1. INTRODUCTION

1.1 What is Statistics?

STATISTICS: a subject dealing with the **collection** and the **use** of data in order to (help to) answer the questions like:

- Do stock markets rise and fall randomly?
- Is global warming really happening?
- Does a certain new drug prolong life for AIDS sufferers?
- Are GCSE and A level examinations standards declining?
- Is the national lottery making us a nation of compulsive gamblers?
- Is the gap between rich and poor widening in Britain?
- Do Persil adverts really make us want to buy Persil?
- .....

Those questions are difficult to study in laboratory, and admit no self-evident axioms

Typically *Data* are subject to **uncertainty** — *random variables*

Statistics: **Let the Data Speak**!

- making guesses about the process generating the data – *Estimation*
- testing the guesses – *Testing hypotheses*
- forecasting future – *Prediction*

Note. Statistics also deals with *experiment design* which will be untouched in this course.

Statistics offers quantitative approaches to solve practical problems. Therefore, **a good knowledge** on the practical world behind data is essential for decent statistics practice.

No unique statistical solution for most practical problems! Therefore

Statistics is also an art

Some guidelines for learning/applying statistics:

- Understand what data say in each specific context. All the methods are just tools to help understand data
- Concentrate on what to do and why, rather than concrete calculation and graphing
- It may take a while to catch the essences of statistics – Keep thinking!

What is the difference between *probability* **and** *statistical inference***?**

Probability — a mathematical subject

Statistics — an applied oriented subject: inference based on data plus 'assumptions'

Note. The assumptions imposed in statistical inference are typically based on probability theory.

Consider a simple experiment: a coin is tossed n times.

Define a random variable X = the number of 'heads'.

Probability Questions:

- What is $P(X = 1)$, $P(X \leq 1)$, $P(X = x)$ or $P(X \leq x)$?
- What is $E(X)$, $\mathrm{Var}(X)$, $E(X^m)$ or $E(e^{tX})$?

Probability Answers

If we **assume** that the tosses are independent with a constant probability of success π, then

- $P(X = 1) = n\pi(1-\pi)^{n-1}$,
  $P(X \leq 1) = (1-\pi)^{n-1}(1-\pi+n\pi)$,
  $P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$, etc
- $E(X) = n\pi$, $\mathrm{Var}(X) = n\pi(1-\pi)$, etc

Statistical Questions

- What is the value of π?
- Is there any evidence that the coin is 'biased' i.e. $\pi \neq 0.5$?
- Are the tosses independent?

Statistical Answers

- Estimate π by the estimator x/n. Give an interval in which π may be expected to fall.
- Reject the Null Hypothesis that $\pi = 0.5$ if x is too far away from $n/2$
- Reject the model of independent trials if there is too much regularity in the observations - e.g. HTHTHTHTHTHTHTHT....

1.2 Terminology and Notation

The classic statistical inference is based on a **sample** (i.e. a set of **data**)

$$\mathbf{X} = (X_1, \cdots, X_n)^T$$

and the **assumption** that $\{X_i\}$ i.i.d. with

$$X_i \sim F_X(x;\theta), \quad \theta \in \Theta,$$

where $F_X(\cdot;\theta)$ is a probability distribution and is known up to an *unknown* **parameter** θ, Θ is the **parameter space**.

Population: $\{F_X(\cdot;\theta),\ \theta \in \Theta\}$, which is assumed to contain the **true** distribution of X_i

Sample space: consisting of all possible values of $\mathbf{X}$

Statistic: any (known) function of $\mathbf{X}$ only

Dual identity of a sample:

- a set of observed real numbers in practice, and
- a set of i.i.d. r.v.s in theoretical exploration.

Goal of statistical inference in this course: estimating the true value of θ or testing the hypotheses about θ

Note. A sample or a random sample refers to a set of i.i.d. observations.

Example 1. Let $X_1, \cdots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Then $(X_1, \cdots, X_n)$ is a sample, R^n is the sample space, and

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \qquad \Theta = R \times R^+.$$

The population is

$$\{N(\mu, \sigma^2),\ (\mu, \sigma^2) \in \Theta\}.$$

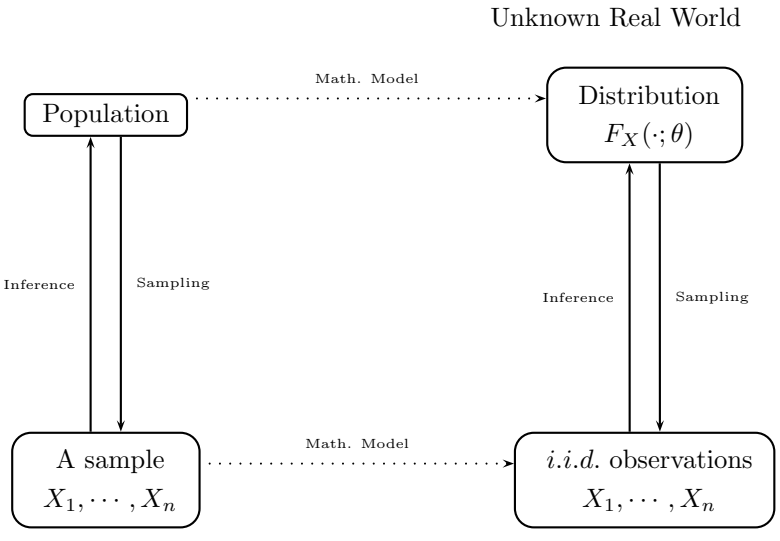
Both sample mean $\bar{X} = n^{-1} \sum_{1 \le i \le n} X_i$ and sample variance $S = \frac{1}{n-1} \sum_{1 \le i \le n} (X_i - \bar{X})^2$ are statistics. But $\bar{X} - \mu$ is not!

Note. The above setting is referred as a **parametric model** since the population is known up to unknown parameters.

In contrast, a **nonparametric model** specifies that the true distribution belongs to a class which cannot be specified by a finite number of parameters only. For example, a population may be

$$\{\text{all continuous density functions in one variable}\}.$$

We deal with parametric models first.



1.3 Statistics from Normal samples

In this section we always assume that $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ be a random sample from $N(0,1)$.

Then the two summary statistics are sample mean and sample variance:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_i - \bar{X})^2.$$

Theorem 1. The following statements hold.

(i) $\bar{X} \sim N(0, 1/n)$,
(ii) $(n-1)S^2 \sim \chi^2_{n-1}$,
(iii) $\bar{X}$ and S^2 are independent.

Note. The distribution of $\sum_{1\le j\le n} X_i^2$ is called **the χ^2-square distribution with n degrees of freedom**, denoted as χ^2_n or $\chi^2(n)$.

Proof. We make a linear transformation from $X_1, \ldots, X_n$ to $\mathbf{Z} = (Z_1, \ldots, Z_n)^\tau$ in such a way that the Z_i's are still i.i.d. $N(0,1)$. This can be achieved by any transformation $\mathbf{Z} = \mathbf{HX}$ with $\mathbf{H}$ being a $n \times n$ orthogonal matrix (i.e. $\mathbf{H}^\tau\mathbf{H} = \mathbf{I}_n$). Let the first row of $\mathbf{H}$ be

$$n^{-1/2}(1, 1, \ldots, 1).$$

Then $Z_1 = \sqrt{n}\bar{X}$. Hence (i) holds. Since

$$\sum_{j=1}^{n} Z_j^2 = \mathbf{X}^\tau\mathbf{H}^\tau\mathbf{H}\mathbf{X} = \mathbf{X}^\tau\mathbf{X} = \sum_{j=1}^{n} X_j^2,$$

it holds that

$$\sum_{j=2}^{n} Z_j^2 = \sum_{j=1}^{n} X_j^2 - Z_1^2 = \sum_{j=1}^{n} X_j^2 - n\bar{X}^2 = (n-1)S^2.$$

Therefore (ii) and (iii) also hold.

Remark. The above proof indicates the following decomposition

$$\begin{array}{ccccc} \sum_{i=1}^n X_i^2 & = & \sum_{i=1}^n (X_i - \bar{X})^2 & + & n\bar{X}^2 \\ \chi^2_n & & \chi^2_{(n-1)} & & \chi^2_{(1)}, \end{array}$$

where the two χ^2 random variables on the RHS are independent.

Example 1. Let $Y_1, \cdots, Y_n$ iid from $N(\mu, \sigma^2)$. Define

$$X_i = (Y_i - \mu)/\sigma.$$

Then $\{X_i\}$ i.i.d. $N(0,1)$, and

$$\bar{Y} = \sigma\bar{X} + \mu, \quad S_y^2 = \sigma^2 S_x^2.$$

Hence

(i) $\bar{Y} \sim N(\mu, \sigma^2/n)$,
(ii) $(n-1)S_y^2/\sigma^2 \sim \chi^2_{n-1}$,
(iii) $\bar{Y}$ and S_y^2 are independent.

Further,

$$\sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \bar{Y})^2}{\sigma^2} + n\frac{(\bar{Y} - \mu)^2}{\sigma^2}.$$

F distribution

Definition. Let $Y \sim \chi^2_n$ and $X \sim \chi^2_m$, and X and Y are independent. Then the random variable

$$U \equiv \frac{X/m}{Y/n} = \frac{n}{m}\frac{X}{Y}$$

has an F-distribution with degrees of freedom (m, n). We write $U \sim F(m, n)$ or $U \sim F_{m,n}$.

Obviously, $1/U \sim F_{n,m}$.

A formal notation: $F_{m,n} = \frac{\chi^2_m/m}{\chi^2_n/n}$.

Typical use of F distribution

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two indep samples from $N(\mu_x, \sigma^2_x)$ and $N(\mu_y, \sigma^2_y)$ respectively. Then

$$\frac{\frac{1}{(m-1)\sigma^2_x}\sum_{i=1}^{m}(X_i - \bar{X})^2}{\frac{1}{(n-1)\sigma^2_y}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \sim F_{(m-1),(n-1)}.$$

Student's t-distribution

Definition. Let $Z \sim N(0, 1)$ and $U \sim \chi^2_{(k)}$, and Z and U be independent. Then the random variable

$$T = \frac{Z}{[\frac{1}{k}U]^{\frac{1}{2}}},$$

has a t-distribution with k degrees of freedom. We write $T \sim t_k$ or $T \sim t(k)$.

Obviously, $T^2 \sim F_{1,k}$.

Also T and $-T$ have the same distribution, i.e. the pdf is symmetric. In fact $ET = 0$ for $k > 1$.

A formal notation: $t_k = \frac{N(0,1)}{\{\chi^2_k/k\}^{1/2}}$.

Typical use of t distribution

If $Y_1, Y_2, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, then

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

and

$$(n-1)S^2_y/\sigma^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

and these two are independent. It follows that

$$\frac{\bar{Y} - \mu}{\sqrt{S^2_y/n}} \sim t_{n-1}.$$

2. MAXIMUM LIKELIHOOD ESTIMATION

2.1 Likelihood

Likelihood is one of the most fundamental concept in all types of statistical inference.

Definition 1 Suppose that $\mathbf{X}$ has density function or probability function $f(\mathbf{x};\boldsymbol{\theta})$. We have observed $\mathbf{X}=\mathbf{x}$. Then the likelihood function with observation $\mathbf{x}$ is defined as

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}).$$

Density/probability function: a function of $\mathbf{x}$, specifying the distribution of random variable $\mathbf{X}$

Likelihood: a function of $\boldsymbol{\theta}$, reflecting information on $\boldsymbol{\theta}$ contained in observation $\mathbf{x}$

Note. A likelihood function represents the uncertainty on a unknown non-random constant $\boldsymbol{\theta}$, and it is **n**ot a density or probability function! It provides

a rational degree of belief, or
an order of preferences

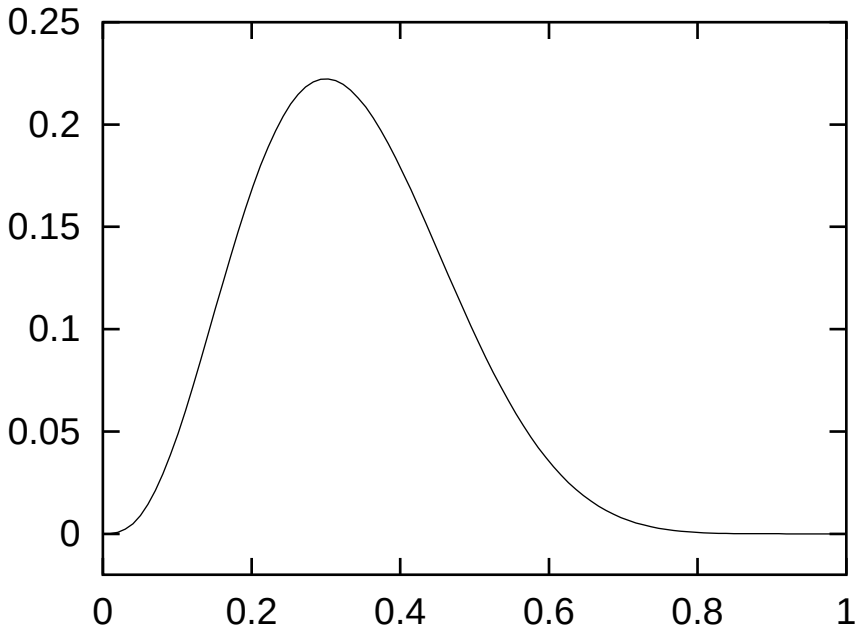
on possible values of the parameter $\boldsymbol{\theta}$. This can be seen more clearly in the simple example on next slide.

In fact, a likelihood function is often defined up to a positive constant — the constant here refers to a quantity independent of $\boldsymbol{\theta}$. But it may depending on $\mathbf{x}$. (Note $\mathbf{x}$ is a given constant.)

Example 1. Suppose that x is the number of successes from a known number n of independent trials with unknown probability of success π. The probability function, and so the likelihood function is

$$L(\pi) = f(x;\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}.$$

The likelihood function $L(\pi;x)$ can be graphed as a function of π. It changes shape for different values of x. A likelihood function for $x=3$ when $n=10$ is shown in the Figure below.



Notice that the likelihood function shown above is *not* a density function. It does not have an area of 1 below it.

We use the likelihood function to compare the plausibility of different possible parameter values. For instance, the likelihood is much larger for $\pi = 0.3$ than for $\pi = 0.8$, that is the data $x = 3$ have a greater probability of being observed if $\pi = 0.3$ than if $\pi = 0.8$. This makes $\pi = 0.3$ much more **l**ikely as the true value for π than 0.8.

Note. In the above argument, we do not need to calculate exact probabilities under different values of $\boldsymbol{\theta}$. Only the order of those quantities matters!

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \boldsymbol{\theta})$. Write $\mathbf{X} = (X_1, \cdots, X_n)^\tau$. Then the likelihood function is

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta}),$$

which is a product of n terms. Then the **l**og-likelihood function is

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) \equiv \log\{L(\boldsymbol{\theta}; \mathbf{X})\} = \sum_{i=1}^{n} \log\{f(X_i, \boldsymbol{\theta})\},$$

which is a sum of n terms.

This explains why log-likelihood functions are often used with independent observations.

Definition 2. *Likelihood Principle*

For two observed values $\mathbf{x}$ and $\mathbf{y}$, if

$$L(\boldsymbol{\theta}; \mathbf{x}) \propto L(\boldsymbol{\theta}; \mathbf{y}),$$

the inferences for $\boldsymbol{\theta}$ based on $\mathbf{x}$ and $\mathbf{y}$ should be the same.

2.2 Sufficiency and Data Compression

If we start with n observations and construct from them a k-dimensional statistic, if $k < n$ we expect that in general some information in the sample about $\boldsymbol{\theta}$ would be lost. The statistics is less complicated than the original sample because it is in a smaller dimension. It reduces the data, and in general will reduce the information about $\boldsymbol{\theta}$. A simple example may help to make clear what is happening.

Example 2 Based on a sample $\{X_1, \cdots, X_n\}$, consider a sequence of statistics of increasing dimensions:

$$\begin{aligned}
T_1(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i\right) \\
T_2(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right) \\
T_3(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3\right) \\
&\cdots \\
T_{n-1}(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3, \ldots, \sum_{i=1}^{n} X_i^{n-1}\right) \\
T_n(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3, \ldots, \sum_{i=1}^{n} X_i^n\right)
\end{aligned}$$

Here we see that T_1 is very simple and in 1 dimension, whereas T_n is in n dimensions. Actually, from T_n one can reconstruct the original observations $X_1, X_2, \ldots, X_n$, except for their ordering. So T_n contains all

the information in the sample about $\boldsymbol{\theta}$. But except in special cases T_1 will not include all the information about $\boldsymbol{\theta}$ in the sample.

If a statistic contains all the information in a sample about $\boldsymbol{\theta}$ we shall say that it is a **sufficient statistic** for $\boldsymbol{\theta}$.

Example 3 If X_1, X_2 are independent observations from $N(\mu, 1)$ then the statistic $X_1 + X_2$ contains all the information in the sample about μ. This is intuitive, because knowing $X_1 + X_2$ and $X_1 - X_2$ is equivalent to having both observations X_1, X_2, but $X_1 - X_2$ has a $N(0, 2)$ distribution and so contains no information about μ. This conclusion is made more secure by the independence of X_1+X_2 and X_1-X_2.

Now we make the idea of sufficiency more precise with one of the great ideas from the work of R.A.Fisher. We seek to define an idea of 'sufficiency' such that a statistic T is 'sufficient' for $\boldsymbol{\theta}$ if it contains all the information about $\boldsymbol{\theta}$ that was in the whole set of observations.

Definition 3. *Sufficient Statistic*
Suppose $\mathbf{X} \sim f(\cdot, \boldsymbol{\theta})$. $\mathbf{T}(\mathbf{X})$ is said to be a **s**ufficient statistic for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

The likelihood principle implies the sufficiency principle which indicates that we only need to use sufficient statistics in inference.

Definition 4. *Sufficiency Principle*
All sufficient statistics based on $\mathbf{X}$ should lead to the same inferences for $\boldsymbol{\theta}$.

Theorem 1. *Factorisation Criterion*
Let $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. Then $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ iff

$$f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}).$$

We will not prove this result. To see how a proof might go, note that it is always possible, but not necessary, to take $h(\mathbf{x})$ as the conditional density of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ and $g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})$ as the density of $\mathbf{T}(\mathbf{X})$.

Example 4. Suppose that we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)^\tau$ from $N(\mu, 1)$. Then the joint density is

$$f_{\mathbf{X}}(\mathbf{X}; \boldsymbol{\theta}) = \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}n(\bar{X} - \mu)^2\right] \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right].$$

So we can take $T(\mathbf{X}) = \bar{X}$ and

$$g(\bar{X}, \boldsymbol{\theta}) = \exp\left[-\frac{1}{2}n(\bar{X} - \mu)^2\right],$$

and

$$h(\mathbf{X}) = \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right].$$

Two samples with the same sample mean should lead to exactly the same inferences about μ if the sufficiency principle holds.

Example 5. Here is another example where a sufficient statistic is of the same dimension as the sample size, but is not the whole sample. Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a sample from $U(-\boldsymbol{\theta}, \boldsymbol{\theta})$ and $\Theta = R^+$. Then the absolute values of the observations $\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid$ are from $U(0, \boldsymbol{\theta})$. One loses the information on the signs of the observations, but retains full information on $\boldsymbol{\theta}$. Here the statistic $\mathbf{T}(X) = (\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid)$ is a sufficient statistic for $\boldsymbol{\theta}$.

In fact the conditional distribution of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X}) = (T_1, T_2, \ldots, T_n) = (\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid)$ is the distribution where independently for all i, $X_i = T_i$ with probability 0.5 and $X_i = -T_i$ with probability 0.5. This conditional distribution does not depend on $\boldsymbol{\theta}$, so the definition is satisfied.

Example 6. (Continuing Example) Another sufficient statistic of $\boldsymbol{\theta}$ is $\max_{1 \le i \le n} |X_i|$. To this end, note that

$$f_{X_i}(x_i) = \frac{1}{2}\boldsymbol{\theta} I_{(-\boldsymbol{\theta}, \boldsymbol{\theta})}(x_i) = \begin{cases} \frac{1}{2\boldsymbol{\theta}} & |x_i| < \boldsymbol{\theta}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$f_{\mathbf{X}}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

$$= \begin{cases} \frac{1}{(2\boldsymbol{\theta})^n} & \max_{1 \le i \le n} |x_i| < \boldsymbol{\theta}, \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{1}{(2\boldsymbol{\theta})^n} I(\max_{1 \le i \le n} |x_i| < \boldsymbol{\theta}).$$

Now the conclusion follows from the factorisation criterion.

How do you estimate $\boldsymbol{\theta}$?

Example 7. Suppose that the $X_1, \cdots, X_n$ are iid $U(0, \boldsymbol{\theta})$, where $\boldsymbol{\theta} > 0$. Then

$$f(x_i; \boldsymbol{\theta}) = \begin{cases} \frac{1}{\boldsymbol{\theta}} & x_i \in [0, \boldsymbol{\theta}] \\ 0 & \text{otherwise} \end{cases}$$

which can be written

$$f(x_i; \boldsymbol{\theta}) = \frac{1}{\boldsymbol{\theta}} \times I_{\{x_i \le \boldsymbol{\theta}\}} I_{\{x_i \ge 0\}}.$$

The joint density is

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\boldsymbol{\theta}^n} \prod_{i=1}^{n} I_{\{x_i \le \boldsymbol{\theta}\}} \prod_{i=1}^{n} I_{\{x_i \ge 0\}}$$

$$= \frac{1}{\boldsymbol{\theta}^n} I_{\{\max_i x_i \le \boldsymbol{\theta}\}} I_{\{\min_i x_i \ge 0\}}$$

$$= g\left(T(\mathbf{x}); \boldsymbol{\theta}\right) \times h(\mathbf{x})$$

where $T(\mathbf{x}) = \max_i x_i$. In this case the (at first sight) rather surprising result is that $\max_i X_i$ is a sufficient statistic for $\boldsymbol{\theta}$.

2.3 Maximum likelihood estimator (MLE)

The MLE is by far the most popular method for deriving estimators.

Definition 5 — **M**LE
A *Maximum Likelihood Estimator* (MLE), $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$, of parameter $\boldsymbol{\theta}$ is an estimator satisfying

$$L(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq L(\boldsymbol{\theta}; \mathbf{X}) \text{ for all } \boldsymbol{\theta} \in \Theta, \text{ or equivalently } l(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq l(\boldsymbol{\theta}; \mathbf{X}) \text{ for all } \boldsymbol{\theta} \in \Theta.$$

Obviously, a maximum likelihood estimator is the most plausible value for $\boldsymbol{\theta}$ as judged by the likelihood function. In many cases where Θ is continuous and the maximum does not occur at a boundary of Θ, $\widehat{\boldsymbol{\theta}}$ **i**s often the solution of the equation

$$s(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{X}) = 0.$$

We call $s(\boldsymbol{\theta}) \equiv s(\boldsymbol{\theta}; \mathbf{X})$ **a** score function.

Example 8. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ where neither μ or σ^2 is known. Then we can find the maximum likelihood estimator from the log-likelihood

$$\begin{aligned} l(\mu, \sigma^2) &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \mu)^2/(2\sigma^2) \\ &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \bar{Y})^2/(2\sigma^2) \\ &- n(\bar{Y} - \mu)^2/(2\sigma^2). \end{aligned}$$

This is maximised by choosing $\mu = \bar{Y}$, so $\widehat{\mu} = \bar{Y}$ is the MLE for μ. It is easy to see

$$E(\widehat{\mu}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

Such a estimator is called **u**nbiased.

Note. The above MLE for μ is equivalent to minimise

$$\sum_{i=1}^n (Y_i - \mu)^2.$$

Therefore $\widehat{\mu}$ is also the **l**east squares estimator (LSE). Note that LSE is derived from a simple empirical/geometric rule, which makes **no use** of the underlying distribution.

The **p**rofile log-likelihood remaining is

$$l(\widehat{\mu}, \sigma^2) = -n \log \sqrt{2\pi} + (n/2)(\log \sigma^{-2} - \widehat{\sigma}^2 \sigma^{-2}),$$

where $\widehat{\sigma}^2 = \sum_1^n (Y_i - \bar{Y})^2/n$. By the lemma below, the MLE for σ^2 is $\widehat{\sigma}^2$. Note that the MLE of σ^2 is *biased* since

$$E(\widehat{\sigma}^2) = (1 - 1/n)\sigma^2 \neq \sigma^2.$$

Lemma. Define $L(x) = \log(x^{-1}) - b/x$, where $n \geq 1$ and $b > 0$ are constants. Then $L(b) \geq L(x)$ for all $x > 0$.

Example 9. Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(π). Then

$$L(\pi) = \prod_{i=1}^n \pi^{X_i} (1 - \pi)^{1 - X_i} = \pi^{n\bar{X}} (1 - \pi)^{n(1 - \bar{X})}.$$

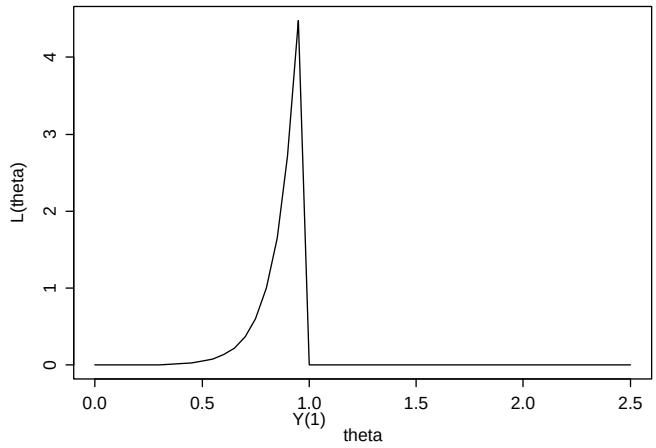
$$l(\pi) = n\bar{X} \log \pi + n(1 - \bar{X}) \log(1 - \pi).$$

Let $s(\pi) = \frac{\partial}{\partial \pi} l(\pi) = 0$, leading to $\widehat{\pi} = \bar{X}$.

Example 10. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from an exponential distribution with density function $e^{-(y-\theta)}$ for $y \geq \theta$. This is the usual exponential distribution shifted to start at θ. The Likelihood is

$$L(\theta; \mathbf{Y}) = e^{-n(\bar{Y}-\theta)} I_{\{(\theta,\infty)\}}(Y_{(1)}),$$

where $Y_{(1)}$ is the smallest observation. This likelihood is zero for $\theta > Y_{(1)}$ and increases in θ for $\theta \leq Y_{(1)}$. So the MLE $\widehat{\theta} = Y_{(1)}$, which is a boundary maximum.



Invariance property of MLEs

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$, and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ is a 1-1 transform. Let $\widehat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{X}, \boldsymbol{\theta}).$$

It is obvious to see that the MLE for $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}).$$

2.4 Numerical computation of MLEs

In modern statistical applications, it is typically difficult to find explicit analytic forms for the maximum likelihood estimators. These estimators are found more often by iterative procedures built into computer software. An iterative scheme starts with some guess at the MLE and then steadily improves it with each iteration. The estimator is considered found when it has become numerically stable. Sometimes the iterative procedures become trapped at a local maximum which is not a global maximum. There may be a very large number of parameters in a model, which makes such local entrapment much more common.

Newton-Raphson Scheme

Suppose that the log-likelihood function function $l(\boldsymbol{\theta})$ is sufficiently smooth. Then

$$s(\widehat{\boldsymbol{\theta}}) = 0,$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE and $s(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} l(\boldsymbol{\theta})$ is the score function. Let

$$\dot{s}(\boldsymbol{\theta}) = \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta \partial \theta^\tau} l(\boldsymbol{\theta}).$$

Suppose $\widehat{\boldsymbol{\theta}}$ is close to the true value $\boldsymbol{\theta}_0$. By a simple Taylor expansion,

$$s(\boldsymbol{\theta}_0) = \dot{s}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) + o_p(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||).$$

This leads to the approximation

$$\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_0 - \{\dot{s}(\boldsymbol{\theta}_0)\}^{-1} s(\boldsymbol{\theta}_0).$$

Since θ_0 is unknown, we use iterative estimators

$$\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_k - \{\dot{s}(\widehat{\boldsymbol{\theta}}_k)\}^{-1} s(\widehat{\boldsymbol{\theta}}_k) \tag{1}$$

for $k = 1, 2, \cdots$, where $\widehat{\boldsymbol{\theta}}_0$ is a prescribed initial value. We define $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_j$ if $\widehat{\boldsymbol{\theta}}_j$ and $\widehat{\boldsymbol{\theta}}_{j-1}$ differ by a small amount.

Fisher Score method: replace $\dot{s}(\widehat{\boldsymbol{\theta}}_k)$ in (1) by $E_{\boldsymbol{\theta}}\{\dot{s}(\boldsymbol{\theta})\}$ under $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_k$.

Like most iterative algorithms, the choice of appropriate **i**nitial values is important to ensure the convergence to right limits. In practice multiple initial values are often used.

The differences between the Newton-Raphson and score methods are subtle. We make observations below

- The convergence of the Newton-Raphson algorithm is often faster when both algorithms converge
- The radius of convergence for the score method is often larger, making the choice of initial values less important for the score method.

2.5 EM algorithms

Goal: to find the MLE $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{Y})$ for $\boldsymbol{\theta}$ from the likelihood based on data $\mathbf{Y}$:

$$L(\boldsymbol{\theta}; \mathbf{Y}) = f_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}),$$

while the '*complete*' data $\mathbf{X}^\tau = (\mathbf{Y}^\tau, \mathbf{Z}^\tau)$ contain a '*missing*' component $\mathbf{Z}$. The likelihood based on the complete data is

$$L(\boldsymbol{\theta}; \mathbf{X}) = f_{\mathbf{X}}(\mathbf{X}, \boldsymbol{\theta}).$$

EM (*E*xpectation and *M*aximisation) **a**lgorithms

- *E-step*: compute the conditional expectation

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\theta}_0) \equiv E\{\log L(\boldsymbol{\theta}; \mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}_0\}$$

- *M-step*: maximise $Q(\boldsymbol{\theta})$ to give an updated value $\boldsymbol{\theta}_1$

then go to the E-step using $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, and keep iterating until convergence. The limit of $\boldsymbol{\theta}_0$ is taken as $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$.

Example 11. The genetic example from (Rao 1973, p.396) assumes that the phenotype data

$$\begin{aligned} \mathbf{Y} &= (Y_1, Y_2, Y_3, Y_4)^\tau \sim \\ &\quad \text{Multinomial}\big(4; \frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\big), \end{aligned}$$

where $\theta \in (0,1)$. Then log-likelihood is

$$l(\theta, \mathbf{Y}) = Y_1 \log(2+\theta) + (Y_2 + Y_3)\log(1-\theta) + Y_4 \log\theta + C,$$

which **d**oes not yield a closed form $\widehat{\theta}$.

Now we treat $\mathbf{Y}$ as incomplete data from $\mathbf{X} = (X_1, \ldots, X_5)^\tau$ with multinomial probabilities

$$(\frac{1}{2},\ \frac{\theta}{4},\ \frac{1-\theta}{4},\ \frac{1-\theta}{4},\ \frac{\theta}{4}).$$

Then

$$Y_1 = X_1 + X_2, \quad Y_i = X_{i+1} \quad \text{for } i = 2,3,4.$$

The log-likelihood of based on $\mathbf{X}$ is

$$l(\theta, \mathbf{X}) = (X_2 + X_5)\log\theta + (X_3 + X_4)\log(1-\theta) + C,$$

which **r**eadily yields

$$\widehat{\theta}(\mathbf{X}) = \frac{X_2 + X_5}{X_2 + X_3 + X_4 + X_5}.$$

Now the E-step is to find

$$\begin{aligned} Q(\theta) &= E\{l(\theta, \mathbf{X})|\mathbf{Y}, \theta_0\} \\ &= \log\theta E(X_2 + X_5|\mathbf{Y}, \theta_0) + \log(1-\theta)E(X_3 + X_4|\mathbf{Y}, \theta_0) \\ &= (\widehat{X}_2 + Y_4)\log\theta + (Y_2 + Y_3)\log(1-\theta), \end{aligned}$$

where $\widehat{X}_2 = E(X_2|\mathbf{Y}, \theta_0)$. Since the conditional distribution of X_2 given $Y_1 (= X_1 + X_2)$ is a binomial distribution with $n = Y_1$ and

$$p = \frac{\theta_0/4}{1/2 + \theta_0/4} = \frac{\theta_0}{2+\theta_0}.$$

Hence

$$\widehat{X}_2 = np = \frac{Y_1\theta_0}{2+\theta_0}. \tag{2}$$

The M-step leads to

$$\theta_1 = \frac{\widehat{X}_2 + Y_4}{\widehat{X}_2 + Y_4 + Y_2 + Y_3}. \tag{3}$$

For $Y = (125, 18, 20, 34)$ and the **initial value**

$$\theta_0 = 4 \times 34/(125 + 18 + 20 + 34)$$

which is a relative frequency estimate, the first 5 iterations between (2) and (3) are 0.690, 0.635, 0.628, 0.627 and 0.627, giving the MLE $\widehat{\theta} = 0.627$.

What can be said about *convergence properties* of the EM algorithm?

Let $\boldsymbol{\theta}_0$ be an arbitrary initial value and $\boldsymbol{\theta}_1$ be the updated value obtained from applying the iteration once. Then it can be shown that

$$L(\boldsymbol{\theta}_1; \mathbf{Y}) \geq L(\boldsymbol{\theta}_0; \mathbf{Y}).$$

Unfortunately, it does not imply that the iterations will always lead to the MLE eventually.

It is important to choose appropriate initial values to ensure the algorithm converges to the MLE. (This is also true for both Gaussian-Raphson and score methods!) In practice, it is a good idea to use **a** variety of initial values.

Further discussion on the convergence of EM algorithms, see Wu (1983) *Annals of Statistics, Vol.11, pp.95-103* and §12.4 of Pawitan (2001).

General comments on EM algorithms:

- The EM algorithm is a general procedure for computing MLEs. It is not really a numerical algorithm. The calculation of M-Step typically involves other numerical algorithms such as Newton-Raphson and the score methods.
- It can be applied when some data are *genuinely* missing. It can also be applied when the missing information is merely a concept based on which we transform a difficult optimisation problem into a sequence of easier problems; see Example 11 above. This is particularly relevant when $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$ is difficult to calculate while $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is easier to obtain.
- The convergence of EM algorithm may be very slow, depending on the amount of missing information. It is often slower than the Newton-Raphson algorithm.

Example 12. Mixture distributions
Let $Y_1, \cdots, Y_n$ be i.i.d. with a mixture pdf

$$f(y) = \sum_{j=1}^{k} \alpha_j f_j(y, \boldsymbol{\lambda}_j),$$

where f_j are pdfs, $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$. The parameter $\boldsymbol{\theta}$ contains all α_j and $\boldsymbol{\lambda}_j$. This model becomes important because

1. it represents heterogeneous data well since each f_j represents one heterogeneous component, and
2. it provides very good approximations to a large class of distributions.

The likelihood based on $\mathbf{Y} = (Y_1, \cdots, Y_n)^\tau$ is

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{n} \Big\{ \sum_{j=1}^{k} \alpha_j f_j(Y_j, \boldsymbol{\lambda}_j) \Big\}.$$

Maximising this likelihood is difficult, due to the presence of the summations, which reflect the fact that we are typically lacking in knowledge of which component any particular sample value comes from. This is the missing information!

Let $\mathbf{X}_i^\tau = (Y_i, \mathbf{Z}_i^\tau)$, where $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ik})^\tau$ is a $k \times 1$ vector with a 1 in the position corresponding to the component of the mixture that Y_i comes from, and 0 elsewhere.

Let $\mathbf{e}_j$ be the $k \times 1$ vector with the j-th component 1 and all the other components 0. The joint probability-density function for $(\mathbf{Z}_1, Y_1)$ is

$$\begin{aligned} & P\{Z_1 = \mathbf{e}_\ell,\ Y_1 \in [y_1, y_1 + dy_1)\}/dy_1 \\ = \ & P(Z_1 = \mathbf{e}_\ell)\, P\{Y_1 \in [y_1, y_1 + dy_1) | Z_1 = \mathbf{e}_\ell\}/dy \\ = \ & \alpha_\ell f_\ell(y_1, \boldsymbol{\lambda}_\ell) \ = \ \prod_{j=1}^{k} \alpha_j^{e_{\ell j}} f_j(y_1, \boldsymbol{\lambda}_j)^{e_{\ell j}}, \end{aligned}$$

where $\mathbf{e}_\ell = (e_{\ell 1}, \cdots, e_{\ell k})^\tau$.

Let $\mathbf{X}^\tau = (\mathbf{X}_1^\tau, \cdots, \mathbf{X}_n^\tau)$.

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \alpha_j^{Z_{ij}} f_j(Y_i, \boldsymbol{\lambda}_j)^{Z_{ij}},$$

$$l(\boldsymbol{\theta};\mathbf{X}) = \sum_{i=1}^{n} \mathbf{Z}_i^\tau \left\{ \begin{pmatrix} \log\alpha_1 \\ \vdots \\ \log\alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i,\boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i,\boldsymbol{\lambda}_k) \end{pmatrix} \right\}.$$

Note that

$$E(\mathbf{Z}_i|\mathbf{Y},\boldsymbol{\theta}_0) = \Big(\frac{\alpha_1^0 f_1(Y_i|\boldsymbol{\lambda}_1^0)}{\sum_{\ell=1}^k \alpha_\ell^0 f_\ell(Y_i|\boldsymbol{\lambda}_\ell^0)}, \cdots, \frac{\alpha_k^0 f_k(Y_i|\boldsymbol{\lambda}_k^0)}{\sum_{\ell=1}^k \alpha_\ell^0 f_\ell(Y_i|\boldsymbol{\lambda}_\ell^0)}\Big)^\tau,$$

which are constants as far as the M-step is concerned.

Now the E-step implies that

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} E(\mathbf{Z}_i^\tau|\mathbf{Y},\boldsymbol{\theta}_0) \left\{ \begin{pmatrix} \log\alpha_1 \\ \vdots \\ \log\alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i,\boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i,\boldsymbol{\lambda}_k) \end{pmatrix} \right\},$$

and the M-step requires to maximise this function, which is much easier than minimizing $l(\theta,\mathbf{Y})$ directly for, for example, normal pdf f_j's.

2.6 Evaluating estimation

To measure the accuracy of an MLE or, more general, any estimation procedure, we need to define some measures for the goodness (or badness) of an estimator.

Let $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ be an estimator of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_o$ be the (unknown) *true value* of $\boldsymbol{\theta}$. Note that

(i) exact estimation error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ is unknown, and
(ii) $\widehat{\boldsymbol{\theta}}$ is a random variable

we have to gauge the error

(i) in terms of a probability average, and
(ii) for all possible values of $\boldsymbol{\theta}_o \in \Theta$.

Let $P_{\boldsymbol{\theta}}$, $E_{\boldsymbol{\theta}}$ and $\mathrm{Var}_{\boldsymbol{\theta}}$ denote the probability distribution, expectation and variance under $\boldsymbol{\theta}_o = \boldsymbol{\theta}$.

Bias: $\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$

Variance: $\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$

Standard error: $\{\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Mean square error (MSE): $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$

Mean absolute error (MAE): $E_{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|$

Note that

- standard error is a meaningful measure of accuracy for unbiased estimators only, and
- MSE (it its squared-root) should be used in general as

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \{\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^2 + \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}).$$

Ideally we would seek for the estimator which minimises MSE or MAE **f**or all $\boldsymbol{\theta} \in \Theta$ over all possible candidate estimators. Unfortunately such a global optimum rarely exists. However if we confine to some subclass of estimators, the MLE is often optimal or asymptotically optimal.

The MSE is most frequently used largely due to is technical tractability while the MAE leads to estimators which is more robust against outliers in observations.

Fisher Information

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. The score function is

$$s(\boldsymbol{\theta}) = \dot{l}(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}.$$

We assume certain regularity conditions so that we can take derivatives under the integral sign.

Mean of $s(\boldsymbol{\theta})$:

$$\begin{aligned} E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} &= \int s(\boldsymbol{\theta}) f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{x}, \boldsymbol{\theta})\} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \ = \ 0. \end{aligned}$$

Variance of $s(\boldsymbol{\theta})$ — **F**isher information matrix:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} \\ &= E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau}\} \\ &= -E_{\boldsymbol{\theta}}\big[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\tau}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}\big], \end{aligned}$$

because

$$\begin{aligned} & E_{\boldsymbol{\theta}}\big\{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\tau}} l(\boldsymbol{\theta})\big\} \\ = & \int \frac{\ddot{L}(\boldsymbol{\theta}) L(\boldsymbol{\theta}) - \dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & \int \ddot{L}(\boldsymbol{\theta}) d\mathbf{x} - \int \frac{\dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & -\int \frac{\dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & -\int s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ = & -E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau}\}. \end{aligned}$$

Fisher information $\mathcal{I}(\boldsymbol{\theta})$ measures **t**he information on $\boldsymbol{\theta}$ contained in data $\mathbf{X}$. Further if $\mathbf{X} = (X_1, \cdots, X_n)^{\tau}$, and $X_1, \cdots, X_n$ are independent,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \mathcal{I}_{X_j}(\boldsymbol{\theta}).$$

For $\boldsymbol{\theta} = \theta$ is a scalar, the Fisher information is

$$\mathcal{I}(\theta) = E_\theta\{s(\theta)^2\} = -E_\theta\{\ddot{l}(\theta)\}.$$

Theorem 2. *(Cramér-Rao inequality)*
Let $\mathbf{X} \sim f(\cdot, \theta)$ which satisfying some regularity conditions. Let $T = T(\mathbf{X})$ be a statistic with $g(\theta) = E_\theta(T)$. Then for any $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T) \geq \{\dot{g}(\theta)\}^2/\mathcal{I}(\theta).$$

The Cramér-Rao inequality specifies **a** lower bound for any *unbiased estimator* for the parameter $g(\theta)$. When the equality holds, T is the **m**inimum variance unbiased estimator (MVUE) of $g(\theta)$.

An MVUE is a member in the class of all unbiased estimators with the minimum variance. *Not all MVUEs can attain the Cramér-Rao lower bounds.*

Example 13. Let $X_1, \cdots, X_n$ be a sample from $N(\mu, \sigma^2)$. We consider estimators for μ, treating σ^2 as known. The score function (for one observation) is

$$\begin{aligned} s(\mu; X_1) &= \frac{\partial}{\partial \mu} \log[e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2}/\sqrt{2\pi\sigma^2}] \\ = \quad & \frac{\partial}{\partial \mu}[-\frac{1}{2\sigma^2}(X_1-\mu)^2] = (X_1-\mu)/\sigma^2. \end{aligned}$$

Note $\ddot{l}(\mu) = \dot{s}(\mu) = -\sigma^{-2}$. Hence the Fisher information based on a single observation is $\mathcal{I}_{X_1}(\mu) = \sigma^{-2}$. Therefore

$$\mathcal{I}(\mu) = \mathcal{I}_{X_1,\cdots,X_n}(\mu) = n/\sigma^2.$$

For any unbiased estimator $\widehat{\mu}$ for μ, it holds that

$$\mathrm{Var}_\mu(\widehat{\mu}) \geq \sigma^2/n,$$

which is the variance of $\bar{X}$. Hence $\bar{X}$ is the MVUE for μ.

Asymptotic properties of MLEs

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \boldsymbol{\theta})$. Write

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^{n} \log f(X_j, \boldsymbol{\theta}).$$

Let $\widehat{\boldsymbol{\theta}}$ be the MLE which maximises $l(\boldsymbol{\theta})$. Suppose f fulfils certain regularity conditions.

(a) **Consistency**.
The MLE is consistent in the sense that as $n \to \infty$,

$$P_{\boldsymbol{\theta}}\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon\} \to 0$$

for any $\varepsilon > 0$.

Consistency requires that an estimator converges to the parameter to be estimated. It is a very mild and modest condition that any reasonable estimator should fulfil. The consistency condition is often used to *rule out bad estimators*.

(b) **Asymptotic normality**
As $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\big(0,\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}\big).$$

For large n, it holds **a**pproximately that

$$\widehat{\boldsymbol{\theta}} \sim N\big(\boldsymbol{\theta},\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}/n\big).$$

Therefore *asymptotically* the MLE is unbiased and attains the Cramér-Rao lower bound. Any estimator fulfilling this condition is called **e**fficient.

An approximate standard error of the j-th component of $\widehat{\boldsymbol{\theta}}$ is the square-root of the (j,j)-th element of $\{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}$ divided by $n^{1/2}$.

2.7 Bootstrapping MSEs

An MSE provides a measure for the accuracy of the estimator. The standard errors of an MLE derived from the asymptotic normality typically depend on unknown parameters, which are replaced by their estimators in practice. Alternatively we may estimate the MSE by a simulation method called Bootstrap.

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \theta)$. Let

$$\widehat{\theta} = T(X_1, \cdots, X_n)$$

be an estimator. The *goal* here is to estimate

$$\nu \equiv \{\mathrm{MSE}_{\theta_o}(\widehat{\theta})\}^{1/2},$$

where θ_o is the true value.

If we *knew* $f(\cdot, \theta_o)$ completely, ν is known in principle, and may be estimated easily via a repeated sampling as follows. We draw B independent samples of size n from $f(\cdot, \theta_o)$. For each sample, we calculate $\widehat{\theta}$, obtaining $\widehat{\theta}_1, \cdots, \widehat{\theta}_B$. Then the sample root-MSE

$$\Big\{\frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b - \theta_o)^2\Big\}^{1/2}$$

is a reasonable estimator for ν. By the LLN, this estimator converges to ν as $B \to \infty$.

The **b**asic idea of bootstrap is to adopt the above sampling procedure in the so-called *bootstrap world*: now the population is $f(\cdot, \widehat{\theta})$ which is **known**. We draw a sample denoted as $(X_1^*, \cdots, X_n^*)$ from this distribution. Define the *bootstrap version* of the estimator

$$\widehat{\theta}^* = T(X_1^*, \cdots, X_n^*).$$

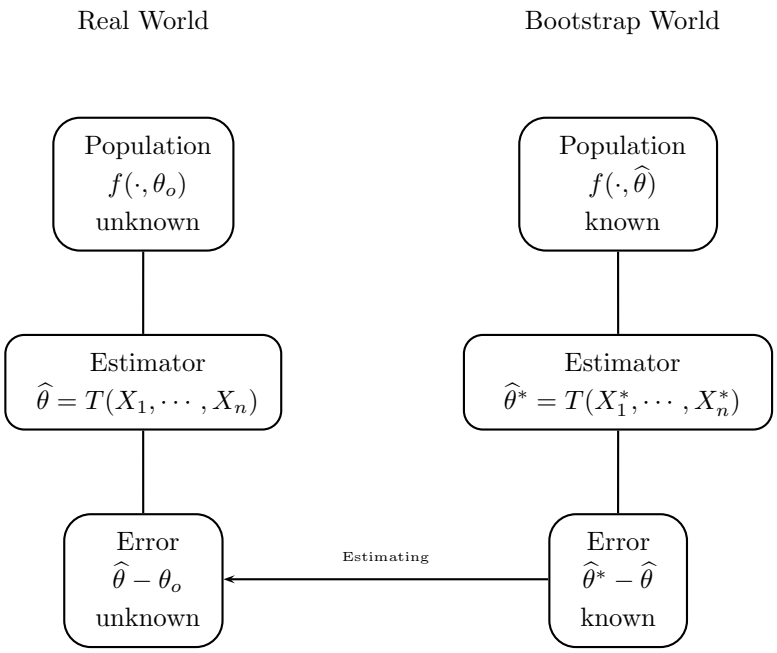
Then the quantity

$$\nu^* = \{\mathrm{MSE}_{\widehat{\theta}}(\widehat{\theta}^*)\}^{1/2}$$

is known in principle since the distribution $f(\cdot, \widehat{\theta})$ is completely known. Define ν^* as a **b**ootstrap estimator for ν.

In practice, we draw B sets samples from $f(\cdot, \widehat{\theta})$, forming B bootstrap versions of estimator $\widehat{\theta}_1^*, \cdots, \widehat{\theta}_B^*$. The ν^* is calculated as

$$\nu^* = \Big\{\frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_j^* - \widehat{\theta})^2\Big\}^{1/2}.$$



Bootstrap is a powerful tool for statistical inference. It has different forms for different applications. The one introduced here is in the form of *parametric bootstrap*. The special monographs on this topic include:

- Hall, P. (1992). The bootstrap and Edgeworth expansion. Springer, New York.
- Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman and Hall, New York.
- Shao, J. and Tu, D. (1995). The jackknife and bootstrap. Springer, New York.
- Davison, A.C. and Hinkley, D.V. (1997). Bootstrap methods and their application. Cambridge University Press, Cambridge.

3. TESTING HYPOTHESES

3.1 Hypothesis testing: basic setting

Statistical estimation and tests address *different kind* of practical problems.

We work with a family of distributions $f(\mathbf{x};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$, and $f(\mathbf{x};\boldsymbol{\theta})$ is completely known apart from $\boldsymbol{\theta}$.

Basic setting:
A statistical test make a decision between two sets of hypotheses on unknown parameter $\boldsymbol{\theta}$, namely

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \Theta_1,$$

where Θ_0, Θ_1 are subsets of Θ. We always assume

$$\Theta_0 \cap \Theta_1 = \emptyset \text{ (empty set)},$$

i.e. H_0 and H_1 are distinct, mutually exclusive, although we do not always require $\Theta_0 \cup \Theta_1 = \Theta$.

Based on a sample $\mathbf{X} = (X_1, \cdots, X_n)^\tau$, a statistical test is to make a decision either

to reject H_0, or
not to reject H_0.

In latter case, we claim that there is **n**o significant difference between the null hypothesis and the underlying distribution according to the observed data.

Remark. The setting is **n**ot symmetric in the sense that the two hypotheses H_0 and H_1 will be treated differently.

Some specific terms used statistical tests are now in order:

- A *Hypothesis* is an assumption made about the value of $\boldsymbol{\theta}$.
- A *Simple Hypothesis* completely specifies $\boldsymbol{\theta}$ to a known constant such as H: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.
- A *Composite Hypothesis* is any hypothesis that is not simple. For instance the hypothesis H: $\boldsymbol{\theta} \in [1, \infty)$.
- A *Test Statistic* is a statistic based on which the decision is made.
- A *Null Hypothesis*, usually denoted as H_0, is the basic or default hypothesis.
- The *Alternative Hypothesis*, usually denoted as H_1, is a hypothesis that is to be compared with the Null Hypothesis.

Mathematically speaking, a statistical test is equivalent to define a binary function

$$\phi(\mathbf{X}) = \begin{cases} 1 & \text{when reject } H_0, \\ 0 & \text{when not reject } H_0. \end{cases}$$

$\phi(\cdot)$ is called a **d**ecision rule, which practically divides the sample space of $\mathbf{X}$ into two subspaces: *rejection region* (which is also called '**c**ritical region') and its compliment which is often **imprecisely** labelled as *'acceptance' region*.

There are two types of errors that can be made in a statistical test, which are displayed in the table below.

		Decision Made	
		H_0 not rejected Θ_0	H_0 rejected Θ_1
True State of Nature	H_0 Θ_0	Correct Decision	Type I Error
	H_1 Θ_1	Type II Error	Correct Decision

Ideally we would like to have a test that minimises the probabilities of making both types of errors, which unfortunately is not feasible.

Construction of a statistical test:
we control the probability of Type I error under a given level, say, α, and then we minimize the probability of Type II error.

Definition. A test is said to have *significance level* α if

$$\sup_{\boldsymbol{\theta}\in\Theta_0} P_{\boldsymbol{\theta}}(H_0 \text{ is rejected}) \leq \alpha.$$

Definition. A test T is said to be of *size* α if

$$\sup_{\boldsymbol{\theta}\in\Theta_0} P_{\boldsymbol{\theta}}(H_0 \text{ is rejected}) = \alpha.$$

Remark. (i) In practice, we usually take the significance level $\alpha = 0.1, 0.05$ or 0.01.
(ii) The size of the test is no greater than the significance level.

Definition. The *power function* of a test is defined as

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(H_0 \text{ is rejected}), \quad \boldsymbol{\theta} \in \Theta_1.$$

In practice, it is useful to extend $\beta(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$. For $\boldsymbol{\theta} \in \Theta_0$, $\beta(\boldsymbol{\theta})$ must be smaller than the prescribed significance level, and is the probability of making a Type I error at the parameter value $\boldsymbol{\theta}$. For $\boldsymbol{\theta} \in \Theta_1$, $1 - \beta(\boldsymbol{\theta})$ is the probability of making a Type II error.

If a test is defined in terms of decision rule ϕ, then

$$\beta(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\phi(\mathbf{X})],$$

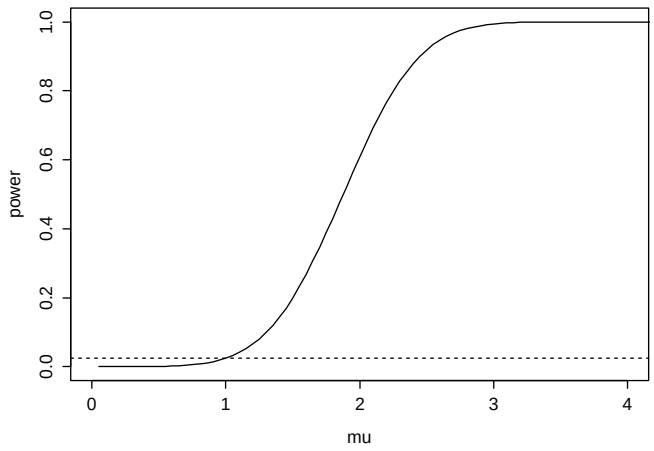
because

$$\begin{aligned} E[\phi(\mathbf{X})] &= \int \phi(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}) d\mathbf{x} \\ &= P_{\boldsymbol{\theta}}\{\phi(\mathbf{X}) = 1\} = P_{\boldsymbol{\theta}}\{H_0 \text{ is rejected}\}. \end{aligned}$$

Example 1. Suppose that we have a random sample of size 5 from a normal distribution with mean μ and variance 1. Let us consider:

$$\begin{aligned} H_0&: \quad \mu = 1 \\ H_1&: \quad \mu > 1 \end{aligned}$$

One possible test statistic is $\bar{X}$. Intuitively we would reject H_0 for large values of $\bar{X}$, say, $\bar{X} > K$. The constant K is determined by the prescribed significance level α. Let $\alpha = 0.025$. Since $\sqrt{5}(\bar{X} - 1) \sim N(0, 1)$, $K = 1 + 1.96/\sqrt{5}$. We reject H_0 if $\bar{X} > 1 + 1.96/\sqrt{5}$. The power function is shown below.



The power is small for $\mu = 1$, and rises quite fast as μ increases past 1.

Question: What should we do if we change the alternative hypothesis to $H_1 : \mu < 1$ or $H_1 : \mu \neq 1$.

The above null hypothesis is simple, because $\Theta_0 = \{1\}$, but the alternative hypothesis is composite because $\Theta_1 = (1, \infty)$. If instead we used the Null Hypothesis H_0: $\mu \leq 1$, that would be a composite null hypothesis. It is convenient to have one value to summarise the probability of Type I error, even when the hull hypothesis is composite.

Question: Is the test constructed above still reasonable for testing

$$H_0 : \ \mu \leq 1 \quad \text{vs} \quad H_1 : \mu > 1 \ ?$$

If so, at which value of μ under H_0 does the probability of type I error obtain the maximum?

3.2 Neyman-Pearson Lemma for testing simple hypotheses

Most powerful test (MPT): among all the tests of size α, the one with maximum power $\beta(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta_1$ is called the MPT. When Θ_1 consists of more than one point, it is often called the uniformly MPT (UMPT).

Note. UMPT may not exist.

Let $L(\boldsymbol{\theta}; \mathbf{X})$ be the likelihood function with observation $\mathbf{X}$.

Theorem 1. (*Neyman-Pearson Lemma*)
The MPT for hypotheses

$$H_0 : \ \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{against} \quad H_1 : \ \boldsymbol{\theta} = \boldsymbol{\theta}_1$$

rejects H_0 when

$$L(\boldsymbol{\theta}_1; \mathbf{X}) > KL(\boldsymbol{\theta}_0; \mathbf{X}),$$

and does not reject H_0 when

$$L(\boldsymbol{\theta}_1; \mathbf{X}) < KL(\boldsymbol{\theta}_0; \mathbf{X}),$$

where constant K is determined by the size of the test.

Remark. In general the MPT rejects H_0 if the **l**ikelihood ratio

$$LR(\mathbf{X}) = \frac{L(\boldsymbol{\theta}_1; \mathbf{X})}{L(\boldsymbol{\theta}_0; \mathbf{X})} > K,$$

which indicates that $\boldsymbol{\theta}_1$ is relatively favoured over $\boldsymbol{\theta}_0$ according to the likelihood function.

Proof. For each test T, we define the decision rule $\phi_T(\mathbf{x})$ as follows

$$\phi_T(\mathbf{X}) = \begin{cases} 1 & \text{Reject } H_0 \\ 0 & \text{Do not Reject } H_0. \end{cases}$$

The power $\beta(\boldsymbol{\theta})$ is equal to $E\phi_T(\mathbf{X})$.

Suppose that T is a test of size α that satisfies the conditions of the Neyman-Pearson Lemma, and S is some other test of size α. Then

$$\begin{aligned}
\beta_S(\boldsymbol{\theta}_1) = E\phi_S(\mathbf{x}) &= \int \phi_S(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}_1) d\mathbf{x} \\
&= \int \phi_S(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}_1) d\mathbf{x} - K \int \phi_S(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}_0) d\mathbf{x} + K\alpha \\
&= \int \phi_S(\mathbf{x}) [f(\mathbf{x}; \boldsymbol{\theta}_1) - K f(\mathbf{x}; \boldsymbol{\theta}_0)] d\mathbf{x} + K\alpha \\
&= \int \phi_S(\mathbf{x}) [L(\boldsymbol{\theta}_1; \mathbf{x}) - K L(\boldsymbol{\theta}_0; \mathbf{x})] d\mathbf{x} + K\alpha \\
&\leq \int_{\{L(\boldsymbol{\theta}_1;\mathbf{x}) > K L(\boldsymbol{\theta}_0;\mathbf{x})\}} \phi_S(\mathbf{x}) [L(\boldsymbol{\theta}_1; \mathbf{x}) - K L(\boldsymbol{\theta}_0; \mathbf{x})] d\mathbf{x} + K\alpha \\
&\leq \int \phi_T(\mathbf{x}) [L(\boldsymbol{\theta}_1; \mathbf{x}) - K L(\boldsymbol{\theta}_0; \mathbf{x})] d\mathbf{x} + K\alpha \\
&= \int \phi_T(\mathbf{x}) [f(\mathbf{x}; \boldsymbol{\theta}_1) - K f(\mathbf{x}; \boldsymbol{\theta}_0)] d\mathbf{x} + K\alpha \\
&= \int \phi_T(\mathbf{x}) f(\mathbf{x}; \boldsymbol{\theta}_1) d\mathbf{x} - K \int \phi_T f(\mathbf{x}; \boldsymbol{\theta}_0) d\mathbf{x} + K\alpha \\
&= \beta_T(\boldsymbol{\theta}_1) - K\alpha + K\alpha \\
&= \beta_T(\boldsymbol{\theta}_1).
\end{aligned}$$

Example 2. Suppose that $X_1, X_2, \ldots, X_n$ are a random sample from $N(\mu, \sigma^2)$, where we treat σ^2 as a known quantity. Suppose we want to test:

$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu = 5.$$

The likelihood ratio is

$$\begin{aligned}
LR &= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n (X_i - 5)^2/(2\sigma^2)\right)}{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^n X_i^2/(2\sigma^2)\right)} \\
&= \exp\left(5n(2\bar{X} - 5)/(2\sigma^2)\right).
\end{aligned}$$

We can look for a most powerful test by noting that $LR > K$ is equivalent to $\bar{X} > K_1$, and *the critical value* K_1 is determined by the size of the test. So, if we find a test of this form which has size α, it will

be a most powerful test of size α. It is easy to see that the test that rejects H_0 iff $\bar{X} > z_\alpha \sigma/\sqrt{n}$ is of this form and is therefore a most powerful test of size α.

Question: If we change the alternative hypothesis to $H_1 : \mu = 50$, what is the MPT then?

Example 3. Let $(X_1, \cdots, X_n)$ be a sample from Poisson distribution with mean μ. To test

$$H_0 : \ \mu = 2 \quad \text{against} \quad H_1 : \ \mu = 6,$$

the likelihood ratio is

$$LR = \frac{6^{\sum X_i} e^{-6n}/(X_1! \cdots X_n!)}{2^{\sum X_i} e^{-2n}/(X_1! \cdots X_n!)} = 3^{\sum X_i} e^{-4n}.$$

According to the N-P lemma, the MPT will reject H_0 if $LR > K$, or equivalently, $\sum_i X_i > K_1$, where the critical value K_1 is determined by the significance level of the test.

Suppose $n = 4$, then $\sum_{i=1}^4 X_i \sim Poisson(8)$ under H_0. From the table for Poisson CDFs, the size of test $\alpha = 0.064$ with $K_1 = 12$, and $\alpha = 0.034$ with $K_1 = 13$.

When $n = 8$, $\sum_{i=1}^8 X_i \sim Poisson(16)$ under H_0. The size of test $\alpha = 0.058$ with $K_1 = 22$, and $\alpha = 0.037$ with $K_1 = 23$.

3.3 Uniformly Most Powerful Tests

Although the Neyman-Pearson Lemma was designed for testing simple hypotheses, it is possible to use it to construct the UMPT for, for example, *one-sided null hypothesis* against *one-sided alternative*.

A general setting: For hypotheses

$$H_0 : \ \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \ \theta \in \Theta_1, \tag{4}$$

$\phi(\mathbf{X})$ is the decision rule of a test at significant level α, i.e.

$$E_\theta\{\phi(\mathbf{X})\} \le \alpha \quad \text{for all } \theta \in \Theta_0.$$

If $\phi(\mathbf{X})$ is also the MPT of size α for *simple hypotheses*

$$H_0 : \ \theta = \theta_0 \quad \text{vs} \quad H_1 : \ \theta = \theta_1$$

for **some** $\theta_0 \in \Theta_0$ and **all** $\theta_1 \in \Theta_1$. Then $\phi(\mathbf{X})$ is the UMPT for hypotheses (4).

We look at a simple scenario first — *UMPTs for simple H_0 and one-sided H_1*.

Suppose the MPT for simple hypotheses

$$H_0 : \ \theta = \theta_0 \quad \text{against} \quad H_1 : \ \theta = \theta_1$$

does not change its form for all $\theta_1 \in \Theta_1$. Then it is also the UMPT for

$$H_0 : \ \theta = \theta_0 \quad \text{against} H_1 : \ \theta \in \Theta_1.$$

Such an UMPT often exists for $\Theta_1 = (\theta_0, \infty)$ or $\Theta_1 = (-\infty, \theta_0)$.

Example 4. If $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, 1)$, then the most powerful test for testing

$$H_0 : \mu = 1 \quad \text{vs} \quad H_1 : \mu = 2$$

is obtained by rejecting H_0 if and only if $\bar{X} > 1 + z_\alpha/\sqrt{n}$. The same test is obtained for

$$H_0 : \mu = 1 \quad \text{vs} \quad H_1 : \mu = \mu_1$$

for every $\mu_1 > 1$. So this test is uniformly most powerful for

$$H_0 : \mu = 1 \quad \text{vs} \quad H_1 : \mu > 1.$$

Although the UMPT does not depend on the value μ_1 specified under H_1, its power varies over $\{\mu|\mu > 1\}$. In fact

$$\begin{aligned} & \beta(\mu) = P_\mu\{\bar{X} > 1 + z_\alpha/\sqrt{n}\} \\ = \; & P_\mu\{\sqrt{n}(\bar{X} - \mu) > \sqrt{n}(1 - \mu) + z_\alpha\} \\ = \; & 1 - \Phi\{z_\alpha - \sqrt{n}(\mu - 1)\} = \Phi\{\sqrt{n}(\mu - 1) - z_\alpha\}, \end{aligned}$$

where $\Phi(\cdot)$ is the CDF of N(0,1). Thus the power increases as μ increases.

Note that the test is not the UMPT for $H_1 : \mu \neq 1$.

Note. The UMPT usually does not exist for two-sided (composite) alternative hypothesis.

Example 5. Let Y have the binomial distribution $Bin(n, p)$. Find the UMPT for testing

$$H_0 : p = p_0 \quad \text{vs} \quad H_1 : p > p_0.$$

Let $p_1 > p_0$. The LR for testing the H_0 above against $H_1 : p = p_1$ is

$$\begin{aligned} LR &= \frac{\binom{n}{Y} p_1^Y (1 - p_1)^{n-Y}}{\binom{n}{Y} p_0^Y (1 - p_0)^{n-Y}} \\ &= \left(\frac{p_1(1 - p_0)}{p_0(1 - p_1)}\right)^Y \left(\frac{1 - p_1}{1 - p_0}\right)^n \end{aligned}$$

Note that $p_1(1-p_0) > p_0(1-p_1)$ since $p_1 > p_0$. Thus $LR > K$ is equivalent to $Y > K_1$. Thus the UMPT rejects H_0 iff $Y > K_1$ with the size

$$P(Y > K_1 | p = p_0).$$

Example 6. Let $(X_1, \cdots, X_n)$ be a random sample from an exponential distribution with mean $1/\lambda$. We are interested in testing

$$H_0 : \; \lambda \leq \lambda_0 \quad \text{vs} \quad H_1 : \; \lambda > \lambda_0.$$

For

$$H_0 : \; \lambda = \lambda_0 \quad \text{vs} \quad H_1 : \; \lambda = \lambda_1,$$

the MPT rejects H_0 iff $\sum_{i=1}^n X_i \leq K$ for any $\lambda_1 > \lambda_0$, where K is determined by

$$P_{\lambda_0}\{\sum_{i=1}^n X_i < K\} = \alpha.$$

It is easy to verify that for $\lambda < \lambda_0$,

$$P_\lambda\{\sum_{i=1}^n X_i < K\} < \alpha.$$

Hence the MPT for the simple null hypothesis against simple alternative is also the UMPT for the composite hypotheses.

3.4 Likelihood Ratio Tests

We now deal with the most popular ways of constructing tests when both null and alternative hypotheses are composite. There are no guaranteed optimality properties in small samples for these tests, but in regular cases they usually have good power for large sample sizes.

Let $\mathbf{X} \sim f(\cdot, \boldsymbol{\theta})$. Consider hypotheses

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \Theta - \Theta_0.$$

The likelihood ratio test will reject H_0 for the large values of the statistic

$$\begin{aligned} LR \;=\; LR(\mathbf{X}) \quad &\equiv \quad \frac{\sup_{\boldsymbol{\theta}\in\Theta} f(\mathbf{X}, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta_0} f(\mathbf{X}, \boldsymbol{\theta})} \\ &= \quad f(\mathbf{X}, \widehat{\boldsymbol{\theta}}) / f(\mathbf{X}, \widetilde{\boldsymbol{\theta}}), \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}$ the (unconstrained) MLE, and $\widetilde{\boldsymbol{\theta}}$ is the constrained MLE under hypothesis H_0.

Remark. (i) It is easy to see that $LR \geq 1$.

(ii) The exact sampling distributions of LR are usually unknown, except in a few special cases.

Example 7. **(**One-sample t-test)
Let $\mathbf{X} = (X_1, \cdots, X_n)^\tau$ be a random sample from $N(\mu, \sigma^2)$. We are interested in testing hypotheses

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0,$$

where μ_0 is given, and σ^2 is unknown and is a nuisance parameter. Now both H_0 and H_1 are composite. The likelihood function is

$$L(\mu, \sigma^2) = C\sigma^{-n} \exp\big\{ -\frac{1}{2}\sigma^2 \sum_{j=1}^{n} (X_i - \mu)^2 \big\}.$$

The unconstrained MLEs are

$$\widehat{\mu} = \bar{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} (X_i - \bar{X})^2,$$

and the constrained MLE is

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} (X_i - \mu_0)^2.$$

The LR-ratio statistic is then

$$LR = \frac{L(\widehat{\mu}, \widehat{\sigma}^2)}{L(\mu_0, \widetilde{\sigma}^2)} = (\widetilde{\sigma}^2/\widehat{\sigma}^2)^{n/2}.$$

Since

$$n\widetilde{\sigma}^2 = n\widehat{\sigma}^2 + n(\bar{X} - \mu_0)^2,$$

it holds that $\widetilde{\sigma}^2/\widehat{\sigma}^2 = 1 + T^2/(n-1)$, where

$$T = \sqrt{n}(\bar{X} - \mu_0) / \big\{ \frac{1}{n-1} \sum_{j=1}^{n} (X_i - \bar{X})^2 \big\}^{1/2}.$$

Note that $T \sim t_{n-1}$ under H_0. The LRT will reject H_0 iff $|T| > t_{n-1,\alpha/2}$, where $t_{k,\alpha}$ is the upper α-point of the t-distribution with k degrees of freedom.

Asymptotic Distribution of Likelihood ratio test statistic
Let $\mathbf{X} = (X_1, \cdots, X_n)^\tau$, and assume certain regularity conditions. Then as $n \to \infty$, the distribution

of $2\log(LR)$ under H_0 converges to the χ^2-distribution with $d - d_0$ degrees of freedom, where d is the 'dimension' of Θ and d_0 is the 'dimension' of Θ_0.

To make the computation of 'dimension' easy, **r**eparametrisation is often adopted. Suppose that the parameter θ may be written in two parts

$$\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$$

where $\boldsymbol{\psi}$ is $k \times 1$ parameter of interest, and $\boldsymbol{\lambda}$ is of little interest and is called *nuisance parameters*. The hypotheses to be tested may be expressed as

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0 \quad \text{vs} \quad H_1 : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0.$$

Now the LR-statistic is of the form

$$LR = \frac{L(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}}; \mathbf{X})}{L(\boldsymbol{\psi}_0, \widetilde{\boldsymbol{\lambda}}; \mathbf{X})},$$

where $(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}})$ is unconstrained MLE while $\widetilde{\boldsymbol{\lambda}}$ is the constrained MLE of $\boldsymbol{\lambda}$ subject to $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Then as $n \to \infty$,

$$2\log(LR) \xrightarrow{D} \chi^2_k \quad \text{under } H_0.$$

Example 8. Let $X_1, \cdots, X_n$ be independent, and $X_j \sim N(\mu_j, 1)$. Consider the null hypothesis

$$H_0 : \mu_1 = \cdots = \mu_n.$$

The likelihood function is

$$L(\mu_1, \cdots, \mu_n) = C \exp\Big\{ -\frac{1}{2}\sum_{j=1}^{n}(X_j - \mu_j)^2 \Big\},$$

where $C > 0$ is a constant independent of μ_j. Then the unconstrained MLE are $\widehat{\mu}_j = X_j$ and the constrained MLE is $\widetilde{\mu} = \bar{X}$. Hence

$$\begin{aligned} LR &= \frac{L(\widehat{\mu}_1, \cdots, \widehat{\mu}_n)}{L(\widetilde{\mu}, \cdots, \widetilde{\mu})} \\ &= \exp\Big\{ \frac{1}{2}\sum_{j=1}^{n}(X_j - \bar{X})^2 \Big\}. \end{aligned}$$

Hence

$$2\log(LR) = \sum_{j=1}^{n}(X_j - \bar{X})^2 \sim \chi^2_{n-1} \quad \text{under } H_0,$$

which is true for any finite n as well.

How to *calculate the degree of freedom*?

Since

$$d = n, \qquad d_0 = 1,$$

the d.f. is $d - d_0 = n - 1$.

Alternatively we may adopt the following
reparametrisation:

$$\mu_j = \mu_1 + \psi_j \quad \text{for } 2 \leq j \leq n.$$

Then the null hypothesis can be expressed as

$$H_0 : \psi_2 = \cdots = \psi_n = 0.$$

Therefore $\boldsymbol{\psi} = (\psi_2, \cdots, \psi_n)^\tau$ has $n-1$ component, i.e. $k = n-1$.

4. INTERVAL ESTIMATION

Interval estimation is more informative than point estimation, and is very important in practice.

Confidence Interval is the most commonly used interval estimation. More general type is *Confidence Set* which may consist of several intervals.

4.1 What is an interval estimator

Example 1. Let us start with a simple example. A random sample $X_1, \cdots, X_n$ are drawn from $N(\mu, 1)$. Then

$$\sqrt{n}(\bar{X} - \mu) \sim N(0, 1).$$

Hence

$$P(-1.96 \leq \sqrt{n}(\bar{X} - \mu) \leq 1.96) = 0.95,$$

or

$$P(\bar{X} - 1.96/\sqrt{n} < \mu < \bar{X} + 1.96/\sqrt{n}) = 0.95.$$

So a 95% confidence interval for μ is

$$(\bar{X} - 1.96/\sqrt{n},\ \bar{X} + 1.96/\sqrt{n}).$$

Suppose $n = 4$, $\bar{X} = 2.25$. Then a 95% C.I. is $(2.25 - 0.98, 2.25 + 0.98) = (1.27, 3.23)$.

Question: what is $P(1.27 < \mu < 3.23)$? — Note μ is a unknown constant!

Answer: $(1.27, 3.23)$ is one realisation of the **r**andom interval $(\bar{X} - 0.98,\ \bar{X} + 0.98)$ which covers μ with probability 0.95.

Definition. Suppose the joint distribution of $\mathbf{X} = (X_1, \cdots, X_n)$ depends some parameter θ.

If $\theta_l(\mathbf{X})$ and $\theta_u(\mathbf{X})$ are two statistics for which

$$P\{\theta_l(\mathbf{X}) < \theta < \theta_u(\mathbf{X})\} = 1 - \alpha,$$

$(\theta_l(\mathbf{X}),\ \theta_u(\mathbf{X}))$ is called a $100(1-\alpha)\%$ *confidence interval* for θ.

Remark. (i) $1 - \alpha$ is called the confidence level, which is usually set at 0.95, 0.99 or 0.999. Naturally for given α, we shall search for the interval with the shortest length $\theta_u(\mathbf{X}) - \theta_l(\mathbf{X})$, which gives the **most accurate** estimation.

(ii) We may have $\theta_l(\mathbf{X}) = -\infty$ or $\theta_u(\mathbf{X}) = \infty$, giving a one-sided interval.

(iii) In general we may use non-interval type set $S(\mathbf{X}) \subset \Theta$ as the estimator for θ, i.e.

$$P\{\theta \in S(\mathbf{X})\} = 1 - \alpha.$$

We call $S(\mathbf{X})$ a $100(1-\alpha)\%$ **c**onfidence set.

4.2 Method of pivotal functions

Set-up: Observations – $\mathbf{X}$, Parameter of interest – θ (possible other parameters).

Definition. A function of $\mathbf{X}$ and θ alone is a *pivotal function* if its distribution does not depend on any unknown parameters (i.e. can be calculated numerically).

Examples of pivotal functions: Let $X_1, \cdots, X_n$ be a random sample.

1. for exponential distribution with mean θ, $\theta^{-1}\sum_j X_j$ is a pivotal,

2. for $N(\mu,1)$, $\bar{X}-\mu$ is a pivotal,

3. for $N(\mu,\sigma^2)$, $\sqrt{n}(\bar{X}-\mu)/S$ is a pivotal.

Construction of C.I. based on a pivotal:

Step 1: Find a pivotal function $T=T(\mathbf{X},\theta)$, and identify its distribution

Step 2: Use the distribution to identify a range of values t_1,t_2 such that

$$P(t_1<T<t_2)=1-\alpha.$$

Step 3: Manipulate the inequalities

$$T>t_1 \quad \text{and} \quad T<t_2$$

to find a set of values for θ. The values included in this set form a $100(1-\alpha)\%$ C.I. for θ.

Example 2. (*Student's t-interval*)
$(X_1,\cdots,X_n)$ is a random sample from $N(\mu,\sigma^2)$. A pivotal function is

$$T=\sqrt{n}(\bar{X}-\mu)/S\sim t_{n-1},$$

where $S^2=\frac{1}{n-1}\sum_j(X_j-\bar{X})^2$.

For $n=10$, we have $P(-2.26<T<2.26)=0.95$. Note that $T>-2.26$ is equivalent to

$$\mu<\bar{X}+2.26S/\sqrt{n}.$$

(Similar for $T<2.26$.) A 95% C.I. for μ is $(\bar{X}-2.26S/\sqrt{n},\ \bar{X}+2.26S/\sqrt{n})$.

Note. (i) In general it is not easy to identify a pivotal function.

(ii) For location parameter μ, $\bar{X}-\mu$ is **likely** to be pivotal.

(iii) The MLE is asymptotically pivotal:

$$\widehat{\theta}\sim N(\theta,\mathcal{I}^{-1}(\theta)).$$

Hence, $\sqrt{\mathcal{I}(\theta)}(\widehat{\theta}-\theta)\sim N(0,1)$ approximately for large n.

4.3 Inverting Tests

One easy way to get a confidence set is to invert a test; see Example 1.

Theorem. Suppose that we have a size α test for the Null Hypothesis $H_0:\ \theta=\theta_0$. For each $\theta_0\in\Theta$, suppose the set $A(\theta_0)$ is the collection of $\mathbf{x}$ for which H_0 is *not rejected*. Then the set

$$S(\mathbf{x})=\{\theta:\mathbf{x}\in A(\theta),\theta\in\Theta\}$$

is a family of confidence sets for θ at confidence level $1-\alpha$.

The proof of this is obvious, because when θ is the true parameter value

$$P_\theta[\theta\in S(\mathbf{X})]=P_\theta[\mathbf{X}\in A(\theta)]=1-\alpha.$$

Example 3. Suppose that $X_1,X_2,\ldots,X_n$ is a random sample from an exponential distribution with mean $1/\lambda$. The UMPT with size 5% for

$$H_0:\ \lambda=\lambda_0 \quad \text{against} \quad H_1:\ \lambda>\lambda_0$$

is to reject H_0 when $\sum X_i < \frac{1}{2\lambda_0}\chi^2_{(2n),0.05}$, where $\chi^2_{(2n),0.05}$ is the lower 5% point of the distribution χ^2_{2n}. The acceptance region here is $\sum X_i > \frac{1}{2\lambda_0}\chi^2_{(2n),0.05}$, which on inversion gives

$$\left(\frac{\chi^2_{(2n),0.05}}{2\sum X_i},\ \infty\right)$$

as the confidence set for λ at the 95% level of confidence.

4.4 Bootstrap confidence intervals

Let $X_1, \cdots, X_n$ be sample from unknown distribution of F. We are interested in constructing confidence intervals for some characteristic $\theta = \theta(F)$ of distribution F.

We adopt the so-called *nonparametric bootstrap method* now. It draws a bootstrap sample

$$X_1^*, \cdots, X_n^*$$

independently from the uniform distribution over n discrete points $\{X_1, \cdots, X_n\}$.

Now in principle the distribution of any statistic based on $(X_1^*, \cdots, X_n^*)$ is known, *conditionally on* $(X_1, \cdots, X_n)$.

4.4.1 Bootstrap percentiles

Let $\widehat{\theta} = T(X_1, \cdots, X_n)$ be an estimator for θ. Define

$$\widehat{\theta}^* = T(X_1^*, \cdots, X_n^*).$$

Let l_α^* and u_α^* be, respectively, the lower and upper α-point of the distribution $\widehat{\theta}^*$, i.e.

$$P\{\widehat{\theta}^* \le l_\alpha^* | X_1, \cdots, X_n\} = \alpha,$$

$$P\{\widehat{\theta}^* > u_\alpha^* | X_1, \cdots, X_n\} = \alpha.$$

Then **the** $(1-\alpha)$**100%-th bootstrap interval** for θ is defined as

$$(l_{\alpha/2}^*,\ u_{\alpha/2}^*].$$

In practice we draw B sets of bootstrap samples for some large B, resulting in B bootstrap estimates $\widehat{\theta}_1^*, \cdots, \widehat{\theta}_B^*$. Then $l_{\alpha/2}^*$ and $u_{\alpha/2}^*$ are, respectively, the $[B\alpha/2]$-th smallest and the $[B\alpha/2]$-th largest values among $\widehat{\theta}_1^*, \cdots, \widehat{\theta}_B^*$.

Note. Both $l_{\alpha/2}^*$ and $u_{\alpha/2}^*$ are statistics depending on $\{X_1, \cdots, X_n\}$.

Intuitively, we **m**ust start with a *good* estimator $\widehat{\theta}$. In fact, if $\widehat{\theta} = T(X_1, \cdots, X_n)$ fulfils some conditions, it holds that as $n \to \infty$,

$$P(l_{\alpha/2}^* < \theta \le u_{\alpha/2}^*) \to 1 - \alpha.$$

4.4.2 Bootstrap-t

Suppose we have a *studentised 'pivot'*

$$V = (\widehat{\theta} - \theta)/\widehat{\sigma},$$

where $\widehat{\theta}$ is an estimator for θ, and $\widehat{\sigma}^2$ is an estimator for $\mathrm{Var}(\widehat{\theta})$. Define its bootstrap version

$$V^* = (\widehat{\theta}^* - \widehat{\theta})/\widehat{\sigma}^*.$$

Let

$$\begin{aligned} \alpha/2 &= P\{V^* \le l^*_{\alpha/2} | X_1, \cdots, X_n\} \\ &= P\{V^* > u^*_{\alpha/2} | X_1, \cdots, X_n\}. \end{aligned}$$

Then the $100(1-\alpha)\%$-th bootstrap-t interval for θ is

$$[\widehat{\theta} - \widehat{\sigma} u^*_{\alpha/2},\ \widehat{\theta} - \widehat{\sigma} l^*_{\alpha/2}).$$

Remark. The estimator $\widehat{\sigma}^2$ may be obtained by bootstrap, then $\widehat{\sigma}^{2*}$ will be calculated via a **n**ested bootstrap.

Ideally we require that the distribution, or at least the asymptotic distribution, of V does not depend on anything unknown. This will make the approximation

$$P\big\{\widehat{\theta} - \widehat{\sigma} u^*_{\alpha/2} \le \theta < \widehat{\theta} - \widehat{\sigma} l^*_{\alpha/2}\big\} \approx \alpha$$

more accurate.

We give a heuristic argument for the above assertion.

If V is an asymptotic pivot in the sense that

$$P(V \le x) = \Phi(x) + \frac{1}{\sqrt{n}} g(x, \theta) + O(\frac{1}{n}), \tag{5}$$

similarly in the bootstrap world we have

$$\begin{aligned} P(V^* \le x | X_1, \cdots, X_n) &= \Phi(x) + \frac{1}{\sqrt{n}} g(x, \widehat{\theta}) + O_p(\frac{1}{n}) \\ &= \Phi(x) + \frac{1}{\sqrt{n}} g(x, \theta) + O_p(\frac{1}{n}). \end{aligned} \tag{6}$$

Hence

$$l_\alpha = l^*_\alpha + O_p(\frac{1}{n}).$$

In contrast if $\Phi(x)$ depends on θ, then $\Phi(x)$ should be replaced by $\Phi(x, \theta)$ in (5), and by $\Phi(x, \widehat{\theta})$ in (6). Now

$$l_\alpha = l^*_\alpha + O_p(\frac{1}{\sqrt{n}}).$$

