Contents

[Unit 1 Introduction to Statistics 1](#_Toc48424962)

[Lecture 2 Probability Redux 1](#_Toc48424963)

[Unit 2 Foundation of Inference 2](#_Toc48424964)

[Lecture 3 Parametric Statistical Models 2](#_Toc48424965)

[Lecture 4 Parametric Estimation and Confidence Intervals 3](#_Toc48424966)

[Lecture 5 Delta Method and Confidence Intervals 3](#_Toc48424967)

[Recitation 2 4](#_Toc48424968)

[Lecture 6 Hypothesis Testing, Type I, II Errors 5](#_Toc48424969)

[Lecture 7 Hypothesis Testing and p-value 5](#_Toc48424970)

[Recitation 4 5](#_Toc48424971)

[Unit 3 Method of Estimation 6](#_Toc48424972)

[Lecture 8 Distance Measures between Distributions 6](#_Toc48424973)

[Lecture 9 Introduction to MLE 7](#_Toc48424974)

[Lecture 10 Consistency of MLE 7](#_Toc48424975)

[Recitation 6 8](#_Toc48424976)

[Lecture 11 Fisher Information, Asymptotic Normality of MLE; Method of Moments 8](#_Toc48424977)

# Unit 1 Introduction to Statistics

## Lecture 2 Probability Redux

1. let be i.i.d. random variables, with then:

**Law of large numbers** (weak and strong):

**Central Limit Theorem:**

1. Markov inequality: For with mean , and , we have
2. Chebyshev inequality: For with mean , variance , and we have
3. Modes of convergence
   1. Almost surely () convergence
   2. Convergence in probability
   3. Convergence in distribution

for all at which the of is continuous

* 1. Convergence in

1. Important properties of convergence in distribution. The following are equivalent:
   1. converges to in probability
   2. , for all continuous and bounded function
   3. for all
2. Convergence ;
3. If then

   2. If then

In general, these rules do not apply to convergence in distribution (d).

1. Slutsky’s theorem

Let be 2 sequences of random variables, such that

where is a random variable, and is a real number.

Then:

* 1. If then

1. Continuous mapping theorem: if is a continuous function, then . Note: not apply to convergence in Lp.

# Unit 2 Parametric Inference

## Lecture 3 Parametric Statistical Models

1. Let the observed outcome of a statistical experiment be a sample of random variables in some measurable space (usually ) and denoted by their common distribution.

A **statistical model** associated to that statistical experiment is a pair

where:

* is called sample space
* is a family of probability measures on   
  is any set, called parameter set

1. A statistical model is **well specified** if such that
2. We often assume for some , and the model is called parametric; sometimes we could have be infinite dimensional, in which case the model is called nonparametric.
3. The parameter is **identifiable** if , i.e., from the distribution we can uniquely determine .

## Lecture 4 Parametric Estimation and Confidence Intervals

1. A statistic is any measurable function of the sample.
2. An estimator of is a statistic whose expression does not depend on .
3. An estimator of is weakly (resp. strongly) consistent if
4. An estimator of is asymptotically normal if

where is called asymptotic variance of

1. The quadratic risk of is

Note: If quadratic risk goes to 0 as , then converges in L2 to , and is (weakly) consistent.

1. Let be a statistical model based on observations , and assume .

Confidence interval (C.I.) of level for : any random interval whose boundaries do not depend on such that

Confidence interval of asymptotic level for : any random interval whose boundaries do not depend on such that

1. Example: construct C.I. of for model . Let

According to CLT, Thus However, it is not a confidence interval because it depends on . We have three solutions.

* 1. Conservative bound: as , we can use a conservative bound by replacing p with
  2. Solve the quadratic function to get exact solution.
  3. By LLN + CMT + Slutsky, . Plug in to replace p, the C.I. is

## Lecture 5 Delta Method and Confidence Intervals

1. If then

However, it is a biased estimator.

Consistency does not imply unbiased.

1. Delta Method: Letbe a sequence of random variables such that

for some and .

Let be continuously differentiable at Then

Informal Proof:

1. Applying Delta method to the previous problem, we can get the asymptotical distribution of

## Recitation 2

1. Claim: If , then

Proof:

1. Claim: If then

Proof: Fix ,

1. Claim: If then

Proof: which is cdf of

1. Claim: If , and

Then: but does not hold.

Proof: we can assume . , so converge in probability.

However, convergence does not hold, because does not converge to any number

## Lecture 6 Hypothesis Testing, Type I, II Errors

1. Consider a statistical model . Let and be disjoint subsets of , and consider the 2 hypotheses:

is null hypothesis, is alternative hypothesis.

1. A test is a statistic
   * If is not rejected
   * If is rejected
2. Rejection region of a test :

Type I error of a test :

Type II error of a test :

Power of a test :

i.e., 1-Type II error

1. A test has level if

A test has asymptotic level if

1. Type I, Type II errors are function of , while level and power are a number

## Lecture 7 Hypothesis Testing and p-value

1. Definition: The (asymptotic) p-value of a test is the smallest (asymptotic) level at which rejects It is random, depends on sample.

## Recitation 4

1. Assume . We want to test:

Solution:

* 2. For ,

Let it be so

* 1. For , Let it be so

For

# Unit 3 Method of Estimation

## Lecture 8 Distance Measures between Distributions

1. If we want to estimate of a distribution , then the basic idea is: find such that and are close to each other. Thus we need to define some measurement of “distance” of two distributions.
2. Let be a statistic model. The total variation distance between two probability measures and is defined by
3. If is discrete, then

If is continuous, then

This is a property that can be derived from definition of

1. Properties of total variation
   1. If , then

is a distance between probability measures.

1. Exercise: If , then what is ?

Solution: As is discrete, and is continuous, if we let be the support of , then . Therefore, although

1. KL divergence (relative entropy)
2. Properties of KL divergence
   1. (can be proved with Jensen’s inequality)
   2. If then
   3. triangle inequality

KL is not a distance. It is divergence.

1. is easier to estimate than since it is an expectation.

By , , thus

1. From KL divergence to MLE

## Lecture 9 Introduction to MLE

1. For multivariate function ,
2. For
3. Multivariate Gaussian: if

## Lecture 10 Consistency of MLE

1. Under mild regularity conditions, we have

Reason: According to LLN, . Under some mild regularity conditions, we can convert the convergence in y-axis to convergence in x-axis.

1. Multivariate CLT: Let be with . Then
2. Multivariate Delta Method: Let be a sequence of random variables in

Let be continuously differentiable at , then

## Recitation 6

Assume . Find an estimator of and find its distribution.

* 1. By multivariate CLT,
  2. By multivariate Delta method, let so

## 

## Lecture 11 Fisher Information, Asymptotic Normality of MLE; Method of Moments

1. Fisher Information: Assume the log-likelihood for one observation as and is twice differentiable. Under some regularity conditions, the Fisher information of the statistical model is:

It is equivalent to

is a matrix.

If , then

1. Proof of for 1-d case:
2. Intuition: Fisher information measures the curvature of the likelihood function (also the curvature of KL divergence as function of ). As the strategy of MLE is to find : if is very curved, then small move of will lead to very large move of (or KL divergence), and thus the variance of would be small; in contrary, if is flat, then can be far from while is still close to the theoretical maximum value, and thus the variance of is large.
3. Theorem: **asymptotic normality of MLE**

Let Assume:

* 1. The parameter is identifiable
  2. For all , the support of does not depend on ( does not apply)
  3. is not on boundary of (for )
  4. is invertible in a neighborhood of
  5. a few more technical conditions

Then satisfies:

1. Informal proof

Denote

Let From KL divergence is non negative, is maximized at , so

Let , then is maximized at from definition of MLE,

According to LLN, for all .

From mean value theorem, for some between and , so

Thus, from central limit theorem,

As between and , , so and .

Therefore,

1. Motivation of Methods of Moments: if is close to , then should be close to for all continuous bounded function . As continuous function on an interval can be arbitrarily well approximated by polynomials, we only need to check .
2. Performance of Methods of Moments: let and assume is invertible, then . Let .
   1. MoM estimator is weakly/strongly consistent
   2. Let M be the matrix of moments of X, if w.r.t. by CLT, then by Delta method, , where

In 1d case, is sample mean, so so)

1. MLE vs Methods of Moments
   1. If model is mis-specified:
      * MLE will still try to find a model closest to truth
      * MoM could be completely off
   2. In general, MLE is more accurate than MoM