北京交通大學

Estimation

Chapter 6 of Probability and Statistics

Yiping Cheng
ypcheng@bjtu.edu.cn
19-Nov-2019

School of Electronic and Information Engineering
Beijing Jiaotong University

Chapter Contents

1. Population and Sample
2. The Concept of Statistics
3. Method-of-Moments Estimation
4. Maximum Likelihood Estimation
5. Properties of Estimators
6. Distributions Derived from the Normal Distribution
7. Confidence Intervals for Normal Distribution Parameters

Population and Sample

Population

A population consists of all the individuals being studied in a statistical inference problem. Imagine that there is a city with 10000 people. The city hall wants to have the data about the health of its citizens, but they do not have the resources needed for physical examination of all the citizens. So they decide to use sampling. They randomly choose 100 people to conduct physical examination, and their data, such as height, weight, blood pressure, eyesight, etc. are collected. The population in this example is the 10000 people, or mathematically, the set $\{0, \dots, 9999\}$, if each citizen is assigned a distinct number in $\{0, \dots, 9999\}$. Normally we use simple probability, i.e., for each i in $\{0, \dots, 9999\}$, the basic event $\{i\}$ is assigned probability $\frac{1}{10000}$.

The health indices of the citizens can then be modelled as random variables over the population. For example, we can use $H(\omega)$ to denote the height of citizen ω, $W(\omega)$ to denote the weight of citizen ω, etc.

Practically, the population is always finite and simple. But theoretically the population can be infinite and even continuous.

Random sample

We just mentioned that 100 people are to be sampled. The sampling method is assumed to be simple random, i.e. each person has identical chance to be selected and the selections are mutually independent. Since we can only discuss independence on the same probability space, the selected persons must be the outcome of a single random experiment. Denote the population by Π and the number of selections by n (in this example $\Pi = \{0, \dots, 9999\}$ and $n = 100$. Then the sample space of the experiment is $\Pi^n := \underbrace{\Pi \times \cdots \times \Pi}_{n \text{ times}}$. For example, if the random experiment is finished and the outcome is $e = \underbrace{(934, 8031, 2507, \dots)}_{100 \text{ entries}}$, then that means that the person numbered 934 is the first person selected, the person numbered 8031 is the second person selected, etc.

Denote the i-th component of e by e_i. Now for every random variable X **over the population** Π, we can define a set of random variables $X_1, \dots, X_n$ **over the experiment sample space** Π^n as follows:

$$X_i(e) = X(e_i) \text{ for } e \in \Pi^n,\ i = 1, \dots, n.$$

Random sample (continued)

It can be proved that if the sampling is simple random, then $X_1, \dots, X_n$ **are mutually independent and all have the same distribution as** X.

So in the sequel, we will say the following:

Definition (random sample)

We say $X_1, \dots, X_n$ are a random sample from X if $X_1, \dots, X_n$ are mutually independent and all have the same distribution as X. We also say $X_1, \dots, X_n$ are a random sample from distribution $F(x)$ if $X_1, \dots, X_n$ are mutually independent and all have distribution $F(x)$.

The obtained data $(x_1, \dots, x_n) = (X_1(e), \dots, X_n(e))$ is called an observation, or observed data, of $X_1, \dots, X_n$.

Remark: We have seen that $X_1, \dots, X_n$ do not live in the same probability space as X, so it makes no sense to talk about whether they are independent, the covariance of X_i and X, and other similar concepts.

Joint distribution of the random sample

If the population random variable has CDF $F(x)$, then the joint CDF of random sample $X_1, \ldots, X_n$ from X is given by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

If X is discrete with PF $p(x)$, then the joint PF of its random sample $X_1, \ldots, X_n$ is given by

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i).$$

If X is continuous with PDF $f(x)$, then the joint PDF of its random sample $X_1, \ldots, X_n$ is given by

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Introduction to statistical inference

Statistical inference is the process of deducing properties of a population probability distribution by analysis of data. Here is a partial list of statistical inference examples:

1. **One Population, Parameter Point Estimation.** An inspector wants to find the defective rate of the products of a company. The number of products of this company is too huge. So she decides to use sampling. She randomly chooses 100 samples from the products and inspects every one of them, and then she has the defective rate of the samples. She thinks the defective rate of the samples is a good approximation of the total defective rate.
2. **One Population, Parameter Point Estimation.** A company that sells electronic components wants to know how long their components are likely to last. They choose to use the family of exponential distributions to model the length of time (in years) from when a component is put into service until it fails. Then they collect data of life length on a sample of components that have been used under typical conditions. Based on the data, the parameter of the exponential distribution is estimated.

3. **One Population, Nonparametric Inference.** The state government wants to have the distribution of annual income of its one million residents. They randomly select a thousand residents and they take the annual income distribution of the selected residents as an approximation of the desired distribution.
4. **One Population, Parameter Interval Estimation.** A TV advertising company wants to find an interval that contains the average time per day of a person in a region spent in watching TV, with a level of confidence. They assume this time per day in watching TV has a normal distribution with unknown mean and variance. They have called a sample of people and collected the data.
5. **Two Populations, Hypothesis Testing.** A female employee sues her employer for discrimination in salary against female employees. The court then invites a statistician to conduct a relevant research. This involves two populations: all male employees and all female employees of the same employer. The statistician's task is to test the hypothesis: the mean salary of the male employees is equal to the mean salary of the female employees, based on sampled data.

Elements of statistical inference problems

Normally a statistical inference problem consists of the following three elements:

1. A **model of the population distribution**. Here by *population distribution* we mean the distribution of the concerned random variable X defined over the population. This distribution is assumed to involve a finite number of parameters (parametric), or no parameters but actually infinitely many parameters (nonparametric).
2. A set of **sample data** $x_1, \ldots, x_n$, which are an observation of random sample $X_1, \ldots, X_n$.
3. The **purpose of inference**, such as parameter point estimation, parameter interval estimation, nonparametric distribution estimation, hypothesis testing, etc.

The conventional approach to statistical inference

We have seen the model of the population distribution involves unknown parameters, whether finitely many or infinitely many. The subsequent statistical inference is therefore to extract information about the parameters from the data.

The conventional approach to statistical inference is to treat the parameters as **unknown but fixed, deterministic** quantities.

In this approach, the data are modelled as a random sample from the distribution, i.e. a sequence of independent random variables $X_1, \ldots, X_n$ that each have the same distribution as the population distribution.

The Concept of Statistics

Statistics and estimators

Definition (statistic)

Let $X_1, \ldots, X_n$ be a random sample from a distribution. A random variable $g(X_1, \ldots, X_n)$ is called a **statistic** if it does not involve any unknown parameters.

Note: if $(x_1, \ldots, x_n)$ is an observation of $(X_1, \ldots, X_n)$, then $g(x_1, \ldots, x_n)$ is an observation of $g(X_1, \ldots, X_n)$.

Example: Suppose $X_1, \ldots, X_n$ are a random sample from distribution $N(\mu, \sigma^2)$, where μ is known but σ^2 is unknown, then $\min(X_1, \ldots, X_n)$, $(X_1 + X_2)/2$, and $\frac{X_1+\cdots+X_n}{n} - \mu$ are statistics, but $\frac{(X_1+X_2)^2}{\sigma^2}$ is not a statistic.

Definition (estimator)

If a statistic is supposed close to a parameter of the distribution, then it is called an estimator of that parameter. An observation of an estimator is called an estimate.

Sampling distributions

Definition

The distribution of a statistic is called a **sampling distribution**.

Remark: The name "sampling distribution" comes from the fact that a statistic depends on a random sample and so the distribution of a statistic is derived from the distribution of the sample.

Some common statistics

Name of statistic	Definition
sample mean	$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
sample variance	$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$
sample standard deviation	$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$
sample k-th moment	$A_k = \frac{\sum_{i=1}^{n} X_i^k}{n}$
sample k-th central moment	$B_k = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^k}{n}$

Observations of the common statistics

Name of observation of statistic	Definition
sample mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
sample variance	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
sample standard deviation	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
sample k-th moment	$a_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$
sample k-th central moment	$b_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n}$

The corresponding population parameters

Name of parameter	Definition
population mean	$\mu = E(X)$
population variance	$\sigma^2 = \mathrm{Var}(X)$
population standard deviation	$\sigma = \sqrt{\mathrm{Var}(X)}$
population k-th moment	$\mu_k = E(X^k)$
population k-th central moment	$E[(X - \mu)^k]$

Sample moments converge in probability to population moments

Theorem

Suppose that $X_1, X_2, \ldots$ form an infinite-size random sample from X whose k-th moment μ_k is finite. For every n, let $A_{k,n}$ be the k-th sample moment of random sample $X_1 \ldots, X_n$. Then

$$A_{k,n} \xrightarrow{P} \mu_k.$$

Proof.

$X_1^k, X_2^k, \ldots$ form an infinite-size random sample from the distribution of X^k whose mean is μ_k, thus by the law of large numbers, $A_{k,n} \xrightarrow{P} \mu_k$. □

Order statistics

Definition (order statistic)

Given vector $(x_1, x_2, \ldots, x_n)$, reorder it to obtain $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ so that $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, then each $x_{(k)}$ is called the k-th ordered number of $x_1, x_2, \ldots, x_n$.

Suppose that $X_1, X_2, \ldots, X_n$ are a random sample from some distribution. Let $X_{(k)}$ be the k-th ordered number of $X_1, X_2, \ldots, X_n$, then $X_{(k)}$ is called the ***k*-th order statistic**.

Remark: The first order statistic (or smallest order statistic) is the minimum of the sample, that is, $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$. The n-th order statistic (or largest order statistic) is the maximum, that is $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$. The statistical range is defined as $X_{(n)} - X_{(1)}$.

Empirical distribution functions

Definition (empirical distribution function)

Let $x_1, \ldots, x_n$ be an observation of a random sample $X_1, \ldots, X_n$. For each number x, define the value $F_n(x)$ as the proportion of observed values in the sample that are less than or equal to x, i.e., if exactly k of the observed values in the sample are less than or equal to x, then $F_n(x) = \frac{k}{n}$. More formally,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{x_i \le x}.$$

The function $F_n(x)$ defined in this way is called an empirical distribution function.

Remark: Different observations of the random sample give rise to different empirical distribution functions. And for any fixed x, $F_n(x)$ is a statistic; it converges in probability to $F(x)$ (the population CDF) as $n \to \infty$.

Method-of-Moments Estimation

The method of moments

Suppose that the population distribution has r unknown parameters which constitute a vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$. We want to estimate them. The method of moments proceeds as follows:

1. Compute the 1-, $\dots$, r-th moments of the population distribution, $\mu_1, \dots, \mu_r$ as a function of $\theta_1, \dots, \theta_r$, i.e. find

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_r \end{bmatrix} = \mathcal{M}(\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_r \end{bmatrix}).$$

2. Find the inverse function $\mathcal{M}^{-1}$, i.e. find

$$\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_r \end{bmatrix} = \mathcal{M}^{-1}(\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_r \end{bmatrix}).$$

3. Let

$$\hat{\theta} = \mathcal{M}^{-1}\left(\begin{bmatrix} A_1 \\ \vdots \\ A_r \end{bmatrix}\right)$$

where $A_1, \ldots, A_r$ are the 1-, $\ldots$, r-th sample moments of random sample $X_1, \ldots, X_n$. Then $\hat{\theta}$ is called a method-of-moments estimator of θ.

The theoretical basis of method-of-moments estimation is: Since $A_1, \ldots, A_r$ converges in probability to $\mu_1, \ldots, \mu_r$, and if $\mathcal{M}^{-1}$ is a continuous mapping (which is usually the case), then $\hat{\theta}$ converges in probability to θ.

If some of the first r moments do not depend on θ, we will have to discard them and use higher order moments.

Method-of-moments estimator of variance

For any distribution with finite mean μ and finite variance σ^2, we have

$$\mu = \mu_1, \quad \sigma^2 = \mu_2 - \mu_1^2.$$

Therefore the method-of-moments estimator of σ^2 is

$$\hat{\sigma^2} = A_2 - A_1^2 = \frac{\sum_{i=1}^n X_i^2}{n} - \overline{X}^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n} \text{ which is } B_2.$$

Proof of the above last equality.

It is equivalent to show $\sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n (X_i - \overline{X})^2$.

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \sum_{i=1}^n X_i^2 - 2\overline{X}\sum_{i=1}^n X_i + n\overline{X}^2$$

$$= \sum_{i=1}^n X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2.$$

□

Method-of-moments estimator of uniform distribution parameters

Example

Let $X_1, \ldots, X_n$ form a random sample from the uniform distribution $U(a, b)$, where $a < b$ are unknown parameters. Find the method-of-moments estimator of a and b.

Solution: We have $\mu_1 = \frac{a+b}{2}$, $\mu_2 = E(X^2) = E^2(X) + \mathrm{Var}(X) = \mu_1^2 + \frac{(b-a)^2}{12}$, and thus $\frac{(b-a)^2}{12} = \mu_2 - \mu_1^2$. Therefore, $a = \mu_1 - \sqrt{3(\mu_2 - \mu_1^2)}$ and $b = \mu_1 + \sqrt{3(\mu_2 - \mu_1^2)}$. Hence, the method-of-moment estimator of (a, b) is

$$\hat{a} = A_1 - \sqrt{3(A_2 - A_1^2)} = \overline{X} - \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{b} = A_1 + \sqrt{3(A_2 - A_1^2)} = \overline{X} + \sqrt{\frac{3}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Maximum Likelihood Estimation

Basic idea of maximum likelihood estimation

Suppose that the population distribution has r unknown parameters which constitute a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$. We want to estimate them. The basic idea behind the method of maximum likelihood estimation is as follows:

Maximum likelihood estimation chooses the estimate of θ as the value for the parameter that makes the observed data most probable.

The likelihood function

Definition

The joint PF or joint PDF $f_n(x_1, \dots, x_n; \theta)$ of the random sample at its particular observation $x_1, \dots, x_n$, regarded as a function of the parameter θ, is called the **likelihood function**. This is denoted by

$$L(\theta; x_1, \dots, x_n) = f_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

or if there will be no confusion arising from the omission, simply

$$L(\theta) = f_n(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Remark: Observed values $(x_1, \dots, x_n)$ are primary independent variables in the joint PF/PDF, but are secondary independent variables in the likelihood function; whereas θ is secondary independent variable in the joint PF/PDF, but is primary independent variable in the likelihood function.

Maximum likelihood estimators

Let $\mathbf{\Theta}$ be the parameter space, i.e. the set of all possible values of the parameter(s).

Definition [maximum likelihood estimate]: Given a particular sample data $x_1, ..., x_n$, let $\hat{\theta}$ be a value of $\theta \in \mathbf{\Theta}$ for which the likelihood function $L(\theta; x_1, \ldots, x_n)$ is a maximum, i.e.

$$\hat{\theta} = \arg\max_{\theta \in \mathbf{\Theta}} L(\theta; x_1, \ldots, x_n).$$

Then $\hat{\theta}$ is called a **maximum likelihood estimate** of θ for this sample data.

Definition [maximum likelihood estimator]: For each possible sample data $x_1, \ldots, x_n$, let $\delta(x_1, \ldots, x_n) \in \mathbf{\Theta}$ be a maximum likelihood estimate of θ for this sample data. Let

$$\hat{\theta} = \delta(X_1, \ldots, X_n).$$

Then $\hat{\theta}$ is called a **maximum likelihood estimator** (M.L.E.) of θ.

MLE example 1

Lifetimes of Electronic Components. Choose a sample of size 3 from the exponential distribution with parameter $\lambda > 0$, the observed data are $(x_1, x_2, x_3) = (3, 1.5, 2.1)$. The PDF of the population distribution is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0, \\ 0 & \text{for } x \le 0. \end{cases}$$

The likelihood function then is

$$L(\lambda) = f(x_1;\lambda)f(x_2;\lambda)f(x_3;\lambda) = \lambda^3 e^{-\lambda(x_1+x_2+x_3)} = \lambda^3 e^{-6.6\lambda}.$$

Since log is an increasing function, the value of λ that maximizes $L(\lambda)$ will be the same as the value of λ that maximizes $\ln L(\lambda)$, which is

$$\ln L(\lambda) = 3\ln\lambda - 6.6\lambda.$$

Its derivative is

$$\frac{d\ln L(\lambda)}{d\lambda} = \frac{3}{\lambda} - 6.6.$$

Setting the derivative to 0, the solution for λ is $3/6.6 = 0.455$. Since $\frac{d^2 \ln L(\lambda)}{d\lambda^2} = -\frac{3}{\lambda^2} < 0$, the solution is a maximizer. Therefore the maximum likelihood estimate of λ is 0.455.

MLE example 2

Suppose that the random variables $X_1, \ldots, X_n$ form a random sample from the Bernoulli distribution with parameter p, which is unknown ($0 \leq p \leq 1$). The PF of the population distribution is

$$f(x;p) = \begin{cases} p & \text{for } x = 1, \\ 1-p & \text{for } x = 0. \end{cases}$$

For observed values $x_1, ..., x_n$, where each x_i is either 0 or 1, the likelihood function is

$$L(p) = \prod_{i=1}^{n} f(x_i;p).$$

1. If $\sum_{i=1}^{n} x_i = 0$, then $L(p) = (1-p)^n$, and L achieves its maximum at $0 = \bar{x}$.
2. If $\sum_{i=1}^{n} x_i = n$, then $L(p) = p^n$, and L achieves its maximum at $1 = \bar{x}$.

3. Suppose $\sum_{i=1}^{n} x_i \notin \{0, n\}$. If $0 < p < 1$, then

$$f(x; p) = p^x (1-p)^{1-x}.$$

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = p^{\sum_{i=1}^{n} x_i} \cdot (1-p)^{n - \sum_{i=1}^{n} x_i}.$$

$$\ln L(p) = \sum_{i=1}^{n} x_i \ln p + [n - \sum_{i=1}^{n} x_i] \ln(1-p).$$

$$\frac{d \ln L(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}.$$

Setting $\frac{d \ln L(p)}{dp} = 0$, the solution is $p = \bar{x}$. Since $\frac{d^2 \ln L(p)}{dp^2} = -\frac{\sum_{i=1}^{n} x_i}{p^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-p)^2} < 0$, L achieves its local maximum at $p = \bar{x}$. Since $L(0) = 0$ and $L(1) = 0$ (why? please check), $p = \bar{x}$ is also a global maximizer of $L(p)$.

Therefore in all cases, the M.L.E. of p is $\overline{X}$.

MLE example 3

Suppose that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ is unknown and the variance σ^2 is known. For observed values $x_1, \ldots, x_n$, the likelihood function is

$$L(\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

$L(\mu)$ will be maximized by the value of μ that minimizes

$$Q(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2$$

We see that Q is a quadratic in μ with positive coefficient on μ^2. The minimizer of Q is the solution of the following equation:

$$\frac{dQ(\mu)}{d\mu} = -2\sum_{i=1}^{n} x_i + 2n\mu = 0,$$

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Therefore the M.L.E. of μ is $\hat{\mu} = \overline{X}$.

MLE example 4

Suppose again that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ and the variance τ are both unknown.

MLE example 4

Suppose again that $X_1, \ldots, X_n$ form a random sample from a normal distribution for which the mean μ and the variance τ are both unknown. For observed values $x_1, \ldots, x_n$, the likelihood function is

$$L(\mu, \tau) = \frac{1}{(2\pi\tau)^{n/2}} \exp\left[-\frac{1}{2\tau}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

The parameter is $\theta = (\mu, \tau)$, where $\tau > 0$.

$$\ln L(\mu, \tau) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\tau - \frac{1}{2\tau}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The maximizer of $L(\mu, \tau)$ is the solution of the following equation:

$$\begin{cases} \frac{\partial \ln L}{\partial \mu} = \frac{1}{\tau}\sum_{i=1}^{n}(x_i - \mu) = 0 \\ \frac{\partial \ln L}{\partial \tau} = -\frac{n}{2\tau} + \frac{1}{2\tau^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases},$$

$$\mu = \bar{x}, \quad \tau = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Thus the M.L.E. of θ is $\hat{\theta} = (\hat{\mu}, \hat{\tau}) = (\overline{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)$.

MLE example 5

Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where $\theta > 0$.

MLE example 5

Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where $\theta > 0$.

The PDF of the population distribution is

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function for sample data $x_1, \ldots, x_n$ is

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n} & \text{for } 0 \le \min(x_1, \ldots, x_n) \le \max(x_1, \ldots, x_n) \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

It can be seen that $\theta = \max(x_1, \ldots, x_n)$ maximizes $L(\theta)$, hence the M.L.E. of θ is $\hat{\theta} = \max(X_1, ..., X_n)$.

Properties of maximum likelihood estimators

Theorem [M.L.E. invariance against parameter transformations]: If $\hat{\theta}$ is the M.L.E. of θ and if g is a one-to-one function, then $g(\hat{\theta})$ is the M.L.E. of $g(\theta)$.

Note: The method-of-moments estimators also have this property.

Theorem [M.L.E. invariance against data transformations]: Suppose that $X_1, \ldots, X_n$ form a random sample from distribution $F(x; \theta)$. Denote the M.L.E. of θ based on this random sample by $\hat{\theta}_F$. Let r be a **one-to-one function that does not depend on θ**, and let $Y_i = r(X_i)$ for $i = 1, \ldots, n$. Then $Y_1, \ldots, Y_n$ form a random sample from distribution $G(y; \theta)$. Denote the M.L.E. of θ based on this random sample by $\hat{\theta}_G$. Then $\hat{\theta}_F = \hat{\theta}_G$.

Note: The method-of-moments estimators do not have this property, so this property can be considered as a unique property of M.L.E.s.

Properties of Estimators

Performance citeria for estimators

1. Consistency
2. Bias and unbiasedness
3. Efficiency

Consistency of estimators

Definition [consistent estimator]: A sequence of estimators $\hat{\theta} = r(X_1, \ldots, X_n)$ of a parameter θ is said to be a consistent sequence of estimators (or informally, a consistent estimator), if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1.$$

Consistency means that the estimator converges to the parameter in probability. It is **required** for an estimator to be practically usable.

Example

Let $X_1, \ldots, X_n$ be a random sample from any distribution with finite mean μ. Then according to the law of large numbers, the sample mean

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$$

is a consistent estimator of population mean μ.

Bias and unbiasedness of estimators

Definition [bias, unbiasedness of an estimator]: An estimator $\hat{\theta} = r(X_1, \ldots, X_n)$ is said to be an **unbiased estimator** of θ, if $E(\hat{\theta}) = \theta$. An estimator that is not unbiased is called a biased estimator. The bias of an estimator $\hat{\theta}$ of θ, is $E(\hat{\theta}) - \theta$.

Example

Let $X_1, \ldots, X_n$ be a random sample from any distribution with finite mean μ and variance σ^2. The sample mean $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ is an unbiased estimator of μ, since $E(\overline{X}) = \mu$. However, $\overline{X}^2$ is a biased estimator of μ^2 unless $\sigma^2 = 0$, because

$$E[\overline{X}^2] = \frac{\sum_{i=1}^{n} E(X_i^2) + \sum_{1 \le i \ne j \le n} E(X_i X_j)}{n^2} = \frac{n(\mu^2 + \sigma^2) + n(n-1)\mu^2}{n^2} = \mu^2 + \frac{\sigma^2}{n}.$$

But this sequence of estimators $\overline{X}^2$ is asymptotically unbiased, as $\frac{\sigma^2}{n} \to 0$.

Bias of B_2 as estimator of variance

As previously derived, the method-of-moments estimator of variance is

$$B_2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2.$$

It follows that

$$E(B_2) = \frac{\sum_{i=1}^{n} E(X_i^2)}{n} - E(\overline{X}^2)$$

$$= (\mu^2 + \sigma^2) - (\mu^2 + \frac{\sigma^2}{n}) = \frac{n-1}{n}\sigma^2.$$

Hence B_2 **is biased estimator of variance**. Correcting for this bias yields the unbiased sample variance (or simply just sample variance)

$$S^2 = \frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}.$$

Efficiency of estimators

Efficiency is usually only considered for unbiased estimators. A precise definition of efficiency is beyond this course. However, it can be said that **the smaller is the variance of an estimator, the higher is its efficiency**.

Example

Let $X_1, \ldots, X_n$ be a random sample from a distribution with finite mean μ and variance σ^2. We now have known that $\overline{X}$ is a consistent and unbiased estimator of μ. Now let us define $\tilde{X} = \frac{\sum_{i=1}^{[n/2]} X_{2i}}{[n/2]}$. We can also verify that $\tilde{X}$ is also a consistent and unbiased estimator of μ. But the two estimators have different variances:

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} < \mathrm{Var}(\tilde{X}) = \frac{\sigma^2}{[n/2]}.$$

Therefore $\overline{X}$ has higher efficiency than $\tilde{X}$.

Mean square error of an estimator

The **mean square error** of an estimator $\hat{\theta}$ of θ is

$$MSE = E[(\hat{\theta} - \theta)^2].$$

We have MSE = variance + bias^2, i.e.

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

Proof.

Let $t = E(\hat{\theta})$. Then

$$E[(\hat{\theta} - \theta)^2] = E[[(\hat{\theta} - t) + (t - \theta)]^2]$$

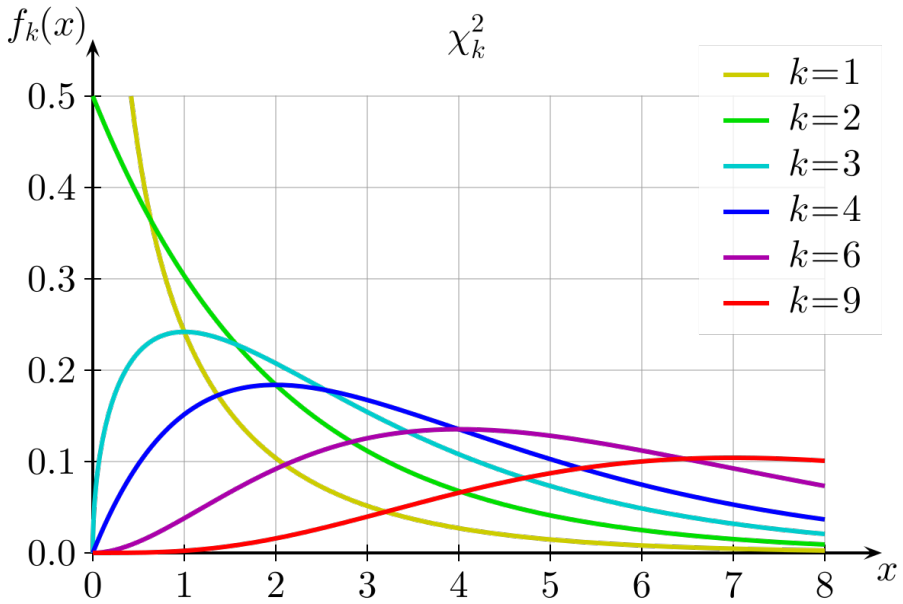
$$= E[(\hat{\theta} - t)^2] + (t - \theta)^2 + 2(t - \theta)\underbrace{E(\hat{\theta} - t)}_{=0} = \mathrm{Var}(\hat{\theta}) + \underbrace{(t - \theta)^2}_{\text{bias}}.$$

Distributions Derived from the Normal Distribution

chi-square distributions

Definition [χ^2 distribution]: Let n be a positive integer. A random variable X is said to have the χ^2 distribution with n degrees of freedom, denoted by $X \sim \chi^2(n)$, if X has continuous distribution with the following PDF:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



Properties of χ^2 distributions

Theorem

1. *If* $Z \sim N(0,1)$, *then* $Z^2 \sim \chi^2(1)$.
2. *If* $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, *and* X, Y *are independent, then* $X + Y \sim \chi^2(m+n)$.
3. *If* $X \sim \chi^2(n)$, *then* $E(X) = n$ *and* $Var(X) = 2n$.

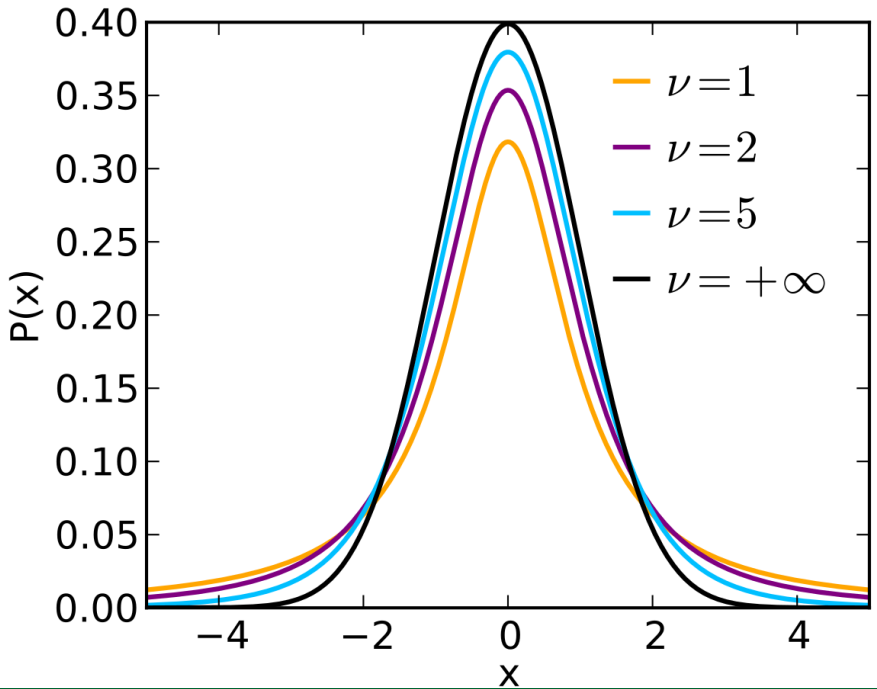
Corollary: Let $Z_1, \ldots, Z_n$ be **independent** random variables with all $Z_i \sim N(0,1)$, then $Z_1^2 + \cdots + Z_n^2$ has the χ^2 distribution with n degrees of freedom, i.e.

$$Z_1^2 + \cdots + Z_n^2 \sim \chi^2(n).$$

t distributions

Definition [t distribution]: Let n be a positive integer. A random variable X is said to have the t distribution with n degrees of freedom, denoted by $X \sim t(n)$, if X has continuous distribution with the following PDF:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



Properties of t distributions

Theorem

Suppose $X \sim N(0,1)$, $Y \sim \chi^2(n)$ *and* X, Y *are independent. Then*

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n).$$

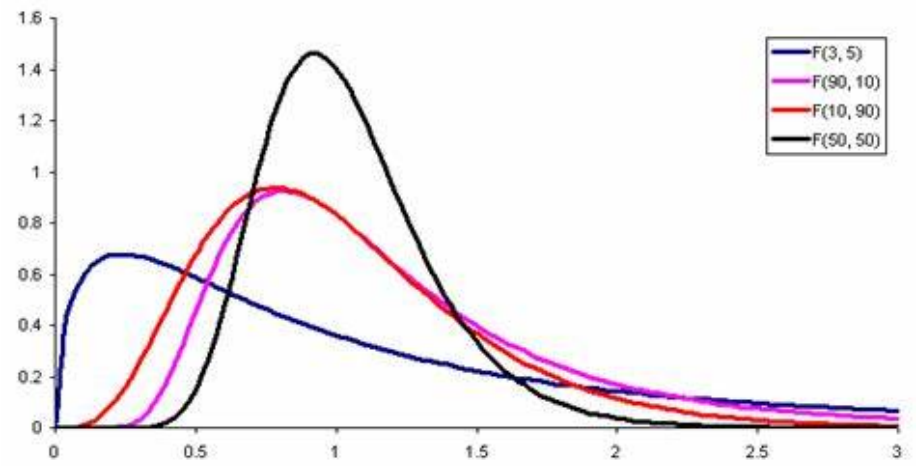
Theorem

When $t \to \infty$, *we have* $t(n) \to N(0,1)$.

F distributions

Definition [F distribution]: Let m, n be a positive integers. A random variable X is said to have the F distribution with m and n degrees of freedom, denoted by $X \sim F(m, n)$, if X has continuous distribution with the following PDF:

$$f(x) = \begin{cases} \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}(\frac{m}{n})^{\frac{m}{2}} x^{\frac{m}{2}-1}(1+\frac{m}{n}x)^{-\frac{m+n}{2}} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$



Properties of F distributions

Theorem

Suppose $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$ and X, Y are independent. Then

$$\frac{X/m}{Y/n} \sim F(m,n).$$

Theorem

If $F \sim F(m,n)$, then $\frac{1}{F} \sim F(n,m)$.

Theorem

If $T \sim t(n)$, then $T^2 \sim F(1,n)$.

Upper quantile for distributions

Suppose X is a random variable, $F(x) = P(X \leq x)$ is its CDF. For any real x, $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$. Sometimes, we need to consider its inverse problem. That is, if $P(X > x) = \alpha$ is given, determine the value of x.

Definition (upper α quantile)

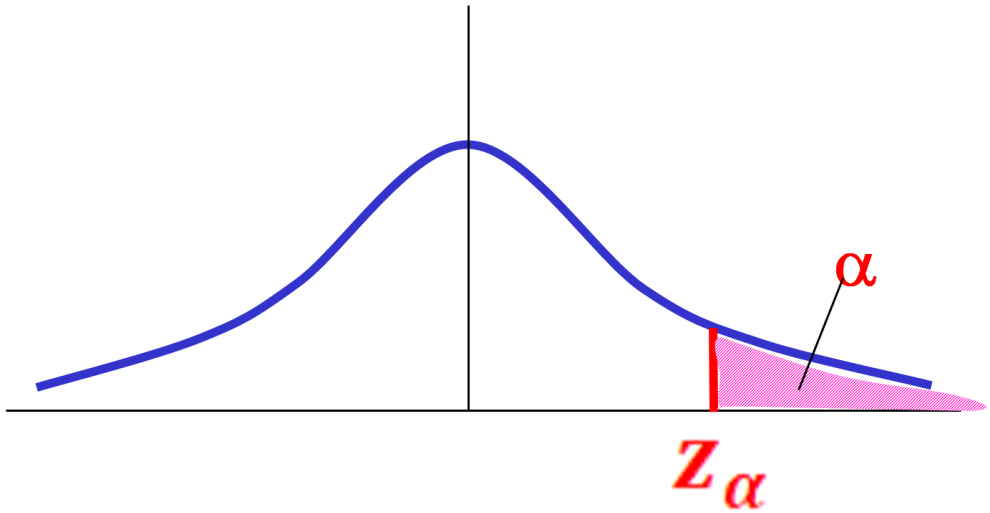
Given a distribution, whose CDF is $F(x)$, and a number α with $0 < \alpha < 1$, the upper α quantile of the distribution is the value x_α such that

$$1 - F(x_\alpha) = \alpha.$$

Upper α quantile for standard normal distribution

The upper α quantile for the standard normal distribution is denoted by z_α. That is, if $Z \sim N(0,1)$, then

$$P(Z > z_\alpha) = \alpha.$$

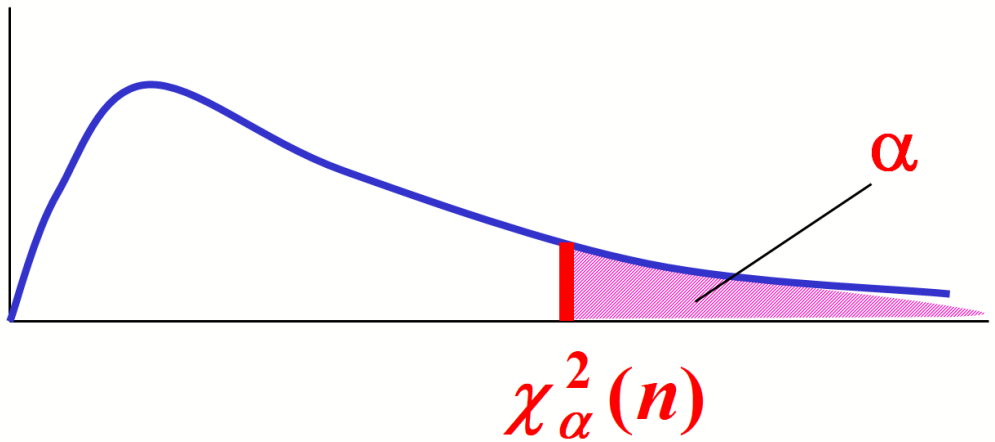


$z_\alpha = \Phi^{-1}(1-\alpha)$ can be found in the standard normal table. By symmetry, $z_{1-\alpha} = -z_\alpha$. For example, $z_{0.05} = 1.645$, $z_{0.95} = -1.645$.

Upper α quantile for χ^2 distributions

The upper α quantile for the $\chi^2(n)$ distribution is denoted by $\chi^2_\alpha(n)$. That is, if $X \sim \chi^2(n)$, then

$$P(X > \chi^2_\alpha(n)) = \alpha.$$

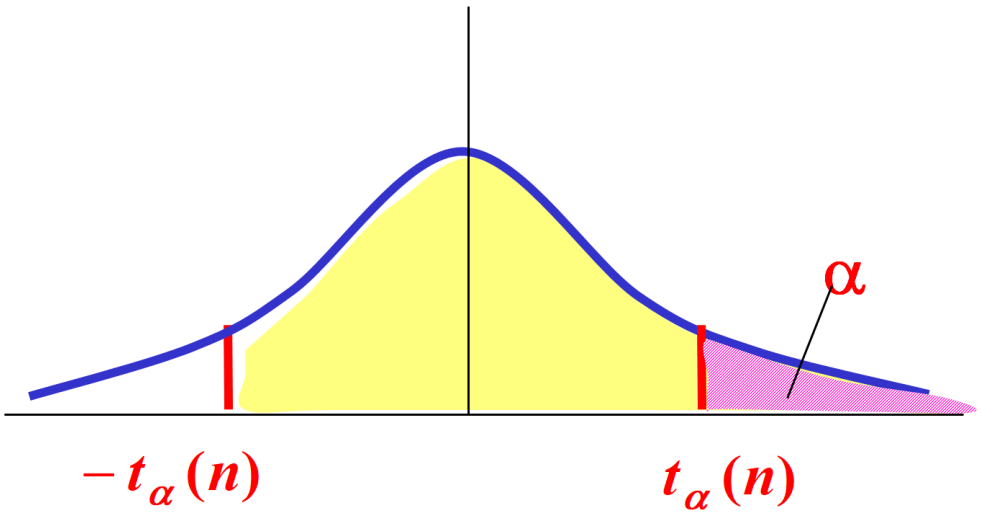


$\chi^2_\alpha(n)$ can be found in the χ^2 table.

Upper α quantile for t distributions

The upper α quantile for the t distribution with n degrees of freedom is denoted by $t_\alpha(n)$. That is, if $T \sim t(n)$, then

$$P(T > t_\alpha(n)) = \alpha.$$

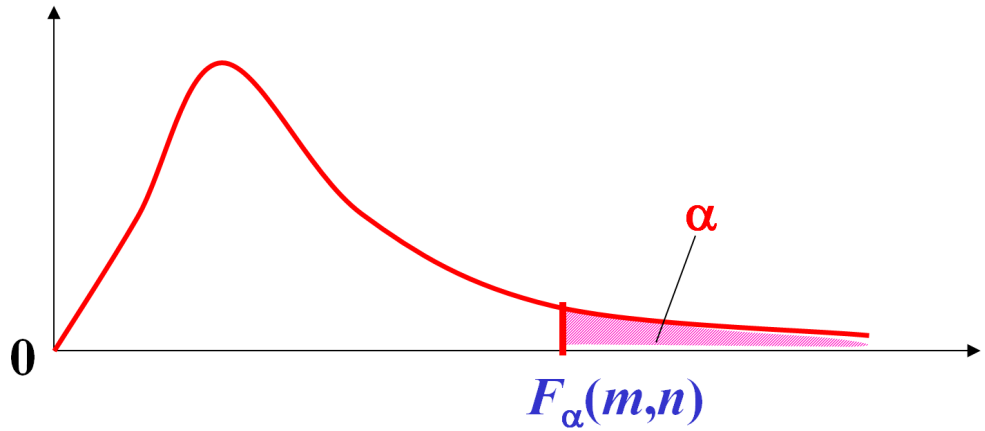


$t_\alpha(n)$ can be found in the t table. By symmetry, $t_{1-\alpha}(n) = -t_\alpha(n)$. For example, $t_{0.05}(4) = 2.1318$, $t_{0.95}(4) = -t_{0.05}(4) = -2.1318$. If $n > 30$, $t_\alpha(n) \approx z_\alpha$.

Upper α quantile for F distributions

The upper α quantile for the F distribution with m and n degrees of freedom is denoted by $F_\alpha(m,n)$. That is, if $F \sim F(m,n)$, then

$$P(F > F_\alpha(m,n)) = \alpha.$$



$F_\alpha(m,n)$ can be found in the F table. By the fact that if $F \sim F(m,n)$, then $\frac{1}{F} \sim F(n,m)$, thus

$$F_{1-\alpha}(m,n) = \frac{1}{F_\alpha(n,m)}.$$

For example, $F_{0.95}(7,4) = \frac{1}{F_{0.05}(4,7)} = \frac{1}{4.12} = 0.2427.$

Confidence Intervals for Normal Distribution Parameters

Student's theorem

Theorem

Suppose that $X_1, \ldots, X_n$ *are a random sample from* $N(\mu, \sigma^2)$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *and* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. *Then*

1. $\overline{X}$ *and* S^2 *are independent;*
2. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$;
3. $\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$;
4. $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

Proof.

See pages 182-183 of the textbook, or see my paper "A constructive algebraic proof of student's theorem"
`https://arxiv.org/pdf/1806.08031.pdf` □

The concept of confidence interval

Definition (confidence interval)

Suppose that $X_1, \ldots, X_n$ are a random sample from $X \sim F(x;\theta)$, where θ is unknown. For pre-specified $0 < \alpha < 1$, an interval estimate gives an interval such that the true value is falling into the interval with a given probability:

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ are statistics. Then

- $(\hat{\theta}_L, \hat{\theta}_U)$ is called a $1-\alpha$ confidence interval for θ.
- $\hat{\theta}_L$ is called the confidence lower limit, and $\hat{\theta}_U$ is called the confidence upper limit.
- $1-\alpha$ is called the confidence level.

Remarks:

1. It is NOT that the parameter is random. The parameter is fixed, although unknown. It is the interval that is random (its two endpoints are random variables).
2. The fact that the confidence interval contains the parameter with a probability $1-\alpha$ should be understood as follows: By the law of large numbers, if we select K simple random samples, and we calculate $1-\alpha$ confidence interval from each sample, then about $K(1-\alpha)$ of those K intervals would contain the true parameter. In practice, we will have only one observation of the confidence interval, and it is determined whether this observed interval contains the true parameter or not. We only do not know the answer. So for a particular observed confidence interval, $1-\alpha$ is not its probability to contain the true parameter; $1-\alpha$ is a **confidence**.
3. It is desirable to have a confidence interval that is short and with a high level of confidence. But the two requirements are usually contradictory to each other.

Confidence interval for μ of $N(\mu, \sigma^2)$, σ^2 known

Theorem

Suppose that $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$, where μ is unknown but σ^2 is known. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then a $1-\alpha$ confidence interval for μ is

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \quad \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}).$$

Proof.

Part (2) of Student's theorem states that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

Therefore

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

□

Confidence interval for μ of $N(\mu, \sigma^2)$, σ^2 unknown

Theorem

Suppose that $X_1, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then a $1-\alpha$ confidence interval for μ is

$$(\overline{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1), \quad \overline{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)).$$

Proof.

Part (4) of Student's theorem states that $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

Therefore $$1-\alpha = P(-t_{\frac{\alpha}{2}}(n-1) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2}}(n-1))$$

$$= P(\overline{X} - \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1) < \mu < \overline{X} + \frac{S}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1)).$$

□

General procedure for finding a confidence interval

In the preceding two theorems, we noticed that $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ and $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ are functions of the random sample and the parameter to be estimated (therefore they are not statistics), and their distributions do not depend on the parameter. They are called **pivotal quantities**. Pivotal quantities are fundamental to the construction of confidence intervals. A confidence interval can be constructed as follows:

1. Find a pivotal quantity $g(X_1, \ldots, X_n, \theta)$.
2. For a specified confidence level, find confidence lower limit and confidence upper limit on the pivotal quantity, that is, numbers a and b such that $P(a < g(X_1, \ldots, X_n, \theta) < b) = 1 - \alpha$.
3. Solve the inequalities to obtain $P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha$, and we then find the confidence interval.

Confidence interval is not unique

The constants a and b are called critical values. They are obtained from a table for the distribution of the pivotal quantity or from a computer program. In general, they are not unique and thus the confidence interval is not unique. A convenient choice is to use equal-tailed critical values as this often leads to the shortest confidence interval.

Example (1)

The cost of the textbools for statistics courses is normally distributed, and the standard deviation is \$100. A sample of 55 textbooks is selected and the mean cost is \$124. Construct the 95% confidence interval for the mean cost of all statistics textbooks.

Solution: Let X be the cost in \$ of the textbooks. Then $X \sim N(\mu, \sigma)$ where $\sigma = 100$ is known. We have $n = 55$, $1 - \alpha = 0.95$, and thus $\alpha = 0.05$ and $z_{\frac{\alpha}{2}} = z_{0.025} = \Phi^{-1}(0.975) = 1.96$. The observed 95% confidence interval for μ is

$$(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{\frac{\alpha}{2}}) = (124 - \frac{100}{\sqrt{55}} \times 1.96, 124 + \frac{100}{\sqrt{55}} \times 1.96)$$

$$\approx (97.6, 150.4).$$

Example (2)

A sample of 23 first-year students at a large university has a mean SAT math score of 525 with a standard deviation of 88. Construct the 90% confidence interval for the mean SAT math score of the entire first-year class.

Solution: Let X be the SAT math score of the first-year class. Then $X \sim N(\mu, \sigma)$ where μ, σ are unknown. We have $n = 23$, $\bar{x} = 525$, $s = 88$. We also have $1 - \alpha = 0.9$, and thus $\alpha = 0.1$, and thus $t_{\frac{\alpha}{2}}(n-1) = t_{0.05}(22) = 1.7171$. The observed 90% confidence interval for μ is

$$(\bar{x} - \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1),\ \bar{x} + \frac{s}{\sqrt{n}} t_{\frac{\alpha}{2}}(n-1))$$

$$= (525 - \frac{88}{\sqrt{23}} \times 1.7171, 525 + \frac{88}{\sqrt{23}} \times 1.7171)$$

$$\approx (493.5, 556.5).$$