Contents

[Unit 1 Introduction to Statistics 2](#_Toc82356054)

[Lecture 2 Probability Redux 2](#_Toc82356055)

[Unit 2 Parametric Inference 3](#_Toc82356056)

[Lecture 3 Parametric Statistical Models 3](#_Toc82356057)

[Lecture 4 Parametric Estimation and Confidence Intervals 3](#_Toc82356058)

[Lecture 5 Delta Method and Confidence Intervals 4](#_Toc82356059)

[Recitation 2 5](#_Toc82356060)

[Lecture 6 Hypothesis Testing, Type I, II Errors 5](#_Toc82356061)

[Lecture 7 Hypothesis Testing and p-value 6](#_Toc82356062)

[Recitation 4 6](#_Toc82356063)

[Unit 3 Method of Estimation 6](#_Toc82356064)

[Lecture 8 Distance Measures between Distributions 6](#_Toc82356065)

[Lecture 9 Introduction to MLE 8](#_Toc82356066)

[Lecture 10 Consistency of MLE 8](#_Toc82356067)

[Recitation 6 8](#_Toc82356068)

[Lecture 11 Fisher Information, Asymptotic Normality of MLE; Method of Moments 9](#_Toc82356069)

[Unit 4 Hypothesis Testing 11](#_Toc82356070)

[Lecture 12 Parametric Hypothesis Testing 11](#_Toc82356071)

[Lecture13 Testing Goodness of Fit 12](#_Toc82356072)

[Unit 5 Regression 14](#_Toc82356073)

[Unit 6 Bayesian Statistics 16](#_Toc82356074)

[Unit 7 Principal Component Analysis 16](#_Toc82356075)

# 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. If the support of P and Q are different, we only need to sum up/integrate over the support of P, because at any point with but ,
2. 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. Fisher Information of normal distribution , so
3. 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
2. Shrinkage estimator of normal distribution by minimizing MSE

Notice that .

First consider the estimator for Start from , then consider the estimator , for . We have thus , which is minimized when .

Similarly, we consider the estimator of for a . The unbiased estimator is . Then let the estimator be . Thus , which is minimized when

1. M-estimators: Assume we want to estimate some parameter (e.g., mean, variance, median, quantiles) of a distribution . The strategy is to find a function such that is minimized at .
   1. If and then
   2. If and , then
   3. If and , then is the median of
   4. If is fixed and , the is a -quantile of . Here check function
   5. If , and , then . MLE is a special case of M-estimator
2. No statistical model needs to be assumed for M-estimator.
3. M-estimator is asymptotically normal under some technical conditions.

# Unit 4 Hypothesis Testing

## Lecture 12 Parametric Hypothesis Testing

1. T test

Let where both and are unknown. We want to test: Then the test statistic

followsdistribution, where

Proof: Cochran’s theorem. , where and are independent.

1. W**ald‘s test**

Consider the hypothesis: . If is true and MLE technical conditions are met, then

Hence,

Thus Wald’s test with asymptotic level :

where is the quantile of

1. Likelihood ratio test

Consider the hypothesis: for some fixed numbers .

Let be the log likelihood function,

Let be the MLE, be the constrained MLE. Then . If is true, then should be close to .

Test statistic . Here Assume is true and MLE technical conditions are met, then

Like ratio test with asymptotic level :

where is the quantile of

1. Implicit hypothesis

Instead of testing , we want to test function of , e.g.,

Idea is to start from distribution of , and apply (multivariate) delta method, and we would know the distribution of , and design tests accordingly.

## Lecture13 Testing Goodness of Fit

1. CDF for :

Empirical CDF for sample :

1. Convergence of empirical CDF to CDF
   1. LLN:
   2. Glivenko-Cantelli Theorem (Fundamental theorem of statistics):

Stronger than law of large numbers: LLN is pointwise, convergence, and GC theorem is uniform convergence.

* 1. CLT:
  2. Donsker’s Theorem: if is continuous, then

Where is a Brownian bridge on

1. Kolmogorov-Smirnov test

Let be real random variables with unknown CDF and let be a continuous CDF. Consider the hypothesis:

Let be the empirical CDF of the . By Donsker’s Theorem, we can design the test statistic:

Let be the reordered sample, then

And the test statistic .

The KS test with asymptotic level

is the quantile of which can be look up in table.

1. Important notes about KS test:
   1. Only applies to continuous distributions
   2. The distribution to be tested must be fully specified. For example, KS can be used to test “whether the data follows ”, but cannot be used to test “whether the data follows normal distribution”, which is a common mistake.

If we just plug in sample mean and sample variance, and apply KS to test whether data is normal, then it will be harder to reject, and conclude the data is normal more often than it should be. (Some people do this on purpose to conclude the data is normal)

* 1. Although the test statistic converges to , the non-asymptotical critical values can be founded in the KS table, which are calculated via simulations.

1. Other goodness of fit tests
   1. Kolmogorov-Smirnov
   2. Cramer-Von Mises
   3. Anderson-Darling
2. What if we want to test if my data is normal? Kolmogorov-Lilliefors test

The test statistic is the same as plug in sample mean and sample variance into KS test, but the distribution of the test statistic is different. Need to look up table to find critical values.

1. Testing for Multinomial distribution

The test statistic

Intuition: Under , we can calculate the Fisher information of MLE: , so It follows that Therefore, Fisher information . In addition, as , the fisher information matrix is not invertible and lose 1 degree of freedom. Therefore,

1. Test for a family of (discrete) distribution

Let be random variables on some finite sample space with some probability measure . Let be a family of probability distributions on .

Consider the hypothesis:

Let be the MLE of under . Then we can calculate the “theoretical” probability under .

Let be the “observed probability“.

Test statistic:

is the size of parameter

1. Note the degree of freedom is different in “Test multinomial” and “Test family of distribution”. When test multinomial, degree of freedom = . When test family of multinomial distribution, degree of freedom = . The reason is in testing multinomial, the true probability is given in null; but in testing family of multinomial, the true parameter/probability is unknown, and is estimated from MLE.

# Unit 5 Regression

1. Gauss Markov Assumptions
   1. Linearity: is well specified, where
   2. is deterministic and
   3. Zero condition mean + Homoscedasticity + no correlation:
2. Under Gauss Markov assumptions + normality of , LSE and MLE will yield the same estimator.

Proof:

* 1. Quadratic risk of :

Proof: As , the quadratic risk is actually

* 1. Prediction error:

Proof: . Let be the projection matrix onto column space of X, and is also a projection matrix onto the left null space of X. Therefore, Thus

As trace equals sum of eigenvalues, we have .

As is a projection matrix, its eigenvalues are 0 and 1, where the number of 1s equals the rank of , which is . Therefore and thus

1. Hypothesis tests for linear regression
   1. Test single explanatory variable is significant

Let be the -th diagonal coefficient of , then we can use test statistic

* 1. Test a group of explanatory variables, with Bonferroni’ test

We cannot test each single variable at level , since if the level (type I error) of , then the level of the joint test would be , where

Bonferroni’s test: Test each single variable at level and if any single variable is rejected at level , we reject at level of (at most) . Here

* 1. Let be a matrix with and .

Under , , and from Lemma below,

If is unknown, then is of the form , which is and is .

Lemma: If for invertible , then .

Proof: If , then , so .

1. Bayesian set up for linear regression. See Hw
2. Heteroskacidascity. Instead of minimizing , we minimize to estimate , which is the same as MLE under the Gaussian noise assumption.

# Unit 6 Bayesian Statistics

1. Conjugate prior: prior distribution and posterior distribution are the same probability distribution family.
   1. If , prior observe data then posterior )
2. An **improper prior** on is a measurable, nonnegative function defined on that is not integrable.

For example, constant prior . If is bounded, then it is uniform prior on ; if is unbounded, then it is an improper prior.

1. Jeffreys prior: where is the Fisher information matrix (provided it exists).

The key property of Jeffreys prior is that it is invariant under a change of coordinates for the parameter vector : If for some one-to-one map , then

In comparison, the uniform prior is not invariant.

1. Bayesian confidence interval/region: For a Bayesian confidence region with level is a random subset of the parameter space , which depends on the prior and the sample such that
2. Bayes estimator: posterior mean, posterior median, etc.

# Unit 7 Principal Component Analysis

1. Let be a d-dimensional random variable. Then the covariance matrix for is
2. Let be matrix, with each row be an independent copy of . The sample covariance

Let . Then is a projection matrix. Take then is actually removing from each element of .Then , which is the same as the definition of covariance.

1. If , then

* is the variance of
* is the sample variance of

1. Eigenvalue decomposition of , where is the orthogonal eigenvector matrix, and D is the diagonal eigenvalue matrix.

Then is the empirical covariance matrix of . In particular, is the empirical variance of , is the empirical variance of , etc.

1. The k orthogonal directions in which the data is the most spread out correspond to exactly the eigenvectors associated with the k largest eigenvalues of S.
2. Algorithm of PCA:
   1. Step 1: calculate the empirical covariance matrix of data
   2. Step 2: Compute the eigenvalue decomposition , where with, and is an orthogonal matrix.
   3. Step 3: Choose and set
   4. Output:

# Unit 8 Generalized Linear Models

1. A Gaussian linear model assumes , and . Essentially is a Gaussian random variable conditional on , and only affects mean of .
2. Generalized linear models relax the model from two components:
   1. may not be Gaussian. It could be any distribution from exponential family, like uniform, Bernoulli, exponential, etc.
   2. The expected value may not be a linear function of . Instead, we have , where called link function, and .
3. Examples
   1. Disease occurring rate: .

Link: Using a log link, .

Model: Since this is a count, we can model as Poisson distribution with expected value .

* 1. Prey capture rate: .

Link: Using a reciprocal link, .

Model: biology experience suggests gamma distribution.

* 1. Binomial response. We need invertible function such that . Any CDF can be an eligible

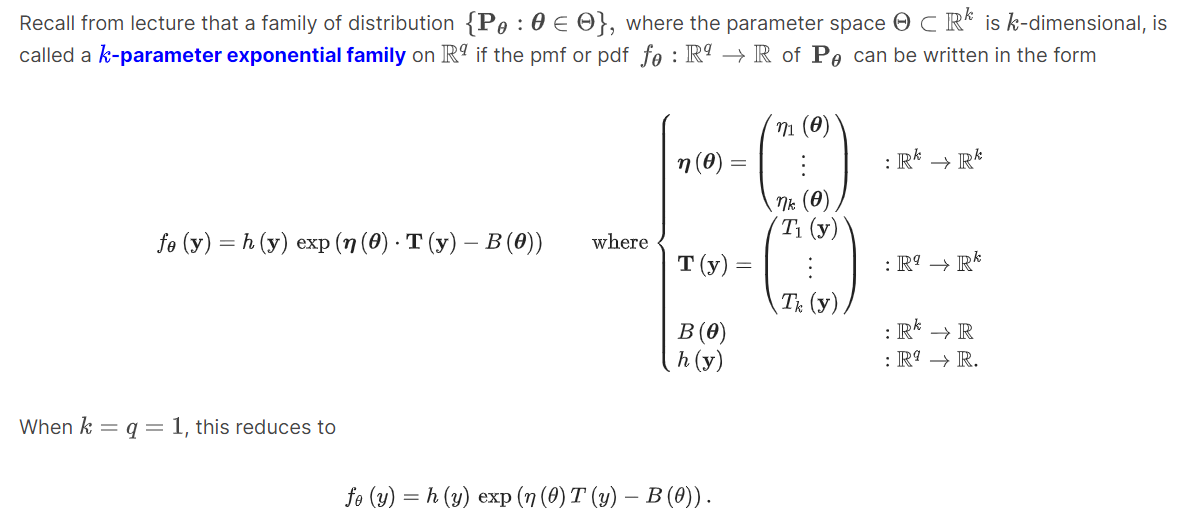
1. Exponential family: A family of distribution is said to be a k-parameter exponential family on , if there exist real valued functions:

* and of
* and of

such that the density function (pmf or pdf) of can be written as

Note: usually is a normalization term to make the density function integrate to 1.

1. Exponential family in vector form



1. Examples of exponential family
   1. Normal distribution with unknown mean and variance. .

Let , we can see it is in exponential family.

* 1. Normal distribution with unknown mean and known variance. .

Let , we can see it’s in exponential family.

* 1. Bernoulli, .

Let

Or:

Let

* 1. Other examples of continuous exponential family distributions: Gamma, Inverse Gamma, Inverse Gaussian, Chi-square, Beta, Binomial, Negative Binomial
  2. Other examples of discrete exponential family distributions: Poisson, Geometric
  3. Examples of what is not in exponential family: uniform distribution, t distribution

1. One parameter canonical exponential family

For , or some known functions and .

If is known, this is a one-parameter exponential family with be the canonical parameter.

If is unknown, this may/may not be a two-parameter exponential family. is called dispersion parameter.

In this class, we always assume is known.

Example: normal distribution with known , Poisson, Bernoulli.

Note: could be a transformation of the original parameter.

is called cumulant generating function

1. Referring to Lecture 11 Fisher Information, if , then we have , where is w.r.t .

For one-parameter canonical exponential family, , so

(1) , thus gives

(2) , so , so

Note: here we only use the fact that integrates to 1. If a density function can be written in the form of a canonical exponential family, then must contain the information of and .

1. Link function is the function that , which should be:
   1. Continuously differentiable
   2. Strictly increase (could also be strictly decreasing, but in this course we require it to be increasing)
2. Canonical link: the choice link function is not unique. A natural choice is the canonical link. the function that links mean to the canonical parameter is called canonical link: . Since , we have .
3. Given canonical parameter and link function , we have , where is defined as .

If is the canonical link, then and thus is the identity function.

1. Likelihood:

If we use the canonical link function, we get

We can use the likelihood function to calculate MLE of .