Notes for the course Principles of Statistical Inference

Prof. Noel Veraverbeke
Yilma Tefera
Legesse Negash
Zeytun Gashaw
Belay Birlie

Master in Biostatistics, South West Ethiopia

This course has been developed within the context of the North-South-South project sponsored by VLIR-UOS, Belgium (see http:\\www.NSSbiostat.ugent.be)and Cross Cutting Initiatives in statistics.

Prof. Noel Veraverbeke
Yilma Tefera
Legesse Negash
Zeytun Gashaw
Belay Birlie

Notes for the course Principles of Statistical Inference

North-South-South project in Biostatistics Series

VLIR-UOS, Belgium

Yilma Tefera
Addis Ababa University
Ethiopia

Legesse Negash
Jimma University
Ethiopia

Zeytun Gashaw
Hawassa University
Ethiopia

Belay Birlie
Jimma University
Ethiopia

Prof. Noel Veraverbeke
Center for Statistics
Hasselt University
Agoralaan - Building D
B-3590 Diepenbeek
Belgium
noel.veraverbeke@uhasselt.be

©2012 North-South-South project in Biostatistics Series

All rights reserved. This work is based on the course notes of Prof. Noel Veraverbeke, that was developed in the context of the Principles of Statistical inference course taught at the Masters of Biostatistics of Hasselt University, Belgium. This work may not be translated or copied in whole or in part without the written permission of the authors.

North-South-South project in Biostatistics Series

The North-South-South project in Biostatistics is sponsored by the Flemish Interuniversity Council - University Development Cooperation (VLIR-UOS) and constitutes a collaboration between different Flemish and Ethiopian universities, and Eduardo Mondlane University, Mozambique.
It aims at starting up new or reinforcing existing Master programmes in Biostatistics in Ethiopia.

The project originates from initiatives taken by Luc Duchateau (UGent) and Paul Janssen (UHasselt) in the context of the Institutional University Collaboration programme (IUC) with Jimma University. One of the objectives of the IUC programme was to establish a Masters programme in Biostatistics at Jimma University. It seemed, however, more appropriate to open up this project to other Ethiopian Universities. After establishing contacts with interested Ethiopian universities, the North-South-South project in Biostatistics was successfully defended.

The North-South-South project in Biostatistics operates in the following way. First, a list of 10 courses that are required in a Master in Biostatistics was made. We then matched each course in the list with an existing course in one of the existing master programmes in (bio)statistics in Flanders. At the same time we compiled for each course a Flemish-Ethiopian writing team. When the course is taught in Flanders, two to three Ethiopian academic staff members of the writing team for that course are invited to come to Belgium to follow the course. During that time, the team reconsiders the original course notes as used in the existing course and adapts the notes for use in the Ethiopean/Southern context (i.e., they include Ethiopian data sets and write programmes for statistical analysis using freeware (R)). These courses will then be used by the Ethiopean/Southern team members in their respective master programmes. All the materials such as the syllabus, the slides, the data sets, the statistical programmes are made available to the different participating universities through http:\\www.NSSbiostat.ugent.be.

Luc Duchateau and Ziv Shkedy
Series Editor

Preface

This textbook aims at providing the theory of statistical inference for students of statistics at Masters level. Prerequisites for the book are calculus, linear algebra, and some knowledge of basic probability.

Chapter 1 provides a quick overview of important concepts and results in distribution theory that is used as tools in statistical inference. Chapter 2 studies the theory and methods in point estimation under parametric models. Chapter 3 covers interval estimation and confidence sets. The last chapter focuses on hypothesis testing. The classical frequentist approach is adopted in this book, although the Bayesian approach is also introduced in sections 2.4.2, 3.5, and 4.3.2.

Although this textbook focuses on the theory of statistical inference, it is often helpful to apply principles and formulas to data to better explore and understand concepts and to visualize results. There are many different software products available for data analysis. Incorporating R into a statistical inference course can facilitate the instruction of many concepts and principles typically covered in this course and allows for expansion to other topics as well. In this material, we will present a few of the ways to use R to allow students to connect, explore, visualize, and expand different concepts in statistical inference. R is a free open-source program for statistical computing and graphics that can be downloaded. R is a very powerful and flexible program that can run on Windows, Linux and Macintosh computers. The commercial version of R is S-plus.

Prof. Noel Veraverbeke,Yilma Tefera, Legesse Negash, Zeytu Gashaw. and Belay Birlie

Acknowledgement

This work has been funded by VLIR-UOS (Flemish University Council - University Development Cooperation, www.vliruos.be) and the Cross cutting Initiative Project in statistics. The course has been developed in Stwo stages.

In the first stage a team consisting of Yilma Tefera (Adis Ababa University), Zeytun Gashaw (Hawassa University) and Legesse Negash (Jimma University) was invited to Hasselt University, Belgium to develop this course under the support of North-South-South project in Biostatistics. This team on the behalf of Jimma University, Department of statistics is grateful to the Flemish Interuniversity Council (VLIR-UOS) and the Directorate General for Belgian Development Cooperation (DGDC) for covering the costs of air tickets to and from Belgium and also covering all expenses during its stay in Belgium. In the second stage, Belay Birlie from Jimma University was invited to Hasselt University Under the support of Cross Cutting Initiatives in statistics, Belgium to work on the course further with Professor Noel Veraverbeke.

Contents

Chapter 1

Introduction

1.1 Parametric and nonparametric statistical inference

Statisticians want to learn as much as possible from a limited amount of **data**. A first step is to set up an appropriate mathematical model for the process which generated the data. Such a model has a probabilistic nature.

Suppose the statistician observes an outcome x of an experiment. This outcome is considered as a value of some random variable (or random vector) X. The distribution function F of X is unknown to the statistician. Using some information about the way the experiment runs, he is mostly able to say that F belongs to some specified family $\mathfrak{F}$ of distribution functions which are appropriate for his experiment. $\mathfrak{F}$ is called the **model**. The statistician would like to know the true F, i.e. that member of $\mathfrak{F}$ that actually governs the experiment.

- Sometimes each member of $\mathfrak{F}$ can be specified by one single real parameter θ, or more general, by a finite number of real parameters, i.e. a vector $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$. In that case, the family $\mathfrak{F}$ can be replaced by the set Θ of all possible parameter values. The set $\Theta \subset I\!R^k$ is called the **parameter space**. In this case , the model $\mathfrak{F}$ can be written as $\mathfrak{F} = \{F_\theta | \theta \in \Theta\}$: a family of distribution functions, indexed by θ. Such a situation is called a **parametric** situation. The statistician wants to know the "true parameter".
- In other cases, the members of $\mathfrak{F}$ cannot be represented by a finite number of parameters. These situations are called **nonparametric**.

Examples

1. An experiment has only two possible outcomes : S and F (Success and Failure).
   Let X be the random variable

$$X = \begin{cases} 0 & \dots \text{if } F \text{ occurs} \\ 1 & \dots \text{if } S \text{ occurs} \end{cases}$$

   Let θ denote the probability that S occurs.
   Model : $X \sim B(1;\theta)$ (Bernoulli, parameter θ)
   Parameter space : $\Theta =]0,1[$.

2. An experiment has k possible outcomes : $O_1, O_2, \dots, O_k$ and is carried out independently n times.
   Let $\underset{\sim}{X} = (X_1, \dots, X_k)$ be the random vector with X_j = the number of times that O_j occurs, $j = 1, \dots, k$.
   Let θ_j denote the probability that O_j occurs $(j = 1, \dots, k)$
   Model : $\underset{\sim}{X} \sim M(n; (\theta_1, \dots, \theta_k))$ (multinomial)
   Parameter space : $\Theta = \{(\theta_1, \dots, \theta_k) | \theta_1 \geq 0, \dots, \theta_k \geq 0, \theta_1 + \dots + \theta_k = 1\}$

3. An experiment consists of measuring the value of a certain constant $\theta \in \mathbb{R}$. Measurements are subject to random error.
   Let X be the random variable which describes the outcome of the experiment. A parametric model could be :

$$X \sim N(\theta; \sigma^2) \quad \text{(normal)}$$

   Parameter space : $\Theta = \{(\theta, \sigma^2) | \theta \in \mathbb{R}, \sigma > 0\}$.

4. A nonparametric model could be : X has a distribution function symmetric about θ.

1.2 Some problems of statistical inference

Statistical inference deals with methods of using the outcomes to obtain information on the "true distribution function" (or the "true parameter") underlying the experiment.

Approaches to statistical inference There are two broad approaches to formal statistical inference. Differences relate to interpretation of probability and objectives of statistical inference.

(i) Frequentist(or classical) approach. Inference from data founded on comparison with datasets from hypothetical repetitions of the experiment generating data, under exactly the same conditions. A central role is played by sufficiency, likelihood and optimality criteria. Inference procedures are taken as decision problems rather than as a summary of data. Optimum inference procedures identified before the data are observed. Optimality defined explicitly in terms of the repeated sampling principle.

(ii) Bayesian approach. The unknown parameter θ treated as random variable. The key in this method is that one has to specify *a* prior distribution about θ before the data analysis. The specification is objective or subjective. Inference is a formalization of how the prior changes to the posterior in the light of data via Bayes' formula.

In this material we will mainly focus on the frequentist approach. We will study methods for obtaining estimation and testing procedures which satisfy certain optimality criteria.

The two most important topics in statistical inference are **estimation** and **hypothesis testing**.

- **Estimation.** The observations are used to calculate an approximation (estimate) for some numerical characteristic of the true distribution function (e.g. the mean, the variance, ...) or, if the model $\mathfrak{F}$ is parametric, the statistician wants to find a numerical approximation for the true parameter.
  The approximation can take the form of one numerical value (**point estimation**) or the form of an interval or set of possible values (**set estimation**).

- **Hypothesis testing.** The observations are used to conclude whether or not the true distribution belongs to some smaller family of distribution functions $\mathfrak{F}_0 \subset \mathfrak{F}$. In the parametric case, the statistician infers whether or not the true parameter belongs to a subset $\Theta_0 \subset \Theta$.

1.3 Random sample

A very common situation is the following : the statistician has available a number of outcomes

$$x_1, x_2, \ldots, x_n$$

which are obtained by repeating the experiment "independently" a number of times. These outcomes can be considered as values of random variables

$$X_1, X_2, \ldots, X_n$$

which are independent and have the same distribution as X.
We will often write

$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} X$$

and say : "$X_1, \ldots, X_n$ are independent and identically distributed as X" or also: $X_1, \ldots, X_n$ is a **random sample from** X. Sometimes X is referred to as the population random variable or as the **population**.

1.4 Statistics

Mostly the statistician does not use the observations $x_1, x_2, \ldots x_n$ as such, but he tries to condense them in some known function (not depending on any unknown parameters) such as

$$t(x_1, x_2, \ldots x_n)$$

If the function t is such that $t(X_1, X_2, \ldots X_n)$ is a random variable, then

$$T_n = t(X_1, X_2, \ldots X_n)$$

is called a **statistic**.

A **k-dimensional statistic** is a vector

$$\underset{\sim}{T}_n = (T_{n1}, T_{n2}, \ldots T_{nk})$$

where, for $j = 1, \ldots k$:

$$T_{nj} = t_j(X_1, \ldots X_n)$$

is a (1-dimensional) statistic.

Important **examples** of statistics are

- the **sample mean** :

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- the **sample variance** :

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

It will be important to calculate, for a given statistic T_n, characteristics such as: $E(T_n), Var(T_n), \ldots$ or the distribution function $P(T_n \leq x), \ldots$ In the next section, we consider the distribution theory for $\overline{X}$ and S^2 in the important case where the sample comes from a normally distributed population random variable X.

1.5 Distribution theory for samples from a normal population

In this section we give distribution theory for two important statistics $\overline{X}$ and S^2 in sampling from a normal population : i.e.

$$X_1, \ldots, X_n \overset{i.i.d}{\sim} X \sim N(\mu; \sigma^2)$$

The reason is that these results play a crucial role in the whole theory of statistics. This is because many populations are normal or can be well approximated by a normal. We restrict attention to the two statistics $\overline{X}$ and S^2. These will turn out to be the only two of interest in sampling from a normal population ($\overline{X}$ and S^2 are "sufficient" statistics).

A crucial fact in normal sampling theory is the following theorem.

Theorem 1

If X is normally distributed, then $\overline{X}$ and S^2 are independent.

Proof

We prefer to give a proof only in the very special case $n = 2$. In this case

$$\overline{X} = \frac{1}{2}(X_1 + X_2)$$

and

$$S^2 = \frac{1}{2}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2] = \frac{1}{4}(X_1 - X_2)^2.$$

So $\overline{X}$ is a function of $X_1 + X_2$ and S^2 is a function of $X_1 - X_2$. Hence it suffices to show that $X_1 + X_2$ and $X_1 - X_2$ are independent, or, equivalently that $Y_1 = X_1 + X_2 - 2\mu$ and $Y_2 = X_1 - X_2$ are independent. By simple calculation it follows :

$$\begin{aligned} E[e^{it_1Y_1+it_2Y_2}] &= e^{-2it_1\mu}E[e^{i(t_1+t_2)X_1}e^{i(t_1-t_2)X_2}] \\ &= e^{-2it_1\mu}E[e^{i(t_1+t_2)X_1}]E[e^{i(t_1-t_2)X_2}] = e^{-\frac{1}{2}(2\sigma^2t_1^2+2\sigma^2t_2^2)} \end{aligned}$$

From this joint characteristic function we see that (Y_1, Y_2) is 2-variate normal with mean vector $(0, 0)$ and variance-covariance matrix $\begin{pmatrix} 2\sigma^2 & 0 \\ 0 & 2\sigma^2 \end{pmatrix}$.
Hence Cov $(Y_1, Y_2) = 0$ and (because of normality) this is equivalent to independence of Y_1 and Y_2. □

We are now able to derive the distribution of $\overline{X}$ and S^2.

Theorem 2

If $X \sim N(\mu; \sigma^2)$, *then*

(a) $\overline{X} \sim N(\mu; \frac{\sigma^2}{n})$
(b) $\frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$

Proof

(a) Since $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a linear combination of $X_1, \ldots, X_n$, we have

$$\overline{X} \sim N(\sum_{i=1}^{n}\frac{1}{n}\mu; \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2) = N(\mu; \frac{\sigma^2}{n})$$

(b)

$$\begin{aligned} nS^2 &= \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2 \end{aligned}$$

$$\frac{nS^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right)^2$$

or

$$\underset{U}{\frac{nS^2}{\sigma^2}} + \underset{V}{\left(\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right)^2} = \underset{W}{\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2}$$

For the characteristic functions of U, V, W we have

$$\begin{aligned} \varphi_W(t) &= \varphi_{U+V}(t) \\ &= \varphi_U(t).\varphi_V(t) \qquad \text{,since } U \text{ and } V \text{ are independent (Th.1).} \end{aligned}$$

Hence :

$$\begin{aligned} \varphi_U(t) &= \frac{\varphi_W(t)}{\varphi_V(t)} \\ &= \frac{(1-2it)^{-n/2}}{(1-2it)^{-1/2}}, \qquad \text{since } W \sim \chi^2(n) \text{ and } V \sim \chi^2(1) \\ &= (1-2it)^{-\frac{n-1}{2}} \end{aligned}$$

By the uniqueness theorem $U = \dfrac{nS^2}{\sigma^2} \sim \chi^2(n-1)$. □

A further useful result is

Theorem 3

If $X \sim N(\mu; \sigma^2)$,*then* $\dfrac{\overline{X} - \mu}{\sqrt{\dfrac{S^2}{n-1}}} \sim t(n-1)$

Proof

$$\frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n-1}}} = \frac{\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{\frac{nS^2}{\sigma^2}}{n-1}}} \sim t(n-1), \text{ since } \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0;1); \frac{nS^2}{\sigma^2} \sim \chi^2(n-1) \text{ ; and}$$

$\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$ and $\frac{nS^2}{\sigma^2}$ are independent (Th. 1). □

Chapter 2

Parametric Point Estimation

2.1 Problems of point estimation

Suppose that the population random variable X has a distribution function that depends on one or more real unknown parameters.

Important remarks concerning notation :

(i) If we want to stress that the distribution function $F(x) = P(X \leq x)$ depends on some parameter θ, then we will also write : $F_\theta(x)$, or $F(x;\theta)$ or $P_\theta(X \leq x)$.

(ii) The term **density function** (or **density**) will be used for both **continuous** and **discrete** random variables. Also the notation will be the same : $f(x)$, or $f_\theta(x)$ or $f(x;\theta)$. Hence, in the continuous case $f(x;\theta)$ is a nonnegative integrable function such that $F(x;\theta) = \int_{-\infty}^{x} f(t;\theta)dt$. In the discrete case : $f(x;\theta) = P_\theta(X = x)$.

(iii) Expectations with respect to a density $f(x;\theta)$ will sometimes be denoted by E_θ.

(iv) The above remarks also apply if X and/or θ is a vector.

Point estimation of an unknown parameter θ is done in the following way :
the observed values $x_1, \ldots, x_n$ of a random sample $X_1, \ldots, X_n$ are used to suggest an approximation for θ of the form $t(x_1, \ldots, x_n)$ where t is such that $t(X_1, \ldots, X_n)$ is a statistic.

It is tradition to call the random variable

$$T_n = t(X_1, \ldots, X_n)$$

an **estimator for** θ
and a numerical value

$$t(x_1, \ldots, x_n)$$

an **estimate for** θ.

If the unknown parameter is multidimensional, say $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$, then an estimator for $\underset{\sim}{\theta}$ is a vector

$$\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$$

where, for $j = 1, \ldots, k : T_{nj} = t_j(X_1, \ldots, X_n)$ is an estimator for θ_j.

Two **problems** immediately arise :

1. What are methods to construct estimators for a given parameter ?
2. How do we construct an estimator which is 'best' in some sense ?

In this chapter we will discuss the maximum likelihood method and other methods for constructing estimators. Their optimality properties will be examined.

We first introduce some general properties of estimators.

2.2 General properties of point estimators

2.2.1 Unbiasedness

Definition

An estimator $T_n = t(X_1, \ldots, X_n)$ is said to be an **unbiased estimator for** θ if

$$E_\theta(T_n) = \theta$$

for all $\theta \in \Theta$.

Example

Let $X_1, \ldots, X_n$ be random sample from X.
Assume $E(X) = \mu$ and $Var(X) = \sigma^2$.
Then

(a) $\overline{X}$ is an unbiased estimator for μ.

(b) $\dfrac{n}{n-1}S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator for σ^2.

Proof

(a) $E(\overline{X}) = E\left(\dfrac{1}{n}\sum_{i=1}^{n} X_i \right) = \dfrac{1}{n}.n.E(X) = \mu$

(b) We have :

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

Hence

$$\begin{aligned} E\left(\frac{n}{n-1}S^2 \right) &= \frac{n}{n-1}E[(X-\mu)^2] - \frac{n}{n-1}E[(\overline{X}-\mu)^2] \\ &= \frac{n}{n-1}Var(X) - \frac{n}{n-1}Var(\overline{X}) \\ &= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\frac{\sigma^2}{n} \\ &= \sigma^2. \end{aligned}$$

□

Generalization to the multidimensional case is straightforward :

Definition

An estimator $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ is an unbiased estimator for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ if for $j = 1, \ldots, k$:

$$E(T_{nj}) = \theta_j$$

for all $\underset{\sim}{\theta} \in \Theta$.

Example

Let $\underset{\sim}{X} = (X_1, \dots X_k) \sim M(n; (\theta_1, \dots, \theta_k))$.
Then the vector of relative frequencies

$$\left(\frac{X_1}{n}, \frac{X_2}{n}, \dots, \frac{X_k}{n}\right)$$

is an unbiased estimator for $(\theta_1, \dots, \theta_k)$.

Indeed : for $j = 1, \dots, k$

$$E\left(\frac{X_j}{n}\right) = \frac{1}{n}E(X_j) = \frac{1}{n}n\theta_j = \theta_j. \qquad \square$$

If T_n is an estimator for θ, then the quantity

$$bias_\theta(T_n) = E_\theta(T_n) - \theta$$

is called the **bias of** T_n. With this definition : T_n is unbiased for θ iff $bias_\theta(T_n) = 0$ for all $\theta \in \Theta$.
If only

$$\lim_{n\to\infty} bias_\theta(T_n) = 0 \qquad , \text{ for all } \theta \in \Theta,$$

then we say that T_n is an **asymptotically unbiased estimator** for θ.

Example

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an asymptotically unbiased estimator for $\sigma^2 = Var(X)$. Indeed :

$$\begin{aligned} bias(S^2) &= E(S^2) - \sigma^2 \\ &= \left(\sigma^2 - \frac{\sigma^2}{n}\right) - \sigma^2 = -\frac{\sigma^2}{n} \to 0, \quad n \to \infty. \end{aligned}$$

Remark: Bias arises due to systematic error and measures, on the average, how far away and in what direction $E_\theta(T_n)$ is from the parameter θ.

Moreover, the property of unbiasedness by itself is not enough. We also need to have a procedure to measure the precision of an estimator. An unbiased estimator with a distribution highly concentrated near θ should be preferable to one with a distribution that is very spread out. The precision of any estimator T_n (biased or unbiased) for θ is measured by its mean square error (MSE) defined by

$$MSE_\theta(T_n) = E_\theta[(T_n - \theta)^2].$$

It is easy to check (Section 2.2.2) that

$$MSE_\theta(T_n) = Var_\theta(T_n) + (bias_\theta(T_n))^2.$$

In particular, if $bias_\theta(T_n) = 0$, then T_n is unbiased for θ and $MSE_\theta(T_n) = Var_\theta(T_n)$. An estimator T_n for θ is preferable above an estimator $T_n^{'}$ for θ if

$$MSE_\theta(T_n) \leq MSE_\theta(T_n^{'}),$$

for all $\theta \in \Theta$ with strict inequality holding for at least one $\theta \in \Theta$. For unbiased estimators T_n and $T_n^{'}$ it simplifies to

$$Var_\theta(T_n) \leq Var_\theta(T_n^{'}),$$

for all $\theta \in \Theta$ with strict inequality for at least one $\theta \in \Theta$. If there exists an unbiased estimator for θ which has the smallest variance among all unbiased estimators it is called a uniformly minimum variance unbiased (UMVU) estimator for θ. We will discuss this later.

Intuitively, if there are two unbiased estimators for θ the one based on sufficient statistic should be preferable to the one that is not based on a sufficient statistic. It would appear therefore that we need only concentrate on estimators based on sufficient statistics. We will discuss this later.

2.2.2 Consistency

Suppose we wish to estimate a parameter θ by an estimator $T_n = t(X_1, X_2, ..., X_n)$ based on a sample of size n. If n is allowed to increase indefinitely, we are practically sampling the whole population. In that case we should expect T_n to be practically equal to θ. That is, if $n \to \infty$ it would be desirable to have T_n converge to θ in some sense. We say that a sequence of estimators T_n is a consistent sequence of estimators for θ if T_n converges in probability to θ

Definition An estimator $T_n = t(X_1, \ldots, X_n)$ is said to be a **weakly consistent estimator for** θ if, for $n \to \infty$:

$$T_n \xrightarrow{P} \theta$$

for all $\theta \in \Theta$. where $\xrightarrow{P}$ denotes convergence in probablity.

Theorem *If T_n is an estimator which is asymptotically unbiased for θ and such that*

$$\lim_{n\to\infty} Var_\theta(T_n) = 0$$

for all $\theta \in \Theta$, then T_n is weakly consistent for θ.

Proof From Chebishev's inequality : for every $\varepsilon > 0$:

$$P_\theta(|T_n - \theta| \geq \varepsilon) \leq \varepsilon^{-2} E_\theta[(T_n - \theta)^2]$$

Now,

$$\begin{aligned} E_\theta[(T_n - \theta)^2] &= E_\theta[(T_n - E_\theta(T_n) + E_\theta(T_n) - \theta)^2] \\ &= Var_\theta(T_n) + (E_\theta(T_n) - \theta)^2 \\ &= Var_\theta(T_n) + (bias_\theta(T_n))^2 \to 0, \text{ as } n \to \infty. \end{aligned}$$

Hence, for all $\varepsilon > 0$:

$$P_\theta(|T_n - \theta)| \geq \varepsilon) \to 0, \text{ for all } \theta \in \Theta. \qquad \square$$

Example Let $X_1, \ldots, X_n$ be a random sample from X.

(a) if $E|X| < \infty$, then
$\overline{X}$ is a weakly consistent estimator for $\mu = E(X)$.

(b) if $E(X^2) < \infty$, then
S^2 is a weakly consistent estimator for $\sigma^2 = Var(X)$.

Proof

(a) $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$ is true by the weak law of large numbers.

(b) $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\overline{X} - \mu)^2$
By the weak law of large numbers :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \xrightarrow{P} E(X - \mu)^2 = \sigma^2$$

(indeed : the random variables $(X_i - \mu)^2$ are i.i.d. and $E(X - \mu)^2 < \infty$)

For the second term we have : $\overline{X} - \mu \xrightarrow{P} 0$, and hence $(\overline{X} - \mu)^2 \xrightarrow{P} 0$.
By Slutsky's theorem : $S^2 \xrightarrow{P} \sigma^2$. □

Note
If $MSE_\theta(T_n) \to 0$ for all $\theta \in \Theta$, then T_n is weakly consistent.

The following facts are often useful and are stated without proof.

1. Let g be a continous function. If $T_n \xrightarrow{P} \theta$ then $g(T_n) \xrightarrow{P} g(\theta) \quad as \quad n \to \infty$.
2. If $T_n \xrightarrow{P} \theta_1$ and $T_n^{'} \xrightarrow{P} \theta_2$ then:
   (a) $T_n \pm T_n^{'} \xrightarrow{P} \theta_1 \pm \theta_2$.
   (b) $T_n T_n^{'} \xrightarrow{P} \theta_1 \theta_2$
   (c) $\frac{T_n}{T_n^{'}} \xrightarrow{P} \frac{\theta_1}{\theta_2} (\theta_2 \neq 0)$

2.2.3 Efficiency

In statistics, efficiency is a term used in the comparision of various statistical procedures and , in particular, it refers to a measure of the desirability of an estimator or of an experimental design. Efficiencies are often defined using the variance or mean square error as the measure of desirability.
The efficiency of an unbiased estimator T_n is defined as

$$eff(T_n) = \frac{1/i(\theta)}{Var(T_n)}$$

where $i(\theta)$ is the Fisher information of the sample. Thus $eff(T_n)$ is the minimum possible variance for an unbiased estimator divided by its actual variance. The Cramer-Rao bound can be used to prove that (see section 2.4.1 about Fisher information and Cramer-Rao bound):

$$eff(T_n) \leq 1$$

$$Var(T_n) \geq \frac{1}{i(\theta)}$$

$$1 \geq \frac{1/i(\theta)}{Var(T_n)} = eff(T_n)$$

Efficient Estimator

If an unbiased estimator of a parameter θ attains $eff(T_n) = 1$ for all values of the parameter, then the estimator is called efficient.

Equivalently, the estimator achieves equality in the Cramer-Rao inequality for all $\theta \in \Theta$.

An efficient estimator is also the uniformily minimum variance unbiased (UMVU) estimator. This is because an efficient estimator maintains equality on the Cramer-Rao inequality for all parameter values, which means it attains the minimum variance for all parameters (the definition of the UMVU estimator). The UMVU estimator, even if it exists, is not necessarily efficient, because "minimum" does not mean equality holds on the Cramer-Rao inequality.

Thus an efficient estimator need not exist, but if it does, it is the UMVU estimator.

Asymptotic efficiency

For some estimators, they can attain efficiency asymptotically and are thus called asymptotically efficient estimators. This can be the case for some maximum likelihood estimators or for any estimators that attain equality of the Cramer-Rao bound asymptotically.

Relative efficiency

If T_n and $T_n^{'}$ are estimators for the parameter θ, then T_n is said to dominate $T_n^{'}$ if:

1. Its mean squared error (MSE) is smaller for at least some value of θ.
2. The MSE does not exceed that of $T_n^{'}$ for any value of θ.

The relative efficiency is defined as

$$eff(T_n, T_n^{'}) = \frac{E[(T_n^{'} - \theta)^2]}{E[(T_n - \theta)^2]}$$

Although efficiency is in general a function of θ, in many cases the dependence drops out; if this is so, efficiency being greater than one would indicate that T_n is preferable, whatever the true value of θ.

2.2.4 Asymptotically normal estimators

Definition

An estimator T_n for θ is said to be **asymptotically normal** if there exists a constant $\sigma(\theta) > 0$ such that, for $n \to \infty$:

$$n^{\frac{1}{2}} \frac{T_n - \theta}{\sigma(\theta)} \xrightarrow{d} Z$$

with $Z \sim N(0;1)$, for all $\theta \in \Theta$.

We will also write or say :

$$n^{\frac{1}{2}} \frac{T_n - \theta}{\sigma(\theta)} \xrightarrow{d} N(0;1)$$

or : $n^{\frac{1}{2}}(T_n - \theta) \xrightarrow{d} N(0; \sigma^2(\theta))$
or : T_n is approximately $N\left(\theta; \frac{\sigma^2(\theta)}{n}\right)$.

Example

Let $X_1, \ldots, X_n$ be a random sample from X.
Assume $E(X^2) < \infty$. Call $E(X) = \mu,\ Var(X) = \sigma^2$.
If $\sigma^2 > 0$, then

$$n^{\frac{1}{2}} \frac{\overline{X} - \mu}{\sigma} \xrightarrow{d} N(0;1)$$

Example

Let $X_1, \ldots, X_n$ be a random sample from X.
Assume $E(X^4) < \infty$. Call $E(X) = \mu,\ Var(X) = \sigma^2$, and $\tau^2 = E[(X-\mu)^4] - \sigma^4$.
If $\tau^2 > 0$, then

$$n^{\frac{1}{2}} \frac{S^2 - \sigma^2}{\tau} \xrightarrow{d} N(0;1)$$

Proof

$$n^{\frac{1}{2}} \frac{S^2 - \sigma^2}{\tau} = \frac{1}{\tau\sqrt{n}} \sum_{i=1}^{n} [(X_i - \mu)^2 - \sigma^2] - \frac{\sqrt{n}}{\tau}(\overline{X} - \mu)^2$$

For the first term we have by the central limit theorem

$$\frac{1}{\tau\sqrt{n}}\sum_{i=1}^{n}[(X_i-\mu)^2-\sigma^2] \xrightarrow{d} N(0;1)$$

(indeed : the random variables are i.i.d., $E(X-\mu)^2=\sigma^2$ and $Var(X-\mu)^2=E[(X-\mu)^4]-\sigma^4=\tau^2>0$)

For the second term we have

$$\frac{\sqrt{n}}{\tau}(\overline{X}-\mu)^2=\frac{\sigma}{\tau}\left[\sqrt{n}\frac{\overline{X}-\mu}{\sigma}\right][\overline{X}-\mu]\xrightarrow{P}0\ ,$$

since $\sqrt{n}\dfrac{\overline{X}-\mu}{\sigma}\xrightarrow{d}Z$ and $\overline{X}-\mu\xrightarrow{P}0$.

By Slutsky's theorem the result follows. □

The multi-dimensional extension is given by

Definition

An estimator $\underset{\sim}{T}_n=(T_{n1},\ldots,T_{nk})$ for $\underset{\sim}{\theta}=(\theta_1,\ldots,\theta_k)$ is said to be **asymptotically multivariate normal** if there exists a symmetric positive definite matrix $\Sigma(\underset{\sim}{\theta})$ such that, for $n\to\infty$

$$\begin{aligned} & n^{\frac{1}{2}}(\underset{\sim}{T}_n-\underset{\sim}{\theta}) \\ = \ & n^{\frac{1}{2}}(T_{n1}-\theta_1,\ldots,T_{nk}-\theta_k)\xrightarrow{d}\underset{\sim}{Z}\end{aligned}$$

with $\underset{\sim}{Z}\sim N_k(\underset{\sim}{0};\Sigma(\underset{\sim}{\theta}))$, for all $\underset{\sim}{\theta}\in\Theta$.

We will also write or say :

$$n^{\frac{1}{2}}(\underset{\sim}{T}_n-\underset{\sim}{\theta})\xrightarrow{d}N_k(\underset{\sim}{0};\Sigma(\underset{\sim}{\theta}))$$

or

$$\underset{\sim}{T}_n \text{ is approximately } N_k(\underset{\sim}{\theta};\frac{1}{n}\Sigma(\underset{\sim}{\theta})).$$

Example

Let $\underset{\sim}{X} = (X_1, \ldots, X_k) \sim M(n; (\theta_1, \ldots, \theta_k))$.
Then, the vector of relative frequencies

$$\left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right)$$

is asymptotically multivariate normal :

$$n^{\frac{1}{2}} \left(\frac{X_1}{n} - \theta_1, \ldots, \frac{X_k}{n} - \theta_k\right) \xrightarrow{d} N_k(\underset{\sim}{0}; \Sigma)$$

where $\Sigma = [\sigma_{ij}]_{i,j=1,\ldots,k}$ and

$$\sigma_{ij} = \begin{cases} \theta_i(1-\theta_i) & \ldots \text{ if } i = j \\ -\theta_i\theta_j & \ldots \text{ if } i \neq j \end{cases}$$

Proof

Define the random vectors

$$\begin{aligned} \underset{\sim}{Y}_1 &= (Y_{11}, Y_{12}, \ldots, Y_{1k}) \\ \underset{\sim}{Y}_2 &= (Y_{21}, Y_{22}, \ldots, Y_{2k}) \\ &\ldots \\ \underset{\sim}{Y}_n &= (Y_{n1}, Y_{n2}, \ldots, Y_{nk}) \end{aligned}$$

where, for $m = 1, \ldots, n$

$\underset{\sim}{Y}_m = (0, \ldots, 0, 1, 0, \ldots, 0)$ if outcome 0_j occurs at the m-th trial.

$j-$th position

We have : $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, \ldots, \underset{\sim}{Y}_n$ are independent random vectors with the same distribution as $\underset{\sim}{Y}$, with mean vector $(\theta_1, \ldots, \theta_k)$ and covariance matrix $\sum = [\sigma_{ij}]_{i,j=1,\ldots,k}$.

Hence

$$n^{\frac{1}{2}}\left(\frac{X_1}{n}-\theta_1,\ldots,\frac{X_k}{n}-\theta_k\right)=\frac{1}{\sqrt{n}}(X_1-n\theta_1,\ldots,X_k-n\theta_k)$$
$$=\quad\frac{1}{\sqrt{n}}\left(\sum_{j=1}^{n}Y_{j1}-n\theta_1,\ldots,\sum_{j=1}^{n}Y_{jk}-n\theta_k\right)$$

$\xrightarrow{d} N_k(\underset{\sim}{0};\Sigma)$, by the multivariate central limit theorem. □

2.2.5 Functions of asymptotically normal estimators

Univariate case

Theorem [Delta Method]

Suppose that T_n is an asymptotically normal estimator for θ :

$$n^{\frac{1}{2}}(T_n-\theta)\xrightarrow{d}N(0;\sigma^2(\theta)).$$

If $g:\mathbb{R}\to\mathbb{R}:x\mapsto g(x)$ is a function, differentiable at $x=\theta$, with $g'(\theta)\neq 0$, then

$$n^{\frac{1}{2}}(g(T_n)-g(\theta))\xrightarrow{d}N(0;(g'(\theta))^2\sigma^2(\theta))$$

Proof

Define the function

$$h(x)\quad=\begin{cases}0 & \ldots \text{ if } x=\theta\\ \dfrac{g(x)-g(\theta)}{x-\theta}-g'(\theta) & \ldots \text{ if } x\neq\theta\end{cases}$$

then

$$g(T_n)-g(\theta)=(T_n-\theta)g'(\theta)+(T_n-\theta)h(T_n)$$

or

$$\frac{n^{\frac{1}{2}}(g(T_n)-g(\theta))}{g'(\theta)\sigma(\theta)} = n^{\frac{1}{2}}\frac{T_n-\theta}{\sigma(\theta)} + n^{\frac{1}{2}}\frac{T_n-\theta}{\sigma(\theta)}h(T_n)\frac{1}{g'(\theta)}$$

The theorem will be proved if the term $n^{\frac{1}{2}}\frac{T_n-\theta}{\sigma(\theta)}.h(T_n)$ tends to zero in probability.
But, $n^{\frac{1}{2}}\frac{T_n-\theta}{\sigma(\theta)} \xrightarrow{d} Z \sim N(0;1)$ and $h(T_n) \xrightarrow{P} 0$ since, by differentiability of g at θ, h is continuous at θ. Apply Slutsky's theorem. □

Example : sample standard deviation

Let $X_1, \ldots, X_n$ be a random sample from X.
Assume $E(X^4) < \infty$. Call $E(X) = \mu$, $Var(X) = \sigma^2$ and $\tau^2 = E[(X-\mu)^4] - \sigma^4$.
If $\tau^2 > 0$, then

$$n^{\frac{1}{2}}\frac{S-\sigma}{\lambda} \xrightarrow{d} N(0;1)$$

where $\lambda^2 = \frac{\tau^2}{4\sigma^2} = \frac{E[(X-\mu)^4]-\sigma^4}{4\sigma^2}$

Indeed : we know that $n^{\frac{1}{2}}(S^2-\sigma^2) \xrightarrow{d} N(0;\tau^2)$.
Apply the theorem with $g(x) = \sqrt{x}$, i.e. $g'(x) = \frac{1}{2\sqrt{x}}$ so that $g'(\sigma^2) = \frac{1}{2\sigma}$.

Application : variance - stabilizing transformations In many cases, an estimator T_n for θ is asymptotically normal, but the asymptotic variance depends on θ :

$$n^{\frac{1}{2}}(T_n-\theta) \xrightarrow{d} N(0;\sigma^2(\theta))$$

Therefore, it may be useful to find a variance-stabilizing transformation : i.e. a suitable function g such that

$$n^{\frac{1}{2}}(g(T_n)-g(\theta)) \xrightarrow{d} N(0;c^2)$$

where c^2 is now a constant, independent of θ.
From the theorem, we have to choose g such that

$$(g'(\theta))^2\sigma^2(\theta) = c^2$$

i.e. g has to satisfy the **differential equation** :

$$g'(\theta) = \frac{c}{\sigma(\theta)}.$$

Example : angular transformation

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1;\theta)$ (Bernoulli with parameter θ). For $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ we have

$$n^{\frac{1}{2}}(\overline{X} - \theta) \xrightarrow{d} N(0;\theta(1-\theta)).$$

The differential equation for g becomes

$$g'(\theta) = \frac{c}{\sqrt{\theta(1-\theta)}}.$$

Solving this equation for $c = \dfrac{1}{2}$ gives : $g(\theta) = \arcsin\sqrt{\theta}$.
Hence :

$$n^{\frac{1}{2}}\left(\arcsin\sqrt{\overline{X}} - \arcsin\sqrt{\theta}\right) \xrightarrow{d} N(0;\frac{1}{4}).$$

Example : square root transformation

Let T_n be the Poisson with parameter $\theta.n$, where $\theta > 0$.
We have

$$n^{\frac{1}{2}}\left(\frac{T_n}{n} - \theta\right) \xrightarrow{d} N(0;\theta)$$

and

$$n^{\frac{1}{2}}\left(\sqrt{\frac{T_n}{n}} - \sqrt{\theta}\right) \xrightarrow{d} N(0;\frac{1}{4})$$

From this one often concludes : if $X \sim P(\lambda)$ with λ large, then $\sqrt{X}$ is approximately $N(\sqrt{\lambda};\frac{1}{4})$.

Multivariate extension

Theorem [Delta Method]

Suppose that $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ *is an asymptotically multivariate normal estimator for* $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ *:*

$$n^{\frac{1}{2}}(\underset{\sim}{T}_n - \underset{\sim}{\theta}) \xrightarrow{d} N_k(\underset{\sim}{0}; \Sigma)$$

If $g : \mathbb{R}^k \to \mathbb{R} : \underset{\sim}{x} \mapsto g(\underset{\sim}{x})$ *is a function whose partial derivatives* $\dfrac{\partial g}{\partial x_i}$ $(i = 1, \ldots, k)$ *are continuous at* $\underset{\sim}{\theta}$ *and not all zero at* $\underset{\sim}{\theta}$*, then*

$$n^{\frac{1}{2}}(g(\underset{\sim}{T}_n) - g(\underset{\sim}{\theta})) \xrightarrow{d} N(0; \underset{\sim}{\Delta}\Sigma\underset{\sim}{\Delta}')$$

where

$$\underset{\sim}{\Delta} = \left(\left.\frac{\partial g}{\partial x_1}\right|_{\underset{\sim}{x}=\underset{\sim}{\theta}}, \ldots, \left.\frac{\partial g}{\partial x_k}\right|_{\underset{\sim}{x}=\underset{\sim}{\theta}} \right).$$

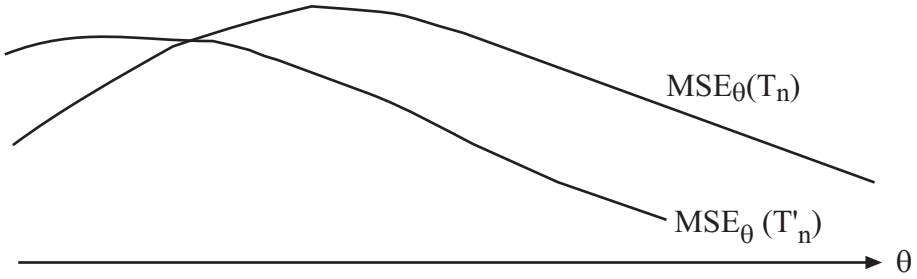
2.3 Uniformly Minimum Variance Unbiased estimators

2.3.1 Introduction

Measuring the closeness of an estimator T_n and the unknown parameter θ can be done by the **mean-squared error** of T_n :

$$MSE_\theta(T_n) = E_\theta[(T_n - \theta)^2].$$

It is then natural to look for an estimator for which the MSE is small. However, in general it will be impossible to compare two estimators T_n and T_n' by comparing the two functions $MSE_\theta(T_n)$ and $MSE_\theta(T_n')$. Indeed, the graphs of such functions are likely to cross.(for a concrete understanding see problem 1 of section 2.8)



In this section we will look for estimators with uniformly minimum MSE within the restricted class of **unbiased** estimators. If an estimator T_n is unbiased for θ, then

$$MSE_\theta(T_n) = Var_\theta(T_n)$$

for all $\theta \in \Theta$.

Definition

As estimator $T_n = t(X_1, \ldots, X_n)$ *is called an* **uniformly minimum variance unbiased (UMVU)** *estimator for* θ *if*

(i) T_n is an unbiased estimator for θ

(ii) For any other unbiased estimator T_n' of θ we have

$$Var_\theta(T_n) \leq Var_\theta(T_n') \qquad \text{for all } \theta \in \Theta.$$

In order to give a discussion on the existence and uniqueness of such estimator, we need to introduce the concept of **sufficiency**.

2.3.2 Sufficient statistics

In making inference about an unknown parameter θ, the statistician makes a **reduction of the data** by using a statistic, i.e. a function $t(X_1, \ldots, X_n)$ of the sample $X_1, \ldots, X_n$. He compresses the n random variables $X_1, \ldots, X_n$ into a single random variable $T_n = t(X_1, \ldots, X_n)$.

We will now formalize the intuitive idea that the statistic should be such that "no information about θ is lost". That is the function $t(X_1, \ldots, X_n)$ of the sample should tell us as much about θ as the sample $X_1, \ldots, X_n$ itself.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim Bernoulli$ with parameter $\theta \in [0, 1]$.
Suppose $x_1, \ldots, x_n$ is a set of observations, i.e. a sequence of 0's and 1's. It is intuitively clear that, in order to say something about $\theta (= P_\theta(X = 1))$, the only useful information is the number of 1's in the sequence (i.e. $\sum_{i=1}^{n} x_i$).
Once we know the sum, it looks like if the information concerning the order of 0's and 1's cannot help us any further. The sum carries all the information the sample has to give about the unknown parameter θ.

This concept in statistics is called **sufficiency** and it has been introduced by Fisher.

Let $X_1, \ldots, X_n$ be a random sample from X. Suppose that X has a (discrete or continuous) density $f(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}$.

Definition

A statistic $T_n = t(X_1, \ldots, X_n)$ is called **sufficient for a parameter** θ if the conditional distribution of $X_1, \ldots, X_n$ given $T_n = c$, does not depend on θ, for all values of c.

Note : In the above definition, θ can be a vector.
Thus, once the value of a sufficient statistic T_n is known, the sample $X_1, \ldots, X_n$ does not contain any further information about the parameter θ. Indeed, the distribution of the sample, given the sufficient statistic, does not depend on θ (and hence cannot be used to learn something about θ).

Example
Consider a random sample of size three from $B(1;\theta)$. Let $S = \sum_{i=1}^{3} X_i$ and $T = \sum_{i=2}^{3} X_i$. Then, we will show that S is sufficient and T is not.

Output	S	T	$f_{X_1,X_2,X_3}\|_S$	$f_{X_1,X_2,X_3}\|_T$
(000)	0	0	1	$(1-\theta)$
(001)	1	1	1/3	$\frac{(1-\theta)}{2}$
(010)	1	1	1/3	$\frac{(1-\theta)}{2}$
(100)	1	0	1/3	θ
(011)	2	2	1/3	$(1-\theta)$
(101)	2	1	1/3	$\frac{\theta}{2}$
(110)	2	1	1/3	$\frac{\theta}{2}$
(111)	3	2	1	θ

Then one can see that the statistic S is sufficient as the conditional distribution of the random sample given the statistic is free of θ, while the statistic T is not as the conditional distribution depends on the parameter θ.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1;\theta)$. $T_n = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ :

$$
\begin{aligned}
& P(X_1 = x_1, \ldots, X_n = x_n | \sum_{i=1}^{n} X_i = c) \\
&= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P\left(\sum_{i=1}^{n} X_i = c\right)} & \ldots \text{ if } \sum_{i=1}^{n} x_i = c \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{\theta^c(1-\theta)^{n-c}}{\binom{n}{c}\theta^c(1-\theta)^{n-c}} & \ldots \text{if } \sum_{i=1}^{n} x_i = c \qquad (\text{since } \sum_{i=1}^{n} X_i \sim B(n;\theta)) \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{1}{\binom{n}{c}} & \ldots \text{ if } x_1, \ldots, x_n = 0 \text{ or } 1 \text{ with } \sum_{i=1}^{n} x_i = c \\ 0 & \ldots \text{ if otherwise} \end{cases}
\end{aligned}
$$

which is independent of θ for all c.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim$ Poisson, with parameter $\theta > 0$.
$T_n = \sum_{i=1}^{n} X_i$ is sufficient for θ :

$$P(X_1 = x_1, \ldots, X_n = x_n | \sum_{i=1}^{n} X_i = c)$$

$$= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P\left(\sum\limits_{i=1}^{n} X_i = c\right)} & \ldots \text{ if } \sum\limits_{i=1}^{n} x_i = c \\ 0 & \ldots \text{ if otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}/x_1! \ldots x_n!}{e^{-n\theta}(n\theta)^c/c!} & \ldots \text{ if } \sum\limits_{i=1}^{n} x_i = c \\ & \qquad (\text{since } \sum\limits_{i=1}^{n} X_i \sim P(n\theta)) \\ 0 & \ldots \text{ if otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{c!}{x_1! \ldots x_n! n^c} & \ldots \text{ if } \sum\limits_{i=1}^{n} x_i = c \\ 0 & \ldots \text{ if otherwise} \end{cases}$$

which is independent of θ, for all c.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\theta; \sigma^2)$ with σ^2 known.
$T_n = \sum\limits_{i=1}^{n} X_i$ is sufficient for θ :

The conditional density function of $X_1, \ldots, X_n$ given that $\sum\limits_{i=1}^{n} X_i = c$ is :

$$\begin{cases} \dfrac{\left(\dfrac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\dfrac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2}}{\dfrac{1}{\sqrt{n}\sigma\sqrt{2\pi}} e^{-\dfrac{1}{2n\sigma^2}(c-n\theta)^2}} & \dots \text{ if } \sum_{i=1}^n x_i = c \\ & \qquad (\text{since } \sum_{i=1}^n X_i \sim N(n\theta; n\sigma^2) \\ 0 & \dots \text{ if otherwise} \end{cases}$$

$$= \begin{cases} \dfrac{\sqrt{n}}{\sigma^{n-1}(\sqrt{2\pi})^{n-1}} e^{-\dfrac{1}{2n\sigma^2}\left(n\sum_{i=1}^n x_i^2 - c^2\right)} & \dots \text{ if } \sum_{i=1}^n x_i = c \\ 0 & \dots \text{ if otherwise} \end{cases}$$

which does not depend on θ.

For some problems it is impossible to find one single sufficient statistic.
However, there will always exist a set of jointly sufficient statistics (in the sense of the following definition).

Definition

The r-dimensional statistic $\underset{\sim}{T}_n = (T_{n1}, \dots, T_{nr})$ is called **sufficient for a parameter** θ if the conditional distribution of $X_1, \dots, X_n$ given $T_{n1} = c_1, \dots, T_{nr} = c_r$ does not depend on θ, for all values of $c_1, \dots, c_r$.

Notes :

1. We also say : the set of statistics $T_{n1}, \dots, T_{nr}$ is **jointly sufficient** for θ.
2. θ can be a vector in the above definition.
3. There always exist two examples of **trivial** jointly sufficient statistics :
   - The sample $X_1, \dots, X_n$ itself is always jointly sufficient.
   - The ordered sample $X_{n:1}, \dots, X_{n:n}$ (where $X_{n:1} \leq X_{n:2} \leq \dots \leq X_{n:n}$ are the **order statistics** of the random sample $X_1, \dots, X_n$) is jointly sufficient : indeed, for given $c_1, \dots, c_n$:

$$P(X_1 = x_1, \ldots, X_n = x_n | X_{n:1} = c_1, \ldots, X_{n:n} = c_n)$$
$$= \begin{cases} \frac{1}{n!} & \ldots \text{ if } (x_1, \ldots, x_n) \text{ is a permutation of } (c_1, \ldots, c_n) \\ 0 & \ldots \text{ if otherwise} \end{cases}$$
which is independent of θ.

2.3.3 Factorization theorem of Fisher and Neyman

Instead of checking sufficiency via the definition it is more convenient to make use of the following theorem which gives a necessary and sufficient condition for a statistic to be sufficient. It also has the advantage that we no longer have to guess what statistic is sufficient.
The (a)-part of the theorem is for a one-dimensional statistic. The (b)-part deals with a multidimensional statistic. In both parts, θ can be a vector.

Theorem [Factorization theorem of Fisher and Neyman]

(a) *The statistic $T_n = t(X_1, \ldots, X_n)$ is sufficient for θ if and only if*

$$f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta) = g(t(x_1, \ldots, x_n); \theta).h(x_1, \ldots, x_n)$$

where g is a nonnegative function depending on θ and on $x_1, \ldots, x_n$ only through $t(x_1, \ldots, x_n)$ and h is a nonnegative function, not depending on θ.

(b) *The set of statistics $T_{n1} = t_1(X_1, \ldots, X_n), \ldots, T_{nr} = t_r(X_1, \ldots, X_n)$ is jointly sufficient for θ if and only if*

$$f(x_1; \theta) f(x_2; \theta) \ldots f(x_n; \theta) = g(t_1(x_1, \ldots, x_n), \ldots t_r(x_1, \ldots, x_n); \theta).h(x_1, \ldots, x_n)$$

where g is a nonnegative function depending on θ and on $x_1, \ldots, x_n$ only through $t_1(x_1, \ldots, x_n), \ldots, t_r(x_1, \ldots, x_n)$ and h is a nonnegative function not depending on θ.

'Proof'

We only give a proof of the (a)-part-in the **discrete** case–

- Suppose that $T_n = t(X_1, \ldots, X_n)$ is sufficient for θ.
  Then,
$$\begin{aligned} & f(x_1, \theta) f(x_2; \theta) \ldots f(x_n; \theta) \\ &= P(X_1 = x_1, \ldots, X_n = x_n) \\ &= P(X_1 = x_1, \ldots, X_n = x_n | T_n = c) P(T_n = c) \end{aligned}$$

The first factor is a function of $x_1, \ldots, x_n$ but does not depend on θ. The second factor depends on θ and on $t(x_1, \ldots, x_n)$.

- Conversely, suppose that the factorization holds.
  Then, $T_n = t(X_1, \ldots, X_n)$ is sufficient, since

$$
\begin{aligned}
&P(X_1 = x_1, \ldots, X_n = x_n | T_n = c) \\
&= P(X_1 = x_1, \ldots, X_n = x_n | t(X_1, \ldots, X_n) = c) \\
&= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T_n = c)} & \ldots \text{ if } t(x_1, \ldots, x_n) = c \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{P(X_1 = x_1, \ldots, X_n = x_n)}{\sum\limits_{\{t(y_1,\ldots,y_n)=c\}} P(X_1 = y_1, \ldots, X_n = y_n)} & \ldots \text{ if } t(x_1, \ldots, x_n) = c \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{g(t(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n)}{\sum\limits_{t(y_1,\ldots,y_n)=c} g(t(y_1, \ldots, y_n); \theta) h(y_1, \ldots, y_n)} & \ldots \text{ if } t(x_1, \ldots, x_n) = c \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{h(x_1, \ldots, x_n)}{\sum\limits_{t(y_1,\ldots,y_n)=c} h(y_1, \ldots, y_n)} & \ldots \text{ if } t(x_1, \ldots, x_n) = c \\ 0 & \ldots \text{ if otherwise} \end{cases}
\end{aligned}
$$

  and this does not depend on θ.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1; \theta), 0 < \theta < 1$.

$$
f(x; \theta) = \begin{cases} \theta^x (1 - \theta)^{1-x} & \ldots \text{ if } x = 0, 1 \\ 0 & \ldots \text{ if otherwise} \end{cases}
$$

$$
\begin{aligned}
\prod_{i=1}^{n} f(x_i; \theta) &= \begin{cases} \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} & \ldots \text{ if } x_1, \ldots, x_n = 0 \text{ or } 1 \\ 0 & \ldots \text{ if otherwise} \end{cases} \\
&= g(\sum x_i; \theta) h(x_1, \ldots, x_n)
\end{aligned}
$$

where

$$\begin{aligned} g(\textstyle\sum x_i;\theta) &= \theta^{\sum x_i}(1-\theta)^{n-\sum x_i} \\ h(x_1,\ldots,x_n) &= \begin{cases} 1 & \ldots \text{ if } x_1,\ldots,x_n = 0 \text{ or } 1 \\ 0 & \ldots \text{ if otherwise} \end{cases} \end{aligned}$$

Hence : $\sum_{i=1}^{n} X_i$ is sufficient for θ.

Example

Let $X_1,\ldots,X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with σ^2 known.
Put $\theta = \mu$.

$$\begin{aligned} \prod_{i=1}^{n} f(x_i;\theta) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2} \\ &= e^{-\frac{1}{2\sigma^2}[-2\theta\sum_{i=1}^{n} x_i + n\theta^2]} \; \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2}}{(\sigma\sqrt{2\pi})^n} \\ &= g\left(\sum_{i=1}^{n} x_i;\theta\right).h(x_1,\ldots,x_n) \end{aligned}$$

Hence : $\sum_{i=1}^{n} X_i$ is sufficient for μ (if σ^2 is known).

Example

Let $X_1,\ldots,X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with μ known.
Put $\theta = \sigma^2$.

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\theta) &= \frac{e^{-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2}}{(\sqrt{\theta}\sqrt{2\pi})^n}\cdot 1 \\ &= g\left(\sum_{i=1}^{n}(x_i-\mu)^2;\theta\right)h(x_1,\ldots,x_n)\end{aligned}$$

Hence : $\sum_{i=1}^{n}(X_i-\mu)^2$ is sufficient for σ^2 (if μ is known).

Example

Let $X_1,\ldots,X_n$ be a random sample from $X\sim N(\mu;\sigma^2)$ with μ and σ^2 both unknown.
Put $\underset{\sim}{\theta}=(\theta_1,\theta_2),\theta_1=\mu;\theta_2=\sigma^2$.

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\underset{\sim}{\theta}) &= \frac{e^{-\frac{1}{2\theta_2}\left[\sum_{i=1}^{n}x_i^2-2\theta_1\sum_{i=1}^{n}x_i+n\theta_1^2\right]}}{(\sqrt{\theta_2}\sqrt{2\pi})^n}\cdot 1 \\ &= g\left(\sum_{i=1}^{n}x_i,\sum_{i=1}^{n}x_i^2;\underset{\sim}{\theta}\right).h(x_1,\ldots,x_n)\end{aligned}$$

Hence : $\left(\sum_{i=1}^{n}X_i,\sum_{i=1}^{n}X_i^2\right)$ is sufficient for (μ,σ^2).

Example

Let $X_1,\ldots,X_n$ be a random sample from $X\sim Un[\theta_1,\theta_2]$.
Then :

(i) if θ_1 is known : $\max(X_1,\ldots,X_n)$ is sufficient for θ_2

(ii) if θ_2 is known : $\min(X_1,\ldots,X_n)$ is sufficient for θ_1

(iii) if θ_1,θ_2 unknown : $(\min(X_1,\ldots,X_n),\ \max(X_1,\ldots,X_n))$ is sufficient for (θ_1,θ_2)

We show (i).

$$\begin{aligned}
\prod_{i=1}^{n} f(x_i;\theta_2) &= \begin{cases} \dfrac{1}{(\theta_2-\theta_1)^n} & \dots \text{ if } \theta_1 \le x_1,\dots,x_n \le \theta_2 \\ 0 & \dots \text{ if otherwise} \end{cases} \\
&= \begin{cases} \dfrac{1}{(\theta_2-\theta_1)^n} & \dots \text{ if } \min(x_i) \ge \theta_1 \text{ and } \max(x_i) \le \theta_2 \\ 0 & \dots \text{if otherwise} \end{cases} \\
&= g(\max(x_i);\theta_2).h(x_1,\dots,x_n)
\end{aligned}$$

with

$$\begin{aligned}
g(\max(x_i);\theta_2) &= \begin{cases} \dfrac{1}{(\theta_2-\theta_1)^n} & \dots \text{ if} \max(x_i) \le \theta_2 \\ 0 & \dots \text{ if otherwise} \end{cases} \\
h(x_1,\dots,x_n) &= \begin{cases} 1 & \dots \text{ if } \min(x_i) \ge \theta_1 \\ 0 & \dots \text{ if otherwise} \end{cases}
\end{aligned}$$

Example

Let $\underset{\sim}{X} = (X_1,\dots,X_k) \sim M(n;(\theta_1,\dots,\theta_k))$

With $\underset{\sim}{x} = (x_1,\dots,x_k)$ and $\underset{\sim}{\theta} = (\theta_1,\dots,\theta_k)$, we have

$$f(\underset{\sim}{x};\underset{\sim}{\theta}) = \begin{cases} \dfrac{n!}{x_1!\dots x_k!}\theta_1^{x_1}\dots\theta_k^{x_k} & \dots \text{ if } \sum_{i=1}^{k} x_i = n \\ 0 & \dots \text{ if otherwise} \end{cases}$$

Hence : $(X_1,\dots,X_k)$ is sufficient for $(\theta_1,\dots,\theta_k)$.

The factorization theorem also immediately implies the following **invariance properties**.

Corollary

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta)$, $\theta \in \Theta$.
Let $T_n = t(X_1, \ldots, X_n)$ be a sufficient statistic for θ.

a) *If φ is a one-to-one function such that $\varphi(T_n)$ is again a statistic, then $\varphi(T_n)$ is a sufficient statistic for θ.*

b) *If ψ is a one-to-one function, then T_n is a sufficient statistic for $\psi(\theta)$.*

Proof

(a) Let $\widetilde{T}_n = \varphi(T_n) = \widetilde{t}\,(X_1, \ldots, X_n)$. Then $T_n = \varphi^{-1}(\widetilde{T}_n)$.
Hence,

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\theta) &= g(t(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n)\\ &= g(\varphi^{-1}(\widetilde{t}\,(x_1,\ldots,x_n));\theta)h(x_1,\ldots,x_n)\\ &= \widetilde{g}\,(\widetilde{t}\,(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n)\\ \text{where } \widetilde{g} &= g\circ\varphi^{-1}\end{aligned}$$

or $\widetilde{T}_n = \widetilde{t}\,(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

(b) Let $\theta^* = \psi(\theta)$. Then $\theta = \psi^{-1}(\theta^*)$.
Hence,

$$\prod_{i=1}^{n} f(x_i;\theta^*) = g(t(x_1,\ldots,x_n);\theta^*)h(x_1,\ldots,x_n)$$

or T_n is sufficient for θ^*. □

Note : in the above θ and/or T_n may be multi-dimensional.

Another consequence of the factorization theorem is the following relation between ML-estimators and sufficient statistics.

Corollary

If $T_n = t(X_1, \ldots, X_n)$ is a sufficient statistic for θ and if the ML-estimator for θ is unique, then the ML-estimator is a function of T_n (i.e. the ML-estimator depends on $X_1, \ldots, X_n$ only through $t(X_1, \ldots, X_n)$.

Proof

If $t(X_1, \ldots, X_n)$ is a sufficient statistic for θ, then by the factorization theorem, we have for the likelihood function L :

$$L(\theta; x_1, \ldots, x_n) = g(t(x_1, \ldots, x_n); \theta)h(x_1, \ldots, x_n).$$

If there is a unique ML-estimator, then there is a unique value of θ that maximizes the left hand side. Since h does not depend on θ, this value also maximizes $g(t(x_1, \ldots, x_n); \theta)$. But then it is seen that this value will depend on $x_1, \ldots, x_n$ only through the function $t(x_1, \ldots, x_n)$. □

Notes

1. The result for the multiparameter case is similar.
2. The result can fail if the ML-estimator is not unique.

Example
Let $X_1, \ldots, X_n$ be a random sample from $X \sim Un[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$

It is seen that $\underset{\sim}{T}_n = (\min(X_i), \max(X_i))$ is a sufficient statistic for θ.

We also know that the ML-estimator is not unique : the class of ML-estimators is given by

$$(1 - c)\left(\max(X_i) - \frac{1}{2}\right) + c\left(\min(X_i) + \frac{1}{2}\right)$$

where $0 \leq c \leq 1$.

It is possible to choose a ML-estimator which is not a function of $\underset{\sim}{T}_n$ alone : e.g.

$$\sin^2(X_1)\left(\max(X_i) - \frac{1}{2}\right) + \cos^2(X_1)\left(\min(X_i) + \frac{1}{2}\right).$$

Note
For a given parameter, there can be more than one set of sufficient statistics. E.g. for the parameter $\underset{\sim}{\theta} = (\theta_1, \theta_2)$ in a $N(\theta_1; \theta_2)$ density, we have

- $(X_1, \ldots, X_n)$ (the sample)
- $(X_{n:1}, \ldots, X_{n:n})$ (the order statistics)
- $(\overline{X}, S^2)$ (see before : one-to-one transformation of $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$)

2.3.4 Minimal sufficient statistics

When we introduced the concept of sufficiency, we said that our objective was to condense the data without losing any information about the parameter. We have seen that there is more than one set of sufficient statistics. For example, in sampling from a normal distribution with both the mean and variance unknown, we have noted three sets of jointly sufficient statistics, namely, the sample $X_1, X_2, ..., X_n$ itself, the order statistics $Y_1, Y_2, ..., Y_n$, and $\overline{X}$ and S^2. We naturally prefer the jointly sufficient set $\overline{X}$ and S^2 since they condense the data more than either of the other two. The question that we might ask is: Does there exist a set of sufficient statistics that condenses the data more than $\overline{X}$ and S^2? The answer is that there does not, but we will not develop the necessary tools to establish this answer. The notion that we are alluding to is that of a minimum set of sufficient statistics, which we label minimum sufficient statistics.

Definition A set of jointly sufficient statistics is defined to be minimal sufficient if and only if it is a function of every other set of sufficient statistics. Like many other definition , this definition is of little use in finding minimal sufficient statistics. If the joint density is properly factored, the factorization criterion will give us minimal sufficient statistics.

2.3.5 Theorem of Rao and Blackwell

This theorem relates the notion of sufficient statistics to the notion of UMVU-estimators. We will make use of **"conditional expectation"** i.e. expectation with respect to the conditional distribution.

Theorem [Rao - Blackwell] *Let U_n be an unbiased estimator for θ.*
Let T_n be a sufficient statistic for θ.
Put $\varphi(t) = E[U_n|T_n = t]$.
Then

(a) *$\varphi(t)$ does not depend on θ (hence : $\varphi(T_n)$ is an estimator for θ).*

(b) *$E[\varphi(T_n)] = \theta$ (hence : $\varphi(T_n)$ is an unbiased estimator for θ).*

(c) *$Var[\varphi(T_n)] \leq Var(U_n)$ (hence : $\varphi(T_n)$ has a variance which is not larger than that of U_n). And $Var[\varphi(T_n)] = Var(U_n)$ if and only if U_n is essentially a function of T_n.*

Proof

(a)

$$\begin{aligned}
& \varphi(t) \\
= \;& E[U_n|T_n = t] \\
= \;& E[u(X_1, \ldots, X_n)|T_n = t] \\
= \;& \begin{cases} \sum_{x_1} \ldots \sum_{x_n} u(x_1, \ldots, x_n) P(X_1 = x_1, \ldots, X_n = x_n | T_n = t) & \text{(discrete case)} \\ \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} u(x_1, \ldots, x_n) f_{X_1,\ldots,X_n|T_n}(x_1, \ldots, x_n|t) dx_1 \ldots dx_n & \text{(continuous case)} \end{cases}
\end{aligned}$$

That this is independent of θ, follows from the definition of sufficiency.

To prove (b) and (c), we restrict to the continuous case.

(b) $\varphi(t) = \int u f_{U_n|T_n}(u|t)du = \dfrac{1}{f_{T_n}(t)} \int u f_{U_n,T_n}(u,t)du$

Hence :

$$\begin{aligned}
\varphi(t) f_{T_n}(t) &= \int u f_{U_n,T_n}(u,t)du \\
E[\varphi(T_n)] &= \int \varphi(t) f_{T_n}(t)dt \\
&= \int [\int u f_{U_n,T_n}(u,t)dt]du \\
&= \int u f_{U_n}(u)du = E(U_n) = \theta
\end{aligned}$$

(c)

$$\begin{aligned}
Var[U_n] &= E[(U_n - \theta)^2] \\
&= E[[(U_n - \varphi(T_n)) + (\varphi(T_n) - \theta)]^2] \\
&= E[(U_n - \varphi(T_n)^2] + E[(\varphi(T_n) - \theta)^2] \\
&\quad +2E[(U_n - \varphi(T_n))(\varphi(T_n) - \theta)]
\end{aligned}$$

The first term is ≥ 0, the second equals $Var[\varphi(T_n)]$ and the third term is zero (show this).
Hence : $Var(U_n) \geq Var[\varphi(T_n)]$.

Note
This theorem shows how to improve on an unbiased estimator U_n by conditioning on a sufficient statistic T_n. The new unbiased estimator is the UMVU-estimator provided the sufficient statistic satisfies an additional property : **completeness**.

2.3.6 Completeness

We define the notion of completeness for a general family of densities. let T be a (discrete or continuous) random variable with corresponding family of densities $\{g(x;\theta)|\theta \in \Theta\}$.

Definition
The family of densities $\{g(x;\theta)|\theta \in \Theta\}$ *of* T *is* **complete** *if for every function* $u(x)$*, not depending on* θ*, it holds that :*
$E_\theta[u(T)] = 0$, *for all* $\theta \in \Theta$
implies
$P_\theta[u(T) = 0] = 1$, *for all* $\theta \in \Theta$.

Demonstrating completeness usually requires application of some theorem of analysis.

Example
Let $T \sim B(n;\theta)$.
The family of densities is $\{g(x;\theta)|0 < \theta < 1\}$, where

$$g(x;\theta) = \begin{cases} \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} & \dots \text{if } x = 0, 1, \dots, n \\ 0 & \dots \text{ if otherwise} \end{cases}$$

$$\begin{aligned} E_\theta[u(T)] &= 0 \text{ for all } 0 < \theta < 1 \\ &\Leftrightarrow \sum_{x=0}^{n} u(x) \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} = 0 \text{ , for all } 0 < \theta < 1 \\ &\Rightarrow \sum_{x=0}^{n} u(x) \begin{pmatrix} n \\ x \end{pmatrix} t^x = 0 \text{ , for all } 0 < t < \infty \left(t = \frac{\theta}{1-\theta} \right) \end{aligned}$$

Now, in order for a polynomial to be zero for all $t > 0$, we must have $u(x) \begin{pmatrix} n \\ x \end{pmatrix} = 0$,

for all $x = 0, 1, \dots, n$. This implies $u(x) = 0$ for all $x = 0, 1, \dots, n$.

Hence : the family is complete.

Example

Let $T \sim Un[0,\theta]$.
The family of densities is $\{g(x;\theta)|\theta > 0\}$, where

$$g(x;\theta) = \begin{cases} \frac{1}{\theta} & \dots \text{ if } 0 \le x \le \theta \\ 0 & \dots \text{ if otherwise} \end{cases}$$

$$\begin{aligned} E_\theta[u(T)] &= 0 \text{ , for all } \theta > 0 \\ &\Leftrightarrow \int_0^\theta u(x)\frac{1}{\theta}dx = 0 \text{ , for all } \theta > 0 \\ &\Rightarrow \int_0^\theta u(x)dx = 0 \text{ , for all } \theta > 0 \\ &\Rightarrow u(x) = 0 \text{ , for all } x > 0 \end{aligned}$$

Hence : complete.

Example

Let $T \sim N(\mu;\sigma^2)$, with σ^2 known.
Put $\theta = \mu$. Then the family is

$$\left\{ \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2} | \theta \in I\!R \right\}$$

$$\begin{aligned} E_\theta[u(T)] &= 0 \text{ , for all } \theta \in I\!R \\ &\Leftrightarrow \int_{-\infty}^{\infty} u(\theta+\sigma z) e^{-\frac{1}{2}z^2} dz = 0 \text{ , for all } \theta \in I\!R \\ &\Rightarrow u(x) = 0 \text{ , for all } x \in I\!R \qquad \text{(no proof)} \end{aligned}$$

Hence : complete family.

Example

Let $T \sim N(\mu; \sigma^2)$, with μ known.
Put $\theta = \sigma^2$. Then the family is

$$\left\{ \frac{1}{\sqrt{\theta}\sqrt{2\pi}} e^{-\frac{1}{2\theta}(x-\mu)^2} | \theta > 0 \right\}$$

$$\begin{aligned} E_\theta[u(T)] &= 0 \text{ , for all } \theta > 0 \\ &\Leftrightarrow \int_{-\infty}^{\infty} u(\mu + \sqrt{\theta}z) e^{-\frac{1}{2}z^2} dz = 0 \text{ , for all } \theta > 0 \end{aligned}$$

This does not imply : $u(x) = 0$ for all $x \in I\!R$.
Take e.g. : $u(x) = x - \mu$.
Hence : this family is **not** complete.

2.3.7 Theorem of Lehmann and Scheffé

We first combine the notions of sufficiency and completeness.

Definition

A statistic T_n is called a **complete sufficient statistic for** θ if

(i) T_n is a sufficient statistic for θ

(ii) The corresponding family of densities $\{g(x;\theta)|\theta \in \Theta\}$ of T_n is complete.

Theorem [Lehmann - Scheffé]

Let T_n be a complete sufficient statistic for θ.
Let φ be a function such that for all $\theta \in \Theta$:

$$E_\theta[\varphi(T_n)] = \theta \text{ and } Var_\theta[\varphi(T_n)] < \infty$$

Then

(a) *$\varphi(T_n)$ is unique*

(b) *$\varphi(T_n)$ is an UMVU-estimator for θ*

Proof

(a) Suppose ψ is another function such that $E_\theta[\psi(T_n)] = \theta$.
Then, $E_\theta[\varphi(T_n) - \psi(T_n)] = 0$ for all $\theta \in \Theta$, and hence
$P_\theta[\varphi(T_n) = \psi(T_n)] = 1$.

(b) $\varphi(T_n)$ is unbiased. Hence, it remains to show that for every unbiased estimator U_n :

$$Var_\theta(\varphi(T_n)) \leq Var_\theta(U_n) \ , \text{ for all } \theta \in \Theta$$

Let $\psi(t) = E_\theta[U_n|T_n = t]$, then from Rao–Blackwell : $E_\theta[\psi(T_n)] = \theta$ and $Var_\theta[\psi(T_n)] \leq Var_\theta(U_n)$
But, by (a) : $P_\theta[\varphi(T_n) = \psi(T_n)] = 1$; hence : $Var_\theta[\varphi(T_n)] \leq Var_\theta(U_n)$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1; \theta)$
We know :

- $\sum_{i=1}^{n} X_i$ is sufficient for θ
- $\sum_{i=1}^{n} X_i \sim B(n; \theta)$, and this family is complete
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is unbiased estimator for θ
  Hence : $\overline{X}$ is UMVU-estimator for θ.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$ with σ^2 known.
We know :

- $\sum_{i=1}^{n} X_i$ is sufficient for μ
- $\sum_{i=1}^{n} X_i \sim N(n\mu; n\sigma^2)$, and this family is complete
- $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is unbiased estimator for μ
  Hence : $\overline{X}$ is UMVU-estimator for μ.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$ with μ known.
We have :

- $\sum_{i=1}^{n}(X_i - \mu)^2$ is sufficient for σ^2
- $\sum_{i=1}^{n}(X_i - \mu)^2 = \sigma^2.T$, where $T \sim \chi^2(n)$
  This family is complete (no proof)
- $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is unbiased estimator for σ^2
  Hence : $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is UMVU-estimator for σ^2.

2.3.8 The Exponential Class

There is a large class of densities for which there is a simple sufficient statistic (or set of sufficient statistics). This is the so called **exponential class** (or **Koopman-Darmois class**).

Definition

The density of a random variable X belongs to a **one-parameter exponential class** if it is of the form

$$f(x;\theta) = c(\theta)e^{q(\theta)t(x)}h(x) \quad , \qquad x \in I\!R$$
$$\theta \in \Theta \subset I\!R$$

where

$c(\theta) > 0$ for all $\theta \in \Theta$

$h(x) > 0$ for all $x \in S = \{x | f(x;\theta) > 0\}$,
and where S is assumed to be independent of θ.

Example
Let $X \sim B(n;\theta)$:

If we use the **notation** $I_A(x) = \begin{cases} 1 & \dots \text{ if } x \in A \\ 0 & \dots \text{ if } x \notin A \end{cases}$

then, we can write

$$\begin{aligned} f(x;\theta) &= \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} I_{\{0,1,\dots,n\}}(x) \quad , 0 < \theta < 1 \\ &= (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^x \begin{pmatrix} n \\ x \end{pmatrix} I_{\{0,1,\dots,n\}}(x) \end{aligned}$$

which is the exponential type with

$$c(\theta) = (1-\theta)^n, q(\theta) = ln\left(\frac{\theta}{1-\theta}\right), t(x) = x, h(x) = \begin{pmatrix} n \\ x \end{pmatrix} I_{\{0,1,\dots,n\}}(x).$$

Example

Let $X \sim N(\theta;\sigma^2), \sigma^2$ known :

$$f(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\theta^2}{2\sigma^2}} e^{\frac{\theta}{\sigma^2}x} e^{-\frac{1}{2\sigma^2}x^2} \quad , \theta \in I\!R$$

is of the exponential type with

$$c(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\theta^2}{2\sigma^2}}, q(\theta) = \frac{\theta}{\sigma^2}, t(x) = x, h(x) = e^{-\frac{1}{2\sigma^2}x^2}.$$

Example

Let $X \sim N(\mu;\theta),\ \mu$ known :

$$f(x;\theta) = \frac{1}{\sqrt{\theta}\sqrt{2\pi}} e^{-\frac{1}{2\theta}(x-\mu)^2} \quad , \theta > 0$$

is of the exponential type with

$$c(\theta) = \frac{1}{\sqrt{\theta}\sqrt{2\pi}}, q(\theta) = \frac{-1}{2\theta}, t(x) = (x-\mu)^2, h(x) = 1.$$

Examples of densities which **cannot** be written in this form are

- $f(x;\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$ $\qquad x \in I\!R, \theta \in I\!R$ (Cauchy)
- $f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \dots \text{ if } x \geq \theta \\ 0 & \dots \text{ if } x < \theta \end{cases}$ $\qquad , \theta \in I\!R$ (truncated exponential)

Properties

1. If $X_1, \dots, X_n$ is a random sample from X with density $f(x;\theta)$ of the form above, then the joint density of $X_1, \dots, X_n$ is given by
$$\prod_{i=1}^{n} f(x_i;\theta) = c^n(\theta) e^{q(\theta)\sum_{i=1}^{n} t(x_i)} h(x_1)\dots h(x_n)$$
2. From the factorization theorem, it follows :
$$T_n = \sum_{i=1}^{n} t(X_i) \text{ is a } \textbf{sufficient statistic for } \theta.$$
3. In both discrete and continuous case one can show that the density of T_n has the form
$$g(x;\theta) = c^n(\theta) e^{q(\theta)x} \widetilde{h}(x)$$
where $\widetilde{h}$ does not depend on θ.
(i.e. again the form of a one-parameter exponential family).
This is easy to see in the discrete case :
(but more difficult to prove in the continuous case)
$$\begin{aligned} g(x;\theta) &= P\left(\sum_{i=1}^{n} t(X_i) = x\right) \\ &= \sum\nolimits^{*} P(X_1 = x_1, \dots, X_n = x_n) \end{aligned}$$
where the sum $\sum^*$ is over all $(x_1, \dots, x_n)$ such that $\sum_{i=1}^{n} t(x_i) = x$.
Thus
$$\begin{aligned} g(x;\theta) &= \sum\nolimits^{*} f(x_1;\theta)\dots f(x_n;\theta) \\ &= \sum\nolimits^{*} c^n(\theta) e^{q(\theta)\sum_{j=1}^{n} t(x_j)} h(x_1)\dots h(x_n) \\ &= c^n(\theta) e^{q(\theta)x} \widetilde{h}(x) \end{aligned}$$

where $\widetilde{h}(x) = \sum^* h(x_1) \dots h(x_n)$.

4. One can also prove that the family of densities $\{g(x;\theta)|\theta \in \Theta\}$ is complete, provided Θ contains an open interval. Hence, in this case :

$$T_n = \sum_{i=1}^{n} t(X_i) \text{ is a } \textbf{complete sufficient statistic for } \theta$$

There exist more general exponential classes. The following definition generalizes to an r-dimensional parameter $\underset{\sim}{\theta} = (\theta_1, \dots, \theta_r)$ with $r \geq 1$, and at the same time to random vectors $\underset{\sim}{X} = (X_1, \dots, X_k)$, with $k \geq 1$.

Definition

The density of a random vector $\underset{\sim}{X} = (X_1, \dots, X_k)$ belongs to an **r-parameter exponential class** if it is of the form

$$f(\underset{\sim}{x};\underset{\sim}{\theta}) = c(\underset{\sim}{\theta}) e^{\sum_{i=1}^{r} q_i(\underset{\sim}{\theta}) t_i(\underset{\sim}{x})} h(\underset{\sim}{x}), \quad \underset{\sim}{x} = (x_1, \dots, x_k) \in I\!R^k$$

$$\underset{\sim}{\theta} = (\theta_1, \dots, \theta_r) \in \Theta \subset I\!R^r$$

where

$c(\underset{\sim}{\theta}) > 0$ for all $\underset{\sim}{\theta} \in \Theta$

$h(\underset{\sim}{x}) > 0$ for all $\underset{\sim}{x} \in S = \{\underset{\sim}{x}|f(\underset{\sim}{x};\underset{\sim}{\theta}) > 0\}$, and where S is assumed to be independent of $\underset{\sim}{\theta}$.

Example

Let $\underset{\sim}{X} = (X_1, \dots, X_k) \sim M(n; (\theta_1, \dots, \theta_k))$:
With $\underset{\sim}{x} = (x_1, \dots, x_k), \underset{\sim}{\theta} = (\theta_1, \dots, \theta_{k-1})$:

$$f(\underset{\sim}{x};\underset{\sim}{\theta}) = (1 - \theta_1 - \dots - \theta_{k-1})^n e^{\sum_{i=1}^{k-1} x_i ln\left(\frac{\theta_i}{1-\theta_1-\dots-\theta_{k-1}}\right)} \frac{n!}{x_1! \dots x_k!} I_A(\underset{\sim}{x})$$

where $A = \{\underset{\sim}{x} = (x_1, \dots, x_k)|x_1 \geq 0, \dots, x_k \geq 0 \ , \sum_{i=1}^{k} x_i = n\}$ is of the exponential type with

$c(\underset{\sim}{\theta}) = (1 - \theta_1 - \dots - \theta_{k-1})^n, q_i(\underset{\sim}{\theta}) = ln\left(\frac{\theta_i}{1-\theta_1-\dots-\theta_{k-1}}\right) \qquad (i = 1, \dots, k-1)$

$t_i(\underset{\sim}{x}) = x_i \ (i = 1, \dots, k-1), h(\underset{\sim}{x}) = \frac{n!}{x_1! \dots x_k!} I_A(\underset{\sim}{x})$.

Example

Let $X \sim N(\theta_1; \theta_2)$:

$$f(x; \underset{\sim}{\theta}) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} e^{-\frac{\theta_1^2}{2\theta_2}} e^{\frac{\theta_1}{\theta_2}x} e^{-\frac{1}{2\theta_2}x^2}, x \in I\!R$$

is of the exponential type with $c(\underset{\sim}{\theta}) = \frac{1}{\sqrt{\theta_2}\sqrt{2\pi}} e^{-\frac{\theta_1^2}{2\theta_2}}, q_1(\underset{\sim}{\theta}) = \frac{\theta_1}{\theta_2}$,
$q_2(\underset{\sim}{\theta}) = \frac{1}{2\theta_2}, t_1(x) = x, t_2(x) = -x^2, h(x) = 1.$

The properties of the 1-parameter exponential class have their analogues in this multi-parameter case.

2.4 General methods of point estimation

2.4.1 Maximum Likelihood Estimation

Introduction

The method of maximum likelihood is a routine procedure for obtaining estimators for unknown parameters from a set of data

$$x_1, x_2, \ldots, x_n.$$

We assume that these data are the observed values of a random sample

$$X_1, \ldots, X_n \overset{iid}{\sim} X.$$

Assume that the distribution function of X depends on some unknown parameter $\theta \in \Theta$. Let E be the event of observing $x_1, \ldots, x_n$. The probability of E can be determined from the model and, in general, it will depend on the unknown parameter θ. Denote it by $P_\theta(E)$.
The maximum likelihood estimate for θ is a value of θ which maximizes $P_\theta(E)$ over Θ. It is a parameter value which is "most likely" in the light of what has been observed.

If X is **discrete** (with density $f(x; \theta) = P_\theta(X = x)$)then

$$\begin{aligned} P_\theta(E) &= P_\theta(X_1 = x_1, \ldots, X_n = x_n) \\ &= \prod_{i=1}^{n} P_\theta(X_i = x_i) = \prod_{i=1}^{n} f(x_i; \theta). \end{aligned}$$

If X is **continuous** (with density $f(x;\theta)$) then $P(X = x) = 0$. Since always (with $F(x;\theta) = P_\theta(X \leq x)$, the distribution function of X)

$$P(X = x) = F(x;\theta) - \lim_{\substack{h\to 0 \\ >}} F(x-h;\theta)$$

we have the approximation for $h > 0$, small

$$\begin{aligned} P(X = x) &\approx F(x;\theta) - F(x-h;\theta) \\ &\approx f(x;\theta)h \end{aligned}$$

Since h does not depend on θ, we have (approximately) that maximization of $P_\theta(E)$ is equivalent to maximizing

$$\prod_{i=1}^{n} f(x_i;\theta)$$

Maximum Likelihood Estimation: One Parameter case

Let $X_1, \ldots, X_n \overset{iid}{\sim} X$. Suppose X has density $f(x;\theta)$ and that $\theta \in \Theta \subset I\!R$.

Definition

The **likelihood function** of $X_1, \ldots, X_n$ is the function

$$L(\theta, \underset{\sim}{x}) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

Example

Let X be a Bernoulli distribution with parameter $\theta \in [0,1]$.
Then : $f(x;\theta) = \theta^x(1-\theta)^{1-x} \qquad (x = 0, 1)$

and

$$L(\theta; \underset{\sim}{x}) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \qquad (\text{all } x_i = 0, 1)$$

Example

Let X be an exponential distribution with parameter $\theta \in]0, \infty[$
Then : $f(x;\theta) = \theta e^{-\theta x} \qquad (x > 0)$

and

$$L(\theta, \underset{\sim}{x}) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \qquad (\text{all } x_i > 0)$$

Definition

A value $\hat{\theta}_n$ (where $\hat{\theta}_n$ is a function of the observations $x_1, \ldots, x_n$; say $\hat{\theta}_n = t(x_1, \ldots, x_n)$) which maximizes the likelihood function $L(\theta; x_1, \ldots, x_n)$ over all $\theta \in \Theta$ is called a **maximum likelihood estimate (ML-estimate) for** θ.
Hence, for all $\theta \in \Theta$:

$$L(t(x_1, \ldots, x_n); x_1, \ldots, x_n) \geq L(\theta; x_1, \ldots, x_n).$$

The random variable $T_n = t(X_1, \ldots, X_n)$ is called a **maximum likelihood estimator (ML-estimator) for** θ.

Remarks

1. From the examples we will see that a ML-estimator is not necessarily **unique**.
2. From the examples we will see that a ML-estimator is not necessarily **unbiased**.
3. In many cases, but not always, the maximum of L can be found by **differentiation** methods.
4. Since L is a product, it is usually more convenient to maximize lnL (which is a sum of terms). Since ln is monotone, any value of θ which maximizes L, also maximizes lnL.

In the following, we will need more terminology.

Definition

The function

$$l(\theta; \underset{\sim}{x}) = l(\theta; x_1, \ldots, x_n) = lnL(\theta; x_1, \ldots, x_n)$$

is called the **log likelihood function** of $X_1, \ldots, X_n$.

Definition

The function

$$S(\theta; \underset{\sim}{x}) = S(\theta; x_1, \ldots, x_n) = \frac{\partial}{\partial \theta} l(\theta; \underset{\sim}{x})$$

is called the **score function** of $X_1, \ldots, X_n$.

Definition

The function

$$\begin{aligned} \mathcal{I}(\theta; \underset{\sim}{x}) = \mathcal{I}(\theta; x_1, \ldots, x_n) &= -\frac{\partial}{\partial \theta} S(\theta; \underset{\sim}{x}) \\ &= -\frac{\partial^2}{\partial \theta^2} l(\theta; \underset{\sim}{x}) \end{aligned}$$

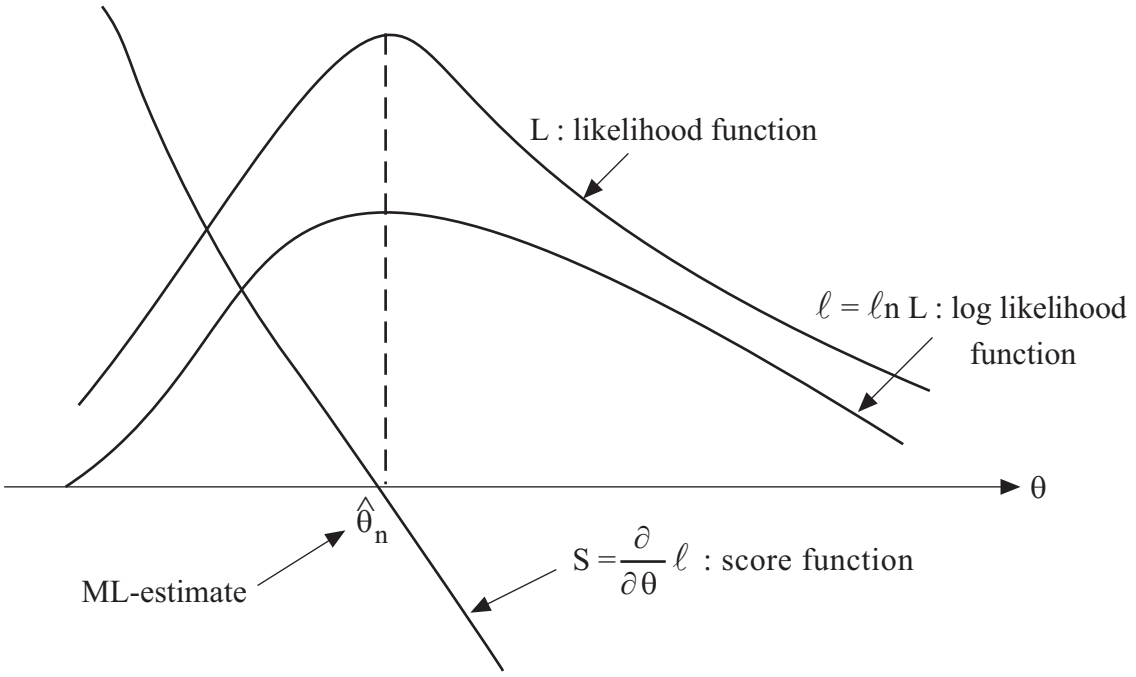
is called the **information function** of $X_1, \ldots, X_n$.

Hence, in many cases, $\hat{\theta}_n$ can be found by solving the **maximum likelihood equation** :

$$S(\theta; \underset{\sim}{x}) = 0$$

and by checking that

$$\mathcal{I}(\hat{\theta}_n; \underset{\sim}{x}) > 0.$$



Example

Let $X_1, \ldots, X_n$ be a random sample from X.
X be a Bernoulli distribution with parameter $\theta \in]0,1[$.
Then :

$$\begin{aligned} l(\theta; \underset{\sim}{x}) &= (\Sigma x_i) ln\theta + (n - \Sigma x_i) ln(1-\theta) \\ S(\theta; \underset{\sim}{x}) &= \frac{\Sigma x_i}{\theta} - \frac{n - \Sigma x_i}{1-\theta} \\ \mathcal{I}(\theta; \underset{\sim}{x}) &= \frac{\Sigma x_i}{\theta^2} + \frac{n - \Sigma x_i}{(1-\theta)^2} \end{aligned}$$

Solving $S(\theta; \underset{\sim}{x}) = 0$ gives $\hat{\theta}_n = \frac{1}{n}\Sigma x_i$ and since $\mathcal{I}(\hat{\theta}_n; \underset{\sim}{x}) > 0$, we have that $\hat{\theta}_n = \frac{1}{n}\Sigma x_i$ is the ML-estimate for θ and that $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ is the ML-estimator for θ.

Example

Let $X_1, \ldots, X_n$ be a random sample from X.
X be an exponential distribution with parameter $\theta \in]0, \infty[$.
Then :

$$\begin{aligned} l(\theta;\underset{\sim}{x}) &= nln\theta - \theta\Sigma x_i \\ S(\theta;\underset{\sim}{x}) &= \frac{n}{\theta} - \Sigma x_i \\ \mathcal{I}(\theta;\underset{\sim}{x}) &= \frac{n}{\theta^2} > 0 \end{aligned}$$

It follows that $\overset{\wedge}{\theta}_n = \frac{n}{\Sigma x_i}$ is the ML-estimate for θ and that $T_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}$ is the ML-estimator for θ.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with σ^2 known.
Put $\theta = \mu$. Then

$$\begin{aligned} l(\theta;\underset{\sim}{x}) &= -nln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2 \\ S(\theta;\underset{\sim}{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\theta) \\ \mathcal{I}(\theta;\underset{\sim}{x}) &= \frac{n}{\sigma^2} \end{aligned}$$

It follows that the ML-estimator for μ is : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with μ known.
Put $\theta = \sigma^2$. Then

$$\begin{aligned} l(\theta;\underset{\sim}{x}) &= -nln(\sqrt{2\pi}) - \frac{n}{2}ln\theta - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\mu)^2 \\ S(\theta;\underset{\sim}{x}) &= -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ \mathcal{I}(\theta;\underset{\sim}{x}) &= \frac{-n}{2\theta^2} + \frac{1}{\theta^3}\sum_{i=1}^{n}(x_i-\mu)^2 \end{aligned}$$

Solving $S(\theta;\underset{\sim}{x}) = 0$ gives $\overset{\wedge}{\theta}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$ and since $\mathcal{I}(\overset{\wedge}{\theta}_n;\underset{\sim}{x}) > 0$ it follows that

$T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is the ML-estimator for σ^2.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim Un[0, \theta], \theta > 0$.
We have :

$$L(\theta; \underset{\sim}{x}) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \ldots \text{if } 0 \leq x_1, \ldots, x_n \leq \theta \\ 0 & \ldots \text{if otherwise} \end{cases}$$

The maximum cannot be found by differentiation. But L is maximized by choosing θ as small as possible, i.e. $\hat{\theta}_n = \max(x_i)$. Hence the ML-estimator for θ is $T_n = \max(X_1, \ldots, X_n)$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim Un[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, with $\theta \in \mathbb{R}$ (uniform distribution with known range)
We have :

$$L(\theta; \underset{\sim}{x}) = \begin{cases} 1 & \ldots \text{if } \theta - \frac{1}{2} \leq x_1, \ldots x_n \leq \theta + \frac{1}{2} \\ 0 & \ldots \text{if otherwise} \end{cases}$$

So, L reaches its maximal value if

$$\theta - \frac{1}{2} \leq x_1, \ldots, x_n \leq \theta + \frac{1}{2}$$

$$\text{i.e. if} \quad \theta - \frac{1}{2} \leq \min(x_i) \text{ and } \max(x_i) \leq \theta + \frac{1}{2}$$

$$\text{or if} \quad \max(x_i) - \frac{1}{2} \leq \theta \leq \min(x_i) + \frac{1}{2}$$

Hence, there is no unique ML-estimator. Each estimator of the form

$$T_n = (1 - c)[\max(X_i) - \frac{1}{2}] + c[\min(X_i) + \frac{1}{2}]$$

with $0 \leq c \leq 1$ is ML-estimator for θ. A particular example is obtained for $c = \frac{1}{2} : T_n = \frac{1}{2}[\max(X_i) + \min(X_i)]$.

Maximum Likelihood Estimation: Multi-Parameter case

In many cases we have a random sample $X_1, \ldots, X_n$ from X with density $f(x; \theta_1, \ldots, \theta_k)$ depending on k real parameters. This means that we now have a density $f(x; \underset{\sim}{\theta})$ with $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset I\!R^k, k \geq 1$.

The generalizations of the definitions to this multi-parameter case are :
The likelihood function :

$$L(\underset{\sim}{\theta}; \underset{\sim}{x}) = L(\theta_1, \ldots, \theta_k; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \underset{\sim}{\theta})$$

The log likelihood function :

$$l(\underset{\sim}{\theta}; \underset{\sim}{x}) = lnL(\underset{\sim}{\theta}; \underset{\sim}{x})$$

The score function : is now a $k \times 1$ vector : **score vector :**

$$\begin{aligned} S(\underset{\sim}{\theta}; \underset{\sim}{x}) &= (S_1(\underset{\sim}{\theta}; \underset{\sim}{x}), \ldots, S_k(\underset{\sim}{\theta}, \underset{\sim}{x})) \\ &= \left(\frac{\partial}{\partial \theta_1} l(\underset{\sim}{\theta}; \underset{\sim}{x}), \ldots, \frac{\partial}{\partial \theta_k} l(\underset{\sim}{\theta}; \underset{\sim}{x}) \right) \end{aligned}$$

The information function : is now a $k \times k$ matrix : **information matrix :**

$$\begin{aligned} \mathcal{I}(\underset{\sim}{\theta}; \underset{\sim}{x}) &= (\mathcal{I}_{ij}(\underset{\sim}{\theta}, \underset{\sim}{x}))_{i,j=1,\ldots,k} \\ &= \left(-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\underset{\sim}{\theta}; \underset{\sim}{x}) \right)_{i,j=1,\ldots,k} \end{aligned}$$

Maximum likelihood estimate for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$: is now a vector $\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nk})$ where

$$\begin{aligned} \hat{\theta}_{n1} &= t_1(x_1, \ldots, x_n) \\ \ldots & \\ \hat{\theta}_{nk} &= t_k(x_1, \ldots, x_n) \end{aligned}$$

are such that for all $\underset{\sim}{\theta} \in \Theta$:

$$L(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{x}) \geq L(\underset{\sim}{\theta}; \underset{\sim}{x}).$$

Maximum likelihood estimator for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$: is the vector $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ where

$$\begin{array}{rcl} T_{n1} & = & t_1(X_1, \ldots, X_n) \\ \ldots & & \\ T_{nk} & = & t_k(X_1, \ldots, X_n) \end{array}$$

In certain cases : $\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{nk})$ is a solution of the k equations (**maximum likelihood equations**) :

$$\left\{ \begin{array}{l} \frac{\partial}{\partial \theta_1} L(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \\ \ldots \\ \frac{\partial}{\partial \theta_k} L(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \end{array} \right. \text{ or } \left\{ \begin{array}{l} \frac{\partial}{\partial \theta_1} l(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \\ \ldots \\ \frac{\partial}{\partial \theta_k} l(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \end{array} \right. \text{ or } \left\{ \begin{array}{l} S_1(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \\ \ldots \\ S_k(\underset{\sim}{\theta}; \underset{\sim}{x}) = 0 \end{array} \right.$$

The condition to have that $\hat{\underset{\sim}{\theta}}_n$ is a maximum is that the matrix $\mathcal{I}(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{X})$ is positive definite.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$ with μ and σ^2 unknown. The parameter $\underset{\sim}{\theta} = (\theta_1, \theta_2)$ with $\theta_1 = \mu, \theta_2 = \sigma^2$ is 2-dimensional :

$$\underset{\sim}{\theta} \in \Theta = \{(\theta_1, \theta_2) | \theta_1 \in I\!R, \theta_2 > 0\} \subset I\!R^2$$

We have

$$l(\underset{\sim}{\theta}; \underset{\sim}{x}) = -nln\sqrt{2\pi} - \frac{n}{2} ln\theta_2 - \frac{1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2$$

$$S(\underset{\sim}{\theta};\underset{\sim}{x}) = \left(\frac{1}{\theta_2}\sum_{i=1}^{n}(x_i-\theta_1) \quad , \quad -\frac{n}{2\theta_2}+\frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2 \right)$$

$$\mathcal{I}(\underset{\sim}{\theta};\underset{\sim}{x}) = \begin{pmatrix} \frac{n}{\theta_2} & \frac{1}{\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1) \\ \frac{1}{\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1) & -\frac{n}{2\theta_2^2}+\frac{1}{\theta_2^3}\sum_{i=1}^{n}(x_i-\theta_1)^2 \end{pmatrix}$$

The ML-equations :

$$\begin{cases} \sum(x_i-\theta_1)=0 \\ -n+\frac{1}{\theta_2}\sum(x_i-\theta_1)^2=0 \end{cases}$$

have a solution $\hat{\underset{\sim}{\theta}}_n = (\hat{\theta}_{n1}, \hat{\theta}_{n2})$ with

$$\begin{cases} \hat{\theta}_{n1} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \\ \hat{\theta}_{n2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = s^2 \end{cases}$$

and this is a maximum, since

$$\mathcal{I}(\hat{\underset{\sim}{\theta}}_n, \underset{\sim}{x}) = \begin{pmatrix} \frac{n}{s^2} & 0 \\ 0 & \frac{n}{2s^4} \end{pmatrix} \text{ is positive definite.}$$

Hence, the ML estimator for (μ, σ^2) is $(\overline{X}, S^2)$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

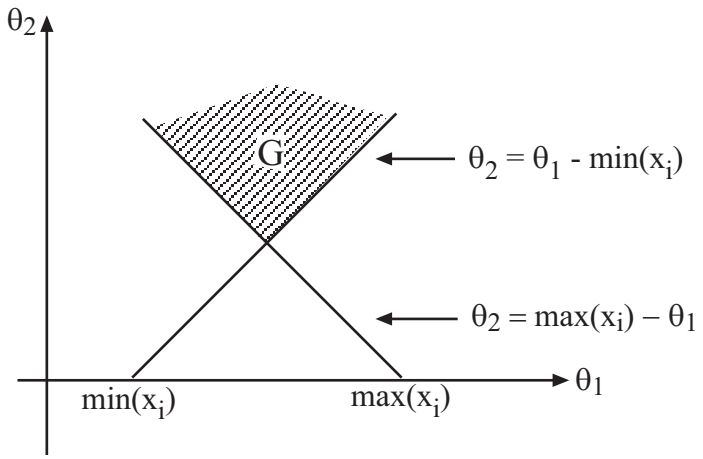
Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim Un[\theta_1-\theta_2, \theta_1+\theta_2]$ with $\theta_1 \in \mathbb{R}, \theta_2 > 0$.

$$L(\theta_1, \theta_2, \underset{\sim}{x}) = \begin{cases} \left(\frac{1}{2\theta_2}\right)^n & \dots \text{ if } \theta_1 - \theta_2 \leq x_1, \dots, x_n \leq \theta_1 + \theta_2 \\ 0 & \dots \text{ if otherwise} \end{cases}$$

$$= \begin{cases} \left(\frac{1}{2\theta_2}\right)^n & \dots \text{ if } \theta_2 \geq \theta_1 - \min(x_i) \text{ and } \theta_2 \geq \max(x_i) - \theta_1 \\ 0 & \dots \text{ if otherwise} \end{cases}$$

$$= \begin{cases} \left(\frac{1}{2\theta_2}\right)^n & \dots \text{ if } (\theta_1, \theta_2) \in G \text{ (see figure below)} \\ 0 & \dots \text{ if otherwise} \end{cases}$$



It is now clear that L is maximal when $\theta_2 = \theta_1 - \min(x_i) = \max(x_i) - \theta_1$, i.e.
$\theta_1 = \frac{1}{2}[\max(x_i) + \min(x_i)]$ and $\theta_2 = \frac{1}{2}[\max(x_i) - \min(x_i)]$.
Hence, the ML-estimator for (θ_1, θ_2) is (T_{n1}, T_{n2}), where $T_{n1} = \frac{1}{2}[\max(X_i) + \min(X_i)]$,
$T_{n2} = \frac{1}{2}[\max(X_i) - \min(X_i)]$.

Example

Let $\underset{\sim}{X} = (X_1, \dots, X_k) \sim M(n; (\theta_1, \dots, \theta_k))$. Here we have a single observation from a multivariate discrete density.
For $\underset{\sim}{x} = (x_1, \dots, x_k), x_1 \geq 0, \dots, x_k \geq 0$, integers, such that $x_1 + \dots + x_k = n$:

$$\begin{aligned} P(\underset{\sim}{X}=\underset{\sim}{x}) &= P(X_1=x_1,\ldots,X_k=x_k) \\ &= \frac{n!}{x_1!\ldots x_k!}\theta_1^{x_1}\theta_2^{x_2}\ldots\theta_k^{x_k} \\ &= \frac{n!}{x_1!\ldots x_k!}\theta_1^{x_1}\theta_2^{x_2}\ldots\theta_{k-1}^{x_{k-1}}(1-\theta_1-\ldots-\theta_{k-1})^{x_k} \\ &\equiv L(\theta_1,\ldots,\theta_{k-1};\underset{\sim}{x}) \end{aligned}$$

$$\begin{aligned} l(\theta_1,\ldots,\theta_{k-1};\underset{\sim}{x}) &= ln\left(\frac{n!}{x_1!\ldots x_k!}\right)+x_1ln\theta_1+\ldots \\ &\quad \ldots+x_{k-1}ln\theta_{k-1}+x_kln(1-\theta_1-\ldots-\theta_{k-1}). \end{aligned}$$

$$S(\theta_1,\ldots,\theta_{k-1};\underset{\sim}{x}) = \left(\frac{x_1}{\theta_1}-\frac{x_k}{1-\theta_1-\ldots-\theta_{k-1}},\quad\ldots\quad,\frac{x_{k-1}}{\theta_{k-1}}-\frac{x_k}{1-\theta_1-\ldots-\theta_{k-1}}\right)$$

$$\mathcal{I}(\theta_1,\ldots,\theta_{k-1};\underset{\sim}{x}) = \begin{pmatrix} \frac{x_1}{\theta_1^2}+\frac{x_k}{(1-\theta_1-\ldots-\theta_{k-1})^2} & & \\ & \ddots & \\ & & \frac{x_{k-1}}{\theta_{k-1}^2}+\frac{x_k}{(1-\theta_1-\ldots-\theta_{k-1})^2} \end{pmatrix}$$

$$\text{all other entries are } = \frac{x_k}{(1-\theta_1-\ldots-\theta_{k-1})^2}$$

ML-equations :

$$\begin{cases} \frac{x_1}{\theta_1}-\frac{x_k}{1-\theta_1-\ldots-\theta_{k-1}} &= 0 \\ \ldots & \\ \frac{x_{k-1}}{\theta_{k-1}}-\frac{x_k}{1-\theta_1-\ldots-\theta_{k-1}} &= 0 \end{cases}$$

Hence :

$$\frac{x_1}{\theta_1}=\frac{x_2}{\theta_2}=\ \ldots\ =\frac{x_{k-1}}{\theta_{k-1}}=\frac{x_k}{1-\theta_1-\ldots-\theta_{k-1}}=\frac{x_1+\ldots+x_k}{1}=n$$

or

$$\hat{\theta}_{n1}=\frac{x_1}{n},\ldots,\hat{\theta}_{n,k-1}=\frac{x_{k-1}}{n}.$$

This is a maximum, since

$$\mathcal{I}(\hat{\theta}_{n1},\ldots,\hat{\theta}_{n,k-1};\underset{\sim}{x}) = \begin{pmatrix} \frac{n^2}{x_1}+\frac{n^2}{x_k} & & \\ & \ddots & \\ & & \frac{n^2}{x_{k-1}}+\frac{n^2}{x_k} \end{pmatrix}$$

$$\text{all other entries } = \frac{n^2}{x_k}$$

is positive definite.
Conclusion : the ML-estimator for $(\theta_1,\ldots,\theta_k)$ is

$$(\frac{X_1}{n},\frac{X_2}{n},\ldots,\frac{X_k}{n}).$$

Example

Suppose we have a random sample of m observations $\underset{\sim}{X}_1=(X_{11},\ldots,X_{1k}),\underset{\sim}{X}_2=(X_{21},\ldots,X_{2k}),\ldots,\underset{\sim}{X}_m=(X_{m1},\ldots,X_{mk})$ from a multinomial distribution with parameters n and $(\theta_1,\ldots,\theta_k)$.
Show that the ML-estimator for $(\theta_1,\ldots,\theta_k)$ is

$$\left(\frac{\sum_{i=1}^{m}X_{i1}}{nm},\ldots,\frac{\sum_{i=1}^{m}X_{ik}}{nm}\right).$$

The Score Statistics

We defined the score function $S(\theta;\underset{\sim}{x})$ in the 1-parameter case and the score vector $S(\underset{\sim}{\theta};\underset{\sim}{x})$ in the k-parameter case. Recall

$$\begin{aligned} S(\theta;\underset{\sim}{x}) &= S(\theta;x_1,\ldots,x_n) \\ &= \frac{\partial}{\partial\theta}l(\theta;\underset{\sim}{x}) = \frac{\partial}{\partial\theta}lnL(\theta;\underset{\sim}{x}) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}lnf(x_i;\theta) \\ S(\underset{\sim}{\theta};\underset{\sim}{x}) &= S(\underset{\sim}{\theta};x_1,\ldots,x_n) \\ &= (S_1(\underset{\sim}{\theta};\underset{\sim}{x}),\ldots,S_k(\underset{\sim}{\theta};\underset{\sim}{x})) \\ &= (\frac{\partial}{\partial\theta_1}l(\underset{\sim}{\theta};\underset{\sim}{x}),\ldots,\frac{\partial}{\partial\theta_k}l(\underset{\sim}{\theta};\underset{\sim}{x})). \end{aligned}$$

Definition

The random variable $S(\theta; \underset{\sim}{X}) = S(\theta; X_1, \ldots, X_n)$ (or the random vector $S(\underset{\sim}{\theta}; \underset{\sim}{X}) = S(\underset{\sim}{\theta}; X_1, \ldots, X_n)$) is called the **score statistic**.

(a) Mean and variance of the score statistic : one parameter case

Since the score statistic is a sum of i.i.d. random variables :

$$S(\theta; \underset{\sim}{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} lnf(X_i; \theta)$$

we have :

$$\begin{aligned} E[S(\theta; \underset{\sim}{X})] &= nE[\frac{\partial}{\partial\theta} lnf(X; \theta)] \\ Var[S(\theta; \underset{\sim}{X})] &= nVar[\frac{\partial}{\partial\theta} lnf(X; \theta)] \end{aligned}$$

Theorem

Under regularity conditions :

(i) $E[\frac{\partial}{\partial\theta} lnf(X; \theta)] = 0$

(ii) $E[(\frac{\partial}{\partial\theta} lnf(X; \theta))^2] = -E[\frac{\partial^2}{\partial\theta^2} lnf(X; \theta)]$

'Proof'

We sketch the proof for the case where X is discrete. (In the continuous case : replace all sums by integrals)

(i) First note that $\frac{\partial}{\partial\theta} lnf(x; \theta) = \frac{\frac{\partial}{\partial\theta} f(x; \theta)}{f(x; \theta)}$

and hence

$$\frac{\partial}{\partial\theta} lnf(x; \theta).f(x; \theta) = \frac{\partial}{\partial\theta} f(x; \theta)$$

Now

$$\begin{aligned} E[\frac{\partial}{\partial\theta}lnf(X;\theta)] &= \sum_x \frac{\partial}{\partial\theta}lnf(x;\theta).f(x;\theta) \\ &= \sum_x \frac{\partial}{\partial\theta}f(x;\theta) \\ &= \frac{\partial}{\partial\theta}\sum_x f(x;\theta) \qquad \text{(under regularity conditions)} \\ &= \frac{\partial}{\partial\theta}(1) = 0. \end{aligned}$$

(ii)

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2}lnf(x;\theta) &= \frac{f(x;\theta)\frac{\partial^2}{\partial\theta^2}f(x;\theta) - (\frac{\partial}{\partial\theta}f(x;\theta))^2}{f^2(x;\theta)} \\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)} - (\frac{\partial}{\partial\theta}lnf(x;\theta))^2 \end{aligned}$$

Now

$$\begin{aligned} E[\frac{\partial^2}{\partial\theta^2}lnf(X;\theta)] &= \sum_x (\frac{\partial^2}{\partial\theta^2}lnf(x;\theta)).f(x;\theta) \\ &= \sum_x \frac{\partial^2}{\partial\theta^2}f(x;\theta) - \sum_x (\frac{\partial}{\partial\theta}lnf(x;\theta))^2.f(x;\theta) \\ &= -E[(\frac{\partial}{\partial\theta}lnf(X;\theta))^2], \qquad \text{since} \end{aligned}$$

$$\begin{aligned} \sum_x \frac{\partial^2}{\partial\theta^2}f(x;\theta) &= \frac{\partial}{\partial\theta}\sum_x \frac{\partial}{\partial\theta}f(x;\theta) \qquad \text{(under regularity conditions)} \\ &= \frac{\partial}{\partial\theta}(0) = 0. \end{aligned}$$

□

Note : as can be seen from the proof, the **regularity conditions** for the above result are concerned with the possibility of interchanging differentiation and summation (or integration).

Definition

The quantity

$$i(\theta) = E[(\frac{\partial}{\partial\theta} lnf(X;\theta))^2]$$

is called the **Fisher information number**.

Corollary

Under regularity conditions :

$$\begin{aligned} E[S(\theta; \underset{\sim}{X})] &= 0 \\ Var[S(\theta; \underset{\sim}{X})] &= ni(\theta) = E[\mathcal{I}(\theta; \underset{\sim}{X})] \end{aligned}$$

Indeed :

$$\mathcal{I}(\theta; \underset{\sim}{x}) = -\frac{\partial^2}{\partial\theta^2} l(\theta; \underset{\sim}{x}) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} lnf(x_i;\theta)$$

so that : $E[\mathcal{I}(\theta; \underset{\sim}{X})] = -nE[\frac{\partial^2}{\partial\theta^2} lnf(X;\theta)] = ni(\theta)$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with σ^2 known.
Put $\theta = \mu$. Then :
$\frac{\partial}{\partial\theta} lnf(x;\theta) = \frac{1}{\sigma^2}(x-\theta); i(\theta) = \frac{1}{\sigma^4}E[(X-\theta)^2] = \frac{1}{\sigma^2}; S(\theta; \underset{\sim}{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \theta) =$
$\frac{n}{\sigma^2}(\overline{X} - \theta)$; hence : $E(S(\theta; \underset{\sim}{X})] = 0$ and $Var[S(\theta; \underset{\sim}{X})] = \frac{n^2}{\sigma^4}.\frac{\sigma^2}{n} = \frac{n}{\sigma^2}$;
$\mathcal{I}(\theta; \underset{\sim}{X}) = \frac{n}{\sigma^2}; E[\mathcal{I}(\theta; \underset{\sim}{X})] = \frac{n}{\sigma^2}$.

(b) Mean vector and variance-covariance matrix of the score vector : multi-parameter case

The analogue for the multi-parameter case of the Fisher information number is the Fisher information matrix.

Definition

The $k \times k$ matrix

$$\begin{aligned} B(\underset{\sim}{\theta}) &= (B_{ij}(\underset{\sim}{\theta}))_{i,j=1,\ldots,k} \\ &= \left(E[\frac{\partial}{\partial\theta_i} lnf(X;\underset{\sim}{\theta}).\frac{\partial}{\partial\theta_j} lnf(X;\underset{\sim}{\theta})]\right)_{i,j=1,\ldots,k} \end{aligned}$$

is called the **Fisher information matrix**.
As before we have :

Theorem

Under regularity conditions :

(i) $E[\frac{\partial}{\partial\theta_i} lnf(X;\underset{\sim}{\theta})] = 0 \qquad$ *for* $i = 1, \ldots, k$

(ii) $E[\frac{\partial}{\partial\theta_i} lnf(X;\underset{\sim}{\theta}).\frac{\partial}{\partial\theta_j} lnf(X;\underset{\sim}{\theta})]$
$= -E[\frac{\partial^2}{\partial\theta_i\partial\theta_j} lnf(X;\underset{\sim}{\theta})] \qquad$ *for* $i, j = 1, \ldots, k.$

Corollary

Under regularity conditions :

- *the mean vector of* $S(\underset{\sim}{\theta};\underset{\sim}{X})$ *is the zero vector*
- *the variance-covariance matrix of* $S(\underset{\sim}{\theta};\underset{\sim}{X})$ *is*

$$nB(\underset{\sim}{\theta}) = \left(E[\mathcal{I}_{ij}(\underset{\sim}{\theta};\underset{\sim}{X})]\right)_{i,j=1,\ldots,k}.$$

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with μ and σ^2 unknown. Put $\underset{\sim}{\theta} = (\theta_1, \theta_2)$, where $\theta_1 = \mu, \theta_2 = \sigma^2$.

$$\begin{aligned} \frac{\partial}{\partial\theta_1} lnf(x;\underset{\sim}{\theta}) &= \frac{1}{\theta_2}(x - \theta_1) \\ \frac{\partial}{\partial\theta_2} lnf(x;\underset{\sim}{\theta}) &= -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2}(x - \theta_1)^2. \end{aligned}$$

Calculate the Fisher information matrix :

$$B(\underset{\sim}{\theta}) = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix}$$

Compare with the expected values of the matrix $\mathcal{I}(\underset{\sim}{\theta}; \underset{\sim}{X})$ that has been calculated before.

Iterative procedures for calculation of ML-Estimates

It is not always possible to obtain a closed expression for the ML-estimate. We therefore present some numerical procedures for solving ML-equations.

Let $X_1, \ldots, X_n$ be a random sample from $X \sim$ Cauchy with parameter θ :

$$\begin{aligned} f(x;\theta) &= \frac{1}{\pi}\,\frac{1}{1+(x-\theta)^2} \\ l(\theta;\underset{\sim}{x}) &= -nln\pi + \sum_{i=1}^{n} ln\left(\frac{1}{1+(x_i-\theta)^2}\right) \\ S(\theta;\underset{\sim}{x}) &= 2\sum_{i=1}^{n}\frac{x_i-\theta}{1+(x_i-\theta)^2}. \end{aligned}$$

The ML-equation does not allow a solution in closed form. Numerical methods are necessary.

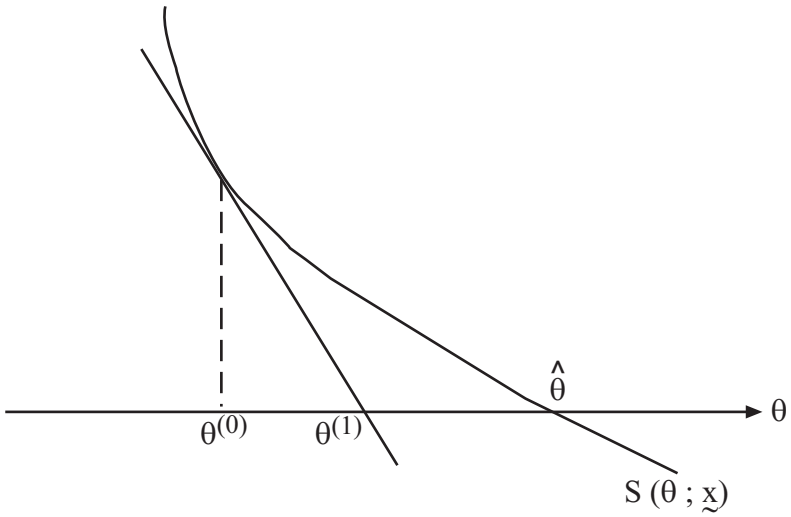
(a) Solving the ML-equation by Newton's method : one-parameter case Suppose we have to find a solution $\hat{\theta}$ of the ML-equation

$$S(\theta;\underset{\sim}{x}) = 0$$

Let $\theta^{(0)}$ be an initial guess, close to $\hat{\theta}$.
By Taylor series expansion :

$$S(\theta;\underset{\sim}{x}) \approx S(\theta^{(0)};\underset{\sim}{x}) + (\theta-\theta^{(0)})(-\mathcal{I}(\theta^{(0)};\underset{\sim}{x}))$$

That is : we approximate $S(\theta;\underset{\sim}{x})$ by a linear function of θ which has the same value and the same slope as $S(\theta;\underset{\sim}{x})$ in $\theta = \theta^{(0)}$.



Solving

$$S(\theta^{(0)}; \underset{\sim}{x}) + (\theta - \theta^{(0)})(-\mathcal{I}(\theta^{(0)}; \underset{\sim}{x})) = 0$$

gives a solution $\theta^{(1)}$ for θ :

$$\theta^{(1)} = \theta^{(0)} + \frac{S(\theta^{(0)}; \underset{\sim}{x})}{\mathcal{I}(\theta^{(0)}; \underset{\sim}{x})}$$

This solution $\theta^{(1)}$ is taken as a new initial guess and the calculations are repeated. That is : we obtain an **iterative procedure** given by

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)}; \underset{\sim}{x})}{\mathcal{I}(\theta^{(i)}; \underset{\sim}{x})} \qquad\qquad i = 0, 1, 2, \ldots$$

In this way we construct a sequence $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$. We stop as soon as $\theta^{(i+1)} \approx \theta^{(i)}$, in which case $S(\theta^{(i)}, \underset{\sim}{x}) \approx 0$, and a root has been found.

Modification: Fisher's scoring method

A simplification is to use $E_{\theta^{(i)}}[\mathcal{I}(\theta^{(i)}; \underset{\sim}{X})]$ instead of $\mathcal{I}(\theta^{(i)}; \underset{\sim}{X})$. From above it also follows that in most cases this also equals $ni(\theta^{(i)})$.

The fully modified procedure then becomes

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)}; \underset{\sim}{X})}{ni(\theta^{(i)})} \qquad i = 0, 1, 2, \ldots$$

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim$ Cauchy with parameter θ; i.e. $f(x;\theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$.
A calculation shows :

$$i(\theta) = \frac{4}{\pi} \int_{-\infty}^{\infty} \frac{t^2}{(1+t^2)^3} dt = \frac{1}{2}$$

which is independent of θ. In this case, the iterative procedure is simply

$$\theta^{(i+1)} = \theta^{(i)} + \frac{2}{n} S(\theta^{(i)}; \underset{\sim}{x}) \qquad i = 0, 1, 2, \ldots$$

(b) Solving the ML-equations by the Newton-Raphson method : multiparameter case We illustrate the method for the 2-parameter case ($k = 2$). We need to find a solution $(\hat{\theta}_1, \hat{\theta}_2)$ of the simultaneous equations

$$\begin{cases} S_1(\theta_1, \theta_2; \underset{\sim}{x}) &= 0 \\ S_2(\theta_1, \theta_2; \underset{\sim}{x}) &= 0 \end{cases}$$

Let $(\theta_1^{(0)}, \theta_2^{(0)})$ be a preliminary guess and consider the linear approximations (by bivariate Taylor expansions) :

$$\begin{aligned} S_1(\theta_1, \theta_2; \underset{\sim}{x}) &\approx S_1(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) + (\theta_1 - \theta_1^{(0)}) \frac{\partial S_1}{\partial \theta_1}(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) + (\theta_2 - \theta_2^{(0)}) \frac{\partial S_1}{\partial \theta_2}(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) \\ S_2(\theta_1, \theta_2; \underset{\sim}{x}) &\approx S_2(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) + (\theta_1 - \theta_1^{(0)}) \frac{\partial S_2}{\partial \theta_1}(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) + (\theta_2 - \theta_2^{(0)}) \frac{\partial S_2}{\partial \theta_2}(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x}) \end{aligned}$$

Solving the system (with obvious abbreviations) :

$$\begin{cases} S_1 + (\theta_1 - \theta_1^{(0)})\dfrac{\partial S_1}{\partial \theta_1} + (\theta_2 - \theta_2^{(0)})\dfrac{\partial S_1}{\partial \theta_2} &= 0 \\ S_2 + (\theta_1 - \theta_1^{(0)})\dfrac{\partial S_2}{\partial \theta_1} + (\theta_2 - \theta_2^{(0)})\dfrac{\partial S_2}{\partial \theta_2} &= 0 \end{cases}$$

or :

$$(\theta_1 - \theta_1^{(0)} \quad \theta_2 - \theta_2^{(0)}) \begin{pmatrix} -\dfrac{\partial S_1}{\partial \theta_1} & -\dfrac{\partial S_2}{\partial \theta_1} \\ -\dfrac{\partial S_1}{\partial \theta_2} & -\dfrac{\partial S_2}{\partial \theta_2} \end{pmatrix} = (S_1 \quad S_2)$$

gives

$$(\theta_1 - \theta_1^{(0)} \quad \theta_2 - \theta_2^{(0)}) = (S_1 \quad S_2) \begin{pmatrix} -\dfrac{\partial S_1}{\partial \theta_1} & -\dfrac{\partial S_2}{\partial \theta_1} \\ -\dfrac{\partial S_1}{\partial \theta_2} & -\dfrac{\partial S_2}{\partial \theta_2} \end{pmatrix}^{-1}$$

or :

$$(\theta_1 \quad \theta_2) = (\theta_1^{(0)} \quad \theta_2^{(0)}) + (S_1 \quad S_2) \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{21} \\ \mathcal{I}_{12} & \mathcal{I}_{22} \end{pmatrix}^{-1}$$

where S_1, S_2, and the $\mathcal{I}_{ij}$ have to be evaluated at $(\theta_1^{(0)}, \theta_2^{(0)}; \underset{\sim}{x})$. This has to be applied repeatedly until convergence is obtained.
The **iterative procedure** has the following form :

$$(\theta_1^{(i+1)} \quad \theta_2^{(i+1)}) = (\theta_1^{(i)} \quad \theta_2^{(i)}) + (S_1 \quad S_2) \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{21} \\ \mathcal{I}_{12} & \mathcal{I}_{22} \end{pmatrix}^{-1}$$

where the S_1, S_2 and the $\mathcal{I}_{ij}$ on the right hand side have to be evaluated at $(\theta_1^{(i)}, \theta_2^{(i)}; \underset{\sim}{x})$.
Modification: Fisher's scoring method

The analogous modification as in the 1-parameter case also applies here : one can replace the $\mathcal{I}_{jk}(\underset{\sim}{\theta}^{(i)}; \underset{\sim}{x})$ by $E_{\underset{\sim}{\theta}^{(i)}}[\mathcal{I}_{jk}(\underset{\sim}{\theta}^{(i)}; \underset{\sim}{X})]$ (i.e. by $nB(\underset{\sim}{\theta}^{(i)})$, in most cases).
The modified procedure then becomes:

$$\underset{\sim}{\theta}^{(i+1)} = \underset{\sim}{\theta}^{(i)} + \frac{S(\underset{\sim}{\theta}^{(i)}; \underset{\sim}{x})}{n} B^{-1}(\underset{\sim}{\theta}^{(i)}) \qquad\qquad i = 0, 1, 2, \ldots$$

(c) The Expectation-Maximization (EM) algorithm The EM algorithm is an iterative procedure for ML-estimation in problems with incomplete data. The term EM was introduced by Dempster, Laird and Rubin (1977) and the algorithm is closely related to the following old and simple iterative idea: (1) replace the missing data by estimated values; (2) estimate parameters; (3) reestimate the missing values assuming the parameters are correct (4) reestimate the parameters, etc... until convergence. Hence each iteration of the algorithm consists of two steps: the E step (**expectation step**) and the M step (**maximization step**).
In the E step, the conditional expectations of the "missing data" are calculated, given the observed data and the current parameter estimates. These expected values are substituted for the "missing data", to complete the set of observations. In the M step, maximum likelihood estimation is done using the completed set of observations.
(the quotes around "missing data" refer to the fact that the missing values themselves are not necessarily being subsituted but rather missing sufficient statistics (certain functions of the missing values)).
Suppose that $\underset{\sim}{x} = (x_1, \ldots, x_n)$ denotes the complete data coming from a sample $\underset{\sim}{X} = (X_1, \ldots, X_n)$. The random vector $\underset{\sim}{X}$ is not observed. Instead we observe a random vector $\underset{\sim}{Y}$ (the incomplete data) that is the image of $\underset{\sim}{X}$ under some many-to-one transformation.

For example

- $\underset{\sim}{X} = (X_1, X_2, X_3, X_4)$ with a multinomial distribution
  $\underset{\sim}{Y} = (X_1, X_2, X_3 + X_4)$ (collapsing of two cells)
- $\underset{\sim}{X} = (X_1, \ldots, X_n) = (\underset{\sim}{X}^{obs}, \underset{\sim}{X}^{mis})$
  (where $\underset{\sim}{X}^{obs}$ and $\underset{\sim}{X}^{mis}$ are the observed, resp. missing part of the sample.)
  $\underset{\sim}{Y} = \underset{\sim}{X}^{obs}$

Let $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = f_{\underset{\sim}{X}}(\underset{\sim}{x}; \theta)$ denote the density of $\underset{\sim}{X}$ and let $f_{\underset{\sim}{Y}}(\underset{\sim}{y}; \theta)$ denote the density of $\underset{\sim}{Y}$. Since

$$f_{\underset{\sim}{X}}(\underset{\sim}{x}; \theta) = f_{\underset{\sim}{Y}}(\underset{\sim}{y}; \theta) f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x}|\underset{\sim}{y}; \theta)$$

we have for the log likelihood functions:

$$l_{\underset{\sim}{X}}(\theta; \underset{\sim}{x}) = l_{\underset{\sim}{Y}}(\theta; \underset{\sim}{y}) + \ln f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x}|\underset{\sim}{y}; \theta)$$

or

$$l_{\underset{\sim}{Y}}(\theta; \underset{\sim}{y}) = l_{\underset{\sim}{X}}(\theta; \underset{\sim}{x}) - \ln f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x}|\underset{\sim}{y}; \theta)$$

A natural estimator for θ is the maximizer $\widehat{\theta}$ of the left hand side (the observed likelihood). The first term of the right hand side is the complete-data log likelihood and the second

term of the right hand side is the missing part of the complete-data log likelihood. It cannot be calculated because $\underset{\sim}{X}$ is unobserved. Therefore we will take the conditional expectation, given the observed data $\underset{\sim}{Y}$ and some preliminary estimate $\theta^{(j)}$ of θ. Denote

$$\begin{aligned} Q(\theta \mid \theta^{(j)}) &= \int \ell_{\underset{\sim}{X}}(\theta; \underset{\sim}{x}) f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x} \mid \underset{\sim}{y}; \theta^{(j)}) d\underset{\sim}{x} \\ H(\theta \mid \theta^{(j)}) &= \int \ln f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x} \mid \underset{\sim}{y}; \theta) f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x} \mid \underset{\sim}{y}; \theta^{(j)}) d\underset{\sim}{x}. \end{aligned}$$

It then follows that

$$Q(\theta \mid \theta^{(j)}) = \ell_{\underset{\sim}{Y}}(\theta; \underset{\sim}{y}) + H(\theta \mid \theta^{(j)})$$

It can be shown that, if $\theta^{(j+1)}$ maximizes $Q(\theta \mid \theta^{(j)})$, then $\ell_{\underset{\sim}{Y}}(\theta^{(j+1)}; \underset{\sim}{y}) \geq \ell_{\underset{\sim}{Y}}(\theta^{(j)}; \underset{\sim}{y})$. This means that $\theta^{(j+1)}$ is a better estimate than $\theta^{(j)}$. This leads to the following algorithm (**EM algorithm**):

Start with initial value $\theta^{(0)}$. Let $\theta^{(j)}$ be the estimate at iteration j. Then iteration $(j+1)$ goes as follows:
E-step*: Calculate the expected complete-data log likelihood $Q(\theta \mid \theta^{(j)})$.*
M-step*: Find $\theta^{(j+1)}$ as maximizer of $Q(\theta \mid \theta^{(j)})$.*

In well behaved cases, the EM algorithm produces a sequence $\theta^{(0)}, \theta^{(1)}, \ldots$ of estimates of the parameter θ, that converges to $\widehat{\theta}$, where $\widehat{\theta}$ is the unique maximizer of the observed log likelihood $\ell_{\underset{\sim}{Y}}(\theta; \underset{\sim}{y})$.

Let us have a closer look at this in the important case where $\underset{\sim}{X} = (X_1, \ldots, X_n)$ is a random sample from a density $f(x; \theta)$ of the "**exponential family**" form

$$f(x; \theta) = e^{\theta t(x) - a(\theta) + b(x)}$$

where $x \in \Re, \theta \in \Theta \subset \Re,\ t(x) > 0$ for all $x \in S = \{x | f(x; \theta) > 0\}$ and where S does not depend on θ.
The log likelihood function of $\underset{\sim}{X}$ is given by

$$l_{\underset{\sim}{X}}(\theta; \underset{\sim}{x}) = \theta T(\underset{\sim}{x}) - na(\theta) + B(\underset{\sim}{x})$$

where $T(\underset{\sim}{x}) = \sum_{i=1}^{n} t(x_i)$ and $B(\underset{\sim}{x}) = \sum_{i=1}^{n} b(x_i)$.
For the E-step:

$$\begin{aligned} Q(\theta \mid \theta^{(j)}) &= \int [\theta T(\underset{\sim}{x}) - na(\theta) + B(\underset{\sim}{x})] f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x} \mid \underset{\sim}{y}; \theta^{(j)}) d\underset{\sim}{x} \\ &= \theta E_{\theta^{(j)}}[T(\underset{\sim}{X}) \mid \underset{\sim}{Y}] - na(\theta) + \int B(\underset{\sim}{x}) f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{x} \mid \underset{\sim}{y}; \theta^{(j)}) d\underset{\sim}{x}. \end{aligned}$$

For the M-step:
Since the integral in the above expression does not involve θ, we have that maximization of $Q(\theta \mid \theta^{(j)})$ w.r.t. θ is the same as the maximization of $\ell_{\underset{\sim}{X}}(\theta; \underset{\sim}{x})$, but where $T(\underset{\sim}{x})$ (th sufficient statistic) has been replaced by $E_{\theta^{(j)}}[T(\underset{\sim}{X}) \mid \underset{\sim}{Y}]$.

Some interpretation
We generally have that

$$E_\theta\left(\frac{\partial}{\partial\theta} l_{\underset{\sim}{X}}(\theta;\underset{\sim}{X})\right) = 0 \text{and } E_\theta\left(\frac{\partial}{\partial\theta}\ln f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{X}|\underset{\sim}{Y},\theta)|\underset{\sim}{Y}\right) = 0.$$

This gives, in our situation,

$$E_\theta(T(X)) = na'(\theta) \text{and} \frac{\partial}{\partial\theta} l_{\underset{\sim}{Y}}(\theta;\underset{\sim}{Y}) = E_\theta[T(\underset{\sim}{X})|\underset{\sim}{Y}] - na'(\theta)$$

and from these two equations it follows that

$$\frac{\partial}{\partial\theta}\ell_{\underset{\sim}{Y}}(\theta;\underset{\sim}{Y}) = E_\theta[T(\underset{\sim}{X})|\underset{\sim}{Y}] - E_\theta[T(\underset{\sim}{X})].$$

Because $\dfrac{\partial}{\partial\theta} l_{\underset{\sim}{Y}}(\theta;\underset{\sim}{Y}) = 0$ for $\theta = \widehat{\theta}$, we have

$$E_{\widehat{\theta}}[T(\underset{\sim}{X})] = E_{\widehat{\theta}}[T(\underset{\sim}{X})|\underset{\sim}{Y}].$$

Hence the solution $\widehat{\theta}$ can be characterized as that value of the parameter under which the conditional expectation of $T(\underset{\sim}{X})$ given $\underset{\sim}{Y}$ is the same as the unconditional expectation.
There is also an interpretation for the sequence of EM approximations $\theta^{(0)}, \theta^{(1)}, \ldots$. If $\theta^{(j)}$ is the current estimate for θ, then the maximizer $\theta^{(j+1)}$ of $Q(\theta \mid \theta^{(j)})$ satisfies

$$E_{\theta^{(j)}}[T(\underset{\sim}{X})|\underset{\sim}{Y}] = na'(\theta^{(j+1)}).$$

Since also

$$E_{\theta^{(j+1)}}[T(\underset{\sim}{X})] = na'(\theta^{(j+1)})$$

we also have that

$$E_{\theta^{(j+1)}}[T(\underset{\sim}{X})] = E_{\theta^{(j)}}[T(\underset{\sim}{X})|\underset{\sim}{Y}].$$

To obtain a graphical interpretation, we now show that both $E_\theta(T(\underset{\sim}{X}))$ and $E_\theta[T(\underset{\sim}{X})|\underset{\sim}{Y}]$ are increasing functions of θ and that the first increases more rapidly than the second. Indeed, from above

$$\begin{aligned}\frac{\partial}{\partial\theta}E_\theta(T(\underset{\sim}{X})) &= na''(\theta)\\ \frac{\partial}{\partial\theta}E_\theta[T(\underset{\sim}{X})|\underset{\sim}{Y}] &= na''(\theta) + \frac{\partial^2}{\partial\theta^2} l_Y(\theta;\underset{\sim}{Y}).\end{aligned}$$

To show that both expressions are positive, we recall that, in general:

$$E_\theta\left(\frac{\partial^2}{\partial\theta^2} l_{\underset{\sim}{X}}(\theta;\underset{\sim}{X})\right) < 0$$

and

$$E_\theta\left(\frac{\partial^2}{\partial\theta^2}\ln f_{\underset{\sim}{X}|\underset{\sim}{Y}}(\underset{\sim}{X}|\underset{\sim}{Y},\theta)|\underset{\sim}{Y}\right) < 0.$$

This gives, in our situation,

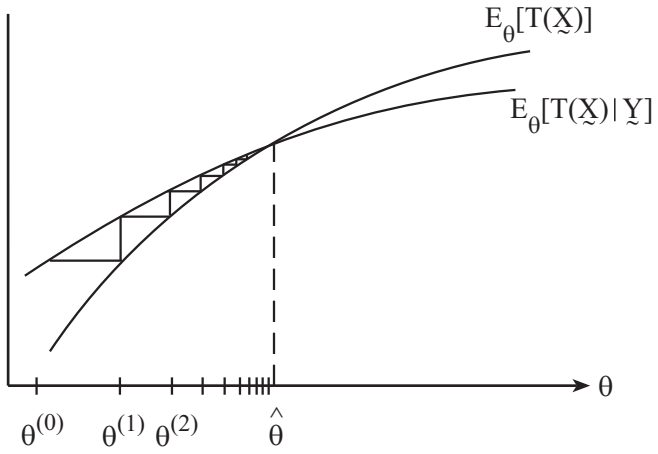
$$na''(\theta) > 0$$

and

$$na''(\theta) + \frac{\partial^2}{\partial\theta^2} l_{\underset{\sim}{Y}}(\theta; \underset{\sim}{Y}) > 0.$$

Moreover, since $\frac{\partial^2}{\partial\theta^2} l_{\underset{\sim}{Y}}(\theta; \underset{\sim}{Y}) < 0$ in some neighborhood of $\widehat{\theta}$, we have that $E_\theta[T(\underset{\sim}{X})]$ has the largest slope.

This is graphically illustrated as follows:



We see that the sequence of EM approximations converges monotonically to $\widehat{\theta}$.

The above method and properties of the EM algorithm also hold (under regularly conditions on the likelihood function) outside the specific situation of exponential families. Also the parameter θ may be multidimensional.
We do not present the theoretical properties on convergence of the EM algorithm, but rather give some (simple) examples.

Example
Let $X_1, \ldots, X_n$ be a random sample from $X \sim Exp(\theta), \theta > 0$.

Suppose $X_1, \ldots, X_m$ are observed
$X_{m+1}, \ldots, X_n$ are missing.
This is the simplest situation: in a random sample some units are missing. It is in fact just a reduction in sample size from n to m.

In our notation:

$$\begin{aligned} \underset{\sim}{X} &= (X_1, \ldots, X_n) \\ \underset{\sim}{Y} &= (X_1, \ldots, X_m) \\ l_{\underset{\sim}{X}}(\theta; \underset{\sim}{x}) &= -\theta \sum_{i=1}^{n} x_i + n \ln \theta \\ T(\underset{\sim}{x}) &= -\sum_{i=1}^{n} x_i \\ a(\theta) &= -\ln \theta. \end{aligned}$$

We have

$$E_\theta(T(\underset{\sim}{X})) = -\frac{n}{\theta}$$

and for $i = m+1, \ldots, n$

$$E_\theta(X_i|\underset{\sim}{Y}) = E(X_i|X_1, \ldots, X_m) = E(X_i) = \frac{1}{\theta}.$$

Let $\theta^{(0)}$ be some initial estimate for θ and $\theta^{(j)} (j = 1, 2, \ldots)$ estimates for θ at the successive iterations.
At iteration j we have

$$\begin{aligned} E_{\theta^{(j)}}[T(\underset{\sim}{X})|\underset{\sim}{Y}] &= E_{\theta^{(j)}}[-\sum_{i=1}^{n} X_i | X_1, \ldots, X_m] \\ &= -\sum_{i=1}^{m} X_i - (n-m)\frac{1}{\theta^{(j)}}. \end{aligned}$$

Now recall that the ML-estimate for θ based on the complete data set is $\dfrac{n}{\sum\limits_{i=1}^{n} x_i}$. Hence, at iteration $j+1$, the ML-estimate for θ is

$$\theta^{(j+1)} = \frac{n}{\sum\limits_{i=1}^{m} X_i + (n-m)\dfrac{1}{\theta^{(j)}}}.$$

Note. Setting $\theta^{(j)} = \theta^{(j+1)} = \widehat{\theta}$ we find that this iteration converges to

$$\widehat{\theta} = \frac{m}{\sum\limits_{i=1}^{m} X_i}$$

which is the ML estimator for θ based on $\underset{\sim}{Y} = \underset{\sim}{X}^{obs}$.
The EM algorithm is unnecessary in this example since the ML-estimator is obtained explicitly.

Example Let

$$\underset{\sim}{X} = (X_1, X_2, X_3, X_4) \sim M(n; \frac{1}{2} - \frac{\theta}{2}, \frac{\theta}{4}, \frac{\theta}{4}, \frac{1}{2})) \qquad (0 \leq \theta \leq 1)$$

Suppose we observe $\underset{\sim}{Y} = (X_1, X_2, X_3 + X_4)$; say $x_1 = 38, x_2 = 34, x_3 + x_4 = 125$.

The complete-data log likelihood function is

$$l(\theta; x) = \ln\left(\frac{(x_1 + x_2 + x_3 + x_4)!}{x_1!x_2!x_3!x_4!}\right) + x_1 \ln\left(\frac{1}{2} - \frac{\theta}{2}\right) + x_2 \ln\left(\frac{\theta}{4}\right) + x_3 \ln\left(\frac{\theta}{4}\right) + x_4 \ln\left(\frac{1}{2}\right).$$

Setting
$S(\theta; \underset{\sim}{x}) = -\frac{x_1}{1-\theta} + \frac{x_2}{\theta} + \frac{x_3}{\theta} = 0$ gives that the ML-estimate for θ for the complete data set is given by

$$\widehat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3}.$$

Note that the essential part of log likelihood function is linear in x_1, x_2, x_3, and x_4.
We have

$$E_\theta(X_3 \mid \underset{\sim}{Y}) = E_\theta(X_3 | X_1 = 38, X_2 = 34, X_3 + X_4 = 125) = 125 \frac{\frac{\theta}{4}}{\frac{1}{2} + \frac{\theta}{4}}$$

$$E_\theta(X_4 \mid \underset{\sim}{Y}) = E_\theta(X_4 | X_1 = 38, X_2 = 34, X_3 + X_4 = 125) = 125 \frac{\frac{1}{2}}{\frac{1}{2} + \frac{\theta}{4}}.$$

If $\theta^{(0)}, \theta^{(1)}, \ldots$ are the successive estimates for θ, then at iteration j we have

$$E_{\theta^{(j)}}(X_3 | X_1 = 38, X_2 = 34, X_3 + X_4 = 125) = \frac{125\left(\frac{\theta^{(j)}}{4}\right)}{\frac{1}{2} + \frac{\theta^{(j)}}{4}}$$

$$E_{\theta^{(j)}}(X_4 | X_1 = 38, X_2 = 34, X_3 + X_4 = 125) = \frac{125\left(\frac{1}{2}\right)}{\frac{1}{2} + \frac{\theta^{(j)}}{4}}.$$

At iteration $j + 1$:

$$\theta^{(j+1)} = \frac{34 + \frac{125\left(\frac{\theta^{(j)}}{4}\right)}{\frac{1}{2} + \frac{\theta^{(j)}}{4}}}{38 + 34 + \frac{125\left(\frac{\theta^{(j)}}{4}\right)}{\frac{1}{2} + \frac{\theta^{(j)}}{4}}}.$$

Using this formula iteratively we obtain the ML-estimate for θ, based on the observed likelihood.

For instance, starting with an initial estimate $\theta^{(0)} = \frac{1}{2}$, we obtain the following values:

$$\begin{aligned}
\theta^{(0)} &= 0.5 \\
\theta^{(1)} &= 0.608247423 \\
\theta^{(2)} &= 0.624321051 \\
\theta^{(3)} &= 0.626488879 \\
\theta^{(4)} &= 0.626777323 \\
\theta^{(5)} &= 0.626815632 \\
\theta^{(6)} &= 0.626820719 \\
\theta^{(7)} &= 0.626821395 \\
&\dots
\end{aligned}$$

Setting $\theta^{(j+1)} = \theta^{(j)} = \widehat{\theta}$ in the iteration formula, we see that $\widehat{\theta}$ satisfies

$$197\widehat{\theta}^2 - 15\widehat{\theta} - 68 = 0$$

The positive root of this quadratic equation is

$$\widehat{\theta} = \frac{15 + \sqrt{15^2 + 4(197)(68)}}{2(197)} = 0.6268215.$$

Note.
Inferences about θ can be based on the observed log likelihood $l(\theta; \underset{\sim}{x}^{obs})$ only if the missing-data mechanism leading to the incomplete data can be ignored. Such a missing-data mechanism can be described by introducing indicator variables $\underset{\sim}{R} = (R_1, \dots, R_n)$ where

$$R_i = \begin{cases} 1 & \text{if} \quad X_i \text{ is observed} \\ \\ 0 & \text{if} \quad X_i \text{ is missing.} \end{cases}$$

The more general model then specifies the joint distribution of $\underset{\sim}{X} = (X_1, \dots, X_n)$ and $\underset{\sim}{R} = (R_1, \dots, R_n)$

$$f_{\underset{\sim}{X},\underset{\sim}{R}}(\underset{\sim}{x}; \underset{\sim}{r}; \theta, \psi) = f_{\underset{\sim}{X}}(\underset{\sim}{x}; \theta) f_{\underset{\sim}{R}|\underset{\sim}{X}}(\underset{\sim}{r}|\underset{\sim}{x}; \psi).$$

The parameter ψ appears in the conditional distribution of $\underset{\sim}{R}$ given $\underset{\sim}{X}$. Sometimes this distribtution is known, and ψ is unnecessary.
The actual observed data are $(\underset{\sim}{X}^{obs}, \underset{\sim}{R})$. So the likelihood function to work with is given by (integrating out $\underset{\sim}{x}^{mis}$)

$$\begin{aligned}
f_{\underset{\sim}{X}^{obs},\underset{\sim}{R}}(\underset{\sim}{x}^{obs}, \underset{\sim}{r}; \theta, \psi) &= \int f_{\underset{\sim}{X}^{obs},\underset{\sim}{X}^{mis},\underset{\sim}{R}}(\underset{\sim}{x}^{obs}, \underset{\sim}{x}^{mis} \underset{\sim}{r}; \theta, \psi) d\underset{\sim}{x}^{mis} \\
&= \int f_{\underset{\sim}{X}^{obs},\underset{\sim}{X}^{mis}}(\underset{\sim}{x}^{obs}, \underset{\sim}{x}^{mis}; \theta) f_{\underset{\sim}{R}|\underset{\sim}{X}^{obs},\underset{\sim}{X}^{mis}}(\underset{\sim}{r}|\underset{\sim}{x}^{obs}, \underset{\sim}{x}^{mis}; \psi) d\underset{\sim}{x}^{mis}.
\end{aligned}$$

The question is now under what conditions we can use the simple $f_{\underset{\sim}{X}^{obs}}(\underset{\sim}{x}^{obs};\theta)$ instead of $f_{\underset{\sim}{X}^{obs},\underset{\sim}{R}}(\underset{\sim}{x}^{obs},\underset{\sim}{r};\theta,\psi)$. (hence ignoring the missing-data mechanism).
The missing data are said to be **missing at random** (MAR) if

$$f_{\underset{\sim}{R}|\underset{\sim}{X}^{obs},\underset{\sim}{X}^{mis}}(\underset{\sim}{r}|\underset{\sim}{x}^{obs},\underset{\sim}{x}^{mis};\psi) = f_{\underset{\sim}{R}|\underset{\sim}{X}^{obs}}(\underset{\sim}{r}|\underset{\sim}{x}^{obs};\psi)$$

i.e. the distribution of the missing-data mechanism does not depend on the missing values $\underset{\sim}{x}^{mis}$. In this case

$$\begin{aligned} f_{\underset{\sim}{X}^{obs},\underset{\sim}{R}}(\underset{\sim}{x}^{obs},r;\theta;\psi) &= f_{\underset{\sim}{R}|\underset{\sim}{X}^{obs}}(\underset{\sim}{r}|\underset{\sim}{x}^{obs};\psi)\int f_{\underset{\sim}{X}^{obs},\underset{\sim}{X}^{mis}}(\underset{\sim}{x}^{obs},\underset{\sim}{x}^{mis};\theta)d\underset{\sim}{x}^{mis} \\ &= f_{\underset{\sim}{R}|\underset{\sim}{X}^{obs}}(\underset{\sim}{r}|\underset{\sim}{x}^{obs};\psi)f_{\underset{\sim}{X}^{obs}}(\underset{\sim}{x}^{obs};\theta) \end{aligned}$$

In many cases the parameters θ and ψ are distinct and hence the likelihood $f_{\underset{\sim}{X}^{obs},\underset{\sim}{R}}(\underset{\sim}{x}^{obs},r;\theta,\psi)$ and $f_{\underset{\sim}{X}^{obs}}(\underset{\sim}{x}^{obs};\theta)$ are proportional.

Large Sample properties of ML-Estimators

We now consider the limiting behaviour as the number n of observations tends to infinity.

(a) Weak consistency and asymptotic normality Under regularity conditions on the family of densities $\{f(x;\underset{\sim}{\theta})|\underset{\sim}{\theta} \in \Theta\}$, the ML-estimators are weakly consistent and asymptotically normal. We do not state these regularity conditions, but, only give a short idea of the proofs.

One-parameter case

Theorem

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta), \theta \in \Theta \subset \mathbb{R}$. Under regularity conditions, the ML-equation provides a ML-estimator T_n which satisfies, as $n \to \infty$:

(i) $T_n \xrightarrow{P} \theta$

(ii) $n^{\frac{1}{2}}(T_n - \theta) \xrightarrow{d} N(0; \frac{1}{i(\theta)})$
where $i(\theta)$ is the Fisher information number.
(i.e. for large n : T_n is approximately $N\left(\theta; \frac{1}{ni(\theta)}\right)$).

'Proof'

$T_n = t(X_1, \ldots, X_n)$ and $\hat{\theta}_n = t(x_1, \ldots, x_n)$ satisfies $S(\hat{\theta}_n; \underset{\sim}{x}) = 0$.
Taylor expansion around θ gives :

$$0 = S(\hat{\theta}_n; \underset{\sim}{x}) \approx S(\theta; \underset{\sim}{x}) + (\hat{\theta}_n - \theta)(-\mathcal{I}(\theta; \underset{\sim}{x}))$$

or:

$$T_n - \theta \approx \frac{S(\theta; \underset{\sim}{X})}{\mathcal{I}(\theta; \underset{\sim}{X})}$$

To prove (i), we write

$$T_n - \theta \approx \frac{\frac{1}{n} S(\theta; \underset{\sim}{X})}{\frac{1}{n}\mathcal{I}(\theta; \underset{\sim}{X})} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} lnf(X_i;\theta)}{\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2} lnf(X_i;\theta)\right)}$$

Application of the weak law of large numbers for sums of i.i.d. random variables, gives, as $n \to \infty$:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}lnf(X_i;\theta) \xrightarrow{P} E\left[\frac{\partial}{\partial\theta}lnf(X;\theta)\right] &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}lnf(X_i;\theta)\right) \xrightarrow{P} E\left[-\frac{\partial^2}{\partial\theta^2}lnf(X;\theta)\right] &= i(\theta) \end{aligned}$$

and hence, as $n \to \infty : T_n - \theta \xrightarrow{P} 0$.

To prove (ii), we write

$$n^{\frac{1}{2}}(T_n - \theta) \approx \frac{\frac{1}{\sqrt{n}}S(\theta; \underset{\sim}{X})}{\frac{1}{n}\mathcal{I}(\theta; \underset{\sim}{X})} = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}lnf(X_i;\theta)}{\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}lnf(X_i;\theta)\right)}$$

As before, we have for the denominator, by the law of large numbers :

$$\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{\partial^2}{\partial\theta^2}lnf(X_i;\theta)\right) \xrightarrow{P} i(\theta).$$

The numerator is a properly normalized sum of i.i.d. random variables with mean

$$E\left[\frac{\partial}{\partial\theta}lnf(X;\theta)\right] = 0$$

and variance

$$E\left[\left(\frac{\partial}{\partial\theta}lnf(X;\theta)\right)^2\right] = i(\theta)$$

Hence, by the central limit theorem for sums of i.i.d. random variables :

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}lnf(X_i;\theta) \xrightarrow{d} N(0; i(\theta))$$

Hence, by Slutsky's theorem :

$$n^{\frac{1}{2}}(T_n - \theta) \xrightarrow{d} N\left(0; \frac{1}{i(\theta)}\right). \qquad \square$$

Multiparameter case

The previous theorem generalizes to the multi-parameter case (in which $\underset{\sim}{\theta}$ is a vector). The basic tools in the previous proof (Taylor's theorem, central limit theorem, law of large numbers) all have their multivariate version.

Theorem

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x; \underset{\sim}{\theta}), \underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$. Under regularity conditions, the ML-equations provide a ML-estimator $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ which satisfies, as $n \to \infty$:

(i) $\underset{\sim}{T}_n \xrightarrow{P} \underset{\sim}{\theta}$

(ii) $n^{\frac{1}{2}}(T_{n1} - \theta_1, \ldots, T_{nk} - \theta_k) \xrightarrow{d} N_k(\underset{\sim}{0}; B^{-1}(\underset{\sim}{\theta}))$ *where $B(\underset{\sim}{\theta})$ is the Fisher information matrix.*
(i.e. for large n : $\underset{\sim}{T}_n$ is approximately $N_k(\underset{\sim}{\theta}; \frac{1}{n} B^{-1}(\underset{\sim}{\theta}))$).

(b) Asymptotic efficiency of ML-estimators Under regularity conditions we obtained that the limiting distribution of the ML-estimator is normal around $\underset{\sim}{\theta}$ and with variance $1/ni(\theta)$ (in the 1-parameter case) or $\frac{1}{n} B^{-1}(\underset{\sim}{\theta})$ (in the multi-parameter case). These expressions are the lower bounds in the **Cramer-Rao inequality** (or information inequality). The fact that the variance of the limiting normal distribution of the ML-estimator achieves this lower bound is usually expressed as : the ML-estimator is **asymptotically efficient** (or **B.A.N.** : best asymptotically normal). To see this connection, we recall the Cramer-Rao inequality :

One-parameter case

Theorem [Cramer-Rao inequality]
Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta), \theta \in \Theta \subset \mathbb{R}$.
Let T_n be any unbiased estimator for θ.
Then, under regularity conditions,

$$Var_\theta(T_n) \geq \frac{1}{ni(\theta)} \qquad , \text{ for all } \theta \in \Theta$$

where $i(\theta)$ is the Fisher information number.

'Proof'

We sketch the proof for the case where X is continuous. (In the discrete case : replace integrals by sums)
Since the estimator $T_n = t(X_1, \ldots, X_n)$ is unbiased for θ, we have

$$\theta = \int \ldots \int t(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i;\theta) dx_1 \ldots dx_n$$

or

$$\theta = \int \ldots \int t(\underset{\sim}{x}) L(\theta; \underset{\sim}{x}) dx_1 \ldots dx_n$$

Differentiation with respect to θ gives (under regularity conditions) :

$$1 = \int \ldots \int t(\underset{\sim}{x}) \frac{\partial}{\partial\theta} L(\theta; \underset{\sim}{x}) dx_1 \ldots dx_n$$

or

$$1 = \int \ldots \int t(\underset{\sim}{x}) S(\theta; \underset{\sim}{x}) L(\theta; \underset{\sim}{x}) dx_1 \ldots dx_n$$

or

$$1 = E_\theta[T_n.S(\theta; \underset{\sim}{X})]$$

or

$$1 = Cov_\theta[T_n, S(\theta; \underset{\sim}{X})]$$

(since, under regularity conditions, we have $E[S(\theta; \underset{\sim}{X})] = 0$ – see before –)

By the Cauchy-Schwarz inequality

$$\begin{aligned} 1 &\leq Var_\theta(T_n).Var_\theta(S(\theta; \underset{\sim}{X})) \\ &= Var_\theta(T_n).(ni(\theta)) \end{aligned}$$

(under regularity conditions – see before –)

Hence : $Var_\theta(T_n) \geq \dfrac{1}{ni(\theta)}$. □

Multi-parameter case

The Cramer-Rao inequality for the variance generalizes to the multi-parameter case. The next theorem states that, in a certain sense, the matrix $\dfrac{1}{n}B^{-1}(\underset{\sim}{\theta})$ is a "lower bound" for the variance-covariance matrix of an unbiased estimator for $\underset{\sim}{\theta}$.

Theorem [information inequality]

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x; \underset{\sim}{\theta}), \underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset \mathbb{R}^k$.
Let $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ be any unbiased estimator for $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$. Denote by $Var_{\underset{\sim}{\theta}}(\underset{\sim}{T}_n)$ the variance-covariance matrix of $\underset{\sim}{T}_n$.
Then, under regularity conditions,

$$Var_{\underset{\sim}{\theta}}(\underset{\sim}{T}_n) - \frac{1}{n}B^{-1}(\underset{\sim}{\theta}) \text{ is positive semidefinite for all } \underset{\sim}{\theta} \in \Theta,$$

where $B(\underset{\sim}{\theta})$ is the Fisher information matrix.

'Proof'

Since $\underset{\sim}{T}_n = (T_{n1}, \dots T_{nk})$ is unbiased for $\underset{\sim}{\theta}$, we have that for each $i = 1, \dots, k$:
$T_{ni} = t_i(X_1, \dots, X_n)$ is unbiased for θ_i, or

$$\int \dots \int t_i(\underset{\sim}{x}) L(\underset{\sim}{\theta}; \underset{\sim}{x}) dx_1 \dots dx_n = \theta_i$$

By differentiation : for all $i, j = 1, \dots, k$:

$$\int \dots \int t_i(\underset{\sim}{x}) \frac{\partial}{\partial \theta_j} L(\underset{\sim}{\theta}; \underset{\sim}{x}) dx_1 \dots dx_n = \delta_{ij} = \begin{cases} 0 \dots \text{ if } i \neq j \\ 1 \dots \text{ if } i = j \end{cases}$$

or

$$\int \dots \int t_i(\underset{\sim}{x}) S_j(\underset{\sim}{\theta}; \underset{\sim}{x}) L(\underset{\sim}{\theta}; \underset{\sim}{x}) dx_1 \dots, dx_n = \delta_{ij}$$

where $S_j(\underset{\sim}{\theta}; \underset{\sim}{x})$ is the j-th component of the score vector. Hence : $Cov(T_{ni}, S_j(\underset{\sim}{\theta}; \underset{\sim}{X})) = \delta_{ij}$.
Consider the variance-covariance matrix of the vector $(T_{n1}, \dots T_{nk}, S_1(\underset{\sim}{\theta}; \underset{\sim}{X}), \dots, S_k(\underset{\sim}{\theta}, \underset{\sim}{X}))$;
it can be written as the following partitioned matrix :

$$\begin{pmatrix} Var_{\underset{\sim}{\theta}}(\underset{\sim}{T}_n) & I \\ \\ \\ I & nB(\underset{\sim}{\theta}) \end{pmatrix}$$

Because this is a variance-covariance matrix, it is positive semidefinite.
It follows that

$$\begin{bmatrix} I & -\frac{1}{n} B^{-1}(\underset{\sim}{\theta}) \end{bmatrix} \begin{bmatrix} Var_{\underset{\sim}{\theta}}(\underset{\sim}{T}_n) & I \\ \\ \\ I & nB(\underset{\sim}{\theta}) \end{bmatrix} \begin{bmatrix} I \\ \\ \\ -\frac{1}{n} B^{-1}(\underset{\sim}{\theta}) \end{bmatrix}$$

is also positive semidefinite.
But this last matrix is just $Var_{\underset{\sim}{\theta}}(\underset{\sim}{T}_n) - \frac{1}{n}B^{-1}(\underset{\sim}{\theta})$. □

Example
Let $X_1, \ldots, X_n$ be a random sample from $X \sim$ Poisson (θ), where $\theta > 0$.

$f(x;\theta) = e^{-\theta}\frac{\theta^x}{x!}$
$lnf(x;\theta) = -lnx! - \theta + xln\theta$
$\frac{\partial}{\partial\theta}lnf(x;\theta) = \frac{x-\theta}{\theta}$
$i(\theta) = \frac{1}{\theta^2}E[(X-\theta)^2] = \frac{1}{\theta^2}Var(X) = \frac{\theta}{\theta^2} = \frac{1}{\theta}$.
The Cramer-Rao bound is $\frac{\theta}{n}$.
Note that $\overline{X}$ is an unbiased estimator which reaches the Cramer-Rao bound : $E(\overline{X}) = \theta$ and $Var(\overline{X}) = \frac{\theta}{n}$.

Example

Let $X_1, \ldots, X_n$ be a random sample form $X \sim$ Bernoulli with parameter $\theta, \theta < \theta < 1$.
Calculation shows that

$$i(\theta) = \frac{1}{\theta(1-\theta)}$$

and hence the Cramer-Rao bound is

$$\frac{\theta(1-\theta)}{n}$$

Check that $\overline{X}$ is an unbiased estimator for θ which reaches the Cramer-Rao bound: $Var(\overline{X}) = \frac{\theta(1-\theta)}{n}$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu;\sigma^2)$ with σ^2 known.
Put $\theta = \mu$.
Check that $i(\theta) = \frac{1}{\sigma^2}$.

Hence the Cramer-Rao bound is $\frac{\sigma^2}{n}$.
This bound is attained by the unbiased estimator $\overline{X}$.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$ with μ known.
Put $\theta = \sigma^2$.
Check that $i(\theta) = \frac{1}{2\theta^2}$.
Hence the Cramer-Rao bound is $\frac{2\theta^2}{n}$.
Check that the unbiased estimator $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ attains this bound.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$ with μ and σ^2 unknown.
Put $\underset{\sim}{\theta} = (\theta_1, \theta_2)$ with $\theta_1 = \mu, \theta_2 = \sigma^2$.

We have

$$B(\underset{\sim}{\theta}) = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix}$$

which has inverse :

$$B^{-1}(\underset{\sim}{\theta}) = \begin{pmatrix} \theta_2 & 0 \\ 0 & 2\theta_2^2 \end{pmatrix}.$$

The "lower bound matrix" for the variance-covariance matrix of any unbiased estimator (T_{n1}, T_{n2}) for (θ_1, θ_2) is

$$\begin{pmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{pmatrix}$$

so that,

if $E(T_{n1}) = \theta_1$, then $Var(T_{n1}) \geq \frac{\theta_2}{n}$

if $E(T_{n2}) = \theta_2$, then $Var(T_{n2}) \geq \frac{2\theta_2^2}{n}$.

Example

Let $\underset{\sim}{X} = (X_1, X_2, X_3) \sim M(n; (\theta_1, \theta_2, \theta_3))$ (**Trinomial distribution**)
Since $\theta_3 = 1 - \theta_2 - \theta_3$, we have in fact a 2-dimensional parameter $\underset{\sim}{\theta} = (\theta_1, \theta_2)$.
With $\underset{\sim}{x} = (x_1, x_2, x_3)$, we have
$f(\underset{\sim}{x}; \underset{\sim}{\theta}) = \frac{n!}{x_1!x_2!x_3!}\theta_1^{x_1}\theta_2^{x_2}(1-\theta_1-\theta_2)^{x_3}$

$$lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) = ln\left(\frac{n!}{x_1!x_2!x_3!}\right) + x_1 ln\theta_1 + x_2 ln\theta_2 + x_3 ln(1-\theta_1-\theta_2)$$

$$\begin{aligned}
\frac{\partial}{\partial\theta_1} lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) &= \frac{x_1}{\theta_1} - \frac{x_3}{1-\theta_1-\theta_2} \\
\frac{\partial}{\partial\theta_2} lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) &= \frac{x_2}{\theta_2} - \frac{x_3}{1-\theta_1-\theta_2} \\
\frac{\partial^2}{\partial\theta_1^2} lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) &= -\frac{x_1}{\theta_1^2} - \frac{x_3}{(1-\theta_1-\theta_2)^2} \\
\frac{\partial^2}{\partial\theta_2^2} lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) &= -\frac{x_2}{\theta_2^2} - \frac{x_3}{(1-\theta_1-\theta_2)^2} \\
\frac{\partial^2}{\partial\theta_1\partial\theta_2} lnf(\underset{\sim}{x}; \underset{\sim}{\theta}) &= -\frac{x_3}{(1-\theta_1-\theta_2)^2}
\end{aligned}$$

Now $E(X_1) = n\theta_1$, $E(X_2) = n\theta_2$, $E(X_3) = n(1-\theta_1-\theta_2)$
This leads to :

$$B(\underset{\sim}{\theta}) = \begin{pmatrix} \frac{n}{\theta_1} + \frac{n}{1-\theta_1-\theta_2} & \frac{n}{1-\theta_1-\theta_2} \\ \frac{n}{1-\theta_1-\theta_2} & \frac{n}{\theta_2} + \frac{n}{1-\theta_1-\theta_2} \end{pmatrix}$$

and hence

$$B^{-1}(\underset{\sim}{\theta}) = \begin{pmatrix} \dfrac{\theta_1(1-\theta_1)}{n} & \dfrac{-\theta_1\theta_2}{n} \\ \\ \dfrac{-\theta_1\theta_2}{n} & \dfrac{\theta_2(1-\theta_2)}{n} \end{pmatrix}.$$

The "lower bound matrix" for the variance-covariance matrix of any unbiased estimators for θ_1 and θ_2 is

$$\begin{pmatrix} \dfrac{\theta_1(1-\theta_1)}{n} & \dfrac{-\theta_1\theta_2}{n} \\ \\ \dfrac{-\theta_1\theta_2}{n} & \dfrac{\theta_2(1-\theta_2)}{n} \end{pmatrix}$$

One may verify that this lower bound is attained by the variance-covariance matrix of the obvious unbiased estimators for θ_1 and θ_2 : $\dfrac{X_1}{n}$ and $\dfrac{X_2}{n}$.

Example

For the general multinomial distribution $M(n; (\theta_1, \ldots, \theta_k))$, we have, with $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_{k-1})$:
Fisher information matrix

$$B(\underset{\sim}{\theta}) = n\left(\frac{\delta_{ij}}{\theta_i} + \frac{1}{1-\theta_1-\theta_2-\ldots-\theta_{k-1}}\right)_{i,j=1,\ldots,k-1}$$

and for its inverse :

$$B^{-1}(\underset{\sim}{\theta}) = \frac{1}{n}\left(\delta_{ij}\theta_i - \theta_i\theta_j\right)_{i,j=1,\ldots,k-1}.$$

2.4.2 Minimax and Bayes Estimation

Decision function. Loss function. Risk function

Suppose we have a random sample from X with density $f(x;\theta)$ where $\theta \in \Theta \subset \mathbb{R}$ is an unknown parameter.

It is convenient to borrow some language from **decision theory**.

1. An estimate for θ, i.e. a function $\delta(x_1,\ldots,x_n)$ of the observations is often called a **decision**.
   The function : $\delta : \mathbb{R}^n \to \mathbb{R}$ is called a **decision function**.

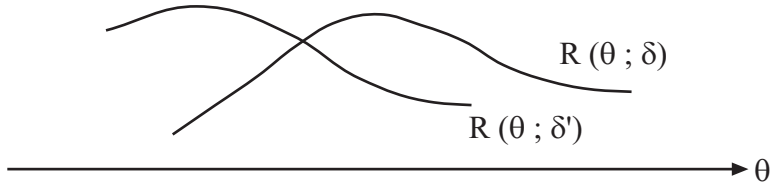
2. If θ is estimated by $\delta(x_1,\ldots,x_n)$, then the error is called the **loss** and a measure for the error is called a **loss function** , i.e. a nonnegative function of θ and $\delta(x_1,\ldots,x_n)$:
   $$l(\theta;\delta(x_1,\ldots,x_n))$$
   Examples
   - $l(\theta;\delta(x_1,\ldots,x_n)) = |\theta - \delta(x_1,\ldots,x_n)|$: **Absolute error loss**
   - $l(\theta;\delta(x_1,\ldots,x_n)) = (\theta - \delta(x_1,\ldots,x_n))^2$: **Squared error loss**

3. Suppose a certain loss function has been chosen. We want to choose an estimate for θ, i.e. a decision function $\delta(x_1,\ldots,x_n)$ such that the **average** loss is small. The average loss is called the **risk function**. It is a function R of θ and $\delta(x_1,\ldots,x_n)$

$$\begin{aligned} & R(\theta;\delta) \\ = \; & E_\theta[l(\theta;\delta(X_1,\ldots,X_n))] \\ = \; & \begin{cases} \sum_{x_1}\ldots\sum_{x_n} l(\theta;\delta(x_1,\ldots,x_n)) \prod_{i=1}^{n} f(x_i;\theta) & \text{(discrete case)} \\ \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} l(\theta;\delta(x_1,\ldots,x_n)) \prod_{i=1}^{n} f(x_i;\theta)dx_1\ldots dx_n & \text{(continuous case)} \end{cases} \end{aligned}$$

   Note
   For the squared error loss function, the corresponding risk function is the mean-squared error **(MSE)**.

4. In trying to compare 2 estimators, say $\delta(X_1,\ldots,X_n)$ and $\delta'(X_1,\ldots,X_n)$ via their risk functions $R(\theta;\delta)$ and $R(\theta;\delta')$, we usually find that comparison is impossible since the risk functions, as functions of θ, cross :



Minimax estimators

A first way to overcome the difficulty of finding an estimator which minimizes the risk function **uniformly** in θ, is to look for an estimator which minimizes the worst that could occur, i.e. which minimizes the maximum (over θ) risk.

So two estimators $\delta(X_1, \ldots, X_n)$ and $\delta'(X_1, \ldots, X_n)$ for θ can be compared by comparing two **numbers**:

$$\sup_{\theta \in \Theta} R(\theta; \delta) \quad \text{and} \quad \sup_{\theta \in \Theta} R(\theta; \delta').$$

Definition

A **minimax estimator for** θ is defined to be the estimator $\delta(X_1, \ldots, X_n)$ for which

$$\sup_{\theta \in \Theta} R(\theta; \delta) \leq \sup_{\theta \in \Theta} R(\theta; \delta')$$

for any other estimator $\delta'(X_1, \ldots, X_n)$ for θ.

A method for finding a minimax estimator will be given at the end of the next section.

Bayes estimators

A second way to overcome the difficulty of comparing risk functions, is to average out $R(\theta; \delta)$ over θ. This leads to the so called Bayes estimators.

Essential in the Bayesian approach is to view the parameter θ as a value of some random variable $\widetilde{\Theta}$ with a known distribution (rather than viewing θ as an unknown constant). This completely specified (discrete or continuous) density over the parameter space Θ is

called the **prior density**. The choice of the prior density reflects past experience about the parameter θ. It expresses the degree of belief in different values of the parameter, **before** the observations are made.

The averaging out of $R(\theta;\delta)$ over the parameter space Θ will be done using the density of $\widetilde{\Theta}$ (= the prior density) as weight function.
Let $\pi(\theta)(\theta \in \Theta)$ be the prior density.

The **Bayes risk** $R(\delta)$ is defined by

$$R(\delta) = \begin{cases} \sum\limits_{\Theta} R(\theta;\delta)\pi(\theta) & \dots \text{ if } \pi \text{ is discrete} \\ \int\limits_{\Theta} R(\theta;\delta)\pi(\theta)d\theta & \dots \text{ if } \pi \text{ is continuous} \end{cases}$$

This gives us a way to compare two estimators $\delta(X_1,\dots,X_n)$ and $\delta'(X_1,\dots,X_n)$: we will compare two **numbers**, nl. their Bayes risks

$$R(\delta) \quad \text{and} \quad R(\delta').$$

Definition

The **Bayes estimator for** θ with respect to the loss function l and the prior density π is defined to be the estimator $\delta(X_1,\dots,X_n)$ for which

$$R(\delta) \leq R(\delta')$$

for any other estimator $\delta'(X_1,\dots,X_n)$ for θ.

For squared error loss, finding the Bayes estimator is easy :

Theorem

The Bayes estimator for θ with respect to the squared eroor loss function and the prior density π is given by $\delta(X_1,\dots,X_n)$, where

$$\delta(x_1,\ldots,x_n) = \begin{cases} \dfrac{\sum\limits_{\Theta} \theta \left[\prod\limits_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)}{\sum\limits_{\Theta} \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)} & \ldots \text{ if } \pi \text{ is discrete} \\ \dfrac{\int\limits_{\Theta} \theta \left[\prod\limits_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta) d\theta}{\int\limits_{\Theta} \left[\prod\limits_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta) d\theta} & \ldots \text{ if } \pi \text{ is continuous} \end{cases}$$

Proof

We give the proof for the case that the prior density π is of the continuous type. For the case of a discrete type prior density : replace all integrals by sums.
We have

$$\begin{aligned} R(\delta) &= \int\limits_{\Theta} R(\theta;\delta)\pi(\theta)d\theta \\ &= \int\limits_{\Theta} \left[\int \ldots \int (\theta - \delta(x_1,\ldots,x_n))^2 \left[\prod_{i=1}^{n} f(x_i;\theta)\right] dx_1 \ldots dx_n\right] \pi(\theta)d\theta \\ &= \int \ldots \int \left[\int\limits_{\Theta} (\theta - \delta(x_1,\ldots,x_n))^2 \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)d\theta\right] dx_1 \ldots dx_n. \end{aligned}$$

$R(\delta)$ will be minimized if the integrand

$$I \equiv \int\limits_{\Theta} (\theta - \delta(x_1,\ldots,x_n))^2 \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)d\theta$$

is minimized for all $(x_1,\ldots,x_n)$. But :

$$\begin{aligned} I &= \int\limits_{\Theta} \theta^2 \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)d\theta \\ &\quad -2\delta(x_1,\ldots,x_n) \int\limits_{\Theta} \theta \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)d\theta \\ &\quad +\delta^2(x_1,\ldots,x_n) \int\limits_{\Theta} \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta)d\theta \end{aligned}$$

and this is minimized for

$$\delta(x_1,\ldots,x_n)=\frac{\int\limits_{\Theta}\theta\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)d\theta}{\int\limits_{\Theta}\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)d\theta}.$$

□

Interpretation

The quantity $\dfrac{\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)}{\sum\limits_{\Theta}\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)}$ (in the discrete case) or $\dfrac{\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)}{\int\limits_{\Theta}\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)d\theta}$ (in the continuous case) is a density as a function of θ, with $x_1,\ldots,x_n$ fixed. It is called the **posterior density of** $\widetilde{\Theta}$. It is the conditional density of $\widetilde{\Theta}$, given that $X_1=x_1,\ldots,X_n=x_n$:

$$f_{\widetilde{\theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n).$$

Indeed, (in the discrete case) :

$$\begin{aligned}
&f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n)\\
&=P[\widetilde{\Theta}=\theta|X_1=x_1,\ldots,X_n=x_n]\\
&=\frac{P[X_1=x_1,\ldots,X_n=x_n|\widetilde{\Theta}=\theta]P[\widetilde{\Theta}=\theta]}{\sum\limits_{\Theta}P(X_1=x_1,\ldots X_n=x_n|\widetilde{\Theta}=\theta]P[\widetilde{\Theta}=\theta]}\\
&\qquad\qquad\qquad\qquad\text{(by }\textbf{Bayes rule}\text{)}\\
&=\frac{\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)}{\sum\limits_{\Theta}\left[\prod\limits_{i=1}^{n}f(x_i;\theta)\right]\pi(\theta)}.
\end{aligned}$$

Thus, we have :

The Bayes estimator for θ with respect to the squared error loss function and the prior π is given by $\delta(X_1, \ldots, X_n)$, where

$$\begin{aligned}\delta(x_1, \ldots, x_n) &= \begin{cases} \sum\limits_{\Theta} \theta f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1, \ldots, x_n) \\ \int\limits_{\Theta} \theta f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1, \ldots, x_n)d\theta \end{cases} \\ &= E[\widetilde{\Theta}|X_1 = x_1, \ldots, X_n = x_n] \\ &= \text{the } \textbf{mean of the posterior distribution.}\end{aligned}$$

This corresponds to the well known fact that, for a random variable X with $E(X^2) < \infty$: $E[(X - a)^2]$ is minimum when $a = E(X)$.

Note
The same interpretation holds for Bayes estimators with respect to other loss functions. E.g. for absolute error loss, the Bayes estimator will be given by **a median of the posterior distribution**. This corresponds to the fact that, for a random variable X with $E|X| < \infty : E|X - a|$ is minimum when a is any median of X.

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1; \theta)$ with $0 < \theta < 1$.
Prior density : $\widetilde{\Theta} \sim Beta(\alpha; \beta)$, i.e.

$$\pi(\theta) = \begin{cases} \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} & \ldots \text{ if } 0 < \theta < 1 \\ 0 & \ldots \text{ if otherwise} \end{cases}$$

- $$\begin{aligned}(1) &= \int_0^1 \theta \left[\prod_{i=1}^{n} f(x_i; \theta)\right] \pi(\theta)d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta.\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+\sum x_i+1)\Gamma(\beta+n-\sum x_i)}{\Gamma(\alpha+\beta+n+1)}\end{aligned}$$

- $$\begin{aligned}(2) &= \int_0^1 \left[\prod_{i=1}^n f(x_i;\theta)\right]\pi(\theta)d\theta \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+\sum x_i)\Gamma(\beta+n-\sum x_i)}{\Gamma(\alpha+\beta+n)}\end{aligned}$$

- $\delta(x_1,\ldots,x_n) = \dfrac{(1)}{(2)} = \dfrac{\sum x_i + \alpha}{n+\alpha+\beta}$

Bayes estimator for θ : $\dfrac{\sum_{i=1}^n X_i + \alpha}{n+\alpha+\beta}$

Special case : If $\alpha = \beta = 1$, then the $Beta(\alpha;\beta)$ becomes the $Un[0,1]$-prior density and the Bayes estimator is

$$\frac{\sum_{i=1}^n X_i + 1}{n+2}.$$

Example

Let $X_1,\ldots,X_n$ be a random sample from $X \sim N(\theta;\sigma^2)$ with σ^2 known; $\theta \in I\!R$.
Prior density : $\widetilde{\Theta} \sim N(\mu_0;\sigma_0^2)$, with μ_0 and σ_0^2 known.

- $$\begin{aligned}(1) &= \int_{-\infty}^{\infty} \theta \left[\prod_{i=1}^n f(x_i;\theta)\right]\pi(\theta)d\theta \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \frac{1}{\sigma_0\sqrt{2\pi}} \int_{-\infty}^{\infty} \theta e^{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2 - \frac{1}{2\sigma_0^2}(\theta-\mu_0)^2} d\theta\end{aligned}$$

  Now algebraic manipulations give

$$\begin{aligned}&\frac{\sum(x_i-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\sigma_0^2} \\ &= C + \frac{\sigma^2+n\sigma_0^2}{\sigma^2\sigma_0^2}\left(\theta - \frac{\sigma^2\mu_0 + \sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}\right)^2\end{aligned}$$

  where C does not depend on θ.
  Hence :

$$(1) = C'.\int_{-\infty}^{\infty} \theta e^{-\frac{1}{2}\frac{\sigma^2+n\sigma_0^2}{\sigma^2\sigma_0^2}\left(\theta - \frac{\sigma^2\mu_0+\sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}\right)^2} d\theta$$

  where C' does not depend on θ.
  And similarly, with the same C' :

- $$\begin{aligned}(2) &= \int_{-\infty}^{\infty} \left[\prod_{i=1}^{n} f(x_i;\theta)\right] \pi(\theta) d\theta \\ &= C' \int_{-\infty}^{\infty} e^{-\frac{1}{2}\frac{\sigma^2+n\sigma_0^2}{\sigma^2\sigma_0^2}\left(\theta - \frac{\sigma^2\mu_0+\sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}\right)^2} d\theta\end{aligned}$$

  Hence :

  $$\delta(x_1,\ldots,x_n) = \frac{(1)}{(2)} = \frac{\sigma^2\mu_0+\sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}$$

  Bayes estimator for θ :

  $$\frac{\sigma^2\mu_0+\sigma_0^2 n\overline{X}}{\sigma^2+\sigma_0^2 n}.$$

Notes :

- The Bayes estimator is a weighted average of the ML-estimator $\overline{X}$ and the prior mean μ_0
- The Bayes estimator gets closer to the ML-estimator as $n \to \infty$.

We finally mention that Bayes estimation is sometimes used to find a **minimax** estimator.

Theorem

If $T_n = \delta(X_1,\ldots,X_n)$ is a Bayes estimator having constant risk (i.e. $R(\theta;\delta)$ is independent of θ), then T_n is a minimax estimator.

Proof (continuous case) :

Since $T_n = \delta(X_1,\ldots,X_n)$ is Bayes :

$$\int_{\Theta} R(\theta;\delta)\pi(\theta)d\theta \leq \int_{\Theta} R(\theta;\delta')\pi(\theta)d\theta$$

for any other estimator $\delta'(X_1,\ldots,X_n)$.
But if $R(\theta;\delta)$ does not depend on θ, we can say

$$R(\theta;\delta) \leq \sup_{\theta\in\Theta} R(\theta;\delta')$$

and

$$\sup_{\theta\in\Theta} R(\theta;\delta) \leq \sup_{\theta\in\Theta} R(\theta;\delta').$$

□

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1;\theta)$ with $0 < \theta < 1$.
For a $Beta(\alpha;\beta)$ prior, we found that the Bayes estimator was given by :

$$\delta(X_1, \ldots, X_n) = \frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta}.$$

The risk :

$$R(\theta;\delta) = E_\theta \left[\left(\frac{\sum_{i=1}^{n} X_i + \alpha}{n + \alpha + \beta} - \theta \right)^2 \right]$$

Since $\sum_{i=1}^{n} X_i \sim B(n;\theta)$, we have

$$E_\theta(\sum_{i=1}^{n} X_i) = n\theta \quad \text{and} \quad E_\theta \left[\left(\sum_{i=1}^{n} X_i \right)^2 \right] = n\theta(1 - \theta + n\theta).$$

Hence,

$$R(\theta,\delta) = \frac{[(\alpha+\beta)^2 - n]\theta^2 - [2\alpha^2 + 2\alpha\beta - n]\theta + \alpha^2}{(n+\alpha+\beta)^2}.$$

If $\alpha = \beta = \frac{\sqrt{n}}{2}$, then $(\alpha+\beta)^2 - n = 0, 2\alpha^2 + 2\alpha\beta - n = 0$ and the risk becomes independent of θ. Hence, the **minimax estimator for** θ is

$$\frac{\sum_{i=1}^{n} X_i + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}.$$

2.5 Other estimation methods

2.5.1 The method of moments

Many parameters of unknown densities are nice functions of one or more moments of the population random variable X, e.g.

$$\theta = \varphi(\mu_1, \ldots, \mu_k) \qquad , k \geq 1$$

where $\mu_r = E(X^r)$.

This suggests the following estimator for θ :

$$\varphi\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i^k \right)$$

where, for $r = 1, \ldots, k$, we replaced the population mean μ_r by the corresponding r-th **sample moment**

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

Example

If $\theta = E(X)$, then the estimator is $\frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$.

Example

If $\theta = Var(X)$, then $\theta = \mu_2 - \mu_1^2$ and the estimator is $\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = S^2$.

Example

If $X_1, \ldots, X_n$ is a random sample from $X \sim Un[\theta_1 - \theta_2, \theta_1 + \theta_2]$ then $E(X) = \theta_1, E(X^2) = \theta_1^2 + \frac{1}{3}\theta_2^2$. Hence, $\theta_1 = E(X), \theta_2 = \sqrt{3[E(X^2) - (E(X))^2]}$ and the moment estimators for θ_1 and θ_2 are :

$$\overline{X} \text{ and } \sqrt{3}S.$$

2.5.2 The method of least squares

The method of least squares is generally used in the estimation of parameters in a **linear model**. In the simplest case, we have n observations $y_1, \ldots, y_n$ made at different (known) values $x_1, \ldots, x_n$. The model is that $y_1, \ldots, y_n$ are values of random variables $Y_1, \ldots, Y_n$ which are independent and such that

$$E(Y_i) = \alpha + \beta x_i \qquad , i = 1, \ldots, n$$

where α and β are unknown parameters. The least squares principle takes estimates a and b (for α and β respectively) in such a way that the sum of the squares of the errors

$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

is minimized. This simple model and more general models are discussed in the courses on **Regression** and **Anova**.

2.6 Cramer-Rao Lower Bound and Uniformly Minimum Variance Unbiased estimation

2.6.1 Univariate case

The theorem and the proof of the Cramer lower bound is already given in Section (2.4). In this section we will discuss the role of Cramer Rao lower bound in constructing a UMVU estimator. The Cramer Rao Lower Bound (CRLB) gives the minimum variance that can be expected from an unbiased estimator. If $E[t(\underset{\sim}{X})] = \tau(\theta)$, where $\tau(\theta)$ is a function of θ then under regularity conditions,

$$var(t(\underset{\sim}{X})) \geq \frac{[\tau'(\theta)]^2}{nE\left(\frac{-\partial^2 ln f_{\underset{\sim}{X}}(\underset{\sim}{x};\theta)}{\partial\theta^2}\right)}$$

Equality holds iff there exists $k(n,\theta)$ such that

$$\sum \frac{\partial ln f_X(x_i;\theta)}{\partial\theta} = k(n,\theta)[t(\underset{\sim}{x}) - \theta] \tag{2.1}$$

$t(\underset{\sim}{X})$ is then UMVU estimator.

Definition
In a regular case of point estimation, the ratio of the CRLB to the actual variance of any unbiased estimator for a parameter is called the efficiency of the estimator.

The Cramer-Rao inequality has two uses:

(i) It gives a lower bound for the variance of unbiased estimators.

(ii) If an unbiased estimator whose variance coincides with the Cramer-Rao lower bound (CRLB) can be found, then this estimator is a UMVU estimator.

Example

Let $X_1 \ldots X_n$ be a random sample from $Exp(\theta)$; i.e.,
$f_X(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$
let $\tau(\theta) = 1/\theta \Rightarrow \tau'(\theta) = -1/\theta^2$.
$CRLB = 1/n(\theta^2)$.

Now consider

$$\begin{aligned}\sum \frac{\partial ln f_X(x_i;\theta)}{\partial \theta} &= \sum_{i=1}^{n}(1/\theta - x_i) \\ &= n/\theta - \sum_{i=1}^{n} x_i \\ &= -n(\overline{x} - 1/\theta)\end{aligned}$$

Therefore, $\overline{X}$ is UMVU estimator of $1/\theta$, since $var(\overline{X}) = 1/n\theta^2$ which is the CRLB for the variance of unbiased estimate of $1/\theta$

Example

Let $X_1 \ldots X_n$ be a random sample from $Poisson(\theta)$; i.e.,
$f_X(x;\theta) = \frac{e^{-\theta}\theta^x}{x!} \ldots$ for x=0,1,...
Let $\tau(\theta) = \theta \Rightarrow \tau'(\theta) = 1$
$\frac{\partial ln f_X(x;\theta)}{\partial \theta} = -1 + x/\theta$
$CRLB = \theta/n$

$$\begin{aligned}\sum \frac{\partial ln f_X(x_i;\theta)}{\partial \theta} &= \sum_{i=1}^{n}(x/\theta - 1) \\ &= \frac{\sum_{i=1}^{n} x_i}{\theta} - n \\ &= n/\theta(\overline{x} - \theta).\end{aligned}$$

Therefore, $\overline{X}$ is UMVU estimator of θ, since $var(\overline{X}) = \frac{\theta}{n} =$ CRLB.

Example
Let S^2 denote the variance of a random sample of size $n > 1$ from a distribution which is $N(\mu, \theta)$, $0 < \theta < \infty$. We know that $E[nS^2/(n-1)] = \theta$. What is the efficiency of the statistic $nS^2/(n-1)$?

Solution:

$$\begin{aligned}lnf(x;\theta) &= -\frac{(x-\mu)^2}{2\theta} - \frac{ln(2\pi\theta)}{2}, \\ \frac{\partial lnf(x;\theta)}{\partial \theta} &= \frac{(x-\mu)^2}{2\theta^2} - \frac{1}{2\theta} \\ \text{and } \frac{\partial^2 lnf(x;\theta}{\partial \theta^2} &= \frac{(x-\mu)^2}{\theta^3} + \frac{1}{2\theta^2} \\ Hence, -E\left[\frac{\partial^2 lnf(x;\theta)}{\partial \theta^2}\right] &= \frac{1}{2\theta^2}\end{aligned}$$

Thus the *Cramér* lower bound is $2\theta^2/n$
We know that $nS^2/\theta \sim \chi^2(n-1)$, so that the variance of nS^2/θ is $2(n-1)$. Accordingly, the variance of $nS^2/(n-1)$ is $2\theta^2/(n-1)$. Thus the efficency of the statistic $nS^2/(n-1)$ is $(n-1)/n$.

2.7 Point estimation using R

Point estimation using the Methods of Moments and the Maximum likelihood

```
## Point Estimation: Method of estimation
## R code for Figure 2.1
theta=10
sampsz=10
nsim=100
moment.estimates=numeric(nsim)
ML.estimates=numeric(nsim)
for(i in 1:nsim)
{ru=runif(n=sampsz,min=0,max=theta)
moment.estimates[i]=2*mean(ru)
ML.estimates[i]=max(ru)}
plot(density(moment.estimates),xlab="",
ylab="",main="",ylim=c(0,0.6),las=1)
abline(v=theta,lty=3)
lines(density(ML.estimates),lty=2)
legend(11,0.5,legend=c("moment","ML"),lty=1:2,cex=0.6)
You should see that the method of moments
unbiased estimates of which many are not in the range space.
The maximum liklihood estimates almost all are less than 10.
```

```
## R code for Figure 2.2
##Normal Moments
##Method of moments estimator of the mean and the variance
of N(14,16)
mu=14
sigma=4
sampsz=10
nsim=100
mu.est=numeric(nsim)
var.est=numeric(nsim)
for(i in 1:nsim){
rn=rnorm(mean=mu,sd=sigma,n=sampsz)
mu.est[i]=mean(rn)
var.est[i]=mean((rn-mean(rn))^2)}
par(mfrow=c(2,1))
plot(density(mu.est))
abline(v=mu,lty=3)
plot(density(var.est))
abline(v=sigma,lty=3)
Note that the Figure 1.2 shows that the sample mean
is unbiased estimate of the population mean while
the sample variance with n in the denominator is not
```

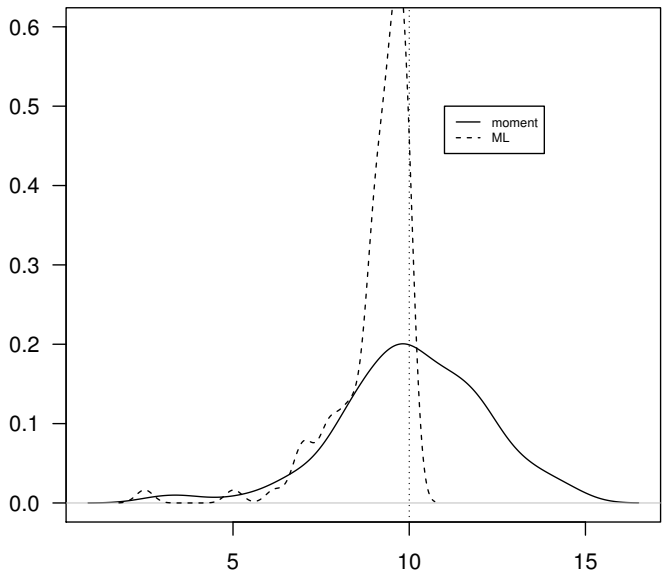


Figure 2.1: Method of Estimation

Consistency

```
> ## Consistency(Figure 2.3)##
> ##To demonstrate that the MLE is consistent for
estimating theta for uniform(0,theta)
> theta=10
> sampsz=10
> nsim=100
> ml.est=numeric (nsim)
> for(i in 1:nsim){
+ ru=runif(n=sampsz,min=0,max=theta)
+ if(i==1) ml.est[i]=max(ru)
+ else ml.est[i]=max(ml.est[i-1],max(ru))}
> plot(ml.est,type="l")
> abline(h=theta,lty=2)
> #Note that as n increases the MLestimate
#approaches the value of the parameter.
```

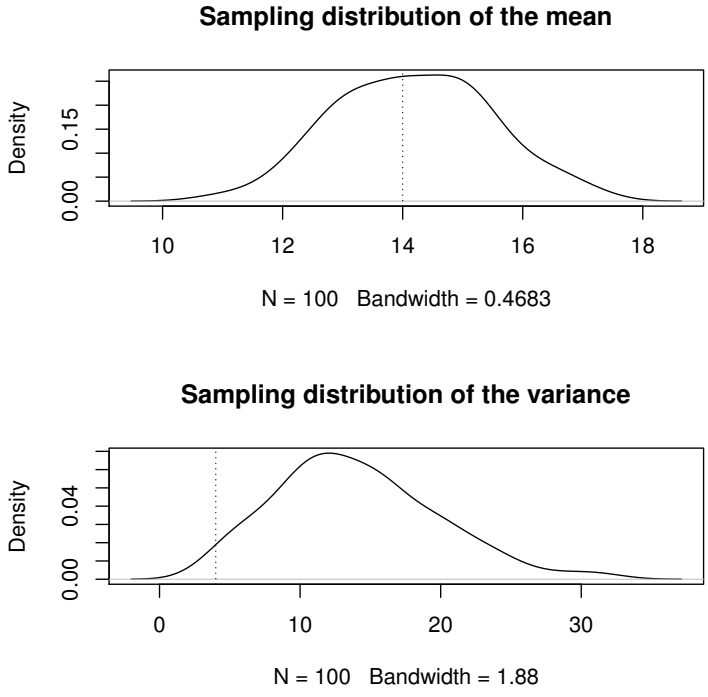


Figure 2.2: Normal Moments

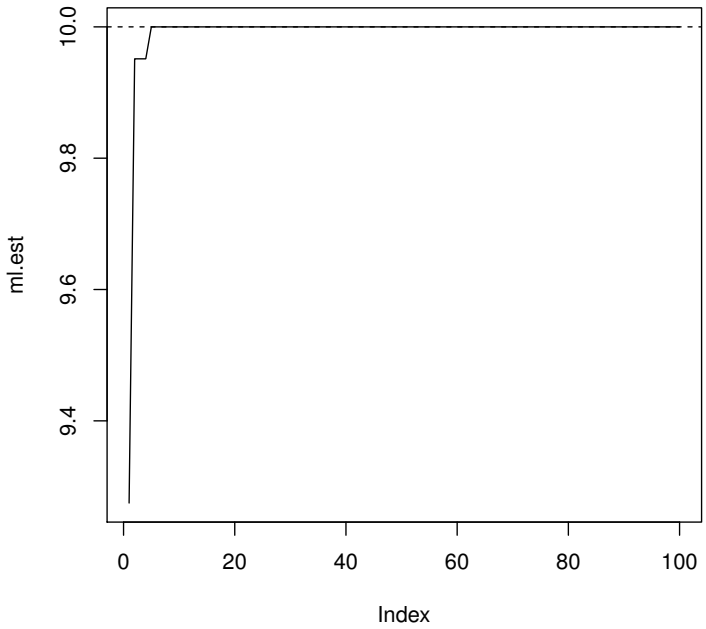


Figure 2.3: consistency

Relative efficiency

```
## MLE of the parameter of the cauchy distribution and relative
#efficiency. Here the true unknown alpha is zero, but we pretend
#we dont know that and see how good the MLE is as an estimator.
> n <- 30
> set.seed(42)
> x <- rcauchy(n)
> mu.start <- median(x) ## median as an estimator
> mu.start
[1] -0.1955062
> out2= mlogl2 <- function(mu, x) {
+ sum(log(1 + (x - mu)^2))
+ }
> out2 <- nlm(mlogl2, mu.start, x = x) ##MLE as estimator
> mu.hat <- out2$estimate
> mu.hat
[1] -0.1816501
#We see for these data, the MLE is slightly better than the sample
# median. But this is just one data set.For random data sometimes
#the MLE will be better and sometimes the sample median will be better.
#As statisticians, what we are interested is in the sampling
distributions of the two estimators, which we can easily study by simulation.
> nsim <- 100
> mu <- 0
> mu.hat <- double(nsim)
> mu.twiddle <- double(nsim)
> for (i in 1:nsim) {
+ xsim <- rcauchy(n, location = mu)
+ mu.start <- median(xsim)
+ out <- nlm(mlogl2, mu.start, x = xsim)
+ mu.hat[i] <- out$estimate
+ mu.twiddle[i] <- mu.start
+ }
> mean((mu.hat - mu)^2)
[1] 0.06203112
> mean((mu.twiddle - mu)^2)
[1] 0.08242236
> #The two numbers reported from the simulation
#are the mean square errors (MSE) of the two estimators.
# Their ratio
> mean((mu.hat - mu)^2)/mean((mu.twiddle - mu)^2)
[1] 0.7526007
> #is the asymptotic relative efficiency (ARE) of the estimators.
# Now we see that the MLE is more accurate, as theory says it must be.
```

MLE of of the parameters of $N(\theta_1; \theta_2)$

```
> ##MLE for the parameters of a normal population
#using sample of observations
> x=c(2,5,3,7,-3,-2,0)
> fn=function(theta,x){
+ sum(0.5*(x-theta[1])^2/theta[2]+0.5*log(theta[2]))}
> op=optim(c(2,9),fn,x=x,hessian=T)
> #c(2,9) are initial values for the parameters
#to be optimized over
> #fn=A function to be maximized, with first argument
> #the vector of parameters over which maximization is to take place.
> #It should return a scalar result
> op
$par
[1] 1.713956 11.347966
$value
[1] 12.00132
$counts
function gradient
45 NA
$convergence
[1] 0
$message
NULL
$hessian
[,1] [,2]
[1,] 6.168506e-01 1.793543e-05
[2,] 1.793543e-05 2.717398e-02
> ## 1.713956 and 11.347966 are MLEs of theta[1] and theta[2]
```

2.8 Exercises

1. Let Y denote the number of successes in n independent Bernoulli trials with parameter θ, and define $T_1 = \frac{Y}{n}$ and $T_2 = \frac{Y+1}{n+2}$. Find and compare the MSE's of T_1 and T_2 when $n = 4$ and when $n = 8$.

2. Show that the function $f_X(x;\alpha,\beta) = \frac{1}{\beta}e^{\frac{-(x-\alpha)}{\beta}}$, $x \geq \alpha$, $\alpha \in R$, $\beta > 0$ is a probability density function.

3. On the basis of a random sample of size n from the density function of exercise 2 above, determine

   (i) MLE of α when β is known.

   (ii) MLE of β when α is known.

   (iii) MLE of α and β when both are unknown.

   (iv) a sufficient statistic for β when α is known.

   (v) a sufficient statistic for α when β is known.

   (vi) a set of sufficient statistics for β and α when they are both unknown.

   (vii) Show that $E(X) = \alpha+\beta$ and $E(X^2) = \alpha^2+2\alpha\beta+2\beta^2$ and calculate $Var(X)$.

   (viii) Derive the moment estimators of β and α

4. If $X \sim Exp(\theta)$, then $E(X) = \frac{1}{\theta}$. So a natural candidate for estimating θ from a random sample of size n is $\widehat{\theta} = \frac{1}{\overline{X}}$.

   (i) Calculate $E(\frac{1}{\overline{X}})$ when $n > 1$.

   (ii) From the result of (a) find an unbiased estimator of the parameter θ and calculate its MSE.

   (iii) Show that the multiple of $\frac{1}{\overline{X}}$ with the smallest MSE in estimating θ is $(n-2)/\sum_{i=1}^{n} X_i$.

5. Let X be a r.v. denoting the life span of an equipment. Then the reliability of the equipment at time x, $R(x)$, is defined as $P(X > x)$. Now suppose X has an exponential distribution with parameter θ. Then:

   (i) Calculate the reliability based on this r.v.

   (ii) Determine the MLE of $R(x)$ on the basis of a random sample of size n from this density.

6. Consider a random sample X of size n from a geometric distribution :
   $f(x|p) = p(1-p)^{x-1}$, $x = 1, 2, ...$ Define the estimator U as the indicator function of the event $X_1 = 1$.

   (i) Show that U is an unbiased estimator of p.

   (ii) Find a sufficient statistic, T.

   (iii) Calculate $E(U|T)$.

7. Let $X_1, ..., X_n$ be a random sample of size n from $U(\theta_1, \theta_2)$ distribution, $\theta_1 < \theta_2$, and let $Y_{(1)}$ and $Y_{(n)}$ be the smallest and the largest order statistics of the X_i^s. Obtain the density function of Y_1, Y_2 and then by calculating $E(Y_{(1)})$ and $E(Y_{(n)})$, Construct unbiased estimators of the mean $\frac{\theta_1+\theta_2}{2}$ and for the range $(\theta_1 + \theta_2)$ depending only on $Y_{(1)}$ and $Y_{(n)}$.

8. For estimating the parameter θ of the uniform distribution on $(0, \theta)$ based on a random sample of size n,

   (i) Find the method of moments estimator.

   (ii) Show that the estimator in (i) is consistent.

9. Let $X_1, ..., X_n$ be a random sample from the Gamma (r, λ) with r known and λ unknown. Let $\frac{1}{\lambda} = \theta$.

   (i) Determine the Fisher information $I(\theta)$.

   (ii) Show that the estimate $U(X_1, X_2, ..., X_n) = \frac{\sum_{i=1}^{n} X_i}{nr}$ is unbiased and calculate its variance.

   (iii) Show that $Var(U) = \frac{1}{n} I(\theta)$, so that U is UMVU estimator of θ.

10. Given: n independent pairs (X_i, Y_i), each with joint density function

$$f_{X,Y}(x, y; \theta) = e^{\frac{-\theta x - y}{\theta}}$$

    for $x > 0, y > 0$, where $\theta > 0$.

    (i) Find the minimal sufficient statistic.

    (ii) Find the MLE of θ. (Is it sufficient?)

11. Let $X_1, ..., X_n$ be a random sample from the geometric p.d.f.

    (i) Show that X is both sufficient and complete.

    (ii) Show that the estimate U defined by: $U(X) = 1$ if $X = 1$, and $U(X) = 0$ if $X = 0$,is unbiased estimate of θ.

    (iii) Conclude that U is the UMVU estimate of θ and also an entirely unreasonable estimate.

    (iv) Prove that the variance of U is uniformly bigger than the Cramer -Rao lower bound.

12. Consider independent observations $Y_1, ..., Y_n$, where each Y_i is $N(\alpha + \beta x_i; 1)$, for given constants $x_1, ..., x_n$. Find the joint MLE of the parameters (α, β).

13. Consider an observation from a density function $f_X(x; \theta) = (1 - \theta)\theta^{x-1}$, $x = 1, 2, ...$, $\theta \in (0, 1)$. Assume that θ has a uniform prior distribution on the interval (0,1). Then,

    (i) Determine the posterior density function of θ, given $X = x$.

    (ii) Obtain the Bayes estimator of θ with respect to squared error loss.

14. Given a random sample of size n from $X \sim Exp(\theta)$ with $\theta > 0$. For an $Exp(\beta)$ prior, find bayes estimate of θ assuming a quadratic loss function.

15. Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\theta; 1)$, $\theta \in I\!R$ and on $I\!R$, consider the density function of θ to be that of $N(\mu; 1)$ with μ known. Then show that the Bayes estimator with respect to squared error loss of θ, is given by :$\frac{n\overline{x}+\mu}{n+1}$.

16. Suppose we observe a Bernoulli process with parameter θ and found that it took 15 trials to get the 4^{th} success. If our prior for θ is $Beta(4; 2)$ and if we assume quadratic loss, find the Bayes estimator of θ?

17. Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\theta; \theta)$,$\theta > 0$.

    (i) Find a complete sufficient statistic if such exists.

    (ii) Argue that $\overline{X}$ is not an UMVU estimator of θ.

    (iii) Is θ either a location or a scale parameter?

18. Let $\overline{X}$ denote the mean of a random sample of size $n = 5$ from $X \sim N(\mu; 1)$. Given that $\mu \sim N(5; 1)$ and $\overline{X} = 4$, find the Bayes estimator assuming absolute error loss.

19. Let $Z_1, ..., Z_n$ be a random sample from $X \sim N(0; \theta^2)$, $\theta > 0$. Define $X_i = |Z_i|$, and consider estimation of θ and θ^2 on the basis of the random sample $X_1, ..., X_n$.

    (i) Find the UMVU estimator of θ^2 if such exists.

    (ii) Find an estimator of θ^2 that has uniformly smaller mean-squared error than the estimator that you found in part (i).

    (iii) Find the UMVU estimator of θ if such exists.

20. Suppose a lot of 10 items has M defective, and a simple random sample of size four includes exactly one defective. Find the Bayes estimator of M, when $M \sim B(10; \frac{1}{2})$, assuming quadratic loss.

21. Let $X_1, ..., X_n$ be a random sample from $f(x; \theta) = e^{-(x-\theta)} I_{[\theta,\infty)}(x)$ for $-\infty < \theta < \infty$.

    (i) Find a sufficient statistic for θ.

    (ii) Find a maximum-likelihood estimator of θ.

    (iii) Find a method of moments estimator of θ.

    (iv) Is there a complete sufficient statistic? If so, find it.

    (v) Find the UMVU estimator of θ if one exists.

    (vi) Using the prior density $g(\theta) = e^{-\theta} I_{0,\infty}(\theta)$, find the Bayes estimator of θ and use quadratic loss.

22. Show that the family of densities with density function

$$f(x; r, s) \propto x^{r-1}(1-x)^{s-1},$$

    $0 < x < 1$, where $r > 0$,$s > 0$, is in the two-parameter exponential family. (The constant of proportionality will involve the parameters r and s.)

23. Let$X_1, ..., X_n$ be a random sample from the density

$$f(x; \alpha, \theta) = (1 - \theta)\theta^{x-\alpha} I_{(\alpha, \alpha+1, ...)}(x),$$

where $-\infty < \alpha < \infty$ and $0 < \theta < 1$.

(i) Find a two-dimensional set of sufficient statistics.

(ii) Find the maximum likelihood estimator of (α, θ)

24. Let X be a r.v. having the Negative Binomial distribution with parameter $\theta \in (0, 1)$. Find the UMVU estimator of $g(\theta) = \frac{1}{\theta}$ and determine its variance.

25. Let $X_1, ..., X_n$ be i.i.d. r.v's from the $U(\theta; 2\theta)$, $\theta \in (0, \infty)$ distribution and set

$$U_1 = \frac{n+1}{2n+1} X_(n)$$

and

$$U_2 = \frac{n+1}{5n+4}[2X_{(n)} + X_{(1)}].$$

Then show that both U_1 and U_2 are unbiased estimators of θ and that U_2 is uniformly better than U_1 (in the sense of variance).

26. Suppose that certain particles are emitted by a radioactive source (whose strength remains the same over a long period of time) according to a Poisson distribution with parameter θ during a unit of time. The source in question is observed for n time units, and let X be the r.v. denoting the number of times that no particles were emitted. Find the MLE of θ in terms of X.

Chapter 3

Interval Estimation

3.1 Introduction

In this chapter we move away from inference based upon the use of a single estimate of an unknown population quantity, focusing instead upon interval estimation, or also known as set estimation.

3.2 Problems with point estimators

An estimator is a statistic and therefore a random variable and it will have a probability distribution function. In this respect the use of a single statistic as a point estimate ignores the inherent variation in the random variable. In addition, for continuous variables the probability that a random variable assumes a single value is zero.
Instead of choosing one plausible point, one may try to determine a plausible **subset** of the parameter space Θ. This is called **set estimation** (**or interval estimation**, in the case that $\Theta \subset \mathbb{R}$).

If $D(x_1, \ldots, x_n)$ is such a subset of Θ (depending on $x_1, \ldots, x_n$, but not on θ) we would like to have that

$$P_\theta(\theta \in D(X_1, \ldots, X_n))$$

(the probability that the **random set contains** θ) is large. Therefore, the statistician chooses a small number $\alpha \in [0, 1]$ (e.g. $\alpha = 0.05$) and tries to construct a set such that

$$P_\theta(\theta \in D(X_1, \ldots, X_n)) = 1 - \alpha \qquad \text{, for all } \theta \in \Theta$$

Such a region $D(x_1, \ldots, x_n)$ is called a $100(1-\alpha)\%$ **confidence region for** θ.

Note

Sometimes, particularly in discrete models, we cannot find a region for which this probability is exactly $1-\alpha$, for a given preassigned α. If so, we try to have **at least** $1-\alpha$ and as close as possible to $1-\alpha$:

$$P_\theta(\theta \in D(X_1,\ldots,X_n)) \geq 1-\alpha \qquad \text{, for all } \theta \in \Theta.$$

3.2.1 Confidence intervals

The general idea from the introduction becomes simple in the case of a single real parameter $\theta \in \Theta \subset I\!R$. In this case, a confidence region $D(x_1,\ldots,x_n)$ is typically of the form

$$[l(x_1,\ldots,x_n), r(x_1,\ldots,x_n)]$$

i.e. an interval with $l(x_1,\ldots,x_n)$ and $r(x_1,\ldots,x_n)$ in Θ. The functions l and r will be such that, for a sample $X_1,\ldots,X_n : l(X_1,\ldots,X_n)$ and $r(X_1,\ldots,X_n)$ are statistics.

Definition

Let $X_1,\ldots,X_n$ be a random sample from X with density $f(x;\theta), \theta \in \Theta \subset I\!R$.
Let $\alpha \in]0,1[$.
If

$$L_n = l(X_1,\ldots,X_n) \qquad \text{and} \qquad R_n = r(X_1,\ldots,X_n)$$

are two statistics satisfying

$$P_\theta(L_n \leq \theta \leq R_n) = 1-\alpha \qquad \text{, for all } \theta \in \Theta$$

then the random interval $[L_n, R_n]$ is called a $100(1-\alpha)\%$ **interval estimator for** θ.
For observations $x_1,\ldots,x_n$, the corresponding interval estimate for θ

$$[l(x_1,\ldots,x_n), r(x_1,\ldots,x_n)]$$

is called a $100(1-\alpha)\%$ **confidence interval for** θ.

Definition: One Sided Lower Confidence Interval

Let $T_1(\underset{\sim}{X}) = t_1(X_1,\ldots,X_n)$ be a statistic such that $P[T_1 \leq \theta] = 1-\alpha$. $[T_1,\infty)$ is a one sided lower $(1-\alpha)100\%$ **confidence interval** for θ.
For observations $x_1,\ldots,x_n$, the corresponding interval estimate for θ

$$[t_1(x_1,\ldots,x_n), \infty)$$

is called a $100(1-\alpha)\%$ **lower confidence interval for** θ.

Definition: One Sided Upper Confidence Interval

Let $T_2(\underset{\sim}{X}) = t_2(X_1, \ldots, X_n)$ be a statistic such that $P[T_2 \geq \theta] = 1 - \alpha$. $(-\infty, T_2]$ is a one sided upper$(1-\alpha)100\%$ **confidence interval** for θ.
For observations $x_1, \ldots, x_n$, the corresponding interval estimate for θ

$$(-\infty, t_2(x_1, \ldots, x_n)]$$

is called a $100(1-\alpha)\%$ **upper confidence interval for** θ.

Example Let $X_1, \ldots, X_n$ be a random sample from Exp(θ). We wish to drive a one sided lower $100(1-\alpha)\%$ confidence interval for θ. We know that $\overline{X}$ is sufficient for θ and also that $2n\overline{X}/\theta \sim \chi^2(2n)$.Thus,

$$\begin{aligned} P[2n\overline{X}/\theta < \chi^2_{2n;1-\alpha}] &= 1-\alpha \\ P[2n\overline{X}/\chi^2_{2n;1-\alpha} < \theta] &= 1-\alpha \end{aligned}$$

Similarly, a one sided upper $100(1-\alpha)\%$ confidence interval is obtained from:

$$P[\theta < 2n\overline{X}/\chi^2_{2n;\alpha}] = 1-\alpha$$

3.2.2 A method for finding confidence interval

Pivotal Quantity
Let $X_1, \ldots X_n$ denote a random sample with common density $f_X(;\theta)$. Let $Q = q(X_1, \ldots, X_n; \theta)$. If Q has a distribution that does not depend on θ, Q is a pivotal quantity.

Example Let$X_1, \ldots, X_n$ be a random sample from $N(\mu; 9)$.
$\overline{X} \sim N(\mu; 9/n)$ is not a pivotal quantity as it depends on μ
$\frac{\overline{X}-\mu}{3/\sqrt{n}} \sim N(0;1)$ is a pivotal quantity
$\overline{X}/\mu \sim N(1; 9/n\mu^2)$ is not a pivotal quantity.

Pivotal Quantity Method

If $Q = q(x;\theta)$ is a pivotal quantity with known probability density function, then for any fixed $0 < (1-\alpha) < 1$,there exists q_1, q_2 depending on $(1-\alpha)$ such that

$$P[q_1 < t(x_1, ..., x_n) < q_2] = 1-\alpha$$

If for each sample realization $(x_1, \ldots, x_n)$

$$q_1 < Q(\underset{\sim}{x};\theta) < q_2$$

iff functions $t_1(x_1, \ldots, x_n) < \theta < t_2(x_1, \ldots, x_n)$ for functions t_1 and t_2, then $(T_1, T_2)$is a $100(1-\alpha)\%$confidence interval for θ.

Note:

(i) q_1 and q_2 are independent of θ.

(ii) For any fixed $(1-\alpha)$ there exists many possible pairs of numbers (q_1, q_2), such that $P[q_1 < Q < q_2] = 1-\alpha$ as we will show below.

(iii) Essential feature of this method is that the inequality $P[q_1 < Q < q_2]$ can be pivoted as

$$[t(.) < \theta < t(.)]$$

for any set of sample values $x_1, \ldots, x_n$.

3.2.3 Criteria for comparing confidence intervals

As mentioned above for any fixed $(1-\alpha)$ there are many possible pairs of numbers q_1 and q_2 that can be selected so that $P(q_1 < Q < q_2) = 1-\alpha$.

Example Let $X_1, \ldots, X_{25}$ be a random sample of size 25 from $N(\theta; 9)$. We wish to construct a 95% C.I. for θ.
$\overline{X}$ is the maximum liklihood estimatior of θ.
$\frac{\overline{X}-\theta}{\sigma/\sqrt{n}} \sim N(0;1) \Rightarrow \frac{\overline{X}-\theta}{\sigma/\sqrt{n}}$ is a pivotal quantity.
For given $(1-\alpha)$, we can find q_1 and q_2 such that

$$\begin{aligned} P[q_1 < \frac{\sqrt{n}(\overline{X}-\theta)}{\sigma} < q_2] &= 1-\alpha \\ P[\overline{X} - \frac{\sigma q_2}{\sqrt{n}} < \theta < \overline{X} - \frac{\sigma q_1}{\sqrt{n}}] &= 1-\alpha \end{aligned}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for θ is $(\overline{X} - \frac{\sigma q_2}{\sqrt{n}}, \overline{X} - \frac{\sigma q_1}{\sqrt{n}})$. Let the sample mean computed from 25 observations be $\overline{x}$ =17.5. Then inserting this value in the inequality above we have the following possible confidence intervals: $CI_1(16.32, 18.68)$ and $CI_2(16.49, 19.12)$.

How does CI_2 compares to CI_1? Obviously CI_1 is superior to CI_2, since the length of $CI_1 = 2.36$ is less than the length of $CI_2 = 2.63$.
We want to select q_1 and q_2 that will make t_1 and t_2 close together. This can be achieved by selecting q_1 and q_2 such that the length of the interval is the shortest, or the average length of the random interval the smallest. Such an interval is desirable since it is more informative. We have to note also that shortest-length confidence intervals do not always exist.
For the previous example, the length of the confidence interval is given by

$$[\overline{X} - q_1(\sigma/\sqrt{n})] - [\overline{X} - q_2(\sigma/\sqrt{n})] = (q_2 - q_1)(\sigma/\sqrt{n})$$

We have to select q_1 and q_2, such that $(q_2 - q_1)$ is minimum under the restriction that $P(q_1 < Q < q_2) = 1-\alpha$. This is true if $q_1 = -q_2$. Such an interval is a $100(1-\alpha)\%$ shortest-length confidence interval based on Q.

Example Let $X_1, \ldots, X_n$ be a random sample from $N(\mu; \sigma^2)$, where σ^2 is known. Consider the pivotal quantity:

$$Q(\underset{\sim}{X}; \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Then

$$P[\overline{X} - q_2(\sigma/\sqrt{n}) < \mu < \overline{X} - q_1(\sigma/\sqrt{n})] = 1 - \alpha$$

The length of the confidence interval is $L = (\sigma/\sqrt{n})(q_2 - q_1)$. We wish to minimize L such that

$$\phi(q_2) - \phi(q_1) = \int_{q_1}^{q_2} f_X(x)dx = 1\alpha$$

where $f_X(x) = 1/\sqrt{2\pi}e^{-x^2/2}$.

$$dL/dq_1 = \frac{\sigma}{\sqrt{n}}(dq_2/dq_1 - 1)$$

and

$$f_X(q_2)\frac{dq_2}{dq_1} - f_X(q_1) = 0$$

which give us

$$dL/dq_1 = \frac{\sigma}{\sqrt{n}}\left[\frac{f_X(q_1)}{f_X(q_2)} - 1\right]$$

The minimum occurs when $f_X(q_1) = f_X(q_2)$,that is, when $q_1 = -q_2$

Note: For some problems, the equal tailed choice of q and $-q$ will provide the minimum expected length, but for others it will not.

3.3 Confidence interval for the parameters of a normal population

3.3.1 The one sample problem

Let $X_1, \ldots, X_n$ be a random sample of X with $X \sim N(\mu; \sigma^2)$

Example [Confidence interval for μ if σ^2 is known]

A natural estimator for μ is the ML-estimator $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We have, by the central limit theorem,

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0; 1).$$

Hence, for any a

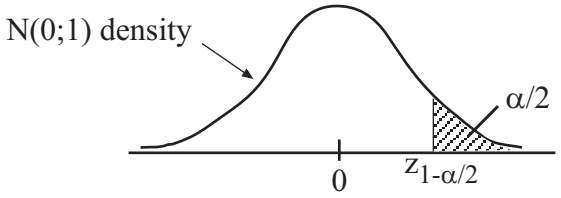
$$P\left(-a \leq \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq a\right) = \Phi(a) - \Phi(-a)$$

or

$$P\left(\overline{X} - a\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{X} + a\sqrt{\frac{\sigma^2}{n}}\right) = \Phi(a) - \Phi(-a)$$

where Φ is the standard normal distribution function.
Let us now choose a such that

$$\begin{aligned} & \Phi(a) - \Phi(-a) = 1 - \alpha \\ \text{or} \quad & 2[1 - \Phi(a)] = \alpha \\ \text{or} \quad & \Phi(a) = 1 - \frac{\alpha}{2} \\ \text{or} \quad & a = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \equiv z_{1-\frac{\alpha}{2}} \qquad \text{(notation)} \end{aligned}$$



Then we have :

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

Conclusion : if $x_1, \ldots, x_n$ are the observed values of a sample from $X \sim N(\mu; \sigma^2)$ with σ^2 known, then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\overline{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right].$$

Example [Confidence interval for μ if σ^2 is unknown]

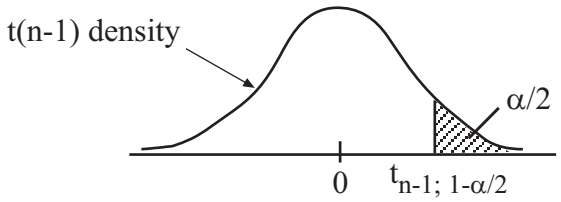
We replace σ^2 in the previous example by the unbiased estimator $\frac{nS^2}{n-1}$. We know :

$$\frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n-1}}} \sim t(n-1).$$

As before, we obtain :

$$P\left(\overline{X} - t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}} \leq \mu \leq \overline{X} + t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}}\right) = 1 - \alpha$$

where $t_{n-1;1-\alpha/2} = F^{-1}(1-\frac{\alpha}{2})$ with F the distribution function of a $t(n-1)$ random variable :



Conclusion : if $x_1, \ldots, x_n$ are the observed values of a sample from $X \sim N(\mu; \sigma^2)$ with σ^2 unknown, then a $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\overline{x} - t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n-1}}, \overline{x} + t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n-1}}\right].$$

Example [Confidence interval for σ^2 if μ is known]

The ML-estimator for σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$ and we know that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2(n).$$

Hence, for all $0 < a < b$:

$$P\left(a \leq \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \leq b\right) = F(b) - F(a)$$

or

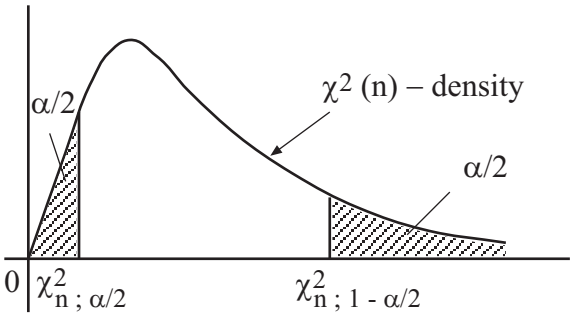
$$P\left(\frac{1}{b}\sum_{i=1}^{n}(X_i-\mu)^2 \leq \sigma^2 \leq \frac{1}{a}\sum_{i=1}^{n}(X_i-\mu)^2\right) = F(b) - F(a)$$

where F is the distribution function of a $\chi^2(n)$ random variable.
In order to obtain a $100(1-\alpha)\%$ confidence interval, we have to choose a and b such that

$$\begin{aligned} F(b) - F(a) &= 1-\alpha \\ \text{or } [1-F(b)] + F(a) &= \alpha \end{aligned}$$

A possible choice is

$$\begin{aligned} 1-F(b) &= F(a) &&= \tfrac{\alpha}{2} \\ \text{i.e.} \qquad a &= F^{-1}\left(\tfrac{\alpha}{2}\right) &&\equiv \chi^2_{n;\alpha/2} \\ b &= F^{-1}\left(1-\tfrac{\alpha}{2}\right) &&\equiv \chi^2_{n;1-\alpha/2} \end{aligned}$$



Conclusion : a $100(1-\alpha)\%$ confidence interval for σ^2 if μ is known is given by

$$\left[\frac{1}{\chi^2_{n;1-\alpha/2}}\sum_{i=1}^{n}(x_i-\mu)^2, \frac{1}{\chi^2_{n;\alpha/2}}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Example [Confidence interval for σ^2 if μ is unknown]
Use the fact that

$$\frac{nS^2}{\sigma^2} \sim \chi^2(n-1).$$

Conclusion : a $100(1-\alpha)\%$ confidence interval for σ^2 if μ is unknown is given by

$$\left[\frac{n}{\chi^2_{n-1;1-\alpha/2}}.s^2, \frac{n}{\chi^2_{n-1;\alpha/2}}.s^2\right].$$

3.3.2 The two sample problem

Let$X_1,\ldots,X_{n1}$ and $Y_1,\ldots,Y_{n2}$ be, respectively two random samples of sizs n_1 and n_2 from the two normal distributions $N(\mu_1;\sigma_1^2)$ and $N(\mu_2;\sigma_2^2)$.

Example [Confidence interval for $\mu_2-\mu_1$, if σ_1^2 and σ_2^2 are known]

A $100(1-\alpha)\%$ confidence interval for $\mu_2-\mu_1$ if σ_1^2 and σ_2^2 are known, is given by

$$\left[\overline{y}-\overline{x}-z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\ ,\ \overline{y}-\overline{x}+z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right].$$

Example [Confidence interval for $\mu_2 - \mu_1$ if $\sigma_1^2 = \sigma_2^2 = \sigma^2$, but unknown]

To construct a confidence interval for $\mu_2 - \mu_1$, we consider the estimator $\overline{Y} - \overline{X}$, where

$$\overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i \ , \ \overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i.$$

Denote the sample variances by

$$S_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2 \ , \ S_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2.$$

We have
$\overline{Y} - \overline{X} \sim N(\mu_2 - \mu_1; \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$

$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$
$\frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$
$\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$

Define the "pooled variance" S_p^2 by

$$S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

Then, we have

$$\frac{\overline{Y} - \overline{X} - (\mu_2 - \mu_1)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{\overline{Y} - \overline{X} - (\mu_2 - \mu_1)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\sqrt{\frac{\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2}}{n_1 + n_2 - 2}}} \sim t(n_1 + n_2 - 2).$$

Conclusion : a $100(1-\alpha)\%$ confidence interval for $\mu_2 - \mu_1$, if $\sigma_1^2 = \sigma_2^2$ but unknown, is given by

$$\left[\overline{y} - \overline{x} - t_{n_1+n_2-2;1-\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \overline{y} - \overline{x} + t_{n_1+n_2-2;1-\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right].$$

Example [Confidence interval for $\mu_2 - \mu_1$, if possibly $\sigma_1^2 \neq \sigma_2^2$]

It is natural to use the distribution function of

$$T = \frac{\overline{Y} - \overline{X} - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}}}$$

but unfortunately, this distribution depends on the unknown σ_1^2 and σ_2^2 for fixed n_1, n_2. This is known as the **Behrens-Fisher problem**.

There are several solutions to this problem. One of them is due to **Welch**

The distribution of T is approximately $t(\overset{\wedge}{\nu})$, where

$$\overset{\wedge}{\nu} = \frac{\left(\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1 - 1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2 - 1}\right)^2}.$$

If $\overset{\wedge}{\nu}$ is not an integer, then we take the degrees of freedom equal to the integer nearest to $\overset{\wedge}{\nu}$.

The idea behind this solution is to approximate the distribution of $\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}$ by that of a $\chi^2(\nu)$ variable, multiplied by $\frac{\sigma^2}{\nu}$, where σ^2 and ν are chosen so that the first two moments of $\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}$ agree with the first two moments of $\frac{\sigma^2}{\nu}.\chi^2(\nu)$.

Now,

$$E\left(\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$Var\left(\frac{S_1^2}{n_1 - 1} + \frac{S_2^2}{n_2 - 1}\right) = \frac{2\sigma_1^4}{(n_1 - 1)n_1^2} + \frac{2\sigma_2^4}{(n_2 - 1)n_2^2}$$

$$E\left(\frac{\sigma^2}{\nu}\chi^2(\nu)\right) = \frac{\sigma^2}{\nu}.\nu = \sigma^2$$

$$Var\left(\frac{\sigma^2}{\nu}\chi^2(\nu)\right) = \frac{\sigma^4}{\nu^2}.2\nu = 2\frac{\sigma^4}{\nu}$$

Hence

$$\begin{cases} \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} & = & \sigma^2 \\ \frac{\sigma_1^4}{(n_1 - 1)n_1^2} + \frac{\sigma_2^4}{(n_2 - 1)n_2^2} & = & \frac{\sigma^4}{\nu} \end{cases}$$

This gives :

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{\sigma_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{\sigma_2^2}{n_2}\right)^2} .$$

The unknown parameters σ_1^2 and σ_2^2 are now replaced by estimates $\frac{n_1 s_1^2}{n_1 - 1}$ and $\frac{n_2 s_2^2}{n_2 - 1}$.
This gives :

$$\overset{\wedge}{\nu} = -\frac{\left(\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1 - 1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2 - 1}\right)^2} .$$

Conclusion : an **approximate** $100(1-\alpha)\%$ confidence interval for $\mu_2 - \mu_1$ in the case of possibly unequal variances σ_1^2 and σ_2^2 is given by

$$\left[\overline{y} - \overline{x} - t_{\overset{\wedge}{\nu};1-\alpha/2}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}, \overline{y} - \overline{x} + t_{\overset{\wedge}{\nu};1-\alpha/2}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}\right] .$$

Notes

- It can be shown that
  $\min(n_1 - 1, n_2 - 1) \leq \overset{\wedge}{\nu} \leq n_1 + n_2 - 2$
- For $n_1 = n_2 = n$ and $\sigma_1^2 = \sigma_2^2 : \overset{\wedge}{\nu} = 2n - 2$

Example [Confidence interval for $\frac{\sigma_2^2}{\sigma_1^2}$, if μ_1 and μ_2 are known]

use that

$$\frac{\frac{\sum_{i=1}^{n_1}(X_i - \mu_1)^2}{\sigma_1^2} \Big/ n_1}{\frac{\sum_{i=1}^{n_2}(Y_i - \mu_2)^2}{\sigma_2^2} \Big/ n_2} \sim F(n_1, n_2)$$

Conclusion : a $100(1-\alpha)\%$ confidence interval for the ratio σ_2^2/σ_1^2 if μ_1 and μ_2 are known is given by

$$\left[F_{n_1,n_2;\alpha/2} \frac{n_1}{n_2} \frac{\sum_{i=1}^{n_2}(y_i-\mu_2)^2}{\sum_{i=1}^{n_1}(x_i-\mu_1)^2}, F_{n_1,n_2;1-\alpha/2} \frac{n_1}{n_2} \frac{\sum_{i=1}^{n_2}(y_i-\mu_2)^2}{\sum_{i=1}^{n_1}(x_i-\mu_1)^2} \right].$$

Example [Confidence interval for $\dfrac{\sigma_2^2}{\sigma_1^2}$, if μ_1 and μ_2 are unknown]

Use that

$$\frac{\dfrac{n_1 S_1^2}{\sigma_1^2} \Big/ (n_1-1)}{\dfrac{n_2 S_2^2}{\sigma_2^2} \Big/ (n_2-1)} \sim F(n_1-1, n_2-1).$$

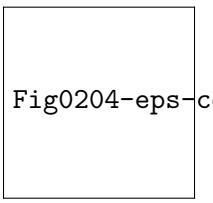
Conclusion : a $100(1-\alpha)\%$ confidence interval for the ratio σ_2^2/σ_1^2 if μ_1 and μ_2 are unknown is given by

$$\left[F_{n_1-1,n_2-1;\alpha/2} \frac{\frac{n_2}{n_2-1}s_2^2}{\frac{n_1}{n_1-1}s_1^2}, F_{n_1-1,n_2-1;1-\alpha/2} \frac{\frac{n_2}{n_2-1}s_2^2}{\frac{n_1}{n_1-1}s_1^2} \right].$$

Here :

$$\begin{aligned} F_{n_1-1,n_2-1;\alpha/2} &\equiv F^{-1}(\alpha/2) \\ F_{n_1-1,n_2-1;1-\alpha/2} &\equiv F^{-1}(1-\alpha/2) \end{aligned}$$

with F the distribution function of a $F(n_1-1, n_2-1)$ random variable.



Example [Confidence Interval for Matched Pairs]

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from bivariate normal distribution with parameters $E(X) = \mu_1, E(Y) = \mu_2, var(X) = \sigma_1^2, var(Y) = \sigma_2^2$ and correlation coefficient$(X, Y) = \rho$. Assume $\sigma_1^2,\ \sigma_2^2$ and ρ known.
Let $D_i = X_i - Y_i$ for $i = 1, 2, \ldots, n$. Then,

$$D_i \sim N(\mu_1 - \mu_2, \underbrace{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2}_{\sigma_D^2})$$

$$\overline{D} \sim N(\mu_1 - \mu_2, \sigma_D^2/n)$$

$$\frac{[\overline{D}-(\mu_1-\mu_2)]}{\sigma_D/\sqrt{n}} \sim N(0,1)$$

$$\sum_{i=1}^n (D_i - \overline{D})^2/\sigma_D^2 \sim \chi^2(n-1)$$

$$\frac{\frac{\sqrt{n}[\overline{D}-(\mu_1-\mu_2)]}{\sigma_D}}{\sqrt{\frac{\sum_{i=1}^n (D_i-\overline{X})^2}{\sigma_D^2(n-1)}}} \sim t(n-1)$$

$$\Leftrightarrow \frac{\sqrt{n(n-1)}[\overline{D}-(\mu_1-\mu_2)]}{\sqrt{\sum_{i=1}^n (D_i-\overline{D})^2}} \sim t(n-1)$$

We can use it as pivotal quantity as the distribution is free of any unknowns.

$$\Rightarrow P\left[-q \le \frac{\sqrt{n(n-1)}[\overline{D} - (\mu_1 - \mu_2)]}{\sqrt{\sum_{i=1}^n (D_i - \overline{D})^2}} \le q\right] = 1 - \alpha$$

Thus, a$100(1-\alpha)\%$ confidence interval for $(\mu_1 - \mu_2)$ is
$[\overline{D} - q\sqrt{\frac{\sum_{i=1}^n (D_i-\overline{D})^2}{n(n-1)}}, \overline{D} + q\sqrt{\frac{\sum_{i=1}^n (D_i-\overline{D})^2}{n(n-1)}}]$

3.4 Other examples of confidence intervals

Example

Let$X_1, \ldots, X_n$ be a random sample from $Un[0, \theta], \theta > 0$.
To construct a confidence interval for θ, we use

$$M_n = \max(X_1, \ldots, X_n)$$

and note that the distribution of $\dfrac{M_n}{\theta}$ does not depend on θ :

$$P\left(\frac{M_n}{\theta} \leq x\right) = \begin{cases} 0 & \dots \text{ if } x \leq 0 \\ x^n & \dots \text{ if } 0 \leq x \leq 1 \\ 1 & \dots \text{ if } x \geq 1 \end{cases}$$

Hence, for all $0 \leq a \leq b \leq 1$:

$$P\left(a \leq \frac{M_n}{\theta} \leq b\right) = b^n - a^n.$$

If

$$b^n - a^n = 1 - \alpha$$

then $P\left(\frac{M_n}{b} \leq \theta \leq \frac{M_n}{a}\right) = 1 - \alpha.$

Since we know that $\theta \geq M_n$, we choose $b = 1$.
Then $a = \alpha^{\frac{1}{n}}$ and

$$P(M_n \leq \theta \leq \alpha^{-\frac{1}{n}} M_n) = 1 - \alpha \ .$$

Conclusion : a $100(1-\alpha)\%$ confidence interval for θ in the $Un[0,\theta]$ distribution is

$$[\max(x_i), \alpha^{-\frac{1}{n}} \max(x_i)].$$

Example

Let $X_1, \dots, X_n$ be a random sample from $Exp(\lambda), \lambda > 0$.
Use characteristic functions to see that

$$2\lambda \sum_{i=1}^{n} X_i \sim \chi^2(2n) \ .$$

Hence

$$P\left(\chi^2_{2n;\alpha/2} \leq 2\lambda \sum_{i=1}^{n} X_i \leq \chi^2_{2n;1-\alpha/2}\right) = 1 - \alpha \ .$$

Conclusion : a $100(1-\alpha)\%$ confidence interval for λ in the $Exp(\lambda)$ distribution is given by

$$\left[\frac{\chi^2_{2n;\alpha/2}}{2\sum_{i=1}^{n} x_i}, \frac{\chi^2_{2n;1-\alpha/2}}{2\sum_{i=1}^{n} x_i}\right].$$

Example

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. r.v.'s from the Beta distribution with $\beta = 1$ and $\alpha = \theta$ unknown.
To construct *a* $100(1-\alpha)\%$ confidence interval we proceed as follows:
$-\sum_{i=1}^{n} lnX_i$ is a sufficient statistic for θ . Consider the transformation $Y_i = -2\theta lnX_i$. It can be easily shown that its p.d.f. is $\frac{1}{2}e^{y_i/2}$,$y_i > 0$ which is the probability density function of $\chi^2(2)$. This shows that

$$T_n = -2\theta \sum_{i=1}^{n} logX_i = \sum_{i=1}^{n} Y_i$$

is distributed as $\chi^2(2n)$, which shows that T_n is a pivotal quantity. Now find l and r $(l < r)$ such that

$$P(l \leq \chi^2(2n) \leq r) = 1 - \alpha \tag{3.1}$$

which give us

$$P(l \leq -2\theta \sum_{i=1}^{n} logX_i \leq r) = 1 - \alpha$$

which is equivalent to

$$P\left(\frac{\chi^2_{2n;\frac{\alpha}{2}}}{\sum_{i=1}^{n} Y_i} \leq \theta \leq \frac{\chi^2_{2n;1-\frac{\alpha}{2}}}{\sum_{i=1}^{n} Y_i} \right) = 1 - \alpha$$

Therefore, *a* $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\frac{\chi^2_{2n;\frac{\alpha}{2}}}{\sum_{i=1}^{n} y_i}, \frac{\chi^2_{2n;1-\frac{\alpha}{2}}}{\sum_{i=1}^{n} y_i} \right]$$

3.5 Bayesian confidence intervals

In Bayesian statistics the estimator for a parameter θ is given by the mean of the posterior distribution (in the case of squared error loss) or by a median of the posterior distribution (in the case of absolute error loss).

In the same spirit we can construct a $100(1-\alpha)\%$ **Bayesian confidence interval for** θ by finding two functions

$$l(x_1, \ldots, x_n) \qquad \text{and} \qquad r(x_1, \ldots, x_n)$$

such that the posterior probability that $\widetilde{\Theta}$ falls in the interval $[l(x_1, \ldots, x_n), r(x_1, \ldots, x_n)]$ equals $1 - \alpha$ (or is at least $1 - \alpha$) :

$$P(l(X_1, \ldots, X_n) \leq \widetilde{\Theta} \leq r(X_1, \ldots, X_n) | X_1 = x_1, \ldots, X_n = x_n) = 1 - \alpha$$

i.e.

$$\sum_{l(x_1,\ldots,x_n)\leq\theta\leq r(x_1,\ldots,x_n)} P(\widetilde{\Theta}=\theta|X_1=x_1,\ldots,X_n=x_n)=1-\alpha$$

in the discrete case

or

$$\int_{l(x_1,\ldots,x_n)}^{r(x_1,\ldots,x_n)} f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n)d\theta=1-\alpha$$

in the continuous case

Example

Let $X_1,\ldots,X_n$ be a random sample from $X\sim N(\theta;\sigma^2)$ with σ^2 known, $\theta\in\mathbb{R}$.
As a prior density, we take $\widetilde{\Theta}\sim N(\mu_0;\sigma_0^2)$ with μ_0 and σ_0^2 known.
For squared error loss, we obtained before that the posterior density is

$$N\left(\frac{\sigma^2\mu_0+\sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}\,;\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)\,.$$

Conclusion : a $100(1-\alpha)\%$ Bayesian confidence interval for θ is given by

$$\frac{\sigma^2\mu_0+\sigma_0^2 n\overline{x}}{\sigma^2+\sigma_0^2 n}\pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}}.$$

Example
Suppose that $X=(X_1,\ldots,Xn)$ is a random sample from the Bernoulli distribution with success parameter p. Moreover, suppose that p has a prior beta distribution with left parameter $a>0$ and right parameter $b>0$. Denote the number of successes by

$$Y=\sum_{i=1}^{n}X_i$$

Recall that for a given value of p, Y has the binomial distribution with parameters n and p.
Given $Y=y$, the posterior distribution of p is beta with left parameter $a+y$ and right parameter $b+(n-y)$.
A $(1-\alpha)$ level Bayesian confidence interval for p is $(l(y),r(y))$, where $l(y)$ is the quantile of order $\alpha/2$ and $r(y)$ is the quantile of order $1-\alpha/2$ for the beta distribution (posterior distribution of p).

Example
Suppose that $X = (X_1, \ldots, X_n)$ is a random sample from $Poisson(\theta)$. Moreover, suppose that θ has a prior $\Gamma(\alpha; \beta)$. The postirior distribution is given by

$$f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) \sim \Gamma[(n+1/\beta)^{-1}, \sum x_i + \alpha]$$

It follows that

$$2(n+1/\beta) f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) \sim \chi^2[2(\sum x_i + \alpha)]$$

and

$$P[\chi^2_{v;\alpha/2} < 2(n+1/\beta) f_{\widetilde{\Theta}|X_1,\ldots,X_n}(\theta|x_1,\ldots,x_n) < \chi^2_{v;1-\alpha/2}] = 1-\alpha$$

where $v = 2(\sum x_i + \alpha)$.
Thus, a $100(1-\alpha)\%$ Baysesian confidence interval for θ is given by

$$\left(\frac{\chi^2_{v;\alpha/2}}{2(n+1/\beta)}, \frac{\chi^2_{v;1-\alpha/2}}{2(n+1/\beta)} \right)$$

3.6 Confidence regions in higher dimensions

The notion of confidence intervals can be extended to confidence regions for a general k-dimensional parameter $\widetilde{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subset I\!R^k$.
The k-dimensional rectangle

$$\{(\theta_1, \ldots, \theta_k) | l_j(x_1, \ldots, x_n) \leq \theta_j \leq r_j(x_1, \ldots, x_n); j = 1, \ldots, k\}$$

is called a $100(1-\alpha)\%$ **confidence rectangle** for $\underset{\sim}{\theta}$ if

$$P(l_j(X_1, \ldots, X_n) \leq \theta_j \leq r_j(X_1, \ldots, X_n); j = 1, \ldots, k) = 1-\alpha \ .$$

Sometimes, multidimensional confidence rectangles can be obtained from **one** dimensional confidence intervals.
Suppose we have confidence intervals for the individual components of $\underset{\sim}{\theta}$: i.e. for $j = 1, \ldots, k$ with

$$L_{jn} = l_j(X_1, \ldots, X_n) \ , \ R_{jn} = r_j(X_1, \ldots, X_n)$$

we have

$$P(L_{jn} \leq \theta_j \leq R_{jn}) = 1 - \alpha_j \ , \text{ say.}$$

If the pairs $(L_{jn}, R_{jn}), j = 1, \ldots, k$ are **independent**, then for the rectangle

$$[L_{1n}, R_{1n}] \times [L_{2n}, R_{2n}] \times \ldots \times [L_{kn}, R_{kn}]$$

we have

$$P(L_{jn} \leq \theta_j \leq R_{jn} \ ; \ j = 1, \ldots, k) = \prod_{j=1}^{k} (1 - \alpha_j).$$

If there is **no independence**, then by **Bonferroni's inequality** ($P(\bigcap_{j=1}^{k} A_j) \geq 1 - \sum_{j=1}^{k} P(A_j^c)$) we only have $P(L_{jn} \leq \theta_j \leq R_{jn}; j = 1, \ldots, k) \geq 1 - \sum_{j=1}^{k} \alpha_j$.
Hence, if $\alpha_j = \frac{\alpha}{k}$ for all $j = 1, \ldots, k$, then

$$P(L_{jn} \leq \theta_j \leq R_{jn} \ ; \ j = 1, \ldots, k) \geq 1 - \alpha.$$

Example

Let $X_1, \ldots, X_n$ be a random sample from $N(\mu; \sigma^2)$.
To set up a $100(1-\alpha)\%$ confidence rectangle for the two-dimensional parameter $\underset{\sim}{\theta} = (\mu, \sigma^2)$, we can use (see before) :

- $P\left(\overline{X} - t_{n-1;1-\alpha/4}\sqrt{\frac{S^2}{n-1}} \leq \mu \leq \overline{X} + t_{n-1;1-\alpha/4}\sqrt{\frac{S^2}{n-1}}\right) = 1 - \frac{\alpha}{2}$
- $P\left(\frac{nS^2}{\chi^2_{n-1;1-\alpha/4}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{n-1;\alpha/4}}\right) = 1 - \frac{\alpha}{2}$.

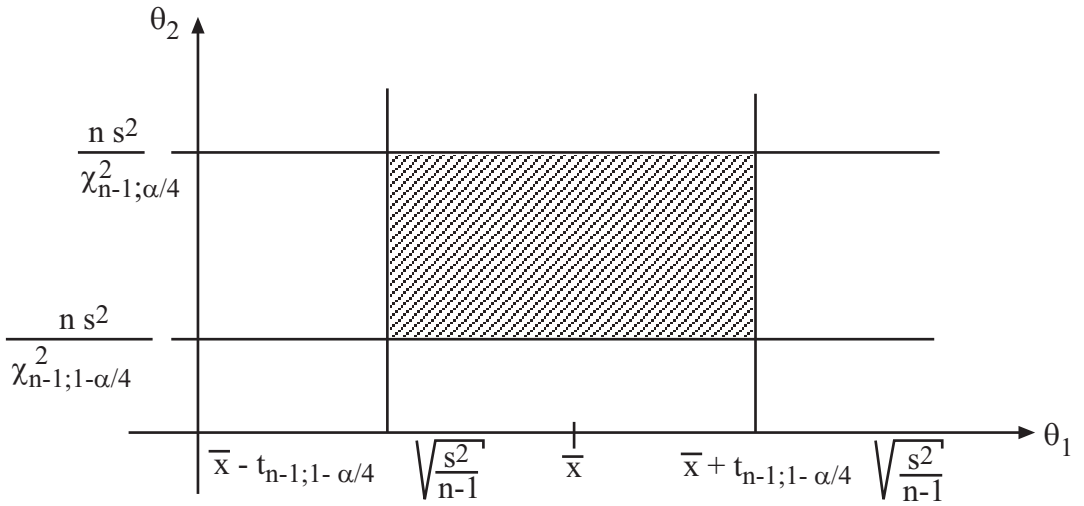
For the resulting rectangle, we can only say

$$P\left(\overline{X} - t_{n-1;1-\alpha/4}\sqrt{\frac{S^2}{n-1}} \leq \mu \leq \overline{X} + t_{n-1;1-\alpha/4}\sqrt{\frac{S^2}{n-1}}, \frac{nS^2}{\chi^2_{n-1;1-\alpha/4}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{n-1;\alpha/4}}\right)$$

$$\geq 1 - \alpha$$

since the two events are not independent.
This rectangular confidence region for $\underset{\sim}{\theta} = (\mu, \sigma^2)$ is

$$\left\{(\theta_1, \theta_2) | \overline{x} - t_{n-1;1-\alpha/4}\sqrt{\frac{s^2}{n-1}} \leq \theta_1 \leq \overline{x} + t_{n-1;1-\alpha/4}\sqrt{\frac{s^2}{n-1}}, \right.$$
$$\left. \frac{ns^2}{\chi^2_{n-1;1-\alpha/4}} \leq \theta_2 \leq \frac{ns^2}{\chi^2_{n-1;\alpha/4}}\right\}$$



A confidence region which is not rectangular can be obtained, using the independence of $\overline{X}$ and S^2.
Indeed, since

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0;1) \quad \text{and} \quad \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

we can determine constants $a > 0, 0 < b < c$ such that

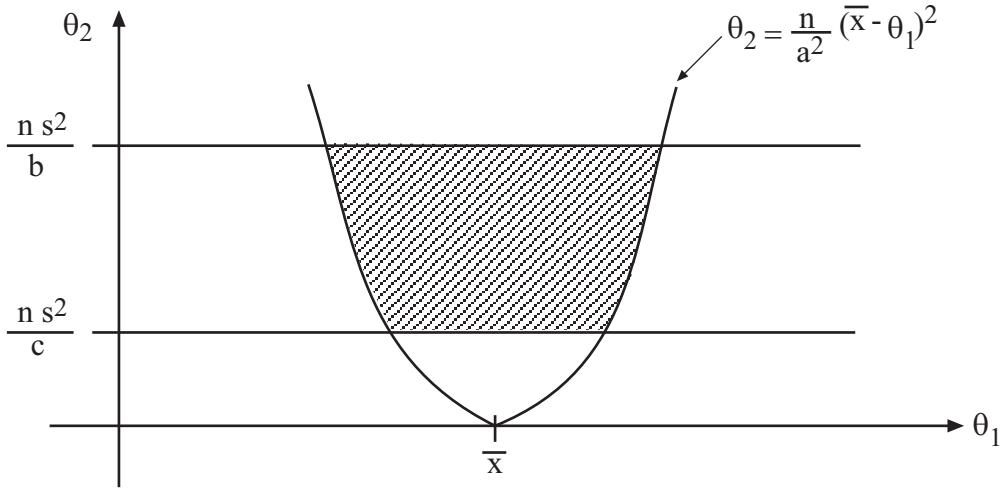
$$P\left(-a \leq \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq a\right) = \sqrt{1-\alpha} \text{ and } P\left(b \leq \frac{nS^2}{\sigma^2} \leq c\right) = \sqrt{1-\alpha}\ .$$

We then have, using independence of $\overline{X}$ and S^2 :

$$P\left(-a \leq \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq a, b \leq \frac{nS^2}{\sigma^2} \leq c\right) = \sqrt{1-\alpha}.\sqrt{1-\alpha} = 1-\alpha.$$

The $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\theta} = (\mu, \sigma^2)$ is :

$$\left\{(\theta_1,\theta_2)|(\overline{x}-\theta_1)^2 \leq \frac{a^2\theta_2^2}{n}\ ,\ \frac{ns^2}{c} \leq \theta_2 \leq \frac{ns^2}{b}\right\} :$$



3.7 Approximate confidence intervals

In all the examples considered up to now (except the Behrens-Fisher problem) the construction of a confidence interval followed from the fact that the distribution of some random variable was **exactly** known (standard normal, $t, \chi^2, F, \ldots$). The use of the large sample **limiting distribution** (as $n \to \infty$) leads to **approximate** $100(1-\alpha)\%$ **confidence intervals.**

We give some examples.

Example [Confidence interval for the mean if the variance is known]

$X_1, \ldots, X_n$: random sample from X with $E(X) = \mu$ and $Var(X) = \sigma^2$ with σ^2 known. Use the central limit theorem :

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} N(0,1) \qquad , n \to \infty$$

and proceed as before.

Conclusion : an approximate $100(1-\alpha)\%$ confidence interval for μ if σ^2 is known is given by

$$\left[\overline{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}, \overline{x} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \right].$$

Example [Confidence interval for the mean if the variance is unknown]

$X_1, \ldots, X_n$: random sample from X with $E(X) = \mu$ and $Var(X) = \sigma^2$.
Because of the central limit theorem in the foregoing example and the fact that $S^2 \xrightarrow{P} \sigma^2$, we have by Slutsky's theorem :

$$\frac{\overline{X} - \mu}{\sqrt{\frac{S^2}{n-1}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty.$$

From this we obtain :

Conclusion : an approximate $100(1-\alpha)\%$ confidence interval for μ if σ^2 is unknown is

$$\left[\overline{x} - z_{1-\alpha/2}\sqrt{\frac{s^2}{n-1}}, \overline{x} + z_{1-\alpha/2}\sqrt{\frac{s^2}{n-1}} \right].$$

Another useful tool in the construction of approximate confidence intervals is the asymptotic normality result of the **maximum likelihood estimator :** for a large sample from X, with sufficiently regular density $f(x;\theta)$, we have that the ML-estimator T_n for θ satisfies

$$\frac{T_n - \theta}{\sqrt{\frac{1}{ni(\theta)}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty$$

where $i(\theta)$ is the Fisher information number.

Example

Let $X_1, \ldots, X_n$be a random sample from $N(0;\sigma^2)$.
Put $\theta = \sigma^2$.
The ML-estimator for θ is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ and $i(\theta) = \frac{1}{2\theta^2}$.
Hence

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \theta}{\sqrt{\frac{2\theta^2}{n}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty.$$

Conclusion : an approximate $100(1-\alpha)\%$ confidence interval for σ^2 in $N(0,\sigma^2)$ is

$$\left[\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{1 + z_{1-\alpha/2}\sqrt{\frac{2}{n}}} \; , \; \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{1 - z_{1-\alpha/2}\sqrt{\frac{2}{n}}} \right].$$

Example [Confidence interval for a proportion]

Let $X_1, \ldots, X_n$ be a random sample from $B(1;\theta)$, where $\theta \in [0,1]$.
The ML-estimator for θ is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $i(\theta) = \frac{1}{\theta(1-\theta)}$.
Hence :

$$\frac{\overline{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty \ .$$

Hence :

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{X} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \leq z_{1-\alpha/2} \right) \approx 1 - \alpha$$

or

$$P\left((\overline{X} - \theta)^2 \leq z^2_{1-\alpha/2} \frac{\theta(1-\theta)}{n} \right) \approx 1 - \alpha$$

or

$$P\left((1 + \ \tfrac{1}{n}\ z^2_{1-\alpha/2})\theta^2 - (2\overline{X} + \ \tfrac{1}{n}\ z^2_{1-\alpha/2})\theta + \overline{X}^2 \leq 0 \right) \approx 1 - \alpha$$

For fixed $\overline{X}$ $(0 \leq \overline{X} \leq 1)$

$$(1 + \ \tfrac{1}{n}\ z^2_{1-\alpha/2})\theta^2 - (2\overline{X} + \ \tfrac{1}{n}\ z^2_{1-\alpha/2})\theta + \overline{X}^2$$

is a quadratic polynomial in θ with 2 real roots. Hence the above is equivalent to : (with $z \equiv z_{1-\alpha/2}$) :

$$P\left(\frac{n\overline{X} + \frac{z^2}{2} - z\sqrt{n\overline{X}(1-\overline{X}) + \frac{z^2}{4}}}{n + z^2} \leq \theta \leq \frac{n\overline{X} + \frac{z^2}{2} + z\sqrt{n\overline{X}(1-\overline{X}) + \frac{z^2}{4}}}{n + z^2} \right) \approx 1 - \alpha.$$

Conclusion : an approximate $100(1-\alpha)\%$ confidence interval for the probability of success θ in $B(1;\theta)$ is

$$\left[\frac{y_n + \frac{z^2}{2} - z\sqrt{\frac{y_n(n-y_n)}{n} + \frac{z^2}{4}}}{n + z^2} \ , \ \frac{y_n + \frac{z^2}{2} + z\sqrt{\frac{y_n(n-y_n)}{n} + \frac{z^2}{4}}}{n + z^2} \right].$$

where $y_n = n\overline{x}$ is the number of successes in n trials and $z = z_{1-\alpha/2}$.

Example

Let $X_1, \ldots, X_n$be a random sample from Poisson (θ), with $\theta > 0$.
The ML-estimator for θ is $\overline{X}$ and $i(\theta) = \frac{1}{\theta}$.
Hence,

$$\frac{\overline{X} - \theta}{\sqrt{\frac{\theta}{n}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty.$$

We obtain, with $z = z_{1-\alpha/2}$:

$$P\left(-z \leq \frac{\overline{X} - \theta}{\sqrt{\frac{\theta}{n}}} \leq z \right) \approx 1 - \alpha$$

or

$$P\left((\overline{X} - \theta)^2 \leq z^2 \frac{\theta}{n} \right) \approx 1 - \alpha.$$

This leads to :

Conclusion : an approximate $100(1-\alpha)\%$ confidence interval for θ in a Poisson (θ) distribution is

$$\left[\overline{x} + \frac{z^2}{2n} - \sqrt{\frac{\overline{x}z^2}{n} + \frac{z^4}{4n^2}} \ , \ \overline{x} + \frac{z^2}{2n} + \sqrt{\frac{\overline{x}z^2}{n} + \frac{z^4}{4n^2}} \right]$$

where $z = z_{1-\alpha/2}$.

The computations needed in the last two examples can be avoided (but lead to a less accurate approximate confidence interval) replacing the asymptotic variance $\frac{1}{ni(\theta)}$ of the ML-estimator by the estimator

$$\frac{1}{ni(T_n)}.$$

We then construct an approximate confidence interval from the fact that, in most cases :

$$\frac{T_n - \theta}{\sqrt{\frac{1}{ni(T_n)}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty.$$

Example [Confidence interval for a proportion]

Let $X_1, \ldots, X_n$ be a random sample from $B(1;\theta)$ with $\theta \in [0,1]$.
If we use that

$$\frac{\overline{X}-\theta}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \xrightarrow{d} N(0;1) \qquad , \; n \to \infty$$

then we obtain the approximate $100(1-\alpha)\%$ confidence interval for θ :

$$\left[\overline{x} - z\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} \; , \; \overline{x} + z\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} \right] .$$

Note Since $X \sim$ Bernoulli :

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

this is also a particular case of the second example in this section.

Example

Let $X_1, \ldots, X_n$ be a random sample from Poisson (θ).
Using

$$\frac{\overline{X}-\theta}{\sqrt{\frac{\overline{X}}{n}}} \xrightarrow{d} N(0;1) \qquad , n \to \infty$$

leads to the approximate $100(1-\alpha)\%$ confidence interval for θ :

$$\left[\overline{x} - z\sqrt{\frac{\overline{x}}{n}} \quad , \quad \overline{x} + z\sqrt{\frac{\overline{x}}{n}} \right]$$

Example

$X_1, \ldots, X_n$: random sample from $X \sim N(0;\sigma^2)$.
Put $\theta = \sigma^2$.
Using

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \theta}{\sqrt{\frac{2}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \theta}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)\sqrt{\frac{2}{n}}} \xrightarrow{d} N(0;1)$$

we obtain as an approximate $100(1-\alpha)\%$ confidence interval for θ :

$$\left[\left(1 - z\sqrt{\frac{2}{n}}\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) \ , \ \left(1 + z\sqrt{\frac{2}{n}}\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)\right] \ .$$

The approximate confidence intervals obtained from the asymptotic normality result of the ML-estimator are **not invariant under transformations of the parameter**.

Example

Let $X_1, \ldots, X_n$ be a random sample from $N(0;\sigma^2)$.
Put $\theta = \sigma$.
The ML-estimator for θ is $\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{1/2}$ and $i(\theta) = \frac{2}{\theta^2}$.
This leads to an approximate $100(1-\alpha)\%$ confidence interval for θ :

$$\left[\frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{1/2}}{1 + z\sqrt{\frac{1}{2n}}} \ , \ \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)^{1/2}}{1 - z\sqrt{\frac{1}{2n}}}\right] \ .$$

Since $\theta > 0$, we could obtain an approximate $100(1-\alpha)\%$ confidence interval for σ^2 by squaring. This would give

$$\left[\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\left(1 + z\sqrt{\frac{1}{2n}}\right)^2} \ , \ \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\left(1 - z\sqrt{\frac{1}{2n}}\right)^2}\right]$$

but, this is not the same as what we found before. Indeed :

$$\left(1 \pm z\sqrt{\frac{1}{2n}}\right)^2 = 1 \pm z\sqrt{\frac{2}{n}} + \frac{z^2}{2n}.$$

A method that produces approximate confidence intervals **invariant** under transformations of the parameter can be deduced from the following fact (see chapter 1) :

$$\frac{S(\theta; \underset{\sim}{X})}{\sqrt{ni(\theta)}} \xrightarrow{d} N(0;1) \qquad , n \to \infty$$

(under regularity conditions on $f(x;\theta)$).

Here

$$S(\theta; \underset{\sim}{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} lnf(X_i;\theta)$$

is the **score statistic** and $i(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} lnf(X;\theta)\right]$.

Let ϕ be a strictly increasing function of θ and let $\phi(\theta) = \theta^*$.

Then

$$\frac{\partial}{\partial\theta} lnf(X;\theta) = \frac{\partial}{\partial\theta^*} lnf(X;\theta).\frac{\partial\phi}{\partial\theta} \ .$$

Hence :

$$E\left[\frac{\partial}{\partial\theta} lnf(X;\theta)\right] = 0 \Rightarrow E\left[\frac{\partial}{\partial\theta^*} lnf(X;\theta)\right] = 0 \ .$$

Also

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2} lnf(X;\theta) &= \frac{\partial^2}{\partial\theta^{*2}} lnf(X;\theta) \left(\frac{\partial\phi}{\partial\theta}\right)^2 \\ &\quad + \frac{\partial}{\partial\theta^*} lnf(X;\theta).\frac{\partial^2\phi}{\partial\theta^2} \ . \end{aligned}$$

Hence :

$$E\left[\frac{\partial^2}{\partial\theta^2} lnf(X;\theta)\right] = \left(\frac{\partial\phi}{\partial\theta}\right)^2 E\left[\frac{\partial^2}{\partial\theta^{*2}} lnf(X;\theta)\right]$$

or

$$i(\theta) = \left(\frac{\partial\phi}{\partial\theta}\right)^2 i(\theta^*) \ .$$

Hence :

$$\frac{S(\theta^*; X)}{\sqrt{ni(\theta^*)}} = \frac{S(\theta; X)}{\sqrt{ni(\theta)}} \ .$$

Example

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(0; \sigma^2)$.
Put $\theta = \sigma$.

Then
$S(\theta; \underset{\sim}{X}) = -\frac{n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} X_i^2$
$i(\theta) = \frac{2}{\theta^2}$

Hence :

$$\frac{-\frac{n}{\theta} + \frac{1}{\theta^3} \sum_{i=1}^{n} X_i^2}{\sqrt{\frac{2n}{\theta^2}}} = \frac{\frac{1}{\theta^2} \sum_{i=1}^{n} X_i^2 - n}{\sqrt{2n}} \xrightarrow{d} N(0; 1)$$

and this produces an approximate $100(1-\alpha)\%$ confidence interval for θ :

$$\left[\sqrt{\frac{\sum_{i}^{n} x_i^2}{n + z\sqrt{2n}}} \ , \ \sqrt{\frac{\sum_{i}^{n} x_i^2}{n - z\sqrt{2n}}} \right] \ .$$

Notes

- This is not the same as in the previous example. But for large n, the difference in negligible.
- If we would have taken σ^2 as the parameter, then this procedure would have given

$$\left[\frac{\sum_{i=1}^{n} x_i^2}{n + z\sqrt{2n}} \ , \ \frac{\sum_{i=1}^{n} x_i^2}{n - z\sqrt{2n}} \right]$$

  and these endpoints are the squares of the above.

For the case of a **multidimensional parameter**, large sample approximate confidence regions can be obtained from the fact that (under regularity conditions) the ML-estimator $\underset{\sim}{T}_n = (T_{n1}, \ldots, T_{nk})$ is asymptoticially normal, with mean $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$ and variance-covariance matrix

$$V = \frac{1}{n} B^{-1}(\underset{\sim}{T}_n)$$

where $B(\underset{\sim}{\theta})$ is the Fisher information matrix.
It follows that

$$(\underset{\sim}{T}_n - \underset{\sim}{\theta}) V^{-1} (\underset{\sim}{T}_n - \underset{\sim}{\theta})'$$

is approximately $\chi^2(k)$ distributed.
Hence, we can find a number c_α, such that for all $\underset{\sim}{\theta}$:

$$P_{\underset{\sim}{\theta}}((\underset{\sim}{T}_n - \underset{\sim}{\theta}) V^{-1} (\underset{\sim}{T}_n - \underset{\sim}{\theta})' \leq c_\alpha) \approx 1 - \alpha.$$

From this, we obtain an approximate $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\theta}$ (a k-dimensional **confidence ellipsoid**).

3.8 Sample size determination

The question of how large the sample size should be to achieve a given accuracy is a very practical one. The answer is not easy. The problem is related to confidence interval estimation. We consider some examples.

3.8.1 Estimation of the mean of a normal population

Let $X_1, \ldots, X_n$ be a random sample from $X \sim N(\mu; \sigma^2)$. Suppose we want a $100(1-\alpha)\%$ confidence interval for μ of **length at most** $2d$, where d is some prescribed number.

- If σ^2 is known, then the length of a $100(1-\alpha)\%$ confidence interval for μ is given by

  $$2z\sqrt{\frac{\sigma^2}{n}}$$

  where $z = z_{1-\alpha/2}$.
  Hence, the width will be $\leq 2d$ if we choose the sample size n as the (smallest) integer satisfying

  $$n \geq \frac{\sigma^2}{d^2} z^2 \ .$$

- If σ^2 is unknown, but if from previous experience some upper bound σ_1^2 is known we can use : $n \geq \dfrac{\sigma_1^2}{d^2} z^2$.

- If σ^2 is unknown and no upper bound is available, then the length of a $100(1-\alpha)\%$ confidence interval is random

$$2t_{n-1;1-\alpha/2}\sqrt{\frac{S^2}{n-1}}$$

and may be arbitrary large.
A way out to achieve a length of at most $2d$ is the following **sequential** procedure : the **two-stage sampling procedure of C.Stein :**

1. Take a first sample of fixed size $n_0 \geq 2$, and compute the sample mean and the sample variance :

$$\begin{aligned} \overline{X}_0 &= \frac{1}{n_0}\sum_{i=1}^{n_0} X_i \\ S_0^2 &= \frac{1}{n_0}\sum_{i=1}^{n_0}(X_i - \overline{X}_0)^2. \end{aligned}$$

2. Take $N - n_0$ further observations where N is the smallest integer satisfying

$$N \geq n_0 + 1$$

and

$$N \geq \frac{\frac{n_0}{n_0-1}S_0^2}{d^2}t^2_{n_0-1;1-\alpha/2}$$

and use as a confidence interval

$$\left[\overline{X}_N - t_{n_0-1;1-\alpha/2}\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}\ ,\ \overline{X}_N + t_{n_0-1;1-\alpha/2}\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}\right]$$

where

$$\begin{aligned} \overline{X}_N &= \frac{1}{N}\sum_{i=1}^{N} X_i \\ &= \frac{\sum_{i=1}^{n_0} X_i + \sum_{i=n_0+1}^{N} X_i}{N} = \frac{n_0}{N}\overline{X}_0 + \frac{1}{N}\sum_{i=n_0+1}^{N} X_i\ . \end{aligned}$$

The length of this confidence interval equals

$$2t_{n_0-1;1-\alpha/2}\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}$$

and this is $\leq 2d$, by the choice of N.
That

$$P\left(\overline{X}_N - t_{n_0-1;1-\alpha/2}\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}} \leq \mu \leq \overline{X}_N + t_{n_0-1;1-\alpha/2}\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}\right) = 1-\alpha$$

follows from the fact that

$$\frac{\overline{X}_N - \mu}{\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}} \sim t(n_0 - 1)$$

(**Note** : N is a random variable)

Proof

$$\begin{aligned} P\left(\frac{\overline{X}_N - \mu}{\sqrt{\frac{\frac{n_0}{n_0-1}S_0^2}{N}}} \leq x\right) &= P\left(\frac{\frac{\overline{X}_N - \mu}{\sqrt{\frac{\sigma^2}{N}}}}{\sqrt{\frac{\frac{n_0 S_0^2}{\sigma^2}}{n_0-1}}} \leq x\right) \\ &= \sum_k P\left(\frac{\frac{\overline{X}_k - \mu}{\sqrt{\frac{\sigma^2}{k}}}}{\sqrt{\frac{\frac{n_0 S_0^2}{\sigma^2}}{n_0-1}}} \leq x \ , \ N = k\right) \ . \end{aligned}$$

Since for $k \geq n_0 + 1$:

$$\overline{X}_k = \frac{n_0}{k}\overline{X}_0 + \frac{1}{k}\sum_{i=n_0+1}^{k} X_i \ .$$

Since X is normal, $\overline{X}_0$ and S_0^2 are independent.
It follows that $\overline{X}_k$ and S_0^2 are independent.
Hence the above equals :

$$\begin{aligned} &= \sum_k P(T \leq x \ , \ N = k) \text{ with } T \sim t(n_0 - 1) \\ &= P(T \leq x). \end{aligned}$$

□

Note: In case that the one-sided $1 - \alpha$ confidence interval is required,then d is specified as the absolute value of the difference between mean ,μ, and the upper or lower limit ,i.e.,

$$|\overline{X} - (\overline{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}| = d$$

Then,

$$Z_\alpha \frac{\sigma}{\sqrt{n}} = \delta$$

which yields

$$n = (Z_\alpha \sigma / \delta)^2$$

.

3.8.2 Estimation of a proportion

Let $X_1, \ldots, X_n$ be a random sample from $X \sim B(1;\theta)$. Suppose we want to determine the sample size needed to obtain an approximate $100(1-\alpha)\%$ confidence interval for θ of length at most $2d$.

- If we use the approximate $100(1-\alpha)\%$ confidence interval

$$\left[\overline{x} - z\sqrt{\frac{\overline{x}(1-\overline{x})}{n}},\ \overline{x} + z\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\ \right]$$

or with $y_n = n\overline{x}$ = the number of successes in n trials :

$$\left[\frac{y_n}{n} - z\sqrt{\frac{\frac{y_n}{n}(1-\frac{y_n}{n})}{n}},\ \frac{y_n}{n} + z\sqrt{\frac{\frac{y_n}{n}(1-\frac{y_n}{n})}{n}}\ \right]$$

we need to have

$$z\sqrt{\frac{\frac{y_n}{n}(1-\frac{y_n}{n})}{n}} \leq d\ .$$

If we use that

$$\frac{y_n}{n}\left(1-\frac{y_n}{n}\right) = \frac{1}{4} - \left(\frac{y_n}{n} - \frac{1}{2}\right)^2 \leq \frac{1}{4}$$

then we obtain

$$z\sqrt{\frac{1}{4n}} \leq d$$

or

$$n \geq \frac{z^2}{(2d)^2}$$

(For $\alpha = 0.05 : z^2 = (1.96)^2 \cong 4$ one sometimes uses $n \approx \frac{1}{d^2}$)

- A similar formula holds if we use the approximate $100(1-\alpha)\%$ confidence interval

$$\frac{y_n + \frac{z^2}{2} \pm z\sqrt{\frac{y_n(n-y_n)}{n} + \frac{z^2}{4}}}{n+z^2}\ .$$

To obtain a length of at most $2d$, we need to have

$$\frac{z}{n+z^2}\sqrt{\frac{y_n(n-y_n)}{n} + \frac{z^2}{4}} \leq d$$

or, again using that $\frac{y_n(n-y_n)}{n} \leq \frac{n}{4}$,

$$z\sqrt{\frac{1}{4(n+z^2)}} \leq d$$

or,

$$n \geq \frac{z^2}{(2d)^2} - z^2$$

- The obtained formulas are crude since they rely on the inequality

$$\theta(1-\theta) \leq \frac{1}{4} \qquad (0 \leq \theta \leq 1)$$

which is only good near $\theta = \frac{1}{2}$.

It is clear that we can do better if we know a priori that $\theta \leq \theta_0 < \frac{1}{2}$ or $\theta \geq \theta_1 > \frac{1}{2}$.

3.8.3 Sampling from a finite population

The lower bounds for the required sample size may be very high. This can be a problem if the size of the population is small.
A way out can be to use the procedure of **sampling without replacement**.

Suppose we have a finite population of size N :

$$\{x_1, x_2, \ldots, x_N\}$$

Denote the sample of size n by

$$X_1, \ldots, X_n$$

where X_i denotes the i-th object sampled.
In sampling without replacement, we have

$$P(X_1 = x_1', X_2 = x_2', \ldots, X_n = x_n') = \frac{1}{N}\frac{1}{N-1}\cdots\frac{1}{N-n+1}$$

for all $\{x_1', \ldots, x_n'\} \subset \{x_1, \ldots, x_N\}$ $(n \leq N)$.
The marginal distribution of each of the X_i $(i = 1, \ldots, n)$ is uniform over $\{x_1, \ldots, x_n\}$:

$$P(X_i = x) = \begin{cases} \frac{1}{N} & \ldots \text{ if } x = x_1, \ldots, x_N \\ 0 & \ldots \text{ if otherwise.} \end{cases}$$

Let us examine some properties of the sample mean $\overline{X}$ and the sample variance S^2 as estimators for the population mean μ and the population variance σ^2.
The population mean and population variance are now

$$\begin{aligned} \mu &= \frac{1}{N}\sum_{i=1}^{N} x_i \\ \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 \ . \end{aligned}$$

We have, for each $i, j = 1, \ldots, n$ $(j \neq i)$:

- $E(X_i) = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$

- $Var(X_i) = E(X_i^2) - (E(X_i))^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2 = \sigma^2$

- $$\begin{aligned} Cov(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} x_i x_j P(X_i = x_i, X_j = x_j) - \mu^2 \\ &= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} x_i x_j - \mu^2 \\ &= \frac{1}{N}\frac{1}{N-1}\sum_{i=1}^{N} x_i \left(\sum_{j=1}^{N} x_j - x_i\right) - \mu^2 \\ &= \frac{1}{N}\frac{1}{N-1}\left[\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right] - \mu^2 \\ &= \frac{1}{N}\frac{1}{N-1}\left[(N\mu)^2 - \sum_{i=1}^{N} x_i^2\right] - \mu^2 \\ &= -\frac{1}{N-1}\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] \\ &= -\frac{\sigma^2}{N-1} \end{aligned}$$

Hence

- $E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$
- $$\begin{aligned} Var(\overline{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} Cov(X_i, X_j) \\ &= \frac{1}{n^2} n\sigma^2 - \frac{n(n-1)}{n^2}\frac{\sigma^2}{N-1} \\ &= \frac{\sigma^2}{n}\left(1 - \frac{n-1}{N-1}\right) \\ &= \frac{\sigma^2}{n}\frac{N-n}{N-1} < \frac{\sigma^2}{n} \end{aligned}$$

  Hence : $\overline{X}$ is still unbiased but with smaller variance.

  The fraction $\frac{N-n}{N-1}$ is called the **finite population correction factor**. This factor becomes negligible if N is large and $\frac{n}{N}$ is small (say, n is less than 5% of N).
  Indeed :
  $$\frac{N-n}{N-1} = \frac{1-\frac{n}{N}}{1-\frac{1}{N}} \approx 1 \ .$$
  The quantity $\frac{n}{N}$ is called the **sampling fraction**.
- $$\begin{aligned} E(S^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i^2) - E(\overline{X}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} [Var(X_i) + (E(X_i))^2] - [Var(\overline{X}) + (E(\overline{X}))^2] \\ &= \frac{1}{n}\sum_{i=1}^{n} (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n}\frac{N-n}{N-1} + \mu^2\right) \\ &= \sigma^2\left(1 - \frac{1}{n}\frac{N-n}{N-1}\right) = \sigma^2\frac{n-1}{n}\frac{N}{N-1} \end{aligned}$$

  Hence : $\frac{N-1}{N}\frac{n}{n-1}S^2$ is unbiased for σ^2.

To find an approximate confidence interval for the population mean, we use that

$$\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}} \quad \text{is approximately } N(0;1).$$

For the case of **proportions** :

Suppose N is the size of the population and that N_1 of them have a certain property S (and $N-N_1$ do not have this property). We want to estimate

$$\theta = \frac{N_1}{N}$$

If we take a sample of size n without replacement, and denote

$$X_i = \begin{cases} 1 & \dots \text{ if the i-th object sampled has property } S \\ 0 & \dots \text{ if otherwise} \end{cases}$$

then

$$Y_n = \sum_{i=1}^{n} X_i = \text{the number of observations with property} S.$$

Since $Y_n = n\overline{X}$ and since

$$\begin{aligned} \mu &= \frac{N_1}{N} = \theta \\ \sigma^2 &= \mu - \mu^2 = \mu(1-\mu) = \frac{N_1}{N}\left(1-\frac{N_1}{N}\right) = \theta(1-\theta) \end{aligned}$$

we have from the above

$$\frac{\frac{Y_n}{n}-\theta}{\sqrt{\frac{\theta(1-\theta)}{n}.\frac{N-n}{N-1}}} \quad \text{is approximately } N(0;1)$$

[In fact the **exact** distribution of Y_n is hypergeometric with parameters N_1, N and n :

$$P(Y_n = x) = \begin{cases} \begin{pmatrix} N_1 \\ x \end{pmatrix}\begin{pmatrix} N-N_1 \\ n-x \end{pmatrix} \Big/ \begin{pmatrix} N \\ n \end{pmatrix} & \dots \text{ if } x = 0,1,\dots,n \\ 0 & \dots \text{ if otherwise . }] \end{cases}$$

Thus :

$$P\left(\frac{Y_n}{n} - z\sqrt{\frac{\theta(1-\theta)}{n}.\frac{N-n}{N-1}} \le \theta \le \frac{Y_n}{n} + z\sqrt{\frac{\theta(1-\theta)}{n}.\frac{N-n}{N-1}}\right) \approx 1-\alpha \ .$$

If we replace θ under the square root sign by $\dfrac{Y_n}{n}$, we obtain :

an approximate $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\frac{y_n}{n} - z\sqrt{\frac{\frac{y_n}{n}\left(1-\frac{y_n}{n}\right)}{n}.\frac{N-n}{N-1}},\ \frac{y_n}{n} + z\sqrt{\frac{\frac{y_n}{n}\left(1-\frac{y_n}{n}\right)}{n}.\frac{N-n}{N-1}} \right] .$$

(If N is large and n/N is small then $\dfrac{N-n}{N-1} \approx 1$ and this interval is like before)

To achieve a length of at most $2d$, we need to have

$$z\sqrt{\frac{\frac{y_n}{n}\left(1-\frac{y_n}{n}\right)}{n}.\frac{N-n}{N-1}} \leq d \ .$$

Using that $\dfrac{y_n}{n}\left(1-\dfrac{y_n}{n}\right) \leq \dfrac{1}{4}$, this gives

$$z\sqrt{\frac{1}{4n}.\frac{N-n}{N-1}} \leq d$$

or, solving for n :

$$n \geq \frac{N}{1+(N-1)\left(\dfrac{2d}{z}\right)^2} \ .$$

For $\alpha = 0.05 : z = 1.96 \approx 2$ one sometimes uses the practical approximation

$$n \approx \frac{N}{1+Nd^2}.$$

3.9 Interval estimation using R

```
> ##Sampling of confidence interval
> ##R code for Figure 3.1##
> n=20;nsim=100;mu=4;sigma=2
> xbar=rep(NA,nsim)
> xsd=rep(NA,nsim)
> SE=rep(NA,nsim)
>
> for(i in 1:nsim){
+ x=rnorm(n,mean=mu,sd=sigma)
+ xbar[i]=mean(x)
+ xsd[i]=sd(x)
+ SE[i]=sd(x)/sqrt(n)
+ alpha=0.05;zstar=qnorm(1-alpha/2)
+ matplot(rbind(xbar-zstar*SE,xbar+zstar*SE),rbind(1:nsim,1:nsim),
type="l",lty=1,lwd=2,xlab = "mean tail length",ylab = "sample run")}
> abline(v=mu)
> cov=sum(xbar-zstar*SE <= mu & xbar+zstar*SE >=mu)
> cov ## Number of intervals that contain the parameter.
[1] 93
#Note that out of 100 intervals constructed 93(i.e. 93%)
# of them contain the mean.
##If we increase the sample size we can bring this percentage close to 95%.
```

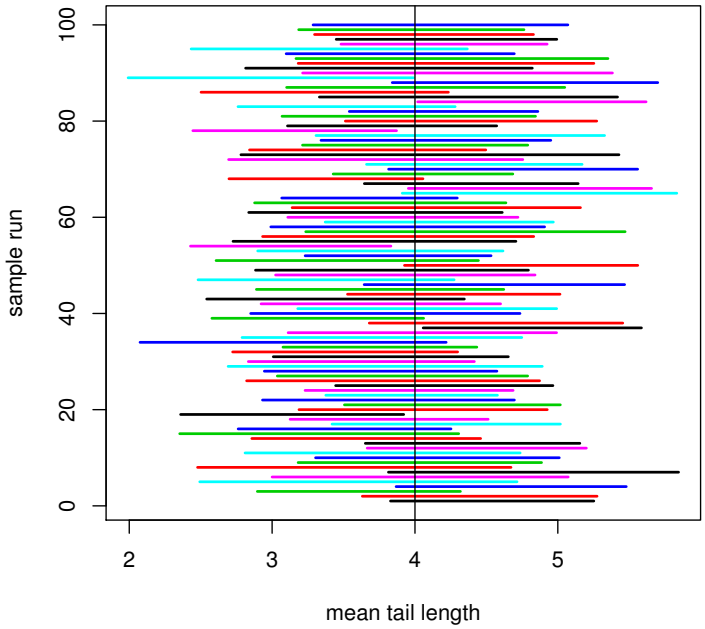


Figure 3.1: sampling of confidence interval

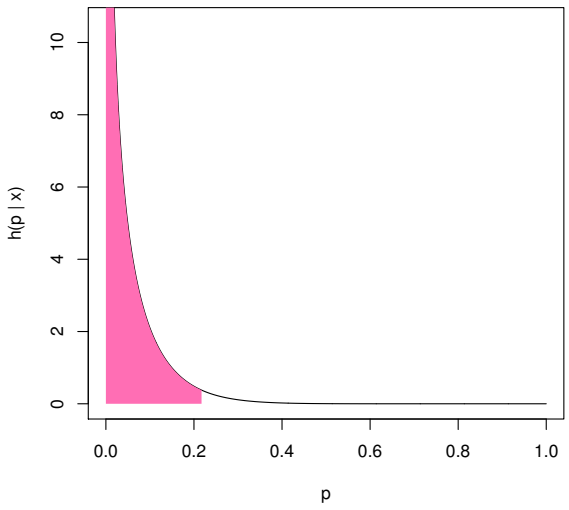


Figure 3.2: Bayseian Interval Estimate

```
##code for Figure3.2##
#Bayesian interval estimate

x = 0
n = 10
alpha1 = 1 / 2
alpha2 = 1 / 2
conf.level = 0.95
alpha = 1 - conf.level

qlow = qbeta(alpha / 2, x + alpha1, n - x + alpha2)
qhig = qbeta(alpha / 2, x + alpha1, n - x + alpha2,
    lower.tail = FALSE)
round(c(qlow, qhig), 4)

eps = 1e-4
theta = seq(0, 1, eps)
y = dbeta(theta, x + alpha1, n - x + alpha2)
ymax = max(y)
if (! is.finite(ymax)) ymax <- max(
    dbeta(0.02, x + alpha1, n - x + alpha2),
    dbeta(0.98, x + alpha1, n - x + alpha2))
qlow = round(qlow / eps) * eps
qhig = round(qhig / eps) * eps
plot(theta, y, type = "l", ylim = c(0, ymax),
    xlab = "p", ylab = "h(p | x)")
tpoly = seq(qlow, qhig, eps)
xpoly = c(tpoly, qhig, qlow)
ypoly = c(dbeta(tpoly, x + alpha1, n - x + alpha2), 0, 0)
ypoly = pmin(ypoly, par("usr")[4])
polygon(xpoly, ypoly, border = NA, col = "hotpink1")
lines(theta, y)
```

confidence interval for the mean of a normal population: Two sided

```
> ## Confidence intervals for the mean of the normal distribution.
> #Two sided confidence interval
> #Let us generate normal data and then find a 95% confidence interval
> #for the mean of a normal population when the variance is known.
> # (The set.seed command resets the random number
> #generator to a specific point so that we can reproduce results
> #if required.)
> set.seed(12345)
> normdata <- rnorm(15, mean=100, sd=20)
> mean(normdata)+c(-1,1)*qnorm(0.975)*20/sqrt(length(normdata))#we used
> # z-distribution.
[1] 90.56137 110.80379
>
> # Let us consider the following data on ozone levels (in ppm)
> # taken on 10 days in a market garden.We wish to construct a 95%
> # confidence interval for the population mean assuming that the
> #observations are taken from a normal population.
> gb=c(5,5,6,7,4,4,3,5,6,5)
> mean(gb)+c(-1,1)*qt(0.975,9)*sd(gb)/sqrt(10)# used t-distribution
[1] 4.173977 5.826023 # A 95% confidence interval for the mean
```

confidence interval for the mean of a normal population: One sided

```
> #One sided lower 95% confidence interval for a normal population
# mean with variance=1.5.
> gb=c(5,5,6,7,4,4,3,5,6,5)
> sigma=1.5
> simple.z.test = function(x,sigma,conf.level=0.95) {
+ n = length(gb);xbar=mean(gb)
+ alpha = 1 - conf.level
+ zstar = qnorm(1-alpha)
+ SE = sigma/sqrt(n)
+ xbar - zstar*SE
+ }
> ## now try it
> simple.z.test(x,sigma)
[1] 4.219777
> #One sided upper 95% confidence interval for a normal population mean
# with variance=1.5.
> gb=c(5,5,6,7,4,4,3,5,6,5)
> sigma=1.5
> simple.z.test = function(x,sigma,conf.level=0.95) {
+ n = length(gb);xbar=mean(gb)
+ alpha = 1 - conf.level
+ zstar = qnorm(1-alpha)
+ SE = sigma/sqrt(n)
+ xbar + zstar*SE
+ }
> ## now try it
> simple.z.test(gb,sigma)
[1] 5.780223
```

Confidence interval for the variance of a normal population

```
> ## Confidence interval for the population variance
> x=rnorm(30,20,4)
> df=length(x)-1
> s2=var(x)
> df*s2/qchisq(c(0.025,0.975),df,lower.tail=FALSE)
[1] 10.54129 30.03488
## Note that in this case we know the true variance and we observe
# that it lies in the interval.
```

Approximate confidence interval for proportion

```
> ## Approximate confidence interval for proportion
> ##You can find the formula used to get this confidence interval
## on page(134).
> m=1;n=20;p=0.5
> xbar=rbinom(m,n,p)/n
> yn=n*xbar
> z=qnorm(0.975)
> c=yn+z*z/2
> b=sqrt((yn*(n-yn))/n +z*z/4)
> l=(c-z*b)/(n+z*z)
> r=(c+z*b)/(n+z*z)
> cat("95%CI is(",l,",",r,")\n",sep="")
95%CI is(0.299298,0.700702)
```

Approximate confidence interval for Poisson parameter

```
> ## Approximate 95% confidence interval for the Poisson parameter.
> ## The formula used to get this confidence interval can be found
on page (135)
> n=2000
> la=2
> z=qnorm(0.975)
> x=rpois(n,la)
> xbar=mean(x)
> c=xbar+(z*z)/(2*n)
> d=sqrt(((xbar*z*z)/n) +(z^4)/(4*n^2))
> l=c-d
> r=c+d
> cat("95%CI is(",l,",",r,")\n",sep="")
95%CI is(1.913378,2.036543)
```

3.10 Exercises

1. Let $X_1, \ldots, X_n$ be a random sample with p.d.f. given by

$$f_X(x;\theta) = e^{-(x-\theta)} I_{(\theta,\infty)}(x), \theta \in \Theta = I\!R,$$

set $Y_1 = X_{(1)}$. Then show that:

(i) The p.d.f. f of Y_1 is given by $f_{Y1}(y_1) = ne^{-n(y-\theta)} I_{(\theta,\infty)}(y)$.

(ii) The random variable $T_n(\theta) = 2n(Y_1 - \theta)$ is distributed as χ^2_2.

(iii) A confidence interval for θ, based on $T_n(\theta)$, with confidence coefficient $1-\alpha$ is of the form $[Y_1 - (b/2n), Y_1 - (a/2n)]$.

2. Let $X_1, \ldots, X_n$ be a random sample from $U(0,\theta)$. Set $R = X_{(n)} - X_{(1)}$. Then:

(i) Find the distribution of R.

(ii) Show that a confidence interval for θ, based on R with confidence coefficient $1-\alpha$ is of the form $[R, R/c]$, where c is a root of the equation $c^{n-1}[n-(n-1)c] = \alpha$.

3. Let $X_1, \ldots, X_n$ be a random sample from Weibull p.d.f. Then show that

(i) The r.v. $T_n(\theta) = 2Y/\theta$ is distributed as χ^2_{2n} where $Y = \sum_{i=1}^n$.

(ii) A confidence interval for θ, based on $T_n(\theta)$, with confidence coefficient $1-\alpha$ is of the form $[2Y/b, 2Y/a]$.

4. Suppose that the random variable X has a geometric probability density function with parameter θ.

(i) Derive a conservative one-sided lower $100(1-\alpha)\%$ confidence limit for θ based on a single observation x.

(ii) If $x = 5$, find a one sided lower 90% confidence limit for θ.

(iii) If $X_1, \ldots, X_n$ is a random sample from a geometric probability density function with parameter θ, describe the form of one sided lower $100(1-\alpha)\%$ confidence limit for θ based on sufficient statistics.

5. Let $X_1, \ldots, X_n$ be a random sample from $Exp(1/\theta)$. Suppose that the prior density of θ is also $Exp(1/\beta)$, where β is known. Then,

(i) Find the posterior distribution of θ.

(ii) Derive $100(1-\alpha)\%$ Bayesian interval estimate of θ.

(iii) Derive $100(1-\alpha)\%$ Bayesian interval estimate of $1/\theta$.

6. If x is a value of a random variable having the exponential distribution, find k so that the interval from 0 to kx is a 1-α confidence interval for the parameter θ.

7. Let X be a single observation from the density

$$f_X(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x)$$

, where $\theta > 0$.

(i) Find a pivotal quantity, and use it to find a confidence- interval estimator of θ.

(ii) Show that $(Y/2, Y)$ is a confidence interval. Find the confidence cofficent.

8. Let$X_1, \ldots, X_n$ be a random sample from $f_X(x;\theta) = I_{(\theta-1/2,\theta+1/2)}(x)$. Let $Y_1 < \ldots < Y_n$ be the corresponding ordered sample. Show that (Y_1, Y_n) is a confidence interval for θ. Find its confidence cofficient.

9. Let $X_1, \ldots, X_n$ be a random sample from $f_X(x;\theta) = (1/\theta)x^{(1-\theta)/\theta}I_{[0,1]}(x)$, where $\theta > 0$. Find the $100(1-\alpha)\%$ interval for θ. Find its expected length.

10. Consider independent random samples from two exponential distributions, $X_i \sim Exp(\theta_1)$ and $Y_j \sim Exp(\theta_2)$; $i = 1, \ldots, n_1, j = 1, \ldots, n_2$.

(i) Show that $(\theta_2/\theta_1)(\overline{X}/\overline{Y}) \sim F(2n_1, 2n_2)$

(ii) Derive a $100(1-\alpha)\%$ CI for θ_2/θ_1.

11. Consider a random sample of size n from $U(0,\theta)$ $\theta > 0$, and let Y_n be the largest order statistic.

(i) Find the probability that the random interval $(Y_n, 2Y_n)$ contains θ.

(ii) Find the constant c such that (y_n, cy_n) is a $100(1-\alpha)\%CI$ for θ.

12. Let$X_1, \ldots, X_n$ be a random sample from a beta$(\theta, 1)$ p.d.f. and assume that θ has a gamma(α, β) prior p.d.f. Find a $1-\alpha$ Bayes interval set for θ.

13. Suppose that $X_1, \ldots, X_n$ is a random sample from a $N(\mu;\sigma^2)$ population.

(i) If σ^2 is known, find a minimum value for n to guarantee that a 0.95 confidence interval for μ will have length no more than $\sigma/4$.

(ii) If σ^2 is unknown, find a minimum value for n to guarantee, with probability 0.90, that a 0.95 confidence interval for μ will have length no more than $\sigma/4$.

14. If X_1 and X_2 are independent random variables having, respectively, binomial distributions with the parameters n_1 and θ_1 and the parameters n_2 and θ_2, construct a $1-\alpha$ large sample confidence interval for $\theta_1-\theta_2$. (Hint: Approximate the distribution of $X_1/n_1 - X_2/n_2$ with a normal distribution.)

15. Let Y denote the sum of of the items of a random sample of size n from a distribution which is $B(1,\theta)$ Assume that the unknown θ is a value of a random variable Θ which has a beta distribution with parameters α and β.

(i) Find the posterior distribution of θ.

(ii) Explain how to find a Bayesian interval estimate of θ subject to the availability of suitable tables of integrals.

16. X is a single observation from $\theta e^{\theta}I_{(0,\infty)}(x)$, where $\theta > 0$.

(i) $(X, 2X)$ is a confidence interval for $1/\theta$. What is the confidence coefficient?

(ii) Find another confidence interval for $1/\theta$ that has the same coefficient but smaller expected length.

17. Let X_1, X_2 be a random sample of size 2 from $N(\theta; 1)$. Let $Y_1 < Y_2$ be the corresponding order sample.

    (i) Determine γ in $P[Y_1 < \theta < Y_2] = \gamma$. Find the expected length of the interval (Y_1, Y_2).

    (ii) Find the confidence interval estimator for θ using $\overline{x} - \theta$ as a pivotal quantity that has a confidence coefficent γ and compare the length with the expected length in part(i).

18. $X_1, \ldots, X_n$ is a random sample from $(1/\theta)x^{(1-\theta)/\theta}I_{[0,1]}(x)$, where $\theta > 0$. Find the $100(1-\theta)\%$ CI for θ. Find the expected length. Find the limiting expected length of your confidence interval. Find n such that $P[length \leq \delta\theta] \geq \rho$ for fixed δ and ρ. (You may use the central limit theorem).

19. Develop a method for estimating the parameter of the Poisson distribution by a confidence interval.

20. Let $X_1, \ldots, X_n$ be a random sample from $f(x/\theta) = \theta x^{\theta-1}I_{(0,1)}(x)$, where $\theta > 0$. Assume that the prior distribution of Θ is given by

$$f_\Theta(\theta) = \frac{\lambda^r \theta^{r-1} e^{-\lambda\theta}}{\Gamma(r)} I_{(0,\infty)}(\theta)$$

    where r and λ are known. Find a 95 percent Bayesian interval estimator of θ.

Chapter 4

Hypothesis Testing

4.1 Introduction

The two major areas of statistical inference are estimation of parameters and testing hypotheses. With regard to estimation we have tried to deal about it in the preceding chapters and the case of hypotheses testing will be discussed in this chapter. The general aim of the chapter is develop general methods for testing hypotheses and to apply those methods to some common problems.
A **hypothesis** is a statement or a claim about some unknown aspect of the state of nature. Scientific investigators, industrial quality control engineers, market researchers, government decision makers, among others, will often have hypotheses about the particular facets of nature of immediate concern to them. They gather data and look to the data for evidence that will help either support or cast doubt on their assertion. A **test** of hypothesis is a procedure, based on sample information, that culminates in an inferential statement about the hypothesis and possibly, in some situations, in a decision what action to take. Let $X_1, \ldots, X_n$ be a random sample from X. Suppose that X has density $f(x;\theta)$ belonging to some family $\{f(x;\theta)|\theta \in \Theta\}$.

Definition

A **statistical hypothesis** is a statement about the distribution of X (in our case : about the parameter θ).

We write :

$$H : \theta \in \Theta_0$$

where Θ_0 is some subset of Θ.

If Θ_0 contains only one member, then we say that the statistical hypothesis H is **simple**. Otherwise H is called **composite**.

Definition

A **test of a statistical hypothesis** H is a rule for deciding whether to **reject** H or not, given the observations.

Such a rule is based on the observed values of $\underset{\sim}{X} = (X_1, \ldots, X_n)$, namely $\underset{\sim}{x} = (x_1, \ldots, x_n)$.

Let $\mathcal{X}$ denote the sample space of all possible values $(x_1, \ldots, x_n)$ of $X_1, \ldots, X_n$. A test of the hypothesis H can be defined as :

Partition the sample space $\mathcal{X}$ into two subsets :

- the set R of outcomes $\underset{\sim}{x}$ not consistent with H : the **rejection region of** H
- the set $R^c = \mathcal{X} \backslash R$ of outcomes, consistent with H : the **acceptance region of** H.

Hence, the rule is :

reject H if and only if $\underset{\sim}{x} = (x_1, \ldots, x_n) \in R$

Such a test is called a **non randomized test**.
It is specified by the rejection region R which is also called the **critical region** of the test.

Note

In a non randomized test, the decision is usually not taken on the basis of $\underset{\sim}{x} = (x_1, \ldots, x_n)$ but on some function $t(\underset{\sim}{x}) = t(x_1, \ldots, x_n)$, where $T_n = t(X_1, \ldots, X_n)$ is some statistic, called **test statistic**. The rejection region R then typically takes the form

$$\begin{array}{ll} & \{\underset{\sim}{x} \mid t(\underset{\sim}{x}) \geq c\} \\ \text{or} & \{\underset{\sim}{x} \mid t(\underset{\sim}{x}) \leq c\} \\ \text{or} & \{\underset{\sim}{x} \mid t(\underset{\sim}{x}) \leq c\} \cup \{\underset{\sim}{x} \mid t(\underset{\sim}{x}) \geq c'\}. \end{array}$$

Purely to overcome some mathematical difficulties (see later) we also have to define **randomized tests**. Suppose we have a function, defined on $\mathcal{X}$, taking values in the interval $[0, 1]$:

$$\begin{array}{rcl} \phi & : & \mathcal{X} \to [0,1] \\ & & \underset{\sim}{x} \mapsto \phi(\underset{\sim}{x}) \end{array}$$

Then a test can be defined as follows :

If $\underset{\sim}{x}$ is observed, then calculate $\phi(\underset{\sim}{x})$, and

- reject H with probability $\phi(\underset{\sim}{x})$
- do not reject H with probability $1 - \phi(\underset{\sim}{x})$.

Such a test is called a **randomized** test. It is specified by the function ϕ which is called the **critical function** of the test.
(We will sometimes use the terminology : "a test ϕ" instead of "a test with critical function ϕ".)

Note
A non randomized test is a particular case of a randomized test. Indeed, if we define

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \dots \text{ if } \underset{\sim}{x} \in R \\ 0 & \dots \text{ if } \underset{\sim}{x} \in R^c \end{cases}$$

then we have a critical function of a non randomized test with critical region R.

4.2 Neyman - Pearson theory

In the general theory of Neyman and Pearson, two statistical hypotheses are involved :

- the **null hypothesis** : $H_0 : \theta \in \Theta_0$

- the **alternative hypothesis** : $H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$.

The hypothesis to be tested is the null hypothesis (and the idea is that, if H_0 is not true, then H_1 is true). We say that we test $\mathbf{H_0}$ **versus** $\mathbf{H_1}$ (or $\mathbf{H_0}$ **against** $\mathbf{H_1}$).

When performing a test of H_0 versus H_1 one may arrive at the correct decision or may commit one of the following two kinds of errors :

- **Type I error** : rejecting H_0 when H_0 is true

- **Type II error** : not rejecting H_0 when H_0 is false.

Associated with any test there are two functions which describe the probabilities of these errors :

- **Type I error probability** : described by a function $\alpha(.)$ on Θ_0 :

$$\begin{aligned} \alpha(\theta) &= P_\theta(\underset{\sim}{X} \in R) \ , \text{ for } \theta \in \Theta_0 \\ [&= E_\theta(\phi(\underset{\sim}{X})) \ , \text{ for } \theta \in \Theta_0]. \end{aligned}$$

- **Type II error probability** : described by a function $\beta(.)$ on Θ_1 :

$$\begin{aligned} \beta(\theta) &= P_\theta(\underset{\sim}{X} \in R^c) \ , \text{ for } \theta \in \Theta_1 \\ &= 1 - P_\theta(\underset{\sim}{X} \in R) \ , \text{ for } \theta \in \Theta_1 \\ [&= 1 - E_\theta(\phi(\underset{\sim}{X})) \ , \text{ for } \theta \in \Theta_1]. \end{aligned}$$

The number $1 - \beta(\theta_1)$, for some $\theta_1 \in \Theta_1$, is called the **power of the test against the alternative** θ_1.
Both power and type I error probability are contained in the **power function of the test**.

Definition

The **power function** of a test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ is defined for all $\theta \in \Theta = \Theta_0 \cup \Theta_1$ as :

$$\begin{aligned} \pi(\theta) &= P_\theta(\underset{\sim}{X} \in R) \ , \text{ for } \theta \in \Theta \\ [&= E_\theta(\phi(\underset{\sim}{X})) \ , \text{ for } \theta \in \Theta]. \end{aligned}$$

Note that, if $\theta \in \Theta_0$, then $\pi(\theta) = \alpha(\theta) =$ the probability of a type I error and if $\theta \in \Theta_1$, then $\pi(\theta) = 1 - \beta(\theta) = 1-$ the probability of a type II error.

It is natural to aim first for a test whose **type I** and **type II** error probabilities are zero. This is usually impossible. Taking $R =$ the empty set (i.e. never reject H_0) gives $\alpha(\theta) = 0$ and $\beta(\theta) = 1$. Taking $R = \mathcal{X}$ (i.e. always reject H_0) makes $\beta(\theta) = 0$ and $\alpha(\theta) = 1$. Therefore we are going to keep these probabilities at an acceptable small level.

The proposal of Neyman and Pearson is as follows :

- control the type I error probability by specifying some small number $0 < \alpha < 1$ and requiring that it should be $\leq \alpha$ for all $\theta \in \Theta_0$:

$$\alpha(\theta) \leq \alpha \ , \text{ for all } \theta \in \Theta_0.$$

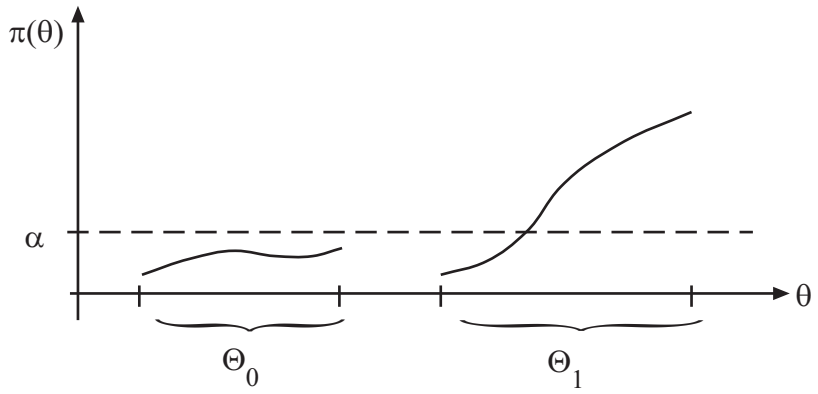
We say that the test has **significance level** α or that the test is a **level-α test**.
Having prescribed α in advance, it is not always the case that this upper bound for

$\alpha(\theta)$ is attained for some $\theta \in \Theta_0$.
The number

$$\sup_{\theta \in \Theta_0} \alpha(\theta)$$

is called the **size of the test**.

- Next we try to arrive at an optimum test. Having restricted to tests of level α, we will select within this class on the basis of the type II error probability $\beta(\theta)$, or equivalently the power function $\pi(\theta) = 1 - \beta(\theta)$. The problem is to select the test so as to maximize the power $\pi(\theta)$ for all $\theta \in \Theta_1$, subject to the condition : $\alpha(\theta) \leq \alpha$ for all $\theta \in \Theta_0$.



The difficulty that arises here is the following : typically, the test that maximizes the power against a certain alternative in Θ_1, depends on this alternative.
There is one important exception : if Θ_1 is simple. If Θ_1 is composite, it may also turn out that the same test maximizes the power for all alternatives in Θ_1. This will be called a uniformly most powerful (UMP) test.

Note
The choice of the significance level α is done by the statistician. Typical choices are : $\alpha = 0.05,\ \alpha = 0.01, \ldots$
If the test is based on a test statistic, then an alternative way of presenting the result of a statistical test is by reporting the ***p*-value of the test**.
The ***p*-value** (probability value) of a test of a null hypothesis $H_0 : \theta = \theta_0$ is the probability, under H_0, of obtaining the observed value of the test statistic or a value that is more extreme in the direction of the alternative hypothesis.

The less the p-value, the less reason there is to believe that H_0 is true. The p-value is commonly used in scientific reporting (e.g. "$p \leq 0.05$", ...) and it is also used in the out prints of statistical computer packages.

If $T_n = t(X_1, \ldots, X_n)$ is the test statistic and if the critical region of the test is of the form $\{\underset{\sim}{x} \mid t(x_1, \ldots, x_n) \geq c\}$, then, for an observed sample $x_1^*, \ldots x_n^*$, the p-value is

$$P_{\theta_0}(T_n \geq t(x_1^*, \ldots, x_n^*)).$$

If $H_0 : \theta \in \Theta_0$ is composite, then the p-value is

$$\sup_{\theta \in \Theta_0} P_\theta(T_n \geq t(x_1^*, \ldots, x_n^*)).$$

Similarly for the other types of critical regions.

4.3 Simple hypotheses versus simple alternative

If the statistical hypothesis completely specifies about the distribution, then it is referred to as **simple hypothesis**, other wise it is called composite hypothesis. Regarding the composite hypothesis, we will discuss about it in the next section detail.

Here our objective is to infer about the parent population from which our samples came from. We assume that we have a sample that came from one of two completely specified distributions. The aim here is to indicate from which population the samples came from. More precisely, assume that a random sample $X_1, \ldots, X_n$ came from $f_0(x)$ or $f_1(x)$ and we want to test $H_o : X_i$ is distributed as $f_0(.)$ versus $H_1 : X_i$ is distributed as $f_1(.)$.

If we had only one observation say x_1, one might quite rationally decide that the observation came from $f_0(.)$ if $f_0(x_1) > f_1(x_1)$, and conversely, decide that the observation came from $f_1(.)$ if $f_1(x_1) > f_0(x_1)$. This simple intuitive method of obtaining a test can be expanded in to a family of tests that, as we will consider will contain some good tests.

Definition Simple likelihood ratio-test

Let $X_1,\ldots,X_n$ be a random sample from either $f_0(.)$ or $f_1(.)$. A test ϕ^* of H_0: $X_i \sim f_0(.)$ versus H_1: $X_i \sim f_1(.)$ is defined to be a simple likelihood ratio-test if ϕ^* is defined by:
Reject H_0, if $\lambda > k$,
Accept H_1, if $\lambda < k$,
Either accept or reject H_0 or randomize if $\lambda = k$,
Where

$$\lambda = \lambda(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_0(x_i)}$$

$$= \frac{L_1(x_1, \ldots, x_n)}{L_0(x_1, \ldots, x_n)}$$

and k is *a* non-negative constant. $L_j = L_j(x_1, \ldots, x_n)$ for $j = 0, 1$ is the likelihood function for sampling from the density $f_j(.)$.

4.3.1 Most powerful test

Suppose $\Theta = \{\theta_0, \theta_1\},\ \Theta_0 = \{\theta_0\},\ \Theta_1 = \{\theta_1\}$. Hence :

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1. \end{aligned}$$

This case is not very useful in practice, but this simplest possible situation illustrates the theory.

How to obtain a test ?

Heuristically : if we have observations $\underset{\sim}{x} = (x_1, \ldots, x_n)$ and we have to decide whether they come from density $f(.;\theta_0)$ or from density $f(.;\theta_1)$, then an intuitive reasoning says that we should reject H_0 if it is more likely that the sample came from $f(.;\theta_1)$ than from $f(.\theta_0)$. That is, we have to compare the likelihood functions

$$L(\theta_1; \underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i; \theta_1) \qquad \text{and} \qquad L(\theta_0; \underset{\sim}{x}) = \prod_{i=1}^{n} f(x_i; \theta_0)$$

and reject H_0 if

$$\frac{L(\theta_0; \underset{\sim}{x})}{L(\theta_1; \underset{\sim}{x})} \text{ is sufficiently small.}$$

How to define a 'best' test ?

Definition

A test ϕ^* of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is a **most powerful (MP) test of size** α $(0 < \alpha < 1)$ if

i) $\pi_{\phi^*}(\theta_0) = \alpha$

ii) $\pi_{\phi^*}(\theta_1) \geq \pi_\phi(\theta_1)$, for any other test ϕ with $\pi_\phi(\theta_0) \leq \alpha$.

Hence : a test ϕ^* is MP of size α if it has size α, and if, among all other tests of size $\leq \alpha$, it has the largest power.

The key result on MP tests in the lemma of Neyman and Pearson. We shall first state and prove a limited version, nl. for nonrandomized tests.

Lemma [Neyman - Pearson]

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$. Let $0 < \alpha < 1$.
Consider the testing problem : $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$.
Suppose that there exists a test with critical region R^* of the form

$$R^* = \left\{ \underset{\sim}{x} \mid \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} \geq k \right\}$$

for some $k \geq 0$ and such that

$$P_{\theta_0}(\underset{\sim}{X} \in R^*) = \alpha$$

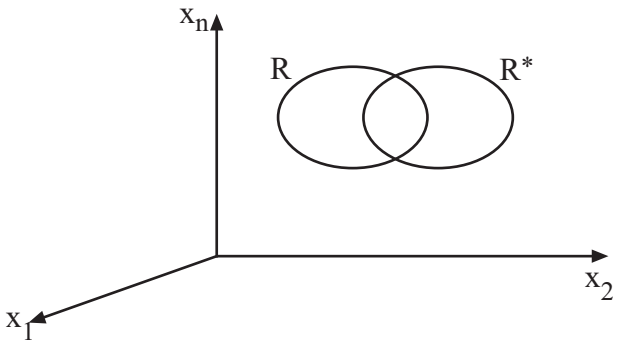
Then this test is MP.

Proof

Consider any other test, with critical region R and with

$$P_{\theta_0}(\underset{\sim}{X} \in R) \leq \alpha$$

Then, we have to show that $P_{\theta_1}(\underset{\sim}{X} \in R^*) \geq P_{\theta_1}(\underset{\sim}{X} \in R)$.
We give the proof for the case of density functions. The same proof holds for discrete densities if we replace integrals by sums.



$$P_{\theta_1}(\underset{\sim}{X} \in R^*) - P_{\theta_1}(\underset{\sim}{X} \in R)$$

$$= \int \dots \int_{R^*} \prod_{i=1}^{n} f(x_i;\theta_1)dx_1 \dots dx_n - \int \dots \int_{R} \prod_{i=1}^{n} f(x_i;\theta_1)dx_1 \dots dx_n$$

$$= \int \dots \int_{R^* \cap R^c} \prod_{i=1}^{n} f(x_i;\theta_1)dx_1 \dots dx_n - \int \dots \int_{R \cap R^{*c}} \prod_{i=1}^{n} f(x_i;\theta_1)dx_1 \dots dx_n$$

Since $\prod_{i=1}^{n} f(x_i;\theta_1) \geq k \prod_{i=1}^{n} f(x_i;\theta_0)$ on $R^* \cap R^c$ and $\prod_{i=1}^{n,} f(x_i;\theta_1) < k \prod_{i=1}^{n} f(x_i;\theta_0)$ on $R \cap R^{*c}$, we have

$$P_{\theta_1}(\underset{\sim}{X} \in R^*) - P_{\theta_1}(\underset{\sim}{X} \in R)$$

$$\geq k \left\{ \int \dots \int_{R^* \cap R^c} \prod_{i=1}^{n} f(x_i;\theta_0)dx_1 \dots dx_n - \int \dots \int_{R \cap R^{*c}} \prod_{i=1}^{n} f(x_i;\theta_0)dx_1 \dots dx_n \right\}$$

$$= k \left\{ \int \dots \int_{R^*} \prod_{i=1}^{n} f(x_i;\theta_0)dx_1 \dots dx_n - \int \dots \int_{R} \prod_{i=1}^{n} f(x_i;\theta_0)dx_1 \dots dx_n \right\}$$

$$= k \left[P_{\theta_0}(\underset{\sim}{X} \in R^*) - P_{\theta_0}(\underset{\sim}{X} \in R) \right] \geq 0,$$

since $P_{\theta_0}(\underset{\sim}{X} \in R^*) = \alpha$ and $P_{\theta_0}(\underset{\sim}{X} \in R) \leq \alpha$. □

Note
The Neyman-Pearson lemma indicates how to choose k : if possible we should choose k such that

$$P_{\theta_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\theta_1)}{\prod_{i=1}^{n} f(X_i;\theta_0)} \geq k\right) = \alpha \ .$$

In the examples below we illustrate how to do this. In discrete variable problems, this is not always possible (see below).

Example [mean of normal with known variance]
$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$ with σ^2 known.
$H_0 : \mu = \mu_0$
$H_1 : \mu = \mu_1 \qquad$ (where $\mu_1 > \mu_0$).

We have :

$$\frac{\prod_{i=1}^{n} f(x_i;\mu_1)}{\prod_{i=1}^{n} f(x_i;\mu_0)} = e^{\frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{n} x_i + n\frac{\mu_0^2-\mu_1^2}{2\sigma^2}} \ .$$

Hence

$$\begin{aligned} R^* &= \left\{ \underset{\sim}{x} \mid \frac{\prod_{i=1}^{n} f(x_i;\mu_1)}{\prod_{i=1}^{n} f(x_i;\mu_0)} \geq k \right\} \\ &= \left\{ \underset{\sim}{x} \mid \frac{\mu_1-\mu_0}{\sigma^2}\sum_{i=1}^{n} x_i + n\frac{\mu_0^2-\mu_1^2}{2\sigma^2} \geq k' \right\}, \text{ for some } k' \\ &= \{\underset{\sim}{x} \mid \overline{x} \geq k''\}, \text{ for some } k''. \end{aligned}$$

Hence :

$$\begin{aligned} & P_{\mu_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\mu_1)}{\prod_{i=1}^{n} f(X_i;\mu_0)} \geq k\right) = \alpha \\ \Leftrightarrow \quad & P_{\mu_0}(\overline{X} \geq k'') = \alpha \\ \Leftrightarrow \quad & P_{\mu_0}\left(\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \geq \frac{k''-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right) = \alpha \end{aligned}$$

$$\Leftrightarrow \quad 1-\Phi\left(\frac{k''-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}\right)=\alpha, \text{ since } \overline{X}\sim N\left(\mu_0;\frac{\sigma^2}{n}\right) \text{ if } H_0 \text{ is true}$$

$$\Leftrightarrow \quad \frac{k''-\mu_0}{\sqrt{\frac{\sigma^2}{n}}}=\Phi^{-1}(1-\alpha)=z_{1-\alpha}$$

$$\Leftrightarrow \quad k''=\mu_0+z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$$

Conclusion : $R^*=\left\{\underset{\sim}{x} \mid \overline{x}\geq \mu_0+z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}\right\}$ is the critical region of a MP test of $H_0:\mu=\mu_0$ versus $H_1:\mu=\mu_1$ $(\mu_1>\mu_0)$.

Example [Variance of normal with known mean]
$X_1,\ldots,X_n$: random sample from $X\sim N(\mu;\sigma^2)$ with μ known.
$H_0:\sigma=\sigma_0$
$H_1:\sigma=\sigma_1$ (where $\sigma_1<\sigma_0$).

We have :

$$\frac{\prod_{i=1}^{n} f(x_i;\sigma_1)}{\prod_{i=1}^{n} f(x_i;\sigma_0)}=\left(\frac{\sigma_0}{\sigma_1}\right)^n e^{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n}(x_i-\mu)^2}$$

Hence

$$P_{\sigma_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\sigma_1)}{\prod_{i=1}^{n} f(X_i;\sigma_0)}\geq k\right)=\alpha$$

$$\Leftrightarrow \quad P_{\sigma_0}\left(\sum_{i=1}^{n}(X_i-\mu)^2\leq k'\right)=\alpha, \text{ for some } k'$$

$$\Leftrightarrow \quad P_{\sigma_0}\left(\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma_0}\right)^2\leq \frac{k'}{\sigma_0^2}\right)=\alpha.$$

Since under $H_0:\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma_0}\right)^2\sim\chi^2(n)$, we have :

Conclusion : $R^* = \left\{ \underset{\sim}{x} \mid \sum_{i=1}^{n}(x_i - \mu)^2 \leq \chi^2_{n;\alpha}.\sigma_0^2 \right\}$ is the critical region of a MP test of $H_0 : \sigma = \sigma_0$ versus $H_1 : \sigma = \sigma_1$ $(\sigma_1 < \sigma_0)$.

Example [Exponential]
$X_1, \ldots, X_n$: random sample from $X \sim Exp(\theta)$.
$H_0 : \theta = \theta_0$
$H_1 : \theta = \theta_1$ (where $\theta_1 > \theta_0$).

We have :

$$\frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} = \left(\frac{\theta_1}{\theta_0}\right)^n e^{-(\theta_1-\theta_0)\sum_{i=1}^{n} x_i}.$$

Hence

$$P_{\theta_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} \geq k \right) = \alpha$$

$$\Leftrightarrow \quad P_{\theta_0}\left(\sum_{i=1}^{n} X_i \leq k' \right) = \alpha, \text{ for some } k'.$$

Since under $H_0 : \sum_{i=1}^{n} X_i \sim \Gamma(n; \frac{1}{\theta_0})$, we have :

Conclusion : $R^* = \left\{ \underset{\sim}{x} \mid \sum_{i=1}^{n} x_i \leq c \right\}$ with c such that

$$\int_0^c \frac{\theta_0^n}{\Gamma(n)} x^{n-1} e^{-\theta_0 x} dx = \alpha$$

is the critical region of a MP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ $(\theta_1 > \theta_0)$.

Example [Bernoulli]

$X_1, \ldots, X_n$: random sample from $X \sim B(1;\theta)$.
$H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1 \qquad$ (where $\theta_1 < \theta_0$)

We have :

$$\frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} = \left(\frac{1-\theta_1}{1-\theta_0}\right)^n \left(\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)}\right)^{\sum_{i=1}^{n} x_i}$$

Since $\theta_1 < \theta_0$, this is a decreasing function of $\sum_{i=1}^{n} x_i$ and hence the problem of finding a k such that

$$P_{\theta_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\theta_1)}{\prod_{i=1}^{n} f(X_i;\theta_0)} \geq k\right) = \alpha$$

is equivalent to that of finding k' such that

$$P_{\theta_0}\left(\sum_{i=1}^{n} X_i \leq k'\right) = \alpha$$

If H_0 is true, then $\sum_{i=1}^{n} X_i \sim B(n;\theta_0)$.
For a given α, it is not always possible to find a value of k' such that this is true.

Example : $n = 3,\ \theta_0 = \frac{3}{4},\ \theta_1 = \frac{1}{4},\ \alpha = 0.05.$
The $B(3;\frac{3}{4})$ distribution is given by

0	1	2	3
0.0156	0.1416	0.4219	0.4219

.

Hence :

- if $0 \leq k' < 1 : P_{\theta_0}\left(\sum_{i=1}^{3} X_i \leq k'\right) = P_{\theta_0}\left(\sum_{i=1}^{3} X_i = 0\right) = 0.0156 < 0.05$

- if $1 \leq k' < 2 : P_{\theta_0}\left(\sum_{i=1}^{3} X_i \leq k'\right) = 0.0156 + 0.1416 > 0.05.$

Example [Poisson]
$X_1, \ldots, X_n$: random sample from $X \sim$ Poisson (λ).
$H_0 : \lambda = \lambda_0$
$H_1 : \lambda = \lambda_1 \qquad$ (where $\lambda_1 > \lambda_0$).

We have :

$$\frac{\prod_{i=1}^{n} f(x_i;\lambda_1)}{\prod_{i=1}^{n} f(x_i;\lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^{\sum_{i=1}^{n} x_i} e^{-n(\lambda_1-\lambda_0)}.$$

Since $\lambda_1 > \lambda_0$, this is an increasing function of $\sum_{i=1}^{n} x_i$ and hence

$$\begin{aligned}
& P_{\lambda_0}\left(\frac{\prod_{i=1}^{n} f(X_i;\lambda_1)}{\prod_{i=1}^{n} f(X_i;\lambda_0)} \geq k\right) = \alpha \\
\Leftrightarrow \quad & P_{\lambda_0}\left(\sum_{i=1}^{n} X_i \geq k'\right) = \alpha, \text{ for some } k'. \\
\Leftrightarrow \quad & \sum_{r=k'}^{\infty} e^{-n\lambda_0}\frac{(n\lambda_0)^r}{r!} = \alpha, \text{ since under } H_0 : \sum_{i=1}^{n} X_i \sim \text{ Poisson } (n\lambda_0).
\end{aligned}$$

For a given α, it is not always possible to find a value of k' such that this equality holds.

Example : $n = 10,\ \lambda_0 = 0.4,\ \alpha = 0.05.$
From tables of the Poisson distribution

$$\begin{aligned}
\sum_{r=8}^{\infty} e^{-4}\frac{4^r}{r!} &= 0.0511 > 0.05 \\
\sum_{r=9}^{\infty} e^{-4}\frac{4^r}{r!} &= 0.0214 < 0.05.
\end{aligned}$$

To overcome the difficulty with discrete distributions (as in the last two examples) we have to introduce randomized tests (see definition in section 1). If we do so, we can prove a more general version of the Neyman-Pearson lemma which says that there **always** exists a MP test of size α for a simple H_0 versus a simple H_1. The randomization is done to get the size equal to α.

Lemma [Neyman-Pearson]

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta)$, $\theta \in \Theta = \{\theta_0, \theta_1\}$. Let $0 < \alpha < 1$.
Consider the testing problem : $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Consider the test with critical function

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \ldots \quad \text{if } \prod_{i=1}^{n} f(x_i;\theta_1) > k \prod_{i=1}^{n} f(x_i;\theta_0) \\ \gamma & \ldots \qquad\qquad = \\ 0 & \ldots \qquad\qquad < \end{cases}$$

where the constants γ $(0 \leq \gamma \leq 1)$ and $k \geq 0$ are determined such that

$$E_{\theta_0}[\phi(\underset{\sim}{X})] = \alpha.$$

Then this test is MP.

Example [Bernoulli]

In the Bernoulli example before, the MP test of $H_0 : \theta = \dfrac{3}{4}$ versus $H_1 : \theta = \dfrac{1}{4}$ would be :

if $\sum_{i=1}^{3} x_i = 0 \ldots$ reject H_0

if $\sum_{i=1}^{3} x_i = 1 \ldots$ reject H_0 with probability 0.245; accept H_0 with probability 0.755

if $\sum_{i=1}^{3} x_i > 1 \ldots$ accept H_0

i.e. with $\underset{\sim}{x} = (x_1, x_2, x_2)$:

$$\phi(\underset{\sim}{x}) = \begin{cases} 1 & \ldots \text{ if } \sum x_i = 0 \\ 0.245 & \ldots \text{ if } \sum x_i = 1 \\ 0 & \ldots \text{ if } \sum x_i > 1 \end{cases} .$$

The number 0.245 has been calculated as to have a type I error probability of exactly 0.05 :

$$\begin{aligned} & E_{\theta_0}[\phi(\underset{\sim}{X})] \\ = \; & P_{\theta_0}(\sum_{i=1}^{3} X_i = 0) + \gamma P_{\theta_0}(\sum_{i=1}^{3} X_i = 1) \\ = \; & 0.0156 + \gamma \; 0.1406 \\ = \; & 0.05 \ldots \text{ if } \gamma = \frac{0.05 - 0.0156}{0.1406} = 0.245. \end{aligned}$$

Note

Randomization is usually not done in practice. Mostly, one changes the α to some level for which a non randomized Neyman-Pearson test can be found.

4.3.2 Minimax and Bayes test

Instead of using the power to set the goodness of the test we could also use a loss function.

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta),\ \theta \in \Theta = \{\theta_0, \theta_1\}$. Consider the testing problem.

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1.$$

Let d_0 be the **decision** that the observed sample comes from $f(x;\theta_0)$ and let d_1 be the **decision** that the observed sample comes from $f(x;\theta_1)$.

A non randomized test with critical region R can be seen as a **decision function** δ with

$$\delta(x_1, \ldots, x_n) = \begin{cases} d_1 & \ldots \text{ if } (x_1, \ldots, x_n) \in R \\ d_0 & \ldots \text{ if } (x_1, \ldots, x_n) \in R^c. \end{cases}$$

Suppose that a certain **loss function** $l(\theta; \delta(x_1, \ldots, x_n)$ has been chosen. Then

$$l(\theta;\delta(x_1,\ldots,x_n)) = \begin{cases} l(\theta;d_1) & \ldots \text{ if } (x_1,\ldots,x_n)\in R \\ l(\theta;d_0) & \ldots \text{ if } (x_1,\ldots,x_n)\in R^c. \end{cases}$$

In the case that θ can assume only two values θ_0 and θ_1,

$l(\theta_0;d_1)$ = the loss when decision d_1 is made and θ_0 is the true parameter
$l(\theta_1;d_1) = 0$
$l(\theta_0;d_0) = 0$
$l(\theta_1;d_0)$ = the loss when decision d_0 is made and θ_1 is the true parameter.

For a test with critical region R, we define the **risk function of the test** as the average loss, i.e.

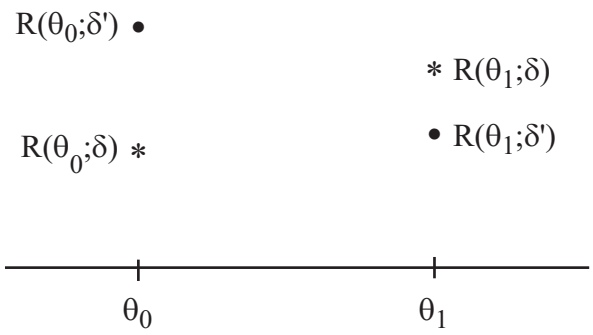
$$\begin{aligned} R(\theta;\delta) &= E_\theta[l(\theta;\delta(X_1,\ldots,X_n))] \\ &= \begin{cases} \sum\ldots\sum l(\theta;\delta(x_1\ldots x_n))\prod_{i=1}^n f(x_i;\theta) \\ \int\ldots\int l(\theta;\delta(x_1,\ldots,x_n))\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \end{cases} \\ &= \begin{cases} \sum\ldots\sum_{R} l(\theta;d_1)\prod_{i=1}^n f(x_i;\theta) + \sum\ldots\sum_{R^c} l(\theta;d_0)\prod_{i=1}^n f(x_i;\theta) \\ \int\ldots\int_{R} l(\theta;d_1)\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n + \int\ldots\int_{R^c} l(\theta;d_0)\prod_{i=1}^n f(x_i;\theta)dx_1\ldots dx_n \end{cases} \\ &= l(\theta;d_1)P_\theta(\underset{\sim}{X}\in R) + l(\theta;d_0)P_\theta(\underset{\sim}{X}\in R^c) \\ &= l(\theta;d_1)\pi(\theta) + l(\theta;d_0)(1-\pi(\theta)) \end{aligned}$$

where π is the power function of the test.
Since θ can only assume two values (θ_0 and θ_1), the risk function can only assume two values :

$$R(\theta_0;\delta) = l(\theta_0;d_1)\pi(\theta_0)$$

$$R(\theta_1;\delta) = l(\theta_1;d_0)(1-\pi(\theta_1)).$$

It is usually impossible to find a test which minimizes the risk function uniformly.
For two tests δ and δ' we typically have the following situation :



A first way to define a 'good' test is to look for a test that minimizes the largest value of the risk function :

Definition

A test δ for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is a **minimax test** if

$$\max\{R(\theta_0;\delta), R(\theta_1;\delta)\} \leq \max\{R(\theta_0;\delta'), R(\theta_1;\delta')\}$$

for any other test δ'.

A second way to define a 'good' test is to look for a test that minimizes the **Bayes risk**, i.e. the average of $R(\theta;\delta)$ over θ, using a **prior** density over the parameter space.
In our case the prior density is the density of a discrete random variable $\widetilde{\Theta}$ with two values θ_0 and θ_1. It is completely determined by

$$p = P(\widetilde{\Theta} = \theta_1)(= 1 - P(\widetilde{\Theta} = \theta_0)).$$

The Bayes risk is given by

$$R(\delta) = (1-p)R(\theta_0;\delta) + pR(\theta_1;\delta).$$

Definition

A test δ for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ is a **Bayes test** with respect to the prior distribution given by $p = P(\widetilde{\Theta} = \theta_1)$ if

$$(1-p)R(\theta_0;\delta) + pR(\theta_1;\delta) \leq (1-p)R(\theta_0;\delta') + pR(\theta_1;\delta')$$

for any other test δ'.

Theorem

The Bayes test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ with respect to the prior distribution given by $p = P(\widetilde{\Theta} = \theta_1)$ has critical region given by

$$R = \left\{ (x_1, \ldots, x_n) \mid \frac{\prod_{i=1}^{n} f(x_i;\theta_1)}{\prod_{i=1}^{n} f(x_i;\theta_0)} \geq \frac{(1-p)l(\theta_0;d_1)}{pl(\theta_1;d_0)} \right\}$$

Proof (continuous case)

$$\begin{aligned} R(\delta) &= (1-p)R(\theta_0;\delta) + pR(\theta_1;\delta) \\ &= (1-p)l(\theta_0;d_1)\pi(\theta_0) + pl(\theta_1;d_0)(1-\pi(\theta_1)) \end{aligned}$$

$$= (1-p)l(\theta_0;d_1) \int \ldots \int \prod_{i=1}^{n} f(x_i;\theta_0)dx_1 \ldots dx_n$$

$$+pl(\theta_1;d_0)\left(1 - \int \ldots \int_R \prod_{i=1}^{n} f(x_i;\theta_1)dx_1 \ldots dx_n\right)$$

$$= pl(\theta_1;d_0)$$

$$+ \int \ldots \int_R [(1-p)l(\theta_0;d_1)\prod_{i=1}^{n} f(x_i;\theta_0) - pl(\theta_1;d_0)\prod_{i=1}^{n} f(x_i;\theta_1)]dx_1 \ldots dx_n$$

and this is minimized if the region R is defined to be all $(x_1, \ldots, x_n)$ for which the integrand is negative. □

4.4 Testing composite hypothesis

We now turn to the more general hypothesis testing problem,
that of testing composite hypothesis. We assume that we have
a random sample from $f(x;\theta)$, $\theta \in \Theta$, and we want to test:
H_0: $\theta \in \Theta_0$ versus
H_1: $\theta \in \Theta_1$
where $\Theta_0 \subset \Theta$,$\Theta_1 \subset \Theta$ and Θ_0
and Θ_1 are disjoint. Usually, $\Theta_1 = \Theta - \Theta_0$.
To perform such tests we use the generalized likelihood ratio test and it
will be described as follows.

4.4.1 Generalized likelihood ratio tests

In the previous section, we considered the problem of testing a simple H_0 versus a simple H_1.
From the Neyman-Pearson lemma we know that the most powerful test rejects H_0 if the **simple likelihood ratio function**

$$\frac{L(\theta_1; \underset{\sim}{x})}{L(\theta_0; \underset{\sim}{x})} = \frac{\prod_{i=1}^{n} f(x_i; \theta_1)}{\prod_{i=1}^{n} f(x_i; \theta_0)}$$

is too large.

For a more general testing problem

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1 = \Theta \backslash \Theta_0$$

where Θ_0 and Θ_1 contain more than one point each, the above idea could lead to the consideration of the ratio

$$\frac{\sup_{\theta \in \Theta_1} L(\theta; \underset{\sim}{x})}{\sup_{\theta \in \Theta_0} L(\theta; \underset{\sim}{x})}$$

or the closely related ratio

$$\frac{\sup\limits_{\theta\in\Theta} L(\theta;\underset{\sim}{x})}{\sup\limits_{\theta\in\Theta_0} L(\theta;\underset{\sim}{x})}.$$

It is more convenient to work with a ratio with values between 0 and 1 rather than between 1 and $+\infty$. Therefore we invert this ratio to

$$\frac{\sup\limits_{\theta\in\Theta_0} L(\theta;\underset{\sim}{x})}{\sup\limits_{\theta\in\Theta} L(\theta;\underset{\sim}{x})}$$

Also, we allow θ to be a vector of parameters : $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k)$.
We finally arrive at the following definition.

Definition

The **generalized likelihood ratio function** for a null hypothesis $H_0 : \underset{\sim}{\theta} \in \Theta_0$ is defined by

$$\lambda(\underset{\sim}{x}) = \lambda(x_1, \ldots, x_n) = \frac{\sup\limits_{\underset{\sim}{\theta}\in\Theta_0} L(\underset{\sim}{\theta};\underset{\sim}{x})}{\sup\limits_{\underset{\sim}{\theta}\in\Theta} L(\underset{\sim}{\theta};\underset{\sim}{x})}.$$

The corresponding statistic

$$\Lambda_n = \lambda(X_1, \ldots, X_n)$$

is called the **generalized likelihood ratio statistic**.

Notes :

(1) We have that $0 \leq \lambda \leq 1$.

(2) if $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, then the generalized likelihood ratio function does not reduce to the simple likelihood ratio function.

(3) For the denominator we have :

$$\sup_{\underset{\sim}{\theta}\in\Theta} L(\underset{\sim}{\theta};\underset{\sim}{x}) = L(\underset{\sim}{\hat{\theta}}_n;\underset{\sim}{x})$$

where $\hat{\underset{\sim}{\theta}}_n$ is the ML-estimate for $\underset{\sim}{\theta}$.

If H_0 is not true, then it is intuitively clear that the numerator in $\lambda(\underset{\sim}{x})$ will tend to be small, so that $\lambda(\underset{\sim}{x})$ will be small. Hence it is reasonable to use this as a test procedure :

Definition

A **generalized likelihood ratio test** for the null hypothesis $H_0 : \underset{\sim}{\theta} \in \Theta_0$, is a test which rejects H_0 if and only if $\lambda(\underset{\sim}{x})$ is small.

The **critical region** of such a test has the form

$$\{\underset{\sim}{x} \mid \lambda(\underset{\sim}{x}) \leq \lambda_0\}$$

where λ_0 is some fixed constant, $0 < \lambda_0 < 1$.

If we want a **size-α test**, then the constant λ_0 has to be determined such that

$$\sup_{\underset{\sim}{\theta} \in \Theta_0} P_{\underset{\sim}{\theta}}(\Lambda_n \leq \lambda_0) = \alpha$$

where Λ_n is the generalized likelihood ratio statistic.

Very often one uses the large sample limiting distribution of Λ_n (see below). This provides **tests with approximate size** α.

4.4.2 Examples generalized likelihood ratio tests

Example [Mean of normal with known variance]

$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$ with σ^2 known.
$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$.

We have : $L(\mu; \underset{\sim}{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$

$$\Theta = I\!R \ , \ \Theta_0 = \{\mu_0\}.$$

$$\sup_{\mu \in I\!R} L(\mu; \underset{\sim}{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$\lambda(\underset{\sim}{x}) = \frac{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2}}{e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2}} = e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2}.$$

Generalized likelihood ratio test :

$$\text{reject } H_0 \text{ if and only if } e^{-\frac{n}{2\sigma^2}(\overline{x}-\mu_0)^2} \leq \lambda_0 \qquad (0 < \lambda_0 < 1)$$

or equivalently,

$$\text{reject } H_0 \text{ if and only if } \left|\frac{\overline{x}-\mu_0}{\sqrt{\sigma^2/n}}\right| \geq c.$$

To set the level equal to α, we choose c such that $P_{\mu_0}\left(\left|\frac{\overline{X}-\mu_0}{\sqrt{\sigma^2/n}}\right| \geq c\right) = \alpha$. Since, under H_0, $\frac{\overline{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0;1)$, we have $c = z_{1-\alpha/2}$.

Example [Exponential]

$X_1, \ldots, X_n$: random sample from $X \sim Exp(\theta)$.
$H_0 : \theta \leq \theta_0$
$H_1 : \theta > \theta_0$.

We have : $L(\theta; \underset{\sim}{x}) = \theta^n e^{-\theta\sum_1^n x_i}$

$$\Theta = \{\theta \mid \theta > 0\} \ , \ \Theta_0 = \{\theta \mid \theta \leq \theta_0\}$$

$$\sup_{\theta\in\Theta} L(\theta;\underset{\sim}{x}) = \left(\frac{n}{\sum x_i}\right)^n e^{-n}$$

$$\sup_{\theta\in\Theta_0} L(\theta;\underset{\sim}{x}) = \begin{cases} \left(\dfrac{n}{\sum x_i}\right)^n e^{-n} & \dots \text{ if } \dfrac{n}{\sum x_i} \le \theta_0 \\ \theta_0^n e^{-\theta_0 \sum x_i} & \dots \text{ if } \dfrac{n}{\sum x_i} > \theta_0 \end{cases}$$

$$\lambda(\underset{\sim}{x}) = \begin{cases} 1 & \dots \text{ if } \dfrac{n}{\sum x_i} \le \theta_0 \\ \dfrac{\theta_0^n e^{-\theta_0 \sum x_i}}{\left(\dfrac{n}{\sum x_i}\right)^n e^{-n}} & \dots \text{ if } \dfrac{n}{\sum x_i} > \theta_0 \end{cases}.$$

Generalized likelihood ratio test :

$$\text{reject } H_0 \text{ if and only if } \frac{n}{\sum x_i} > \theta_0 \text{ and } \left(\frac{\theta_0 \sum x_i}{n}\right)^n e^{-\theta_0 \sum x_i + n} \le \lambda_0$$

for some $0 < \lambda_0 < 1$.
Or, with $\overline{x} = \dfrac{\sum x_i}{n}$:

$$\text{reject } H_0 \text{ if and only if } \theta_0\overline{x} < 1 \text{ and } (\theta_0\overline{x})^n e^{-n(\theta_0\overline{x}-1)} \le \lambda_0.$$

Now : $y^n e^{-n(y-1)}$ is maximal for $y = 1$. Hence :

$$\text{reject } H_0 \text{ if and only if } \theta_0\overline{x} \le c$$

where $0 < c < 1$ is such that $c^n e^{-n(c-1)} = \lambda_0$.

To set the size equal to α we have to choose c such that

$$\sup_{\theta\in\Theta_0} P_\theta(\theta_0\overline{X} \le c) = \alpha$$

$$\text{or} \quad : \quad \sup_{0<\theta\le\theta_0} P_\theta\left(\sum_{i=1}^n X_i \le \frac{nc}{\theta_0}\right) = \alpha$$

$$\text{or} \quad : \quad \sup_{0<\theta\leq\theta_0} \int_0^{\frac{nc}{\theta_0}} \frac{\theta^n}{\Gamma(n)} x^{n-1} e^{-\theta x} dx = \alpha \qquad \left(\text{since } \sum_{i=1}^{n} X_i \sim \Gamma(n; \frac{1}{\theta})\right)$$

$$\text{or} \quad : \quad \sup_{0<\theta\leq\theta_0} \int_0^{\frac{nc\theta}{\theta_0}} \frac{1}{\Gamma(n)} t^{n-1} e^{-t} dt = \alpha$$

or : c has to satisfy the equation :

$$\int_0^{nc} \frac{1}{\Gamma(n)} t^{n-1} e^{-t} dt = \alpha.$$

Example [Mean of normal with unknown variance]

$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$ with μ and σ^2 unknown.
$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$
(where μ_0 is known and σ^2 is unspecified).

We have : $\underset{\sim}{\theta} = (\mu, \sigma^2)$

$$L(\underset{\sim}{\theta}; \underset{\sim}{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

$$\Theta = \{\underset{\sim}{\theta} = (\mu, \sigma^2) \mid \mu \in I\!R,\ \sigma^2 > 0\}$$

$$\Theta_0 = \{\underset{\sim}{\theta} = (\mu, \sigma^2) \mid \mu = \mu_0,\ \sigma^2 > 0\}$$

$$\sup_{\underset{\sim}{\theta}\in\Theta} L(\underset{\sim}{\theta}; \underset{\sim}{x}) = \left(\frac{ne^{-1}}{2\pi\sum(x_i-\overline{x})^2}\right)^{\frac{n}{2}}$$

(replacing μ and σ^2 by the ML estimates
$\overline{x} = \frac{1}{n}\sum x_i$ and $\frac{1}{n}\sum(x_i-\overline{x})^2$)

Also, since

$$L((\mu_0, \sigma^2); \underset{\sim}{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu_0)^2}$$

we have

$$\sup_{\underset{\sim}{\theta}\in\Theta_0} L(\underset{\sim}{\theta};\underset{\sim}{x}) = \left(\frac{ne^{-1}}{2\pi\sum(x_i-\mu_0)^2}\right)^{\frac{n}{2}}$$

Hence :

$$\lambda(\underset{\sim}{x}) = \left(\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\mu_0)^2}\right)^{\frac{n}{2}}$$

or since

$$\sum_{i=1}^{n}(x_i-\mu_0)^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\mu_0)^2 :$$

$$\lambda(\underset{\sim}{x}) = \frac{1}{\left(1+\frac{n(\overline{x}-\mu_0)^2}{\sum(x_i-\overline{x})^2}\right)^{\frac{n}{2}}} = \frac{1}{\left(1+\frac{t^2(\underset{\sim}{x})}{n-1}\right)^{\frac{n}{2}}}$$

where

$$t(\underset{\sim}{x}) = \frac{\overline{x}-\mu_0}{\sqrt{\frac{\sum(x_i-\overline{x})^2}{n(n-1)}}}.$$

The generalized likelihood ratio test rejects H_0 if and only if $\lambda(\underset{\sim}{x})$ is small or, equivalently if and only if $|t(\underset{\sim}{x})|$ is large.

Under $H_0 : \mu = \mu_0$, the corresponding statistic

$$T_n = t(X_1,\ldots,X_n) = \frac{\overline{X}-\mu_0}{\sqrt{\frac{\sum(X_i-\overline{X})^2}{n(n-1)}}} = \frac{\overline{X}-\mu_0}{\sqrt{\frac{S^2}{n-1}}}$$

has a Student's t distribution with $(n-1)$ degrees of freedom.

Conclusion : the size-α generalized likelihood ratio test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ rejects H_0 if and only if $|t(\underset{\sim}{x})| \geq t_{n-1,1-\alpha/2}$, i.e.

$$\left| \frac{\overline{x} - \mu_0}{\sqrt{s^2/n - 1}} \right| \geq t_{n-1,1-\alpha/2}.$$

This is called the **two-sided t-test**.

Note

The example can be modified to other testing problems, such as for instance :

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0. \end{aligned}$$

Since now

$$\begin{aligned} \Theta &= \{\underset{\sim}{\theta} = (\mu, \sigma^2) \mid \mu \in I\!R,\ \sigma^2 > 0\} \\ \Theta_0 &= \{\underset{\sim}{\theta} = (\mu, \sigma^2) \mid \mu \leq \mu_0,\ \sigma^2 > 0\}, \end{aligned}$$

we obtain :

$$\lambda(\underset{\sim}{x}) = \begin{cases} 1 & \dots \text{ if } \overline{x} \leq \mu_0 \\ \dfrac{1}{\left(1 + \dfrac{t^2(\underset{\sim}{x})}{n-1}\right)^{\frac{n}{2}}} & \dots \text{ if } \overline{x} \geq \mu_0 \end{cases}$$

This leads to : reject H_0 if and only if

$$|t(\underset{\sim}{x}) \mid \text{ is large and } \overline{x} \geq \mu_0$$

i.e. if and only if

$$\frac{\overline{x} - \mu_0}{\sqrt{\dfrac{s^2}{n-1}}} \geq t_{n-1,1-\alpha}.$$

This is called a **one-sided t-test**.

Example [Variance of normal with unknown mean]
$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$ with μ and σ^2 unknown.
$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$
(where σ_0^2 is known and μ is unspecified).

Show that the generalized likelihood ratio function is given by

$$\lambda(\underset{\sim}{x}) = \left(\frac{w(\underset{\sim}{x})}{n}\right)^{\frac{n}{2}} e^{-\left(\frac{w(\underset{\sim}{x})}{2} - \frac{n}{2}\right)}$$

where

$$w(\underset{\sim}{x}) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

This is not a monotone function of $w(\underset{\sim}{x})$, but $\lambda(\underset{\sim}{x})$ has a single maximum. Therefore, a generalized likelihood ratio test is given by : reject H_0 if and only if

$$w(\underset{\sim}{x}) \leq c_1 \qquad \text{or} \qquad w(\underset{\sim}{x}) \geq c_2$$

To construct a size-α test, c_1 and c_2 have to be determined from the fact that the statistic

$$W_n = w(X_1, \ldots, X_n) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is $\chi^2(n-1)$ distributed under the null hypothesis.
This is called a **two-sided chi-squared test**.

Example [Comparing means]

Two independent samples :

$$X_1, \ldots, X_{n_1} : \text{ from } X \sim N(\mu_1; \sigma_1^2)$$
$$Y_1, \ldots, Y_{n_2} : \text{ from } Y \sim N(\mu_2; \sigma_2^2)$$

Assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (unknown).

We want to test :
$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$.

The joint density of $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$ is given by

$$\left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+n_2} \frac{1}{\sigma_1^{n_1}} \frac{1}{\sigma_2^{n_2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2 - \frac{1}{2\sigma_2^2}\sum_{i=1}^{n_2}(y_i-\mu_2)^2}.$$

We have : $\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma^2)$

$$L(\underset{\sim}{\theta}; \underset{\sim}{x}, \underset{\sim}{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+n_2} \frac{1}{\sigma^{n_1+n_2}} e^{-\frac{1}{2\sigma^2}[\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{i=1}^{n_2}(y_i-\mu_2)^2]}$$

$$\Theta = \{\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma^2) \mid \mu_1, \mu_2 \in I\!R,\ \sigma^2 > 0\}$$

$$\Theta_0 = \{\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma^2) \mid \mu_1 = \mu_2 \in I\!R,\ \sigma^2 > 0\}$$

One can calculate :

- the supremum of L over Θ is obtained for $\mu_1 = \overline{x},\ \mu_2 = \overline{y}$, and
  $\sigma^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2\right]$
  and hence :

$$\sup_{\underset{\sim}{\theta}\in\Theta} L(\underset{\sim}{\theta}; \underset{\sim}{x}, \underset{\sim}{y}) = \left(\frac{(n_1+n_2)e^{-1}}{2\pi\left[\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2\right]}\right)^{\frac{n_1+n_2}{2}}.$$

- the supremum of L over Θ_0 is obtained for $\mu_1 = \mu_2 = \frac{n_1\overline{x}+n_2\overline{y}}{n_1+n_2} = \overset{\wedge}{\mu}$ and
  $\sigma^2 = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(x_i-\overset{\wedge}{\mu})^2 + \sum_{i=1}^{n_2}(y_i-\overset{\wedge}{\mu})^2\right]$.

  Hence :

$$\sup_{\underset{\sim}{\theta}\in\Theta_0} L(\underset{\sim}{\theta}; \underset{\sim}{x}, \underset{\sim}{y}) = \left(\frac{(n_1+n_2)e^{-1}}{2\pi\left[\sum_{i=1}^{n_1}(x_i-\overset{\wedge}{\mu})^2 + \sum_{i=1}^{n_2}(y_i-\overset{\wedge}{\mu})^2\right]}\right)^{\frac{n_1+n_2}{2}}.$$

It follows that :

$$\lambda(\underset{\sim}{x},\underset{\sim}{y}) = \left(\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2}{\sum_{i=1}^{n_1}(x_i-\hat{\mu})^2 + \sum_{i=1}^{n_2}(y_i-\hat{\mu})^2} \right)^{\frac{n_1+n_2}{2}}.$$

But :

$$\begin{aligned}
& \sum_{i=1}^{n_1}(x_i-\hat{\mu})^2 + \sum_{i=1}^{n_2}(y_i-\hat{\mu})^2 \\
= & \sum_{i=1}^{n_1}(x_i-\overline{x}+\overline{x}-\hat{\mu})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y}+\overline{y}-\hat{\mu})^2 \\
= & \sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2 + n_1(\overline{x}-\hat{\mu})^2 + n_2(\overline{y}-\hat{\mu})^2 \\
= & \sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2 + \frac{n_1 n_2}{n_1+n_2}(\overline{x}-\overline{y})^2.
\end{aligned}$$

Hence :

$$\begin{aligned}
\lambda(\underset{\sim}{x},\underset{\sim}{y}) &= \left(1 + \frac{\frac{n_1 n_2}{n_1+n_2}(\overline{x}-\overline{y})^2}{\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2} \right)^{-\frac{n_1+n_2}{2}} \\
&= \left(1 + \frac{t^2(\underset{\sim}{x},\underset{\sim}{y})}{n_1+n_2-2} \right)^{-\frac{n_1+n_2}{2}}
\end{aligned}$$

where

$$t(\underset{\sim}{x},\underset{\sim}{y}) = \frac{\overline{x}-\overline{y}}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad , \quad s_p^2 = \frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_1+n_2-2}.$$

The test rejects H_0 if and only if $\lambda(\underset{\sim}{x},\underset{\sim}{y})$ is small, i.e. if and only if $|t(\underset{\sim}{x},\underset{\sim}{y})|$ is large. The corresponding statistic is

$$T_n = t(X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

(where $S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}$) and under H_0 it is $t(n_1 + n_2 - 2)$-distributed.
This is called the **two-sample t-test**.

Example [Comparing variances]
Two independent samples

$$X_1, \ldots, X_{n_1} : \text{ from } X \sim N(\mu_1; \sigma_1^2)$$

$$Y_1, \ldots, Y_{n_2} : \text{ from } Y \sim N(\mu_2; \sigma_2^2)$$

We want to test :
$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = \tau$

$H_1 : \frac{\sigma_2^2}{\sigma_1^2} \neq \tau$

(where τ is known; $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ unspecified).

We have : $\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$

$$L(\underset{\sim}{\theta}; \underset{\sim}{x}, \underset{\sim}{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+n_2} \frac{1}{\sigma_1^{n_1}} \frac{1}{\sigma_2^{n_2}} e^{-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(x_i - \mu_1)^2} e^{-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n_2}(y_i - \mu_2)^2}$$

$$\Theta = \{\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \mid \mu_1, \mu_2 \in I\!R, \sigma_1^2 > 0, \sigma_2^2 > 0\}$$

$$\Theta_0 = \{\underset{\sim}{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \mid \mu_1, \mu_2 \in I\!R, \sigma_2^2 = \tau\sigma_1^2\}$$

One finds :

$$\sup_{\underset{\sim}{\theta} \in \Theta} L(\underset{\sim}{\theta}; \underset{\sim}{x}, \underset{\sim}{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+n_2} \left(\frac{n_1 e^{-1}}{\sum_{i=1}^{n_1}(x_i - \overline{x})^2}\right)^{\frac{n_1}{2}} \left(\frac{n_2 e^{-1}}{\sum_{i=1}^{n_2}(y_i - \overline{y})^2}\right)^{\frac{n_2}{2}}$$

$$\sup_{\ell\in\Theta_0} L(\ell;\underset{\sim}{x},\underset{\sim}{y}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+n_2} \frac{1}{\tau^{\frac{n_2}{2}}} \left(\frac{(n_1+n_2)e^{-1}}{\sum_{i=1}^{n_1}(x_i-\overline{x})^2 + \frac{1}{\tau}\sum_{i=1}^{n_2}(y_i-\overline{y})^2}\right)^{\frac{n_1+n_2}{2}}$$

$$\lambda(\underset{\sim}{x},\underset{\sim}{y}) = \frac{(n_1+n_2)^{\frac{n_1+n_2}{2}}}{n_1^{\frac{n_1}{2}}\, n_2^{\frac{n_2}{2}}} \frac{\left(\frac{1}{\tau}\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}\right)^{\frac{n_2}{2}}}{\left(1+\frac{1}{\tau}\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}\right)^{\frac{n_1+n_2}{2}}}.$$

Or, with

$$f(\underset{\sim}{x},\underset{\sim}{y}) = \frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2/(n_2-1)}{\tau\sum_{i=1}^{n_1}(x_i-\overline{x})^2/(n_1-1)} :$$

$$\lambda(\underset{\sim}{x},\underset{\sim}{y}) = \frac{(n_1+n_2)^{\frac{n_1+n_2}{2}}}{n_1^{\frac{n_1}{2}}\, n_2^{\frac{n_2}{2}}} \frac{\left(\frac{n_2-1}{n_1-1}f(\underset{\sim}{x},\underset{\sim}{y})\right)^{\frac{n_2}{2}}}{\left(1+\frac{n_2-1}{n_1-1}f(\underset{\sim}{x},\underset{\sim}{y})\right)^{\frac{n_1+n_2}{2}}}.$$

The test rejects H_0 if and only if $\lambda(\underset{\sim}{x},\underset{\sim}{y})$ is small, i.e. if and only if

$$\frac{\left(\frac{n_2-1}{n_1-1}f(\underset{\sim}{x},\underset{\sim}{y})\right)^{\frac{n_2}{2}}}{\left(1+\frac{n_2-1}{n_1-1}f(\underset{\sim}{x},\underset{\sim}{y})\right)^{\frac{n_1+n_2}{2}}} \text{ is small.}$$

Since this, as a function of $f(\underset{\sim}{x},\underset{\sim}{y})$, has a single maximum, the test is given by : reject H_0 if and only if

$$f(\underset{\sim}{x},\underset{\sim}{y}) \le c_1 \qquad \text{or} \qquad f(\underset{\sim}{x},\underset{\sim}{y}) \ge c_2.$$

To construct a size-α test, c_1 and c_2 have to be determined from the fact that the statistic

$$F_n = f(X_1,\ldots,X_{n_1},Y_1,\ldots,Y_{n_2}) = \frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2/(n_2-1)}{\tau\sum_{i=1}^{n_1}(X_i-\overline{X})^2/(n_1-1)}$$

$$= \frac{\frac{\sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{\sigma_2^2}/(n_2-1)}{\frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2}{\sigma_1^2}/(n_1-1)} \qquad \text{(under } H_0 : \frac{\sigma_2^2}{\sigma_1^2}=\tau\text{)}$$

has a $F(n_2-1;n_1-1)$-distribution.

This is called an F**-test**.

4.4.3 Uniformly most powerful tests

The search for an optimum test is more difficult in the case of composite hypotheses. The reason is that, if Θ_1 contains more than one element, we cannot simply compare the **numbers** $\pi_\phi(\theta_1)$ and take that test with the largest such number, but we must compare **functions** $\pi_\phi(\theta)$, $\theta\in\Theta_1$. Their graphs may 'cross' and not yield an uniformly best one.

We define an optimum property that such a test may possess.

Definition

A test ϕ^* of $H_0 : \theta\in\Theta_0$ versus $H_1 : \theta\in\Theta_1=\Theta\backslash\Theta_0$ is a **uniformly most powerful (UMP) test of size** α $(0<\alpha<1)$ if

(i) $\sup_{\theta\in\Theta_0}\pi_{\phi^*}(\theta)=\alpha$

(ii) $\pi_{\phi^*}(\theta)\geq\pi_\phi(\theta)$, for all $\theta\in\Theta_1$, and for any other test ϕ with size $\leq\alpha$.

Hence : a test ϕ^* is UMP of size α if it has size α and if among all other tests of size $\leq\alpha$, it has the largest power function for **all** alternative values of θ.

Unfortunately it is not so often the case that a UMP test exists.

Example [Mean of normal with known variance]

$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$ with σ^2 known.
$H_0 : \mu = \mu_0$
$H_1 : \mu > \mu_0$.

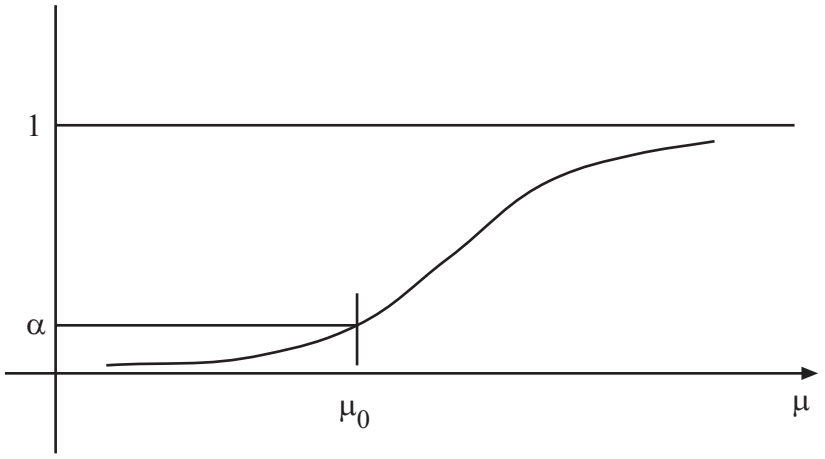
From a previous example we know that for every $\mu_1 > \mu_0$, the test with critical region

$$\left\{ \underset{\sim}{x} \mid \overline{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}} \right\}$$

is MP for testing $\mu = \mu_0$ againts $\mu = \mu_1$.
But, since we get the same MP test for any value $\mu_1 > \mu_0$, this test is also UMP.

The **power** of this test :

$$\begin{aligned}
\pi(\mu) &= P_\mu(\underset{\sim}{X} \in R) = P_\mu(\overline{X} \geq \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}) \\
&= P_\mu(\overline{X} - \mu \geq \mu_0 - \mu + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}) \\
&= P_\mu\left(\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \geq \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{n}}} + z_{1-\alpha} \right) \\
&= 1 - \Phi\left(\frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} + z_{1-\alpha} \right).
\end{aligned}$$



Example [Exponential]
$X_1, \ldots, X_n$: random sample from $X \sim Exp(\theta)$.
$H_0 : \theta = \theta_0$
$H_1 : \theta > \theta_0$.

From a previous example we know that for every $\theta_1 > \theta_0$, the test with critical region

$$\left\{ \underset{\sim}{x} \mid \sum_{i=1}^{n} x_i \leq c \right\}, \text{ with } c \text{ such that } \int_0^c \frac{\theta_0^n}{\Gamma(n)} x^{n-1} e^{-\theta_0 x} dx = \alpha$$

is MP for testing $\theta = \theta_0$ against $\theta = \theta_1$. Since this test does not depend on θ_1, it is also UMP.

A general **theorem** which sometimes can be used to find (one-sided) UMP tests is the following.

Theorem

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x;\theta)$, $\theta \in \Theta =$ some interval.

If the ratio $\dfrac{\prod_{i=1}^{n} f(x_i;\theta')}{\prod_{i=1}^{n} f(x_i;\theta'')}$ *is a* *nonincreasing* *[nondecreasing]* *function of $t(x_1, \ldots, x_n)$ for every $\theta' < \theta''$,*

then

- *UMP test of size α for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ has critical region $\{\underset{\sim}{x} \mid t(x_1, \ldots, x_n) > c\}$*

  [<]

  with $P_{\theta_0}(t(X_1, \ldots, X_n) > c) = \alpha$

  [<]

- *UMP test of size* α *for* $H_0 : \theta \geq \theta_0$ *versus* $H_1 : \theta < \theta_0$ *has critical region* $\{\underset{\sim}{x} \mid t(x_1, \ldots, x_n) \ < \ c\}$

[>]

with $P_{\theta_0}(t(X_1, \ldots, X_n) \ < \ c) = \alpha$.

[>]

Example [Exponential]

$$\frac{\prod_{i=1}^{n} f(x_i; \theta')}{\prod_{i=1}^{n} f(x_i; \theta'')} = \left(\frac{\theta'}{\theta''}\right)^n e^{-(\theta' - \theta'') \sum_{i=1}^{n} x_i}$$ is monotone nondecreasing in $\sum_{i=1}^{n} x_i$ for every $\theta' < \theta''$.

4.5 Summary of tests on the parameters of a normal distribution

In this section we give some tables which summarize the customary procedures for testing about mean and variance of a normal distribution (one sample problem) and comparison of means or variances (two sample problem). The tests are generalized likelihood tests (or slight modifications).

ONE SAMPLE PROBLEM
$X_1, \ldots, X_n$: random sample from $X \sim N(\mu; \sigma^2)$.

TESTS ABOUT THE MEAN

		σ^2 known	σ^2 unknown
H_0	H_1	reject H_0 if	
$\mu \leq \mu_0$	$\mu > \mu_0$	$\overline{x} \geq \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$	$\overline{x} \geq \mu_0 + t_{n-1,1-\alpha}\sqrt{\frac{s^2}{n-1}}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\overline{x} \leq \mu_0 + z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}}$	$\overline{x} \leq \mu_0 + t_{n-1,1-\alpha}\sqrt{\frac{s^2}{n-1}}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert\overline{x} - \mu_0\rvert \geq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}$	$\lvert\overline{x} - \mu_0\rvert \geq t_{n-1,1-\frac{\alpha}{2}}\sqrt{\frac{s^2}{n-1}}$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

TESTS ABOUT THE VARIANCE

		μ known	μ unknown
H_0	H_1	reject H_0 if	
$\sigma \leq \sigma_0$	$\sigma > \sigma_0$	$\sum_{i=1}^{n}(x_i - \mu)^2 \geq \chi^2_{n,1-\alpha}\sigma_0^2$	$s^2 \geq \chi^2_{n-1,1-\alpha}\frac{\sigma_0^2}{n}$
$\sigma \leq \sigma_0$	$\sigma < \sigma_0$	$\sum_{i=1}^{n}(x_i - \mu)^2 \leq \chi^2_{n,\alpha}\sigma_0^2$	$s^2 \leq \chi^2_{n-1,\alpha}\frac{\sigma_0^2}{n}$
$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$\sum_{i=1}^{n}(x_i - \mu)^2 \leq \chi^2_{n,\alpha/2}\sigma_0^2$ or $\sum_{i=1}^{n}(x_i - \mu)^2 \geq \chi^2_{n,1-\alpha/2}\sigma_0^2$	$s^2 \leq \chi^2_{n-1,\alpha/2}\frac{\sigma_0^2}{n}$ or $s^2 \geq \chi^2_{n-1,1-\alpha/2}\frac{\sigma_0^2}{n}$

TWO SAMPLE PROBLEM

Independent samples : $X_1, \ldots, X_{n_1}$: from $X \sim N(\mu_1; \sigma_1^2)$
$Y_1, \ldots, Y_{n_2}$: from $Y \sim N(\mu_2; \sigma_2^2)$.

TESTS ABOUT THE DIFFERENCE OF THE MEANS

H_0	H_1	σ_1^2, σ_2^2 known	σ_1^2, σ_2^2 unknown, $\sigma_1^2 = \sigma_2^2$
(δ : known)		reject H_0 if	
$\mu_2 - \mu_1 \leq \delta$	$\mu_2 - \mu_1 > \delta$	$\overline{y} - \overline{x} \geq \delta + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\overline{y} - \overline{x} \geq \delta + t_{n_1+n_2-2,1-\alpha}.s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$\mu_2 - \mu_1 \geq \delta$	$\mu_2 - \mu_1 < \delta$	$\overline{y} - \overline{x} \leq \delta - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\overline{y} - \overline{x} \leq \delta - t_{n_1+n_2-2,1-\alpha}.s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$\mu_2 - \mu_1 = \delta$	$\mu_2 - \mu_1 \neq \delta$	$\lvert\overline{y} - \overline{x} - \delta\rvert \geq z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\lvert\overline{y} - \overline{x} - \delta\rvert \geq t_{n_1+n_2-2,1-\frac{\alpha}{2}}.s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \overline{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \quad s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2 \quad s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i - \overline{y})^2 \quad s_p^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

If the variances σ_1^2 and σ_2^2 are possibly unequal **(Behrens - Fisher** problem), then one of the solutions is **Welch's** t**-test** with rejection regions of the form :

$$\overline{y} - \overline{x} \geq \delta + t_{\overset{\wedge}{\nu},1-\alpha}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

$$\overline{y} - \overline{x} \leq \delta - t_{\overset{\wedge}{\nu},1-\alpha}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

$$\lvert\overline{y} - \overline{x} - \delta\rvert \geq t_{\overset{\wedge}{\nu},1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}}$$

with $\overset{\wedge}{\nu} = \dfrac{\left(\frac{s_1}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}\right)^2}{\frac{1}{n_1}\left(\frac{s_1^2}{n_1 - 1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2 - 1}\right)^2}$

rounded up to nearest integer

TESTS ABOUT THE RATIO OF THE VARIANCES

H_0	H_1	μ_1, μ_2 known	μ_1, μ_2 unknown
(τ : known)		reject H_0 if	
$\frac{\sigma_2^2}{\sigma_1^2} \leq \tau$	$\frac{\sigma_2^2}{\sigma_1^2} > \tau$	$\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2/n_2}{\tau\sum_{i=1}^{n_1}(x_i-\overline{x})^2/n_1} \geq F_{n_2,n_1;1_\alpha}$	$\frac{\frac{n_2}{n_2-1}s_2^2}{\tau\frac{n_1}{n_1-1}s_1^2} \geq F_{n_2-1,n_1-1;1-\alpha}$
$\frac{\sigma_2^2}{\sigma_1^2} \geq \tau$	$\frac{\sigma_2^2}{\sigma_1^2} < \tau$	$\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2/n_2}{\tau\sum_{i=1}^{n_1}(x_i-\overline{x})^2/n_1} \leq F_{n_2,n_1;\alpha}$	$\frac{\frac{n_2}{n_2-1}s_2^2}{\tau\frac{n_1}{n_1-1}s_1^2} \leq F_{n_2-1,n_1-1;1-\alpha}$
$\frac{\sigma_2^2}{\sigma_1^2} = \tau$	$\frac{\sigma_2^2}{\sigma_1^2} \neq \tau$	$\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2/n_2}{\tau\sum_{i=1}^{n_1}(x_i-\overline{x})^2/n_1} \leq F_{n_2,n_1;\alpha/2}$ or $\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2/n_2}{\tau\sum_{i=1}^{n_1}(x_i-\overline{x})^2/n_1} \geq F_{n_2,n_1;1-\alpha/2}$	$\frac{\frac{n_2}{n_2-1}s_2^2}{\tau\frac{n_1}{n_1-1}s_1^2} \leq F_{n_2-1,n_1-1-\alpha/2}$ or $\frac{\frac{n_2}{n_2-1}s_2^2}{\tau\frac{n_1}{n_1-1}s_1^2} \geq F_{n_2-1,n_1-1;1-\alpha/2}$

Warning : the procedures for variances are very sensitive to departures from the normality assumptions.

4.6 Comparing several means

We assume that we have available k random samples, one from each k normal populations; that is, Suppose $X_{11}, \ldots, X_{jn_j}$ be a random sample from the j^{th} normal population. For $j = 1, 2, \ldots, k$. Assume that the j^{th} population has mean μ_j and variance σ^2. Further assume that the k random samples are independent.

Our objective is to test the null hypothesis that all the population means are equal versus the alternative proposes that not all the means are equal. That is:
H_0: μ_1=$\mu 2$=$\ldots$=μ_k versus
H_1: Not all $\mu j's$ are equal.
To test the proposed hypothesis, we consider generalized likelihood ratio-test.

$$L\left(\mu_1, \ldots, \mu_k, \sigma^2; x_{11}, \ldots, x_{1n_1}, \ldots, x_{k1}, \ldots, x_{kn_k}\right)$$

$$= \prod_{j=1}^{k} \prod_{i=1}^{n_j} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}\left[\frac{\left(x_{ji}-\mu_j\right)}{\sigma_j}\right]^2}$$

$$= \left(2\pi\sigma^2\right)^{\frac{-n}{2}} e^{\left[\frac{-1}{2\sigma^2}\sum_j \sum_i (x_{ji}-\mu j)^2\right]}$$

where

$$n = \sum_{j=1}^{k} n_j$$

The maximum likelihood estimates of $\mu_1, \ldots, \mu_k, \sigma^2$ are given by;

$$\widehat{\mu_j} = \bar{x_{j.}} = \frac{1}{n} \sum_{i=1}^{n_j}, j = 1, 2, \ldots, k$$

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \bar{x}_{j.})^2$$

Hence,

$$L_0 = \left[\frac{2\pi \sum\limits_j \sum\limits_i (x_{ji} - \bar{x})^2}{n}\right]^{\frac{-n}{2}} e^{-n}$$

$$L_1 = \left[\frac{2\pi \sum\limits_j \sum\limits_i (x_{ji} - \bar{x}_{j.})^2}{n}\right]^{\frac{-n}{2}} e^{-n}$$

The generalized likelihood ratio test is

$$\lambda = \frac{L_0}{L_1}$$

$$= \left[\frac{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x})^2}{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x}_{j.})^2}\right]^{\frac{-n}{2}}$$

$$= \left[\frac{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x}_j + \bar{x}_j - \bar{x})^2}{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x})^2}\right]^{\frac{-n}{2}}$$

$$= \left[\frac{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x}_j)^2 + \sum\limits_j n_j (\bar{x}_j - \bar{x})^2}{\sum\limits_j \sum\limits_i (x_{ji} - \bar{x})^2}\right]^{\frac{-n}{2}}$$

$$= \left[1 + \frac{k-1}{n-k} \frac{\frac{\sum_j n_j(\bar{x}_j - \bar{x})^2}{k-1}}{\frac{\sum_j \sum_i (x_{ji} - \bar{x}_j)^2}{n-k}}\right]^{\frac{-n}{2}}$$

$$= \left[1 + \frac{k-1}{n-k} r\right]^{\frac{-n}{2}}$$

where

$$r = \frac{\frac{\sum_j n_j(\bar{x}_j - \bar{x})^2}{k-1}}{\frac{\sum_j \sum_i (x_{ji} - \bar{x}_j)^2}{n-k}}$$

A generalized likelihood-ratio test is described as follows:
Reject H_0 if $\lambda \leq k$, but $\lambda \leq k$ if and only if,
$r \geq c$, where c is some constant.

Note The ratio r is sometimes referred to as the variance ratio or F-ratio.
The constant c is determined so that the test will have size α.
The two quantities both in numerator and denominator of r are independent and if we divide both by the common variance, we have the ratio of two independent chi-square distribution which gives us the F-distribution.
As a result we can choose the constant c is $(1-\alpha)^{th}$
quantile of the F-distribution with $(k-1)$ and $(n-k)$ degrees of freedom.
Therefore, for large value of F, we have small value of the likelihood ratio, that means they are inversely related. As a result we reject the null hypothesis for small value of λ or for large value of F.

4.7 The relationship between two-sided tests of hypotheses and confidence interval

Suppose $X_1, \ldots, X_n$, be a random sample from a normal population and consider testing
H_0: $\mu = \mu_0$ versus H_1: $\mu \neq \mu_0$
The test can be carried out at specified level of significance, say α percent. If we have

given that the variance is known, then reject H_0

$$\begin{aligned}
&\Leftrightarrow \quad \left|\frac{\overline{x}-\mu_0}{\sqrt{\sigma^2/n}}\right| \geq z_{1-\alpha/2} \\
&\Leftrightarrow \quad \frac{\overline{x}-\mu_0}{\sqrt{\sigma^2/n}} \geq z_{1-\alpha/2} \text{ or } \frac{\overline{x}-\mu_0}{\sqrt{\sigma^2/n}} \leq z_{1-\alpha/2} \\
&\Leftrightarrow \quad \mu_0 \leq \overline{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \text{ or } \mu_0 \geq \overline{x} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \\
&\Leftrightarrow \quad \mu_0 \text{ outside } 100(1-\alpha)\% \text{ the confidence interval.}
\end{aligned}$$

If the variance σ^2 is unknown, then reject H_0

$$\begin{aligned}
&\Leftrightarrow \quad \left|\frac{\overline{x}-\mu_0}{\sqrt{s^2/n-1}}\right| \geq t_{n-1;1-\alpha/2} \\
&\Leftrightarrow \quad \frac{\overline{x}-\mu_0}{\sqrt{s^2/n-1}} \geq t_{n-1;1-\alpha/2} \text{ or } \frac{\overline{x}-\mu_0}{\sqrt{s^2/n-1}} \leq t_{n-1;1-\alpha/2} \\
&\Leftrightarrow \quad \mu_0 \leq \overline{x} - t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n-1}} \text{ or } \mu_0 \geq \overline{x} - t_{n-1;1-\alpha/2}\sqrt{\frac{s^2}{n-1}} \\
&\Leftrightarrow \quad \mu_0 \text{ outside } 100(1-\alpha)\% \text{ the confidence interval.}
\end{aligned}$$

Also,consider testing H_0:$\sigma^2 = \sigma_0^2$ versus H_1: $\sigma^2 \neq \sigma_0^2$
Suppose μ is unknown, we reject H_0

$$\begin{aligned}
&\Leftrightarrow \quad \frac{ns^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha/2} \text{ or } \frac{ns^2}{\sigma_0^2} \geq \chi^2_{n-1;1-\alpha/2} \\
&\Leftrightarrow \quad \sigma_0^2 \geq \frac{ns^2}{\chi^2_{n-1;\alpha/2}} \text{ or } \sigma_0^2 \leq \frac{ns^2}{\chi^2_{n-1;1-\alpha/2}} \\
&\Leftrightarrow \quad \sigma_0^2 \text{ outside } 100(1-\alpha)\% \text{ the confidence interval.}
\end{aligned}$$

Note The usefulness of the strong relationship between two-sided tests of hypothesis and confidence sets is one can be used to construct the other but also in the result that often an optimal property of one carries over the other. That is, if one can find a test that is optimal in some sense, then the corresponding constructed confidence is also optimal in some sense.

4.8 Large sample distribution of generalized likelihood ratio

The exact distribution of the generalized likelihood ratio statistic Λ_n is usually difficult to obtain. An example where the exact distribution of Λ_n is known is that of the normal distribution $N(\theta;\sigma^2)$, with σ^2 known. For the null hypothesis $H_0 : \theta = \theta_0$ we found (see before)

$$\Lambda_n = e^{-\frac{1}{2}\left(\frac{\overline{X}-\theta_0}{\sqrt{\sigma^2/n}}\right)^2}$$

Hence,

$$-2ln\Lambda_n = \left(\frac{\overline{X} - \theta_0}{\sqrt{\sigma^2/n}}\right)^2$$

and, under H_0, this is exactly $\chi^2(1)$-distributed.
It turns out that, in general, the statistic

$$D_n = -2ln\Lambda_n$$

is more convenient for asymptotic considerations.

Theorem

Let $X_1, \ldots, X_n$ be a random sample from X with density $f(x; \underset{\sim}{\theta})$ where $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$ and Θ is a k-dimensional subset of $\mathbb{R}^k$.
Consider the null hypothesis

$$H_0 : \theta_1 = \theta_1^0, \theta_2 = \theta_2^0, \ldots, \theta_r = \theta_r^0$$

where $\theta_1^0, \ldots, \theta_r^0$ are known constants $(1 \leq r \leq k)$ and where $\theta_{r+1}, \ldots, \theta_k$ are unspecified.
Under regularity conditions, we have under H_0, as $n \to \infty$:

$$D_n = -2ln\Lambda_n \xrightarrow{d} \chi^2(r)$$

Note that the number of degrees of freedom of the limiting χ^2-distribution equals **the number of parameters specified by** H_0.

'Proof'

(We only give a sketch of the proof in the one parameter case ($k = 1$) and for a simple null hypothesis $H_0 : \theta = \theta_0$)).
In this case (with notations from ML theory) :

$$\lambda(\underset{\sim}{x}) = \frac{L(\theta_0; \underset{\sim}{x})}{L(\hat{\theta}_n; \underset{\sim}{x})}$$

where $\hat{\theta}_n = t(x_1, \ldots, x_n)$ is the ML estimate for θ.
Hence :

$$-2ln\lambda(\underset{\sim}{x}) = 2[l(\hat{\theta}_n; \underset{\sim}{x}) - l(\theta_0; \underset{\sim}{x})].$$

Taylor expansion of $l(\theta_0; \underset{\sim}{x})$ around $l(\hat{\theta}_n; \underset{\sim}{x})$ gives :

$$l(\theta_0; \underset{\sim}{x}) - l(\hat{\theta}_n; \underset{\sim}{x}) \approx (\theta_0 - \hat{\theta}_n)S(\hat{\theta}_n; \underset{\sim}{x}) - \frac{1}{2}(\theta_0 - \hat{\theta}_n)^2\mathcal{I}(\hat{\theta}_n; \underset{\sim}{x}).$$

Since $S(\hat{\theta}_n; \underset{\sim}{x}) = 0$:

$$-2ln\lambda(\underset{\sim}{x}) \approx (\hat{\theta}_n - \theta_0)^2\mathcal{I}(\hat{\theta}_n; \underset{\sim}{x}).$$

From this one can prove (with $T_n = t(X_1, \ldots, X_n)$ the ML estimator for θ) :

$$\begin{aligned} D_n &\approx (T_n - \theta_0)^2\mathcal{I}(T_n; \underset{\sim}{X}) \\ &\approx (T_n - \theta_0)^2\mathcal{I}(\theta_0; \underset{\sim}{X}) \\ &= \left(\frac{\sqrt{n}(T_n - \theta_0)}{\sqrt{\frac{1}{i(\theta_0)}}} \right)^2 \frac{1}{i(\theta_0)} \frac{\mathcal{I}(\theta_0; \underset{\sim}{X})}{n}. \end{aligned}$$

By the asymptotic normality of the ML estimator

$$\left(\frac{\sqrt{n}(T_n - \theta_0)}{\sqrt{\frac{1}{i(\theta_0)}}} \right)^2 \xrightarrow{d} \chi^2(1)$$

Also :

$$\frac{\mathcal{I}(\theta_0; \underset{\sim}{X})}{n} \xrightarrow{P} i(\theta_0).$$

Hence, by Slutsky's theorem : $D_n \xrightarrow{d} \chi^2(1)$. □

We now give some **examples** of the D_n-statistic in the case of a simple null hypothesis $H_0 : \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$. We have :

$$D_n = 2[l(\underset{\sim}{T}_n; \underset{\sim}{X}) - l(\underset{\sim}{\theta}_0, \underset{\sim}{X})]$$

where $l = lnL$ is the loglikelihood function and $\underset{\sim}{T}_n$ is the ML estimator for $\underset{\sim}{\theta}$.

Example [Normal]

$X \sim N(\theta; \sigma^2)$, σ^2 known.
$H_0 : \theta = \theta_0$

$$L(\theta; \underset{\sim}{x}) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

$$l(\theta; \underset{\sim}{x}) = -nln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2$$

$$T_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$D_n = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(X_i-\theta_0)^2 - \sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \frac{n}{\sigma^2}(\overline{X}-\theta_0)^2$$

$$D_n = \left(\frac{\overline{X}-\theta_0}{\sqrt{\frac{\sigma^2}{n}}}\right)^2$$

Example [Binomial]

$X \sim B(1; \theta)$
$H_0 : \theta = \theta_0$

$$L(\theta; \underset{\sim}{x}) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

$$l(\theta; \underset{\sim}{x}) = (\textstyle\sum x_i) ln\theta + (n - \sum x_i) ln(1-\theta)$$

$$T_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$D_n = 2n\left[ln\left(\frac{\overline{X}}{\theta_0}\right)\overline{X} + (1-\overline{X})ln\left(\frac{1-\overline{X}}{1-\theta_0}\right)\right].$$

Example [Poisson]

$X \sim P(\theta)$
$H_0 : \theta = \theta_0$

$$l(\theta; \underset{\sim}{x}) = -ln(x_1! \ldots x_n!) - n\theta + (\textstyle\sum x_i) ln\theta$$

$T_n = \overline{X}$
$D_n = 2n\left[\overline{X} ln\left(\frac{\overline{X}}{\theta_0}\right) + \theta_0 - \overline{X}\right].$

Example [Multinomial]

$\underset{\sim}{X} \sim M(n; (\theta_1, \ldots, \theta_k))$
$H_0 : \theta_i = \theta_i^0 \ , i = 1, \ldots, k$
$l(\theta_1, \ldots, \theta_k; \underset{\sim}{x}) = ln\left(\frac{n!}{x_1! \ldots x_k!}\right) + \sum_{i=1}^{k} x_i ln\theta_i$
$\underset{\sim}{T}_n = \left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}\right)$
$D_n = 2\left[l\left(\frac{X_1}{n}, \ldots, \frac{X_k}{n}; \underset{\sim}{X}\right) - l(\theta_1^0, \ldots, \theta_k^0; \underset{\sim}{X})\right]$
$D_n = 2\sum_{i=1}^{k} X_i ln\left(\frac{X_i}{n\theta_i^0}\right)$

Two other statistics related to $D_n = -2ln\Lambda_n$

In the one parameter case we had

$$D_n \approx (T_n - \theta_0)^2 \mathcal{I}(T_n; \underset{\sim}{X})$$

In the k-parameter case, this becomes

$$D_n \approx (\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)\mathcal{I}(\underset{\sim}{T}_n; \underset{\sim}{X})(\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)'$$

where $\underset{\sim}{T}_n$ is the ML estimator and $\mathcal{I}(\underset{\sim}{\theta}; \underset{\sim}{x})$ is the information matrix.

- Replace the elements of the matrix $\mathcal{I}(\underset{\sim}{\theta}; \underset{\sim}{X})$ by their expected values. This amounts in replacing $\mathcal{I}(\underset{\sim}{\theta}; \underset{\sim}{X})$ by $nB(\underset{\sim}{\theta})$, where $B(\underset{\sim}{\theta})$ is the Fisher-information matrix.

  This leads to

  $$W_n = n(\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)B(\underset{\sim}{T}_n)(\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)'$$

  which is the statistic, introduced by **Wald**.

- Recall that the ML estimate $\hat{\underset{\sim}{\theta}}_n$ satisfies the equations $S(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{x}) = \underset{\sim}{0}$, i.e.

  $$(S_1(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{x}), \ldots, S_k(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{x})) = (0, \ldots, 0).$$

For $j = 1, \ldots, k$, we have by multivariate Taylor expansion :

$$\begin{aligned} 0 &= S_j(\hat{\underset{\sim}{\theta}}_n; \underset{\sim}{x}) \\ &\approx S_j(\underset{\sim}{\theta}_0; \underset{\sim}{x}) + \sum_{i=1}^{k} (\hat{\theta}_{ni} - \theta_{0i}) \frac{\partial S_j}{\partial \theta_i}(\underset{\sim}{\theta}_0; \underset{\sim}{x}) \\ &= S_j(\underset{\sim}{\theta}_0; \underset{\sim}{x}) - \sum_{i=1}^{k} (\hat{\theta}_{ni} - \theta_{0i}) \mathcal{I}_{ij}(\underset{\sim}{\theta}_0; \underset{\sim}{x}) \end{aligned}$$

where $\mathcal{I}(\theta_0; \underset{\sim}{x}) = [\mathcal{I}_{ij}(\underset{\sim}{\theta}_0; \underset{\sim}{x})]_{i,j=1,\ldots,k}$ is the information matrix.
Hence, for $j = 1, \ldots, k$:

$$S_j(\underset{\sim}{\theta}_0; \underset{\sim}{x}) \approx \sum_{i=1}^{k} (\hat{\theta}_{ni} - \theta_{0i}) \mathcal{I}_{ij}(\underset{\sim}{\theta}_0; \underset{\sim}{x}).$$

Thus :

$$S(\underset{\sim}{\theta}_0; \underset{\sim}{x}) \approx (\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \mathcal{I}(\underset{\sim}{\theta}_0; \underset{\sim}{x}).$$

Or :

$$\hat{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0 \approx S(\underset{\sim}{\theta}_0; \underset{\sim}{x}) \mathcal{I}^{-1}(\underset{\sim}{\theta}_0; \underset{\sim}{x}).$$

From this one can prove

$$\underset{\sim}{T}_n - \underset{\sim}{\theta}_0 \approx S(\underset{\sim}{\theta}_0; \underset{\sim}{X}) \mathcal{I}^{-1}(\underset{\sim}{\theta}_0; \underset{\sim}{X}).$$

Then :

$$\begin{aligned} D_n &\approx (\underset{\sim}{T}_n - \underset{\sim}{\theta}_0) \mathcal{I}(\underset{\sim}{T}_n; \underset{\sim}{X}) (\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)' \\ &\approx (\underset{\sim}{T}_n - \underset{\sim}{\theta}_0) \mathcal{I}(\underset{\sim}{\theta}_0; \underset{\sim}{X}) (\underset{\sim}{T}_n - \underset{\sim}{\theta}_0)' \\ &\approx S(\underset{\sim}{\theta}_0; \underset{\sim}{X}) \mathcal{I}^{-1}(\underset{\sim}{\theta}_0; \underset{\sim}{X}) S'(\underset{\sim}{\theta}_0; \underset{\sim}{X}). \end{aligned}$$

This is the statistic introduced by **C.R. Rao** :

$$V_n = \frac{1}{n} S(\underset{\sim}{\theta}_0; \underset{\sim}{X}) B^{-1}(\underset{\sim}{\theta}_0) S'(\underset{\sim}{\theta}_0; \underset{\sim}{X}).$$

Note that the computation of this statistic does not require computation of the ML-estimator.

Theorem

Under regularity conditions, we have under $H_0 : \underset{\sim}{\theta} = \underset{\sim}{\theta}_0$, *as* $n \to \infty$:

$$D_n, W_n \text{and } V_n \text{ each converge in distribution to } \chi^2(k).$$

Example [Multinomial]

$\underset{\sim}{X} = (X_1, \ldots, X_k) \sim M(n; (\theta_1, \ldots, \theta_k))$
The 3 statistics for testing $H_0 : \theta_i = \theta_i^0$ $(i = 1, \ldots, k)$ can be calculated :

$$\begin{aligned} D_n &= 2\sum_{i=1}^{k} X_i ln\left(\frac{X_i}{n\theta_i^0}\right) \\ W_n &= \sum_{i=1}^{k} \frac{(X_i - n\theta_i^0)^2}{X_i} \\ V_n &= \sum_{i=1}^{k} \frac{(X_i - n\theta_i^0)^2}{n\theta_i^0}. \end{aligned}$$

Each of these statistics has $\chi^2(k-1)$ as limiting distribution (there are only $k-1$ functionally independent parameters, since $\theta_1 \ldots + \theta_k = 1$).
The statistics W_n and V_n are often called **goodness of fit** statistics. V_n is known as the **Pearson statistic**.

4.9 Hypothesis testing using R

Some of the R codes discussed in this section are used to test hypothesis as well as to construct confidence intervals

Test on the mean of a normal population

```
> ##Testing a mean of a normal population
> ## Assume that the observations are taken from a normal population
>
> ## Compute the t statistic. Note we assume mu=25 under H_0 and mu<25
> ##under the alternative
> xbar=22;s=1.5;n=10
> t = (xbar-25)/(s/sqrt(n))
> t
[1] -6.324555
> ## use pt to get the distribution function of t
> pt(t,df=n-1)
[1] 6.846828e-05
> ##This is a small p-value (0.000068). Thus H_0 is rejected.
```

Confidence interval for proportion using in-built R function Note that this R function is also used for constructing confidence intervals for proportions

```
> ##Proportion test. In-built test
> #One sample problem
> #General form:
> #prop.test(x,n,p=NULL,alternative=c("two.sided","less","greater",
> #conf.level=1-alpha,correct=TRUE)
> #x=a vector of counts of successes or a matrix with 2 columns giving
> # the counts of successes and failures, respectively.
> #a vector of counts of trials
> #a vector of probabilities of success
> #alternative= a character string specifying the alternative hypothesis
> #conf.level =confidence level of the returned confidence interval
> #correct a logical indicating whether Yates continuity correction
> # should be applied.
> #Consider a simple survey. You ask 100 people (randomly chosen)
> #and 42 say "yes" to your question. Does this
> #support the hypothesis that the true proportion is 50%?
> prop.test(42,100,p=.5)
1-sample proportions test with continuity correction
data: 42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.1336
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.3233236 0.5228954
sample estimates:
p
0.42
> #Note the p-value of 0.1336.
> #The p-value reports how likely we are to see this data or
> #worse assuming the null hypothesis.
> #Now, the p-value is not so small as to make an observation
> #of 42 seem unreasonable in 100 samples assuming the
> #null hypothesis. Thus, one would "accept" the null hypothesis.
> #Note also that this R-code provided a 95% confidence interval
> #for the proportion of "yes" answer.

> ##We can also make one sided tests and construct
#one sided confidence interval using the following R commands
> prop.test(42,100,p=.5,alternative="less")
1-sample proportions test with continuity correction
data: 42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.06681
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
0.0000000 0.5072341
sample estimates:
p
0.42
> prop.test(42,100,p=.5,alternative="greater")
1-sample proportions test with continuity correction
data: 42 out of 100, null probability 0.5
X-squared = 2.25, df = 1, p-value = 0.9332
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
0.3372368 1.0000000
sample estimates:
p
```

Two-sample tests and confidence intervals of mean

```
> ##Two-sample tests and confidence interval
# using built-in R function.
> # Equal variances is assumed.
> #Suppose the recovery time for patients taking a new
# drug is measured (in days). A placebo group is also used
> #to avoid the placebo effect. The data are as follows
> #with drug: 15 10 13 7 9 8 21 9 14 8
> #placebo: 15 14 12 8 14 7 16 10 15 12
> #A one-sided test for equivalence of means using the t-test is needed.
> #This tests the H_0 of equal means against H_1 that the drug group
> # has a smaller mean. (mu1-mu2 < 0).
> x=c(15,10,13,7,9,8,21,9,14,8)
> y=c(15,14,12,8,14,7,16,10,15,12)
> t.test(x,y,alt="less",var.equal=TRUE)
Two Sample t-test
data: x and y
t = -0.5331, df = 18, p-value = 0.3002
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf 2.027436
sample estimates:
mean of x mean of y
11.4 12.3
> #We fail to reject H_0 based on this test.
> # Note also the one sided upper 95% confidence interval for mu1-mu2.
> #Instead of assuming the equality of the variances we can test for
> #their equality
> qf(0.975,9,9)
[1] 4.025994
> var(x)
[1] 18.93333
> var(y)
[1] 9.566667
> F.ratio<-var(x)/var(y)
> F.ratio
[1] 1.979094
> #Since F.ratio is smaller than the tabulated value of F we "accept"
> #the hypothesis that the two variances are equal and proceed with
> #the previous test.
```

Two-sample t-test when the variances are not equal

```
>#  Two sample t-test when the variances are not equal.
#Consider now the following problem:
> #Let us consider the following data on ozone levels (in ppm)
> # taken on 10 days in two market gardens. Assuming that the
> #observations are taken from normal populations we wish to compare
> # the population means.
> gardenA <- c(3,4,4,3,2,3,1,3,5,2)
>
> gardenB <- c(5,5,6,7,4,4,3,5,6,5)
> #This time let us compare the population variances using built
> # in function of R.
> var.test (gardenA, gardenC)
F test to compare two variances
data: gardenA and gardenC
F = 0.0938, num df = 9, denom df = 9, p-value = 0.001624
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.02328617 0.37743695
sample estimates:
ratio of variances
0.09375
> #In this case the variances are not equal. Hence we can not use
> #the previous method. Instead we use two-sample t-test developed
# by Welch.
> #This test is an approximate solution to the Behrens-Fisher problem.
> t.test(gardenA,gardenC)
Welch Two Sample t-test
data: gardenA and gardenC
t = -1.6036, df = 10.673, p-value = 0.1380
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.7554137 0.7554137
sample estimates:
mean of x mean of y
3 5
> #Conclusion: The ozone levels in this two gardens are the "same".
## One sided test and confidence interval (Two sample problem)
> prop.test(c(45,56),c(45+35,56+47),alternative="less")
2-sample test for equality of proportions with
continuity correction
data: c(45, 56) out of c(45 + 35, 56 + 47)
X-squared = 0.0108, df = 1, p-value = 0.5414
alternative hypothesis: less
95 percent confidence interval:
-1.0000000 0.1517323
sample estimates:
prop 1 prop 2
0.5625000 0.5436893
> prop.test(c(45,56),c(45+35,56+47),alternative="greater")
2-sample test for equality of proportions with
continuity correction
data: c(45, 56) out of c(45 + 35, 56 + 47)
X-squared = 0.0108, df = 1, p-value = 0.4586
alternative hypothesis: greater
95 percent confidence interval:
-0.1141109 1.0000000
```

Two-sample tests and confidence interval of proportion

```
> ##Two-sample tests and confidence interval of proportion .
> #General form:
> #prop.test(c(x1,x2),c(n1-x1,n2-x2),p=NULL,alternative=
> #c("two.sided","less","greater",conf.level=1-alpha,correct=TRUE)
> #where x1= the number of success in the first sample
> # x2=the number of success in the second sample
> #A survey is taken two times over the course of two weeks.
> #The first week you asked 80 people (randomly chosen)
> #and 45 say "yes" to your question.
> #The second week you asked the same question 103 people
> # and 56 answered "yes".
> #The standard hypothesis test is H0 : P1 equal P2 against the
> # alternative (two-sided) H1 : P1 not equal P2
> prop.test(c(45,56),c(45+35,56+47))
2-sample test for equality of proportions with
continuity correction
data: c(45, 56) out of c(45 + 35, 56 + 47)
X-squared = 0.0108, df = 1, p-value = 0.9172
alternative hypothesis: two.sided
95 percent confidence interval:
-0.1374478 0.1750692
sample estimates:
prop 1 prop 2
0.5625000 0.5436893
```

4.10 Exercises

1. Let X have a Bernoulli distribution, where $P[X = 1]$=θ, $P(X = 0)$=1-θ, for a random sample of size n=10, test $H_0 : \theta \leq \frac{1}{2}$ versus $H_1 : \theta > \frac{1}{2}$. Use the critical region $\sum_{i=1}^{n} x_i \geq 6$, find the power function and the size of the test.

2. Let X be a single observation from the density;

$$f(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x)$$

Find the generalized likelihood ratio test of size α of $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$.

3. Let $X_1, \ldots, X_n$ be a random sample from the poisson distribution,

$$f_{x;\theta} = \frac{e^{-\theta}(\theta)^x}{x!}, x = 0, 1, 2, \ldots$$

(a) Find the UMP test of $H_0 : \theta = \theta_0$ versus $H_0 : \theta > \theta_0$

(b) Test $H_0 : \theta = \theta_0$ versus $H_0 : \theta \neq \theta_0$
Find the general form of the critical region corresponding to the test arrived at using the generalized likelihood ratio principle. (The critical region should be defined in terms of $\sum_i X_i$)

4. Let $X_1, \ldots, X_m$ be a random sample from the density $\theta_1 x^{\theta_1 - 1} I_{(0,1)}(x)$
and $Y_1, \ldots, Y_n$ be a random sample from the density $\theta_2 x^{\theta_2 - 1} I_{(0,1)}(y)$
Assume that the samples are independent. Set $U_i = -lne^{X_i}, i = 1, 2, \ldots, m$ and $V_j = -lne^{Y_j}, j = 1, 2, \ldots, n$.

(a) Find the generalized likelihood ratio for testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$

(b) Show that the generalized likelihood ratio test can be expressed in terms of the statistic

$$T = \frac{\sum_i U_i}{\sum_i U_i + \sum_j V_j}$$

5. Find the generalized likelihood ratio test of size α for testing $H_0 : \theta \leq 1 versus H_1 : \theta >$ 1 on the basis of a random sample $X_1, \ldots, X_n$ from $f(x;\theta) = \theta e^{-\theta x} I_{(0,\infty)}(x)$.

6. Let X be a single observation from the density $f(x;\theta) = (1+\theta)x^{\theta} I_{(0,1)}(x)$ where $\theta > -1$.

(a) Find the most powerful size α test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$

(b) Is there a uniformly most powerful size-α test of $H_0 < 0$ versus $H_1 : \theta > 0$? If so what is it?

7. Let $X_1, \ldots, X_n$ be a random sample from $\theta_1 e^{-\theta x} I_{(0,\infty)}(x)$ and let $Y_1, \ldots, Y_n$ be a random sample from $\theta_2 e^{-\theta y} I_{(0,\infty)}(y)$. Assume that the two samples are independent.

(a) find the generalized likelihood ration for testing $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$.

(b) Show that the generalized ratio test can be expressed in terms of the statistic $T = \frac{\sum_i X_i}{\sum_i X_i + \sum_j Y_j}$

8. Use the confidence interval technique to derive a test of $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ in sampling from the bivariate normal distribution. Such a test is often called paired t-test.

9. Given the sample -4,4,4,2,-4.8 from a normal population with variance 4, and the sample 6,1,3.2,-0.4 from a normal population with variance 5, test at 0.05 level that the means differ by no more than one unit.

10. A metallurgist made four determination of the melting point of manganese:1269,1271,1263, and 1265 degree centigrade. Test the hypothesis that the mean μ this of population is with in 5 degree centigrade of the published value of 1260.(use $\alpha = 0.05$,assume normality and $\sigma^2 = 5$)

11. Let $X_1, \ldots, X_n$ be a random sample of size n from a normal density with known variance. what is the best critical region for testing the null hypothesis that the mean is 6 against the mean is 4?

12. Derive a test of $H_0 : \sigma^2 < 10$ against $H_1 : \sigma^2 \geq 10$ for a sample of size n from a normal population with a mean of zero.

13 A cigarette manufacturer sent each of two laboratories presumably identical samples of tobacco. each made five determination of the nicotine content in milligrams as follows: (i) 24,27,26,21,and 24 (ii) 27,28,23,31 and 26. Were the two laboratories measuring the same thing? (Assume normality and common variance).

14. Given the samples 1.8,2.9,1.4,1.1 and 5,8.6,9.2 from normal populations, test whether the variances are equal at 0.05 level.

15. If $X_1, \ldots, X_n$ are observations from normal populations with known variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, how would one test whether their means are all equal?

16. A prominent baseball player's batting average dropped from 0.313 in one year to 0.280 in the following year.He was at bat 374 times during the first year and 268 times the second year. Is the hypothesis tenable at the 0.05 level that his hitting was the same during the two years?

17. Find the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ using a random sample of size n from exponential density with parameter θ.

18. A study of pain and activity for good and poor sleepers yielded the following statistics: Among 28 good sleepers, the hours of activity had mean $\overline{x} = 10.7$ and $s.d. = 4.8$. For 70 poor sleepers the statistics were $\overline{y} = 8.6$ and $s.d = 4.8$. determine an approximate Z-score for the mean difference that one could use in testing the hypothesis $\mu_X = \mu_Y$.

19. A manufacturer claims that the life time of a certain brand of batteries produced by his factory has a variance of $5,000(hours^2)$. A sample of size 26 has a variance of $7200(hours^2$. Assuming that it is reasonable to treat these data as a random sample from a normal population.Test the manufacturer's claim of variance at 2 percent of level of significance.

20. The rainfall at a certain station during a year may be assumed to be normally distributed random variable with, $\sigma = 3$ inches and unknown mean μ. For the past ten years, a record provides the following rainfalls: $x_1 = 30.5, x_2 = 34.1, x_3 = 27.9, x_4 = 29.4, x_5 = 35.0, x_6 = 26.9, x_7 = 30.2, x_8 = 28.3, x_9 = 31.7, x_{10} = 25.8$. Test the hypothesis $H_0 : \mu = 30$ $versus$ $H_1 : \mu < 30$ at 5 percent of level of significance.

21. A manufacturer claims that packages of certain goods contain 18 ounces. In order to check his claim, 100 packages are chosen at random from a large lot and it is found that $\sum_{j=1}^{100} x_j = 1752$ and $\sum_{i=1}^{100} x_i^2 = 31,157$. Make the appropriate assumptions and test the hypothesis H_0 that the manufacturers claim is correct against the appropriate alternative at level of significance $\alpha = 0.01$.

22. The number X of fatal traffic accidents in a certain city during a year may assumed to be a random variable distributed as Poisson with parameter λ. For the latest year $x = 14$, where as for the past several years the average was 20. Test whether it has been an improvement at one percent of level of significance.

Related references

1. BICKEL P.J., DOKSUM K.A. : Mathematical Statistics. Holden-Day, 1977.
2. Mc CULLAGH P., NELDER J.A. : Generalized linear models. Chapman and Hall, 1983. (2nd edition : 1989)
3. KALBFLEISCH J.G. : Probability and Statistical Inference. Vol. 1 : Probability. Vol. 2 : Statistical Inference. Springer, 1985. (2nd edition)
4. LEHMANN E.L. : Testing statistical hypotheses. John Wiley, 1959. (2nd edition : 1986)
5. LEHMANN E.L. : Nonparametrics. Holden-Day/McGraw-Hill, 1975.
6. LEHMANN E.L. : Theory of Point Estimation. John Wiley, 1983.
7. DOBSON A.J. : An introduction to generalized linear models. Chapman and Hall, 1990.
8. MOOD A.M., GRAYBILL F.A., BOES D.C. : Introduction to the theory of statistics. McGraw-Hill, 1986.
9. RAO C.R. : Linear statistical inference and its applications. John Wiley, 1973.
10. ROUSSAS G.G. : A first course in mathematical statistics. Addison-Wesley, 1973.
11. SERFLING R.J. : Approximation theorems of mathematical statistics. John Wiley, 1980.
12. SILVEY S.D. : Statistical Inference. Chapman and Hall, 1975.

Appendix A

Mathematical addendum

A.1 Introduction

The purpose of this appendix is to provide the reader with a ready reference to some mathematical results that are used in the book. This appendix is divided into two main sections: The first section, gives results that are, for the most part, combinatorial in nature, and the last gives results from calculus.

A.2 Non-calculus

1. A companion series, the finite **geometric series**, or progression, is given by

$$\sum_{j=0}^{n-1} ar^j \;=\; a\frac{1-r^n}{1-r}$$

and an **arithmetic series** or progression, is given by

$$\sum_{j=1}^{n}[a+(j-1)d] \;=\; na+\frac{d}{n}n-1$$

.

2. A product of a positive integer n by all the positive integers smaller than it is usually denoted by $n!$ (read "**n factorial**"). Thus

$$n \;=\; n(n-1)(n-2),\ldots,1 \;=\; \prod_{j=0}^{n-1}(n-j)$$

By convention we take $0! \;=\; 1$

3. A product of a positive integer n by the next k - 1 smaller positive integers is usually denoted by $(n)_k$. Thus

$$(n)_k \;=\; n(n-1),\ldots,(n-k+1) \;=\; \prod_{j=1}^{k}(n-j+1)$$

Binomial coefficient: $\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$ for positive n and $k \leq n$

4. Stirling's Formula

In finding numerical values of probabilities, one is often confronted with the evaluation of long factorial expressions which can be troublesome to compute by direct multiplication. Much labor may be saved by using Stirling's formula, which gives an approximate value of $n!$. Stirling's formula is

$$n! \approx (2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}}$$

or

$$n! \approx (2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}} e^{\frac{r(n)}{12n}},$$

where

$$1 - \frac{1}{12n+1} < r(n) < 1$$

5. The binomial and multinomial theorems

The binomial theorem is often given as

$$(a+b)^n = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) a^j b^{n-j}$$

for n a positive integer. The binomial theorem explains why the $\left(\begin{array}{c} n \\ j \end{array} \right)$ are sometimes called binomial coefficients. Four special cases are noted in the following remark. Remark

$$(1+t)^n = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) t^j,$$

$$(1-t)^n = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) (-1)^j t^j,$$

$$2^n = \sum_{j=0}^{n} \left(\begin{array}{c} n \\ j \end{array} \right)$$

and

$$0 = \sum_{j=0}^{n} (-1)^j \left(\begin{array}{c} n \\ j \end{array} \right).$$

Expanding both sides of

$$(1+x)^a(1+x)^b \;=\; (1+x)a+b$$

and then equating coefficients of x to the n^{th} power gives

$$\sum_{j=0}^{n}\begin{pmatrix} a \\ j \end{pmatrix}\begin{pmatrix} b \\ n-j \end{pmatrix} \;=\; \begin{pmatrix} a+b \\ n \end{pmatrix}$$

a formula that is particularly useful in considerations of the hypergeometric distribution. The following formula are also useful;

$$\sum_{j=0}^{\infty}\begin{pmatrix} n+j-1 \\ j \end{pmatrix} a^j \;=\; (1-a)^{-n}$$

$$\sum_{j=0}^{\infty} xt^x \;=\; \frac{1}{(1-t)^2}$$

$$\sum_{j=0}^{\infty} x(x-1)t^x \;=\; \frac{2t^2}{(1-t)^3}$$

A generalization of the binomial theorem is the **multinomial theorem**, which is

$$(\sum_{j=1}^{k} a_j)^n \;=\; \sum \frac{n!}{\prod_{i=1}^{k} n_i!}\prod_{i=1}^{k} a_i^{ni},$$

where the summation is overall nonnegative integers $n_1, n_2, \ldots, n_k$ which sum to n. A special case is

$$(\sum_{j=1}^{k} a_j)^2 = (\sum_{i=1}^{k} a_i)(\sum_{j=1}^{k} a_j) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j.$$

Also note that

$$(\sum_{i=1}^{m} a_i)(\sum_{j=1}^{n} b_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i a_j.$$

A.3 Calculus

1. It is assumed that the reader is familiar with the concepts of limits, continuity, differentiation, integration, and infinite series. A particular limit that is referred to several times is the limit expression for **the number e** ; that is,

$$\lim_{x\to 0}(1+x)^{\frac{1}{x}} = e$$

There are a number of variations of the above equation, for instance,

$$\lim_{x\to\infty}(1+\frac{1}{x})^x = e$$

$$\lim_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda}$$

and

$$\lim_{x\to 0}(1+x\lambda)^{\frac{1}{x}} = e^{\lambda}$$

for a constant λ.

2. Another rule that we might use is **Leibniz rule** for differentiating an integral: Let

$$I(t) = \int_{g(t)}^{h(t)} f(x;t)\,\mathrm{d}x$$

where $f(.;.), g(.)$, and $h(.)$ are assumed differentiable. Then

$$\frac{dI}{dt} = \int_{g(t)}^{h(t)} \frac{\partial f}{\partial x}\,\mathrm{d}x + f(h(t);t)\frac{dh}{dt} - f(g(t);t)\frac{dg}{dt}$$

Several important special cases derive from Leibniz rule; for example, if the integrand $f(x;t)$ does not depend on t, then

$$\frac{d}{dt}\left[\int_{g(t)}^{h(t)} f(x)\,\mathrm{d}x\right] = f(h(t))\frac{dh}{dt} - f(g(t))\frac{dg}{dt};$$

in practice, if $g(t)$ is constant and $h(t) = t$, the above equation simplifies to

$$\frac{d}{dt}\left[\int_{c}^{t} f(x)\,\mathrm{d}x\right] = f(t)$$

3. The **Taylor Series** for $f(x)$ about x=a is defined as

$$f(x) = f(a) + f^{(1)}(a)(x-a) + \frac{f^{(2)}(a)(x-a)^2}{2!} + \ldots + \frac{f^{(n)}(a)(x-a)^n}{n!} + R_n$$

where

$$f^{(1)}(a) = \frac{d^i f(x)}{dx^i}|_{x=a}; \quad R(n) = \frac{f^{(n+1)}(c)(x-a)^{n+1}}{(n+1)!}$$

and $a \le c \le x$. R_n is the remainder. $f(x)$ is assumed to have derivatives of at least order $n+1$.
The Taylor series for functions of one variable given above can be generalized to the Taylor series for functions of several variables. For example, the Taylor series for $f(x,y)$ about $x = a$ and $y = b$ can be written as

$$f(x,y) = f(a,b) + f_x(a,b)(x-a) + f_y(a,b)(y-b) +$$

$$\frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b) + f_{yy}(a,b)(y-b)^2\right] + ...,$$

where

$$f_x(a,b) = \frac{\partial f}{\partial x}|_{x=a,y=b},$$

$$f_{xy}(a,b) = \frac{\partial^2 f}{\partial x \partial y}|_{x=a,y=b},$$

and similarly for the others.

4. The Gamma and Beta functions

The **gamma function**, denoted by $\Gamma(.)$, is defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, \mathrm{d}x$$

for t¿0.

$\Gamma(t)$ is nothing more than a notation for the definite integral that appears on the right hand side of the above equation. Integrating by part yields

$$\Gamma(t+1) = t\Gamma(t),$$

and, hence, if t=n (an integer),

$$\Gamma(n+1) = n!$$

If n is an integer,

$$\Gamma(n+\frac{1}{2}) = \frac{1.3.5...(2n-1)}{2^n}\sqrt{\pi}$$

and, in particular,

$$\Gamma(\frac{1}{2}) = 2\Gamma(\frac{3}{2}) = \sqrt{\pi}.$$

The **beta function**, denoted by B(.,.)is defined by

$$Beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\, \mathrm{d}x$$

For $a > 0,\ b > 0$

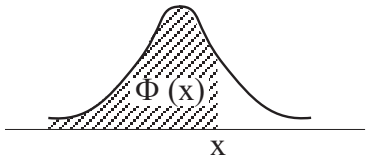
Again, $Beta(a,b)$ is just a notation for the definite integral that appears on the right-hand side of the above equation. A simple variable substitution gives $Beta(a,b) = Beta(b,a)$. The beta function is related to the gamma function according to the following formula:

$$Beta(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Appendix B

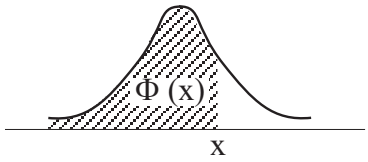
Tables

Table 1 : Standard normal distribution



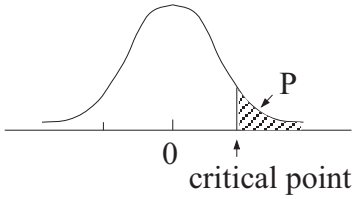
	$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$									
x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706

Table 1 : Standard normal distribution



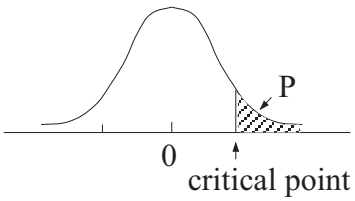
	$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$									
x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Table 2 : Critical points of Student t-distributions



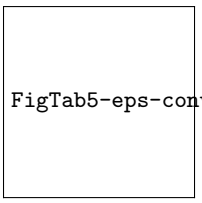
P / d.f.	.25	.10	.05	.025	.010	.005	.001
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	.687	1.325	1.725	2.086	2.528	2.845	3.552

Table 2 : Critical points of Student *t*-distributions



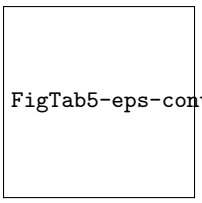
P d.f.	.25	.10	.05	.025	.010	.005	.001
21	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.674	1.282	1.645	1.960	2.326	2.576	3.090

Table 3 : Critical points of χ^2-distributions



P d.f.	.250	.100	.050	.025	.010	.005	.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	23.8	28.4	31.4	34.2	37.6	40.0	45.3

Table 3 : Critical points of χ^2-distributions



P d.f.	.250	.100	.050	.025	.010	.005	.001
21	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	77.6	85.5	90.5	95.0	100	104	112
80	88.1	96.6	102	107	112	116	125
90	98.6	108	113	118	124	128	137
100	109	118	124	130	136	140	149