) \left(-\log \ell_n(\theta) + \log \frac{q(\theta)}{\pi(\theta)} \right) d\theta \\
&= \int q(\theta) \log \frac{q(\theta)}{\ell_n(\theta)\pi(\theta)} d\theta = -\int q(\theta) \log \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta \\
&\geq -\log \left(\int q(\theta) \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta \right) = -\log \left(\int \ell_n(\theta)\pi(\theta) d\theta \right),
\end{aligned}$$

where the inequality follows from Jensen's inequality.

Posterior via a Density Optimization Problem: Proof

Note that the inequality (by Jensen's inequality)

$$-\int q(\theta) \log \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta \geq -\log \left(\int q(\theta) \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta \right)$$

becomes the equality

$$-\int q(\theta) \log \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta = -\log \left(\int q(\theta) \frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} d\theta \right)$$

if and only if

$$\frac{\ell_n(\theta)\pi(\theta)}{q(\theta)} = \text{constant}$$

i.e.,

$$q(x) = p_\pi(\theta|D_n) \propto \ell_n(\theta)\pi(\theta),$$

- Thus, the posterior $q = p_\pi(\cdot|D_n)$ makes the objective function minimal; this complete the proof. □

Posterior via a Density Optimization Problem: Interpretation

- Let's make an interpretation of the result

$$p_\pi(\cdot|D_n) = \arg\min_{q \in \mathcal{P}} -\int q(\theta) \log \ell_n(\theta) d\theta + KL(q\|\pi),$$

- The first term in the objective function is the expected negative log likelihood (which can be seen as an expected loss):

$$\int q(\theta)(-\log \ell_n(\theta)) d\theta,$$

where the expectation is with respect to the density $q(\theta)$.

- This can be written as

$$
\begin{aligned}
&-\int q(\theta)\log \ell_n(\theta)d\theta = -\int q(\theta)\log\prod_{i=1}^{n} p_\theta(X_i)d\theta \\
&= -\int q(\theta)\sum_{i=1}^{n}\log p_\theta(X_i)d\theta = -\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta
\end{aligned}
$$

Posterior via a Density Optimization Problem: Interpretation

- Dividing by n, the optimization problem can be rewritten as

$$
p_\pi(\cdot|D_n) = \arg\min_{q\in\mathcal{P}} -\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta + \frac{1}{n}KL(q\|\pi)
$$

- Since $X_1,\ldots,X_n \sim p$ (i.i.d.), the first term would converge to the integral

$$
\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta \\
&= -\int q(\theta)\left(\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i)\right)d\theta \\
&\approx -\int q(\theta)\left(\int p(x)\log p_\theta(x)dx\right)d\theta
\end{aligned}
$$

Posterior via a Density Optimization Problem: Interpretation

- Thus, for a large enough n, we may have the following approximation for the posterior:

$$
\begin{aligned}
p_\pi(\cdot|D_n) &= \arg\min_{q\in\mathcal{P}} -\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta + \frac{1}{n}KL(q\|\pi) \\
&\approx \arg\min_{q\in\mathcal{P}} -\int q(\theta)\left(\int p(x)\log p_\theta(x)dx\right)d\theta + \frac{1}{n}KL(q\|\pi) \\
&= \arg\min_{q\in\mathcal{P}} -\int q(\theta)\left(\int p(x)\log p_\theta(x)dx\right)d\theta \\
&\quad + \int p(x)\log p(x)dx + \frac{1}{n}KL(q\|\pi),
\end{aligned}
$$

where we added $\int p(x)\log p(x)dx$ in the last line (which does not change the optimization problem, since it does not depend on q).

Posterior via a Density Optimization Problem: Interpretation

- Noting that $\int q(\theta)d\theta = 1$, we have

$$\begin{aligned}
&-\int q(\theta)\left(\int p(x)\log p_\theta(x)dx\right)d\theta + \int p(x)\log p(x)dx \\
&= -\int q(\theta)\left(\int p(x)\log p_\theta(x)dx\right)d\theta + \int q(\theta)\left(\int p(x)\log p(x)dx\right)d\theta \\
&= \int q(\theta)\left(\int p(x)[-\log p_\theta(x) + \log p(x)]dx\right)d\theta \\
&= \int q(\theta)\left(\int p(x)\log\frac{p(x)}{p_\theta(x)}dx\right)d\theta \\
&= \int q(\theta)KL(p\|p_\theta)d\theta.
\end{aligned}$$

Posterior via a Density Optimization Problem: Interpretation

- Thus, for a large enough n, we would have an approximation

$$\begin{aligned}
p_\pi(\cdot|D_n) &= \arg\min_{q\in\mathcal{P}} -\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta + \frac{1}{n}KL(q\|\pi) \\
&\approx \arg\min_{q\in\mathcal{P}} \int q(\theta)KL(p\|p_\theta)d\theta + \frac{1}{n}KL(q\|\pi)
\end{aligned}$$

- The first term

$$\int q(\theta)KL(p\|p_\theta)d\theta$$

measures the average deviation of the model p_θ from the true density p_θ, where the average is taken with respect to $\theta \sim q$.

— This can be interpreted as a loss term for q.

Posterior via a Density Optimization Problem: Interpretation

$$\begin{aligned}
p_\pi(\cdot|D_n) &= \arg\min_{q\in\mathcal{P}} -\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta + \frac{1}{n}KL(q\|\pi) \\
&\approx \arg\min_{q\in\mathcal{P}} \int q(\theta)KL(p\|p_\theta)d\theta + \frac{1}{n}KL(q\|\pi)
\end{aligned}$$

- On the other hand, the second term

$$\frac{1}{n}KL(q\|\pi)$$

measures the discrepancy of q from the prior π.

— This can be interpreted as a regularization term for q.

Posterior via a Density Optimization Problem: Interpretation

$$p_\pi(\cdot|D_n) = \arg\min_{q\in\mathcal{P}} -\frac{1}{n}\sum_{i=1}^{n}\int q(\theta)\log p_\theta(X_i)d\theta + \frac{1}{n}KL(q\|\pi)$$
$$\approx \arg\min_{q\in\mathcal{P}} \underbrace{\int q(\theta)KL(p\|p_\theta)d\theta}_{loss} + \underbrace{\frac{1}{n}KL(q\|\pi)}_{regularization}$$

- Thus, the posterior $p_\pi(\cdot|D_n)$ can be interpreted as the solution to the regularized loss-minimization problem.

— This interpretation is valid in the misspecified setting where

$$p \notin \mathcal{P}_\Theta = \{p_\theta \mid \theta \in \Theta\}$$

- Note that the effect of the prior π (the regularization term) diminishes as $n \to \infty$.

Chapter 7

Bayesian Hypothesis Testing

7.1 Bayesian Hypothesis Testing

Hypothesis Testing

- Assume that we have data $X_1, \ldots, X_n \sim p$ (i.i.d.) from an unknown density p on $\mathcal{X} \subset \mathbb{R}^d$.

- Suppose that we define a parametric model

$$\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}$$

with Θ being a set of parameters (e.g., $\Theta \subset \mathbb{R}^q$).

- Assume the correctly specified case where there exists a "true parameter" $\theta^* \in \Theta$ such that

$$p = p_{\theta^*}.$$

- Then we can consider hypothesis testing regarding the true parameter θ^* from available data $D_n = (X_1, \ldots, X_n)$.

Hypothesis Testing

- Specifically, we can define a null hypothesis H_0 and an alternative hypothesis H_1 as

$$H_0 : \theta^* \in \Theta_0, \quad H_1 : \theta^* \in \Theta_1,$$

where $\Theta_0, \Theta_1 \subset \Theta$ are disjoint subsets, i.e., $\Theta_0 \cap \Theta_1 = \phi$.

e.g., in the case of a point null hypothesis, we may define

$$\Theta_0 := \{\theta_0\}, \quad \Theta_1 := \Theta \backslash \Theta_0 = \{\theta \in \Theta \mid \theta \neq \theta_0\}$$

- The task of hypothesis testing is to decide H_0 or H_1 based on the data D_n.

Priors on Hypotheses

- In Bayesian hypothesis testing, we define prior probabilities on the hypotheses H_0 and H_1, denoted by $\pi(H_0)$ and $\pi(H_1)$, respectively.

- If we don't have a prior knowledge/belief regarding which of H_0 or H_1 should be true, we define the prior as

$$\pi(H_0) = \pi(H_1) = 1/2.$$

- π can be seen as a probability density function on a set $\mathcal{M} = \{H_0, H_1\}$, which consists of two elements (symbols) H_0 and H_1.

Priors on the Parameter Space - Since the null H_0 and alternative H_1 are defined as

$$H_0 : \theta^* \in \Theta_0, \quad H_1 : \theta^* \in \Theta_1$$

we need to define

- a prior density function $\pi_0(\theta) =: \pi(\theta|H_0)$ on Θ_0 for the null H_0; and

  - This represents a prior assumption about the location of $\theta^* \in \Theta_0$ under H_0.

- a prior density function $\pi_1(\theta) =: \pi(\theta|H_1)$ on Θ_1 for the alternative H_1.

  - This represents a prior assumption about the location of $\theta^* \in \Theta_1$ under H_1.

Posteriors on Hypotheses

- After observing data D_n, we apply Bayes' rule to obtain the posterior probabilities of H_0 and H_1, respectively:

- Specifically, for $k = 0, 1$, the posterior $p_\pi(H_k|D_n)$ of hypothesis H_k is

$$p_\pi(H_k|D_n) = \frac{p(D_n|H_k)\pi(H_k)}{\sum_{\ell=0}^{1} p(D_n|H_\ell)\pi(H_\ell)},$$

where $p(D_n|H_k)$ is the marginal likelihood (or the model evidence) of the data D_n given the model H_k:

$$p(D_n|H_k) = \int p(D_n|\theta)\pi(\theta|H_k)d\theta,$$

where $\pi(\theta|H_k)$ is the prior density of the parameter θ under H_k.

Bayes Factor

- Bayesian hypothesis testing is done by computing the Bayes Factor, originally by Harold Jeffreys Jeffreys (1948).

— See also Kass and Raftery (1995) for a good introduction.

Definition: Bayes Factor

- The Bayes factor in favor of H_1 against H_0 is defined by the ratio between the posterior odds $p_\pi(H_0|D_n)/p_\pi(H_1|D_n)$ and prior odds $\pi(H_0)/\pi(H_1)$:

$$B_{10} := \frac{p_\pi(H_0|D_n)/p_\pi(H_1|D_n)}{\pi(H_0)/\pi(H_1)} = \frac{p_\pi(H_0|D_n)\pi(H_1)}{p_\pi(H_1|D_n)\pi(H_0)}$$

- This can be understood as how much the data D_n supports the null H_0 compared to the alternative H_1 relative to the prior assumption.

Bayes Factor

- If $\pi(H_0) = \pi(H_1) = 1/2$, then the Bayes factor becomes the ratio between the posterior probabilities or the ratio between the marginal likelihoods

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} = \frac{p(D_n|H_0)}{p(D_n|H_1)}$$

- If $B_{10} > 1$, then the data D_n supports more H_0 than H_1.

- If $B_{10} < 1$, then the alternative hypothesis H_1 is supported more than the null H_0 by the data (and vice versa).

- Different from p-values in classical testing, the Bayes factor can be interpreted as (the ratio of) the posterior probabilities of the hypotheses.

How to Interpret the Bayes Factor

- How can we interpret the Bayes factor?

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} = \frac{p(D_n|H_0)}{p(D_n|H_1)}$$

- The following is Jeffreys' original explanation (Jeffreys, 1948, Appendix), with minor modifications (e.g., notation).

- *We do not need B_{10} much accuracy.*

  - *Its importance is that if $B_{10} > 1$ the null hypothesis H_0 is supported by the evidence;*
  - *if B_{10} is much less than 1 the null hypothesis H_0 may be rejected.*

How to Interpret the Bayes Factor

- *But B_{10} is not a physical magnitude.*

  - *Its function is to grade the decisiveness of the evidence.*
  - *It makes little difference to the null hypothesis H_0 whether the odds* (i.e., B_{10}) *are 10 to 1 or 100 to 1 against it,*

- *and in practice no difference at all whether they are* 10^4 *or* 10^{10} *to* 1 *against it.*

- *In any case whatever alternative is, most strongly supported will be set up as the hypothesis for use until further notice.*

How to Interpret the Bayes Factor

- $B_{10} = 10^{-1/2}$ *represents only about* 3 *to* 1 *odds, and would be hardly worth mentioning its support of a new discovery* (i.e., alternative H_1).
- *It is at* $B_{10} = 10^{-1}$ *and less that we can have strong confidence that a result* (i.e., H_1) *will survive future investigation .*

How to Interpret the Bayes Factor

- *We may group the values into grades, as follows.*

- Grade 0. $B_{10} > 1$: Null hypothesis H_0 supported.
- Grade 1. $1 > B_{10} > 10^{-1/2}$: Evidence against H_0, but not worth more than a bare mention.
- Grade 2. $10^{-1/2} > B_{10} > 10^{-1}$: Evidence against H_0 substantial.
- Grade 3. $10^{-1} > B_{10} > 10^{-3/2}$: Evidence against H_0 strong.
- Grade 4. $10^{-3/2} > B_{10} > 10^{-2}$: Evidence against H_0 very strong.
- Grade 5. $10^{-2} > B_{10}$: Evidence against H_0 decisive.

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} = \frac{p(D_n|H_0)}{p(D_n|H_1)}$$

How to Interpret the Bayes Factor

- Note that the above grading can be seen as a general guideline, and should not be seen as definite criteria.

- One reason is that the Bayes factor (and the posteriors) depends on the size n of data D_n (also mentioned by Jeffreys.);

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} = \frac{p(D_n|H_0)}{p(D_n|H_1)}$$

— As $n \to \infty$, we have in general $B_{10} \to 0$ or $B_{10} \to \infty$.

- See also Kass and Raftery (1995) for another well-known grading of the Bayes factor.

Computation of the Bayes Factor

- The Bayes factors require the computation of marginal likelihoods.

$$p(D_n|H_k) = \int p(D_n|\theta)\pi(\theta|H_k)d\theta, \quad k = 0, 1.$$

- This integral is not analytically available in general, so we need to approximately compute it in practice.

- Many methods have been proposed (See e.g. Kass and Raftery (1995)):

- Numerical integration or sampling (Monte Carlo, etc.)
- Asymptotic approximation (Bayesian Information Criterion, etc.)

7.2 Example: Testing the Location of a Gaussian Mean

Example: Testing the Location of a Gaussian Mean

- Let's consider testing the location of a Gaussian mean.

- Assume that we have data $X_1, \ldots, X_n \sim p$ (i.i.d.), where

$$p(x) = p_{\text{gauss}}(x; \mu, \sigma^2)$$

is a Gaussian density with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$.

- Our task is to test whether μ is a specified value μ_0 or not (e.g., $\mu_0 = 0$).

- Thus, the null hypothesis H_0 and an alternative H_1 can be defined as

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

- In the previous notation, this corresponds to

$$\Theta := \mathbb{R}, \quad \Theta_0 := \{\mu_0\}, \quad \Theta_1 := \mathbb{R}\backslash\{\mu_0\}.$$

Example: Testing the Location of a Gaussian Mean

- We set equal prior probabilities to H_0 and H_1:

$$\pi(H_0) = \pi(H_1) = 1/2.$$

- For H_0 and H_1, we need to set priors π_0 and π_1 on the parameter spaces Θ_0 and Θ_1, respectively.

- For the null H_0, the parameter space is a singleton set $\Theta_0 = \{\mu_0\}$, and we set

$$\pi_0(\mu_0) = 1.$$

Marginal Likelihood for the Null Hypothesis

- Then the marginal likelihood of the data $D_n = (X_1, \ldots, X_n)$ is

$$\begin{aligned} p_\pi(D_n|H_0) &= \int_{\Theta_0} p(D_n|\mu)\pi_0(\mu)d\mu \\ &= \prod_{i=1}^n p_{\text{gauss}}(X_i; \mu_0, \sigma^2) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2\sigma^2/n}\right), \end{aligned}$$

where $\bar{X}$ is the empirical average of $X_1, \ldots, X_n$:

$$\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i.$$

Marginal Likelihood for the Null Hypothesis

- The last expression follows from

$$\begin{aligned} &\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu_0)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n 2(X_i - \bar{X})(\bar{X} - \mu_0) + \sum_{i=1}^n (\bar{X} - \mu_0)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2, \end{aligned}$$

where the last line follows from

$$\sum_{i=1}^n 2(X_i - \bar{X})(\bar{X} - \mu_0) = 2\left((\sum_{i=1}^n X_i) - n\bar{X}\right)(\bar{X} - \mu_0) = 0.$$

Example: Testing the Location of a Gaussian Mean

- For the alternative H_1, the parameter space is

$$\Theta_1 = \mathbb{R}\backslash\{\mu_0\} = (-\infty, \mu_0) \cup (\mu_0, \infty).$$

- We may define a prior density π_1 as a Gaussian with mean μ_0 and variance $\tau^2 > 0$.

$$\pi_1(\mu) := p_{\text{gauss}}(\mu; \mu_0, \tau^2), \quad \mu \in \mathbb{R} \backslash \{\mu_0\}$$

- Note that this is a valid density function on Θ_1, because

$$\int_{\Theta_1} \pi_1(\mu) d\mu = \lim_{\varepsilon \to +0} \int_{\mu_0+\varepsilon}^{\infty} \pi_1(\mu) d\mu + \int_{-\infty}^{\mu_0-\varepsilon} \pi_1(\mu) d\mu = \frac{1}{2} + \frac{1}{2} = 1.$$

- By setting a large variance τ^2, $\pi_1(\mu)$ tends to "noninformative".

Marginal Likelihood for the Alternative Hypothesis

- It can be shown that the marginal likelihood for the alternative H_1 is

$$\begin{aligned} p_\pi(D_n|H_1) &= \int_{\Theta_1} p(D_n|\mu)\pi_1(\mu) d\mu \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{\sigma^2}{\sigma^2 + \tau^2 n}\right)^{1/2} \\ &\quad \times \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}\right), \end{aligned}$$

where

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i.$$

- See Appendix for the derivation.

Bayes Factor

- Thus, the Bayes factor is given by

$$\begin{aligned} B_{10} &= \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \\ &= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{X}-\mu_0)^2}{2\sigma^2/n}\right)}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{\sigma^2}{\sigma^2+\tau^2 n}\right)^{1/2} \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{X}-\mu_0)^2}{2(\sigma^2/n+\tau^2)}\right)} \\ &= \left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2\sigma^2/n} + \frac{(\bar{X} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}\right) \\ &= \left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2} \left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right), \end{aligned}$$

where

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i.$$

Interpreting the Bayes Factor

- Let's study the obtained Bayes factor:

$$B_{10} = \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)}$$
$$= \left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right).$$

- First notice that, for any possible values of σ^2, τ^2 and n,

$$\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right] > 0.$$

Interpreting the Bayes Factor

- Assume that the empirical mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is very different from the specified value μ_0.

- Then $(\bar{X} - \mu_0)^2$ becomes very large, and thus

$$\exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right)$$

becomes close to 0, implying $B_{10} = \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \approx 0$,

- thus we have a strong evidence against the null $H_0 : \mu = \mu_0$.

(Recall that μ is the unknown true mean of the data distribution).

Interpreting the Bayes Factor: Large Sample Limits

- Assume that the alternative hypothesis $H_1 : \mu \neq \mu_0$ is true: i.e., there exists $\mu_1 \neq \mu_0$ such that

$$X_1, \ldots, X_n \sim p_{\text{gauss}}(x; \mu_1, \sigma^2) \quad (i.i.d.).$$

- Then we have $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \to \mu_1$ as $n \to \infty$, and therefore

$$\frac{(\bar{X} - \mu_0)^2}{2} \to \frac{(\mu_1 - \mu_0)^2}{2} > 0 \quad \text{as} \quad n \to \infty.$$

- On the other hand,

$$\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right] \to +\infty \quad \text{as} \quad n \to \infty$$

- Thus,

$$\exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Interpreting the Bayes Factor: Large Sample Limits

- On the other hand,

$$\left(1+\frac{\tau^2 n}{\sigma^2}\right)^{1/2} \to \infty \quad \text{as} \quad n \to \infty.$$

- But the exponential function decays faster than the polynomial growth, so overall the Bayes factor decays to 0 as $n \to \infty$ exponentially fast:

$$\begin{aligned} B_{10} &= \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \\ &= \underbrace{\left(1+\frac{\tau^2 n}{\sigma^2}\right)^{1/2}}_{\to \infty} \underbrace{\exp\left(-\frac{(\bar{X}-\mu_0)^2}{2}\left[\frac{1}{\sigma^2/n}-\frac{1}{\sigma^2/n+\tau^2}\right]\right)}_{\to 0} \\ &\to 0 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Interpreting the Bayes Factor: Large Sample Limits - Thus, when the alternative $H_1 : \mu \neq \mu_0$ is true, $B_{10} \to 0$ as $n \to \infty$.

- This implies that, if we have large enough data, we will eventually have a strong evidence against the null $H_0 : \mu = \mu_0$.

Interpreting the Bayes Factor: Large Sample Limits

- Assume next that the null hypothesis $H_0 : \mu = \mu_0$ is true.

- In this case, we have (recall the lecture on "Mean Estimation")

$$\mathbb{E}(\bar{X}-\mu_0)^2 = \frac{\sigma^2}{n},$$

and thus

$$\begin{aligned} &\mathbb{E}\left[\frac{(\bar{X}-\mu_0)^2}{2}\left[\frac{1}{\sigma^2/n}-\frac{1}{\sigma^2/n+\tau^2}\right]\right] \\ &= \frac{\sigma^2}{2n}\left[\frac{n}{\sigma^2}-\frac{n}{\sigma^2+\tau^2 n}\right] \\ &= \frac{1}{2}-\frac{\sigma^2}{2(\sigma^2+\tau^2 n)} \to 1/2 \quad \text{as} \quad n \to \infty. \end{aligned}$$

Interpreting the Bayes Factor: Large Sample Limits

- Therefore, we have the following convergence (in probability) as $n \to \infty$:

$$\lim_{n\to\infty} \exp\left(-\frac{(\bar{X}-\mu_0)^2}{2}\left[\frac{1}{\sigma^2/n}-\frac{1}{\sigma^2/n+\tau^2}\right]\right) = \exp\left(-\frac{1}{2}\right) > 0.$$

- Therefore, as $n \to \infty$, the Bayes factor diverges to infinity polynomially fast

$$\begin{aligned} B_{10} &= \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \\ &= \underbrace{\left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2}}_{\to \infty} \underbrace{\exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right)}_{\to \exp(-1/2)} \\ &\to \infty \quad \text{as} \quad n \to \infty. \end{aligned}$$

- Thus, when the null $H_0 : \mu = \mu_0$ is true, $B_{10} \to \infty$ as $n \to \infty$.

- This implies that the null will be supported as we observe more data.

Interpreting the Bayes Factor: Large Sample Limits

- Note that

- the speed of $B_{10} \to 0$ when the alternative H_1 is true is **exponential**
- the speed of $B_{10} \to \infty$ when the null H_0 is true is **polynomial**

- This means that we may need more data to conclude that the null H_0 is true when H_0 is true (compared to the data size for concluding that the alternative H_1 is true when H_1 is true).

- This **asymmetry** is due to the use of the Gaussian prior in the alternative $H_1 : \mu \neq \mu_0$:

$$\pi_1(\mu) = p_{\text{gauss}}(\mu; \mu_0, \tau^2)$$

- This motivated Johnson and Rossell (2010) to consider non-local alternatives that behave more nicely; see the paper if you are interested.

Interpreting the Bayes Factor: Changing the Prior Variance - Next, let's study the behavior of the Bayes factor

$$\begin{aligned} B_{10} &= \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \\ &= \left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right). \end{aligned}$$

when we change the prior variance τ^2 of the alternative $H_1 : \mu \neq \mu_0$:

$$\pi_1(\mu) = p_{\text{gauss}}(\mu; \mu_0, \tau^2)$$

- First, consider the limit $\tau^2 \to 0$ (the prior shrinks to a point mass at μ_0).

- In this limit, the alternative H_1 approaches the null $H_0 : \mu = \mu_0$, and the Bayes factor converges to 1 (trivial case)

$$B_{10} \to 1, \quad \text{as} \quad \tau^2 \to 0$$

Interpreting the Bayes Factor: Changing the Prior Variance

- Let's consider the other limit $\tau^2 \to \infty$;

- In this case, the prior $\pi_1(\mu) = g_{\text{gauss}}(\mu; \mu_0, \tau^2)$ is getting flat and thus becoming "noninformative" as $\tau^2 \to \infty$.

- However, the Bayes factor diverges to infinity as $\tau^2 \to \infty$:

$$\begin{aligned} B_{10} &= \frac{p_\pi(D_n|H_0)}{p_\pi(D_n|H_1)} \\ &= \left(1 + \frac{\tau^2 n}{\sigma^2}\right)^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2}\left[\frac{1}{\sigma^2/n} - \frac{1}{\sigma^2/n + \tau^2}\right]\right) \\ &\to \infty \quad \text{as} \quad \tau^2 \to \infty. \end{aligned}$$

- Thus, if we use the noninformative prior, the Bayes factor always supports the null hypothesis $H_0 : \mu = \mu_0$, regardless of what data $X_1, \ldots, X_n$ are.

Interpreting the Bayes Factor: Changing the Prior Variance

- This happens because the marginal likelihood for the alternative H_1 converges to 0 as $\tau^2 \to \infty$:

$$\begin{aligned} p_\pi(D_n|H_1) &= \int_{\Theta_1} p(D_n|\mu)\pi_1(\mu)d\mu \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \left(\frac{\sigma^2}{\sigma^2 + \tau^2 n}\right)^{1/2} \\ &\quad \times \exp\left(-\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{2\sigma^2}\right) \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}\right) \\ &\to 0 \quad \text{as} \quad \tau^2 \to \infty. \end{aligned}$$

- In other words, the marginal likelihood $p_\pi(D_n|H_1)$ becomes 0 if the prior $\pi_1(\mu)$ is noninformative.

- As shown in this example, noniformative priors cannot be used in Bayesian testing in general (different from Bayesian estimation).

7.3 Hypotheses as Models

How to Define Priors for Model Parameters?

- As we saw, Bayesian testing requires specifying prior probabilities $\pi_0(\theta)$ and $\pi_1(\theta)$ for the null H_0 and alternative H_1, respectively.

- How can we choose these priors?

- Let's look at explanations by Harold Jeffreys, who proposed the approach Jeffreys (1948).

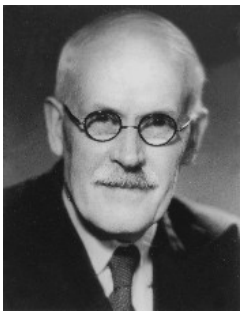


Figure 7.1: Harold Jeffreys (from Wikipedia)

How to Define Priors for Model Parameters?

- To this end, consider again the Gaussian example where observations are given as

$$X_1, \ldots, X_n \sim p(x; \mu, \sigma^2) \quad (i.i.d.)$$

with unknown mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$, and hypotheses are given by

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

for some specified value $\mu_0 \in \mathbb{R}$.

- For the alternative H_1, the parameter space is

$$\Theta_1 = \mathbb{R} \backslash \{\mu_0\} = (-\infty, \mu_0) \cup (\mu_0, \infty).$$

- We may define a prior density π_1 as a Gaussian with mean μ_0 and variance $\tau^2 > 0$.

$$\pi_1(\mu) := p_{\text{gauss}}(\mu; \mu_0, \tau^2), \quad \mu \in \mathbb{R} \backslash \{\mu_0\}$$

Jeffreys' Explanation (Jeffreys, 1948, pp. 225-226)

- *The situation appears to be that when a suggestion arises that calls for a significance test there may be very little information or a great deal.*

- i.e., when people are interested in testing, there would usually be some prior information (very little or a great deal) about a possible alternative H_1.

- *In sampling problems the suggestion that the whole class is of one type may arise before any individual at all has been examined.*

- i.e., when we don't have any prior information, any value $\mu \neq \mu_0$ seems likely to be true. $\Longrightarrow \tau^2$ should be large.

Jeffreys' Explanation (Jeffreys, 1948, pp. 225-226)

- *In the establishment of Kepler's laws, several alternatives had to be discussed and found to disagree wildly with observation before the right solutions were found, and by the time when perturbations began to be investigated theoretically, the extent of departures from Kepler's laws was reasonably well known, and well beyond the standard error of one observation.*

- Kelpler's laws = μ_0;
- alternatives = possible alternative values $\mu_1 \neq \mu_0$
- the extent of departures = $|\mu_1 - \mu_0|$;
- ... was well known and well beyond the standard error of observation $\Longrightarrow$ we have prior information that $|\mu_1 - \mu_0| > \sigma$.
- $\Longrightarrow$ we may set $\tau^2 = C\sigma^2$ for some $C > 1$.

Jeffreys' Explanation (Jeffreys, 1948, pp. 225-226)

- *In experimental physics it usually seems to be expected that there will be systematic error comparable with the standard error of one observation*

- $\Longrightarrow |\mu_1 - \mu_0| \approx \sigma$.
- $\Longrightarrow$ we may set $\tau^2 = C\sigma^2$ for some $C \approx 1$.

- *... In any of these cases it would be perfectly possible to give a form of $\pi_1(\mu)$ that would express the previous information satisfactorily, ...*

Hypotheses as Models

- Jeffreys' point is that the prior $\pi_1(\mu)$ should be decided based on prior information about possible alternative hypotheses.

- e.g., if we know that the mean μ_1 of the alternative hypothesis should satisfy

$$(\mu_1 - \mu_0)^2 \leq C\sigma^2 \quad \text{for some} \quad C > 0,$$

then we could determine a reasonable value for the prior variance τ^2 based on this prior information.

- In this sense, we can regard the prior $\pi_1(\mu)$ as a part of the model under the alternative H_1.

Hypotheses as Models

- Indeed, the Bayes factor is given as the ratio of two marginal likelihoods $p(D_n|H_0)$, $p(D_n|H_1)$ (assuming $\pi(H_0) = \pi(H_1) = 1/2$):

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} = \frac{p(D_n|H_0)}{p(D_n|H_1)} = \frac{\int_{\Theta_0} p(D_n|\theta)\pi_0(\theta)d\theta}{\int_{\Theta_1} p(D_n|\theta)\pi_1(\theta)d\theta}$$

- Thus, Bayesian testing may be understood as a comparison two marginal density functions on $\mathcal{X}$.

$$p_{H_0}(x) := \int_{\Theta_0} p_\theta(x)\pi_0(\theta)d\theta$$
$$p_{H_1}(x) := \int_{\Theta_1} p_\theta(x)\pi_1(\theta)d\theta,$$

where $\pi_0(\theta)$ and $\pi_1(\theta)$ are prior densities under H_0 and H_1, respectively.

(Note that we use the notation for the generic formulation again here).

Hypotheses as Models

- These marginal densities p_{H_0} and p_{H_1} can be thought of as competing models for the true unknown density p.

- Recall that our data $D_n = (X_1, \ldots, X_n)$ are assumed to be generated as

$$X_1, \ldots, X_n \sim p \quad (i.i.d.).$$

- From now on, we don't assume that the model

$$\mathcal{P}_\Theta := \{p_\theta \mid \theta \in \Theta\}$$

is correctly specified; we allow for the misspecified case where

$$p \notin \mathcal{P}_\Theta.$$

Large Sample Behavior of the Bayes Factor

- To see how the Bayes factor is related to the marginal densities p_{H_0}, p_{H_1}, let's study the asymptotic behavior of the posterior density

$$p_\pi(\cdot|D_n) \quad \text{as} \quad n \to \infty$$

- Note that $p_\pi(\cdot|D_n)$ is a probability density function on the set

$$\mathcal{M} := \{H_0, H_1\}$$

which consists only of two elements (or two symbols).

- Let $\mathcal{P}$ be the set of all probability density functions on $\mathcal{M}$; i.e., each $q \in \mathcal{P}$ satisfies

$$q(H_0) + q(H_1) = 1, \quad 0 \le q(H_0), q(H_1) \le 1.$$

Large Sample Behavior of the Bayes Factor

- Then, for a large enough n, the posterior $p_\pi(\cdot|D_n)$ may be given as an approximation to the following optimization problem on $\mathcal{P}$:

$$p_\pi(\cdot|D_n) \approx \arg\min_{q \in \mathcal{P}} \sum_{k=0}^{1} q(H_k) KL(p\|p_{H_k}) + \frac{1}{n} KL(q\|\pi),$$

(recall the lecture on "Bayesian Inference I"), where

$$KL(p\|p_{H_k}) := \int p(x) \log \frac{p(x)}{p_{H_k}(x)} dx,$$
$$KL(q\|\pi) := \sum_{k=0}^{1} q(H_k) \log \frac{q(H_k)}{\pi(H_k)}$$

are KL divergences.

- Note that the effect of the second term $\frac{1}{n} KL(q\|\pi)$ decreases as $n \to \infty$.

Large Sample Behavior of the Bayes Factor

- Thus, as n increases, $p_\pi(\cdot|D_n)$ would approach the following minimizer

$$p_\pi(\cdot|D_n) \to q^* := \arg\min_{q \in \mathcal{P}} \sum_{k=0}^{1} q(H_k) KL(p\|p_{H_k}) \quad \text{as} \quad n \to \infty$$

- Assume that the true density p is closer to p_{H_0} than to p_{H_1} in terms of the KL divergence:

$$KL(p\|p_{H_0}) < KL(p\|p_{H_1})$$

- Then the minimizer q^* should satisfy

$$q^*(H_0) = 1, \quad q^*(H_1) = 0.$$

- Therefore, in this case we would have

$$p_\pi(H_0|D_n) \to 1, \quad p_\pi(H_1|D_n) \to 0, \quad \text{as} \quad n \to \infty.$$

Large Sample Behavior of the Bayes Factor

- Thus the Bayes factor would diverge to infinity

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} \to \infty$$

supporting the null hypothesis H_0.

- On the other hand, if

$$KL(p\|p_{H_0}) > KL(p\|p_{H_1})$$

we would have

$$B_{10} = \frac{p_\pi(H_0|D_n)}{p_\pi(H_1|D_n)} \to 0 \quad \text{as} \quad n \to \infty$$

supporting the alternative H_1.

Bayesian Hypothesis Testing as Model Comparison - From these discussions, we can make the following observations:

- In Bayesian testing, we compare two marginal density functions (or models) p_{H_0} and p_{H_1} in terms of the **goodness-of-fit** to the true unknown density function p, as measured by the KL divergence:

$$KL(p\|p_{H_0}) \quad \text{v.s.} \quad KL(p\|p_{H_1})$$

- The hypothesis H_k $(k = 0, 1)$ that gives better goodness of fit to the true density p will be supported by the test.

- See Chib and Kuffner (2016); Johnson and Rossell (2010) and references therein for more formal theoretical results on Bayes factor consistency.

7.4 Further Topics and Recommended Reading

Further Topics and Recommended Reading

- There are several extensions of Bayesian testing and model comparison, such as

- Incorporation of the information of losses (or utility) associated with possible errors.
- Straightforward extension to testing with multiple hypotheses.
- Comparison of models with different structures.

- See (Berger, 1985, Section 4), (Gelman et al., 2013, Part II) and Kass and Raftery (1995).

- For comparisons between Bayesian and classical Testing, see Berger and Sellke (1987); Berger et al. (2003).

7.5 Appendix for the Gaussian Example

Marginal Likelihood for the Alternative

- The marginal likelihood $p_\pi(D_n|H_1)$ is given by

$$\begin{aligned}
&p_\pi(D_n|H_1) \\
&= \int_{\Theta_1} p(D_n|\theta)\pi_1(\theta)d\theta = \int_{\Theta} p(D_n|\theta)\pi_1(\theta)d\theta \\
&= \int_{\Theta} \left(\prod_{i=1}^{n} p_{\text{gauss}}(X_i; \mu', \sigma^2)\right) p_{\text{gauss}}(\mu'; \mu_0, \tau^2)d\mu' \\
&= \int_{\Theta} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \mu')^2}{2\sigma^2}\right) (2\pi\tau^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu' - \mu_0)^2}{2\tau^2}\right) d\mu' \\
&= (2\pi)^{-\frac{n+1}{2}} \sigma^{-n}\tau^{-1} \int_{\Theta} \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \mu')^2}{2\sigma^2} - \frac{(\mu' - \mu_0)^2}{2\tau^2}\right) d\mu'
\end{aligned}$$

Marginal Likelihood for the Alternative

- Let $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical average.

- Then

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu')^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu')^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}2(X_i-\bar{X})(\bar{X}-\mu')+\sum_{i=1}^{n}(\bar{X}-\mu')^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu')^2,\end{aligned}$$

where the last line follows from

$$\sum_{i=1}^{n}2(X_i-\bar{X})(\bar{X}-\mu') = 2\left((\sum_{i=1}^{n}X_i)-n\bar{X}\right)(\bar{X}-\mu') = 0.$$

Marginal Likelihood for the Alternative

- Thus,

$$\begin{aligned}&\int \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\mu')^2}{2\sigma^2}-\frac{(\mu'-\mu_0)^2}{2\tau^2}\right)d\mu'\\ &= \int \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu')^2}{2\sigma^2}-\frac{(\mu'-\mu_0)^2}{2\tau^2}\right)d\mu'\\ &= \int \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2}-\frac{(\bar{X}-\mu')^2}{2\sigma^2/n}-\frac{(\mu'-\mu_0)^2}{2\tau^2}\right)d\mu'\\ &= \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2}\right)\int \exp\left(-\frac{1}{2}\left[\frac{(\bar{X}-\mu')^2}{\sigma^2/n}-\frac{(\mu'-\mu_0)^2}{\tau^2}\right]\right)d\mu'\end{aligned}$$

Marginal Likelihood for the Alternative

- Therefore

$$\begin{aligned}&p_\pi(D_n|H_1)\\ &= (2\pi)^{-\frac{n+1}{2}}\sigma^{-n}\tau^{-1}\int_\Theta \exp\left(-\frac{\sum_{i=1}^{n}(X_i-\mu')^2}{2\sigma^2}-\frac{(\mu'-\mu_0)^2}{2\tau^2}\right)d\mu'\\ &= (2\pi)^{-\frac{n+1}{2}}\sigma^{-n}\tau^{-1}\exp\left(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2}\right)\\ &\quad\times\int \exp\left(-\frac{1}{2}\left[\frac{(\bar{X}-\mu')^2}{\sigma^2/n}-\frac{(\mu'-\mu_0)^2}{\tau^2}\right]\right)d\mu'\end{aligned}$$

Marginal Likelihood for the Alternative

- Let $\eta^2 := \sigma^2/n$ and

$$\rho := \frac{1}{\eta^2} + \frac{1}{\tau^2} = \frac{\eta^2 + \tau^2}{\eta^2 \tau^2}$$

- Then,

$$\begin{aligned}
&\frac{(\bar{X} - \mu')^2}{\sigma^2/n} + \frac{(\mu' - \mu_0)^2}{\tau^2} \\
&= \frac{(\bar{X} - \mu')^2}{\eta^2} + \frac{(\mu' - \mu_0)^2}{\tau^2} \\
&= \left(\frac{1}{\eta^2} + \frac{1}{\tau^2}\right)\mu'^2 - 2\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\mu' + \left(\frac{\bar{X}^2}{\eta^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&= \rho\left[\mu'^2 - \frac{2}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\mu'\right] + \left(\frac{\bar{X}^2}{\eta^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&= \rho\left[\mu' - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\right]^2 - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)^2 + \left(\frac{\bar{X}^2}{\eta^2} + \frac{\mu_0^2}{\tau^2}\right)
\end{aligned}$$

Marginal Likelihood for the Alternative - Then,

$$\begin{aligned}
&\frac{(\bar{X} - \mu')^2}{\sigma^2/n} + \frac{(\mu' - \mu_0)^2}{\tau^2} \\
&= \rho\left[\mu' - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\right]^2 - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)^2 + \left(\frac{\bar{X}^2}{\eta^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&= \rho\left[\mu' - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\right]^2 + \frac{(\bar{X} - \mu_0)^2}{\eta^2 + \tau^2},
\end{aligned}$$

where the last line can be show by a straightforward calculation.

Marginal Likelihood for the Alternative - Thus,

$$\begin{aligned}
&\int \exp\left(-\frac{1}{2}\left[\frac{(\bar{X} - \mu')^2}{\sigma^2/n} - \frac{(\mu' - \mu_0)^2}{\tau^2}\right]\right) d\mu' \\
&= \int \exp\left(-\frac{\rho}{2}\left[\mu' - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\right]^2 - \frac{(\bar{X} - \mu_0)^2}{2(\eta^2 + \tau^2)}\right) d\mu' \\
&= \int \exp\left(-\frac{\rho}{2}\left[\mu' - \frac{1}{\rho}\left(\frac{\bar{X}}{\eta^2} + \frac{\mu_0}{\tau^2}\right)\right]^2\right) d\mu' \times \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2(\eta^2 + \tau^2)}\right) \\
&= (2\pi\rho^{-1})^{1/2} \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2(\eta^2 + \tau^2)}\right) \\
&= (2\pi)^{1/2}(\sigma^2 + \tau^2 n)^{-1/2}\tau\sigma \exp\left(-\frac{(\bar{X} - \mu_0)^2}{2(\sigma^2/n + \tau^2)}\right).
\end{aligned}$$

Marginal Likelihood for the Alternative - Therefore

$$
\begin{aligned}
& p_\pi(D_n|H_1) \\
&= (2\pi)^{-\frac{n+1}{2}}\sigma^{-n}\tau^{-1}\exp\left(-\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{2\sigma^2}\right) \\
&\quad \times \int \exp\left(-\frac{1}{2}\left[\frac{(\bar{X}-\mu')^2}{\sigma^2/n}-\frac{(\mu'-\mu_0)^2}{\tau^2}\right]\right)d\mu' \\
&= (2\pi)^{-\frac{n+1}{2}}\sigma^{-n}\tau^{-1}\exp\left(-\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{2\sigma^2}\right) \\
&\quad \times (2\pi)^{1/2}(\sigma^2+\tau^2 n)^{-1/2}\tau\sigma\exp\left(-\frac{(\bar{X}-\mu_0)^2}{2(\sigma^2/n+\tau^2)}\right) \\
&= (2\pi\sigma^2)^{-n/2}\sigma(\sigma^2+\tau^2 n)^{-1/2} \\
&\quad \times \exp\left(-\frac{\sum_{i=1}^n(X_i-\bar{X})^2}{2\sigma^2}\right)\times\exp\left(-\frac{(\bar{X}-\mu_0)^2}{2(\sigma^2/n+\tau^2)}\right).
\end{aligned}
$$

Bibliography

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Berger, J. O. et al. (2003). Could fisher, jeffreys and neyman have agreed on testing? *Statistical Science*, 18(1):1–32.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5):1103.

Chib, S. and Kuffner, T. A. (2016). Bayes factor consistency. *arXiv preprint arXiv:1607.00292*.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Eguchi, S. and Copas, J. (2006). Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, 97(9):2034–2040.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368.

Fisher, R. A. (1934). *Statistical Methods for Research Workers (Fifth Edition)*. Oliver & Boyd (Edinburgh).

Fisher, R. A. (1937). *Design of Experiments (second edition)*. Macmillan.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gray, R. M. (2011). *Entropy and Information Theory*. Springer Science & Business Media.

Jeffreys, H. (1948). *The Theory of Probability (Second Edition)*. OUP Oxford.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Lehmann, E. L. (1993). The fisher, neyman-pearson theories of testing hypotheses: one theory or two? *Journal of the American statistical Association*, 88(424):1242–1249.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses (Third Edition)*. Springer Science & Business Media.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley New York.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall.

Sugiyama, M. and Kawanabe, M. (2012). *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, MA, USA.

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70(2).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25.