Lecture Notes on Statistical Inference

This lecture note is divided into five parts. We start the investigation of point estimation in the first section, the interval estimation in the second section, then the hypothesis testing in the third section. The Bayesian approach to statistical inference is provided in the fourth section, as a complementary material to the first three. We introduce also the linear regression in the last section.

Table of Contents

Review on Point Estimation

Tianyu Zhang[1]

Abstract:

In this monograph we offer a discription about the point estimators. Starting with the evaluation of the point estimations, we introduce some important properties one point estimator may be equipped with. We introduce then the methods of finding point estimators, where, there are four common ways, the MLE, the Bayesian, the method of moments, and the expectation maximization; we treat the MLE in the fourth section and we leave the Bayesian approach in a separater paper.

Table of Contents

1. Introduction

In this section we are going to offer a generalization of the discussion of point estimators, seeing the MLE and Bayesian estimators in the previous two sections, let us now generalize the methodology in both finding and evaluting point estimators, the materials are mainly from [1], some supplementary literature are drawn from [2].

Recall that the main assumption of the mathematical statistics is that the sequence given by

$$X = X_1, \cdots, X_n \text{ for } n \in \mathbb{N},$$

has a cumulative distribution function, namely $F(x, \theta)$ where θ is the unknown parameter, which can be any number (resp. vector) in Θ. The main task is to obtain some information about this parameter.

We shall also assume that for each X_i where $1 \le i \le n$ are independent and identically distributed (i.i.d.), i.e. each X_i has the same distribution as others and they are independent of one another.

Example 1.1: Life Time of Smartphones

Look at the sample of X_i for $i = 1,2,\cdots,n$, where each X_i is a lifetime of a

[1] YMSC, BIMSA, bidenbaka@gmail.com, obamalgb@cantab.net

smartphone and model X_i as an exponential random variable with mean θ. Potentiall, this θ can be any number in $\Theta = (0,\infty)$. Our task is for a specific realization of random variables X_i derive a conclusion about the parameter θ.

Our assumption means that the density of X_1 is $f_{X_1}(x_1) = \frac{1}{\theta}e^{x_1/\theta}$, the density of X_2 is $f_{X_2}\frac{1}{\theta}e^{x_2/\theta}$, so on and so forth.

The joint density of independent datapoints is simply product of the individual densities for each datapoint. In our example, we have

$$f_{X_1,X_2,\cdots,X_n}(x_1,x_2,\cdots,x_n) = \frac{1}{\theta}e^{x_1/\theta}\cdot\frac{1}{\theta}e^{x_2/\theta}\cdot\cdots\cdot\frac{1}{\theta}e^{x_n/\theta} = \frac{1}{\theta^n}e^{(\sum_{i=1}^n x_i)/\theta}. \quad ||$$

In statistics, if we think about the joint density as a function of the model parameter θ, we call it the likelihood function and denote it by

$$L(\theta, x) = \frac{1}{\theta^n}e^{(\sum_{i=1}^n x_i)/\theta}.$$

Now we want to get some information about the parameter θ from the x. For example, we could look for a function of x which would be close to θ. This is called the point estimation problem since we try to find a point (an estimator) which would be close to θ. In fact, this example naturally derives the definition of the point estimator.

Notation:

If θ is a parameter to be estimated, then $\hat{\theta}$ denotes its estimator or a value of the estimator for a given sample. More carefully it is a function of the data $\hat{\theta} := \hat{\theta}(X_1,\cdots,X_n)$.

Note that $\hat{\theta} = \hat{\theta}(X_1,\cdots,X_n)$ is random since its value changes from sample to sample.

Definition: Point Estimator

A point estimator is any function $\hat{\theta}(X_1,\cdots,X_n)$ of a sample; i.e. any statistic is a point estimator.

In this definition we applied the terminology called the statistic which is defined by the following convention.

Definition: Statistic

Let $X_1,\cdots,X_n$ be a random sample of size n from a population and let $T(x_1,\cdots,x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1,\cdots,X_n)$. Then the random variable or random vector $Y := T(X_1,\cdots,X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.

Remark:

A function of the dat sample is called a statistic hence an estimator is a statistic. ||

Most of the terminologies we have encountered so far are statistics, e.g. recall the mean μ and the variance σ^2. We now generalize these concepts to the form, that as a function of the random variable (resp. random vector), μ and σ^2 are themselves random variables.

There are many different approaches to find the estimators, one of the very many we have seen previously are the maximum likelihood function (MLE) and Bayesian estimators. In fact we can also use the method of moments, the expectation maximization (EM) to find estimators, we shall go through the first one in details later.

Now we discuss the evaluation of the estimator. First we shall need to know when to call an estimator a good one. Second we need to compare different estimators. Note that the comparability is based on the "distance" to the θ, the less the better. For an estimator $\hat{\theta}$, however, the distance $d(\theta, \hat{\theta})$ is not a metric.

The distribution of this random variable $\hat{\theta}$ depends on the true value of the parameter θ. One of the things that we can ask from the estimator is that its expected value equal to the true value of the parameter. This is called the unbiasedness. In symbols it is defined by

Definition: Bias

The bias of a point estimator $\hat{\theta}$ of a parameter θ is the difference between the expected value of $\hat{\theta}$ and θ. That is, Bias$\hat{\theta} := \mathbb{E}\hat{\theta} - \theta$.

Definition: Unbiased

An estimator whose bias is identically (in θ) equal to 0 is called unbiased and satisfies $\mathbb{E}\hat{\theta} = \theta \, \forall \theta \in \Theta$.

The second useful property is that when we increase the size of the sample, the estimator converges to the true value of the parameter in the sense of convergence inprobability. This is called consistency.

2. Evaluation of Point Estimators

In this section we are going to introduce some methods in evaluating the point estimator. We shall discuss the biasedness and variance in 2.1, the consistency in 2.2, and then in 2.3 we shall prove that the existence of unbiased estimators are not always valid. In 2.4 we are goin to introduce the asymptotc normality, which is mostly done by CLT, or sometimes Slutsky's Theorem. Then in 2.5 we introduce the risk function with only introduction, the detailed treatment could be seen in the previous chapter. In 2.6 we shall introduce the concept of sufficient statistics and use sufficient statistic sto derive the BUE.

2.1 Biasedness and Variance

The bias can depend on the true value of the parameter. A good estimator should have zero or at least small bias for values of the true parameter.

Example 2.1:

Consider our previous example about the lifetime of smartphones. What is the bias of the following two estimators: $\hat{\theta} = \overline{X}$ and $\hat{\theta} = X_1$?

In fact, $\overline{X}$ appear to be better than X_1. The reason is that the variance of $\overline{X}$ decreases as the sample size grows, while the variation of X_1 does not depend on the size of the sample. ||

This example naturally derives the definition of variance of a given estimator $\hat{\theta}$.

Definition: Variance

Var$\hat{\theta} = \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 = \mathbb{E}\hat{\theta}^2 - (\mathbb{E}\hat{\theta})^2$.

We of course want that $\mathrm{Var}\hat{\theta}$ to be small for all values of the true parameter θ. Ideally, both the bias and the variance of the estimator should be small. Sometimes we value unbiasedness more than anything else. We want to make sure that an estimator is unbiased and only after this condition is satisfied we compare the variance with the principle being smaller is better.

However, sometimes we can tolerate that an estimator is a bit biased. In fact, this is a trade-off, it depends on the practical terms, whether we value the unbiasedness more or we value the variance more. Moreover, in some cases it is very difficult or even impossible to find an unbiased estimator. In this case, it is useful to define a combined measure of the quality of an estimator.

Definition: Mean Squared Error (MSE)

The mean squared error of an estimator $\hat{\theta}$ of a parameter θ is the function of θ defined by $\mathrm{MSE}(\hat{\theta}) := \mathbb{E}(\hat{\theta}-\theta)^2 = \mathrm{Var}\hat{\theta} + (\mathrm{Bias}\hat{\theta})^2$.

This definition is good since it combines the two different perspective in measuring the performance of $\hat{\theta}$. However, we must point out its advantage before we dive deeper.

Disadvantage: MSE

It can be argued that the MSE, while being reasonable for location parameter, is not reasonable to scale parameters since MSE penalizes equally for overestimation and underestimation, which is fine in the location case; in the scale case however, 0 is a natural lower bound, so the estimation is not symmetric. ||

We now prove that the definition of MSE is well defined.

Theorem 2.1: MSE Decomposition

$\mathrm{MSE}(\hat{\theta}) := \mathbb{E}(\hat{\theta}-\theta)^2 = \mathrm{Var}\hat{\theta} + (\mathrm{Bias}\hat{\theta})^2$.

Proof:

Since the expectation is a linear operator, it preserves scalar multiplication and vector addition, hence it follows that

$$\begin{aligned}\mathbb{E}\big((\hat{\theta}-\theta)^2\big) &= \mathbb{E}\big((\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta)^2\big)\\ &= \mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta})^2 + 2\mathbb{E}\big((\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\big) + (\mathbb{E}\hat{\theta}-\theta)^2\\ &= \mathrm{Var}\hat{\theta} + (\mathrm{Bias}\hat{\theta})^2 + 2\mathbb{E}\big((\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}-\theta)\big).\end{aligned}$$

Since $\mathbb{E}\hat{\theta}-\theta$ is a scalar hence we can, by liearity of $\mathbb{E}$, plug it out

$$= \mathrm{Var}\hat{\theta} + (\mathrm{Bias}\hat{\theta})^2 + 2(\mathbb{E}\hat{\theta}-\theta)\mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta})).$$

Then by the fact that $\mathbb{E}(\hat{\theta}-\mathbb{E}\hat{\theta}) = \mathbb{E}\hat{\theta}-\mathbb{E}\hat{\theta} = 0$, result follows.

□

If one finds a biased estimator $\hat{\theta}$, one can sometimes easily corrects the bias to get an unbiased estimator. However, e.g. if we tried an estimator $\hat{\theta}$ and found that it has $\mathbb{E}\hat{\theta} = \sqrt{\theta}$, so we cannot correct the bias by simply taking the square of $\hat{\theta}$. The new estimator $\tilde{\theta} := \hat{\theta}^2$ will not be unbiased for θ. If we call the formula for the second moment of the random variable, then in this particular case we can even compute the bias

$$\mathbb{E}\hat{\theta}^2 = (\mathbb{E}\hat{\theta})^2 + \mathrm{Var}\hat{\theta} = \theta + \mathrm{Var}\hat{\theta},$$

so the bias of the estimator $\hat{\theta}^2$ equals $\mathrm{Var}\hat{\theta}$. In general, it is often quite difficult to find an unbiased estimator.

We will offer some other alternatives to the MSE later. Now we shall introduce another perspective in measuring the performance of a given estimator.

2.2 Consistency

The consistency is actually defined by the convergence, a topological property. Before introducing this concept we need some important results. Let $(X, \mathcal{S}, \mu)$ be a measure space where X is an arbitrary set and $\mathcal{S}$ is the σ-algebra generated by X, μ is the corresponding measure. Then in saying that, given a sequence of measurable functions (random variables) $\{f_n\}$ of almost everywhere finite valued (i.e. there are only finitely many points make f_n infinite), f_n convergest to a measurable function f, if $\forall \varepsilon > 0$, $\lim\limits_{n} \mu\left(\left\{x \,\middle|\, |f_n(x) - f(x)| \geq \varepsilon\right\}\right) = 0$. Since probability is a special case of measure, we have the definition of convergence in probability.

Definition: Converge in Probability

> A sequence of random variables $\{X_n\}$ is said to be convergent in probability to a random varaible X if $\forall \varepsilon > 0$ one has $\lim\limits_{n} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$.

Note that the sequence $\{f_n\}$ needs not to be countable, but we are dealing in most cases a countable sequence of random variables, so it does not lose any generality by just denoting $\lim\limits_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$. In fact, we shall later on use $X_n \xrightarrow{\text{Prob}} X$ to denote that X_n converges to X in probability as n being sufficiently large.

Moreover, when we speak about an estimator $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$, in fact **the distribution of the estimator depends on** n, so it would be more correctly speak about a sequence of random variables $\hat{\theta}_n$.

Usually, we expect that when the size of the sample becomse larger, i.e. as $n \to \infty$, the distribution of the estimator $\hat{\theta}_n$ become concentrated more and more around the true value of the parameter θ. **This is the minimal requirement that we can impose on the family of estimators that depend on the sample size.** If this requirement is not satisfied, then the estimator is not very useful. Technically this property of an estimator is called consistency.

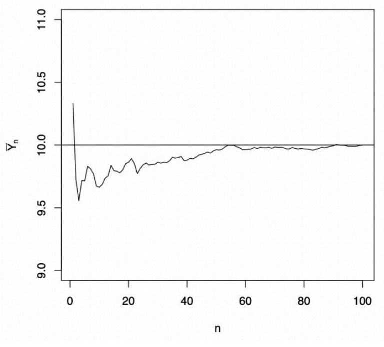

Figure 2.1

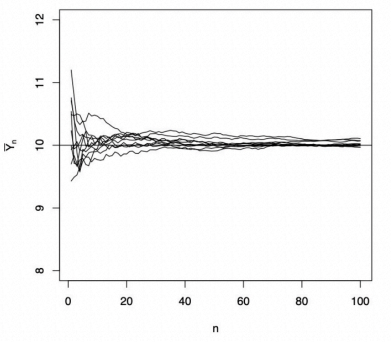

Figure 2.2

A very good intrepretation for the consistency is that it combines the WLLN and SLLN in a natural way. A sample $X_1, X_2, \cdots$ from the distribution $n(\theta, 1/4)$ was generated with $\theta = 10$ and we compute $\hat{\theta}_k = (X_1 + \cdots + X_k)/k$. **Figure 2.1** shows a path of $\hat{\theta}_k$. It suggest that if $k \to \infty$, $\hat{\theta}_k$ converges to the true value θ. In fact, this is a consequence of the Strong Law of Large Numbers (SLLN), which says that this behavior is observed with probability 1. On the other hand, **Figure 2** shows that as $k \to \infty$, $\hat{\theta}_k$ converges to the true value θ asymptotically. This is the consequence of the **Weak Law of Large Numbers** (WLLN). Since we have used these two results, we state, without proof, as facts.

Theorem 2.2: SLLN

Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$, define $\overline{X}_n := \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Then $\forall \varepsilon > 0$, one has that $\mathbb{P}(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon) = 1$, i.e. $\overline{X}_n$ converges almost surely to μ.

Theorem 2.3: WLLN

Let $X_1, X_2, \cdots$ be i.i.d. random variables with $\mathbb{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 < \infty$. Define $\overline{X}_n := \dfrac{1}{n}\sum_{i=1}^{n} X_i$. Then $\forall \varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| < \varepsilon) = 1$; i.e. $\overline{X}_n \xrightarrow{\text{Prob}} \mu$.

Remark:

We shall denote that f_n converges to f almost surely by the notation $f_n \xrightarrow{\text{a.s.}} f$.

Note that

almost surely convergence $\Rightarrow$ Convergence in Probability
almost surely convergence $\not\Leftarrow$ Convergence in Probability. ||

Now we offer the formal definition of consistency.

Definition: Consistent

An estimator $\hat{\theta}_n$ is said to be a consistent estimator of θ if $\hat{\theta}_n$ converges in probability to θ, i.e. $\hat{\theta}_n \xrightarrow{\text{Prob}} \theta$.

Consistency describes a property of the estimator in the $n \to \infty$ limit. Unlike unbiasedness, it is NOT meant to describe the property of the estimator for a fixed n, it is a tendency. Moreover, since the constisency is defined under the convergence in measure (in fact, convergence in probability measure), hence the consistency is entirely determined by the underlying topological structure. Analyzing the consistency therefore falls in to the field of functional analysis.

Note that an unbiased estimator can be inconsistent and a biased estimator can be consistent. Consistency is more important than unbiasedness since it ensures that if the data size is sufficiently large, then we will eventually learn the true value of the parameter.

We now offer a criterion in determining whether a given MSE is consistent.

Theorem 2.4: MSE Being Consistent

If $\mathrm{MSE}(\hat{\theta}_n) \to 0$ as $n \to \infty$ then the estimator $\hat{\theta}_n$ is consistent.

Inspired by the asymptotically convergent, we derive a weaker property than unbiasedness.

Definition: Asymptotically Unbiased

An estimator $\hat{\theta}_n$ is said to be asymptotically unbiased if $\mathrm{Bias}(\hat{\theta}_n) \to 0$ as $n \to \infty$.

Therefore another way to interpret **Theorem 2.4** is that: Any estimator which is asymptotically unbiased and has its variance converging to 0 as $n \to \infty$ is consistent.

However, it is sometimes cumbersome to calculate MSE of an estimator. There are some other tools to establish consistency of an estimator. We will talk about them later.

Unbiasedness vs Consistency:

- Unbiasedness:
  concerns expectation;
  for fixed n.
- Consistency:
  concerns bias and variance (and whether they vanish for large n);
  for $n \to \infty$;
  However, does not necessarily imply unbiasedness for finite n.
- Biased Estimator can be Consistent and Unbiased estimator can be Inconsistent.

We shall now introduce some common unbiased estimators. Let us now assume that $Y_1, \cdots, Y_n$ is a random sample of n i.i.d. observations from a popula-tion with mean μ and variance σ^2.

An Estimator for the Population Mean:

Estimator: $\hat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Variance: $\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$.

MSE: $\mathrm{MSE}\hat{\mu} = \mathrm{Var}\hat{\mu} + (\mathrm{Bias}\hat{\mu})^2 = \frac{\sigma^2}{n}$.

An Estimator for the Variance:

Estimator: $S^2 := \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$.

The variance of this estimator is more complicated to derive and we will not perform it here. However, it turns out that it goes to 0 as $n \to \infty$. In particular this estimator is consistent.

Since consistency is all about convergence in probability, here are some properties of this mode of convergence of random variables.

Theorem 2.5:

Suppose that $\hat{\theta}_n \xrightarrow{\text{Prob}} \theta$ and $\hat{\theta}'_n \xrightarrow{\text{Prob}} \theta'$. Then

(i) $\hat{\theta}_n + \hat{\theta}'_n \xrightarrow{\text{Prob}} \theta + \theta'$.

(ii) $\hat{\theta}_n \times \hat{\theta}'_n \xrightarrow{\text{Prob}} \theta \times \theta'$.

(iii) $\hat{\theta}_n / \hat{\theta}'_n \xrightarrow{\text{Prob}} \theta / \theta'$ provided $\theta' \neq 0$.

(iv) For any continuous function g, $g(\hat{\theta}_n) \xrightarrow{\text{Prob}} g(\theta)$.

(v) For any continuous bifunction g, $g(\hat{\theta}_n, \hat{\theta}'_n) \xrightarrow{\text{Prob}} g(\theta, \theta')$.

(vi) For $\{a_n\}_{n\in\mathbb{N}}$ a collection of numbers such that $a_n \to a$ implies $a_n \xrightarrow{\text{Prob}} a$, where a_n are viewed as special random variables.

If an estimator is not consistent, then it will not produce the correct estimation even if we are given the unlimited amount of data. Hence consistency is very important in evaluating if an estimator is "good". However, consistency does not necessarily guarantee the good performance.

2.3 The Non-Existence of Unbiased Estimators

We have seen above the several natural parameters have unbiased estimators. So it is natural to ask whether it is always possible to find an unbiased estimator for a parameter of interest, i.e. can the existence of the unbiased estimator be guaranteed? The answer is no and we offer a counterexample in this subsection.

Example 2.2: Counterexample to the Existence of Unbiased Estimator

In this example, each observation is taken from Bernoulli distribution with parameter p. That is, $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$. Of course, there is an unbiased estimator for p, namely $\hat{p} = \overline{X}$. The twist of this example is that we try to estimate $\theta = -\ln p \in \Theta = (0,\infty)$. Suppose, by seeking contradiction, that $\hat{\theta}$ is an unbiased estimator of θ and therefore, $\mathbb{E}\hat{\theta} = \theta = -\ln p$. Rewrite it by definition

$$\mathbb{E}\hat{\theta} = \sum_{x_1=0}^{1} \cdot \cdots \cdot \sum_{x_n=0}^{1} \hat{\theta}(x_1, \cdots, x_n)\mathbb{P}(X_1 = x_1, \cdots, x_n = x_n).$$

For Bernoulli random variable we can write $\mathbb{P}(X_i = x_i) = p^{x_i}(1 - o)^{1-x_i}$, where x_i can only take two values, 0 or 1. By independence of random variables $X_1, \ldots, X_n$, one has

$$\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = p^{\sum_i = 1^n x_i}(1 - p)^{n - \sum_{i=1}^n x_i}.$$

So, if $\hat{\theta}$ is unbiased, then

$$-\ln p = \sum_{x_1=0}^{1} \cdots \sum_{x_n=0}^{1} \hat{\theta}(x_1, \cdots, x_n) p^{\sum_{i=1}^n x_i}(1 - p)^{n - \sum_{i=1}^n x_i}, \qquad (3.1)$$

and this should be true for every $p \in (0,1)$ since the estimator is assumed to be unbiased for every $-\ln p \in (0,\infty)$. However, this means that the logarithmic function of p equals to a polynomial in p. This is impossible, e.g. the limit of the LHS in (3.1) for $p \to 0$ is ∞ while the RHS is finite.

We got a contradiction, so that means there is no unbised estimator of $\theta = -\ln p$. ||

2.4 Asymptotic Normality

Definition: Asymptotically Normal

An estimator $\hat{\theta}_n$ is said to be asymptotically normal if $\dfrac{(\hat{\theta}_n - \theta)}{\mathrm{Var}\hat{\theta}_n}$ converges in distribution to the standard normla distribution $N(0,1)$.

Typically, $\mathrm{Var}(\hat{\theta}_n) \sim \sigma^2/n$, and the constant σ^2 is called the asymptotic variance of the estimator. Intuitively, as n grows, the error of the estimator becomes more and more like a normal random variable with variance σ^2/n.

In order to prove the asymptotic normality of a given estimator, we usually use the **Central Limit Theorem** (CLT). Recall that

Theorem 2.6: CLT

Let $X_1, X_2, \cdots$ be a sequence of i.i.d. random variables whose mgfs exist in a neighbourhood of 0. Let $\mathbb{E}X_i = \mu$ and $\mathrm{Var}X_i = \sigma^2 > 0$ be both finite. Define $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and let $G_n(x)$ denote the cdf of $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$. Then, $\forall -\infty < x < \infty$, one has that $\lim_{n\to\infty} G_n(x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$, i.e. $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$ has a limiting standard normal distribution.

CLT is valid in much more general way than it is stated. The only assumption on the parent distribution is that it has finite variance.

An approximation tool that can be used in conjunction with the CLT is known as the Slutsky's Theorem.

Theorem 2.7: Slutsky's Theorem

If $X_n \to X$ in distribution and $Y_n \xrightarrow{\text{Prob}} a$ where a is a constant. Then

(i) $Y_n X_n \to aX$ in distribution.

(ii) $X_n + Y_n \to X + a$ in distribution.

(iii) $X_n / Y_n \to X/c$ in distribution providede $c \neq 0$.

2.5 Risk Functions and Comparison of Point Estimators

We have seen how to tell an estimator is good by describing its unbiasedness, its consistency, and its asymptotic normality. Now we need to know given more than one estimators, how do we tell which one is better.

Recall that the mean squared error of a point estimator $\hat{\theta}$ is given by

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathbb{E}(\hat{\theta} - \theta)^2.$$

We wrote it here as a function of θ to emphasize that the MSE depends on the true value of θ.

This is the special case of the risk function of an estimator. More generally,

Definition: Risk Function

The risk function of an estimator $\hat{\theta}$ is given by $R_{\hat{\theta}}(\theta) := \mathbb{E}\big(u(\hat{\theta} - \theta)^2\big)$.

The function u in the above definition is called the loss function, which is non-negative and might depend on a particular application. So intuitively the risk function is the expected loss from a mistake made while predicting the parameter θ. In the case of MSE, the loss function is simply the quadratic function $u(x) := x^2$.

The discussion of the risk function is detailed in the previous chapter, we shall ignore the detailed treatment here.

2.6 Sufficient Statistics

Recall that in studying linear algebra, it is sometimes hard to deal with rather big vector spaces, even its vector subspaces; to that end, we find it useful to work only through a small collection of elements that contain all the information of the vector space, hence we introduced the basis, as well as subbasis.

Same problems may arise when we are dealing with a big set of data. We wish, therefore, to use a small collection that contains all the information of the original data. However, not every data reduction methods could discard no information, so we wish to have one that preserve as much as possible. We shall introduce three data reduction methods in this subsection. The sufficiency principle promotes a method that preserve the information while achieving summrization of the data. The likelihood principle describes a a function of the parameter, determined by the observed sample, that contains all the information about θ that is available from the sample.

Definition: Sufficient statistic

A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

Theorem 2.8: Criterion for Sufficient Statistic

If $p(x \,|\, \theta)$ is the joint pdf or pmf of X and $q(t \,|\, \theta)$ is the pdf or pmf of $T(X)$, then $T(X)$ is a sufficient statistic for θ if $\forall x \in X$, $\dfrac{p(x \,|\, \theta)}{q(T(x) \,|\, \theta)}$ is constant as a function of θ.

Theorem 2.9: Factorization Theorem

Let $f(x \,|\, \theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for $\theta \Leftrightarrow$ there exist functions $g(t \,|\, \theta)$ and $h(x)$ such that, for all sample points x and all parameter points $\theta, f(x \,|\, \theta) = g(T(x) \,|\, \theta)h(x)$.

It is easy to find a sufficient statistic for an exponential family of distributions using the factorization theorem. Recall that the exponential family is defined by

Definition: Exponential Family

A family of pdfs or pmfs is called an exponential family if it can be expressed as

$$f(x \,|\, \theta) = h(x)c(\theta)\exp\Big\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\Big\},$$

where $h(x) \geq 0$, $t_1(x), \cdots, t_k(x)$ are real-valued functions of the observation x

(they cannot depend on θ), and $c(\theta) \geq 0$, $w_1(\theta), \cdots, w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x).

Remark:

The continuous families — normal, gamma, and beta, the discrete families — binomial, Poisson, and negative binomial, are all exponential families. ||

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still remaining all the information about θ might be considered preferable. The definition of such a statistic is the minimal sufficient statistic.

Definition: Minimal Sufficient Statistic

A sufficient statistic $T(X)$ is called a minimal sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(x)$ is a function of $T'(X)$.

That is to say, $T'(x) = T'(y) \Rightarrow T(x) = T(y)$, or, equivalently, if $\{B_{t'} \,|\, t' \in \mathscr{T}'\}$ are the partition sets of $T'(X)$ and $\{A_t \,|\, t \in \mathscr{T}\}$ are the partition sets for $T(x)$, then every $B_{t'}$ is a subset of A_t. Thus, the partition associated with a minimal sufficient statistic, is the coarsest possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Theorem 2.10: Criterion for Minimal Sufficient Statistic

Let $f(x\,|\,\theta)$ be the pmf or pdf of a sample X. Suppose that there exist a function $T(x)$ such that for every two sample points x and y, the ratio $\dfrac{f(x\,|\,\theta)}{f(y\,|\,\theta)}$ is constant as a function of $\theta \Leftrightarrow T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

However, a minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

Definition: Complete Statistic

Let $f(t\,|\,\theta)$ be a family of pdfs or pmfs for a statistic $T(X)$. The family of distributions is called complete if $\mathbb{E}_\theta g(T) = 0 \, \forall \theta$ then $\mathbb{P}_\theta(g(T) = 0) = 1$ $\forall \theta$. Equivalently, $T(X)$ is called a complete statistic.

Theorem 2.11: Complete Statistic in the Exponential Family

Let $X_1, \cdots, X_n$ be i.i.d. observations from an exponential family with pdf or pmf of the form $f(x\,|\,\theta) = h(x)c(\theta)\exp\big\{\sum_{j=1}^{k} w_j(\theta)t_j(x)\big\}$, where $\theta = (\theta_1, \cdots, \theta_k)$.

Then the statistic $T(X) := \big(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\big)$ is complete if $\big\{\big(w_1(\theta), \cdots, w_k(\theta)\big) \,\big|\, \theta \in \Theta\big\}$ contains an open set in $\mathbb{R}^k$.

The proof of this theorem depends on the uniqueness of a Laplace transform. It should be noted that the minimality of the sufficient statistic was not used in the proof of Basu's theorem. Indeed, the theorem is true with this word omitted, since a fundamental property of a complete statistic is that it is minimal. However, the condition that it contains an open set is necessarily needed.

Theorem 2.12:

If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

So even though the word "minimal" is redundant in the statement of Basu's theorem, it was stated in this way as a reminder that the statistic $T(X)$ in the theorem is a minimal sufficient statistic.

In many cases, the MSEs of two estimators will cross each other, showing that ea-ch estimator is better with respect to the other in only a small portion of the parameter space. However, even this partial information can sometimes provide guidelines for choosing between given estimators. In some worse cases however, only more information is gathered but no absolute answer is obtained.

One of the reason is that the class of all estimators is too large as a class. So instead of stucking in MSE, we have another alternative that is to reduce the size of this class. A popular way of restricting the class of estimators is to consier only unbiased estimators.

If W_1 and W_2 are both unbiased estimators of a parameter θ, i.e. $\mathbb{E}_\theta W_1 = \mathbb{E}_\theta W_2 = \theta$ then their MSE are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance — a best unbiased estimator — then we are done.

Suppose that there is an estimator W^* of θ with $\mathbb{E}_\theta W^* = \tau(\theta) \neq \theta$ and we are interested in investigating the worth of W^*. Consider the class of estimators given by

$$C_\tau := \{W \mid \mathbb{E}_\theta W = \tau(\theta)\}.$$

For all the choice of $W_1, W_2 \in C_\tau$, $\text{Bias}_\theta(W_1) = \text{Bias}_\theta(W_2)$ so one has

$$\mathbb{E}_\theta(W_1 - \theta)^2 - \mathbb{E}_\theta(W_2 - \theta)^2 = \text{Var}_\theta(W_1) - \text{Var}_\theta(W_2)$$

and MSE comparisons, within the class C_τ, can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators with the same expected value $\tau(\theta)$.

Definition: Best Unbiased Estimator (BUE)

An estimator W^* is a best unbiased estimator of $\tau(\theta)$ if it satisfies $\mathbb{E}_\theta W^* = \tau(\theta) \forall \theta$, and for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$.

Definition: Uniform Minimum Variance Unbiased Estimators (UMVUE)

A BUE W^* is said to be a uniform minimum variance unbiased estimator if for any other estimator W with $\mathbb{E}_\theta W = \tau(\theta)$, one always has $\text{Var}_\theta W^* \leq \text{Var}_\theta W \ \forall \theta$.

Suppose that, for estimating a parameter $\tau(\theta)$ of a distribution $f(x \mid \theta)$, we can specify the lower bound, say $B(\theta)$, on the variance of any unbiased estimator of $\tau(\theta)$. If we can find an unbiased estimator W^* such that $\text{Var}_\theta W^* = B(\theta)$, then we have found the BUE. This is the approach taken with the use of the Cramér-Rao lower bound.

Theorem 2.13: Cramér-Rao Inequality

Let $X_1, \cdots, X_n$ be a sample with pdf $f(x \mid \theta)$, and let $W(X) = W(X_1, \cdots, X_n)$ be any estimator satisfying

(i) $\dfrac{d}{d\theta}\mathbb{E}_\theta W(X) = \displaystyle\int_{\Omega_X} \frac{\partial}{\partial\theta} W(x) f(x \mid \theta) dx$

(ii) $\mathrm{Var}_\theta W(X) < \infty$.

Then $\mathrm{Var}_\theta W(X) \geq \dfrac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(X)\right)^2}{\mathbb{E}_\theta\Big(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\Big)}$.

If we add the assumption of independent samples, the calculatin of the lower bound could be simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary implies.

Corollary 2.13.1: Cramér-Rao Inequality, i.i.d. case

Let $X_1, \cdots, X_n$ be an i.i.d. sample with pdf $f(x \mid \theta)$ and let $W(X) := W(X_1, \cdots, X_n)$ be any estimator such that

(i) $\dfrac{d}{d\theta}\mathbb{E}_\theta W(X) = \displaystyle\int_{\Omega_X} \frac{\partial}{\partial\theta} W(x) f(x \mid \theta) dx$

(ii) $\mathrm{Var}_\theta W(X) < \infty$.

Then $\mathrm{Var}_\theta W(X) \geq \dfrac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(X)\right)^2}{n\mathbb{E}_\theta\Big(\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\Big)}$.

Note that the Cramér-Rao lower bound does not only work for the continuous random variables but also the discrete ones. The quantity $\mathbb{E}_\theta\Big(\left(\dfrac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\Big)$ is called the information number, or Fisher information of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the BUE of θ. As the information number increases, the bound on the variance of BUE gets smaller.

For any differentiable function $\tau(\theta)$, we now have a lower bound on the variance of any estimator W such that $\mathbb{E}_\theta W = \tau(\theta)$. The bound depends only on $\tau(\theta)$ and $f(x \mid \theta)$ and is a uniform lower bound for the variance. Any candidate estimator satisfying $\mathbb{E}_\theta W = \tau(\theta)$ and attaining this lower bound is a BUE of $\tau(\theta)$.

Remark:

Even if the Cramér-Rao is applicable, there is no guarantee that the bound is sharp. That is to say, the value of the Cramér-Rao lower bound may be strictly smaller than the variance of any unbiased estimator.

In fact, the most we can say by applying Cramér-Rao is that there exists a parameter $\tau(\theta)$ with an unbiased estimator that achieves the Cramér-Rao lower bound; however, in other typical situations, for other parameters, the bound may not be attainable. Hence we need results dealing with its attainment.

Corollary 2.14: Attainment of Cramér-Rao Lower Bound

Let $X_1, \cdots, X_n$ be i.i.d. $f(x \mid \theta)$ where $f(x \mid \theta)$ satisfies the conditions of Cramér-Rao Theorem. Let $L(\theta \mid x) := \displaystyle\prod_{i=1}^{n} f(x_i \mid \theta)$ denote the likelihood function. If

$W(X) = W(X_1, \cdots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(X)$ attains the Cramér-Rao lower bound if and only if

$$a(\theta)\big(W(x) - \tau(\theta)\big) = \frac{\partial}{\partial\theta} \log L(\theta\,|\,x)$$

for some function $a(\theta)$.

The attainment of the Cramér-Rao lower bound still leaves some questions. Firstly, what if the $f(x\,|\,\theta)$ does not satisfy the assumptions of the **Cramér-Rao Theorem**? Secondly, what if the bound is still unattainable for legal estimators?

One way of answering these questions is to search for methods that are more widely applicable and yield sharper (i.e. greater) lower bounds. Much research has been done on this topic, with perhaps the most famous one is Chapman and Robbins (1951). We leave this to interested readers and we now introduce the study of BUE from another view, using the concept of sufficiency.

In the previous discussion, the concept of sufficiency was not used in our search for unbiased estimates. We will now see the consideration of sufficiency is a powerful tool indeed. The main result of this method relates the sufficient statistic to unbiased estimate. Recall that $\mathbb{E}X = \mathbb{E}\big(\mathbb{E}(X\,|\,Y)\big)$ and $\mathrm{Var}X = \mathrm{Var}\big(\mathbb{E}(X\,|\,Y)\big) + \mathbb{E}\big(\mathrm{Var}(X\,|\,Y)\big)$.

Theorem 2.15: Rao-Blackwell

Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\varphi(T) := \mathbb{E}(W\,|\,T)$. Then

(i) $\mathbb{E}_\theta\varphi(T) = \tau(\theta)$.

(ii) $\mathrm{Var}_\theta\varphi(T) \le \mathrm{Var}_\theta W \;\forall\theta$.

That is , $\varphi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$.

Therefore, conditioning any unbiased estimator on a sufficient statistic will result in a uniform improvement, so we need consider only statistics that are functions of a sufficient statistic in our search for best unbiased estimator.

In fact, conditioning on anything will result in an improvement, but the problem is that the resulting quantity will probably depend on θ and therefore not be an estimator.

We now state and prove a powerful result stating that a best unbiased estimator is unique.

Theorem 2.16:

If W is a best unbiased estimator of $\tau(\theta)$ then W is unique.

Proof:

Suppose that W' is another best unbiased estimator, and consider the estimator $W^* = \dfrac{1}{2}(W + W')$. Note that $\mathbb{E}_\theta W^* = \tau(\theta)$ and

$$\begin{aligned}\mathrm{Var}_\theta W^* &= \mathrm{Var}_\theta(\frac{1}{2}W + \frac{1}{2}W') = \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}\mathrm{Cov}_\theta(W, W')\\ &\le \frac{1}{4}\mathrm{Var}_\theta W + \frac{1}{4}\mathrm{Var}_\theta W' + \frac{1}{2}\big(\mathrm{Var}_\theta W \cdot \mathrm{Var}_\theta W'\big)^{\frac{1}{2}} \quad \text{(Cauchy-Schwartz)}\\ &= \mathrm{Var}_\theta W. \qquad (\mathrm{Var}_\theta W = \mathrm{Var}_\theta W' \text{ by assumption})\end{aligned}$$

But if the above inequality is strict, then the best unbiasedness of W is contradicted, so we must have equality for all θ. Since the inequality is an application of Cauchy-Schwartz we can have equality only if $W' = a(\theta)W + b(\theta)$. Now applying properties of covariance, we have

$$\begin{aligned}\mathrm{Cov}_\theta(W, W') &= \mathrm{Cov}_\theta\big(W, a(\theta)W + b(\theta)\big)\\ &= \mathrm{Var}_\theta(W, a(\theta)W) = a(\theta)\mathrm{Var}_\theta W,\end{aligned}$$

but $\mathrm{Cov}_\theta(W, W') = \mathrm{Var}_\theta W$ hence $a(\theta) = 1$. Since $\mathbb{E}_\theta W' = \tau(\theta)$ we must have $b(\theta) = 0$ therefore $W = W'$, uniqueness follows.

□

To see when an unbiased estimator is best unbiased, we might ask how could we improve upon a given unbiased estimator? The relationship of an unbiased estimator W with unbiased estimators of 0 (i.e. $\mathbb{E}_\theta U = 0 \forall\theta$) is crucial in evaluating whether W is best unbiased. This relationship, in fact, characterizes the best unbiasedness.

Theorem 2.17:

If $\mathbb{E}_\theta W = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta) \Leftrightarrow W$ is uncorrelated with all unbiased estimators of 0.

Remark: Random Noise

Note that an unbiased estimator of 0 is nothing more than random noise; i.e. there is no information in an estimator of 0. Therefore, if an estimator could be improved by adding random noise to it, the estimator probably is defective. ||

Although we now have an interesting characterization of BUEs, its usefulness is limited in application. It is often a difficult task to verify that an estimator is uncorrelated with all unbiased estimators of 0 since it is usually difficult to describe all unbiased estimators of 0.

It is worthwhile to note once again that what is important is the completeness of the family of distributions of the sufficient statistic. Completeness of the original family is of no consequence. This follows from the Rao-Blackwell Theorem, which says that we can restrict attention to functions of a sufficient statistic, so all expectations will be taken with respect to its distribution.

We sum up the relationship between completeness and best unbiasedness in the following theorem.

Theorem 2.18:

Let T be a complete sufficient statistic for a parameter θ and let $\varphi(T)$ be any estimator based only on T. Then $\varphi(T)$ is the unique BUE of its expected value.

In many situations, there will be no obvious candidate for an unbiased estimator of a function $\tau(\theta)$, much less a candidate for BUE. However, in the presence of completeness, **Theorem 2.18** tells us that if we can find any unbiased estimator, then we can find the best unbiased estimator.

Theorem 2.19: Lehmann-Scheffé

Unbiased estimators based on complete sufficient statistics are unique.

3. Methods in Finding Point Estimators

We have offered two methods in finding point estimators in the first two chapters. In the first chapter we see how can we find the MLE and in the second chapter we see a way to draw the Bayesian estimators. In this section we are going to introduce the methods of moments and the Expectation Maximization method.

3.1 Methods of Moments

The method of moments is, perhaps, the oldest method of finding point estimators, it has the virtue of being quite simple to use and almost always yields some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when old methods prove intractable.

Algorithm 3.1: Methods of Moments

Let $X_1, \cdots, X_n$ be a sample from population with pdf or pmf $f(x \mid \theta_1, \cdots, \theta_k)$. Methods of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resulting system of simultaneously equations. More precisely, define

$$m_1 := \frac{1}{n}\sum_{i=1}^{n} X_i', \mu_1' = \mathbb{E}X,$$
$$m_2 := \frac{1}{n}\sum_{i=1}^{n} X_i^2, \mu_2' := \mathbb{E}X^2,$$
$$\cdots\cdots$$
$$m_k := \frac{1}{n}\sum_{i=1}^{n} X_i^k, \mu_k' := \mathbb{E}X^k.$$

The population moments μ_j' will typically be function of $\theta_1, \cdots, \theta_k$, namely $\mu_j'(\theta_1, \cdots, \theta_k)$. The method of moments estimators $(\tilde{\theta}_1, \cdots, \tilde{\theta}_k)$ of $(\theta_1, \cdots, \theta_k)$ is obtained by solving the following system of equations for $(\theta_1, \cdots, \theta_k)$ in terms of $(m_1, \cdots, m_k)$.

$$m_1 = \mu_1'(\theta_1, \cdots, \theta_k),$$
$$m_2 = \mu_2'(\theta_1, \cdots, \theta_k),$$
$$\cdots\cdots$$
$$m_k = \mu_k'(\theta_1, \cdots, \theta_k).$$

The method of moments can be very useful in obtaining approximations to the distribution of statistics. This technique, is sometimes called the moment matching, gives us an approximation that is based on matching moments of distributions. In theory, the moments of distribution of any statistics could be matched, however, in practical terms, it is best to have distributions that are similar.

We now illustrate some examples in using moments to find point estimators.

Example 3.1: Normal Method of Moments

Suppose that $X_1, \ldots, X_n$ are iid $N(\theta, \sigma^2)$. In the preceding notation, $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. We have $m_1 = \frac{1}{n}\sum X_i = \overline{X}$ and $m_2 = \frac{1}{n}\sum X_i^2$, and $\mu_1' = \theta, \mu_2' = \theta^2 + \sigma^2$.

Solving for θ and σ^2 yields the Method of Moments estimator:

$\tilde{\theta} = m_1 = \overline{X}$ and $\tilde{\sigma}^2 = m_2 - m_1^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2 = \frac{1}{n}\sum (X_i - \overline{X})^2.$ ||

In this simple example, the Method of Moments solution coincides with our intuition and perhaps gives some credence to both. The method is somewhat more helpful, however, when no obvious estimator suggests itself.

Example 3.2: Binomial Method of Moments

Let $X_1, \ldots, X_n$ be i.i.d. Binomial(k, p), i.e.,

$$\mathbb{P}(X_i = x \,|\, k, p) = \binom{k}{x} p^x (1-p)^{k-x}, x = 0,1,...,k.$$

Here we assume that both k and p are unknown and we desire point estimators for both parameters.

Equating the first two sample moments to those of the population yields the system of equations:

$\overline{X} = kp, \dfrac{1}{n}\sum X_i^2 = kp(1-p) + k^2p^2$, which must now be solved for k and p.

After a little algebra, we obtain the Method of Moments estimators:

$$\overline{X}^2 = k^2p^2 \Longrightarrow \frac{1}{n}\sum X_i^2 - \overline{X}^2 = kp(1-p) = kp - kp^2 = \overline{X} - \frac{\overline{X}^2}{k}.$$

Therefore, $\tilde{k} = \dfrac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum (X_i - \overline{X})^2}$, and $\tilde{p} = \dfrac{\overline{X}}{\tilde{k}}$.

Admittedly, these are not the best estimators for the population parameters. In particular, it is possible to get negative estimates of k and p, which, of course, must be positive numbers(This is the case where the range of the estimators does not coincide with the range of the parameter it is estimating.) However, in fairness to the Method of Moments, note that negative estimates will occur only when the sample mean is smaller than the sample variance, indicating a large degree of variability in the data.

The Method of Moments has, in this case, at least given us a set of candidates for point estimators of k and p.

Although our intuition may have given us a candidate for an estimator of p, coming up with an estimator of k is much more difficult. ||

The method of moments can be very useful in obtaining approximations to the distributions of statistics. This technique, is sometimes called “moment matching”, gives us an approximation that is based on matching moments of distributions. In theory, the moments of the distribution of any statistic can be matched to those of any distribution but, in practice, it is best to use distributions that are similar.

3.2 Expectation Maximization

The last method that we will look at for finding estimators is inherently different in its approach and specifically designed to find MLEs. Rather than detailing a procedure for solving for the MLE, we specify an algorithm that is guaranteed to converge to the MLE. This algorithm is called Expectation Maximization (EM) algorithm. It is based on the idea of replacing one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem. It is particularly suited to “missing data” problems, as the very fact that there are missing data can sometimes make calculations cumbersome. However, filling in the “missing data” will often make the calculation go more smoothly.

In using the EM algorithm we consider two different likelihood problems. The first problem that we are interested in solving is the “incomplete data” problem and the problem that we actually solve is the “complete-data”problem. Depending on the situation, we can start with either problem.

Expectation Maximization(EM) Algorithm is that conditions for convergence to the incomplete-data MLEs are known, although this topic has obtained an additional bit of folklore. We shall not dive in to deep to this algorithm and we offer some readings for those has interests, [4], [5], and [6], also some lecture notes [7], and a well-organized one [8].

4. Maximum Likelihood Estimation

Serving as a method in searching for estimators, the method of maximum likelihood is, perhaps, by far the most popular technique for deriving estimators. Recall that if $X_1, \cdots, X_n$ are an i.i.d. sample from a pupulation with pdf or pmf $f(x \mid \theta_1, \cdots, \theta_k)$, the likelihood function is defined by

$$L(\theta \mid x) := L(\theta_1, \cdots, \theta_k \mid x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta_1, \cdots, \theta_k).$$

Definition: Maximum Likelihood Estimator (MLE)

For each sample point x, let $\hat{\theta}(x)$ be a paramater value at which $L(\theta \mid x)$ attains its maximum as a function of θ, with x fixed. A maximum likelihood estimator (MLE) of the parameter θ based on a sample X is $\hat{\theta}(X)$. In short, it is the value of θ that maximizes the likelihood function.

Notice that, by this construction, the range of the MLE coincides with the range of the parameter. We also use the abbreviation MLE to stand for Maximum Likelihood Estimate when we are talking about the realized value of the estimator. Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely. In general, the MLE is a good point estimator, processing some of the optimality properties.

One good interpretation for the fact that maximizing over the likelihood function gives as a more accurate estimate is coming from [10]. Recall that the likelihood function is defined as

Definition: Likelihood Function

Let $f(x \mid \theta)$ denote the joint pdf or pmf of the sample $X = (X_1, \cdots, X_n)$. Then, given that $X = x$ is observed, the function of θ defined by $L(\theta \mid x) = f(x \mid \theta)$ is called the likelihood function.

Clearly $L(\theta \mid x)$ depends on the data $X = (X_1, \cdots, X_n)$, but they’re treated as functions of θ only. The likelihood function is not the pdf or pmf of θ so it does not make any sense to integrate over θ values. What we are really interested in is the shape of the likelihood curve or, equivalently, the relative comparisons of the $L(\theta \mid x)$ for different θ’s. That is to say:

Remark: Interpratation for Likelihood Functions

If $L(\theta_1 \mid x) > L(\theta_2 \mid x)$ (resp. $\log L(\theta_1 \mid x) > \log L(\theta_2 \mid x)$), then θ_1 is more likely to have been responsible for producing the observed $X = (X_1, \cdots, X_n)$. ||

In statistical theory, one of the most important concern to the MLE, as it stated in this way, is the optimization. The most natural treatment on this topic is by differentiation; we know that if the first derivative of a function f vanishes at a point x_0, then the function has much more probability in gaining a local extrema at x_0. Unfortunately, what really interest us is the global extrema, this sometimes could be troublesome since in guaranteeing the global behavior we need to check all the possible points, this is a tedius work. To that end, the convex optimization offers some insights in determining the global extema, for example, in [9], if we konw that a point x is where f has finite local minimum and $0 \in \partial f(x)$, then f has its global minimum at x.

Properties of MLE:

(i) Translation Invariant
(ii) Consistent
(iii) Asymptotic Normal

We have proved the translation invariant property, so we just restate it here. The consistency and asymptotic normality are from both [10] and [11].

Theorem 4.1: Translation Invariant

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Proof:

Let $\hat{\eta}$ be the value that maximizes the induced likelihood function $L^*(\eta \,|\, x)$.

WTS I: $L^*(\hat{\eta} \,|\, x) = L^*(\tau(\hat{\theta}) \,|\, x)$.

By definition, the maximum of L and L^* coincide, therefore, it follows that

$$L^*(\hat{\eta} \,|\, x) = \sup_{\eta} \sup_{\{\theta | \tau(\theta) = \eta\}} L(\theta \,|\, x) = \sup_{\theta} L(\theta \,|\, x) = L(\hat{\theta} \,|\, x),$$

where the last equality is by the definition of $\hat{\theta}$. On the other hand, we have

$$\begin{aligned} L(\hat{\theta} \,|\, x) &= \sup_{\{\theta | \tau(\theta) = \tau(\hat{\theta})\}} L(\theta \,|\, x) && (\hat{\theta} \text{ is the MLE}) \\ &= L^*(\tau(\hat{\theta}) \,|\, x). && (\text{Definition of } L^*) \end{aligned}$$

Hence, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ and the invariance follows.

□

Remark:

The invariance property for MLE is still valid for the multivariate case. ||

In stating the consistency, we need first to clarify what do we mean in saying that an MLE is consistent. Before that recall two important results in probability theory, the **Law of Large Numbers** (LLN) and the **Central Limit Theorem** (CLT), we assume the readers are already familiar with these two facts so we state them without proof.

Fact 4.2: LLN

If the distribution of i.i.d. sample $X_1, \cdots, X_n$ is such that $|\mathbb{E}X_1| < \infty$, then

$$\overline{X}_n := \frac{X_1 + \cdots + X_n}{n} \xrightarrow{\text{Prob}} \mathbb{E}X_1, \text{ i.e. } \mathbb{P}(|\overline{X}_n - \mathbb{E}X_1| > \varepsilon) > 0 \text{ as } n \to \infty.$$

Fact 4.3: CLT

If the distribution of i.i.d. sample $X_1, \cdots, X_n$ is such that $|\mathbb{E}X_1| < \infty$ and

$\sigma^2 = \text{Var}X < \infty$ then $\sqrt{n}(\overline{X}_n - \mathbb{E}X_1) \xrightarrow{\text{distribution}} N(0,\sigma^2)$, that is to say $\mathbb{P}\left(\sqrt{n}(\overline{X}_n - \mathbb{E}X_1) \in [a,b]\right) \to \int_a^a \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-x^2}{2\sigma^2}} dx \ \forall a,b \in \mathbb{R}$. That is to say, $\sqrt{n}(\overline{X}_n - \mathbb{E}X_1)$ behaves as normal random variable as n being sufficiently large.

The terminology "consistency" follows directly from these two facts.

Definition: Consistent

We say that an estimate $\hat{\theta}_n$ is consistent if $\hat{\theta} \xrightarrow{\text{Prob}} \theta^*$ as $n \to \infty$ where n is the sample size and θ^* is the "true" but unknown value. That is, $\lim\limits_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta^*| > \varepsilon) = 0 \forall \varepsilon > 0$.

Rule 1: Identifiability

If $\theta \neq \theta'$ then f_θ and $f_{\theta'}$ share different distributions.

Rule 2:

The support of f_θ, i.e. $\text{supp} f_\theta := \{x \,|\, f_\theta(x) > 0\}$ is the same $\forall \theta \in \Theta$.

Rule 3:

θ^* is an interior point of Θ.

Theorem 4.4: Consistent

Under **Rule 1**, **Rule 2**, and **Rule 3**, the MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta^*$ as $n \to \infty$.

Corollary 4.4.1:

If **Rule 1** and **Rule 2** hold, then for any $\theta \neq \theta^*$, $\lim\limits_{n\to\infty} \mathbb{P}\left(L(\theta^* \,|\, x) > L(\theta \,|\, x)\right) = 1$.

The consequence of **Corollary 4.4.1** is that, the likelihood function at the true θ^* tends to be larger than any other likelihood value. So if we estimate θ by maximizing the likelihood, that maximizer ought to be close to θ^*.

We close our first section by introducing the asymptotic normality of the MLE. We want to show that $\sqrt{n}(\hat{\theta} - \theta_n) \xrightarrow{\text{distribution}} N(0,\sigma^2_{\text{MLE}})$ and then compute σ^2_{MLE}. This asymptotic variance in some sense measures the quality of MLE. First we introduce the notion called Fisher Information.

Definition: Fisher Information

Denote $\ell(X \,|\, \theta) := \log f(X \,|\, \theta)$ and by saying $\ell'(X \,|\, \theta)$ we mean the first derivative with respect to θ. The Fisher information of a random variable X with distribution $\mathbb{P}_{\theta_n}$ form the family $\{\mathbb{P}_\theta \,|\, \theta \in \Theta\}$ is defined by

$$I(\theta_n) := \mathbb{E}_{\theta_n}\left(\ell'(X \,|\, \theta_n)\right)^2 = \mathbb{E}_{\theta_n}\left(\frac{\partial}{\partial\theta} \log f(X \,|\, \theta)\Big|_{\theta=\theta_n}\right)^2.$$

Remark:

Let us give a very informal interpretation of the Fisher Information. The derivative $\ell'(X \,|\, \theta_n) = (\log f(X \,|\, \theta_n))' = \dfrac{f'(X \,|\, \theta_n)}{f(X \,|\, \theta_n)}$ can be interpreted as a measure of how quickly the distribution density will change when we slightly change the parameter θ near θ_n. When we square this and take expectation (so we get the form in the definition) we get an averaged version of this measure.

(i) When Fisher Information is large, this means that the distribution will change quickly when we move the parameter, i.e. a small change with respect to the parameter leads to a huge perturbation.

(ii) When the Fisher Information is small, on the other hand, the distribution is very similar for either at θ_n or the points near θ_n. ||

We now state a result without proof for which could relax our computation on the Fisher Information.

Lemma 4.5:

$$\mathbb{E}_{\theta_n}\ell''(X\,|\,\theta_n) = \mathbb{E}_{\theta_n}\frac{\partial^2}{\partial\theta^2}\log f(X\,|\,\theta_n) = -I(\theta_n).$$

We now state but without proof of the asymptotic normality of the MLE. For those readers who are interested in this topic one may consult [11].

Theorem 4.6: Asymptotic Normal

We have, for the MLE $\hat{\theta}$, that $\sqrt{n}(\hat{\theta}-\theta_n) \to N\big(0, \frac{1}{I(\theta_n)}\big)$ as $n \to \infty$.

It should be pointed out that the MLE does not always exists. For the conditions the MLE exists, we found the article [12] very useful, we state it here and one could consult [13] for detailed treatment.

Condition for the Existence of MLE:

The MLE exists if the parameter space Θ is compact and the likelihood function $L(\theta\,|\,x)$ is continuous $\forall\theta \in \Theta$.

Condition for the Uniqueness of MLE:

The MLE is unique if the parameter space Θ is convex and the likelihood function $L(\theta\,|\,x)$ is concave.

Reference:

[1]: Vladislav Kargin, *Lecture Notes for Math 448 Statistics*, available online.

[2]: George Casella, Roger L. Berger, *Statistical Inference, Second Edition*, Wadsworth Group, pp. 311-355.

[3]: Paul R. Halmos, *Measure Theory*, Springer-Verlag, pp. 84-94.

[4]: Dempster, Laird, Rubin's(1977), *Original Proof of the Convergence Had a Flaw*.

[5]: Boyles(1983), Wu(1983), *Valid Convergence Proofs*.

[6]: Also for [5], Finch, Mendell, and Thode(1989).

[7]: Tengyu Ma and Andrew Ng, CS229 Lecture notes, available online.

[8]: Stanley H. Chan, *Expectation-Maximization Algorithm*, available online.

[9]: R.T. Rockafellar, *Convex Analysis, Section 27, The Minimum of a Convex Function*.

[10]: Ryan Martin, *Stat 411 — Lecture Notes 03, Likelihood and Maximum Likelihood Estimation*, available online.

[11]: Dmitry Panchenko, *Lecture 3, Properties of MLE: consistency, asymptotic normality. Fisher Information*, available online.

[12]: When does a maximum likelihood estimate fail to exist? Answer on Math StackExchange, available online.

[13]: Fumio Hayashi, *Ecnometircs, 7.1 Estremum Estimator, Lemma 7.1*, pp. 446-453, available online.

Review on Interval Estimation

Tianyu Zhang[2]

Abstract:

Serving different approach as the point estimation, the interval estimation provides us a way to describe the error size and a confidence level for the estimation to coincide with the realized values. In this short survey we introduce the methods in finding, and in evaluating the interval estimations. Since we have treated the Bayesian Interval Estimation independently, we shall not offer the Bayesian Approach to the interval estimation, hence the optimization in the decision theory (a way to evaluation).

Table of Contents

1. Introduction

While point estimation gives a single value as the best estimate for a parameter, interval estimation provides a range of values to account for the uncertainty associated with the estimation process. Interval estimation is valuable for capturing the potential variability in the estimate and conveying the level of confidence in the result.

Definition: Interval Estimator/Estimate

> An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \cdots, x_n)$ and $U(x_1, \cdots, x_n)$, of a sample that satisfy $L(x) \leq U(x) \forall x \in \mathcal{X}$. If $X = x$ is observed, the inference $L(x) \leq \theta \leq U(x)$ is made. The random variable $[L(X), U(X)]$ is called an interval estimator.

The purpose of using an interval estimator rather than a point estimator is to have some guarantee of capturing the parameter of interest. The centainty of this guarantee is quantified by the following definitions.

Definition: Coverage Probability

> For an interval estimator $[L(X), U(X)]$ of a parameter θ, the coverage probability of $[L(X), U(X)]$ is given by $\mathbb{P}_\theta\big(\theta \in [L(X), U(X)]\big)$ or $\mathbb{P}\Big(\theta \in [L(X), U(X)]\Big|\theta\Big)$.

[2] BIMSA, bidenbaka@gmail.com, obamalgb@cantab.net

Definition: Confidence Coefficient

For an interval estimator $[L(X), U(X)]$ of a parameter θ, the confidence coefficient of $[L(X), U(X)]$ is $\inf_{\theta} \mathbb{P}_\theta\big(\theta \in [L(X), U(X)]\big)$.

We may also use the value $1-\alpha = \mathbb{P}_\theta\big(\theta \in [L(X), U(X)]\big)$ to denote the $(1-\alpha)$ 100% confidence interval $[L(X), U(X)]$, where $1-\alpha$ is called the degree of confidence while $L(X)$ and $U(X)$ are called the lower and upper confidence limits, respectively. For instance, if $\alpha = 0.05$ then the degree of confidence is 95%.

In the next section we shall discuss the estimation of means and variances. Then we proceed to discuss the method in finding the general interval estimation, where we are going to talk about (i) Inverting a test statistic, (ii) Pivotal Quantities, and (iii) Bayesian intervals. Lastly we talk about some evaluation methods, which are (1) Size and coverage probability, (2) Test-related optimization, (3) Bayesian optimization, and (4) Loss function optimization.

2. Pivot Method

In this section we are going to introduce some special estimation results. In 2.1 we will introduce the interval estimation for normal means, with either known or unknown variance; both the univariate and the bivariate (hence multivariate) cases are discussed in this section. In 2.2 we shall proceed to the talk about the interval estimation for the variance of the normal populations; both the univariate and the bivariate (hence multivariate) cases are discussed in this section. In 2.3 we will offer the treatment for the interval estimation in binomial case, still, botht the univariate and the bivariate (hence multivariate) cases are performed. Serving as specific examples for the first three subsections, in 2.4 we introduce the pivot method in finding the interval estimation, which generalizes the first three parts.

Recall the z-values we introduced before. For a non-standard normal random variable $X \sim N(\mu, \sigma^2)$, the standardizing involves the change-of-variable $Z := \dfrac{X-\mu}{\sigma}$, where the denominator is σ since otherwise it would not scale the deviation correctly, and the resulting standardized variable would not have the desired properties.

The $z_{\alpha/2}$ is often valuable in finding the interval estimators. The reason is that, suppose we are going to find an interval estimator $[L(X), U(X)]$ of a parameter θ with confidence coefficient 0.95, then what we really do is to calculate

$$\mathbb{P}_\theta(\theta < L(X)) = \mathbb{P}_\theta(\theta > U(X)) = 0.025,$$

where $\mathbb{P}_\theta\big(\theta \in [L(X), U(X)]\big) = 0.95$.

Recall also the t-distribution. The t-distribution arises in statistical inference when dealing with small sample sizes or when the population standard deviation is unknown. It is commonly used in hypothesis testing and constructing confidence intervals for the mean.

Remark:

The t-distribution is symmetric and bell-shaped, similar to the normal distribution. The shape is determined by the degrees of freedom. As the degrees of freedom increases, the t-distribution approaches the standard normal distribution. ||

Lastly we refer to the F-distribution. The F-distribution is used in statistical hypothesis testing, specifically in the context of comparing variances. It arises when comparing the variability of two independent samples.

Fact 1: F-Distribution

If S_1^2 and S_2^2 are the variances of independent random samples of sizes n_1 and n_2 from normal populations with the variances σ_1^2 and σ_2^2, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

is a random variable having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

2.1 Estimation of Normal Means

To illustrate how the possible size of errors can be appraised in point estimation, suppose that the mean of a random sample is to be used to estimate the mean of a normal population with the unknown mean μ and known variance σ^2. Then the sampling distribution of $\overline{X}$ is $N(\mu, \frac{\sigma^2}{n})$. One has

$$\mathbb{P}(|Z| < z_{\alpha/2}) = 1 - \alpha,$$

where $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ and $z_{\alpha/2}$ is such that the integral of the standard normal density from $z_{\alpha/2}$ to ∞ equals to $\alpha/2$. It follows that

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{2.1}$$

This is summarized in the following theorem:

Theorem 2.1: Coverage Probability

If $\overline{X}$, the mean of a random sample of size n from a normal population with the known variance σ^2, is to be used as an esimator of the mean of the population, then the probability is $1 - \alpha$ that the error will be less thatn $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.

To construct a confidence interval for estimating the mean of a normal population with the known variance σ^2, one can rewrite (2.1) as

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{2.2}$$

We generalize this into the following result, which offers a way in finding the interval estimator with the desired coverage probability.

Theorem 2.2: Estimation for Mean, Known Variance

If $\bar{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, then

$$\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is a $(1 - \alpha)$ 100% confidence interval for the mean of the population.

When we are dealing with a random sample from a normal population, e.g. $n < 30$, and σ is unknown. Then **Theorem 2.1** and **Theorem 2.2** cannot be used. Instead, we make use of the fact that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}},$$

is a random variable having the t-distribution with $n - 1$ degrees of freedom. Substituting $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ for T in

$$\mathbb{P}\big(- t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}\big) = 1 - \alpha,$$

summarizing, we have the following result.

Theorem 2.3: Estimation for Mean, Unkown Variance

> If $\bar{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population, then
>
> $$\bar{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$
>
> is a $(1 - \alpha)$ 100% confidence interval for the mean of the population.

The method by which we constructed confidence intervals in this subsection consisted essentially of finding a suitable random variable whose values are determined by the sample data as well as the population parameters, yet whose distribution does not involve the parameter we are trying to estimate. This method of confidence interval construction is called the pivotal method and it is widely used in finding interval estimators.

Now we introduce some results in finding the interval estimation for the difference between means, i.e. rather than estimating only $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$, we shall proceed to the

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

which has the standard normal distribution. If we substitute this expression for Z into the pivotal method yields the following result.

Theorem 2.4: Estimation for Difference between Means, Known Variance

> If $\bar{x}_1$ and $\bar{x}_2$ are the values of the means of independent random samples of sizes n_1 and n_2 from normal populations with the known variances σ_1^2 and σ_2^2, then
>
> $$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
>
> is a $(1 - \alpha)$ 100% confidence interval for the difference between the two population means.

By virtue of the **Central Limit Theorem**, this confidence interval formula can also be used for independent random samples from non-normal populations with known variances when n_1 and n_2 are large.

For the unknown variance case, we have the following result:

Theorem 2.5: Estimation for Difference between Means, Unknown Variance

If $\bar{x}_1, \bar{x}_2, s_1$, and s_2 are the values of the means and the standard deviations of independent random samples of sizes n_1 and n_2 from normal populations with equal variances s_p, then

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2, n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2$$
$$< (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2, n_1+n_2-2} \cdot s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is a $(1 - \alpha)$ 100% confidence interval for the difference between the two population means.

Since this confidence interval formula is used mainly when n_1 and/or n_2 are small, e.g. less than 30, we refer to it as a small sample confidence interval for $\mu_1 - \mu_2$.

2.2 Estimation for Normal Variances

Given a random sample of size n from a normal distribution, we can obtain a $(1 - \alpha)$ 100% confidence interval for σ^2 by making use of the fact that $\frac{(n-1)S^2}{\sigma^2}$ is a random variable having a chi-square distribution with $n - 1$ degrees of freedom. Thus,

$$\mathbb{P}\left(\chi^2_{1-\alpha/2, n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2, n-1}\right) = 1 - \alpha,$$

by simple calculation we have

$$\mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}}\right) = 1 - \alpha.$$

Summarizing, we have the following result.

Theorem 2.6: Estimation for Variance

If s^2 is the value of the variance of a random sample of size n from a normal population. Then

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}}$$

is a $(1 - \alpha)$ 100% confidence interval for σ^2.

Corresponding $(1 - \alpha)$ 100% confidence limits for σ can be obtained by taking the square roots of the confidence limits for σ^2.

If S_1^2 and S_2^2 are the variances of independent random samples of sizes n_1 and n_2 from normal populations, then, according to **Fact 1** and

$$f_{1-\alpha/2,n_1-1,n_2-1} = \frac{1}{f_{\alpha/2,n_2-1,n_1-1}},$$

we have the following result.

Theorem 2.7: Estimation for Ratio of Variances

If s_1^2 and s_2^2 are the values of the variances of independent random samples of sizes n_1 and n_2 from normal populations, then

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot f_{\alpha/2,n_2-1,n_1-1}$$

is a $(1-\alpha)$ 100% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

Corresponding $(1-\alpha)$ 100% confidence limits for $\frac{\sigma_1}{\sigma_2}$ can be obtained by taking the square roots of the confidence limits for $\frac{\sigma_1^2}{\sigma_2^2}$.

2.3 The Estimation of Proportions (Binomials)

In many problems we must estimate proportions, probabilities, percentages, or the rates. In many of these it is reasonable to assume that we are sampling a binomial population and, hence, that our problem is to estimate the binomial parameter θ. Thus, we can make use of the fact that for large n the binomial distribution can be approximated with a normal distribution, i.e.

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

can be treated as a random sample having approximately the standard normal distribution. Substituting the expression for Z into

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

one has

$$\mathbb{P}\Big(- z_{\alpha/2} < \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} < z_{\alpha/2}\Big) = 1 - \alpha.$$

Summarizing, we have the following result.

Theorem 2.8: Interval Estimation.

If $X \sim$ Binomial(n, θ). If n is large and $\hat{\theta} = \frac{x}{n}$, then

$$\hat{\theta} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} < \theta < \hat{\theta} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

is an approximate $(1-\alpha)$ 100% confidence interval for θ. Moreover, the error, with $(1-\alpha)$ 100%, is less than $z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$.

In many problems we must estimate the difference between the binomial parameters θ_1 and θ_2 on the basis of independent random samples of sizes n_1 and n_2 from two binomial populations.

Theorem 2.9: Estimation for Differences between Proportions

If X_1 is a binomial random variable with the parameters n_1 and θ_1, X_2 is a binomial random variable with the parameters n_2 and θ_2. If n_1 and n_2 are large and $\hat{\theta}_1 := \dfrac{X_1}{n_1}$ and $\hat{\theta}_2 := \dfrac{X_2}{n_2}$, then

$$(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} < \theta_1 - \theta_2$$

$$< (\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

is an approximate $(1-\alpha)$ 100% confidence interval for $\theta_1 - \theta_2$.

2.4 The Pivotal Method

The use of pivotal quantities for confidence set construction, resulting in what has been called pivotal inference, is mainly due to Barnard (1949, 1980) but can be traced as far back as Fisher (1930), whose used the term inverse probability. Closely related is D.A.S. Fraser's theory of structural inference (Fraser 1968, 1979). An interesting discussion of the strengths and weakness of these methods is given in Berger and Wolpert (1984).

Definition: Pivotal Quantity (Pivot)

A random variable $Q(X, \theta) = Q(X_1, \cdots, X_n, \theta)$ is a pivotal quantity (or pivot) if the distribution of $Q(X, \theta)$ is independent of all parameters. That is, if $X \sim F(x \mid \theta)$, then $Q(X, \theta)$ has the same distribution for all values of θ.

The function $Q(x, \theta)$ will usually explicitly contain both parameters and statistics, but for any set A, $\mathbb{P}_\theta(Q(X, \theta) \in A)$ cannot depend on θ. The technique of constructing confidence set from pivots relies on being able to find a pivot and a set A so that the set $\{\theta \mid Q(x, \theta) \in A\}$ is a set estimate of θ.

We have seen in the previous subsections methods in finding pivots, for estimating means (resp. difference between means) for known/unknown variances, for estimating variances (resp. ratio between variances), and estimation for binomials (resp. difference between binomials). The$f(x-\mu)$techniques we used is mainly by the rewriting the pdf, or, rooting by the location-scale property. We summarize this method in the following table.

Form of PDF	Type of PDF	Pivotal Quantity
$f(x-\mu)$	Location	$\overline{X} - \mu$
$\frac{1}{\sigma} f(\frac{x}{\sigma})$	Scale	$\frac{\overline{X}}{\sigma}$
$\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$	Location-Scale	$\frac{\overline{X} - \mu}{S}$

(Table 2.1)

Alternatively, if we base our confidence interval construction for a parameter θ on a real-valued statistic T with cdf $F_T(t\,|\,\theta)$. We will first assume that T is a continuous random variable. The situation where T is discrete is similar but has a few additional technical details to consider. We therefore state the discrete case in a separate theorem. To do so we need a new definition.

Definition: Stocahstically Increasing (Decreasing)

A family of cdfs $F(t\,|\,\theta)$ is stochastically increasing (resp. decreasing) in θ if for each $t \in \mathscr{T}$, the sample space of T, $F(t\,|\,\theta)$ is a decreasing (resp. decreasing) function of θ.

In what follows, we need only the fact that F is monotone, either increasing or decreasing. The more statistical concepts of stochastic increasing or decreasing merely serve as interpretations.

Theorem 2.10: Pivoting a Continuous CDF

Let T be a statistic with continuous cdf $F_T(t\,|\,\theta)$. Let $\alpha_1 + \alpha_2 =: \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathscr{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows:

(i) If $F_T(t\,|\,\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $F_T(t\,|\,\theta_U(t)) = \alpha_1$ and $F_T(t\,|\,\theta_L(t)) = 1 - \alpha_2$.

(ii) If $F_T(t\,|\,\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $F_T(t\,|\,\theta_U(t)) = 1 - \alpha_2$ and $F_T(t\,|\,\theta_L(t)) = \alpha_1$.

Then the random inverval $\big(\theta_L(T), \theta_U(T)\big)$ is a $1 - \alpha$ confidence interval for θ.

As for the discrete case.

Theorem 2.11: Pivoting a Discrete CDF

Let T be a discrete statistic with cdf $F_T(t\,|\,\theta) = \mathbb{P}(T \le t\,|\,\theta)$. Let $\alpha_1 + \alpha_2 =: \alpha$ with $0 < \alpha < 1$ be fixed values. Suppose that for each $t \in \mathscr{T}$, the functions $\theta_L(t)$ and $\theta_U(t)$ can be defined as follows:

(i) If $F_T(t\,|\,\theta)$ is a decreasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $\mathbb{P}(T \le t\,|\,\theta_U(t)) = \alpha_1$ and $\mathbb{P}(T \ge t\,|\,\theta_L(t)) = \alpha_2$.

(ii) If $F_T(t\,|\,\theta)$ is an increasing function of θ for each t, define $\theta_L(t)$ and $\theta_U(t)$ by $\mathbb{P}(T \ge t\,|\,\theta_U(t)) = \alpha_1$ and $\mathbb{P}(T \le t\,|\,\theta_L(t)) = \alpha_2$.

Then the random inverval $\big(\theta_L(T), \theta_U(T)\big)$ is a $1 - \alpha$ confidence interval for θ.

3. Inverting a Test Statistic

There is a very strong correspondence between hypothesis testing and interval estimation. In fact, we can say in general that every confidence set corresponds to a test statistic and vice versa.

The acceptance region of the hypothesis test, the set in the sample space for which $H_0 : \mu = \mu_0$ is accepted, is given by

$$A(\mu_0) := \Big\{(x_1, \cdots, x_n) \Big| \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x} \le \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\Big\},$$

and the confidence interval, the set in the parameter space with plausible values of μ, is given by

$$C(x_1, \cdots, x_n) = \{\mu : \bar{x} - z_{\alpha_2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\}.$$

These sets are connected to each other by the relation

$$(x_1, \cdots, x_n) \in A(\mu_0) \Leftrightarrow \mu_0 \in C(x_1, \cdots, x_n). \tag{3.1}$$

We now summarize this correspondence in the following theorem.

Theorem 2.12:

$\forall\theta_0 \in \Theta$, let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. $\forall x \in \mathcal{X}$, define $C(x) := \{\theta_0 | \theta_0 \in A(x)\}$. Then the random set $C(x)$ is a $1 - \alpha$ confidence set. Conversely, let $C(X)$ be a $1 - \alpha$ confidence set. $\forall\theta_0 \in \Theta$, define $A(\theta_0) := \{x | \theta_0 \in C(x)\}$. Then $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Proof:

Since $A(\theta_0)$ is assumed to be the acceptance region of a level α test, therefore, one has $\mathbb{P}_\theta(X \in [A(\theta_0)]^c) \leq \alpha \Leftrightarrow \mathbb{P}_\theta(X \in A(\theta_0)) \geq 1 - \alpha$. Since θ_0 is arbitrary, w.l.o.g., we may use θ to replace θ_0, then one has, by the definition of $C(x)$, $\mathbb{P}_\theta(\theta \in C(X)) = \mathbb{P}_\theta(X \in A(\theta)) \geq 1 - \alpha \Rightarrow C(X)$ is a $1 - \alpha$ confidence set.

Conversely, the Type I Error probability for the test of $H_0 : \theta = \theta_0$ with acceptance region $A(\theta_0)$ is:

$$\mathbb{P}_\theta(X \in [A(\theta_0)]^c) = \mathbb{P}_\theta(\theta \in [C(X)]^c) \leq \alpha,$$

result follows.

□

This makes it clear why we really have a family of tests, one for each value of $\theta_0 \in \Theta$, that we invert to obtain one confidence set.

The fact that tests can be inverted to obtain a confidence set and vice verse is an interesting theoretical task, but in fact only the first part of the theorem is at the most usefulness. Constructing a level α acceptance region could be an easy task, but constructing a confidence set is sometimes, in fact, most of the times, a more difficult task. Therefore, the method of finding(obtaining) such a set by inverting an acceptance region is useful; all the techniques we used to find tests could offer help in finding(constructing) confidence sets.

One practice, when constructing a confidence set by test inversion, we will have in mind an alternative hypothesis such as $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$. The alternative will dictate the form of $A(\theta_0)$ that is reasonable, and the form of $A(\theta_0)$ will determine the shape of $C(x)$.

In fact, this inverting process could be affected by the property. For example, unbiased tests, when inverted, will produce unbiased confidence sets. As it menti-oned, we can use sufficient statistics to find good confidence sets.

Example 3.1: Inverting a Normal Set

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ and considering testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. For a fixed α level, a reasonable test(in fact the most powerful unbiased test) has rejection region $\left\{x \middle| \, |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\right\}$.

Note that H_0 is accepted for sample points with $|\bar{x} - \mu| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, or, equivalently,

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Since the test has size α, this means that $\mathbb{P}(H_0 \text{ rejected } | \mu = \mu_0) = \alpha$, or, stated in another way,

$$\mathbb{P}(H_0 \text{ accepted} \,|\, \mu = \mu_0) = 1 - \alpha.$$

Combining this with the above characterization of the acceptance region, one has that:

$$\mathbb{P}\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \middle| \mu = \mu_0\right) = 1 - \alpha.$$

But this probability statement is valid $\forall \mu_0$, hence one has

$$\mathbb{P}\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

is true. The interval

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

is obtained by inverting the acceptance region of the level α test, is a $1 - \alpha$ confidence interval. ||

We close this section by introducing finding the interval estimator by inverting an LRT (likelihood Ratio Test).

Example 3.2:

Suppose that we want a confidence interval for the mean λ_1 of an Exponential(λ) population. We can obtain such an interval by inverting a level α test of

$$H_0 : \lambda = \lambda_0 \text{ versus } H_1 : \lambda \neq \lambda_0.$$

If we take a random sample $X_1, \ldots, X_n$, the LRT statistic is

$$\frac{\frac{1}{\lambda_0^n}\exp\{-\sum x_i/\lambda_0\}}{\sup_\lambda \frac{1}{\lambda^n}\exp\{-\sum x_i/\lambda\}} = \frac{\lambda_0^{-n}\exp\{-\sum x_i/\lambda_0\}}{\frac{1}{(\sum x_i/n)^n}e^{-n}} = (\frac{\sum x_i}{n\lambda_0})^n e^n e^{-\sum x_i/\lambda_0}.$$

For feed λ_0, the acceptance region is given by

$$A(\lambda_0) = \{x \middle| (\frac{\sum x_i}{n\lambda_0})^n e^{-\sum x_i/\lambda_0} \ge k^*\},$$

where k^* is a constant chosen such that $\mathbb{P}_{\lambda_0}(X \in A(\lambda_0)) = 1 - \alpha$. Inverting this acceptance region gives the $1 - \alpha$ confidence set:

$$C(x) = \{\lambda \middle| (\frac{\sum x_i}{\lambda})^n e^{-\sum x_i/n} \ge k^*\} \qquad (3.2)$$

The expression defines C depends on x only through $\sum x_i$, so the confidence interval can be expressed in the form:

$$C(\sum x_i) = \{\lambda \Big| L(\sum x_i) \leq \lambda \leq U(\sum x_i)\},$$

where L and U are functions determined by the constraints in (3.2) with probability $1-\alpha$ and

$$\left(\frac{\sum x_i}{L(\sum x_i)}\right)^n e^{-\sum x_i / L(\sum x_i)} = \left(\frac{\sum x_i}{U(\sum x_i)}\right)^n e^{-\sum x_i / U(\sum x_i)}.$$

Without loss of generality, we may set

$$\frac{\sum x_i}{L(\sum x_i)} =: a \text{ and } \frac{\sum x_i}{U(\sum x_i)} =: b$$

such that $a > b$ are constants. Then one has $a^n e^{-a} = b^n e^{-b}$ which yields easily to numerical solution.

To work with details, let $n=2$ and note that $\sum X_i \sim \text{Gamma}(2,\lambda)$ and $\sum X_i/\lambda \sim \text{Gamma}(2,1)$. Hence the confidence interval becomes

$$\left\{\lambda | \frac{1}{a}\sum x_i < \lambda < \frac{1}{b}\sum x_i\right\},$$

where a and b satisfy

$$\mathbb{P}_\lambda(\frac{1}{a}\sum X_i \leq \lambda \leq \frac{1}{b}\sum X_i) = \mathbb{P}(b \leq \frac{\sum X_i}{\lambda} \leq a) = 1-\alpha$$

and $a^2e^{-a} = b^2e^{-b}$, thus,

$$\mathbb{P}(b \leq \frac{\sum X_i}{\lambda} \leq a) = \int_a^b te^{-t}dt = e^{-b}(b+1) - e^{-a}(a+1).$$

To get, e.g., a 90 % confidence interval, we must simultaneously satisfy the probability condition and constraints, to the third decimal, say $a = 5.480$ and $b = 0.441$, with confidence coefficient 0.90006. Thus

$$\mathbb{P}_\lambda(\frac{1}{5.480}\sum X_i \leq \lambda \leq \frac{1}{0.441}\sum X_i) = 0{,}90006. \qquad \|$$

4. Methods in Evaluating Interval Estimators

We now have seen many methods for deriving confidence sets and, in fact, we can derive different confidence sets for the same problem. In such situations we would, of course, want to choose a best one. Therefore, we now examine some methods and criteria for evaluating set estimators.

One of the most straightforward one may be to increase the coverage probability and reduce the size of interval estimator. We will also talk about some optimization results, either optimization with respect to the loss or with respect to the correspo-nding test statistics. We start with the coverage probability and size and then we discuss the optimization with respect to the corresponding test statistics. The optimization with respect to the loss function, together with the Bayesian interval estimation, is discussed in our second lecture notes, hence omitted here.

4.1 Size and Coverage Probability

We consider what appears to be a simple, constrained minimization problem. For a given, specified coverage probability find the confidence interval with the shortest length. We first consider an example.

Example 4.1: Optimizing Length

Let $X_1, \dots, X_n$ be iid $N(\mu, \sigma^2)$, where σ is known. From the fact that $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is a pivot with a standard normal distribution, any a and b such that $\mathbb{P}(a \le Z \le b) = 1 - \alpha$ will give a $1 - \alpha$ confidence interval

$$\left\{\mu \,|\, \bar{x} - b\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} - a\frac{\sigma}{\sqrt{n}}\right\}.$$

It is natural to ask which choice of a and b is the best? More formally, which choice of a and b will minimize the length of the confidence interval while preserving the $1 - \alpha$ coverage? Notice that the length of the confidence interval is equal to $(b - a)\sigma/\sqrt{n}$, since the factor $\sigma/\sqrt{n}$ is part of each interval length, it can be ignored and therefore the length turns out to be $(b - a)$. Thus, we want to find a pair of numbers of a and b such that $\mathbb{P}(a \le Z \le b) = 1 - \alpha$ and minimizes $b - a$.

Take $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, but no mention of optimality. If we take $1 - \alpha = 0.9$. Then

a	b	Probability	$b-a$
-1.34	2.33	$\mathbb{P}(Z < a) = 0.09, \mathbb{P}(Z > b) = 0.01$	3.67
-1.44	1.96	$\mathbb{P}(Z < a) = 0.075, \mathbb{P}(Z > b) = 0.025$	3.40
-1.65	1.65	$\mathbb{P}(Z < a) = 0.05, \mathbb{P}(Z > b) = 0.05$	3.30

This numerical study suggest that the choice of $(a, b) = (-1.65, 1.65)$ gives the best interval, and, in fact, it does. In this case splitting α equally is the best strategy. ||

The strategy of splitting α equally, which is optimal in the above case, is not always optimal. What makes the equal α split optimal in the above case is the fact that the height of the pdf is the same at $-z_{\alpha/2}$ and $z_{\alpha/2}$. We now prove a theorem that will demonstrate this fact, a theorem that is applicable in some generality, needing only the assumption that the pdf is unimodal.

Definition: unimodal

A pdf $f(x)$ is said to be unimodal if there exists x^* such that

$$f(x) \text{ is } \begin{cases} \text{non-decreasing, if } x \le x^* \\ \text{non-increasing, if } x \ge x^* \end{cases}.$$

Theorem 4.1:

Let $f(x)$ be a unimodal PDF. If the interval $[a, b]$ satisfies that

(i) $\displaystyle\int_a^b f(x)dx = 1 - \alpha$

(ii) $f(a) = f(b) > 0$

(iii) $a \le x^* \le b$, where x^* is a mode of $f(x)$.

Then $[a, b]$ is the shortest among all intervals satisfying (i).

Proof:

Without loss of generality, we may assume that $[a', b']$ is any other interval such that $b' - a' < b - a$.

WTS: $\displaystyle\int_{a'}^{b'} f(x)dx < 1 - \alpha.$

The result will be proved only for $a' \le a$, the proof being similar if $a < a'$. Also, two cases need to be considered, $b' \le a$ and $b' > a$.

(i) If $b' \le a \Rightarrow a' \le b' \le a \le x^*$ and

$$\begin{aligned}
\int_{a'}^{b'} f(x)dx &\le f(b')(b' - a') && (x \le b' \le x^* \Rightarrow f(x) \le f(b'))\\
&\le f(a)(b' - a') && (b' \le a \le x^* \Rightarrow f(b') \le f(a))\\
&\le f(a)(b - a) && (b' - a' < b - a \text{ and } f(a) > 0)\\
&\le \int_a^b f(x)dx && ((\text{ii}), (\text{iii}), \text{unimodality} \Rightarrow f(x) \ge f(b))\\
&= 1 - \alpha.
\end{aligned}$$

(ii) If $b' > a \Rightarrow a' \le a < b' < b$ for, if $b' \ge b$ then $b' - a' \ge b - a$. In this case, one writes

$$\begin{aligned}
\int_{a'}^{b'} f(x)dx &= \int_a^b f(x)dx + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right]\\
&= (1 - \alpha) + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right].
\end{aligned}$$

[Claim]: $\displaystyle\int_{a'}^{a} f(x) < \int_{b'}^{b} f(x)dx.$

Using the unimodality of f, the ordering $a' \le a < b' < b$ and by assumption (ii), one has

$$\int_{a'}^{a} f(x)dx \le f(a)(a' - a) \text{ and } \int_{b'}^{b} f(x)dx \ge f(b)(b' - b).$$

Thus, one has,

$$\begin{aligned}
\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx &\le f(a)(a' - a) - f(b)(b' - b)\\
&= f(a)[(a - a') - (b - b')]\\
&= f(a)[(b - a') - (b - a)]
\end{aligned}$$

where the first equality holds since $f(a) = f(b)$ by assumption, and the last expression is negative if $(b' - a') < (b - a)$ and $f(a) > 0$.

□

If more assumptions are applied to the theorem, e.g., continuity of f, will simplify the proof. Moreover, the equal α split will be optimal for any symmetric unimodal pdf. Furthermore, this theorem may even apply when the optimality criterion is somewhat different from the minimum length.

4.2 Test-Related Optimization

Since there is a one-to-one correspondence between confidence sets and the tests of hypotheses, there is some correspondence between optimality on them. The probability of covering the false values, of the probability of false coverage, indirectly measures the size of a confidence set.

Definition: Uniformly Most Accurate(UMA) confidence set

A $1-\alpha$ confidence set that minimizes the value of false coverage over a class of $1-\alpha$ confidence sets is called a uniformly most accurate (UMA) confidence set.

Remark:

UMA confident sets are constructed by inverting the acceptance regions of UMP tests. UMA confidence sets, unfortunately, exists only in a small range of circumstances. UMP is usually one-sided ⇒ so are UMA intervals. ||

Theorem 4.2: UMA Lower Confidence Bound

Let $X \sim f(x\,|\,\theta)$, where θ is a real-valued parameter. For each $\theta_0 \in \Theta$, let $A^*(\theta_0)$ be the UMP level α acceptance region of a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Let $C^*(x)$ be the $1-\alpha$ confidence set formed by inverting the UMP acceptance regions. Then, for any other $1-\alpha$ confidence set C, $\mathbb{P}_\theta(\theta' \in C^*(X)) \le \mathbb{P}_\theta(\theta' \in C(X)) \forall \theta' < \theta$.

Proof:

Let θ' be any value smaller than θ. Let $A(\theta')$ be the acceptance region of the level α test of $H_0 : \theta = \theta'$ obtained by inverting C. Since $A^*(\theta')$ is the UMP acceptance region for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ by assumption and since $\theta > \theta'$, one has:

$$\begin{aligned}\mathbb{P}_\theta(\theta' \in C^*(X)) &= \mathbb{P}_\theta(X \in A^*(\theta')) && \text{(Invert the confidence set)}\\ &\le \mathbb{P}_\theta(X \in A(\theta')) && \text{(Since } A^* \text{ is UMP and true for any } A)\\ &= \mathbb{P}_\theta(\theta' \in C(X)) && \text{(Invert } A \text{ to obtain } C)\end{aligned}$$

Notice that the above inequality is "≤" because we are working with probabilities of acceptance regions. This is 1−power, so UMP tests will minimize these acceptance region probabilities. Therefore, we have established that for $\theta' < \theta$, the probability of false coverage is minimized by the interval obtained from inverting the UMP test.

□

The UMA confidence set in the above theorem is constructed by inverting the family of tests for the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta > \theta_0,$$

where the form of confidence set is governed by the alternative hypothesis. The above alternative hypothesis, which specify that θ_0 is less than a particular value, they are of the form $[L(X), \infty)$.

Example 4.2: UMA Confidence Bound

Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, where σ^2 is known. The interval $C(\bar{x}) := \{\mu \,|\, \mu \geq \bar{x} - z_{\alpha/2}\sigma/\sqrt{n}\}$ is a $1-\alpha$ UMA lower confidence bound since it can be obtained by inverting the UMP test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$. The more common two-sided interval, $C(\bar{x}) := \{\mu \,|\, \bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\}$ is not UMA, since it is obtained by inverting the two-sided acceptance region from the test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, hypothesis for which no UMP test exists. ||

In the testing problem, when considering two-sided tests, we found the property of unbiasedness to be both compelling and useful. In the confidence interval problem, similar ideas apply. When we deal with two-sided confidence intervals, it is reasonable to restrict considerations to unbiased confidence sets. Remember that an unbiased test is one in which the power in the alternative is always greater than the power of the null.

Definition: Unbiased $1-\alpha$ Confidence Set

We say a $1-\alpha$ confidence set $C(x)$ is unbiased if $\mathbb{P}_\theta(\theta' \in C(X)) \leq 1-\alpha$ $\forall \theta \neq \theta'$.

Thus, for an unbiased confidence set, the probability of the false coverage is never more than the minimum probability of true coverage. Unbiased confidence sets can be obtained by inverting the unbiased sets. That is, if $A(\theta_0)$ is an unbiased level α acceptance region of a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ and $C(x)$ is the $1-\alpha$ confidence set formed by inverting the acceptance regions, then $C(x)$ is an unbiased $1-\alpha$ confidence set.

Sets that minimize the probability of false coverage are also called Neymann shortest. The fact that there is a length connotation to this name is somewhat satisfied by the following theorem:

Theorem 4.3: Pratt

Let X be a real-valued random variable with $X \sim f(x\,|\,\theta)$, where θ is a real-valued parameter. Let $C(x) := [L(x), U(x)]$ be a confidence interval for θ. If $L(x)$ and $U(x)$ are both increasing functions of x. Then, for any values of θ^*, one has

$$\mathbb{E}_{\theta^*}[\text{Length}(C(X))] = \int_{\theta \neq \theta^*} \mathbb{P}_{\theta^*}(\theta \in C(X))d\theta.$$

Pf:

From the definition of the expected values, one has

$$\begin{aligned}\mathbb{E}_{\theta^*}[\text{Length}(C(X))] &= \int_{\mathcal{X}} \text{Length(C(X))} \cdot f(x\,|\,\theta^*)dx \\ &= \int_{\mathcal{X}} [U(X) - L(X)] \cdot f(x\,|\,\theta^*)dx \quad \text{(def of Length)} \\ &= \int_{\mathcal{X}} \left(\int_{L(x)}^{U(x)} \right) f(x\,|\,\theta^*)dx \quad \text{(test } \theta \text{ as dummy variable)}\end{aligned}$$

$$= \int_{\Theta} \left(\int_{U^{-1}(x)}^{L^{-1}(x)} f(x \mid \theta^*) dx \right) d\theta \qquad (4.1)$$

(invert the order of integration)

$$= \int_{\Theta} \left[\mathbb{P}_{\theta^*} \left(U^{-1}(\theta) \le X \le L^{-1}(\theta) \right) \right] d\theta$$

(by definition)

$$= \int_{\theta \ne \theta^*} \mathbb{P}_{\theta^*}(\theta \in C(X)) d\theta$$

The last equality holds by the fact that removing the point $\theta = \theta^*$ does not change the value of the integration(a measure 0 set removed). In step (4.1), the interchange of integrals is formally justified by **Fubini's Theorem** but is easily seen to be justified as long as all of the integrands are finite.
Moreover, the inversion of the confidence interval is standard, where we use the relationship

$$\theta \in \{\theta \mid L(x) \le \theta \le U(x)\} \Leftrightarrow x \in \{x \mid U^{-1}(\theta) \le x \le L^{-1}(\theta)\},$$

which is valid because of the assumption that L and U are both increasing functions $\forall x \in \mathcal{X}$. Furthermore, the theorem could be modified to apply to an integral with decreasing endpoints.

□

This theorem says that the expected length of $C(x)$ is equal to the sum(integral) of the probability of false coverage, where the sum(integral) is taking over all false values of the parameter.

Reference:

[1]: George Casella, Roger L. Berger, *Statistical Inference, Second Edition*, pp. 417-447.
[2]: Irwin Miller, Marylees Miller, *John E. Freund's Mathematical Statistics with Applications, Eighth Edition*, pp. 317-337.
[3]: Dennis D. Wackerly, William Mendenhall III, Richard L. Scheaffer, *Mathematical Statistics with Applications, Seventh Edition*, pp. 406-435.

Review on Hypothesis Testing

Tianyu Zhang[3]

Abstract:

In this short monograph we offer a review on hypothesis testing. Serving as a complementary estimation other than point estimators, the interval estimations could offer us a way in describing the error and the chance of success. We introduce the methods of evaluating the tests and then proceed to talk about the methods in finding them.

Table of Contents:

1. Introduction

Once we have an estimator for a parameter θ, it is vital to know how good (or bad) this estimator perform. The performance is evaluated by either the biasedness and variance, the consistency, the translation invariance property, or sometimes the asymptotic normality. Since the existence of the UMVE does not always exist, nor even the unbiased estimators, then given a collection of estimators we should be able to have them comparable one with another, this is done by the loss function with the corresponding risk. Comparing the loss function leads to a "wise" choice, or at least offers us a way to optimize the estimators.

This methodology does not only suit for the point estimators, serving as a special type of point estimation, the interval estimators find themselves fitted too. We also know that one of a way to find the interval estimation is by inverting the test statistic. This natural correspondence leads to the investigation of hypothesis testing. In fact, after making a prediction, we need to know if our prediction is reasonable, hence we use a test statistic to describe its behavior, serving this purpose, a good test statistic also has the important information about the parameter.

Our investigation of the test statistic follows the same structure as we introduced the others. We start with the method of evaluating test statistics, then we introduce the methods in finding them. We do not discuss the Bayesin hypothesis testing and for interested readers may consult [1].

[3] YMSC, BIMSA, bidenbaka@gmail.com

2. Methods in Evaluating Tests

Definition: Hypothesis

A hypothesis is a statement about a population parameter.

The definition of a hypothesis is rather general, but the improvement point is that a hypothesis makes a statement about the population. The goal of a hypothesis test is to decide, based on a sample from the population, which of two complementary hypotheses is true.

Definition: Null and Alternative Hypothesis

The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis and the alternative hypothesis. They are denoted by H_0 and H_1, respectively.

In a hypothesis testing problem, after observing the sample the experimenter must decide either to accept H_0 as true or to reject H_0 as false and decide H_1 is true.

Definition: Hypothesis Testing Procedure/ Hypothesis Test

A hypothesis testing procedure or hypothesis test is a rule that specifies

(i) For which sample values the decision is made to accept H_0 as true.

(ii) For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the rejection region or critical region. The complement of the rejection region is called the acceptance region.

In deciding to accept or reject the null hypothesis H_0, an experimenter might be making a mistake. Usually, hypothesis tests are evaluated and compared through their probabilities of making mistakes. In this subsection we discuss how these error probabilities can be controlled. In some cases, it can even be determined which tests have the smallest possible error probabilities.

We will go through five methods in this subsection, in 2.1 we introduce the (1) Error Probabilities and Power Function, then in 2.2 we treat the (2) Most Powerful Tests, next in 2.3 we discuss the (3) p-Values to close this section.

2.1 Error Probabilities and Power Function

Suppose that R denotes the rejection region for a test. Then for $\theta \in \Theta_0$, the test will make a mistake if $x \in R$, so the probability of a Type I Error is $\mathbb{P}_\theta(X \in R)$. For $\theta \in \Theta_0^c$, the probability of a Type II Error is $\mathbb{P}_\theta(X \in R^c)$. This switching from R to R^c is a bit confusing but if we realize that $\mathbb{P}_\theta(X \in R^c) = 1 - \mathbb{P}_\theta(X \in R)$. This consideration leads to the following definition of the power function.

Definition: Power Function

The power function of a hypothesis test with rejection region R is the function of θ defined by

$$\beta(\theta) := \mathbb{P}_\theta(X \in R) = \begin{cases} \text{probability of a Type I Error}, \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error}, \theta \in \Theta_0^c \end{cases}.$$

Remark:

The ideal power function is 0 $\forall \theta \in \Theta_0$ and 1 $\forall \theta \in \Theta_0^c$. Except in trivial situations, this ideal cannot be attained. Qualitively, a good test has power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0$. ||

Example 2.1: Binomial Power Function

Let $X \sim$ Binomial(5,θ). Consider testing $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$.

Consider first the test that rejects $H_0 \Leftrightarrow$ all "success" are observed. The power function for this test is:

$$\beta_1(\theta) = \mathbb{P}_\theta(X \in R) = \mathbb{P}_\theta(X = 5) = \theta^5.$$

Although the probability of a Type I Error is reasonable low, i.e.,

$$\beta_1(\theta) \leq (\frac{1}{2})^5 = 0.312 \ \forall \theta \leq \frac{1}{2},$$

the probability of a Type II Error is too high, i.e., $\beta_1(\theta)$ is too small for most $\theta > \frac{1}{2}$. The probability of Type II Error is less than $\frac{1}{2}$ only if

$$\theta > (\frac{1}{2})^{\frac{1}{5}} = 0.87.$$

To achieve smaller Type II Error probabilities, we might consider using the test that rejects H_0 if $X = 3,4$, or 5. The power function then will be:

$$\beta_2(\theta) = \mathbb{P}_\theta(X \in \{3,4,5\}) = \binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{4}\theta^4(1-\theta) + \binom{5}{5}\theta^5.$$

The second test achieves a smaller Type II Error than the first test, but as a consequence, it has bigger Type I Error than the first test. ||

Therefore, when choosing the test, sometimes we are facing a trade-off problem, whether deciding which side the optimization occurs, the de-optimization inevitably occurs on the other side, so the researchers should be careful in choosing the test in achieving their goals.

Typically, the power function of a test will depend on the sample size n. If n can be chosen by the experimenter, consideration of the power function might be helpful in determining what sample size is appropriate for an experiment.

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the Type I Error probability at a specified level. Within this class of tests we then search for tests that have Type II Error probability that is as small as possible. The following two terms are useful when discussing tests that control Type I Error probabilities.

Definition: Size α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

Definition: Level α Test

For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size α test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

Some authors do not make distinction between these two definitions. We made the distinction here to stress out the fact that sometimes having a size α test is difficult, so in practical terms, one should make compromises with the alternative level α test.

Remark:

Typical α level tests use $\alpha = 0.01, 0.05$, and 0.10, **but be aware that in fixing the level α test, the experimenter is controlling only the Type I Error.** An LRT is one rejects H_0 if $\lambda(X) \leq c$, for example. ||

Other than α levels, there are other features of a test that might also be of concern. For example, we would like a test to be more likely to reject H_0 if $\theta \in \Theta_0^c$ than if $\theta \in \Theta_0$. This property is called unbiased.

Definition: Unbiased Power Function

A test with power function $\beta(\theta)$ is unbiased if $\beta(\theta') \geq \beta(\theta'') \forall \theta' \in \Theta_0^c$ and $\forall \theta'' \in \Theta_0$.

In most problems there are many unbiased tests. Likewise, there are many size α tests, LRTs, etc. In some cases we have imposed enough restrictions to narrow the consideration to one test. In other cases there remain many tests from which to choose. We discussed only the one that rejects H_0 for large values of T. In the following discussion we will discuss other criteria for selecting one out of a class of tests, criteria that are all related to the power functions of the tests.

2.2 The Uniform Most Powerful Tests

We have seen that the α tests could control the probability of a Type I Error, i.e. level α tests have Type I Error probabilities at most α for all $\theta \in \Theta_0$. A good test in such a class would also have a small Type II Error probability, i.e. a large power function for $\theta \in \Theta_0^c$. If one test has a smaller Type II Error probability than all other tests in the class, it would certainly be a strong contender for the best test in the class, a notion that is formalized in the next definition.

Definition: Uniformly Most Powerful (UMP) Test

Let $\mathscr{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathscr{C}$, with power function $\beta(\theta)$, is a uniformly most powerful class $\mathscr{C}$ test if $\beta(\theta) \geq \beta'(\theta) \forall \theta \in \Theta_0^c$ and $\forall \beta' \in \mathscr{C}$.

In this subsection, the class $\mathscr{C}$ will be the class of all level α tests. The test described in the above definition is then called a UMP level α test. For this test to be interesting, restriction to the class $\mathscr{C}$ must involve some restriction on the Type I Error probability. A minimization of the Type II Error probability without some control of the Type I Error is not very interesting.

The requirements in this definition are so strong that UMP does not exist in many realistic problems. But in problems that have UMP tests, a UMP test might well be considered the best test in the class. Thus, we would like to be able to identify UMP tests if they exist. The following famous theorem clearly describes which tests are UMP level α tests in the situation where the null and alternative hypotheses both consist of only one probability distribution for the sample.

Theorem 2.1: Neymann-Pearson Lemma

Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(x\,|\,\theta_i)$, $i = 0,1$, using a test with rejection region R such that

(i) $x \in R$, if $f(x\,|\,\theta_1) > kf(x\,|\,\theta_0)$,

(ii) $x \in R^c$, if $f(x\,|\,\theta_1) < kf(x\,|\,\theta_0)$,

for some $k \geq 0$ and $\alpha = \mathbb{P}_{\theta_0}(X \in R)$. Then

(a) Any test that satisfies (i) and (ii) is a UMP level α test. (Sufficiency)

(b) If there exists a test satisfies (i) and (ii) with $k > 0$, then every UMP level α test is a size α test and every UMP level α test satisfies the first condition except perhaps on a set with probability measure 0, i.e. on a set A such that $\mathbb{P}_{\theta_0}(X \in A) = \mathbb{P}_{\theta_1}(X \in A) = 0$. (Necessity)

Proof:

We will prove the theorem for the case that $f(x\,|\,\theta_0)$ and $f(x\,|\,\theta_1)$ are PDFs of continuous random variables. The proof of discrete random variables can be accomplished by replacing integrals with sums.

Note first that any test satisfies $\alpha = \mathbb{P}_{\theta_0}(X \in R)$ is a size α and, hence, a level α test because $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(X \in R) = \mathbb{P}_\theta(X \in R) = \alpha$, since Θ_0 has only one point.

WTS I: (a) is true.

To ease notion, we define a test function, a function such ϕ defined by $\phi(x) = \begin{cases} 1, x \in R \\ 0, x \in R^c \end{cases}$, i.e., it is the indicator function of the rejection region.

Let $\phi'(x)$ be the test function of any other level α test and let $\beta(\theta)$, $\beta'(\theta)$ be the corresponding power function of $\phi(x)$ and $\phi'(x)$, respectively.

Since $0 \leq \phi'(x) \leq 1$, by the first assumption,

$x \in R \Rightarrow \phi(x) = 1 \geq \phi'(x)$, as well as $f(x\,|\,\theta_1) > kf(x\,|\,\theta_0)$, hence

$[\phi(x) - \phi'(x)] \cdot [f(x\,|\,\theta_1) - kf(x\,|\,\theta_0)] \geq 0 \ \forall x$.

Then we apply the integration and obtain:

$$0 \leq \int [\phi(x) - \phi'(x)] \cdot [f(x\,|\,\theta_1) - kf(x\,|\,\theta_0)]dx$$
$$= \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)].$$

Since ϕ' is a level α test and ϕ is a size α test, then one has

$\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0$; moreover, $k \geq 0$ by assumption, hence

$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)] \leq \beta(\theta_1) - \beta'(\theta_1)$

$\Rightarrow \beta(\theta_1) \geq \beta'(\theta_1) \ \Rightarrow \phi$ has greater power than ϕ'.

Since ϕ' is an arbitrary level α test and θ_1 is the only point in Θ_0^c, then ϕ is a UMP level α test.

WTS II: (b) is true.

Let ϕ' now be the test function for any UMP level α test. By part (a), the test satisfies the assumptions is also a UMP level α test, thus

$\beta(\theta_1) = \beta'(\theta_1)$.

Since $0 \leq \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)]$ and $k > 0$ by assumption, then one has $0 \leq \beta(\theta_1) - \beta'(\theta_1) - k[\alpha - \beta'(\theta_0)]$
$\Rightarrow \alpha - \beta'(\theta_0) = \beta(\theta_0) - \beta'(\theta_0) \leq 0$ since $\beta(\theta_1) = \beta'(\theta_1)$.
Moreover, ϕ' is a level α test, then $\beta'(\theta_0) \leq \alpha$. Thus $\beta'(\theta_0) = \alpha$ since $\beta'(\theta_0) \geq \alpha$ by the above inequality, thus, ϕ' is a size α test, and this further implies that $0 = \beta(\theta_1) - \beta'(\theta_1) - k[\alpha - \beta'(\theta_0)]$
$\Rightarrow \alpha - \beta'(\theta_0) = \beta(\theta_0) - \beta'(\theta_0) = 0 \Leftrightarrow \alpha = \beta'(\theta_0)$.
But the nonnegative integrand $\int [\phi(x) - \phi'(x)] \cdot [f(x\,|\,\theta_1) - kf(x\,|\,\theta_0)]$ will have a zero integral only if ϕ' satisfies the first assumption except on a set A with $\int_A f(x\,|\,\theta_i)dx = 0$.

□

The following corollary connects the **Neyman-Pearson Lemma** to sufficiency.

Corollary 2.1.1:

Under the same settings as in **Theorem 2.1**. Suppose that $T(X)$ is a sufficient statistic for θ and $g(t\,|\,\theta_i)$ is the pdf or pmf of T corresponding to θ_i for $i = 0,1$. Then any test based on T with rejection region S is a UMP level α test if it satisfies

(1) $t \in S$, if $g(t\,|\,\theta_1) > kg(t\,|\,\theta_0)$,
(2) $t \in S^c$, if $g(t\,|\,\theta_1) < kg(t\,|\,\theta_0)$,

for some $k \geq 0$, where $\alpha = \mathbb{P}_{\theta_0}(T \in S)$.

Proof:

In terms of the original sample X the test bound on T has the rejection region $R = \{x\,|\,T(x) \in S\}$. By the **Factorization Theorem**, the PDF or PMF of X can be written as

$$f(x\,|\,\theta_i) = g(T(x)\,|\,\theta_i)h(x),\ i = 0,1,$$

for some nonnegative function $h(x)$. Multiply with the assumptions, one has:

$$\begin{cases} x \in R & , \text{if} f(x\,|\,\theta_1) = g\big(T(x)\,|\,\theta_1\big)h(x) > kg\big(T(x)\,|\,\theta_0\big)h(x) = kf(x\,|\,\theta_0) \\ x \in R^c & , \text{if} f(x\,|\,\theta_1) = g\big(T(x)\,|\,\theta_1\big)h(x) < kg\big(T(x)\,|\,\theta_0\big)h(x) = kf(x\,|\,\theta_0) \end{cases}$$

By the second assumption that $\alpha = \mathbb{P}_{\theta}(T \in S)$, one has

$$\mathbb{P}_{\theta_0}(X \in R) = \mathbb{P}_{\theta_0}(T(X) \in S) = \alpha.$$

Now all the conditions of the first part of **Neyman-Pearson Lemma** are met, it follows that the test based on T is a UMP level α test.

□

Hypotheses, such as H_0 and H_1 in the **Neyman-Pearson Lemma**, that specify only one possible distribution for the sample X are called simple hypotheses. In most realistic problems however, the hypotheses of interest specify more than one possible distribution for the sample. Such hypotheses are called composite hypotheses. Since the definition of UMP requires the test to be most powerful against each individual

$\theta \in \Theta_0^c$, the **Neyman-Pearson Lemma** can be used to find UMP tests in problems involving composite hypotheses.

In particular, hypotheses that assert that a univariate parameter is large, for example, $H : \theta \geq \theta_0$, or small, e.g. $H : \theta < \theta_0$, are called one-sided hypotheses. Hypotheses that assert that a parameter is either large or small, e.g. $H : \theta \neq \theta_0$, are called two-sided hypotheses. A large class of problems that admit UMP level α test involve one-sided hypotheses and pdfs or pmfs with the monotone likelihood raito property, which is given below.

Definition: Monotone Ratio Likelihood Ratio (MLR)

A family of pdfs or pmfs $\{g(t\,|\,\theta)\,|\,\theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a monotone likelihood ratio (MLR) if, for every $\theta_2 > \theta_1$, $g(t\,|\,\theta_2)/g(t\,|\,\theta_1)$ is monotone (nonincreasing or nondecreasing) function of t on $\{t\,|\,g(t\,|\,\theta_1) > 0 \text{ or } g(t\,|\,\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $0 < c$.

Many common families of distributions have an MLR. For example, the normal (known variance, unknown mean), the Poisson, and binomial all have an MLR. Indeed, any regular exponential family with $g(t\,|\,\theta) = h(t)c(\theta)e^{w(\theta)t}$ has an MLR if $w(\theta)$ is a nondecreasing function.

Theorem 2.2: Karlin-Rubin

Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t\,|\,\theta)\,|\,\theta \in \Theta\}$of T has an MLR then for any t_0, the test that rejects $H_0 \Leftrightarrow T > t_0$ is a UMP level α test where $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$.

By an analogous argument, it can be shown that under the conditons of Karlin-Rubin, the test that rejects $H_0 : \theta \geq \theta_0$ in favor of $H_1 : \theta < \theta_0 \Leftrightarrow T < t_0$ is a UMP level α test with $\alpha = \mathbb{P}_{\theta_0}(T < t_0)$.

However, the UMP does not always exist.

Example 2.2: Nonexistence of UMP test

Let $X_1, \ldots, X_n$ be iid $N(0,\sigma^2)$ with σ^2 known. Consider the test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. For a simplified value of α, a level α test in this problem is any test such that $\mathbb{P}_\theta(\text{reject } H_0) \leq \alpha$.

Consider an alternative parameter point $\theta_1 < \theta_0$. Among all tests that satisfy $\mathbb{P}_\theta(\text{reject } H_0) \leq \alpha$, the test that rejects H_0 if $\overline{X} < -\sigma Z_\alpha/\sqrt{n} + \theta_0$ has the highest possible power at θ_1. Call this Test I.

Furthermore, by part (b) of **Neyman-Pearson Lemma**, any other level α test that has as high a power as Test I at θ, must have the same rejection region as Test I except perhaps for a set A such that

$$\int_A f(x\,|\,\theta_i)dx = 0.$$

Thus, if a UMP level α test exists for this problem, it must be Test I because no other test has as high a power as Test I at θ_1.

Alternatively, we may consider a Test II, which rejects H_0 if $\overline{X} > \sigma Z_\alpha/\sqrt{n} + \theta_0$

The Test II is also a level α test. Let $\beta_i(\theta)$ denote the power function of Test I. For any $\theta_2 > \theta_0$, one has

$$\beta_2(\theta_2) = \mathbb{P}_{\theta_2}\Big(\overline{X} > \frac{\sigma Z_\alpha}{\sqrt{n}} + \theta_0\Big) = \mathbb{P}_{\theta_2}\Big(\frac{\overline{X} - \theta_2}{\sigma/\sqrt{n}} > Z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\Big)$$

$$= \mathbb{P}_{\theta_2}(Z > z_\alpha) \text{ Since } Z \sim N(0{,}1), \qquad (\text{“>” since } \theta_0 - \theta_2 < 0)$$

$$= \mathbb{P}(Z < -z_\alpha) \ > \mathbb{P}_{\theta_2}\Big(\frac{\overline{X} - \theta_2}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}\Big)$$

$$= \mathbb{P}_{\theta_2}\Big(\overline{X} < -\frac{\sigma z_\alpha}{\sqrt{n}} + \theta_0\Big) = \beta_1(\theta_2).$$

Thus Test I is not a UMP level α test because Test II has a bigger power than Test I at θ_2. Earlier we showed that if there were a UMP level α test, it would have to be Test I. Therefore, UMP level α test does not exist in this problem. ||

2.3 The p-Values

After a hypothesis test is done, the conclusions must be reported in some statistically meaningful way. One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. The size of the test carrise important information. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing since the test has a large probability of incorrectly making that decision. Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called a p-value.

Definition: p-Value

A p-value $p(X)$ is a test statistic satisfying $0 \le p(x) \le 1$ for every sample point x. Small values of $p(X)$ give evidence that H_1 is true. A p-value is valid if $\forall \theta \in \Theta_0$ and every $0 \le \alpha \le 1$, $\mathbb{P}_\theta(p(X) \le \alpha) \le \alpha$.

If $p(X)$ is valid it is then easy to construct a level α test based on $p(X)$. The test that rejects H_0 if and only if $p(X) \le \alpha$ is a level α test. An advantage to reporting a test result via a p-value is that each reader can choose the α and then can compare the reported $p(x)$ to α and know whether these data lead to acceptance or rejection of H_0. Morover, the smaller the p-value, the stronger the evidence for rejecting H_0. Hence, a p-value reports the results of a test on a more continuous scale, rather than just accepting H_0 or Rejecting H_0.

The most common way to define a valid p-value is given by the following result.

Theorem 4.5: Valid p-Value

Let $W(X)$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point x, define $p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(X) \ge W(x))$. Then, $p(X)$ is valid.

Proof:

Fix $\theta \in \Theta_0$, let $F(\omega)$ be the CDF of $-W(X)$. Define

$$p_\theta(x) := \mathbb{P}_\theta(W(X) \geq w(x)) = \mathbb{P}_\theta(-W(X) \leq -w(x)) = F_\theta(-W(x)).$$

Hence, by the **Probability Integral Transformation**, the distribution of $p_\theta(X)$ is stochastically greater or equal to Uniform(0,1) distribution. That is,

$$\forall 0 \leq \alpha \leq 1, \mathbb{P}_\theta(p_\theta(X) \leq \alpha) \leq \alpha,$$

because

$$p(x) = \sup_{\theta' \in \Theta_0} p_{\theta'}(x) \geq p_\theta(x) \forall x,$$

hence

$$\mathbb{P}_\theta(p(X) \leq \alpha) \leq \mathbb{P}(p_\theta(X) \leq \alpha) \leq \alpha.$$

This is true $\forall \theta \in \Theta_0$ and $\forall 0 \leq \alpha \leq 1$, then by the definition of p-value, $p(X)$ is a valid p-value.

□

3. Methods in Finding Tests

In this section we are going to introduce some methods in finding the hypothesis testing. In 3.1 we shall treat the Likelihood Ratio Test, then in 3.2 we treat the UIT and IUT, i.e. the Union-Intersection Tests and the Intersection-Union Tests. Instead of talking the evaluation of IUT and UIT in 2, we introduce the evaluation of these tests in 3.3.

3.1 The Likelihood Ratio Tests

The likelihood ratio method of hypothesis testing is related to the maximum likelihood estimators and likelihood ratio tests are as widely applicable as maximum likelihood estimation. Recall that if $X_1, \cdots, X_n$ is a random sample from a population with pdf or pmf $f(x|\theta)$ (θ may be a vector), the likelihood function is defined as

$$L(\theta|x_1, \cdots, x_n) = L(\theta|x) = f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

Let Θ denote the entire parameter space. Likelihood ratio tests are defined as follows.

Definition: Likelihood Ratio Test Statistic

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(x) := \frac{\sup_{\Theta_0} L(\theta|x)}{\sup_{\Theta} L(\theta|x)}.$$

Definition: Likelihood Ratio Test (LRT)

A likelihood ratio test (LRT) is any test that has a rejection region of the form $\{x | \lambda(x) \leq c\}$ where c is any constant such that $0 \leq c \leq 1$.

Recall that in the MLE, the maximization of the likelihood function is, not about making the data itself more probable but rather about finding the parameter values that make the observed data most consistent with the assumed model. The motivation for the LRT is quite the same.

Example 3.1: Normal LRT

Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta,1)$ population. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Here θ_0 is a number fixed by the experimenter prior to the experiment. Since there is only one value of θ specified by H_0, the numerator of $\lambda(x)$ is $L(\theta_0 \mid x)$. We know that the unrestricted MLE of θ was found to be $\overline{X}$, the sample mean. Thus, the denominator of $\lambda(x)$ is $L(\bar{x} \mid x)$. So the LRT statistic is:

$$\lambda(x) = \frac{(2\pi)^{-n/2}\exp\left\{-\sum_{i=1}^{n}(x_i-\theta_0)^2/2\right\}}{(2\pi)^{-n/2}\exp\left\{-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\right\}}$$

$$= \exp\left\{-\frac{\left(\sum_{i=1}^{n}(x_i-\theta_0)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2\right)}{2}\right\}.$$

The expression could be simplified by noting that

$$\sum_{i=1}^{n}(x_i-\theta_0)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2.$$

Thus the LRT statistic is

$$\lambda(x) = \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2}\right].$$

An LRT is a test that rejects H_0 for small values of $\lambda(x)$, then the rejection region $\{x \mid \lambda(x) \leq c\}$ can be written as

$$\left\{x \,\middle|\, |\bar{x}-\theta_0| \geq \sqrt{-\frac{2(\log c)}{n}}\right\}.$$

As c ranges between 0 and 1, $\sqrt{\dfrac{-2\log c}{n}}$ ranges between 0 and ∞.

Thus, the LRTs are just those tests that reject $H_0 : \theta = \theta_0$ if the sample mean differ from the hypothesized value θ_0 by more than a specified amount. ||

It coule be best interpreted in the situation in which $f(x \mid \theta)$ is a pmf of a discrete random variable. In this case, the numeraotr is maximized over the whole parameter space Θ while the denominator is maximized over the Θ_0. The less the ratio is shows that more consistent our model is.

Connection with MLEs:

If we think of maximizing over both the entire parameter space and a subset of the parameter space, then the correspondence between the LRTs and MLEs become very clear. Suppose that $\hat{\theta}$, an MLE of θ, exists; $\hat{\theta}$ is obtained by doing an unrestricted maximization of $L(\theta \mid x)$. We can also consider the MLE of θ, call it $\hat{\theta}_0$, obtained by doing the restriced maximization, assuming that Θ_0 is the parameter space. That is, $\hat{\theta}_0 = \hat{\theta}_0(x)$ is the value of $\theta \in \Theta_0$ that maximizes $L(\theta \mid x)$. Then, the LRT statistics is given by $\lambda(x) = \dfrac{L(\hat{\theta}_0 \mid x)}{L(\hat{\theta} \mid x)}$.

For a sufficient statistic of a random sample X, namely $T(X)$, we know that all the information about θ could be found in $T(X)$, the test based on T should be as good as the test based on the complete sample X. In fact, the tests are equivalent.

Theorem 3.1:

If $T(X)$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(x)$ are the LRT statistics based on T and X, respectively. Then $\lambda^*(T(x)) = \lambda(x)\ \forall x \in \Omega_X$.

Proof:

According to the **Factorization Theorem**, the pdf or pmf of X can be written as $f(x\,|\,\theta) = g(T(x)\,|\,\theta)h(x)$, where $g(t\,|\,\theta)$ is the pdf or pmf of T and $h(x)$ does not depend on θ. Thus,

$$\begin{aligned}
\lambda(x) &:= \frac{\sup_{\Theta_0} L(\theta\,|\,x)}{\sup_{\Theta} L(\theta\,|\,x)} = \frac{\sup_{\Theta_0} f(x\,|\,\theta)}{\sup_{\Theta} f(x\,|\,\theta)} \\
&= \frac{\sup_{\Theta_0} g(T(x)\,|\,\theta)h(x)}{\sup_{\Theta} g(T(x)\,|\,\theta)h(x)} && (T \text{ is sufficient}) \\
&= \frac{\sup_{\Theta_0} g(T(x)\,|\,\theta)}{\sup_{\Theta} g(T(x)\,|\,\theta)} && (h \text{ does not depend on } \theta) \\
&= \frac{\sup_{\Theta_0} L^*(\theta\,|\,T(x))}{\sup_{\Theta} L^*(\theta\,|\,T(x))} && (g \text{ is the pdf or pmf of } T) \\
&=: \lambda^*(T(x)).
\end{aligned}$$

□

LRTs are also useful in situations where there are nuisance parameters, i.e., parameters that are present in a model but are not of direct inferential interest. The presence of such nuisance parameters does not affect the construction of the LRT but, as might expected, the presence of nuisance parameters might lead to a different test.

3.2 The UIT and the IUT

In some situations, tests for complicated null hypothesis can be developed from tests for simpler null hypothesis. There are two corresponding methods, the UIT and the IUT, standing for the Union-Intersection test and the Intersection-Union test, respectively. The motivation for these two methods is very straightforward, in practical problems we often see the null hypothesis is expressed under the set operations.

Algorithm 3.2: Union-Intersection Method

The Union-Intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection, namely

$$H_0 : \theta \in \bigcap_{\gamma\in\Gamma} \Theta_\gamma,$$

where Γ is an arbitrary index set.

Suppose that tests are available for each of the problems of testing

$$H_{0\gamma} : \theta \in \Theta_\gamma \text{ versus } H_{1\gamma} : \theta \in \Theta_\gamma^c.$$

Say the rejection region for the test of $H_{0\gamma}$ is $\{x\,|\,T_\gamma(x) \in R_\gamma\}$. Then the rejection region for the Union-Intersection test is

$$\bigcup_{\gamma\in\Gamma} \{x\,|\,T_\gamma(x) \in R_\gamma\}.$$

The rationale is simple. If any one of the hypothesis $H_{0\gamma}$ is rejected then H_0 should be rejected. On the other hand, H_0 is true one if each of the hypothesis

$H_{0\gamma}$ is accepted as true. ||

When θ_0 is defined to be the intersection of some subsets of the parameter space, instead of checking each $\theta_{0\lambda}$ to be true which is the only way for θ_0 to be true, we take the union of each rejection region and proceed with rejecting θ_0 as long as $\theta_{0\lambda}$ is false for some λ, with accepting otherwise. In some situations a simple expression for the rejection region of a Union-Intersection test has a rejection region of the form $\{x \mid T_\gamma(x) > c\}$, where c does not depend on γ. The rejection region for the Union-Intersection test could be expressed as

$$\bigcup_{\gamma\in\Gamma}\left\{x \mid T_\gamma(x) > c\right\} = \left\{x \mid \sup_{\gamma\in\Gamma} T_\gamma(x) > c\right\}.$$

Thus the test statistic for testing H_0 is $T(x) = \sup\limits_{\gamma\in\Gamma} T_\gamma(x)$.

Example 3.2: Normal Union-Intersection Test

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ population. Consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, where μ_0 is a specified number. We can write H_0 as the intersection of two sets: $H_0 : \{\mu \mid \mu \leq \mu_0\} \cap \{\mu \mid \mu \geq \mu_0\}$.

The LRT of $H_{0L} : \mu \leq \mu_0$ versus $H_{1L} : \mu > \mu_0$ is rejecting $H_{0L} : \mu \leq \mu_0$ in favor of $H_{1L} : \mu > \mu_0$ if $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq t_L$. Similarly, the LRT of $H_{0L} : \mu \geq \mu_0$ versus $H_{1\mu} : \mu < \mu_0$ is rejecting $H_{0U} : \mu \geq \mu_0$ in favor of $H_{1U} : \mu < \mu_0$ if $\dfrac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq t_U$. Thus, the Union-Intersection test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ formed from these two LRTs is

$$\text{Reject } H_0 \text{ if } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq t_L \text{ or } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq t_U.$$

If $t_L = -t_U \geq 0$, the Union-Intersection test can be more simply expressed as the form:

$$\text{Reject } H_0 \text{ if } \frac{|\overline{X} - \mu_0|}{S/\sqrt{n}} \geq t_L.$$

It turns out that this Union-Intersection test is also the LRT for this problem and is called the two-sided *t* test. ||

The analogous Intersection-Union method is formulated in a similar way.

Algorithm 3.3: Intersection-Union Method

Suppose we wish to test the null hypothesis $H_0 : \theta \in \bigcup\limits_{\gamma\in\Gamma} \Theta_\gamma$. Suppose that

$$\forall \gamma \in \Gamma, \{x \mid T_\gamma(x) \in R_\gamma\}$$

is the rejection region for a test of

$$H_{0\gamma} : \theta \in \Theta_\gamma \text{ versus } H_{1\gamma} : \theta \in \Theta_\gamma^c.$$

Then the rejection region for the Intersection-Union test of H_0 versus H_1 is

$$\bigcap_{\gamma\in\Gamma}\{x\,|\,T_\gamma(x)\in R_\gamma\}.$$

H_0 is false $\Leftrightarrow$ all of the $H_{0\gamma}$ is false, so H_0 can be rejected $\Leftrightarrow$ each of the individual hypothesis $H_{0\gamma}$ can be rejected. ||

Again, the test can be greatly simplified if the rejection region for the individual hypothesis are all of the form $\{x\,|\,T_\gamma(x)\geq c\}$, where c is independent of γ. In such cases, the rejection region of H_0 is

$$\bigcap_{\gamma\in\Gamma}\{x\,|\,T_\gamma(x)\geq c\}=\{x\,|\,\inf_{\gamma\in\Gamma}T_\gamma(x)\geq c\}.$$

Here, the Intersection-Union test statistic is $\inf_{\gamma\in\Gamma}T_\gamma(x)$, and the test rejects H_0 for large values of this statistic.

3.3 Evaluation on UIT and IUT

Because of the simple way in which they are constructed, the sizes of the UIT and the IUT can often bebounded above by the sizes of some other tests. Such bounds are useful if a level α test is wanted, but the size of the UIT or IUT is too difficult to evaluate. In this subsection we discuss the bounds and give examples in which the bounds are sharp, i.e. the size of the test is equal to the bound.

First consider UITs. Recall that, in this situation, we are testing a null hypothesis of the form $H_0:\theta\in\Theta_0$ where $\Theta_0=\bigcap_{\gamma\in\Gamma}\Theta_\Gamma$. To be specific, let $\lambda_\gamma(x)$ be the LRT statistic for testing $H_{0\gamma}:\theta\in\Theta_\gamma$ versus $H_{1\gamma}:\theta\in\Theta_\gamma^c$, and let $\lambda(x)$ be the LRT statistic for testing $H_0:\theta\in\Theta_0$ versus $H_1:\theta\in\Theta_0^c$. Then we have the following relationships between the overall LRT and the UIT based on $\lambda_\gamma(x)$.

Theorem 3.4:

Consider testing $H_0:\theta\in\Theta_0$ versus $H_1:\theta\in\Theta_0^c$ where $\Theta_0:=\bigcap_{\gamma\in\Gamma}\Theta_\gamma$ and $\lambda_\gamma(x)$ is defined as above. Define $T(x):=\inf_{\gamma\in\Gamma}\lambda_\gamma(x)$, and form the UIT with rejection region $\{x\,|\,\lambda_\gamma(x)<c \text{ for some } \gamma\in\Gamma\}=\{x\,|\,T(x)<c\}$. Also consider the usual LRT with rejection region $\{x\,|\,\lambda(x)<c\}$. Then

(a) $T(x)\geq\lambda(x)$ for all x.

(b) If $\beta_T(x)$ and $\beta_\lambda(x)$ are the power functions for the tests based on T and λ, respectively, then $\beta_T(\theta)\leq\beta_\lambda(\theta)$ for every $\theta\in\Theta$.

(c) If the LRT is a level α test, then the UIT is a level α test.

Proof:

Since $\Theta_0=\bigcap_{\gamma\in\Gamma}\Theta_\gamma\subseteq\Theta_\gamma$, then by the definition of LRT, one has $\lambda_\gamma(x)\geq\lambda(x)$ $\forall x\,\forall\gamma\in\Gamma$.

Because the region of maximization is bigger for the individual λ, then

$T(x) := \inf_{\gamma \in \Gamma} \lambda_\gamma(x) \geq \lambda(x)$, then (a) follows.

By (a), $\{x \mid T(x) < c\} \subseteq \{x \mid \lambda(x) < c\}$, therefore one has $\beta_T(\theta) := \mathbb{P}_\theta(T(X) < c) \leq \mathbb{P}_\theta(\lambda(X) < c) =: \beta_\lambda(\theta)\ \forall \theta \in \Theta$ then (b) follows.

Since (b) holds $\forall \theta \in \Theta$, therefore, $\sup_{\theta \in \Theta} \beta_T(\theta) \leq \sup_{\theta \in \Theta} \beta_\lambda(\theta) \leq \alpha$ by assumption, therefore (c) holds.

□

Since the LRT is uniformly more powerful in the above theorem than UIT, we might ask why we should use the UIT. One reason is that UIT has a smaller Type I Error probability for every $\theta \in \Theta_0$. Moreover, if H_0 is rejected, we may wish to look at the individual tests of $H_{0\gamma}$ to see why, for which UIT provides us an access.

We now investigate the sizes of IUTs. A simple bound for the size of an IUT is related to the sizes of the individual tests that are used to define the IUT. Recall that in this situation the null hypothesis is expressible as a union, i.e. we are testing

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_0^c, \text{ where } \Theta_0 = \bigcup_{\gamma \in \Gamma} \Theta_\gamma.$$

An IUT has a rejection region of the form $R = \bigcap_{\gamma \in \Gamma} R_\gamma$ where R_γ is the rejection region for a test of $H_{0\gamma} : \theta \in \Theta_\gamma$.

Theorem 3.5:

Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region R_γ. Then the IUT with rejection region $R = \bigcap_{\gamma \in \Gamma} R_\gamma$ is a level $\alpha = \sup_{\gamma \in \Gamma} \alpha_\gamma$ test.

Proof:

Let $\theta \in \Theta_0$. Then $\theta \in \Theta_\gamma$ for some $\gamma \in \Gamma$ and one has $\mathbb{P}_\theta(X \in R) \leq \mathbb{P}_\theta(X \in R_\gamma) \leq \alpha_\gamma \leq \alpha$. Since $\alpha := \sup_{\gamma \in \Gamma} \alpha_\gamma$. Since $\theta \in \Theta_0$ was chosen arbitrarily, then the IUT is a level α test.

□

Typically, the individual rejection regions R_γ are chosen so that $\alpha_\gamma = \alpha\, \forall \gamma$. In such a case, **Theorem 3.5** states that the resulting IUT is a level α test. Moreover, this theorem provides an upper bound for the size of an IUT, is somewhat more useful than **Theorem 3.4**, which provides an upper bound for the size of a UIT.

Remark:

Theorem 3.4 applied only to UITs constructed from LRTs while **Theorem 3.5** applies to any IUT. ||

The bound in **Theorem 3.4** is the size of the LRT, which, in a complicated problem, may be difficult to compute. In **Theorem 3.5** however, the LRT need not be used to obtain the upper bound. Any test $H_{0\gamma}$ with unknown size α_γ can be used, and then the upper bound on the size of the IUT is given in terms of the known sizes $\alpha_\gamma, \gamma \in \Gamma$.

The IUT in **Theorem 3.5** is a level α test. But the size of the IUT may be much less than α; the IUT may be very **conservative**. The following theorem gives conditions under which the size of the IUT is exactly α and the IUT is not conservative in this case.

Theorem 3.6:

Consider testing $H_0 : \theta \in \bigcup_{j=1}^{n} \theta_j$, where k is a finite possible integer. For each $j = 1,...,k$, let R_j be the corresponding rejection region of a level α test of H_{0j}. Suppose that for some $i = 1,...,k$, there exists a sequence of parameter points, $\theta_l \in \Theta_i$, for $l = 1,2,...$, such that:

(i) $\lim_{l\to\infty} \mathbb{P}_{\theta_l}(X \in R_i) = \alpha$.

(ii) $\forall j \neq i, \lim_{l\to\infty} \mathbb{P}_{\theta_l}(X \in R_j) = 1$.

Then, the IUT with rejection region $R = \bigcap_{j=1}^{k} R_j$ is a size α test.

Proof:

To show that the IUT with rejection region $R = \bigcap_{j=1}^{k} R_j$ is a size α test is to show that $\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X \in R) = \alpha$.

"$\leq$":

By **Theorem 3.5** and **Bonferroni's Inequality**, R is a level α test, i.e., $\sup_{\theta\in\Theta} \mathbb{P}_\theta(X \in R) \leq \alpha$.

"$\geq$":

Because all the parameter points θ_l satisfy $\theta_l \in \Theta_i \subseteq \Theta_0$, therefore, one has

$$\begin{aligned}\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(X \in R) &\geq \lim_{l\to\infty} \mathbb{P}_{\theta_l}(X \in R)\\ &= \lim_{l\to\infty} \mathbb{P}_{\theta_l}(X \in \bigcap_{j=1}^{k} R_j)\\ &\geq \lim_{l\to\infty} \sum_{j=1}^{k} \mathbb{P}_{\theta_l}(X \in R_j) + (1-k)\\ &= (k-1) + \alpha - (k-1) \text{ by (i) and (ii)}\\ &= \alpha.\end{aligned}$$

□

Reference:

[1]: George Casella, Roger L. Berger, *Statistical Inference, Second Edition*, Duxbury Thomson Learning.

[2]: Fan Yang, *Lecture Notes on Statistical Theory,* Lecture Given to Tsinghua University at Fall 2022, available online.

[3]: Tianyu Zhang, *Lecture Notes on Probability on Banach spaces*, unpublished version, available online.

[4]: Xiangdong Li, *Lecture Notes on Optimal Transportation Problems*, Lecture Given to Tsinghua University at Spring 2023, available online.

[5]: Shu Cherng Fang, Sarat Puthenpura, *Linear Optimization and Extensions: Theory and Algorithms*.

[6]: Gordan Zitkovic, *Theory of Probability I — Weak Convergence*, available online.

Appendix:

Theorem A: Probability Integral Transformation

Let U follow a uniform distribution and if F^{-1} is the quantil function of X then $F^{-1}(U)$ and X has the same distribution.

Theorem B: Factorization Theorem

Let $f(x\,|\,\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient statistic for $\theta \Leftrightarrow$ there exist functions $g(t\,|\,\theta)$ and $h(x)$ such that, for all sample points x and all parameter points $\theta, f(x\,|\,\theta) = g(T(x)\,|\,\theta)h(x)$.

Appendix C: Bonferroni's Inequalities

Let $A_1, \ldots, A_n$ be events in a probability space $(\Omega, \sum, \mathbb{P})$ and let $A := \bigcup_{i=1}^{n} A_i$.

Then one has $\mathbb{P}(\bigcap_{i=1}^{n} A_i) \geq (1-n) + \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Proof:

For $n = 1$, $\mathbb{P}(A_1) \geq (1-1) + \mathbb{P}(A_1)$ always holds. Without loss of generality, we may assume that $k \geq 1$ and the inequality holds for k, i.e.,

$$\mathbb{P}(\bigcap_{i=1}^{k} A_i) \geq (1-k) + \sum_{i=1}^{k} \mathbb{P}(A_i).$$

WTS: $\mathbb{P}(\bigcap_{i=1}^{k+1} A_i) \geq (1-k) + \sum_{i=1}^{k+1} \mathbb{P}(A_i)$

$$\begin{aligned}
LHS &= \mathbb{P}\left[\left(\bigcap_{i=1}^{k} A_i\right) \cap A_{k+1}\right] \\
&= \mathbb{P}(\bigcap_{i=1}^{k} A_i) + \mathbb{P}(A_{k+1}) - \mathbb{P}\left[\left(\bigcap_{i=1}^{k} A_i\right) \cup A_{k+1}\right] \\
&\geq \sum_{i=1}^{k} \mathbb{P}(A_i) + \mathbb{P}(A_{k+1}) - \mathbb{P}\left[\left(\bigcap_{i=1}^{k} A_i\right) \cup A_{k+1}\right] + (1-k) \\
&\qquad\qquad \text{(By assumption)} \\
&= \sum_{i=1}^{k+1} \mathbb{P}(A_i) - \mathbb{P}\left[\left(\bigcap_{i=1}^{k} A_i\right) \cup A_{k+1}\right] + (1-k).
\end{aligned}$$

By definition, $\mathbb{P}\left[\left(\bigcap_{i=1}^{k} A_i\right) \cup A_{k+1}\right] \leq 1$, therefore,

$$\geq \sum_{i=1}^{k+1} \mathbb{P}(A_i) - 1 + 1 - k = \sum_{i=1}^{k+1} \mathbb{P}(A_i) - k.$$

Result follows by induction.

□

On The Bayesian Statistical Approach

Tianyu Zhang[4]

Abstract:

We introduced the Bayesian statistical appraoch as a complementary to the long-frequentist approach. The three major ingredients of Bayesian analysis, the data, the prior, and the loss functions, except the first one which is assumed to be well-behaved, are discussed in this review. Before having the data the experimenters based on their beliefs offer a prior distribution, the good priors should meet some standards we discuss in the second section. After having the prior, we need to know how much the deviation between esimation and realized values is, the derived loss function, along with some often-used forms are disucssed in the third section. The evaluation, hypothesis testing, and the interval estimation are treated in the fourth section. We discuss also the validity of Bayesian and we introduce the Bayes convolution to close this review.

Table of Contents:

1. Introduction to Bayesian

We start with a short review of the long-run frequentists in 1.1, then we introduce the Bayesian approach in 1.2. Regarding the data collection is well performed, three

[4] YMSC, BIMSA, bidenbaka@gmail.com, obamalgb@cantab.net

ingredients of the Bayesian statistics: data, prior, and loss function are introduced in this section. In 1.3 we introduce the prior and posterior, and we close this section with the loss function introduced in 1.4.

1.1 Long-Run Frequensits

Practically, in all statistics courses, we learn how to make decisions under uncertaity. Formally, we are looking for a decision δ that belongs to an action space $\mathscr{A}$ — a set of all possible decisions that we are allowed to take.

We also know that statisticians collect random samples of data and do their statistics based on them. So, their decisions are functions of data, namely,

$$\delta = \delta(\text{data}) = \delta(X_1, \cdots, X_n). \tag{1.1}$$

This is the frequentist approach. According to it, uncertainty comes from a random sample and its distribution. The only considered distributions, expectations, and variances are distributions, expectations, and variances of data and various statistics computed from data. Population parameters are considered fixed. Statistical procedures are based on the distribution of data given these parameters,

$$f(x \,|\, \theta) = f(X_1, \cdots, X_n \,|\, \theta). \tag{1.2}$$

Properties of these procedures can be stated in terms of long-run frenquencies. For example:

Example 1.1: Long-Run Frenquencies

(i) An estimator $\hat{\theta}$ is unbiased if in a long run of random samples, it averages to the parameter θ.

(ii) A test has significance level α if in a long run of random samples, 100 % of times the true hypothesis is rejected.

(iii) An interval has confidence level $(1-\alpha)$ if in a long run of random samples, $(1-\alpha)$ 100% of obtained confidence intervals contain the parameter. ||

However, there are many situations when using only the data is not sufficient for reasonable decisions. Also, the frequentist concept of a long run may inadequately reflect performance of statistical procedures.

Summary:

Frequentist statistical decision making takes into account only the uncertainty of the data. Statistical decisions are based on the data only, and their performance is evaluated in terms of a "long-run". However, there are situations where such an approach is deficient, unnatural, or even misleading in various ways.

1.2 Bayesian Approach

Different from the long-frequentist approach, there is another method, the famous Bayesian approach. According to this perspective of view, uncertainty is attributed not only to the data but also to the unknown parameter θ. Some values of θ are more likely than others. Then, as long as we talk about the likelihood, we can define a whole distribution of values of θ. We call it the prior distribuion, and it reflects our ideas, beliefs, and past experiences about the parameter before we collect and use the data.

One benefit of this approach is that we no longer have to explain our results in terms of a “long-run”. Often we collect just one sample for our analysis and don’t experience any long run of samples. Instead, with the Bayesian approach, we can state the result in terms of the distribution of parameter θ. For example, we can clearly state the probability for a parameter to belong to a certain interval, or the probability that the hypothesis is true. This would have been impossible under the frequentist approach.

Another benefit is that we can use both pieces of information, the data and the prior, to make better decisions. In Bayesian statistics, decisions are

$$\delta = \delta(\text{data, distribution}). \tag{1.3}$$

1.3 Prior and Posterior

Now we have two sources of information to use in our Bayesian inference:

(i) collected and observed data;

(ii) prior distribution of the parameter.

These two pieces are combined via the Bayes formula

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}. \tag{1.4}$$

Prior to the experiment, our knowledge about the parameter θ is expressed in terms of the prior distribution (prior pdf or pmf) $\pi(\theta)$. The observed sample of data $X = (X_1, \cdots, X_n)$ has distribution (pmf or pdf)

$$f(x|\theta) = f(x_1, \cdots, x_n|\theta). \tag{1.5}$$

This distribution is conditional on θ. That is, different values of the parameter θ generate different distributions of data, and thus, conditional probabilities about X generally depend on the condition θ.

Observed data add information about the parameter. The updated knowledge about θ can be expressed as the posterior disribution, namely,

$$\pi(\theta|x) = \pi(\theta|X = x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}, \tag{1.6}$$

where $m(x)$ represents the unconditional distribution of data X. This is the marginal distribution (pdf or pmf) of the sample X. Being unconditional means that it is constant for different values of the parameter θ. It can be computed by

$$m(x) = \begin{cases} \sum_\theta f(x|\theta)\pi(\theta), & \text{for discrete piror distirbutions } \pi \\ \int_\theta f(x|\theta)\pi(\theta)d\theta, & \text{for continuous prior distributions } \pi \end{cases}.$$

Note that the posterior distribution of the parameter θ is now conditioned on data $X = x$. Naturally, conditional distributions $f(x|\theta)$ and $\pi(\theta|x)$ are related via the Bayes rule (1.4).

Notation:

$\pi(\theta)$	=	prior distribution
$\pi(\theta|x)$	=	posterior distribution
$f(x|\theta)$	=	distribution of data (model)

$m(x)$ = marginal distribution of the data

X = $(X_1, \cdots, X_n)$, sample of data

x = $(x_1, \cdots, x_n)$, observed values of $X_1, \cdots, X_n$.

Summary:

Bayesian approach presumes a prior distribution of the unknown parameter. Adding the observed data, the Bayes Theorem converts the prior distirbution into the posterior which summarizes all we know about the parameter after seeing the data. Bayesian decisions are based on this posterior, and thus, they utilize both the data and the prior.

1.4 Loss Function

Besides the data and the prior distribution, is there any other information that can appear useful in our decision making?

How about anticipating possible consequences of making an error? We know that under uncertainty, there is always a chance of making inaccurate decisions. The third component of the Bayesian Decision Theory is the loss function, defined as below.

Definition: Loss Function

The loss function $L : \Theta \times \mathscr{A} \to \mathbb{R}$, or sometimes $L : \Theta \times \mathscr{A} \to [0,\infty)$ which is a map $(\theta, \delta) \mapsto c \in \mathbb{R}$ (resp. $c \in [0,\infty)$.

This penalty may be 0 if the decision is perfect, for example, if we accept the true null hypothesis or estimate a parameter θ with no error.

Equipped with the loss function, we are looking for optimal statistical decisions. Those that minimize the loss. But minimize with respect to what? The loss function $L(\theta, \delta) = L(\theta, \delta(X))$ has uncertainty — unknown parameter θ and $X = (X_1, \cdots, X_n)$.

Definition: Risk

The risk, or frequentist risk is the expected loss over all possible samples, given a parameter θ. It is defined by $R(\theta, \delta) := \mathbb{E}_\theta^X L(\theta, \delta(X))$, there $\mathbb{E}_\theta^X$ means the expectation depends both on X and θ. Note that

$$\mathbb{E}_\theta^X L(\theta, \delta(X)) = \sum_X L(\theta, \delta(x))\mathbb{P}(x) \text{ or } \int_{-\infty}^{\infty} L(\theta, \delta(x))f(x)dx.$$

With respect to the risk, the optimal decisions are still not clear since $R(\theta, \delta(X))$ depends on the unknown parameter θ. However, it is clear which rules we should not use. Moreover, by this convention, we can tell which action is better than another by just comparing the corresponding risks, this leads to a natural comparability.

Definition: R-better

We say decision δ_1 is R-better than decision δ_2 if either

(i) $R(\theta, \delta_1) \leq R(\theta, \delta_2) \forall \theta$, or

(ii) $R(\theta, \delta_1) < R(\theta, \delta_2)$ for some θ.

With this partial odering, we can further deduce for what decisions are acceptable and for what decisions are not.

Definition: Inadmissible, Admissible

Decision δ is inadmissible if there exists a decision R-bettern than δ.

Alternatively, decision δ is admissible if it is not inadmissible.

It turns our that we can minimize the bad influence brought up by the worst case, i.e. we minimize the maximum of the risk, this leads to a natural application of the minimax theory.

Definition: Minimax Decision

Decision δ is said to be minimax if it minimizes $\inf_{\theta} R(\theta, \delta)$, the worst possible risk over all $\theta \in \Theta$. That is, $\sup_{\theta} R(\theta, \delta_{\text{minimax}}) = \inf_{\delta \in \mathscr{A}} \sup_{\theta} R(\theta, \delta)$.

Note that minimax decisions are **conservative** because they protect against the worst situation where the risk is maximized. They are the best decisions in a game against an intelligent opponent who will always like to give you the worst case. In statistical games, the players know that they are acting against intelligent opponents, and therefore, they devise minimax strategies. This is one of the interest in game theory.

So far we have introduced the loss and risk without letting the prior distribution being involved. Now let us define when the case it is not excluded.

Definition: Bayes Risk

The Bayes risk is the expected frequentist risk

$$r(\pi, \theta) = \mathbb{E}^{\pi(\theta)} R(\theta, \delta) = \mathbb{E}_{\theta}^{X} L(\theta, \delta),$$

where the expectation is taken over the prior distribution $\pi(\theta)$. So it is the loss function averaged over all possible samples of data and all possible parameters.

As we have already seen, the Bayes decisions are based on the posterior distribution, that is, conditioned on the known data X.

Definition: Posterior Risk

The posterior risl is the expected loss, where the expectation is taken over the posterior distribution of parameter θ,

$$\rho(\pi, \delta \,|\, X) = \mathbb{E}_{X}^{\pi(\theta|X)} L(\theta, \delta) = \mathbb{E}\big(L(\theta, \delta) \,|\, X\big).$$

So, the posterior risk is the loss function averaged over parameters θ, given known data X.

Definition: Bayes Decision Rules

The Bayes decision rules minimize the Bayes risk and, as we'll see pretty soon, they also minimize the posterior risk. That is,

$$\rho(\pi, \delta_{\text{Bayes}} \,|\, X) = \inf_{\delta \in \mathscr{A}} \rho(\pi, \delta \,|\, X)$$

for every sample X and $r(\pi, \theta) = \inf_{\delta \in \mathscr{A}} r(\pi, \delta)$.

Summary:

Simple frequentist statistics are based on just the observed data. Decision theory takes into account the data and the loss function. Bayesian statistics is based on the data and the prior distribution. Thus, the Bayes decision rules are based on three components:

(i) The data,
(ii) The Prior disribution,
(iii) The Loss.

Bayes rules minimze the posterior risk, given the observed data. Minimax rules minimize the largest or the worst possible risk.

2. Choice of a Prior Distribution

Recall that Bayes decision rules are based on three components, the data, the prior distrubtions, and the loss. That is, $\delta_{\text{Bayes}} = \delta(X, \pi, L)$. So, for Bayesian decision making, we need

(i) To collect data $X = (X_1, \cdots, X_n)$
(ii) To choose a prior distribution of unknown parameters $\pi(\theta)$.
(iii) To choose a loss function $L(\theta, \delta)$.

We mainly focus on the introduction to (ii). There are perhaps four general ways to choose a prior distribution:

(1) Quantify your personal beliefs, express your uncertainty about the parameter θ in a form of a distribution. (Subjectively)
(2) Let the data suggest the prior distribution. People often use historical data or data on similar cases. (Empirically)
(3) Take a convenient form of the prior $\pi(\theta)$ in order to get a mathematically tractable posterior distribution $\pi(\theta \mid X)$. (Conveniently)
(4) In the absence of any information about the parameter prior to the experiment, which prior distribution would most fairly reflect this situation? (Non-Informatively)

We offer a short treatment of (1) in 2.1, then we move to the discussion of the empirical Bayes solutions in 2.2, where the parametric, non-parametric, and the hierarchy (i.e. we use Bayesian statistical approach to the prior) Bayesian are discussed. We introduce the important terminology “conjugate family” in 2.3, where we offer some conjugate relationships between familiar distribution families. We give a brief introduction to the non-informative Bayes in 2.4, we also generalize the Bayes rules in this subsection.

2.1 Subjective Choice of a Prior Distribution

Subjectively determined prior does not have a direct mathematical formula. It is just an attemp to express one’s original beliefs about the unknown parameter and one’s uncertainty about it in a usable mathematical form.

Often we can determine a few related probabilities and fit a distribution of them. Sometimes we can compare probabilities of different values of θ or probabilities of intervals. Usually, it is easy to compare the chances of events and their complements like $\mathbb{P}(\theta \in [a, b])$ and $\mathbb{P}(\theta \notin [a, b])$ — which one is more likely? Sometimes one can determine some percentiles, say, with probability 25%, parameter θ does not exceed what…

2.2 Empirical Bayes Solutions

The general idea of empirical Bayes analysis is to estimate the prior distribution from the data. This can be done in several ways.

- **Parametric Empirical Bayes**. A family of prior distributions $\pi(\theta \mid \lambda)$ is chosen, but its parameter(s) λ is unknown. This λ will then be estimated from the marginal distribution $m(x) = m(x \mid \lambda)$.

- **Nonparametric Empirical Bayes**. No family of prior distribution is assumed. Thus, there is no form of the posterior as well. Instead, the form of a Bayes decision rule is obtained directly, bypassing the posterior.
- **Hierarchical Bayes**. We can also take a "completely Bayes" approach, assume a family of prior distributions $\pi(\theta \mid \lambda)$ and estimate unknown λ in the Bayesian way. That is, we put a prior distribution $\rho(\lambda)$ on λ and estimate it, before estimating the parameter of interest θ. This second-level prior is called a hyperprior, and parameter λ is a hyperparameter. Sometimes this hierarchy of priors and their parameters has more than two levels (but countably many…)

2.3 Conjugate Priors

Let us focus on the mathematically convenient families of prior distribution. A suitably chosen prior distribution of θ may lead to a very tractable form of the posterior.

Definition: Conjugate

A family of prior distributions π is conjugate to the model $f(x \mid \theta)$ if the posterior distribution belongs to the same family.

Recall in our lecture notes on statistical inference, we have the concept that

Definition: Conjugate Family

Let $\mathscr{F}$ denote the class of pdfs or pmfs $f(x \mid \theta)$ indexed by θ. A class Π of prior distributions is a conjugate family for $\mathscr{F}$ if the posterior distribution is in the class Π $\forall f \in \mathscr{F}$, all priors in Π, and all $x \in X$.

We state some conjugate relationships without proving.

- **Gamma family is conjugate to the Poisson model**.
  Having observed a Poisson sample $X = x$, we update the Gamma(α, λ) prior distribution of θ to the Gamma$(\alpha + \Sigma x_i, \lambda + n)$ posterior.
- **Beta family is conjugate to the Binomial model**.
  Posterior parameters are $\alpha_x = \alpha + \Sigma x_i$ and $\beta_x = \beta + nk - \Sigma x_i$.
- **Normal family is conjugate to the Normal model**.
  Posterior parameters are $\mu_x = \dfrac{n\overline{X}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}$ and $\tau_x^2 = \dfrac{1}{n/\sigma^2 + 1/\tau^2}$.

2.4 Non-Informative Prior Distributions and Generalized Bayes Rules

One of the main arguments of the non-Bayesians against Bayesians was the subjective choice of the prior distribution. Indeed, it is not always trivial to come up with a realistic distribution that truly reflects our prior knowledge and uncertainty about the unknown parameter.

An extreme case, what can we do if we have no prior knowledge whatsover? No information about the parameters until we see the actual data… However, we still would like to use Bayesian methods because of its nice properties. It is natural to ask the question that is there a "fair" prior distribution that reflects our absence of the knowledge. Such a distribution would be called a non-informative prior distribution.

However, this approach has two possible problems, the first is that there are transformations not "preserving" the distributions. That is,

Example 2.1: Bad Tranformations

Consider estimation of parameter θ of Binomial(k, θ) distribution. We know that $\theta \in [0,1]$, and suppose that nothing els is known about θ. Then, we should choose a prior distribution that gives equal weights to all values of θ, making them all "equally likely". So let $\pi(\theta) \sim$ Uniform(0,1)?

Indeed this seems to be the most natural non-informative choice. However, any non-linear transformation of θ, its reparametrization, appears non-Uniform. ||

Another problem is that, since high prior variance means a lot of uncertainty about the unknown parameter; if we consider the extreme case under normal distribution, when Var$\theta =: \tau^2 \to \infty$, we are infinitely uncertain about θ. Moreover, as $\tau^2 \to \infty$, the Normal(μ, τ) prior density becomes more and more flat, converging to a constant.

This leads to a serious problem! There is no constant density on $\mathbb{R}$. There is a constant measure, the Lebesgue measure, with $\pi(\theta) \equiv 1 \,\forall \theta \in \mathbb{R}$ but

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta = \int_{-\infty}^{\infty} d\theta = \infty,$$

hence it is not a probability measure.

Nevertheless, Lebesgue measure gives us a fine posterior distribution

$$\pi(\theta|x) \sim f(x|\theta)\pi(\theta) \sim \exp\left\{ \left(\theta\overline{X} - \frac{\theta^2}{2}\right)\frac{n}{\sigma^2} \right\} \sim \exp\left\{ -\frac{(\theta - \overline{X})^2}{2\sigma^2/n} \right\} \sim \mathrm{N}(\overline{X}, \sigma^2/n).$$

Therefore, without any prior information, after seeing the data, we have exactly as much as uncertainty about θ as the data contain. This is quite reasonable. So, such a prior worth consideration, even though it is not, strictly speaking, a distribution.

Definition: Improper Prior Distribution

An imporper prior distribution is an infinite measure on the parameter space Θ (i.e. $\int_{\Theta} d\pi(\theta) = \infty$) which produces a proper posterior distribution.

Definition: Generalized Bayes Rule

Decision that minimizes the posterior risk under an improper prior is called a generalized Bayes rule. Generalized Bayes rules are limits of proper Bayes rules.

3. Standard Loss Functions and Corresponding Bayes Decision Rules

In this section we shall introduce some common loss functions, they are squared-error loss in 3.1, absolute-error loss in 3.2, and the zero-one loss in 3.3. Note that one deci-sion remains better than another in Bayes or minimax sense if the entire loss function is increased by a constant or multiplied by a positive coefficient. Therefore,

- We can drop constant coefficients and shifts.
- We can only consider the non-negative losses.

One interesting finding is that, the posterior mean is the Bayes decision with respect to the squared-error loss, the posterior median is the Bayes decision with respect to the absolute-error loss, and the posterior mode is the Bayes decision with respect to the zero-one loss. We start with the first one.

3.1 Squared-Error Loss

Definition: Squared-Error Loss Function

The squared-error loss function is defined to be $L(\theta, \delta) := (\theta - \delta)^2$, where θ is the parameter and δ is its estimator.

The corresponding squared-error posterior risk is given by

$$\rho(\pi, \delta|X) = \mathbb{E}^{\theta}\left((\theta-\delta)^2\middle|X\right) = \mathbb{E}^{\theta}\left((\theta-\mu_X)^2\middle|X\right) + \left(\mu_X - \delta(X)\right)^2,$$

which can be interpreted as the posterior variance of θ plus posterior bias squared. Here $\mu_X := \mathbb{E}(\theta|X)$ is the posterior mean of θ.7

The Bayes decision with respect to the squared-error loss is the one that minimizes the posterior risk which we shall write as

$$\rho(\pi, \delta|X) = \mathbb{E}^{\theta}\left((\theta-\delta)^2\middle|X\right) = \delta^2 - 2\delta\mathbb{E}(\theta|X) + \mathbb{E}(\theta^2|X).$$

Therefore, the minimum of the Bayes risk is attained at

$$\delta_{\text{Bayes}} = \frac{-2\mathbb{E}(\theta|X)}{2} = \mathbb{E}(\theta|X) = \mu_X.$$

Summary:

The posterior mean of θ is the Bayes decision with respect to the squared-error loss.

3.2 Absolute-Error Loss

Definition: Absolute-Error Loss Function

The absolute-error loss function is defined to be $L(\theta, \delta) := |\theta - \delta|$, where θ is the parameter and δ is its estimator.

Unlike the squared-error loss function, the absolute-error loss function does not penalize as much for large deviations of the estimator δ from the parameter θ. We now state and prove our main result in this subsection.

Theorem 3.1:

The posterior median is the Bayes decision with respect to the absolute-error loss.

Proof:

Consider M, the median of $\pi(\theta|X)$, and δ, some decision, and compare their losses.

Case I: $\delta < M$.

In this case, the difference of losses is given by

$$\begin{aligned} L(\theta,\delta) - L(\theta, M) &= |\theta - \delta| - |\theta - M| \\ &= \begin{cases} M - \delta, & \text{if } \theta \geq M \\ \text{a linear continuous function}, & \text{if } \delta \leq \theta \leq M. \\ -(M-\delta), & \text{if } \theta \leq \delta \end{cases} \end{aligned}$$

Taking expected values with respect to the posterior distribution $\pi(\theta|X)$, we obtain the difference of the posterior risks,

$$\begin{aligned}\rho(\pi,\delta\,|\,X)-\rho(\pi,M\,|\,X) &\geq -(M-\delta)\mathbb{P}(\theta<M\,|\,X)+(M-\delta)\mathbb{P}(\theta\geq M\,|\,X)\\ &=(M-\delta)\big(\mathbb{P}(\theta\geq M\,|\,X)-\mathbb{P}(\theta<M\,|\,X)\big).\end{aligned}$$

Since M is defined to be the median of $\pi(\theta\,|\,X)$, one has

$$\mathbb{P}(\theta\geq M\,|\,X)\geq\mathbb{P}(\theta<M).$$

It follows that

$$\rho(\pi,\delta\,|\,X)-\rho(\pi,M\,|\,X)\geq 0.$$

Case II: $\delta\geq M$

Analogous to *Case I*.

□

Summary:

The posterior median is the Bayes decision with respect to the absolute-error loss.

3.3 Zero-One Loss

Generally, the zero-one loss function gives a penalty of 1 for any wrong decision and no penalty for the correct decision. This makes sense in hypothesis testing — Type I and Type II errors are penalized while correct acceptance and correct rejections are not.

In estimation, a zero-one loss can be defined as,

Definition: Zero-One Loss Function

The zero-one loss function is defined to be

$$L(\theta,\delta)=I(\theta\neq\delta)=\begin{cases}1, \text{if}\,\theta\neq\delta\\ 0, \text{if}\,\theta=\delta\end{cases}.$$

However, this only makes sense in discrete cases, when $\theta=\delta$ with a non-zero probability. Then, in considering the Bayes decision under this loss function, we compute the posterior risk,

$$\rho(\pi,\delta\,|\,X)=\mathbb{E}^{\theta}\big(I(\theta\neq\delta)\big|X\big)=\mathbb{P}(\theta\neq\delta\,|\,X)=1-\pi(\delta\,|\,X).$$

Now, the Bayes rule should minimize this posterior risk. To that end, it maximizes $\pi(\delta\,|\,X)$. The Bayes rule is the point of maximum of the posterior distribution (pmf) $\pi(\theta\,|\,X)$. In fact, this is the posterior mode.

Summary:

Posterior mode is the Bayes decision with respect to the zero-one loss.

In the continuous case, on the other hand, the probability of $\theta=\delta$ is 0, so there is noting to maximize. For this reason, the 0-1 loss function is often defined as

$$L(\theta,\delta)=I\big(\,|\theta-\delta|>\varepsilon\big)=\begin{cases}1, \text{ if } |\theta-\delta|>\varepsilon\\ 0, \text{ if } |\theta-\delta|\leq\varepsilon\end{cases}$$

allowing the estimator δ to differ from the paramter θ by at most a small ε. The Bayes decision δ in this case maximizes the probability $\mathbb{P}(\delta-\varepsilon\leq\theta\leq\delta+\varepsilon\,|\,X)$, and it converges to the posterior mode as we send ε to 0 (since ε is chosen arbitrarily). This lead to a natural refinement, or we should call the generalized maximum likelihood estimator.

Definition: Generalized Maximum Likelihood Estimator

The generalized MLE of the parameter θ is the posterior mode, the value of θ

that maximizes the pdf or pmf.

4. Bayesian Inference: Estimation, Hypothesis Testing, and Prediction

By now we have obtained all the three components needed for the Bayesian decision making. We collected data, and we determine the prior distribution and the loss function. Then, combining the data and the prior, we obtained the posterior distribution. All the knowledge about the unknown parameter is now included in the posterior, and that is what we shall see in this section.

In 4.1 we introduce the unbiasedness and the variance of the Bayesian estimation, then in 4.2 we discuss the Bayesian interval estimation, where we talk about the HPD region in 4.3 we offer a treatment on the Bayesian hypothesis testing, we introduce the Bayesian prediction problem in 4.4 to close this section.

4.1 Bayesian Estimation and Precision Evaluation

As we have seen in the last section, the Bayesian estimator may take in different forms. The most common one among them is of course the posterior mean, which is given by

$$\hat{\theta}_{\text{Bayes}} := \mathbb{E}(\theta|X) = \begin{cases} \sum_\theta \theta\pi(\theta|X) = \frac{\sum \theta f(X|\theta)\pi(\theta)}{\sum f(X|\theta)\pi(\theta)}, & \text{if } \theta \text{ is discrete} \\ \int_\theta \theta\pi(\theta|X)d\theta = \frac{\int \theta f(X|\theta)\pi(\theta)d\theta}{\int f(X|\theta)\pi(\theta)d\theta}, & \text{if } \theta \text{ is continuous} \end{cases}$$

Posterior mean is the conditional expectation of θ given data X. In abstract terms, the Bayes estimator $\hat{\theta}_{\text{Bayes}}$ is what we expect θ to be, after we observed a sample.

A natural question to ask is that "how accurate is such an estimator?" Among all estimators, $\hat{\theta}_{\text{Bayes}} = \mathbb{E}(\theta|X)$ has the lowest squared-error posterior risk

$$\mathbb{E}\left((\hat{\theta}-\theta)^2\middle|X\right)$$

and also the lowest Bayes risk $\mathbb{E}\mathbb{E}(\hat{\theta}-\theta)^2$, where this double expectation is taken over the joint distribution of X and θ.

For the Bayes estimator $\hat{\theta}_{\text{Bayes}}$, posterior risk equals posterior variance, which shows variability of θ around $\hat{\theta}_{\text{Bayes}}$.

A parameter θ is estimated by an estimator $\delta = \hat{\theta}$. How accurate is this decision? A frequentist measure of precision is the mean-squared error (MSE) given by

$$\text{MSE}(\delta) := \mathbb{E}_\theta^X(\delta-\theta)^2 = \mathbb{E}_\theta(\delta-\mathbb{E}_\theta\delta)^2 + (\mathbb{E}_\theta\delta-\theta)^2 = \text{Var}(\delta) + \left(\text{Bias}(\delta)\right)^2.$$

Here all the expectations are taken in the frenquentist way, i.e. they integrate over all samples of X, given a fixed parameter θ.

Following from the Bayesian appraoch, the same expectations are taken with respect to the posterior distribution of θ, given a fixed, already observed sample X. This is called posterior variance of estimator δ,

$$\begin{aligned} V_X(\delta) &:= \mathbb{E}_X^{\pi(\theta|X)}(\delta-\theta)^2 \\ &= \mathbb{E}\left((\theta-\mu_X)^2\middle|X\right) + (\delta-\mu_X)^2 \\ &= \tau_X^2 + (\delta-\mu_X)^2, \end{aligned}$$

where $\mu_X := \mathbb{E}(\theta \,|\, X)$ is the mean and $\tau_X^2 := \mathrm{Var}(\theta \,|\, X)$ is the variance of the posterior distribution $\pi(\theta \,|\, X)$.

There are two variance components in the posterior variance of δ. One is the posterior variance of the parameter θ, and the other is the squared deviation of estimator δ from the posterior mean μ_X. So, the total variability of our estimation consists of the variability of θ around its mean and the distance between that posterior mean and our estimator, this is very reasonable.

Definition: Posterior Variance

$\mathrm{Var}(\theta \,|\, X)$ or $\tau_X^2 := \mathbb{E}\big((\theta - \mu_X)\,\big|\,X\big)$, the posterior variance of the parameter θ, variance of the posterior distribution θ.

$V_X(\delta) := \tau_X^2 + (\delta - \tau_X)^2$, posterior variance of the estimator δ.

4.2 Bayesian Credible Sets

Bayesian and frequentist approaches to the confidence estimation are quite different. In Bayesian analysis, having a posterior distribution of θ, we no longer have to explain the confidence level $(1-\alpha)$ in terms of a long run of samples. Instead, we can give an interval $[a, b]$ or a set C that has a posterior probability $(1-\alpha)$ and state that the parameter θ belongs to this set with probability $(1-\alpha)$. Such a statement was impossible before we considered prior and posterior distributions. This set is sometimes called a $(1-\alpha)$ 100% credible set.

Definition: $(1-\alpha)$ 100% Credible Set

A set C is said to be a $(1-\alpha)$ 100% credible set for the parameter θ if the posterior probability for θ to belong to C is $(1-\alpha)$. That is,

$\mathbb{P}(\theta \in C \,|\, X) = \displaystyle\int_C \pi(\theta \,|\, X) d\theta = 1 - \alpha$. Note that such a set may not be unique.

Hyndman mentioned in [8] a region possessing the minimized size but with highest probability. This is the notion of Highest Density Regions (HDR), and in Bayesian approach we call it the highest posterior density regions (HPD), or sometimes HPD sets.

Definition: Highest Density Regions (HDR)

- The region covering the sample space for a given probability $1-\alpha$, should have the smallest possible volume.
- Every point inside the region should have probability density at least as large as every point outside the region.

Then such a region is called the highest density region (HDR).

One of the most distinctive property of HDR's is that of all possible regions of probability coverage, the HDR has the smallest region possible in the sample space. "Smallest" mean with respect to some simple measure such as the usual Lebesgue measure; in the one-dimensional continuous case that would be the shortest interval, and in the two-dimensional case that would be the smallest area of the surface.

Similar to this idea, minimizing the length (resp. area) of the set C among all the $(1-\alpha)$ 100% credible sets, we just have to include all the points θ with a high posterior density $\pi(\theta\,|\,X)$, namely,

Definition: Highest Posterior Density (HPD)

The HPD is the set of the form $C := \{\theta\,|\,\pi(\theta\,|\,X) \geq c\}$ for some constants c.

One very useful example is the $\mathrm{N}(\mu_X, \tau_X)$ posterior distribution of θ, the $(1-\alpha)$ 100% HPD set is given by

$$\mu_x \pm z_{\alpha/2}\tau_x = [\mu_x - z_{\alpha/2}\tau_x, \mu_x + z_{\alpha/2}\tau_x]. \tag{4.1}$$

In fact, all the HPD are Bayesian decitions under the loss function given by

$$L(\theta, C) = \lambda\,|\,C\,| + I_{\theta \notin C}, \tag{4.2}$$

where $|\,C\,|$ is the size, usually the length of C, and λ is a coefficient.

4.3 Bayesian Hypothesis Testing

Bayesian hypothesis testing is very easy to interpret. We can compute the prior and the posterior probabilities for the hypothesis H_0 and alternative H_A to be true and decide from there which on to accpet or to reject.

Computing such probabilities was not possible without prior and posterior distributions of the parameter θ. In non-Bayesian statistics, θ is not random, thus H_0 and H_A were either true (with probability 1) or false (with probability 1).

For the Bayesian tests, on the other hand, in order for H_0 and H_A to have meaningful, non-zero probabilities, they often represent disjoint sets of parameter values, namely,

$$H_0 : \theta \in \Theta_0 \text{ versus } H_A : \theta \in \Theta_1. \tag{4.3}$$

(Which makes sense because exact equality $\theta = \theta_0$ is unlikely to hold anyway, and in practice it is understood as $\theta \approx \theta_0$).

Comparing poseterior probabilities of H_0 and H_A yields $\mathbb{P}(\Theta_0\,|\,X)$ and $\mathbb{P}(\Theta_1\,|\,X)$, we decide whether $\mathbb{P}(\Theta_1\,|\,X)$ is large enough to present significant evidence and to reject the null hypothesis. One can again compare it with the $(1-\alpha)$ such as 0.90, 0.95, and 0.99, or state the result in terms of likelihoods, "the null hypothesis is this much likely to be true".

Often one can anticipate the consequences of Type I and Type II errors in hypothesis testing and assign a loss $L(\theta, a)$ associated with each possible error. Here θ is the parameter, and a is our action, the decision on whether we accept or reject the null hypothesis.

Each decision then has its posterior risk $\rho(a)$, defined as the expected loss computed under the posterior distribution. The action with the lower posterior risk is our Bayes decision.

Suppose that the Type I error causes the loss given by

$$w_0 = \text{Loss(Type I error)} = L(\theta,\ \text{reject}\, H_0),\ \text{ for } \theta \in \Theta_0,$$

and the Type II error causes the loss given by

$$w_1 = \text{Loss(Type II error)} = L(\theta,\ \text{accept}\, H_0),\ \text{ for } \theta \in \Theta_1.$$

That is the zero-w_i loss function (for $i = 1$, or 2), and it generalizes the zero-one loss.

Posterior risks of each possible action are then computed as

$$\rho(\pi,\ \text{reject}\,H_0\,|\,X) = w_0\pi(\Theta_0\,|\,X),$$
$$\rho(\pi,\ \text{accept}\,H_0\,|\,X) = w_1(\pi(\Theta_1\,|\,X).$$

Now we can determine the Bayesian decision. If $w_0\pi(\Theta_1\,|\,X) \le w_1\pi(\Theta_0\,|\,X)$, the Bayesian action is to accept H_0, if $w_1\pi(\Theta_1\,|\,X) \le w_0\pi(\Theta_0\,|\,X)$, the Bayesian action is to reject H_0.

Following this algorithm, the Bayesian approach to hypothesis testing naturally generalizes to the case of more than two hypotheses. Instead of classifying the unknown parameters into eitehr Θ_0 (accept H_0) or Θ_1 (reject H_0), it can be classified into one of disjoint subsets of Θ, i.e. for a partition $\{\Theta_1, \cdots, \Theta_n\}$ of Θ,

$$H_1 : \theta \in \Theta_1, \cdots, H_n : \theta \in \Theta_n. \tag{4.4}$$

A loss function will then include $L(\theta, \delta) = w_{ij}$, a penalty for accepting hypothesis H_j whereas hypothesis H_i is true.

Similarly, we can include an action of making no decision and concluding that there is not enough information for or against either hypothesis. This also carries a predetermined penalty w_{0j}, and sometimes "no action" may be the optimal comparing with the penalty of accepting a wrong hypothesis; note this happens mostly when we have no access in gaining further information, sometimes it may happen that even with enough details, we arrive at doing nothing all the same.

A popular tool for the Bayesian hypothesis testing is Bayes factors.

Definition: Bayesian Factor

The Bayes factor is defined to be $B := \dfrac{\pi(\Theta_0\,|\,X)/\pi(\Theta_1\,|\,X)}{\pi(\Theta_0)/\pi(\Theta_1)}$, the posterior odds ratio divided by the prior odds ratio.

Often the Bayes factors are quite stable or insensitive to the choice of prior probabilities $\pi(\Theta_0)$ and $\pi(\Theta_1)$.

How do we use the Bayes factors for hypothesis testing? Given the Bayes factor B, anyone multiplies it by the prior odds ratio $\pi(\Theta_0)/\pi(\Theta_1)$ and obtains the posterior odds ratio $\pi(\Theta_0\,|\,X)/\pi(\Theta_1\,|\,X) = B\big(\dfrac{\pi(\Theta_0)}{\pi(\Theta_1)}\big)$ and use it decide H_0 or H_1.

4.4 Prediction

We are going to predict a random variable Z that is somehow related to the unknown parameter θ to close this section. We first formulate the problem.

Bayesian Prediction Problem:

A sample $X = (X_1, \cdots, X_n)$ is observed from $f(X\,|\,\theta)$. Unknown parameter θ has a prior distribution $\pi(\theta)$. Random variable Z has distribution given by $g(z\,|\,\theta)$. We wish to find the predictive density of Z, which is $p(z\,|\,X)$, the distribution of Z given the observed X.

As always in the Bayesian analysis, the solution to the Bayesian prediction problem is based on the posterior distirbution of θ.

First, combine the posterior of θ with the distribution of Z given θ to obtain the joint distribution of Z and θ,

$$g(z,\theta|X) = g(z|\theta)\pi(\theta|X).$$

Then integrate the joint density over θ to obtain the marginal distribution of Z,

$$p(z|X) = \int g(z,\theta|X)d\theta = \int g(z|\theta)\pi(\theta|X)d\theta. \tag{4.5}$$

This is the prediction density of Z. This integral also represents the expectation $\mathbb{E}\left(g(z|\theta)\middle|X\right)$, taken with respect to the posterior distribution of θ, i.e.

$$\mathbb{E}\left(g(z|\theta)\middle|X\right) = (4.5).$$

If X and Z are, unfortunately, not independent, given θ, then the density of Z is given by $g(z|\theta,X)$.

5. Validity for Bayesian Approach — A Short Discussion

A probability formula was used by Bayes ([10]) to combine a mathematical prior with a model plus data; it gave just a mathematical posterior, with no consequent objective properties. An analogy provided by Bayes did have a real and descriptive prior, but it was not part of the problem actually being examined.

A familiar Bayes example uses a special model, a location model; and the resulting intervals have attractive properties, as viewed by many in statistics.

Fisher ([11]) and Neyman ([12]) defined confidence. And the Bayes intervals in the location model case are seen to satisfy the confidence derivation, thus providing an explanation for the attractive properties.

In [9], D.A.S. Fraser showed that the proportion of true statements in the Bayes case depends critically on the presence of linearity in the model; and with departure from this linearity the Bayes approach can be a poor approximation and be seriously misleading. Beyesian integration of weighted likelihood thus provides a first-order linear approximation to confidence, but without linearity can give substantially incorrect results.

The only source of variation available to support a Bayes posterior probability calculation is that provided by the model, which is what confidence uses.

Lindely ([13]) examined the probability formula argument and the confidence argument and found that they generated the same result only in the Bayes location model case; he then judged the confidence argument to be wrong.

If the model, however, is not location and, thus, the variable is not linear with respect to the parameter, then a Bayes interval can produce correct answers at a rate quite different from that claimed by the Bayes probability calculation; thus, the Bayes posterior may be an unreliable presentation, an unreliable approximation to confidence, and can thus be judged to be wrong.

The failure to make true assertions with a promised reliability can be extreme with the Bayes use of mathematical priors (Stainforth et al., [14]; Heinrich, [15]).

The claim of a probability status for a statement that can fail to be approximate confidence is mis-representation. In other areas of science much false claims would be treated seriously.

Using weighted likelihood, however, can be a fruitful way to explore the information available from just a likelihood function. But the failure to have even a confidence interpretation deserves more than just gentle caution.

A personal or a subjective or an elicited prior may record useful background to recorded in parallel with a confidence assessment. But to use them to do the analysis and just get approximate or biased confidence seem to overextend the excitement of exploratory procedures.

6. Bayesian Convolution

A general convolution theorem within a Bayesian framework is presented in this section as a result of [16]. Consider estimation of the Euclidean parameter θ by an estimator T within a parametric model. Let W be a prior distribution for θ and define G as the W-average of the distribution $T-\theta$ under the parameter θ. In some cases, for any estimator T the distribution G can be written as a convolution $G = K \star L$ with K a distribution depending only on the model, i.e. on W and the distribution under θ of the observations. In such a Bayesian convolution result optimal estimators exist, satisfying $G = K$.

Before we proceed, we introduce some basic results from Fourier analysis. We assume the readers are already familiar with the concepts, so we shall not perform the proofs of these results, for those readers who are not familiar with this topic may concult [17].

If f is an integrable function given on an interval $[a,b] \subseteq \mathbb{R}$ such that $b-a=L$, then the nth Fourier coefficient of f is defined by

$$\hat{f}(n) := \frac{1}{L}\int_a^b f(x)e^{-2\pi inx/L}dx, n \in \mathbb{N}. \tag{6.1}$$

The Fourier series of f is given formally by

$$\sum_{n=-\infty}^{+\infty} \hat{f}(n)e^{2\pi inx/L}. \tag{6.2}$$

Whenever we use a_n we refer to the nth Fourier coefficients of f, denoted by

$$f(x) \sim \sum_{n=-\infty}^{+\infty} a_n e^{2\pi inx/L}. \tag{6.3}$$

to indicate that the series on the RHS is the Fourier series of f.

Given two 2π-periodic integrable functions f and g on $\mathbb{R}$, we define their convolution, denoted by $f \star g$, on $[-\pi,\pi]$, by

$$(f \star g)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(y)g(x-y)dy. \tag{6.4}$$

Also, since the functions are assumed to be periodic, we can rewrite it as

$$(f \star g)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-y)g(y)dy. \tag{6.5}$$

Theorem 6.1: Properties of Convolution

Suppose that f, g, and h are 2π-periodic integrable functions. Then

(i) $f \star (g+h) = (f \star g) + (f \star h)$.

(ii) $(cf) \star g = c(f \star g) = f \star (cg)\ \forall c \in \mathbb{C}$.
(iii) $f \star g = g \star f$.
(iv) $(f \star g) \star h = f \star (g \star h)$.
(v) $f \star g$ is continuous.
(vi) $\widehat{f \star g}(n) = \hat{f}(n)\hat{g}(n)$.

Suppose that we have a random element $X \sim \mathscr{P} := \{\mathbb{P}_\theta | \theta \in \Theta \subseteq \mathbb{R}^k\}$, on a measurable space $(X, \mathscr{A})$, where $\mathscr{A}$ is the σ-algebra generated by the subsets of X. On the basis of this sample we want to estimate the parameter θ. In a Bayesian set-up we choose a weight function or prior W on $\mathbb{R}^k$, putting its mass in Θ, and we consider the average distribution function

$$G(y) = \int_{\mathbb{R}^k} \mathbb{P}_\theta(T - \theta \leq y) dW(\theta), y \in \mathbb{R}^k, \tag{6.6}$$

where $T := t(X)$ is an estimator of θ.

The following observation tells us that for general dimension k the average distribution G is the convolution of a distribution, which depends on $\mathscr{P}$ and W, but which does not depend on the estimator T, and any other distributions.

Let $\psi : X \to \mathbb{R}^k$ be a measurable function such that $\psi(X) - \vartheta$ and X are independent, where $\vartheta \sim W$. Then $\psi(X) - \vartheta$ and $t(X) - \psi(X)$ are also independent. Since $T - \vartheta$ could be rewritten as

$$T - \vartheta = t(X) - \vartheta = \{\psi(X) - \vartheta\} + \{t(X) - \psi(X)\},$$

we may conclude that $G(\cdot) = \mathbb{P}(T - \vartheta \leq \cdot)$ is a convolution of the distribution of $\vartheta(X) - \vartheta$, i.e. $K(\cdot) = \mathbb{P}(\vartheta(X) - \vartheta \leq \cdot)$, which indeed does not depend on T. Consequently, there exists a distribution L such that $G = K \star L$. We will call this identity the Bayes Convolution Theorem. Furthermore, we will call $T = \psi(X)$ the best estimator in the sense $G = K \star L$, since this choice makes L degenerate at 0. We summarize these into the following result.

Theorem 6.2: Bayes Convolution Theorem

Let ϑ be a random variable taking values in $\Theta \subseteq \mathbb{R}^k$ and let the conditional distribution of X given $\vartheta = \theta$ be $\mathbb{P}_\theta$ on $(X, \mathscr{A})$. If the measurable function $\psi : X \to \mathbb{R}^k$ is such that $\psi(X) - \vartheta$ and X are independent, then the distribution $G(\cdot) = \mathbb{P}(T - \vartheta \leq \cdot)$ of $T - \vartheta$ is the convolution of the distribution $K(\cdot) = \mathbb{P}(\psi(X) - \vartheta \leq \cdot)$ and some other distribution L, i.e. $G = K \star L$.

If such a ψ exists, the best estimator with respect to the Bayes risk is $\psi(X) + c$, for $c \in \mathbb{R}^k$ may depend on the loss function. Indeed,

$$\inf_T \mathbb{E}L(T - \vartheta) = \inf_c \mathbb{E}L(\psi(X) - \vartheta + c) \tag{6.7}$$

holds for all convex loss functions L in view of the **Conditional Jensen inequality**.

If for $k = 1$ the distribution of $\psi(X) - \vartheta$ is strongly unimodal, then (6.7) holds for all loss functions which are decreasing-increasing, and again the best estimator is determined by ψ and does not depend on the loss function apart from a shift by c. This may be seen as follows. First, note that the convolution with K strongly unimodal implies that G is at least as spread out as K, i.e. that the quantiles of G are at least as far apart as those of K. Subsequently, if there exists a $u_0 \in [0,1]$ with $G^{-1}(u_0) = 0$ then this spread property

$$\begin{aligned}\mathbb{E}L(T-\vartheta) &= \int_0^1 L\big(G^{-1}(u) - G^{-1}(u_0)\big)du \\ &\geq \int_0^1 L\big(K^{-1}(u) - K^{-1}(u_0)\big)du \\ &= \mathbb{E}L\big(\psi(X) - \vartheta - K^{-1}(u_0)\big). \qquad (6.8)\end{aligned}$$

Finally, note that (6.8) may be adapted to the case where 0 is not a quantile of G.

Reference:

[1]: John E. Freund's, *Mathematical Statistics with Applications*, Eighth Edition, Pearson Education Limited 2014.

[2]: Michael Baron, *Stat 618 Bayesian Statistics Lecture Notes*, available online.

[3]: Lester Mackey, Bayes Estimators and Average Risk Optimality, available online.

[4]: Peter Orbanz, *Lecture Notes on Bayesian Nonparametrics*, available online.

[5]: Heuvel, Edwin R. van den, and Chris A. J. Klaassen. "Bayes Convolution." *International Statistical Review / Revue Internationale de Statistique* 67, no. 3 (1999): 287–99. https://doi.org/10.2307/1403707.

[6]: Elias M. Stein, Rami Shakarchi, *Fourier Analysis, An Introduction*, Princeton University Press, pp. 69-95.

[7]: Andrew Gelman, John B. Carlin, et al. *Bayesian Data Analysis, Third Edition*, CRC Press, pp.

[8]: Hyndman, Rob J. "Computing and Graphing Highest Density Regions." *The American Statistician* 50, no. 2 (1996): 120–26. https://doi.org/10.2307/2684423.

[9]: D.A.S. Fraser, *Is Bayes Posterior just Quick and Dirty Confidence?* Statist. Sci. 26(3): 299-316 (August 2011). DOI: 10.1214/11-STS352.

[10]: Bayes T (1763), *An Essay Towards Solving a Problem in the Doctrine of Chances*, Philos. Trans. R. Soc. Lond. 53 370-418; 54 296-325. Reprinted in Biometrika 45 (1958) 293-315.

[11]: Fisher R.A. (1935), *The Fiducial Argument in Statistical Inference*, Ann. Eugenics B 391-398.

[12]: Neyman, J. (1937), *Outline of a Theory of Statistical Estimation based on the Classical Theory of Probability*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 237 333-380.

[13]: Lindley, D. V. (1958). *Fiducial distributions and Bayes' theorem*. J. Roy. Statist. Soc. Ser. B 20 102–107. MR0095550.

[14]: Stainforth, D. A. , Allen, M. R. , Tredger, E. R. and Smith, L. A. (2007). *Confidence, uncertainty and decisionsupport relevance in climate predictions*. Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 365 2145–2162.

[15]: Heinrich, J. (2006). *The Bayesian approach to setting limits: What to avoid?* In Statistical Problems in Particle Physics , Astrophysics and Cosmology (L. Lyons and U. M. Karag¨oz, ¨ eds.) 98–102. Imperial College Press, London. MR2270215.

[16]: Heuvel, Edwin R. van den, and Chris A. J. Klaassen. "Bayes Convolution." *International Statistical Review / Revue Internationale de Statistique* 67, no. 3 (1999): 287–99. https://doi.org/10.2307/1403707.

[17]: Elias M. Stein, Rami Shakarchi, *Fourier Analysis, An Introduction*, Princeton Lectures in Analysis, pp. 29-57, Chapter 2.

Review on Elmentary Linear Regression

Tianyu Zhang[5]

Abstract:

In this short review article we present both the univariate linear regression and bivariate linear regression, along with the properties and methods in evaluation. Special cases of inferences are also provided in the fourth section.

1. Introduction

If we are given the joint distribution of two random variables X and Y, and X is known to take on the value x, the basic problem of bivariate regression is that of determining the conditional mean $\mu_{Y|x}$, i.e. the "average" value of Y for the given value of X. The term "regression", as it is used here, dates back to Francis Galton, who used it to indicate certain relationships in the theory of heredity. In problems involving more than two random variables, i.e. the multiple regression, we are concerned with quantities cuh as $\mu_{Z|x,y}$, the mean of Z for given values of X and Y.

Definition: Bivariate Regression

If $f(x,y)$ is the value of the joint density of two random variables X and Y, the bivariate regression consisits of determining the conditional density of Y, given $X = x$, and then evaluating the integral

$$\mu_{Y|x} = \mathbb{E}(Y\,|\,x) = \int_{-\infty}^{\infty} y \cdot w(y\,|\,x)dy,$$

where $w(y\,|\,x)$ is the conditional distribution. The resulting equation is called the regression equation of Y on X. Alternatively, the regression equation of X on Y is given by

$$\mu_{X|y} = \mathbb{E}(X\,|\,y) = \int_{-\infty}^{\infty} x \cdot f(x\,|\,y)dy.$$

Example 1.1:

If X and Y have the multinomial distribution

$$f(x,y) = \binom{n}{x, y, n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}$$

for $x = 0,1,\cdots,n$ and $y = 0,1,\cdots,n$ with $x + y \le n$. Find the regression equation for Y on X.

Solution:

The marginal distribution of X is given by

$$\begin{aligned} g(x) &= \sum_{y=0}^{n-x} \binom{n}{x, y, n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y} \\ &= \binom{n}{x} \theta_1^x (1-\theta_1)^{n-x} \end{aligned}$$

for $x = 0,1,\cdots,n$, which we recognize as a bimomial distribution with the

[5] BIMSA, bidenbaka@gmail.com

parameters n and θ_1. Therefore, one has

$$w(y\,|\,x) = \frac{f(x,y)}{g(x)} = \frac{\binom{n-x}{y}\theta_2^y(1-\theta_1-\theta_2)^{n-x-y}}{(1-\theta_1)^{n-x}}$$

for $y = 0,1,\cdots,n-x$, rewriting the formula yields

$$w(y\,|\,x) = \binom{n-x}{y}\left(\frac{\theta_2}{1-\theta_1}\right)^y\left(\frac{1-\theta_1-\theta_2}{1-\theta_1}\right)^{n-x-y}.$$

We find by inspection that the conditional distribution of Y given $X = x$ is a binomial distribution with parameters $n-x$ and $\dfrac{\theta_1}{1-\theta_1}$, so the regression equation of Y on X is $\mu_{Y|x} = \dfrac{(n-x)\theta_2}{1-\theta_1}$. ||

An important feature of **Example 1.1** is that the regression equation is linear; i.e. it is of the form

$$\mu_{Y|x} = \alpha + \beta x,$$

where α and β are constants, called the regression coefficients.

We now introduce a model in which one variable X affect another one Y and the relation is assumed to be linear up to a random vector. That is to say, we have n observations of variables X and Y: $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ and we assume that they satisfy the folloing model:

$$y_i = \beta_0 + \beta_i x_i + \varepsilon_i. \tag{1.1}$$

Here β_0 and β_1 are unknown parameters that we want to estimate. The quantities x_i, for $i = 1,\cdots,n$ are known as parameters, which are called explanatory variables, or independent random variables as we do in high school algebra. The variables ε_i are error terms. They are responsible for the randomness of the model. They are always assumed to have zero mean: $\mathbb{E}\varepsilon_i = 0$. They are also often but not always assumed to have unknown variance σ_2 that does not depend on the index i, i.e. $\mathbb{E}\epsilon_i = \sigma_2$. Even more restrictively, they are often assumed to be normally distributed $\varepsilon_i \sim N(0,\sigma_2)$.

The values y_i are random since they are functions of ε_i (We could write them Y_i following our usual convention about the random variables.) They are usually called the response variables or dependent random variabels as we did in high school algebra. Therefore, $y_1,\cdots,y_n$ are n independent observations of the response variable Y.

In fact, (1.1) is called the regression model. It is often written in a short form that omits the subscript index i.

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{1.2}$$

A general linear regression model includes more than one explanatory variable:

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_n x_i^b + \varepsilon_i. \tag{1.3}$$

Or in short the notation:

$$Y = \beta_0 + \beta_1 x^1 + \cdots + \beta_n x^n + \varepsilon. \tag{1.4}$$

This model is very flexible and can be used to model non-linear dependencies as well. For example, if we believe Y depends on X as a polynomial of degree 3, we can add explanatory variables that corresponds to squares and cubes of X. Then we only need to estimate the regression model given by

$$Y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 + \varepsilon. \tag{1.5}$$

In some other cases we can only consider a transformation of random variable Y so as to get a more suitable distribution for random error terms ε.

As usual, we aim to estimate the estimators $\beta_0, \cdots, \beta_n$ and test some hypothesis correpsonding to their values. There is another goal which we have not seen before. We might be interested in predicting the response Y for some other values of x. In addition we might be intereseted in having some kinds of confidence interval for our prediction.

2. Simple Linear Regression

In this subsection we shall introduce the linera regression model for one variable, namely $Y = \beta_0 + \beta_1 x^1 + \cdots + \beta_n x^n$, which is a polynomial of degree n.

2.1 Least Squares Estimator

Here we look at the simple linear regression given by $y = \beta_0 + \beta_1 x + \varepsilon$, although the methods are also applicable to the general linear regression models.

In order to estimate the parameters β_0 and β_1 we could use the MLE by writing the likelihood function of the random quantities y_i and maximizing it with respect to both β_0 and β_1. It turns out that for normally distributed $\varepsilon \sim N(0,\sigma^2)$ this method gives the same estimates as a simple method describes below.

This simple method aims to minimize the deviation of the fitted values given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, from the observed values of y_i, by a choice of the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Specifically, the method of least squres aims to minimize the sum of squared Errors (SSE), which is defined by

Definition: Sum of Squared Errors (SSE)

SSE:= $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, which is minimized by a choice of $\hat{\beta}_0$ and $\hat{\beta}_1$.

As usual, this minimization can be done by using the First Order Conditions. The first-order condition is obtained by setting the derivative (or gradient) of the log-likelihood function equal to zero.

Definition: Ordinary Least Squared Estimators

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which solve the SSE are called the ordinary Least Square estimators of the linear regression model.

The estimators are called ordinary LS estimators, because sometimes in the definition of SSE the terms have different weights. In this case the solution is called the weighted least squared estimators.

It is a bit simpler to do it for a modified model, in which the explanatory variables are centered by subtractiing their mean:

$$y_i = \alpha_0 + (x_i - \bar{x}) + \varepsilon_i. \tag{2.1}$$

Clearly, this model is equivalent to the original simple regression with $\beta_0 = \alpha_0 - \beta_1\bar{x}$. It is also clear that the LS estimators in these regression problems are related by the similar equations $\hat{\beta}_0 = \hat{\alpha}_0 - \hat{\beta}_1\bar{x}$.

Theorem 2.1:

The least squares estimators are given by the following formulas:

$\hat{\alpha}_0 = \bar{y}, \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$, where $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ and $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

This implies that for our original problem, we have also the following least squares estimator for the parameter β_0:

$$\hat{\beta}_0 = \hat{\alpha}_0 - \hat{\beta}_1\bar{x} = \bar{y} - \hat{\beta}_1\bar{x}.$$

Proof:

Step I: $\hat{\alpha}_0 = \bar{y}$.

Taking the partial derivative on SSE with respect to α_0 yields

$$\frac{\partial SSE}{\partial \hat{\alpha}_0} = \frac{\partial\left\{ \sum_{i=1}^{n}\left[y_i - \left(\hat{\alpha}_0 + \hat{\beta}_1(x_i - \bar{x})\right)\right]^2\right\}}{\partial \hat{\alpha}_0}$$

$$= 2\sum_{i=1}^{n}\left[y_i - \left(\hat{\alpha}_0 + \hat{\beta}_1(x_i - \bar{x})\right)\right]\cdot(-1).$$

Setting this to 0 gives us

$$0 = \sum_{i=1}^{n}\left[y_i - \left(\hat{\alpha}_0 + \hat{\beta}_1(x_i - \bar{x})\right)\right] = \sum_{i=1}^{n} y_i - n\hat{\alpha}_0 + \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x}).$$

Since by (5.6) we have centered it hence the red part vanishes and this gives us

$$0 = \sum_{i=1}^{n} y_i - n\hat{\alpha}_0 \Rightarrow \hat{\alpha}_0 = \bar{y}.$$

Step II: $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

Similarly, taking the partial derivative on SSE with respect to $\hat{\beta}_1$ yields the result, we leave the proof as an exercise.

□

2.2 Properties of LS Estimators

We aim to calculate the expectation and the variance of LS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. This information is important for calculation of the bias of the estimators and for construction of confidence interval.

We start with $\hat{\beta}_1$, which is typically more useful in practice since β_1 measures the effect of X on Y.

Theorem 2.2:

Assume that the error terms in the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ have the properties $\mathbb{E}\varepsilon_i = 0$ and $\mathrm{Var}\varepsilon_i = \sigma^2$. Then

(i) $\mathbb{E}\hat{\beta}_1 = \beta_1$.

(ii) $\mathrm{Var}\hat{\beta}_1 = \sigma^2 / S_{xx}$.

If, in addition, ε_i are normal, then $\hat{\beta}_1$ is also normal.

Before proving this theorem, let us derive some consequences. First, we see that $\hat{\beta}_1$ is an unbiased estimator of β_1. Second, if $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \to \infty$ as $n \to \infty$, then $\hat{\beta}_1$ is a consistent estimator of β_1. The condition $S_{xx} \to \infty$ as $n \to \infty$ means that as the samlpe grows we continue to observe sufficient variation in explanatory variables x_i.

This consequence may look surprising since it states that the larger the deviation between $\bar{x}$ and x_i is, the better performance of $\hat{\beta}_1$ is guaranteed. One good interpretation for this anti-intuitive result may be that, **loosely speaking**, consider x_i are from the probability space $(\Omega_X, \mathscr{S}_X, \mathbb{P}_X)$ and y_i are from $(\Omega_Y, \mathscr{S}_Y, \mathbb{P}_Y)$ along with a mapping $f : \Omega_X \to \Omega_Y$, then the "larger" the space Ω_X is, i.e. the more values we can take for x, the more possible we can find a $\hat{\beta}_1$ to achieve our requirement, or, equivalently, the more confident we are at this $\hat{\beta}_1$. Now we proceed to the proof.

Proof of Theorem 2.2:

It is convenient to work with the modified form (2.1), i.e.

$$y_i = \alpha_0 + (x_i - \bar{x}) + \varepsilon_i.$$

Note that x_i, $\bar{x}$, and $\bar{y}$ are not random. One has

$$\begin{aligned} S_{xy} &:= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= \sum (x_i - \bar{x}) y_i - \bar{y} \sum (x_i - \bar{x}) \quad (\bar{y} \text{ is not random}) \\ &= \sum (x_i - \bar{x}) y_i. \qquad \text{(sum of variance is always 0)} \end{aligned}$$

Step I: $\mathbb{E}\hat{\beta}_1 = \beta_1$.

$$\begin{aligned} \mathbb{E}\hat{\beta}_1 &= \mathbb{E}\frac{S_{xy}}{S_{xx}} \qquad (\text{by } \textbf{Theorem 2.1}) \\ &= \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \mathbb{E} y_i \\ &= \frac{1}{S_{xx}} \sum (x_i - \bar{x}) \big(\alpha_0 + \beta_1 (x_i - \bar{x})\big) \qquad (\text{by (2.1)}) \\ &= \frac{1}{S_{xx}} \alpha_0 \sum (x_i - \bar{x}) + \sum \frac{1}{S_{xx}} \beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) \\ &= \sum \frac{1}{S_{xx}} \beta_1 \sum (x_i - \bar{x})(x_i - \bar{x}) \quad \text{(sum of variance is always 0)} \\ &= \beta_1 \qquad (\text{by the definition of } S_{xx}) \end{aligned}$$

Step II: $\mathrm{Var}\hat{\beta}_1 = \sigma^2 / S_{xx}$.

Similarly, we calculate the variance, it is helpful to denote $\mathrm{Var} y_i = \mathrm{Var} \varepsilon_i = \sigma^2$. One has, taking the variance operation with respect to $\hat{\beta}_1$

$$\mathrm{Var}\hat{\beta}_1 = \mathrm{Var}\frac{S_{xy}}{S_{xx}} \qquad (\text{by } \textbf{Theorem 2.1})$$

$$= \frac{1}{S_{xx}^2}\sum (x_i - \overline{x})^2 \mathrm{Var} y_i$$
$$= \frac{1}{S_{xx}^2} S_{xx}\sigma^2 \qquad \text{(definition of } S_{xx} \text{ and } \mathrm{Var} y_i = \sigma^2\text{)}$$
$$= \sigma^2 / S_{xx}.$$

Step III: ε_i normal $\Rightarrow \hat{\beta}_1$ normal.

Finally, if ε_i are normal, then y_i are also normal. Note that $\hat{\beta}_1$ is a weighted sum of y_i and the coefficients in this sum are non-random. We know that this implies that the sum itself is also normal.

□

In addition, we need to point out that the variance of ε_i in **Theorem 2.2**, defined as σ^2, is necessarily to be finite. For other estimators we have the similar results.

Theorem 2.3:

Assume that the error terms in the simple linear regression model $y_i = \alpha_0 + \beta_1(x_i - \overline{x}) + \varepsilon_i$ has the property that $\mathbb{E}\varepsilon_i = 0$ and $\mathrm{Var}\varepsilon_i = \sigma^2$. Then,

(i) $\mathbb{E}\hat{\alpha}_0 = \alpha_0$.

(ii) $\mathrm{Var}\hat{\alpha}_0 = \sigma^2 / n$.

(iii) $\mathrm{Cov}(\hat{\alpha}_0, \hat{\beta}_1) = 0$.

If, in addition, ε_i is normal, then $\hat{\alpha}_0$ is also normal.

Proof:

(i):

For the expectation, one has

$$\mathbb{E}\hat{\alpha}_0 = \overline{y} \qquad \text{(}\textbf{Theorem 2.1}\text{)}$$
$$= \frac{1}{n}\sum \mathbb{E}y_i \qquad \text{(Definition)}$$
$$= \frac{1}{n}\sum \big(\alpha_0 + \beta_1(x_i - \overline{x})\big) \qquad (\mathbb{E}\varepsilon_i = 0)$$
$$= \alpha_0. \qquad \text{(sum of variance is always 0)}$$

(ii):

For the variance, simple calculation yields

$$\mathrm{Var}\hat{\alpha}_0 = \mathrm{Var}\overline{y} \qquad \text{(}\textbf{Theorem 2.1}\text{)}$$
$$= \frac{1}{n^2}\sum \mathrm{Var} y_i \qquad \text{(Definition and property of Var)}$$
$$= \frac{1}{n^2} n\sigma^2 \qquad \text{(Variance provided by } \varepsilon_i\text{'s)}$$
$$= \sigma^2 / n.$$

(iii):

Finally, if ε_i are normal, then y_i are also normal, and since $\hat{\alpha}_0$ is the average of y_i, $\hat{\alpha}_0$ is also normal.

□

Therefore, $\hat{\alpha}_0$ is an unbiased and consistent estimator of α_0. Since $\hat{\beta}_0 = \hat{\alpha}_0 - \hat{\beta}_1\bar{x}$ and both $\hat{\alpha}_0$ and $\hat{\beta}_1$ are unbiased and consistent estimators, we can conclude the same result for $\hat{\beta}_0$.

Theorem 2.4:

Assume that the error terms of the simple linear regression model $y_i = \alpha_0 + \beta_1(x_i - \bar{x}) + \varepsilon_i$ has the property that $\mathbb{E}\varepsilon_i = 0$ and $\text{Var}\varepsilon_i = \sigma^2$. Then

(i) $\mathbb{E}\hat{\beta}_0 = \beta_0$.

(ii) $\text{Var}\hat{\beta}_0 = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$.

(iii) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \sigma^2\left(\frac{\bar{x}}{S_{xx}}\right)$.

In addition, if ε_i are normal, then $\hat{\beta}_0$ is also normal.

We state one last result to conclude this subsection.

Theorem 2.5:

$\hat{\sigma}^2 := \frac{1}{n-1}SSE \equiv \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \equiv \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is an unbiased estimator of σ^2. If the error terms are normal, then $\hat{\sigma}^2$ is independent from $\hat{\beta}_1$, $\hat{\alpha}_0$, and $\hat{\beta}_0$ and $(n-2)\hat{\sigma}^2/\sigma^2$ has the χ^2 distribution with $n-2$ degrees of freedom.

The reason we define $\hat{\sigma}^2 = \frac{1}{n-1}SSE$ is that as a consequence of **Theorem 2.4** (ii) we see that the original definition of S^2 as an estimator of σ^2 is not appropriate since y_i are no longer identically distributed. The fact that we have $n-2$ in the denominator instead of $n-1$ could be interpreted as we are now treating two parameters instead of one and so lost two degress of freedom.

We now summarize the properties of the Least-Squares estimators for simple linear regression to close this subsection.

Properties of LS Estimators:

(1) The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e. $\mathbb{E}\hat{\beta}_i = \beta_i$ for $i = 0{,}1$.

(2) $\text{Var}\hat{\beta}_0 = c_{00}\sigma^2$, where $c_{00} := \frac{\Sigma x_i}{nS_{xx}}$.

(3) $\text{Var}\hat{\beta}_1 = c_{11}\sigma^2$, where $c_{11} = \frac{1}{S_{xx}}$.

(4) $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$, where $c_{01} = \frac{-\bar{x}}{S_{xx}}$.

(5) An unbiased estimator of σ^2 is $S^2 = \frac{\text{SSE}}{(n-2)}$, where $\text{SSE} := S_{yy} - \hat{\beta}_1 S_{xy}$ and $S_{yy} := \sum(y_i - \bar{y})^2$.

If, in addition, the ε_I, for $i = 1,2,\cdots,n$ are normally distributed.

(6) Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.

(7) The random variable $\dfrac{(n-2)S^2}{\sigma^2}$ has a χ^2 distribution with degree of freedom $n-2$.

(8) The statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

2.3 Confidence Intervals and Hypothesis Tests for Coefficients

Once we know the variances of the parameters, it is easy to construct the confidence intervals. The procedure is essentially the same as what we did when we estimate the mean of a sample.

For example, a large sample two-sided confidence interval for the parameter β_1 can be written as follows

$$\left(\hat{\beta}_1 - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{\beta}_1 + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right), \tag{2.2}$$

where α is the confidence level.

If the sample is small, on the other hand, but we assume that the error terms are normal, we can use our previous theorems to come to conclusion that

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}$$

is a pivotal quantity (i.e. the distribution is dependent on all parameters) that has t distribution with $n-2$ degrees of freedom. In this case an appropriate confidence interval is

$$\left(\hat{\beta}_1 - t_{\alpha/2}^{(n-2)}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}, \hat{\beta}_1 + t_{\alpha/2}^{(n-2)}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right). \tag{2.3}$$

Similarly, if the null hypothesis is $\beta_1 = \beta_1^{(0)}$ we can form the test statistic as

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\hat{\sigma}/\sqrt{S_{xx}}} \tag{2.4}$$

and use the test statistic to test the null hypothesis against various alternative.

Summary:

> If the sample is large (e.g. $n > 30$) then T is distributed as a standard normal random variable. If the sample is small, then we rely on the assumption that ε_i have normal distribution and then T has the t distribution with degrees of freedom being $n-2$.

Similar procedures can be easily established for other parameters, that is for α_0 or β_0. We only need to use the appropriate variance of the estimator instead of the $\hat{\sigma}^2/S_{xx}$.

2.4 Statistical Inference for the Regression Mean

In applications we sometimes want to make some inferences about linera combinations of parameters. In this section we study a particular example of this problem. Suppose that we want to build the confidence interval for the regression mean of Y, when x is equal to some specific value, namely x^*,

$$\mathbb{E}(Y\,|\,x^*) = \beta_0 + \beta_1 x^*. \tag{2.5}$$

The natural estimator for this quantity is the predicted value given by

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

This estimator is unbiased since both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. In order to build the confidence interval, we also need to calculate its variance. It is more convenient to use the other form of the regression for this task:

$$\hat{y}^* = \hat{\alpha}_0 + \hat{\beta}_1(x^* - \bar{x}). \tag{2.6}$$

That is, centering x^* around its mean. Then

$$\begin{aligned}\mathrm{Var}\hat{y}^* &= \mathrm{Var}\hat{\alpha}_0 + (x^* - \bar{x})^2\mathrm{Var}\hat{\beta}_1 + 2(x^* - \bar{x})\mathrm{Cov}(\hat{\alpha}_0, \hat{\beta}_1) \qquad \text{(By (2.6))}\\ &= \sigma^2\Big(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\Big). \qquad (2.7)\end{aligned}$$

Using this information we can build the confidence interval for y^*. For example, if the sample size is large then the two-sided confidence interval with significance level α is

$$\hat{y}^* \pm z_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}},$$

Where $\hat{\sigma} = \sqrt{\dfrac{SSE}{n-2}}$ is the estimate for $\sigma = \mathrm{Var}\varepsilon_i$.

If the sample is small, on the other hand, but we assume that ε_i are normal, then we can use the t distribution with $n-2$ degrees of freedom and the confidence interval because

$$\hat{y}^* \pm t_{\alpha/2}^{(n-2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

And for testing hypothesis $H_0 : y^* = y_0$, we use the statistic

$$T = \frac{y^* - y_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})}{S_{xx}}}}.$$

2.5 Prediction Interval

When predicting Y we are often interested not in variation of our predictions $\hat{y}$ around the true regression mean but rather in variations of the actual quantities y around the true regression mean. The random quantity y has larger variation than $\hat{y}$ since in addition to uncertainty due to the error in parameter estimation it also includes the variation due to the error terms ε_i.

We define the prediction interval with confidence level $1 - \alpha$ as a random interval:

Definition: Prediction Interval

> The prediction interval with confidence level $1 - \alpha$ is given by the random interval (L, U) such that $\mathbb{P}(L \le y_i \le U) = 1 - \alpha$, where L and U are some statistics, so they msut be computable from data.

In order to construct the prediction interval we use the pivotal quantity technique and consider

$$T = \frac{y^* - \hat{y}^*}{SE(y^* - \hat{y}^*)},$$

where SE stands for “stand error”. Here y^* is a new observation which we try to predict and $\hat{y}^*$ is the prediction.

Note that

$$\begin{aligned} y^* - \hat{y}^* &= \beta_0 + \beta_1 x^* + \varepsilon^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x^* + \varepsilon^*. \end{aligned}$$

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1, we see that this quantity has expectation 0.

Moreover, if ε_i are normal then we see that $y^* - \hat{y}^*$ is also normal. What is the standard error of $y^* - \hat{y}^*$? Note that we have

$$\mathrm{Var}(y^* - \hat{y}^*) = \mathrm{Var}(\beta_0 + \beta_1 x^* + \varepsilon^* - \hat{y}^*) = \mathrm{Var}\varepsilon^* + \mathrm{Var}\hat{y}^*,$$

because the new error term ε^* is uncorrelated with the prediction $\hat{y}^*$. Indeed, the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ were estimated using the old error terms ε_i and x^* is not random.

We calculated the variance of $\hat{y}^*$ in the previous subsection, so we have

$$\begin{aligned} \mathrm{Var}(y^* - \bar{y}) &= \sigma^2 + \sigma^2 \Big(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \Big) \qquad \text{(by (2.7))} \\ &= \sigma^2 \Big(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \Big). \end{aligned}$$

It follows that

$$Z = \frac{y^* - \hat{y}^*}{\sigma \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

has the standard normal distribution.

It can be shown that if we use the estimator $\hat{\sigma} = \sqrt{\dfrac{SSE}{n-2}}$ instead of the unknown σ, then the quantity

$$T = \frac{y^* - \hat{y}^*}{\hat{\sigma} \sqrt{\frac{1 + [1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

has the t distribution with $n-2$ degrees of freedom.

So it follows that the prediction interval for y^* can be written as

$$\hat{y}^* \pm t_{\alpha/2}^{(n-2)} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2}^{(n-2)} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

The interpretation is that with probability $1 - \alpha$ the deviation of our prediction $\hat{y}^*$ from the actual realization of y^* will be smaller than the value

$$t_{\alpha/2}^{(n-2)}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}.$$

2.6 Correlation and R-Squared

Sometimes, x_i can be interpreted as observed values of some random quantity X. That is, we have n observations (x_i, y_i) sampled from the joint distribution of the random quantities X and Y. In this case, the coefficient β_1 in the regression $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ can be interpreted as a measure of dependence between X and Y.

On the other hand, we know that another measure of dependence between Y and X is the correlation coefficient given by

$$\rho := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}X\,\mathrm{Var}Y}}, \tag{2.8}$$

and we can estimate it as

$$R := \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{2.9}$$

Since $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ we see that we have the following relation between the estimates of correlation coefficient ρ and linear regression parameter β:

$$R = \beta_1\sqrt{\frac{S_{yy}}{S_{xx}}}. \tag{2.10}$$

So there is a clear relationship between these two measures of association.

The statistic r^2 (called R-squared) has another useful interpreation, which will be later generalized for multiple linear regression model. Namely, it measures the goodness of fit in the simple linear regression model.

Indeed, it is possible to derive the following useful formula.

$$\begin{aligned} SSE &:= \sum_i (y_i-\hat{y}_i)^2 = \sum_i \left(y_i-\bar{y}-\hat{\beta}_1(x_i-\bar{x})\right)^2 \\ &= \sum_i (y_i-\bar{y})^2 - 2\hat{\beta}_1\sum_i (y_i-\bar{y})(x_i-\bar{x}) + \hat{\beta}_1^2\sum_i (x_i-\bar{x})^2 \\ &= S_{yy} - \hat{\beta}_1^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

Now $S_{yy} = \sum_i (y_i-\bar{y})^2$ can be thought as the variation in the response variable if no explanatory variable is used, and SSE is the variation in the response after the explanatory variable is used. So the difference is the reduction in the variation due to the explanatory variable X. In particular, one has

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \tag{2.11}$$

Being the reduction measured in percentage terms. To summarize, R^2 is the proportion of response variable variation that is explained by the explanatory variable X once it is brought into observation.

3. Multiple Linear Regression

A more general version of linear regression reads:

$$y = \beta_0 + \beta_1 x^{(1)} + \cdots + \beta_p x^{(p)} + \varepsilon,$$

where we have p explanatory variables. In fact this stands for n separate equations, one for each observation

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \cdots + \beta_p x_i^{(p)} + \varepsilon_i.$$

Once we treat

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix},$$

and the vectors of coefficients and error terms

$$\beta = \left[\beta_0, \beta_1, \cdots, \beta_p\right]^T \text{ and } \varepsilon = [\varepsilon_1, \varepsilon_2 \cdots, \varepsilon_n]^T.$$

Then we can write our model as

$$y = X\beta + \varepsilon. \tag{3.1}$$

The sum of squared errors can also be written very simply in the matrix notation:

$$SSE\hat{\beta} = \left[y - X\hat{\beta}\right]^T \left[y - X\hat{\beta}\right]^T. \tag{3.2}$$

To summarize $SSE\hat{\beta}$, we need to write the first order conditions, which can also be written in matrix form. Namely, for each $j = 1, \cdots, p$ we have

$$\frac{\partial SSE(\hat{\beta})}{\partial \beta_j} = -2\sum_{i=1}^{n} x_{ij}\left\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip})\right\} = 0.$$

If we stack these $p+1$ equations together, we obtain the matrix form of these system of equations

$$\frac{\partial SSE(\hat{\beta})}{\partial \hat{\beta}} = -2X^T[y - X\hat{\beta}] = -2X^T y + 2X^T X\hat{\beta} = 0.$$

Or, re-arranging the terms and simplifying

$$X^T X\hat{\beta} = X^T y.$$

This system of $(p+1)$ equations in $p+1$ unknowns $\hat{\beta}_i$ is called the normal equations. In matrix form, its solution can be written as

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y.$$

3.1 Properties of LS Estimators

Recall that $\hat{\beta}$ is a p-vector $[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p]^T$ and $\hat{\beta} = (X^T X)^{-1} X^T y$. We have

Theorem 5.6: Expectaion and Variance of β

The LS estimator of β is unbiased, i.e. $\mathbb{E}\hat{\beta} = \beta$. Its variance matrix is the $(p+1) \times (p+1)$ matrix $\text{Var}\hat{\beta} = \sigma(X^T X)^{-1}$.

If, in addition, $\varepsilon_i \sim N(0,\sigma)$, then it can be shown that $\hat{\beta}$ is the multivariate normal with mean β and variance $\sigma^2(X^TX)^{-1}$.

Now it is clear how to build confidence intervals and test the hypothesis for the parameters β_i. We simply notice that

$$\mathrm{Var}\beta_i = \sigma^2 c_{ii},$$

where c_{ii} is the ith element on the main diagonal of the matrix $(X^TX)^{-1}$ given by

$$c_{ii} = \left[(X^TX)^{-1}\right]_{ii}. \tag{3.3}$$

So if σ^2 is known, then the confidence interval for β_i is

$$\hat{\beta}_i \pm z_{\alpha/2}\sigma\sqrt{c}_{ii}.$$

In practice, σ^2 is not known and have to be estimated from data. We can do it using SSE, which is defined similarly to the case of the simple linear regression:

$$SSE := \sum_i (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ are fitted values for the response variable.

Theorem 3.1:

$$\hat{\sigma}^2 := \frac{SSE}{n-p-1} \text{ is an unbiased estimator of } \sigma^2.$$

Moreover, if ε_i are independent normal random variables and $\varepsilon_i \sim N(0,\sigma^2)$. It follows that

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{c}_{ii}}$$

has the t distribution with $n-p-1$ degrees of freedom. Therefore, in this case the confidence interval is given by

$$\hat{\beta}_i \pm t_{\alpha/2}^{(n-p-1)}\hat{\sigma}\sqrt{c_{ii}}. \tag{3.4}$$

3.2 Confidence Interval

If we have p parameters we might be interested in finding the confidence interval for the linear combination

$$a_0\beta_0 + a_1\beta_1 + \cdots + a_p\beta_p,$$

which can be written as $a^T\beta$ where a is the column vector. The confidence interval should be centered at $a^T\hat{\beta}$ and the main question is about the standard error of this estimator.

Since $\mathrm{Var}\sum_{i=0}^{p} a_i\hat{\beta}_i = \sum_{i,j} a_ia_j\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j)$, we have

$$\mathrm{Var}a^T\hat{\beta} = a^T\mathrm{Var}\hat{\beta}a = \sigma^2 a^T(X^TX)^{-1}a, \tag{3.5}$$

where we used the formula for the variance-covariance matrix of the estimator β.

It follows that the confidence interval for $a^T\beta$ can be written as

$$a_T\hat{\beta} \pm z_{\alpha/2}\sigma\sqrt{a^T(X^TX)^{-1}a}. \tag{3.6}$$

Since σ is unknown, we susbstitute it with its estimator. In the case of normal errors, it gives the following confidence interval

$$a^T\hat{\beta} \pm t_{\alpha/2}^{(n-p-1)}\hat{\sigma}\sqrt{a^T(X^TX)^{-1}a}.$$

3.3 Prediction

Suppose that we obtained a new observation with predictor variables $x_*^1, x_*^2, \cdots, x_*^p$ and we want to predict the response variables y_*.

The natural predicor is given by

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*^1 + \cdots + \hat{\beta}_p x_*^p = x_*^T\hat{\beta},$$

where x_* is the column vector $[1, x_*^1, \cdots, x_*^p]^T$.

This expected value of this predictor equals the regression mean $\mathbb{E}y_*$,

$$\mathbb{E}\hat{y}_* = x_*^T\mathbb{E}\hat{\beta} = x_*^T\beta.$$

Let us define the prediction error as the difference between the prediction and the realized response variable, namely,

$$\varepsilon_* = y_* - \hat{y}_*.$$

Then the expected value of error is zero and it is easy to comput its variance

$$\mathrm{Var}\varepsilon_* = \sigma^2 + \sigma^2 x_*^T(X^TX)^{-1}x_*.$$

This allows us to write the prediction interval

$$x_*^T\hat{\beta} \pm t_{\alpha/2}^{(n-p-1)}\hat{\sigma}\sqrt{1 + x_*^T(X^TX)^{-1}x_*}.$$

4. Some Inference Results

In this section we introduce some inference results, we start with the inference concerning the parameters β_i.

Test of Hypothesis for β_i:

$H_0 : \beta_i = \beta_{i0}$.

$$H_a : \begin{cases} \beta_i > \beta_{i0} \text{ upper-tail rejection region} \\ \beta_i < \beta_{i0} \text{ lower-tail rejection region}. \\ \beta_i \neq \beta_{i0} \text{ two-tailed rejection region} \end{cases}$$

Test statistic $T = \dfrac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha \text{ upper-tail alternative} \\ t < -t_\alpha \text{ lower-tail alternative} \\ |t| > t_{\alpha/2} \text{ two-tailed alternative} \end{cases},$$

where $c_{00} = \dfrac{\Sigma x_i^2}{nS_{xx}}$ and $c_{11} = \dfrac{1}{S_{xx}}$. Notice that t_α is based on $n-2$ degrees of freedom.

A $(1-\alpha)$ 100% Confidence Interval for β_i:

$\hat{\beta}_i \pm t_{\alpha/2} S\sqrt{c_{ii}}$, where $c_{00} = \dfrac{\Sigma x_i^2}{nS_{xx}}$ and $c_{11} = \dfrac{1}{S_{xx}}$.

As for the inferences concerning the linear functions of the model parameters, i.e. the simple linear regression, $\theta = a_0\beta_0 + a_1\beta_1$, we have the following results.

A Test for $\theta = a_0\beta_0 + a_1\beta_1$:

$H_0 : \theta = \theta_0$,

$$H_a : \begin{cases} \theta > \theta_0, \\ \theta < \theta_0, \\ \theta \neq \theta_0 \end{cases}$$

Test statistic: $$T = \frac{\hat{\theta} - \theta_0}{S\sqrt{\dfrac{a_0^2 \frac{\Sigma x_i^2}{n} + a_1^2 - 2a_0a_1\bar{x}}{S_{xx}}}}.$$

Rejection Region: $$\begin{cases} t > t_\alpha, \\ t < -t_\alpha, \\ |t| > t_{\alpha/2} \end{cases}$$

Here t_α and $t_{\alpha/2}$ are based on $n-2$ degrees of freedom.

The corresponding $(1-\alpha)$ 100% confidence interval for $\theta = a_0\beta_0 + a_1\beta_1$ is given by

A $(1-\alpha)$ **100% Confidence Interval for** $\theta = a_0\beta_0 + a_1\beta_1$:

$\hat{\theta} \pm t_{\alpha/2} S\sqrt{\dfrac{a_0^2\frac{\Sigma x_i^2}{n} + a_1^2 - 2a_0a_1\bar{x}}{S_{xx}}}$, where $t_{\alpha/2}$ is based on $n-2$ degrees of freedom.

One useful application of the hypothesis-testing and confidence interval techniques just presented is to the problem of estimating $\mathbb{E}Y$, the mean of Y, for a fixed value of the independent variable x. In particular, if x^* denotes a specific value of x that is of interest, then

$$\mathbb{E}Y = \beta_0 + \beta_1 x^*. \tag{4.1}$$

Notice that $\mathbb{E}Y$ is a special case of $a_0\beta_0 + a_1\beta_1$, with $a_0 = 1$ and $a_1 = x^*$. Thus, an inference about $\mathbb{E}Y$ when $x = x^*$ can be made by using the techniques developed earlier for general linear combinations of the β's.

A $(\alpha - 1)$ **100% Confidence Interval for** $\mathbb{E}Y = \beta_0 + \beta_1 x^*$:

$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{S_{xx}}}$, where $t_{\alpha/2}$ is based on $n-2$ degrees of freedom.

Assume that a linear model of the form $Y = \beta_0 + \beta_1 x + \varepsilon$ is in the interest of our inference, then:

A $(1-\alpha)$ **100% Prediction Interval for** Y **when** $x = x^*$:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}.$$

Reference:

[1]: George Casella, Roger L. Berger, *Statistical Inference, Second Edition*, Wadsworth Group, pp. 417-451.

[2]: Vladislav Kargin, *Lecture Notes for Math 448 Statistics*, available online.

[3]: John E. Freund’s, *Mathematical Statistics with Applications*, Eighth Edition, Pearson Education Limited 2014, pp. 563-633