An introduction to Mathematical Statistics
Lecture notes, Spring 2026
(Version: May 31, 2025)

Pu-Zhao Kow

Department of Mathematical Sciences, National Chengchi University
Email address: `pzkow@g.nccu.edu.tw`

Preface

We will use this lecture note as the main teaching material for the course *Statistics* (701007001), for undergraduate levels, during Spring 2026 (114-2). Since the target audiences are undergraduate students from Department of Mathematical Sciences, this lecture note will focus on mathematical statistics (inferential statistics) rather than descriptive statistics. We will follow the outline in [DBC21][1] and we will also explain the mathematical principle based on [Dur19][2]. One also can take a look on supplementary materials [SD15][3] (Exercises with solutions can be found in [HSPW14][4]) as well as [LC98][5]. One can download the above mentioned three monographs under National Chengchi University's IP. If it feels too difficult, you can consult some free materials in MIT (Massachusetts Institute of Technology) Open Course Ware[6] as well as some commercial textbooks [LM21, WMS07], but unfortunately you have to pay for these two commercial textbooks. The lecture note may updated during the course.

Title. Statistics (Spring 2025, 3 credits)

Lectures. TBA

Language. Chinese and English. Materials will be prepared in English.

Instructor. Pu-Zhao Kow (Email: `pzkow@g.nccu.edu.tw`)

Office hour. TBA

Completion. Homework Assignments 30%, Midterm Exam 40%, Final Exam 40%

[1] Access provided by National Chengchi University

[2] Access provided by National Chengchi University. See also authors' page https://services.math.duke.edu/~rtd/PTE/pte.html

[3] Access provided by Taiwan TAEBC eBook Consortium 2015

[4] Access provided by Taiwan TAEBDC eBook Consortium 2014

[5] Access provided by National Chengchi University

[6] https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/download/

Contents

Probability versus statistics

Probability and statistics are two closely related fields in mathematics that are sometimes combined for academic purposes. Despite both probability and statistics about random processes, however they are two different subjects, see the following table:

Probability	Statistics
a branch of pure mathematics	a branch of applied mathematics
studies the consequences of mathematical definitions	tries to make sense of observations in the real world
a logically self contained theory to compute probabilities	make probabilistic inferences/reasoning based on samples (real/experimental data)
predicting the likelihood of future events	analyzing the frequency of past events

TABLE 1. Probability versus Statistics

For example, if one tosses a given coin for 100 times, how many times of head/tail you expect to get? It is possible that, says, I got 51 heads and 49 tails, but you may obtain 48 heads and 52 tails. After several experiments, we may expect that one should obtain 50 heads and 50 tails, and believe that the coin is "fair" in the sense of "the probability of obtaining head/tail is $1/2$". What we have done above is *analyzing the frequency of past events*, or *quantify uncertainties*, therefore the *statistics*.

An a comparison, if you given a coin, and it was tosses for 100 times, how many heads/tails you predicted? If we *assume that* the coin is "fair", i.e. the probability of obtaining head/tail is $1/2$, then we predict that 50 heads will be obtained. If we assume that the coin is "biased", says, the probability of obtaining head is $3/4$, then we predict that 75 heads will be obtained. Some *predictions* are made, but the coin is not really tossed, therefore the *probability*. In summary, probability theory enables us to find the consequences of a given ideal world (some believes), while statistical enables us to *quantify uncertainties* in real world (no single correct answer).

CHAPTER 1

Review of differentiation

1.1. Limits and continuity

We also recall some fact in calculus, see e.g. my lecture note [Kow24] and the references therein for more details.

DEFINITION 1.1.1. A subset $\Omega \subset \mathbb{R}^N$ is said to be *open* if for each $x \in \Omega$ there exists $\varepsilon = \varepsilon(x) > 0$ such that $B_\varepsilon(x) \subset \Omega$. Here and after, the open ball $B_R(x)$ is defined by

$$B_R(x) := \left\{ \boldsymbol{y} = (y_1, \cdots, y_N) \in \mathbb{R}^N : |\boldsymbol{x} - \boldsymbol{y}| < \varepsilon \right\},$$

where $|\boldsymbol{z}| = \sqrt{z_1^2 + \cdots + z_N^2}$ for all $\boldsymbol{z} = (z_1, \cdots, z_N) \in \mathbb{R}^N$.

DEFINITION 1.1.2. Let Ω be an open set in $\mathbb{R}^N$ with $\boldsymbol{x}_0 \in \Omega$ and we consider a function $f : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$.

(1) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = L$ exists if the following holds: Given any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } |f(\boldsymbol{x}) - L| < \varepsilon.$$

In this case, we also say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x})$ exists in $\mathbb{R}$.

(2) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = +\infty$ exists if the following holds: Given any $M > 0$, there exists $\delta = \delta(M) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } f(\boldsymbol{x}) > M.$$

(3) We say that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x}) = -\infty$ exists if the following holds: Given any $M > 0$, there exists $\delta = \delta(M) > 0$ such that

$$0 < |\boldsymbol{x} - \boldsymbol{x}_0| < \delta \text{ implies } f(\boldsymbol{x}) < -M.$$

We also unify the above notions by saying that the *limit* $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} f(\boldsymbol{x})$ exists in $[-\infty, +\infty]$.

The following are some basic properties of limits:

LEMMA 1.1.3. *Let Ω be an open set in $\mathbb{R}^N$ with $\boldsymbol{x}_0 \in \Omega$ and we consider functions $g_1 : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$ and $g_2 : \Omega \setminus \{\boldsymbol{x}_0\} \to \mathbb{R}$. If both limits $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} g_1(\boldsymbol{x})$ and $\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} g_2(\boldsymbol{x})$ exist in $\mathbb{R}$, then the following holds true:*

(1) *for each* $c_1 \in \mathbb{R}$ *and* $c_2 \in \mathbb{R}$ *the limit* $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(c_1 g_1(\boldsymbol{x}) + c_2 g_2(\boldsymbol{x}))$ *exists in* $\mathbb{R}$ *and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(c_1 g_1(\boldsymbol{x}) + c_2 g_2(\boldsymbol{x})) = \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x}) + \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \quad \textit{(linearity)}.$$

(2) *if* $g_1(\boldsymbol{x}) \leq g_2(\boldsymbol{x})$ *for all* $\boldsymbol{x} \in B_\varepsilon(\boldsymbol{x}_0)$ *for some* $\varepsilon > 0$*, then*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x}) \leq \lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \quad \textit{(monotonicity)}.$$

(3) *the limit* $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(g_1(\boldsymbol{x})g_2(\boldsymbol{x}))$ *exists in* $\mathbb{R}$ *and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0}(g_1(\boldsymbol{x})g_2(\boldsymbol{x})) = \left(\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x})\right)\left(\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x})\right).$$

(4) *if we additionally assume that* $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x}) \neq 0$*, then the limit* $\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})}$ *exists in* $\mathbb{R}$ *and satisfies*

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} \frac{g_1(\boldsymbol{x})}{g_2(\boldsymbol{x})} = \frac{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_1(\boldsymbol{x})}{\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} g_2(\boldsymbol{x})}.$$

We now ready to introduce the following notion:

DEFINITION 1.1.4. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}_0} f(\boldsymbol{x}) = f(\boldsymbol{x}_0) \in \mathbb{R} \quad \text{for some } \boldsymbol{x}_0 \in \Omega,$$

then we say that f is *continuous at* $\boldsymbol{x}_0$.

(1) If there exists an open set $\boldsymbol{x}_0 \in U \subset \Omega$ such that f is continuous at all point $\boldsymbol{x} \in U$, then we say that f is *continuous near* $\boldsymbol{x}_0$.

(2) If f is continuous at all points in Ω, then we say that f is *continuous on* Ω.

1.2. First order derivatives

For the case when $N = 1$, the differentiation of the function $f : \Omega \subset \mathbb{R}^1 \to \mathbb{R}$ can be simply define by the limit

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}. \tag{1.2.1}$$

However, the above definition cannot be directly extended to higher dimensional case by simply replace x and h by vectors $\boldsymbol{x}$ and $\boldsymbol{h}$, since the division of vectors are not well-defined. Let's us observe the above equation holds if and only if

$$\begin{aligned} 0 &= \lim_{h\to 0} \left| \frac{f(x+h) - f(x)}{h} - f'(x) \right| \\ &= \lim_{h\to 0} \left| \frac{f(x+h) - f(x) - f'(x)h}{h} \right| = \lim_{h\to 0} \frac{|f(x+h) - f(x) - f'(x)h|}{|h|}. \end{aligned}$$

Now it is natural to consider the differentiation of the function $f : \Omega \subset \mathbb{R}^N \to \mathbb{R}$ as follows:

DEFINITION 1.2.1. Let Ω be an open set in $\mathbb{R}^N$ and let $f:\Omega\to\mathbb{R}$ be a function. We say that f is *differentiable* at $\boldsymbol{x}_0\in\Omega$ if there exists a vector $\boldsymbol{L}\in\mathbb{R}^N$ such that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}=0, \tag{1.2.2}$$

where $\boldsymbol{L}\cdot\boldsymbol{h}=L_1h_1+\cdots+L_Nh_N$. In this case, the total derivative $\boldsymbol{D}f(\boldsymbol{x}_0)$ of f at $\boldsymbol{x}_0$is defined by the vector $\boldsymbol{D}f(\boldsymbol{x}_0):=\boldsymbol{L}$.

(1) If there exists an open set $\boldsymbol{x}_0\in U\subset\Omega$ such that f is differentiable at all point $\boldsymbol{x}\in U$, then we say that f is *differentiable near* $\boldsymbol{x}_0$.
(2) If f is differentiable at all points in Ω, then we say that f is *differentiable on* Ω.

It is not so obvious that whether the vector $\boldsymbol{L}$ in (1.2.2) is unique or not. Suppose that (1.2.2) holds true for $\boldsymbol{L}=\boldsymbol{L}_1$ and $\boldsymbol{L}=\boldsymbol{L}_2$, then

$$\begin{aligned}|\boldsymbol{L}_1-\boldsymbol{L}_2| &= \frac{|(f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_1\cdot\boldsymbol{h})-(f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_2\cdot\boldsymbol{h})|}{|\boldsymbol{h}|}\\ &\le \frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_1\cdot\boldsymbol{h}|-|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}_2\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}\to 0 \quad \text{as } |\boldsymbol{h}|\to 0_+,\end{aligned}$$

which concludes that $\boldsymbol{L}_1=\boldsymbol{L}_2$.

EXERCISE 1.2.2. Show that each differentiable function is also continuous.

After exhibiting an abstract definition of differentiation of a function f, we are now asking how to compute its total derivative $\boldsymbol{D}f(\boldsymbol{x})$ at each point $\boldsymbol{x}$. Let $\boldsymbol{e}_i$ be the i^{th} column of the identity matrix, that is,

$$\begin{aligned}\boldsymbol{e}_1&=(1,0,0,\cdots,0),\\ \boldsymbol{e}_2&=(0,1,0,\cdots,0),\\ &\vdots\\ \boldsymbol{e}_N&=(0,\cdots,0,0,1).\end{aligned}$$

If f is differentiable at $\boldsymbol{x}_0\in\Omega$, then we may restrict the limit (1.2.2) on the straight line $\{h\boldsymbol{e}_i : h\in\mathbb{R}\}$ to see that

$$\begin{aligned}0=\lim_{\boldsymbol{h}\to\boldsymbol{0}}\frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}&=\lim_{\boldsymbol{h}=h\boldsymbol{e}_i\to\boldsymbol{0}}\frac{|f(\boldsymbol{x}+\boldsymbol{h})-f(\boldsymbol{x})-\boldsymbol{L}\cdot\boldsymbol{h}|}{|\boldsymbol{h}|}\\ =\lim_{h\to 0}\frac{|f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})-L_ih|}{|h|},&\end{aligned}$$

which implies

$$L_i=\lim_{h\to 0}\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h},$$

which is exactly identical to the definition of the differentiation of functions with only one variable (1.2.1). Now it is natural to consider the following notion.

DEFINITION 1.2.3. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. For each $i = 1, \cdots, N$, the i^{th} partial derivative of f at $\boldsymbol{x}_0 = (x_1, \cdots, x_N)$ is defined by

$$\frac{\partial}{\partial x_i} f(x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_N)\bigg|_{x=x_i} = \partial_i f(\boldsymbol{x}_0) := \lim_{h\to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{e}_i) - f(\boldsymbol{x}_0)}{h}.$$

If all partial derivatives at $\boldsymbol{x}_0$ exist, then we define the gradient $\nabla f(\boldsymbol{x}_0) := (\partial_1 f(\boldsymbol{x}_0), \cdots, \partial_N f(\boldsymbol{x}_0)) \in \mathbb{R}^N$.

We now put the above discussions in the following lemma.

THEOREM 1.2.4. *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If f is differentiable at $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{D} f(\boldsymbol{x}_0) = \nabla f(\boldsymbol{x}_0)$.*

REMARK 1.2.5. The existence of the gradient $\nabla f(\boldsymbol{x}_0)$ does not guarantee the differentiability of f at $\boldsymbol{x}_0$.

The following sufficient condition is often been used to check the differentiability of a function.

THEOREM 1.2.6 ([**Apo74**, Theorem 12.11]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a function. If the point $\boldsymbol{x}_0 \in \Omega$ satisfies the following two conditions:*

- *there exists $\varepsilon > 0$ such that all partial derivatives $\partial_1 f, \cdots, \partial_N f$exist on $B_\varepsilon(\boldsymbol{x}_0)$; and*
- *all partial derivatives $\partial_1 f, \cdots, \partial_N f$ are continuous at $\boldsymbol{x}_0$;*

then $f$is differentiable at $\boldsymbol{x}_0$.

The above theorem suggested the following definition.

DEFINITION 1.2.7. Let Ω be an open set in $\mathbb{R}^N$. We denote $C^1(\Omega)$ be the collection of differentiable functions $f : \Omega \to \mathbb{R}$ such that all partial derivatives $\partial_1 f, \cdots, \partial_N f : \Omega \to \mathbb{R}$ are continuous.

We often use the following consequence of Theorem 1.2.6 since it is much easy to remember:

COROLLARY 1.2.8. *Let Ω be an open set in $\mathbb{R}^N$. If $f \in C^1(\Omega)$, then $f : \Omega \to \mathbb{R}$ is differentiable and hence $\boldsymbol{D} f(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \Omega$.*

The above corollary allows us to solve multidimensional case by using technique in 1-dimensional case.

1.3. Differentiation rules

Here we only give some special cases which are often used in practical applications. One can refer to the monographs [**Apo74, Rud87**] for the results which are much more optimal.

LEMMA 1.3.1. *Let Ω be an open set in $\mathbb{R}^N$ and let $f_1, f_2 : \Omega \to \mathbb{R}$.*

(a) ***Linearity.*** *If both $f_1, f_2 \in C^1(\Omega)$, then for each $c_1, c_2 \in \mathbb{R}$, the function*

$$c_1 f_1 + c_2 f_2 : \Omega \to \mathbb{R}, \quad (c_1 f_1 + c_2 f_2)(\boldsymbol{x}) := c_1 f_1(\boldsymbol{x}) + c_2 f_2(\boldsymbol{x}) \textit{ for all } \boldsymbol{x} \in \Omega$$

is also in $C^1(\Omega)$, and satisfying

$$\nabla(c_1 f_1 + c_2 f_2)(\boldsymbol{x}) = c_1 \nabla f_1(\boldsymbol{x}) + c_2 \nabla f_2(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega.$$

(b) ***Product rule.*** *If both $f_1, f_2 \in C^1(\Omega)$, then the function*

$$f_1 f_2 : \Omega \to \mathbb{R}, \quad (f_1 f_2)(\boldsymbol{x}) := f_1(\boldsymbol{x}) f_2(\boldsymbol{x}) \textit{ for all } \boldsymbol{x} \in \Omega$$

is also in $C^1(\Omega)$, and satisfying

$$\nabla(f_1 f_2)(\boldsymbol{x}) = f_2(\boldsymbol{x}) \nabla f_1(\boldsymbol{x}) + f_1(\boldsymbol{x}) \nabla f_2(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega.$$

LEMMA 1.3.2 (chain rule [Apo74, Theorem 12.7]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f_1, \cdots, f_m : \Omega \to \mathbb{R}$ be functions which is differentiable at a point $\boldsymbol{x}_0 \in \Omega$. We denote the vector valued function*

$$\boldsymbol{f} : \Omega \to \mathbb{R}^m, \quad \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \cdots, f_m(\boldsymbol{x})) \textit{ for all } \boldsymbol{x} \in \Omega,$$

and its range is defined by $\boldsymbol{f}(\Omega) := \{\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m : \boldsymbol{x} \in \Omega\}$. Let U be an open set in $\mathbb{R}^m$ such that $U \supset \boldsymbol{f}(\Omega)$ and let $g : U \to \mathbb{R}$ be a function which is differentiable at $\boldsymbol{f}(\boldsymbol{x}_0)$. Then the composition of functions

$$g \circ \boldsymbol{f} : \Omega \to \mathbb{R}, \quad g \circ \boldsymbol{f}(\boldsymbol{x}) := g(\boldsymbol{f}(\boldsymbol{x})) \quad \textit{for all } \boldsymbol{x} \in \Omega$$

is also differentiable at $\boldsymbol{x}_0 \in \Omega$ and its partial derivatives are given by

$$\begin{aligned} \frac{\partial}{\partial x_i}(g \circ \boldsymbol{f})(\boldsymbol{x}) &= \nabla_{\boldsymbol{y}} g(\boldsymbol{y})|_{\boldsymbol{y}=f(\boldsymbol{x})} \cdot \frac{\partial}{\partial x_i} \boldsymbol{f}(\boldsymbol{x}) \\ &= \sum_{j=1}^m \frac{\partial}{\partial y_j} g(\boldsymbol{y}) \bigg|_{\boldsymbol{y}=\boldsymbol{f}(\boldsymbol{x})} \frac{\partial}{\partial x_i} f_j(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega. \end{aligned} \tag{1.3.1}$$

In practice, we often use the following corollary, which says that the composition of C^1 functions is also in C^1:

COROLLARY 1.3.3. *Let Ω be an open set in $\mathbb{R}^N$ and let $\boldsymbol{f} \in (C^1(\Omega))^N$. Let U be an open set in $\mathbb{R}^m$ such that $U \supset \boldsymbol{f}(\Omega)$ and let $g \in C^1(U)$. Then the composition of functions $g \circ \boldsymbol{f} \in C^1(\Omega)$ and satisfies (1.3.1).*

1.4. Second order derivatives

Let Ω be an open set in $\mathbb{R}^N$. The first order derivative of $f : \Omega \to \mathbb{R}$ at each $\boldsymbol{x} \in \Omega$ is given by the vector

$$\nabla f(\boldsymbol{x}) = (\partial_1 f(\boldsymbol{x}), \cdots, \partial_N f(\boldsymbol{x})) \in \mathbb{R}^N.$$

We now further assume that $\partial_i f : \Omega \to \mathbb{R}$ is differentiable for all $i = 1, \cdots, N$, then the first order derivative of each $\partial_i f$ at each point $\boldsymbol{x} \in \Omega$ is given by the vector

$$\nabla \partial_i f(\boldsymbol{x}) = (\partial_1 \partial_i f(\boldsymbol{x}), \cdots, \partial_N \partial_i f(\boldsymbol{x})) \in \mathbb{R}^N.$$

This suggests that the second order derivative of $f : \Omega \to \mathbb{R}$ at each $\boldsymbol{x} \in \Omega$ should be the following 2-tensor (i.e. matrix)

$$\nabla^{\otimes 2} f(\boldsymbol{x}) \equiv \nabla \otimes \nabla f(\boldsymbol{x}) := \begin{pmatrix} \partial_1\partial_1 f(\boldsymbol{x}) & \partial_1\partial_2 f(\boldsymbol{x}) & \cdots & \partial_1\partial_N f(\boldsymbol{x}) \\ \partial_2\partial_1 f(\boldsymbol{x}) & \partial_2\partial_2 f(\boldsymbol{x}) & \cdots & \partial_2\partial_N f(\boldsymbol{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_N\partial_1 f(\boldsymbol{x}) & \partial_N\partial_2 f(\boldsymbol{x}) & \cdots & \partial_N\partial_N f(\boldsymbol{x}) \end{pmatrix},$$

that is,

$$\left(\nabla^{\otimes 2} f(\boldsymbol{x})\right)_{ij} := \partial_i \partial_j f(\boldsymbol{x}) \quad \text{for all } i, j = 1, \cdots, N.$$

We call $\nabla^{\otimes 2} f(\boldsymbol{x})$ the *Hessian matrix*. The notation $\otimes$ comes from the juxtaposition $\boldsymbol{u} \otimes \boldsymbol{v}$ of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is defined by $\boldsymbol{u} \otimes \boldsymbol{v} := \boldsymbol{u}\boldsymbol{v}^\intercal$, that is,

$$(\boldsymbol{u} \otimes \boldsymbol{v})_{ij} := u_i v_j.$$

REMARK 1.4.1. Similarly, the third derivatives of f should be the 3-tensor $\nabla^{\otimes 3} f(\boldsymbol{x})$ given by

$$\left(\nabla^{\otimes 3} f(\boldsymbol{x})\right)_{ijk} \equiv (\nabla \otimes \nabla \otimes \nabla f(\boldsymbol{x}))_{ijk} := \partial_i \partial_j \partial_k f(\boldsymbol{x}) \quad \text{for all } i, j, k = 1, \cdots, N.$$

Inductively, the m^{th} derivatives of f should be the m-tensor $\nabla^{\otimes m} f(\boldsymbol{x})$ given by

$$\left(\nabla^{\otimes m} f(\boldsymbol{x})\right)_{i_1 i_2 \cdots i_m} := \partial_{i_1} \partial_{i_2} \cdots \partial_{i_m} f(\boldsymbol{x}) \quad \text{for all } i_1, i_2, \cdots, i_m = 1, \cdots, N.$$

In practice, it is not convenient to work with nonsymmetric Hessian matrix $\nabla^{\otimes 2} f(\boldsymbol{x})$. Luckily this is not the usual case:

THEOREM 1.4.2 ([Apo74, Theorem 12.13]). *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a differentiable function. If there exists $\boldsymbol{x}_0 \in \Omega$ and $i, j \in \{1, \cdots, N\}$ such that both $\partial_i \partial_j f : \Omega \to \mathbb{R}$ and $\partial_j \partial_i f : \Omega \to \mathbb{R}$ exist and continuous at $\boldsymbol{x}_0$, then*

$$\partial_i \partial_j f(\boldsymbol{x}_0) = \partial_j \partial_i f(\boldsymbol{x}_0).$$

This theorem suggested us to consider the following space:

DEFINITION 1.4.3. Let Ω be an open set in $\mathbb{R}^N$. We denote $C^2(\Omega)$ be the collection of functions $f : \Omega \to \mathbb{R}$ such that all partial derivatives

$$\partial_i f : \Omega \to \mathbb{R}, \quad \partial_i \partial_j f : \Omega \to \mathbb{R} \text{ for all } i, j = 1, \cdots, N$$

exist and continuous.

In practice, we often use the following corollary of Theorem 1.4.2.

COROLLARY 1.4.4. *Let Ω be an open set in $\mathbb{R}^N$. If $f \in C^2(\Omega)$, then for each $\boldsymbol{x} \in \Omega$ the Hessian matrix $\nabla^{\otimes 2} f(\boldsymbol{x})$ is symmetric.*

1.5. Extreme values

We first begin with some definitions.

DEFINITION 1.5.1. Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a differentiable function. If $\nabla f(\boldsymbol{x}_0)$ is a zero vector for some $\boldsymbol{x}_0 \in \Omega$, then we refer such point $\boldsymbol{x}_0$ a *critical point* or *stationary point*.

DEFINITION 1.5.2. Let S be a set (not necessarily open) in $\mathbb{R}^N$, let $f : S \to \mathbb{R}$ be a function and let $\boldsymbol{x}_0 \in S$.

(1) If there exists $\varepsilon > 0$ such that $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S \cap B_\varepsilon(\boldsymbol{x}_0)$, then we call $\boldsymbol{x}_0$ a *local minimizer* of $f : S \to \mathbb{R}$.
(2) If there exists $\varepsilon > 0$ such that $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S \cap B_\varepsilon(\boldsymbol{x}_0)$, then we call $\boldsymbol{x}_0$ a *local maximizer* of $f : S \to \mathbb{R}$.
(3) If $f(\boldsymbol{x}_0) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S$, then we call $\boldsymbol{x}_0$ a *global minimizer* of $f : S \to \mathbb{R}$.
(4) If $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S$, then we call $\boldsymbol{x}_0$ a *global maximizer* of $f : S \to \mathbb{R}$.

It is easy to prove the followings:

LEMMA 1.5.3. *Let Ω be an open set in $\mathbb{R}^N$ and let $f : \Omega \to \mathbb{R}$ be a differentiable function. If f has a local maximum or local minimum at $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a critical point.*

It is also possible to define the notion for "positive" symmetric matrices:

DEFINITION 1.5.4. Let $A \in \mathbb{R}^{N \times N}$ be a symmetric matrix.

(1) We say that A is *positive definite*, denoted as $A \succ 0$, when $\boldsymbol{\xi}^\intercal A \boldsymbol{\xi} > 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}$.
(2) We say that A is *negative definite*, denoted as $A \prec 0$, when $\boldsymbol{\xi}^\intercal A \boldsymbol{\xi} < 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\}$.

REMARK 1.5.5. The above definition also make sense when $N = 1$ by simply identify $\mathbb{R} \cong \mathbb{R}^{1 \times 1}$

It is remarkable that the second derivative test also works for higher dimensional case as well:

THEOREM 1.5.6. *Let Ω be an open set in $\mathbb{R}^N$ and let $f \in C^2(\Omega)$.*

(a) *If $\nabla f(\boldsymbol{x}_0) = \boldsymbol{0}$ and $\nabla^{\otimes 2} f(\boldsymbol{x}_0) \succ 0$ hold for some $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a local minimizer of $f : \Omega \to \mathbb{R}$.*
(b) *If $\nabla f(\boldsymbol{x}_0) = \boldsymbol{0}$ and $\nabla^{\otimes 2} f(\boldsymbol{x}_0) \prec 0$ hold for some $\boldsymbol{x}_0 \in \Omega$, then $\boldsymbol{x}_0$ is a local maximizer of $f : \Omega \to \mathbb{R}$.*

CHAPTER 2

Review of probability

2.1. Definition of probability and its properties

In probability, an *experiment* refers to any action or activity whose outcome is subject to uncertainty.

DEFINITION 2.1.1. The *sample space* Ω of an experiment is the set which including all possible outcomes of that experiment, and a *outcome* x is a point in Ω, which we denoted as $x \in \Omega$. An *event* A is a subset of Ω, which we denoted as $A \subset \Omega$.

REMARK 2.1.2. Here Ω is not necessary exactly the set of all possible outcomes. Despite Ω may larger than the set of all possible outcomes, we still refer all elements in Ω an outcome. For example, if all possible outcome of an experiment is $\{1,2,4\}$, one may choose $\Omega = \{1,2,3,4\}$ and we can intuitively set the probability of the outcome "3" as 0.

REMARK 2.1.3. One should not abuse the notation "$\in$" and "$\subset$", for example, the Russell's paradox

$$\{X : X \text{ is a set and } X \notin X\}.$$

In practical, we usually refer the set consists of other sets as a *collection*. For example, let $\mathscr{P}$ be the collection of all subsets in $\{a,b\}$ means that

$$\mathscr{P} = \{\emptyset, \{a\}, \{b\}, \{a,b\}\}.$$

We intuitively view $\emptyset, \{a\}, \{b\}, \{a,b\}$ as "level-1" objects, and view $\mathscr{P}$ as "level-2" object. The elements in sets also can be viewed as "level-0" objects. We usually refer the set consists of collections (i.e. "level-2" objects) as a superset, which is natural to be labeled as "level-3" object. We distinguish between "$\in$" and "$\subset$ as follows:

- We write $x \in X$ for "level-0" object x (point) and for "level-1" object X (set); we write $X \in \mathscr{P}$ for "level-1" object X (set) and for "level-2" object $\mathscr{P}$ (collection), and so on.
- We write $X \subset Y$ for two "level-1" objects X and Y (sets); we write $\mathscr{P} \subset \mathscr{Q}$ for two "level-2" objects $\mathscr{P}$ and $\mathscr{Q}$ (collections).

Here we reserve the notation $\emptyset$ for empty set ("level-1" object). According to this notation system, we remind the readers that $\{\emptyset\}$ is a *nonempty* collection, which consists one element called $\emptyset$.

The *complement* of an event A is defined by $A^{\complement} := \Omega \setminus A$, which is the set of all outcomes in Ω that are not contained in A. The *intersection* of two events A and B is defined by $A \cap B :=$

$\{x \in \Omega : x \in A \text{ and } x \in B\}$, which is the event consisting of all outcomes that are in both A and B. The *union* of two events A and B is defined by $A \cup B := \{x \in \Omega : x \in A \text{ or } x \in B\}$, which is the event consists of all outcomes that are either in A or in B or in both events. We say that the events A and B are disjoint if $A \cap B = \emptyset$, where $\emptyset$ denotes the event consisting of no outcomes whatsoever (i.e. empty event).

A collection $\{A_i\}_{i\in I}$ is said to be *finite* if there exists a bijection between I and $\{1,\cdots,N\}$ for some $N \in \mathbb{N}$. A collection $\{A_i\}_{i\in I}$ is said to be *infinite countable* if there exists a bijection between I and $\mathbb{N}$. A collection $\{A_i\}_{i\in I}$ is said to be *countable* if either I is finite or infinite countable. Let $\{A_i\}_{i\in I}$ be a countable collection of sets, and similarly we denote

$$\bigcup_{i\in I} A := \{x \in \Omega : x \in A_i \text{ for some } i \in I\},$$
$$\bigcap_{i\in I} A := \{x \in \Omega : x \in A_i \text{ for all } i \in I\}.$$

We say that $\{A_i\}_{i\in I}$ is a *collection of disjoint events* if $A_i \cap A_j = \emptyset$ for all $i \neq j$. Given an experiment and its sample space Ω, we denote its power set 2^Ω which consists of all events $A \subset \Omega$. We want to assign a number $\mathbb{P}(A)$ for each event $A \subset \Omega$ according to the following three axioms:

(1) $\mathbb{P}(A) \geq 0$ for any event $A \subset \Omega$;
(2) $\mathbb{P}(\Omega) = 1$;
(3) If $\{A_i\}_{i=1}^\infty$ is countable collection of disjoint events, then $\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$.

However, in many cases, one may fail to assign $\mathbb{P}(A)$ for arbitrary event $A \subset \Omega$. For example, if we consider $\Omega = [0,1]$, it is natural to assign

$$\mathbb{P}(A) := \text{the Lebesgue measure of } A \text{ (i.e. the area of } A).$$

However there exists a set $V \subset [0,1]$, called the Vitali set, which is not measurable (i.e. the Lebesgue measure of V is not well-defined). Therefore we need to restrict the probability on some suitable collection $\mathscr{F} \subset 2^\Omega$, which is *$\sigma$-field* (also known as *$\sigma$-algebra*) on Ω, i.e. a nonempty collection of subsets of Ω that satisfy

(1) $A^\complement \in \mathscr{F}$ if and only if $A \in \mathscr{F}$;
(2) if $\{A_i\}_{i\in I} \subset \mathscr{F}$ is countable, then $\bigcup_{i\in I} A_i \in \mathscr{F}$.

DEFINITION 2.1.4 (Axioms for probability). Given an experiment and its sample space Ω, and let $\mathscr{F}$ be a σ-field on Ω. We now assign a number $\mathbb{P}(A)$ for each event $A \in \mathscr{F}$ according to the following three axioms:

(1) $\mathbb{P}(A) \geq 0$ for any event $A \in \mathscr{F}$;
(2) $\mathbb{P}(\Omega) = 1$;
(3) If $\{A_i\}_{i=1}^\infty \subset \mathscr{F}$ is countable collection of disjoint events, then $\mathbb{P}\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} \mathbb{P}(A_i)$.

We called the triple $(\Omega,\mathscr{F},\mathbb{P})$ a *probability space*, $\mathscr{F}$ the *set of events* and the mapping $\mathbb{P}:\mathscr{F}\to[0,1]$ the *probability*.

If there is no ambiguity, we shall not explicitly mention the sample space Ω and the set of events $\mathscr{F}$ if there is no any ambiguity.

LEMMA 2.1.5. *The probability $\mathbb{P}$ defined by Definition 2.1.4 satisfies the following properties:*

(1) $\mathbb{P}(\emptyset)=0$.
(2) $\mathbb{P}(A)=1-\mathbb{P}(A^{\complement})$ *for any event A.*
(3) $\mathbb{P}(A)\leq 1$ *for any event A.*
(4) $\mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)$ *for any events A and B.*

EXERCISE 2.1.6 (Inclusion-exclusion formula). Given any $N\in\mathbb{N}$ and events $A_1,\cdots,A_N$, show that

$$\mathbb{P}\left(\bigcup_{i=1}^{N}A_i\right)=\sum_{k=1}^{N}(-1)^{k+1}\left(\sum_{1\leq i_1<\cdots<i_k\leq N}\mathbb{P}(A_{i_1}\cap\cdots\cap A_{i_k})\right).$$

2.2. Conditional probability and independence

In many cases, we are interested to the probability of events provided some apriori information is known. Given that B has occurred, the relevant sample space is no longer Ω but consists of just outcomes in B, and A has occurred if and only if one of the outcomes in the intersection $A\cap B$. So the conditional probability of A given B should, logically, be the ratio of the likelihood of these two events. This leads the following definition.

DEFINITION 2.2.1. For any two events A and B with $\mathbb{P}(B)>0$, the *conditional probability* $\mathbb{P}(A|B)$ *of A given that B has occurred* is defined by

$$\mathbb{P}(A|B):=\frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}.$$

REMARK 2.2.2. Here we shall point out that Definition (2.2.1) is actually not good enough in practical applications. For example, we consider the collection $\mathscr{F}$ of measurable sets in $\Omega=[0,1]\times[0,1]$ with probability

$$\mathbb{P}(A):=\text{Lebesgue measure of } A \text{ (i.e. the area of } A\text{)}.$$

We shall believe that the conditional probability of $A=[0,1/2]\times\{0\}$ given that $B=[0,1]\times\{0\}$ has occurred, should be $1/2$, but however it is not possible to formulate this phenomena using Definition (2.2.1) because $\mathbb{P}(B)=0$. One may refer to [**Dur19**] for a more general framework.

Frequently the nature of an experiment suggests that two events A and B should be assumed independent, in other words, we may expect the outcome of second experiment should not affected by the first experiment. This suggests us to say that A and B is independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$ provided $\mathbb{P}(B) > 0$. In fact, one has the followings:

$$\mathbb{P}(A|B) = \mathbb{P}(A) \text{ if and only if } \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

This fact suggests us to consider the following definition:

DEFINITION 2.2.3. We say that two events A and B are *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$, and are *dependent* otherwise.

The notion of independence of two events can be extended to collections of more than two events.

DEFINITION 2.2.4. The events $A_1, \cdots, A_m$ are *independent* if whenever $I \subset \{1, \cdots, m\}$ we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i). \tag{2.2.1}$$

The infinitely countably many events $A_1, A_2, \cdots$ are *independent* if whenever *finite* set $I \subset \mathbb{N}$ we have (2.2.1).

2.3. Random variable

Given an (abstract) probability space $(\Omega, \mathscr{F}, \mathbb{P})$, given a function $X : \Omega \to \mathbb{R}$ (or a "mechanism"), we want to study the event of the form

$$\{X \in B\} := \{x \in \Omega : X(x) \in B\}$$

which is called the *preimage* of B with respect to X. However, as we mentioned above (before Definition 2.1.4), not all subsets in $\mathbb{R}$ is measurable, and the preimage of non-measurable sets are not "meaningful". Let $\mathscr{B}$ be the collection of Borel sets in $\mathbb{R}$, that is,

$$\mathscr{B} := \left\{ B \subset \mathbb{R} : \begin{array}{l} B \text{ can be constructed by countable union, countable intersection} \\ \text{or complenents of open sets in } \mathbb{R} \end{array} \right\}.$$

All elements in $\mathscr{B}$ are Lebesgue measurable, that is, the Lebesgue measure (or simply "volume") of each elements in $B \in \mathscr{B}$ is well-defined. Despite there exists Lebesgue measurable set which is not in $\mathscr{B}$, we usually consider Borel sets $\mathscr{B}$ in practical applications. Let $X : \Omega \to \mathbb{R}$ be a given function, and we are now interesting in the sets of the form

$$\{X \in B\} := \{x \in \Omega : X(x) \in B\} \quad \text{for } B \in \mathscr{B},$$

Now it is natural to quantify abstract probability space $(\Omega, \mathscr{F}, \mathbb{P})$ via the following definition.

DEFINITION 2.3.1. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a given function. If $\{X \in B\} \in \mathscr{F}$ (that is, the probability of all sets $\{X \in B\}$ are well-defined), then we refer such function $X : \Omega \to \mathbb{R}$ a *random variable*.

It is also natural to mention the convergence of a sequence of random variables.

DEFINITION 2.3.2. Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of random variables, and we consider the event $\Omega_0 := \{x \in \Omega : \lim_{k\to\infty} X_k(x) \text{ exists in } [-\infty,\infty]\}$. If $\mathbb{P}(\Omega_0) = 1$, then we say that $\{X_k\}_{k\in\mathbb{N}}$ converges *almost surely* (later we use the abbreviation "a.s.").

In view of Definition 2.2.4, now it is natural to introduce the following definition as well.

DEFINITION 2.3.3. Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space. We say that the random variables $X_1,\cdots,X_N : \Omega \to \mathbb{R}$ are *independent* if

$$\mathbb{P}\left(\bigcap_{i=1}^{N}\{X_i \in B_i\}\right) = \prod_{i=1}^{N}\mathbb{P}(X_i \in B_i) \quad \text{for all } B_1,\cdots,B_N \in \mathscr{B}.$$

We say that $X_1,\cdots,X_N : \Omega \to \mathbb{R}$ are *pairwise independent* if for each $i \neq j$ the random variables X_i and X_j are independent.

REMARK 2.3.4. If $X_1,\cdots,X_N$ are independent (resp. pairwise independent), then $f_1(X_1),\cdots,f_N(X_N)$ are independent (resp. pairwise independent).

DEFINITION 2.3.5. We say that the infinitely countably many random variables $X_1,X_2,\cdots :$ $\Omega \to \mathbb{R}$ are *independent* if whenever *finite* set $I \subset \mathbb{N}$ we have

$$\mathbb{P}\left(\bigcap_{i\in I}\{X_i \in B_i\}\right) = \prod_{i\in I}\mathbb{P}(X_i \in B_i) \quad \text{for all } B_i \in \mathscr{B} \text{ with } i \in I.$$

We say that $X_1,X_2,\cdots : \Omega \to \mathbb{R}$ are *pairwise independent* if for each $i \neq j$ the random variables X_i and X_j are independent.

We first exhibit a trivial, but useful, type of random variable.

EXAMPLE 2.3.6. Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space. Fixing $A \in \mathscr{F}$ and we consider a random variable $X : \Omega \to \{0,1\}$ defined by

$$\mathbb{1}_A(x) := \begin{cases} 1 & ,x \in A, \\ 0 & ,x \notin A. \end{cases}$$

DEFINITION 2.3.7. The random variable $\mathbb{1}_A$ is called the *indicator function* or *test function* on A. Analysts call this object the *characteristic function* on A.

We now restrict ourselves for the case when $n = 1$. For each $x \in \mathbb{R}$, we define $x_+ := \max\{x,0\}$ and $x_- := -\min\{x,0\}$. One sees that $x = x_+ - x_-$ and $|x| = x_+ + x_-$. We first introduce the following definition.

DEFINITION 2.3.8. Given a random variable $X : (\Omega,\mathscr{F},\mathbb{P}) \to [0,\infty)$, we define its expectation by

$$\mathbb{E}X := \int_\Omega X \,\mathrm{d}\mathbb{P},$$

which is always exist in $[0,\infty]$. For general random variable $X : (\Omega,\mathscr{F},\mathbb{P}) \to [0,\infty]$, if either $\mathbb{E}X_+ < +\infty$ or $\mathbb{E}X_- < +\infty$, then we say that the *expectation/mean* $\mathbb{E}X$ of X exists with values

$$\mu \equiv \mathbb{E}X := \mathbb{E}X_+ - \mathbb{E}X_-.$$

If $\mathbb{E}X^2 < \infty$, then the *variance* of X is defined to be

$$\mathrm{var}\,(X) := \mathbb{E}(X-\mu)^2 = \mathbb{E}X^2 - \mu^2.$$

The following are some basic properties:

LEMMA 2.3.9. *If either one of the followings hold:*

- $X \geq 0$ *and* $Y \geq 0$*; or*
- $\mathbb{E}|X| < \infty$ *and* $\mathbb{E}|Y| < \infty$*;*

then

(1) $\mathbb{E}(aX+bY+c) = a\mathbb{E}X + b\mathbb{E}Y + c$ *for all* $a,b,c \in \mathbb{R}$*;*
(2) *If* $X \geq Y$*, then* $\mathbb{E}X \geq \mathbb{E}Y$*.*

REMARK 2.3.10. If $\mathbb{E}X^2 < \infty$, then for each $a,b \in \mathbb{R}$ one can compute that

$$\begin{aligned}\mathrm{var}\,(aX+b) &= \mathbb{E}(aX+b-\mathbb{E}(aX+b))^2 \\ &= \mathbb{E}(aX - a\mathbb{E}(X))^2 = a^2\mathbb{E}(X-\mathbb{E}(X))^2 = a^2\mathrm{var}\,(X).\end{aligned}$$

The expectation operator $\mathbb{E}$ immediate suggests the following definition:

DEFINITION 2.3.11. We say that a sequence of random variables $\{X_k\}_{k\in\mathbb{N}}$ converges to a random variable X in *mean square/in quadratic mean* if

$$\lim_{k\to\infty} \mathbb{E}\left((X_k - X)^2\right) = 0.$$

By using Chebyshev's inequality [**Dur19**, (1.6.1)], which is a special case of Markov inequality [**Dur19**, Theorem 1.6.4], one sees that

$$\mathbb{P}\left(|X_k - X| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\mathbb{E}\left((X_k - X)^2\right).$$

This strongly suggests the following definition:

DEFINITION 2.3.12. We say that a sequence of random variables $\{X_k\}_{k\in\mathbb{N}}$ converges to a random variable X *in probability* if the following holds: Given any $\varepsilon > 0$, one has

$$\lim_{k\to\infty} \mathbb{P}\left(|X_k - X| > \varepsilon\right) = 0.$$

By using Fatou's lemma [**Dur19**, Theorem 1.6.5], one sees the following important theorem.

THEOREM 2.3.13. *If* $X_k \to X$ *a.e.* *(Definition 2.3.2), then* $X_k \to X$ *in probability (Definition 2.3.12).*

DEFINITION 2.3.14. The *cumulative distribution function* (later we use the abbreviation "c.d.f.") $F : \mathbb{R} \to [0,1]$ of a random variable X (not necessarily discrete or continuous) is defined for every number x by

$$F(x) := \mathbb{P}\left((X^{-1}((-\infty,x])\right) \equiv \mathbb{P}\left(\{y \in \Omega : X(y) \leq x\}\right),$$

and we usually simply denote as $F(x) = \mathbb{P}(X \leq x)$. In other words, $F(x)$ is the probability that the observed value of X will be at most X.

EXAMPLE 2.3.15. One can compute that $\mathbb{E}(\mathbb{1}_A \circ X) = \int_\Omega \mathbb{1}_A \circ X \, \mathrm{d}\mathbb{P} = \mathbb{P}(X \in A)$ for any $A \in \mathscr{B}$. By choosing $A = (-\infty, x]$ for any $x \in \mathbb{R}$, one also sees that

$$\mathbb{E}(\mathbb{1}_{(-\infty,x]} \circ X) = \mathbb{P}(X \leq x) = F(x).$$

THEOREM 2.3.16 ([**Dur19**, Theorem 1.2.1]). *Each c.d.f. $F : \mathbb{R} \to [0,1]$ described in Definition 2.3.14 has the following properties:*

(1) *F is nondecreasing;*
(2) *$\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$;*
(3) *F is right continuous, i.e. $\lim_{x\to x_0+} F(x) = F(x_0)$ for each $x_0 \in \mathbb{R}$;*
(4) *If we denote $F(x_0-) := \lim_{x\to x_0-} F(x)$, then $F(x_0-) = \mathbb{P}(X < x)$;*
(5) *$\mathbb{P}(X = x) = F(x) - F(x-)$ for all $x \in \mathbb{R}$.*

The next result shows that we have found more than enough properties to characterize distribution functions.

THEOREM 2.3.17 ([**Dur19**, Theorem 1.2.2]). *If $F : \mathbb{R} \to [0,1]$ satisfies properties (1)–(3) in Theorem 2.3.16, then it is the distribution function of some random variable.*

The following lemma shows that the moments of nonnegative random variables can be expressed in terms of its distribution functions.

LEMMA 2.3.18 ([**Dur19**, Lemma 2.2.13]). *If $X \geq 0$ and $p > 0$ (not necessarily an integer), then*

$$\mathbb{E}Y^p = \int_0^\infty p y^{p-1} \mathbb{P}(Y > y) \, \mathrm{d}y.$$

REMARK 2.3.19. In the case when $F(y) := \mathbb{P}(Y \leq y)$ is differentiable, by using integration by parts one sees that

$$\begin{aligned} \mathbb{E}Y^p &= \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}y}(y^p)(1 - \mathbb{P}(Y \leq y)) \, \mathrm{d}y = -\int_0^\infty y^p \frac{\mathrm{d}}{\mathrm{d}y}(1 - \mathbb{P}(Y \leq y)) \, \mathrm{d}y \\ &= \int_0^\infty y^p \frac{\mathrm{d}}{\mathrm{d}y}(\mathbb{P}(Y \leq y)) \, \mathrm{d}y. \end{aligned}$$

Despite the random variable mentioned in Theorem 2.3.17 may differ (depending on the probability space chosen), Theorem 2.3.16 and Theorem 2.3.17 strongly suggests us to study the probability distributions of random variables. This leads the following definition:

DEFINITION 2.3.20. Let $X : (\Omega_1, \mathscr{F}_1, \mathbb{P}_1) \to \mathbb{R}$ and $Y : (\Omega_2, \mathscr{F}_2, \mathbb{P}_2) \to \mathbb{R}$ be random variables. We say that X and Y *have the same distribution* if

$$\mathbb{P}_1(X \leq x) = \mathbb{P}_2(Y \leq x) \quad \text{for all } x \in \mathbb{R}.$$

In this case, we denote $X \overset{\mathrm{d}}{=} Y$.

In other words, this means that X and Y are basically “identical” with just different “representations/labels/notations”. This gives a reason which we can simply omit the notation $(\Omega, \mathscr{F}, \mathbb{P})$ if there is no confusion. In view of Definition 2.3.2, we need the following notion:

DEFINITION 2.3.21. A sequence of random variables $\{X_k\}_{k\in\mathbb{N}}$ is said to be *converge in distribution* to a limit X_∞ if their distribution functions $F_k(x) = \mathbb{P}(X_k \leq x)$ converges weakly to a limit F_∞, that is, $F_k(x) \to F_\infty(x)$ for all x that are continuity points[1] of F_∞.

In practical application, it is quite often to see the following definitions.

DEFINITION 2.3.22. Given random variables $X_1, \cdots, X_N$. If $X_1, \cdots, X_N$ are independent (Definition 2.3.3) and $X_i \overset{\mathrm{d}}{=} X_j$ for all $i, j \in \{1, \cdots, N\}$, then we say that $X_1, \cdots, X_N$ are *independent and identically distributed random variables* (later we use the abbreviation “i.i.d.”).

DEFINITION 2.3.23. Given infinitely countably many random variables $X_1, X_2, \cdots$. If $X_1, X_2, \cdots$ are independent (Definition 2.3.3) and $X_i \overset{\mathrm{d}}{=} X_j$ for all $i, j \in \mathbb{N}$, then we say that $X_1, X_2, \cdots$ are *independent and identically distributed random variables* (later we use the abbreviation “i.i.d.”).

The following theorem was proved by Etemadi in 1981 [**Ete81**]:

THEOREM 2.3.24 (Strong law of large numbers [**Dur19**, Theorem 2.4.1]). *Let $X_1, X_2, \cdots$ be pairwise independent identically distributed random variables with $\mathbb{E}|X_i| < \infty$. Let $\mathbb{E}X_i = \mu$ and $\overline{X}_N = \frac{X_1+\cdots+X_N}{N}$. Then $\overline{X}_N \to \mu$ a.s. (Definition 2.3.2).*

It is also possible to state the law of large numbers for random variables without the existence of expectations :

THEOREM 2.3.25 (Weak law of large numbers [**Dur19**, Theorem 2.2.12]). *Let $X_1, X_2, \cdots$ be i.i.d. random variables with*

$$x\mathbb{P}(|X_i| > x) \to 0 \quad \text{as } x \to \infty.$$

Let $\overline{X}_N := \frac{X_1+\cdots+X_N}{N}$ and let $\mu_N := \mathbb{E}(X_1 \mathbb{1}_{(|X_1| \leq N)})$. Then $\overline{X}_N - \mu_N \to 0$ in probability (Definition 2.3.12).

We now recall the following result:

[1] In fact, F_∞ is right continuous and the discontinuities of F_∞ are at most a countable set.

THEOREM 2.3.26 ([**Dur19**, Theorem 2.1.13]). *If $X_1, \cdots, X_N$ are independent and if either one of the following holds: (a) $X_i \geq 0$ for all i; or (b) $\mathbb{E}|X_i| < \infty$ for all i; then $\mathbb{E}\left(\prod_{i=1}^N X_i\right)$ exists and*

$$\mathbb{E}\left(\prod_{i=1}^N X_i\right) = \prod_{i=1}^N \mathbb{E}X_i.$$

The above theorem strongly suggests the following concept which is related to independence:

DEFINITION 2.3.27. Two random variables X and Y with $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$ that have $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$ are said to be *uncorrelated*.

REMARK 2.3.28. The expectation $\mathbb{E}(XY)$ is well-defined by the Cauchy-Schwartz inequality:

$$\mathbb{E}|XY| \leq (\mathbb{E}X^2)^{1/2}(\mathbb{E}Y^2)^{1/2} < \infty.$$

EXAMPLE 2.3.29. If X and Y are independent, then Theorem 2.3.26 says that X and Y are uncorrelated. However, the converse may not true, one can see [**Dur19**, Example 2.1.14] for a simple counterexample.

EXERCISE 2.3.30. Let $X_1, \cdots, X_N$ are *independent* random variables such that $\mathbb{E}X_i^2 < \infty$ for all $i = 1, \cdots, N$. Show that

$$\mathrm{var}\,(a_1X_1 + \cdots + a_NX_N) = a_1^2\mathrm{var}\,(X_1) + \cdots + a_N^2\mathrm{var}\,(X_N)$$

for any $a_1, \cdots, a_N \in \mathbb{R}$.

We also recall the following fact (see e.g. [**Goo60**]):

LEMMA 2.3.31. *Let X and Y are two independent random variables with $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$. The variance of their product is*

$$\begin{aligned}\mathrm{var}\,(XY) &= (\mathbb{E}X)^2\mathrm{var}\,(Y) + (\mathbb{E}Y)^2\mathrm{var}\,(X) + \mathrm{var}\,(X)\mathrm{var}\,(Y)\\ &= \mathbb{E}\left(X^2\right)\mathbb{E}\left(Y^2\right) + (\mathbb{E}X)^2(\mathbb{E}Y)^2.\end{aligned}$$

In particular, if $\mathbb{E}X = \mathbb{E}Y = \mu$ and $\mathrm{var}\,(X) = \mathrm{var}\,(Y) = \sigma^2$, then

$$\mathrm{var}\,(XY) = \sigma^2\left(2\mu^2 + \sigma^2\right).$$

2.4. Conditional expectation

Let X be a random variable. The conditional expectation of X given a σ-algebra $\mathscr{F}$ is any random variable $\tilde{X}$ which is measurable with respect to $\mathscr{F}$ and

$$\int_A X\,\mathrm{d}\mathbb{P} = \int_A \tilde{X}\,\mathrm{d}\mathbb{P} \quad \text{for all } A \in \mathscr{F}.$$

By using Radon-Nikodym theorem and [**Dur19**, Theorem 4.1.2], there exists a conditional expectation of X given $\mathscr{F}$ which is unique a.e., and therefore we may denote

$$\mathbb{E}(X|\mathscr{F}) := \tilde{X},$$

which is well-defined a.s. (here and after, we will not explicitly denote "a.s."). If $\mathscr{F}_1 \subset \mathscr{F}_2$, from [**Dur19**, Theorem 4.1.12] we know that

$$\mathbb{E}\left(\mathbb{E}(X|\mathscr{F}_1)|\mathscr{F}_2\right) = \mathbb{E}(X|\mathscr{F}_1) = \mathbb{E}\left(\mathbb{E}(X|\mathscr{F}_2)|\mathscr{F}_1\right).$$

In the case when $\mathbb{E}X^2 < \infty$, we define $\operatorname{var}(X|\mathscr{F}) := \mathbb{E}(X^2|\mathscr{F}) - \mathbb{E}(X|\mathscr{F})^2$. It was showed in [**Dur19**, Exercise 4.1.7] the following *total variance formula* holds:

$$\operatorname{var}(X) = \mathbb{E}\left(\operatorname{var}(X|\mathscr{F})\right) + \operatorname{var}\left(\mathbb{E}(X|\mathscr{F})\right).$$

The definition of conditional expectation given a σ-field contains conditioning on a random variable as a special case: Let X and Y are random variables. The conditional expectation of X on Y is defined by

$$\mathbb{E}(X|Y) := \mathbb{E}(X|\sigma(Y)),$$

where $\sigma(Y)$ is the σ-field generated by Y, i.e. the smallest σ-field for which Y is measurable with respect to $\sigma(Y)$. This definition immediately gives us the *conditional expectation formula*:

$$\mathbb{E}X = \mathbb{E}\left(\mathbb{E}(X|Y)\right), \tag{2.4.1}$$

and we now have the *conditional variance formula*:

$$\operatorname{var}(X) = \mathbb{E}\left(\operatorname{var}(X|Y)\right) + \operatorname{var}\left(\mathbb{E}(X|Y)\right). \tag{2.4.2}$$

Suppose that X and Y have joint density $f(x,y)$ (can be p.d.f. or p.m.f.). Given any function g, it was showed in [**Dur19**, Example 4.1.6] that $\mathbb{E}(g(X)|Y)$ is a random variable satisfying

$$\mathbb{E}(g(X)|Y=y) = \int g(x)\mathbb{P}(X=x|Y=y)\,\mathrm{d}x,$$

where $\mathbb{P}(X=x|Y=y)$ is the conditional probability given by

$$\mathbb{P}(X=x|Y=y) = \frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(Y=y)} = \frac{f(x,y)}{\int f(x,y)\,\mathrm{d}y}.$$

2.5. Discrete random variable

We now introduce discrete random variables, which will be helpful to construct some examples to let us have a better understanding while learning statistics.

EXAMPLE 2.5.1. Let Ω be a countable set in $\mathbb{R}$ and let $\mathscr{F} := 2^{\Omega}$ be the collections of all subsets of Ω. Let $p : \Omega \to [0,1]$ be any function satisfies

$$\sum_{x\in\Omega} p(x) = 1.$$

If we define $\mathbb{P}(A) := \sum_{x\in A} p(x)$ for all $A \in \mathscr{F}$, then $(\Omega,\mathscr{F},\mathbb{P})$ forms a probability space. In this case, any function $X : \Omega \to \mathbb{R}$ is a random variable.

DEFINITION 2.5.2. The probability space $(\Omega,\mathscr{F},\mathbb{P})$ in Example 2.5.1 is called a *discrete probability space*. The random variable $X:\Omega\to\mathbb{R}$ in Example 2.5.1 is called a *discrete random variable*. The function $p=p_X:\Omega\to[0,1]$ in Example 2.5.1 is called a *probability mass function* (later we use the abbreviation "p.m.f."). The support of $p(x)$ is the subset of Ω defined by

$$\mathrm{supp}\,(p) := \{x\in\Omega : p(x)>0\}\,.$$

We will display a p.m.f. for the values in its support, and it always understood that $p(x)=0$ otherwise.

REMARK 2.5.3. In fact, each discrete random variable X has a unique p.m.f. p.

THEOREM 2.5.4 (change of variable formula, a special case of [**Dur19**, Theorem 1.6.9]). *Let X be a discrete random variable with p.m.f. f and let $\phi:\mathbb{R}\to\mathbb{R}$ satisfies either one of the followings:*

$$\phi\geq 0 \quad \textit{or} \quad \sum_{x\in\Omega}|\phi(x)|p(x)<\infty.$$

Then

$$\mathbb{E}(\phi(X))=\sum_{x\in\Omega}\phi(x)p(x).$$

REMARK 2.5.5. One can check that $\mathbb{E}|\phi(X)|=\sum_{x\in\Omega}|\phi(x)|p(x)$. Similar as above, one can easily compute k^{th} moment $\mathbb{E}X^k=\sum_{x\in\Omega}x^k p(x)$.

EXAMPLE 2.5.6 (Bernuolli random variable). The *Bernuolli random vaviable* (with parameter $0<\alpha<1$) is a discrete random variable with density

$$p(x)=\begin{cases}1-\alpha & \text{if } x=0,\\ \alpha & \text{if } x=1.\end{cases}$$

Its k^{th} moment is

$$\sum_{x\in\{0,1\}}x^k p(x)=p(1)=\alpha.$$

From this, one can easily see that its mean (first moment) is α and its variance is

$$\sum_{x\in\{0,1\}}x^2p(x)-\left(\sum_{x\in\{0,1\}}xp(x)\right)^2=\alpha-\alpha^2=\alpha(1-\alpha).$$

For example, the Bernuolli random variable can be used to described the probability of getting a head ($x=1$) with probability α while tossing a coin.

EXAMPLE 2.5.7 (geometric random variable). Given a coin described in Example 2.5.6. The probability of getting the first head at the n^{th}-attempt (providing obtaining $(n-1)$ tails before), assuming that each trial is independent, is

$$p(n):=(1-\alpha)^{n-1}\alpha. \tag{2.5.1}$$

Note that the function $p : \mathbb{N} \to [0,1]$ satisfies

$$\sum_{n=1}^{\infty} p(n) = \alpha \sum_{n=1}^{\infty} (1-\alpha)^{n-1} = 1,$$

therefore a p.m.f.. The *geometric random variable* (with parameter $0 < \alpha < 1$) X is a discrete random variable with density $p : \mathbb{N} \to [0,1]$ given in (2.5.1), see also [**LM21**, Section 4.4]. Since the mapping $\alpha \mapsto \sum_{n=1}^{\infty} (1-\alpha)^{n-1}$ is a power series centered at 1 with radius of converge 1 (see e.g. my other lecture notes [**Kow23, Kow24**]), then we may differentiate "term-by-term" to see that

$$\begin{aligned} \sum_{n=1}^{\infty} n(1-\alpha)^{n-1} &= -\frac{\mathrm{d}}{\mathrm{d}\alpha} \left(\sum_{n=0}^{\infty} (1-\alpha)^n \right) = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \left((1-\alpha) \sum_{n=1}^{\infty} (1-\alpha)^{n-1} \right) \\ &= -\frac{\mathrm{d}}{\mathrm{d}\alpha} \left(\frac{1}{\alpha} - 1 \right) = \frac{1}{\alpha^2}. \end{aligned} \tag{2.5.2}$$

From this, we now can compute the mean of the geometric random variable:

$$\mathbb{E}X = \sum_{n=1}^{\infty} np(n) = \alpha \sum_{n=1}^{\infty} n(1-\alpha)^{n-1} = \frac{1}{\alpha}.$$

By acting the differential operator $-\frac{\mathrm{d}}{\mathrm{d}\alpha}$ on (2.5.2), we now see that

$$\sum_{n=2}^{\infty} n(n-1)(1-\alpha)^{n-2} = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \left(\sum_{n=1}^{\infty} n(1-\alpha)^{n-1} \right) = -\frac{\mathrm{d}}{\mathrm{d}\alpha} \left(\frac{1}{\alpha^2} \right) = \frac{2}{\alpha^3},$$

then we see that

$$\mathbb{E}X(X-1) = \sum_{n=2}^{\infty} n(n-1)p(n) = \alpha(1-\alpha) \sum_{n=2}^{\infty} n(n-1)(1-\alpha)^{n-2} = \frac{2(1-\alpha)}{\alpha^2},$$

therefore its variance is

$$\begin{aligned} \mathrm{var}\,(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X(X-1) + \mathbb{E}X - (\mathbb{E}X)^2 \\ &= \frac{2(1-\alpha)}{\alpha^2} + \frac{1}{\alpha} - \frac{1}{\alpha^2} = \frac{2(1-\alpha)+\alpha-1}{\alpha^2} = \frac{1-\alpha}{\alpha^2}. \end{aligned}$$

EXAMPLE 2.5.8 (binomial random variable). Given a coin described in Example 2.5.6. Given $n \in \mathbb{N}$, the probability of getting exactly k heads in n independent of trials is

$$p(k) = \binom{n}{k} (1-\alpha)^{n-k} \alpha^k, \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{2.5.3}$$

One can use binomial theorem to see that

$$\sum_{k=1}^{n} p(k) = 1,$$

therefore a p.m.f.. The *binomial random variable* (with parameter n,α) is a discrete random variable X with density $p:\mathbb{N}\to[0,1]$ given in (2.5.3). We usually denoted as $X\sim\mathscr{B}(n,\alpha)$. Note that $\mathscr{B}(1,\alpha)$ is exactly the Bernuolli random variable in Example 2.5.6.

EXERCISE 2.5.9. Let $X\sim\mathscr{B}(n,\alpha)$. Prove that $\mathbb{E}X=n\alpha$.

EXAMPLE 2.5.10 (Poisson random variable). Let $X\sim B(n,\alpha)$. In fact, for each fixed $\lambda>0$, one can show that

$$\lim_{\substack{n\to\infty\\ \alpha\to 0\\ n\alpha=\lambda}}\mathbb{P}(X=k)=\lim_{\substack{n\to\infty\\ \alpha\to 0\\ n\alpha=\lambda}}\binom{n}{k}(1-\alpha)^{n-k}\alpha^k=\frac{e^{-\lambda}\lambda^k}{k!},$$

which is the well-known Poisson limit [LM21, Theorem 4.2.1]. But then, in 1898, von Bortkiewicz [vB98] transform Poisson's limit into Poisson random variable: The random variable Y is said to be *Poisson distribution* if

$$p(k)=\mathbb{P}(Y=k)=\frac{e^{-\lambda}\lambda^k}{k!}\quad\text{for } k=0,1,2,\cdots,$$

where λ is a positive constant. One can check that

$$\sum_{k=0}^{\infty}p(k)=1,$$

and thus such Y is a discrete random variable with p.m.f. $p:\mathbb{Z}_{\geq 0}\to[0,1]$. In this case, we write $Y\sim\mathscr{P}(\lambda)$.

EXERCISE 2.5.11. Let $Y\sim\mathscr{P}(\lambda)$. Prove that $\mathbb{E}Y=\lambda$.

2.6. Continuous random variable

It is now natural to introduce the following definition.

DEFINITION 2.6.1. A random variable is said to be *continuous* if its c.d.f. (Definition 2.3.14) is continuous.

By using a well-known theorem from measure theory [WZ15, Theorem 7.29], one has the following fact.

THEOREM 2.6.2. *Let X be a continuous random variable and denote $F:\mathbb{R}\to[0,1]$ be its c.d.f. (Definition 2.3.14). Then the following are equivalent:*

(1) *F is absolutely continuous;*

(2) *The fundamental theorem of calculus holds true for F, more precisely, its derivative*[2] *$f\equiv F'$ exists a.e. in $\mathbb{R}$, $f\in L^1_{\rm loc}(\mathbb{R})$ and*

$$F(x)=\int_{\infty}^{x}f(t)\,\mathrm{d}t\quad\textit{for all } x\in\mathbb{R}.$$

[2] also known as the Radon-Nikodym derivative.

REMARK 2.6.3 (Cantor-Lebesgue function). In fact, there exists a nondecreasing continuous function $\phi : [0,1] \to [0,1]$ with $\phi(0) = 0$, $\phi(1) = 1$ and $\phi' = 0$ a.e. in $[0,1]$. In this case, one sees that

$$\phi(1) - \phi(0) = 1 \neq 0 = \int_0^1 \phi'(t)\,\mathrm{d}t.$$

DEFINITION 2.6.4. Let X be a continuous random variable. If its c.d.f. (Definition 2.3.14) is absolutely continuous, then we call its derivative f the *probability density function* (later we use the abbreviation "p.d.f."). In this case, we say that X is a *continuous random variable with p.d.f. f*.

THEOREM 2.6.5 (change of variable formula, a special case of [Dur19, Theorem 1.6.9]). *Let X be a continuous random variable with p.d.f. f and let $\phi : \mathbb{R} \to \mathbb{R}$ satisfies either one of the followings:*

$$\phi \geq 0 \quad \textit{or} \quad \int_{\mathbb{R}} |\phi(t)| f(t)\,\mathrm{d}t < \infty.$$

Then

$$\mathbb{E}(\phi(X)) = \int_{\mathbb{R}} \phi(t) f(t)\,\mathrm{d}t.$$

REMARK 2.6.6. One can check that $\mathbb{E}|\phi(X)| = \int_{\mathbb{R}} |\phi(t)| f(t)\,\mathrm{d}t$.

EXAMPLE 2.6.7. Let $k \in \mathbb{N}$. By choosing $\phi(t) = t^k$ for all $t \in \mathbb{R}$, one sees that the expectation can be easily computed by the k^{th} *moment* of a random variable X by

$$\mathbb{E}X^k = \int_{\mathbb{R}} t^k f(t)\,\mathrm{d}t.$$

Note that mean $\mathbb{E}X$ is exactly the first moment of X. The variance can be expressed as

$$\mathrm{var}\,(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \int_{\mathbb{R}} t^2 f(t)\,\mathrm{d}t - \left(\int_{\mathbb{R}} t f(t)\,\mathrm{d}t \right)^2.$$

2.6.1. Exponential distribution. Since

$$\int_0^\infty e^{-t}\,\mathrm{d}t = -e^{-t}\Big|_{t=0}^{t\to\infty} = 1,$$

then we see that $f(t) = \chi_{t\geq 0} e^{-t}$ is a p.d.f. (Definition 2.6.4). By using Theorem 2.3.17, there exists a continuous random variable X with p.d.f. $f(t) = \chi_{t\geq 0} e^{-t}$.

EXERCISE 2.6.8. Let X be a continuous random variable with p.d.f. $f(t) = \chi_{t\geq 0} e^{-t}$. Use Example 2.6.7 and mathematical induction to check that

$$\mathbb{E}X^k = \int_0^\infty t^k e^{-t}\,\mathrm{d}t = k!$$

for all $k \in \mathbb{N}$.

For each $\lambda > 0$, we see that the random variable $Y = X/\lambda$ has k^{th} moment (i.e. choose $\phi(t) = (t/\lambda)^k$ in Theorem 2.6.5)

$$\mathbb{E}Y^k = \mathbb{E}(X^k/\lambda^k) = \frac{1}{\lambda^k}\mathbb{E}X^k = \frac{k!}{\lambda^k} \tag{2.6.1}$$

without explicitly computed the p.d.f. (Definition 2.6.4) of Y. From this, it is easy to see that

$$\mathbb{E}Y = \frac{1}{\lambda}, \quad \operatorname{var}(Y) = \mathbb{E}Y^2 - (\mathbb{E}Y)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Indeed, one can easily compute the p.d.f. of Y as follows:

$$\begin{aligned}\mathbb{P}(Y \leq x) = \mathbb{P}(X \leq \lambda x) &= \begin{cases} \int_0^{\lambda x} e^{-t}\,\mathrm{d}t = \int_0^x \lambda e^{-\lambda s}\,\mathrm{d}s & ,x > 0, \\ 0 & ,x \leq 0, \end{cases} \\ &= \int_{-\infty}^x \chi_{t\geq 0}\lambda e^{-\lambda t}\,\mathrm{d}t,\end{aligned}$$

which shows that $Y = X/\lambda$ is a continuous random variable with p.d.f. $f(t) = \chi_{t\geq 0}\lambda e^{-\lambda t}$.

DEFINITION 2.6.9. The above mentioned random variable Y is called the *exponential distribution with parameter* $\lambda > 0$, and we denote $Y \sim \mathscr{E}(\lambda)$.

2.6.2. Normal distribution. By using the Fubini's theorem for Lebesgue integral[3], one sees that

$$\begin{aligned}\left(\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t\right)^2 &= \left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{t^2}{2}}\,\mathrm{d}t\right)\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{s^2}{2}}\,\mathrm{d}s\right) \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^2} e^{-\frac{t^2+s^2}{2}}\,\mathrm{d}t\,\mathrm{d}s = \frac{1}{2\pi}\int_0^{2\pi} \overbrace{\left(\int_0^\infty e^{-\frac{r^2}{2}} r\,\mathrm{d}r\right)}^{=1}\mathrm{d}\theta = 1,\end{aligned}$$

which shows that $f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$ is a p.d.f. (Definition 2.6.4). By using Theorem 2.3.17, there exists a continuous random variable X with p.d.f. $f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. Since f is an even function, then

$$\mathbb{E}X = \int_{\mathbb{R}} t f(t)\,\mathrm{d}t = 0.$$

EXERCISE 2.6.10. Let X be a continuous random variable with p.d.f. $f(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. Prove that $\operatorname{var}(X) = 1$.

For each $\mu \in \mathbb{R}$ and $\sigma > 0$, let $Y = \sigma X + \mu$. Then one can easily compute

$$\mathbb{E}Y = \sigma\mathbb{E}X + \mu = \mu, \quad \operatorname{var}(Y) = \operatorname{var}(\sigma X + \mu) = \sigma^2\operatorname{var}(X) = \sigma^2$$

[3] Lebesgue integral has many properties that holds true on unbounded region, which is better than Riemann integral, which only can be well-defined on bounded region with sufficient smooth boundary.

without explicitly computed the p.d.f. (Definition 2.6.4) of Y. Indeed, one can easily compute the p.d.f. of Y as follows:

$$\mathbb{P}(Y \leq x) = \mathbb{P}\left(X \leq \frac{x-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-\mu)^2}{2\sigma^2}}\,\mathrm{d}s,$$

which shows that $Y = \sigma X + \mu$ is a continuous random variable with p.d.f. $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$.

DEFINITION 2.6.11. The above mentioned random variable Y is called the *normal distribution with mean* μ *and variance* σ^2, and we denote $Y \sim \mathcal{N}(\mu, \sigma^2)$. We often refer $Z \sim \mathcal{N}(0,1)$ the *standard normal distribution*.

We now ready to state an important theorem, which can be found in [**Dur19**, Theorem 3.4.1].

THEOREM 2.6.12 (central limit theorem for i.i.d. sequences). *Let* $X_1, X_2, \cdots$ *be i.i.d. random variables with* $\mathbb{E}X_i = \mu$ *and* $\mathrm{var}\,(X_i) = \sigma^2 \in (0,\infty)$. *If* $\overline{X}_N = \frac{X_1+\cdots+X_N}{N}$ *then*

$\sigma^{-1}N^{1/2}(\overline{X}_N - \mu)$ *converges in distribution to the standard normal distribution as* $N \to \infty$.

In other words,

$$\lim_{N\to\infty} \mathbb{P}\left(\sigma^{-1}N^{1/2}(\overline{X}_N - \mu) \leq x\right) = \mathbb{P}(Z \leq x) \quad \textit{for all } x \in \mathbb{R}.$$

EXERCISE 2.6.13. Let $X_1, X_2, \cdots$ be i.i.d. random variables in $\mathcal{N}(\mu, \sigma^2)$. Show that $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$ and

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1).$$

It is also possible to establish central limit theorem for some non-i.i.d. sequences, see e.g. the Lindeberg-Feller theorem [**Dur19**, Theorem 3.4.10].

2.6.3. Chi-squared (χ^2) distributions , t distributions and F distributions. Central limit theorem shows the importance of normal distribution, therefore we now also introduce three distributions closely related to normal: the chi-squared (χ^2) distributions, t distributions and F distributions. These distributions will then be used to describe the sampling variability of several statistics on which important inferential procedures are based.

EXAMPLE 2.6.14. For a positive integer ν, let $Z_1, \cdots, Z_\nu$ are i.i.d. $\mathcal{N}(0,1)$. Then the chi-squared distribution with ν degree of freedom (later we use the abbreviation "d.f.") is defined to be the distribution of the random variable

$$X = Z_1^2 + \cdots + Z_\nu^2.$$

This will sometimes be denoted by $X \sim \chi_\nu^2$. Its p.d.f. is

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} \quad \text{for all } x > 0,$$

with expectation $\mathbb{E}X = \nu$ and variance $\operatorname{var}(X) = 2\nu$. It is also worth to mention that

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2} \text{ if } X_1 \sim \chi^2_{\nu_1} \text{ and } X_2 \sim \chi^2_{\nu_2} \text{ are independent}$$

as well as

$$X_1 - X_2 \sim \chi^2_{\nu_1-\nu_2} \text{ if } X_1 \sim \chi^2_{\nu_1} \text{ and } X_2 \sim \chi^2_{\nu_2} \text{ are independent provided } \nu_1 > \nu_2.$$

EXAMPLE 2.6.15 (Gosset's theorem [**Gos08**]). Let $Z \sim \mathscr{N}(0,1)$ and let $Y \sim \chi^2_\nu$ be independent random variables. Then the *t distribution* with ν d.f. is defined to be the distribution of the ratio

$$T = \frac{Z}{\sqrt{Y/\nu}}.$$

We will sometimes abbreviate this distribution by $T \sim t_\nu$. Its p.d.f. is

$$f(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\nu/2)} \frac{1}{(1+t^2/\nu)^{\frac{\nu+1}{2}}} \quad \text{for all } t \in \mathbb{R}.$$

In view of the (strong) law of large number (Theorem 2.3.24), it make sense that the t distribution would be "close" to the standard normal for large ν. In addition,

(1) $\mathbb{E}T = 0$ for $\nu > 1$, otherwise undefined;
(2) $\operatorname{var}(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$, $\operatorname{var}(T) = +\infty$ for $1 < \nu \leq 2$, and otherwise undefined.

EXAMPLE 2.6.16. Let $Y_1 \sim \chi^2_{\nu_1}$ and $Y_2 \sim \chi^2_{\nu_2}$ are independent random variables. The F distribution with ν_1 numerator d.f. and ν_2 denominator d.f. is defined to be the distribution of the ratio

$$F = \frac{Y_1/\nu_1}{Y_2/\nu_2}.$$

We will sometimes abbreviate this distribution by $F \sim \mathscr{F}_{\nu_1,\nu_2}$. It is not difficult to see that $t^2_\nu = \mathscr{F}_{1,\nu}$. Its p.d.f. is

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{x^{\frac{\nu_1}{2}}}{(1+\frac{\nu_1}{\nu_2}x)^{\frac{\nu_1+\nu_2}{2}}} \quad \text{for all } x > 0.$$

In addition,

(1) $\mathbb{E}F = \frac{\nu_2}{\nu_2-2}$ for $\nu_2 > 2$, otherwise undefined;
(2) $\operatorname{var}(F) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$, otherwise undefined.

2.6.4. Gamma distributions. We now introduce a versatile two-parameter family of continuous probability distributions. The exponential distributions and the χ^2 distributions are special cases of the gamma distributions. To define the family of gamma distributions, we first need to introduce a function that plays an important role in many branches of mathematics.

EXERCISE 2.6.17. Show that $\int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x < \infty$ if and only if $\alpha > 0$.

Suggested by the above exercise, it is natural to consider the following definition.

DEFINITION 2.6.18. For $\alpha > 0$, the gamma function $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, \mathrm{d}x.$$

In fact, $\Gamma(1/2) = \sqrt{\pi}$. The most important properties of the gamma function are the followings:

EXERCISE 2.6.19. Show that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$ for all $\alpha > 1$. In addition, show that $\Gamma(n) = (n-1)!$ for all $n \in \mathbb{N}$.

It is possible to extend $\Gamma : \mathbb{C} \setminus \mathbb{Z}_{\leq 0} \to \mathbb{C}$ as an analytic function, with pole of order 1 at each point of $\mathbb{Z}_{\leq 0}$, see e.g. my lecture note on complex analysis [**Kow23**]. The following fact will prove useful for several computations that follow.

EXERCISE 2.6.20. Show that[4]

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta} \, \mathrm{d}x = \beta^\alpha \Gamma(\alpha)$$

for all $\alpha, \beta > 0$.

Now it is natural to consider the following definition.

DEFINITION 2.6.21. A continuous random variable X is said to have a *gamma distribution* if the p.d.f. of X is

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad \text{for all } x > 0,$$

with parameters $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter). In this case, we denote $X \sim \mathrm{Gamma}(\alpha, \beta)$. When $\beta = 1$, X is said to have a *standard gamma distribution*.

We see that the χ^2 distributions are special cases of the Γ distribution via the formula $\chi_\nu^2 = \mathrm{Gamma}(\frac{\nu}{2}, 2)$, and the exponential distributions are special cases of the Γ distribution via the formula $\mathscr{E}(\lambda) = \mathrm{Gamma}(1, 1/\lambda)$.

EXERCISE 2.6.22. Let $X \sim \mathrm{Gamma}(\alpha, \beta)$. Show that

$$\mathbb{E}X^k = \beta^k \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \quad \text{for all } k \in \mathbb{N}.$$

Therefore, $\mathbb{E}X = \alpha\beta$ and $\mathrm{var}\,(X) = \alpha\beta^2$.

EXERCISE 2.6.23. The c.d.f. of $X \sim \mathrm{Gamma}(\alpha, 1)$ is

$$G(x) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)} \, \mathrm{d}y \quad \text{for all } x > 0,$$

which is called the *incomplete gamma function*. Suppose that $Y \sim \mathrm{Gamma}(\alpha, \beta)$, show that its c.d.f. is

$$\mathbb{P}(Y \leq x) = G(x/\beta).$$

[4] It is natural to define the fractional Laplacian $(-\Delta)^\alpha := \frac{1}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(-\Delta)^{-1}} \, \mathrm{d}x$. In fact, such definition of fractional Laplacian is equivalent (in some sense) to the one defined using Fourier transform.

2.7. Joint probability distribution

The joint probability distribution is the probability distribution of all possible pairs of outputs of two random variables that are defined on the same probability space. In other words, given random variables $X_1, \cdots, X_N$ and Y, we are now interested in the event of the form

$$\{(X_1, \cdots, X_N) \in B\} := \left\{(x_1, \cdots, x_N) \in \Omega^N : (X_1(x_1), \cdots, X_N(x_N)) \in B\right\} \quad \text{for Borel sets } B \text{ in } \mathbb{R}^N.$$

Accordingly, the joint c.d.f. F of random variables $X_1, \cdots, X_N$ is given by

$$F(x_1, \cdots, x_N) := \mathbb{P}(X_1 \leq x_1, \cdots, X_N \leq x_N) \equiv \mathbb{P}\left(\bigcap_{i=1}^{N} \{X_i \leq x_i\}\right). \tag{2.7.1}$$

If $X_1, \cdots, X_N$ are independent, then we see that $F(x_1, \cdots, x_N) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq x_i)$.

(1) If $X_1, \cdots, X_N$ are all discrete random variables, the *joint p.m.f.* of the variables is the function

$$p(x_1, \cdots, x_N) = \mathbb{P}\left(\bigcap_{i=1}^{N} \{X_i = x_i\}\right).$$

(2) Let $X_1, \cdots, X_N$ be continuous random variables. If there exists $f \in L^1(\mathbb{R}^n)$ with $\int_{\mathbb{R}^n} f = 1$ such that

$$\mathbb{P}\left((X_1, \cdots, X_N) \in B\right) = \int_B f(x_1, \cdots, x_N) \, \mathrm{d}(x_1, \cdots, x_N)$$

for all Borel sets B in $\mathbb{R}^N$. Then we call such f the *joint p.d.f.* of $X_1, \cdots, X_N$. If $f \in C^\infty(\mathbb{R}^N)$, then it is also worth to mention that

$$f(x_1, \cdots, x_N) = \partial_1 \cdots \partial_N F(x_1, \cdots, x_N) \quad \text{for all } (x_1, \cdots, x_N) \in \mathbb{R}^N,$$

where F is the joint c.d.f. of $X_1, \cdots, X_N$ and $\partial_i = \frac{\partial}{\partial x_i}$ is the i^{th} partial derivative.

EXAMPLE 2.7.1. If X and Y have joint p.d.f. $f(x,y)$, then

$$\mathbb{P}(X < Y) = \int_{\{x<y\}} f(x,y) \, \mathrm{d}(x,y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(x,y) \, \mathrm{d}x \, \mathrm{d}y.$$

It is important to mention the following fact:

THEOREM 2.7.2 (Law of the unconcious statistician I). *Let $X_1, \cdots, X_N$ be jointly distribution discrete random variables with joint p.m.f. p. Then the expected value of $h(X_1, \cdots, X_N)$ is given by*

$$\mathbb{E}h(X_1, \cdots, X_N) = \sum_{x_1} \cdots \sum_{x_N} h(x_1, \cdots, x_N) p(x_1, \cdots, x_N).$$

THEOREM 2.7.3 (Law of the unconcious statistician II). *Let $X_1, \cdots, X_N$ be jointly distribution continuous random variables with joint p.d.f. f. Then the expected value of $h(X_1, \cdots, X_N)$ is given by*

$$\mathbb{E}h(X_1, \cdots, X_N) = \int_{\mathbb{R}^N} h(x_1, \cdots, x_N) f(x_1, \cdots, x_N) \, \mathrm{d}(x_1, \cdots, x_N).$$

DEFINITION 2.7.4. Let X and Y be discrete random variables with joint p.m.f. p. The *marginal p.m.f.* of X and Y, denoted by $p_X(x)$ and $p_Y(x)$, respectively, are given by

$$p_X(x) = \sum_y p(x,y), \quad p_Y(y) = \sum_x p(x,y).$$

DEFINITION 2.7.5. Let X and Y be continuous random variables with joint p.d.f. f. The *marginal p.d.f.* of X and Y, denoted by $f_X(x)$ and $f_Y(x)$, respectively, are given by

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,\mathrm{d}y, \quad f_Y(y) = \int_{\mathbb{R}} f(x,y)\,\mathrm{d}x.$$

REMARK 2.7.6. If X and Y have joint p.d.f., then both X and Y are absolutely continuous random variables (see Theorem 2.6.2 above).

It is important to mention that X and Y are independent if and only if

$$\mathbb{P}(X \leq x, Y \leq y) = \mathbb{P}(X \leq x)\mathbb{P}(X \leq y) \quad \text{for all } x,y \in \mathbb{R}.$$

Similarly,

THEOREM 2.7.7 ([**HPS71**, page 143]). *Let X and Y be two discrete random variables with joint p.m.f. p. The following are equivalent:*

(1) *X and Y are independent;*
(2) *$p(x,y) = p_X(x)p_Y(y)$ for all x,y, where p_X and p_Y are p.m.f. of X and Y, respectively.*
(3) *there exist p.m.f. p_1 and p_2 such that $p(x,y) = p_1(x)p_2(x)$ for all x,y.*

THEOREM 2.7.8 ([**HPS71**, page 143]). *Let X and Y be two absolutely continuous random variables (see Theorem 2.6.2 above) with joint p.m.f. f. The following are equivalent:*

(1) *X and Y are independent;*
(2) *$f(x,y) = f_X(x)f_Y(y)$ for all $x,y \in \mathbb{R}$, where f_X and f_Y are p.d.f. of X and Y, respectively.*
(3) *there exist p.d.f. f_1 and f_2 such that $f(x,y) = f_1(x)f_2(x)$ for all $x,y \in \mathbb{R}$.*

CHAPTER 3

Point estimation

3.1. Point Estimation and sample distribution

We now begin to make the transition between probability and inferential statistics. Given a population, at least when then population is finite, in principle we believe that its mean, median, standard deviation, and various other characteristics can be defined and computed. However, this is not plausible in practical since the population may be very large, or even infinite. Therefore, we take N samples from the population and study the sample, where N is much smaller than the number of population. However, the values of the individual sample observations vary from sample to sample, so in general the value of any quantity computed from sample data, and the value of a sample characteristic used as an estimate of the corresponding population characteristics, will virtually never coincide with what is being estimated. A statistic is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, we now introduce the following definition as in [**DBC21**, Section 6.1]:

DEFINITION 3.1.1. A *statistic* is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

EXAMPLE 3.1.2. For example, the sample mean, regarded as a statistic (before a sample has been selected or an experiment has been carried out), is denoted by $\overline{X}$ (or $\overline{X}_N$ to emphasize that there are N samples), see (3.1.1) below; the calculated value of this statistic from a particular sample is $\overline{x}$.

Any statics, being a random variable, has a probability distribution. The probability distribution of any particular statistic depends not only on the population distribution (e.g. normal, uniform, etc.) and the sample size N but also on the method of sampling. Our next definition describes a sampling method often encountered, at least approximately, in practice.

DEFINITION 3.1.3. The random variables $X_1, X_2, \cdots, X_N$ are said to form a (simple) random sample of size N if they are i.i.d. (Definition 2.3.23).

Given a sample of N observations from a population, we will be calculating estimates of the population mean, median, standard deviation, and various other population characteristics (*parameters*). When discussing general concepts and methods of inference, it is convenient to have a generic symbol for the parameter of interest. We will use the Greek letter θ for this purpose.

A point estimate of a parameter θ is a single number that can be regarded as a sensible value of θ. A point estimate is obtained by selecting a suitable statistic and determining its value from the given sample data. The selected statistic $\hat{\theta}$ is called the *point estimator* of θ. For example, given samples $X_1, \cdots, X_N$, in view of the law of large numbers (Theorem 2.3.24 and Theorem 2.3.25), it is natural to select the sample mean

$$\hat{\theta} \equiv \overline{X} := \frac{X_1 + \cdots + X_N}{N} \tag{3.1.1}$$

as a point estimator of the population mean. As the sample size N increases, according to the central limit theorem (Theorem 2.6.12) the sampling distribution of $\overline{X}$ becomes increasingly normal, irrespective of the population distribution from which values were samples.

3.1.1. Assessing Estimators: Accuracy and Precision. However, as we mentioned above, it is not practical to take $N \to \infty$; in practice the sample size N is much smaller than the population size. We now want to give some "norms" to measure "how good" is a chosen point estimator $\hat{\theta}$ of a parameter θ which we believe that is a characteristic of a population. First of all, we hope that our estimators are "accurate" in the following sense:

DEFINITION 3.1.4 (unbiased estimator). Let $\hat{\theta}$ be a point estimator (a random variable) of a parameter θ. The difference $\mathbb{E}\hat{\theta} - \theta$ is called the *bias* of $\hat{\theta}$. Accordingly, we say that a point estimator $\hat{\theta}$ is said to be an *unbiased* estimator of θ if $\mathbb{E}\hat{\theta} = \theta$ for every possible value of θ.

Let $X_1, X_2, \cdots, X_N$ are random sample with $\mathbb{E}X_i = \mu$. Then the sample mean (3.1.1) is an unbiased estimator:

$$\mathbb{E}\overline{X} = \frac{\mathbb{E}X_1 + \cdots + \mathbb{E}X_N}{N} = \mu, \tag{3.1.2}$$

regardless of the value of μ *and the sample size* N. This estimator may not good if the samples X_i have large variance. In this case, it is possible that two statisticians conclude two very different estimators even using the same estimator. Therefore, we also hope that our estimators are "precise" in the following sense:

DEFINITION 3.1.5 (unbiased standard error). Let $\hat{\theta}$ be a point estimator (a random variable) of a parameter θ. The *standard error* of $\hat{\theta}$ is its standard deviation:

$$\sigma_{\hat{\theta}} := \sqrt{\mathrm{var}\,(\hat{\theta})}.$$

EXAMPLE 3.1.6. Let $X \sim \mathscr{B}(N, \theta)$ be the binomial random variable (Example 2.5.8). We consider the *sample proportion* $\hat{P}$ given by

$$\hat{P} := \frac{X}{N}.$$

One sees that

$$\mathbb{E}\hat{P} = \frac{\mathbb{E}X}{N} = \theta,$$

which shows that $\hat{P}$ is an unbiased estimator (Definition 3.1.4) of the parameter θ, *regardless of the value of θ and the sample size N*. The standard error of the estimator is

$$\sigma_{\hat{P}} = \sqrt{\mathrm{var}\left(\frac{X}{N}\right)} = \sqrt{\frac{1}{N^2}\mathrm{var}X} = \sqrt{\frac{\theta(1-\theta)}{N}}.$$

Since θ is unknown (else why estimate?), we could substitute it by $\tilde{\theta} = x/N$, where x is an observed value of X after performing experiments, yielding the estimated standard error

$$s_{\hat{P}} = \sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{N}}.$$

The above example suggests the following notation.

DEFINITION 3.1.7 (unbiased standard error, continued). If the standard error $\sigma_{\hat{\theta}}$ (Definition 3.1.5) itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the *estimated standard error* of $\hat{\theta}$, which can be denoted by $s_{\hat{\theta}}$.

EXAMPLE 3.1.8. Let $X_1, X_2, \cdots, X_N$ are random sample with $\mathbb{E}X_i = \mu$ and $\mathrm{var}(X_i) = \sigma$, provided that $\mathbb{E}X_i^2 < \infty$ for all $i = 1, \cdots, N$. As demonstrated in (3.1.2), the sample mean $\overline{X}$ given in (3.1.1) is an unbiased estimator of the parameter μ, *regardless of the value of μ and the sample size N*. By using Exercise 2.3.30, one also sees that

$$\sigma_{\overline{X}} = \sqrt{\mathrm{var}(\overline{X})} = \sqrt{\frac{\mathrm{var}(X_1)}{N}} = \frac{\sigma}{\sqrt{N}},$$

which shows that the standard error of the sample mean decreases (its precision improves) with increasing sample size. Again, since the value of σ is almost always unknown, we can estimate the standard error of $\overline{X}$ by $s_{\overline{X}} = s/\sqrt{N}$, where s denotes the sample standard deviation.

For an unbiased estimator, some samples will yield estimates that exceed θ and other samples will yield estimates smaller than θ, otherwise θ would not possibly be the "center" of the estimator's distribution. In practice, sometimes "natural" estimators also can "biased".

EXAMPLE 3.1.9. Let $X_1, X_2, \cdots, X_N$ are random sample with $\mathbb{P}(0 \leq X_i \leq \theta) = 1$ for all $i = 1, \cdots, N$. We now want to estimate the unknown parameter θ. It is natural to consider the estimator

$$\hat{\theta}_b := \max_{1 \leq i \leq N} X_i. \tag{3.1.3}$$

However, our proposed estimator $\hat{\theta}_b$ never overestimate θ since the largest sample value cannot exceed the largest population value, and will underestimate θ unless the largest sample value equal to θ (you are lucky). Since

$$\mathbb{P}(\hat{\theta}_b \leq y) = \mathbb{P}(X_1 \leq y, \cdots, X_N \leq y) = \prod_{i=1}^{N} \mathbb{P}(X_i \leq y) = \mathbb{P}(X_1 \leq y)^N,$$

thus we sees that $\mathbb{P}(0 \leq \hat{\theta}_b \leq \theta) = 1$, therefore from Lemma 2.3.18 we see that

$$\mathbb{E}\hat{\theta}_b = \int_0^\theta \overbrace{\mathbb{P}(\hat{\theta}_b > y)}^{\leq 1} \,\mathrm{d}y \leq \theta,$$

which strongly suggests that $\hat{\theta}_b$ is a biased estimator (hence the subscript "b"). If we additionally assume that

$$\lim_{N\to\infty} \mathbb{P}(X_1 \leq y)^N = 0 \text{ for all } y \in (0,\theta),$$

then by using the dominated convergence theorem one sees that

$$\mathbb{E}\hat{\theta}_b = \int_0^\theta \left(1 - \mathbb{P}(X_1 \leq y)^N\right) \mathrm{d}y \to \int_0^\theta 1 \,\mathrm{d}y = \theta \quad \text{as } N \to \infty,$$

the bias approaches 0 as N increases and is negligible for large N. Given any $\varepsilon > 0$, by using the Markov inequality [Dur19, Theorem 1.6.4], one has

$$\mathbb{P}(|\theta - \hat{\theta}_b| \geq \varepsilon) \leq \frac{1}{\varepsilon}\mathbb{E}|\theta - \hat{\theta}_b| = \frac{1}{\varepsilon}\mathbb{E}(\theta - \hat{\theta}_b) \to 0 \quad \text{as } N \to \infty$$

because $\theta - \hat{\theta}_b \geq 0$ a.s., that is, $\hat{\theta}_b \to \theta$ in probability.

The above example strongly suggest the following definition.

DEFINITION 3.1.10. Let $X_1, \cdots, X_N$ be random samples from a distribution that depends on a parameter θ. Then an estimator $\hat{\theta}$ of θ is said to be *consistent* if $\hat{\theta}_b \to \theta$ in probability.

EXAMPLE 3.1.11. Let $\alpha > 0$ and let $X_1, X_2, \cdots, X_N$ are random sample with (the case $\alpha = 1$ corresponding to uniform distribution on $(0,\theta)$)

$$\mathbb{P}(X_i \leq y) = \begin{cases} 0 & ,y \leq 0, \\ y^\alpha/\theta^\alpha & ,0 \leq y \leq \theta, \\ 1 & ,y \geq \theta, \end{cases}$$

then the expectation of (3.1.3) is given by

$$\begin{aligned} \mathbb{E}\hat{\theta}_b &= \int_0^\theta \left(1 - \mathbb{P}(\hat{\theta}_b \leq y)\right) \mathrm{d}y = \int_0^\theta \left(1 - \mathbb{P}(X_1 \leq y)^N\right) \mathrm{d}y \\ &= \int_0^\theta \left(1 - \frac{y^{\alpha N}}{\theta^{\alpha N}}\right) \mathrm{d}y = \frac{\alpha N}{\alpha N + 1}\theta, \end{aligned}$$

which means that the bias

$$\mathbb{E}\hat{\theta}_b - \theta = -\frac{1}{\alpha N + 1}\theta \text{ is negative.}$$

We say that such estimator $\hat{\theta}_b$ is *biased low*, meaning that it systematically underestimates the true value of θ, but it is still consistent (Definition 3.1.10). As we mentioned above, in practice the sample size N is much smaller than the population size, therefore taking $N \to \infty$ may not plausible.

Therefore, we still interested to obtain an unbiased estimator of θ: this can be done by choosing

$$\hat{\theta}_u := \frac{\alpha N + 1}{\alpha N} \max_{1 \le i \le N} X_i,$$

which satisfies $\mathbb{E}\hat{\theta}_u = \theta$, *regardless of the value of θ and the sample size N*, that is, $\hat{\theta}_u$ is an unbiased estimator of θ. This also demonstrates that the unbiased estimator of θ *depends on the distributions of random samples*.

EXERCISE 3.1.12. Let $\hat{\theta}_b$ be the biased estimator in Example 3.1.11. Compute $\operatorname{var}(\hat{\theta}_b)$.

EXAMPLE 3.1.13. Now we want to estimating population variance σ^2 based on a random sample $X_1, \cdots, X_N$ with $\mathbb{E}X_i^2 < \infty$ and $\operatorname{var} X_i = \sigma^2$ for all $i = 1, \cdots, N$. In view of the definition of variance, one may consider the *sample variance* estimator

$$S_b^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \overline{X})^2,$$

where $\overline{X}$ is the sample mean (3.1.2). Since (left as an exercise)

$$\sum_{i=1}^N (X_i - \overline{X})^2 = \sum_{i=1}^N X_i^2 - \frac{1}{N} \left(\sum_{i=1}^N X_i \right)^2 \tag{3.1.4}$$

and $X_1, \cdots, X_N$ are i.i.d., one sees that

$$\begin{aligned}
\mathbb{E}S_b^2 &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}X_i^2 - \frac{1}{N^2} \mathbb{E} \left(\sum_{i=1}^N X_i \right)^2 \\
&= \frac{1}{N} \sum_{i=1}^N \left(\sigma^2 + (\mathbb{E}X_1)^2 \right) - \frac{1}{N^2} \left(\operatorname{var} \left(\sum_{i=1}^N X_i \right) + \left(\mathbb{E} \left(\sum_{i=1}^N X_i \right) \right)^2 \right) \\
&= \sigma^2 + (\mathbb{E}X_1)^2 - \frac{1}{N^2} \left(N\sigma^2 + N^2 (\mathbb{E}X_1)^2 \right) \quad \text{(using Exercise 2.3.30)} \\
&= \frac{N-1}{N} \sigma^2,
\end{aligned}$$

which means that it systematically underestimates the true value of σ^2 since

$$\mathbb{E}S_b^2 - \sigma^2 = -\frac{1}{N}\sigma^2 \text{ is negative.}$$

Therefore, similar to Example 3.1.11, one sees that the estimator

$$S_u^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \overline{X})^2 \tag{3.1.5}$$

is unbiased in the sense of $\mathbb{E}S_u^2 = \sigma^2$, *regardless of the value of σ^2 and the sample size N*. Therefore, we called the estimator (3.1.5) an *unbiased sample variance*.

Suppose that we have a random sample of M random samples $X_1,\cdots,X_M$ from $\mathscr{N}(\mu_1,\sigma_1^2)$ and independent random samples $Y_1,\cdots,Y_N$ from $\mathscr{N}(\mu_2,\sigma_2^2)$. We are now interested to estimate the ratio of variance. In view of Example 3.1.13, it is natural to estimate σ_1^2/σ_2^2 by the consistent estimator $S_{1,u}^2/S_{2,u}^2$ (but may biased), where

$$S_{1,u}^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i-\overline{X})^2 \quad \text{and} \quad S_{2,u}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i-\overline{X})^2.$$

In fact, we have the followings:

THEOREM 3.1.14. *Let $X_1,\cdots,X_N$ be random samples from $\mathscr{N}(\mu,\sigma^2)$. Then $\overline{X} \sim \mathscr{N}(\mu,\sigma/\sqrt{N})$ and $\frac{N-1}{\sigma^2}S_u^2 \sim \chi_{N-1}^2$ are independent.*

Therefore, according to the definition of the F distribution (Definition 2.6.16), we see that

$$\frac{S_{1,u}^2/\sigma_1^2}{S_{2,u}^2/\sigma_2^2} = \frac{\frac{(M-1)S_{1,u}^2/\sigma_1^2}{M-1}}{\frac{(N-1)S_{2,u}^2/\sigma_2^2}{N-1}} \sim \mathscr{F}_{M-1,N-1}.$$

One sees that F distribution may be used to compare the variances from two independent group.

3.1.2. Assessing Estimators: Mean squared error. Another way to measure "how good" is a chosen point estimator $\hat{\theta}$ of a parameter θ is we directly compute its error:

DEFINITION 3.1.15. The *mean squared error* (MSE) of an estimator $\hat{\theta}$ is $\mathbb{E}\left((\hat{\theta}-\theta)^2\right)$.

We see that

$$\begin{aligned}
\mathbb{E}\left((\hat{\theta}-\theta)^2\right) &= \mathbb{E}\left(\left((\hat{\theta}-\mathbb{E}\hat{\theta})+(\mathbb{E}\hat{\theta}+\theta)\right)^2\right) \\
&= \mathbb{E}\left((\hat{\theta}-\mathbb{E}\hat{\theta})^2+(\mathbb{E}\hat{\theta}+\theta)^2+2(\hat{\theta}-\mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta}+\theta)\right) \\
&= \mathbb{E}\left((\hat{\theta}-\mathbb{E}\hat{\theta})^2\right)+(\mathbb{E}\hat{\theta}+\theta)^2 \\
&= \operatorname{var}(\hat{\theta})+(\mathbb{E}\hat{\theta}+\theta)^2.
\end{aligned}$$

In other words,

$$\text{MSE} = \text{variance of estimator} + (\text{bias})^2. \tag{3.1.6}$$

In particular, for any unbiased estimator of θ, its MSE and variance are equal.

EXAMPLE 3.1.16. Let us return to the problem of estimating population variance σ^2 based on a random sample $X_1,\cdots,X_N$ with $\mathbb{E}X_i=\mu$ and $\mathbb{E}X_i=\sigma^2$ as in Example 3.1.13. We now consider an estimate of the form

$$S^2 = c\sum_{i=1}^{N}(X_i-\overline{X})^2,$$

where $c = c(N)$ is a constant. Note that S_b^2 corresponds to $c = \frac{1}{N}$ and S_u^2 corresponds to $c = \frac{1}{N-1}$. As showed in Example 3.1.13, since $X_1, \cdots, X_N$ are i.i.d., one can compute

$$\mathbb{E}\left(S^2\right) = c(N-1)\sigma^2.$$

On the other hand, we also compute

$$\begin{aligned}\operatorname{var}(S^2) &= c^2 \operatorname{var}\left(\sum_{i=1}^N (X_i - \overline{X})^2\right) = c^2 \operatorname{var}\left(\sum_{i=1}^N X_i^2 - \frac{1}{N}\left(\sum_{i=1}^N X_i\right)^2\right) \\ &= c^2 \operatorname{var}\left(\sum_{i=1}^N X_i^2 - \frac{1}{N}\sum_{i=1}^N X_i^2 - \frac{1}{N}\sum_{i\neq j} X_i X_j\right) \\ &= c^2 \operatorname{var}\left(\frac{N-1}{N}\sum_{i=1}^N X_i^2 - \frac{1}{N}\sum_{i\neq j} X_i X_j\right).\end{aligned}$$

The above expression is difficult to compute since the random variables $X_i X_j$ may not uncorrelated (Definition 2.3.27), therefore one cannot use Lemma 2.3.31 to compute $\operatorname{var}(S^2)$. In this case, by using Example 2.6.14, we see that $\operatorname{var}(S^2) = 2c^2\sigma^4(N-1)$. By using (3.1.6), we see that

$$\begin{aligned}\mathbb{E}\left((S^2 - \sigma^2)^2\right) &= 2c^2\sigma^4(N-1) + (c(N-1)-1)^2\sigma^4 \\ &= \left(2c^2(N-1) + (c(N-1)-1)^2\right)\sigma^4.\end{aligned}$$

By differentiating the mapping $c \mapsto 2c^2(N-1) + (c(N-1)-1)^2$, one sees that the choice

$$c = \frac{1}{N+1}$$

minimizes the MSE $\mathbb{E}\left((S^2 - \sigma^2)^2\right)$, which yields a rather unnatural estimator

$$S_{\mathrm{MSE}}^2 = \frac{1}{N+1}\sum_{i=1}^N (X_i - \overline{X})^2,$$

which has bias

$$\mathbb{E}S_{\mathrm{MSE}}^2 - \sigma^2 = -\frac{2}{N+1}\sigma^2 < -\frac{1}{N}\sigma^2 = \mathbb{E}S_b^2 - \sigma^2 \quad \text{when sample size } N \geq 2.$$

In practice, we usually do not use the (unnatural) estimator S_{MSE}^2.

3.1.3. Unbiased estimation. As demonstrated in Example 3.1.16, we would prefer an unbiased rather than the biased one, even if the latter has a smaller MSE. This is sometimes referred as the *principle of unbiased estimation*. If there were a unique unbiased estimator for a parameter, then the situation is simple. However this is not the case in general. Finding an estimator whose mean squared error is smaller than that of every other estimator for all values of the parameter is sometimes not feasible. One common approach is to restrict the class of estimators under consideration in some way, and then seek the estimator that is best in that restricted class.

According to the principle of unbiased estimation, it is common to restrict ourselves on unbiased estimator.

DEFINITION 3.1.17. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If $\text{var}(\hat{\theta}_1) \leq \text{var}(\hat{\theta}_2)$, then we say that $\hat{\theta}_1$ is more *efficient* than $\hat{\theta}_2$.

Among unbiased estimator, it is natural to consider the one with least variance.

DEFINITION 3.1.18. Let $\hat{\theta}$ be an unbiased estimator of θ. If $\text{var}(\hat{\theta}) \leq \text{var}(\tilde{\theta})$ for all unbiased estimators $\tilde{\theta}$ of θ, then we say that $\hat{\theta}$is the *minimum variance unbiased estimator* (later we use the abbreviation “m.v.u.e.”) of θ.

In other words, m.v.u.e. is the accurate estimator with best precision.

EXAMPLE 3.1.19. Let $X_1, X_2, \cdots, X_N$ are random sample with uniform distribution on $(0,\theta)$, that is, its c.d.f. is given by

$$\mathbb{P}(X_i \leq y) = \begin{cases} 0 & ,y \leq 0, \\ y/\theta & ,0 \leq y \leq \theta, \\ 1 & ,y \geq \theta. \end{cases}$$

In Example 3.1.11 (with $\alpha = 1$), we see that

$$\hat{\theta}_u := \frac{N+1}{N} \max_{1 \leq i \leq N} X_i,$$

is an unbiased estimator of θ. However, this is not the unique unbiased estimator of θ. For example, we can easily verify that

$$\hat{\theta} := 2\overline{X} = \frac{2}{N}(X_1 + \cdots + X_N)$$

is also an unbiased estimator of θ, but however,

$$\text{var}(\hat{\theta}) = \frac{4}{N^2}\sum_{i=1}^{N} \text{var}(X_i) = \frac{\theta^2}{3N} > \frac{\theta^2}{N(N+2)} = \text{var}(\hat{\theta}_u),$$

which shows that $\hat{\theta}_u$ is a better estimator in terms of “precision”. In fact, $\hat{\theta}_u$ is the m.v.u.e. (Definition 3.1.18) of θ.

We now list some examples of m.v.u.e.:

THEOREM 3.1.20. *Let $X_1, X_2, \cdots, X_N$ are random sample with $\mathscr{N}(\mu, \sigma^2)$, then*

(1) *the m.v.u.e. of the mean μ is $\overline{X} = \frac{X_1 + \cdots + X_N}{N}$.*
(2) *the m.v.u.e. of variance σ is $S_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$.*

In view of the unbiased estimator S_u^2 of the population variance σ^2, one may estimate the population standard error σ using the estimator

$$S_u = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}.$$

However, in general $\mathbb{E}S_u$ does not equal to σ, i.e. the estimator may biased. In fact[1], if X_i are normal distributions with variance σ^2, then one can check that $K_N S_u$ is an unbiased of σ, in the sense of $\mathbb{E}(K_N S_u) = \sigma$ *regardless of the value of* σ *and the sample size* N, with standard error

$$\sigma_{K_N S_u} = \sqrt{\operatorname{var}(K_N S_u)} = \sigma K_N \sqrt{\frac{V_N}{N-1}},$$

where Γ is the Gamma function,

$$K_N = \sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N}{2})} = \overbrace{\sqrt{\frac{N-1}{2}}\exp\left(\ln\Gamma\left(\frac{N-1}{2}\right) - \ln\Gamma\left(\frac{N}{2}\right)\right)}^{\text{more numerically stable for large } N}$$

and

$$V_N = 2\left(\frac{N-1}{2} - \frac{\Gamma^2(\frac{N}{2})}{\Gamma^2(\frac{N-1}{2})}\right),$$

see [**DBC21**, Exercise 52 in Section 6.4] or [**LC98**, page 92]. In fact:

THEOREM 3.1.21. *Let* $X_1, X_2, \cdots, X_N$ *are random sample with* $\mathcal{N}(\mu, \sigma^2)$*, then the MVUE of standard error* σ *is* $K_N S_u$*.*

The central limit theorem (Theorem 2.6.12) says that $\sigma^{-1}N^{1/2}(\overline{X}_N - \mu)$ would be "close" to the standard normal for large N. Despite the estimator S_u of σ is slightly biased, but it still consistent (Definition 3.1.10), therefore it is natural to replace σ by S_u in the expression $X := \sigma^{-1}N^{1/2}(\overline{X}_N - \mu)$ to obtain the estimator $T := \frac{\overline{X}-\mu}{S_u/\sqrt{N}}$. If $X_1, X_2, \cdots, X_N$ are random sample with $\mathcal{N}(\mu, \sigma^2)$, since

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{(N-1)S_u^2}{\sigma^2} \sim \chi^2_{N-1} \text{ (see Example 2.6.14)},$$

then, by writing

$$\frac{\overline{X}-\mu}{S_u/\sqrt{N}} = \frac{(\overline{X}-\mu)/(\sigma/\sqrt{N})}{\sqrt{(N-1)^{-1}\frac{(N-1)S_u^2}{\sigma^2}}},$$

we see (Example 2.6.15) the following result, which was originally discovered in 1908 by William Sealy Gosset (known as "Student"), a statistician at the Guinness Brewery in Dublin, Ireland:

THEOREM 3.1.22 ([**Gos08**]). *Let* $X_1, X_2, \cdots, X_N$ *are random sample with* $\mathcal{N}(\mu, \sigma^2)$*, then*

$$T := \frac{\overline{X}-\mu}{S_u/\sqrt{N}} \sim t_{N-1}.$$

We have showed in (3.1.2) that $\overline{X}$ is an unbiased estimator of μ, regardless the distribution of each sample. In view of Theorem 3.1.20, it is natural to ask whether $\overline{X}$ is still a MVUE or not if we sample according to other distributions rather than $\mathcal{N}(\mu, \sigma^2)$. In fact, if the random sample comes

[1] https://web.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf

from a Cauchy distribution with p.d.f.

$$f(x) = \frac{1}{\pi(1+(x-\theta)^2)}, \quad x \in \mathbb{R},$$

then in fact $\overline{X}$ is a terrible estimator for μ, since it is very sensitive to outlying observations, and the heavy tails of the Cauchy distribution make a few such observations likely to appear in any sample. Given an integer $1 \leq m < N/2$, the *truncated mean* is defined by

$$\overline{X}_{\mathrm{tr}} := \frac{1}{N-2m+2} \sum_{i=m}^{N-m+1} \max\{X_1, \cdots, X_i\}.$$

Despite $\overline{X}_{\mathrm{tr}}$ is not the best estimator, but it produces (with truncated proportion $\frac{m-1}{N}$ between 10 and 20%) reasonably behaved estimates over a very wide range of possible population models. For this reason, such a truncated mean is said to be a *robust estimator*.

3.2. The method of moments

The estimator we mentioned above is obtained via guessing according to "common sense". We are now interested in some systematic methods to find an estimator. First of all, we introduce the *methods of moments*, which basic idea is to equate certain simple characteristics, such as the sample mean, to the corresponding population expected values. Then solving these equations for unknown parameter values yields the estimators. Let $m \in \mathbb{N}$ and let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) with finite moments $\mathbb{E}|X_i|^m < \infty$ for all $1 \leq k \leq m$. As mentioned above, we see that

$$\frac{1}{N} \sum_{i=1}^{N} X_i^k \tag{3.2.1}$$

is an unbiased estimator of the k^{th} *moment of the distribution* $\mathbb{E}(X_i^k)$, which also called the k^{th} *population moment*. Therefore it is natural to introduce the following definition.

DEFINITION 3.2.1. Let $m \in \mathbb{N}$ and let $X_1, \cdots, X_N$ be random samples with finite moments $\mathbb{E}|X_i|^m < \infty$ for all $1 \leq k \leq m$. For each $k \in \mathbb{N}$, the k^{th} sample moment is the random variable (3.2.1).

Thus, the first population moment is $\mathbb{E}(X_i) = \mu$ and the first sample moment is $\frac{1}{N}\sum_{i=1}^{N} X_i = \overline{X}$. The second population and sample moments are $\mathbb{E}(X_i^2)$ and $\frac{1}{N}\sum_{i=1}^{N} X_i^2$, respectively. The population moments will be functions of any unknown parameters $\theta_1, \theta_2, \cdots$.

DEFINITION 3.2.2. Let $X_1, \cdots, X_N$ be a random sample from a distribution depending on parameters $\theta_1, \cdots, \theta_m$ whose values are unknown. Then the *method of moments estimators* (later we use the abbreviation "m.m.e.") $\hat{\theta}_1, \cdots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \cdots, \theta_m$.

By using the *strong* law of large number (Theorem 2.3.24), we see that for each $k \in \mathbb{N}$ that

$$\frac{1}{N}\sum_{i=1}^{N} X_i^k \to \mathbb{E}(X_i^k) \text{ a.s.}$$

therefore the m.m.e. is always consistent (Definition 3.1.10).

EXAMPLE 3.2.3. Let $X_1, \cdots, X_N$ are random samples from $\mathscr{E}(\lambda) = \text{Gamma}(1, 1/\lambda)$. We first want to estimate the parameter $\theta = 1/\lambda$. Note that

$$\theta = \mathbb{E}X_i.$$

By substituting $\mathbb{E}X_i$ using $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ in the above equation, we see that the m.m.e. of θ is then $\hat{\theta} = \overline{X}$. We see that

$$\mathbb{E}\overline{X} = \mathbb{E}X_i = 1/\lambda = \theta,$$

which shows that the m.m.e. of the parameter $\theta = 1/\lambda$ is an unbiased estimator.

EXAMPLE 3.2.4. Let $X_1, \cdots, X_N$ are random samples from $\mathscr{E}(\lambda) = \text{Gamma}(1, 1/\lambda)$. We now want to estimate the parameter $\theta = \lambda$. Note that

$$\theta = \frac{1}{\mathbb{E}X_i}.$$

By substituting $\mathbb{E}X_i$ using $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ in the above equation, we see that the m.m.e. of θ is then $\hat{\theta} = 1/\overline{X}$. Since the mapping $t \mapsto 1/t$ is strictly convex, then Jensen's inequality shows that

$$\mathbb{E}\hat{\theta} = \mathbb{E}(1/\overline{X}) > 1/\mathbb{E}\overline{X} = \lambda,$$

which shows that the estimator $\hat{\theta}$ is biased high. In fact, by using [**DBC21**, Exercise 13 in Section 7.1] we have

$$\mathbb{E}\hat{\theta} = \frac{N}{N-1}\lambda,$$

which shows that $\hat{\theta}_u = \frac{N-1}{N}\hat{\theta}$ is an unbiased estimator of θ.

Example 3.2.3 and Example 3.2.4 demonstrate that the m.m.e. just suggest some estimators, but may biased.

EXAMPLE 3.2.5. Let $X_1, \cdots, X_N$are random samples with variance $\text{var}(X_i) = \sigma^2$. We now want to estimate the parameter $\theta = \sigma^2$. Note that

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

By substituting $\mathbb{E}X_i$ (resp. $\mathbb{E}X_i^2$) using $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ (resp. $\mathbb{E}X_i^2 \to \frac{1}{N}\sum_{i=1}^{N} X_i^2$) in the above equation, we see that the m.m.e. of θ is

$$\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)^2 \overset{(3.1.4)}{=} \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2 = S_b^2,$$

which is exactly the biased estimator mentioned in Example 3.1.13.

EXAMPLE 3.2.6. Let $X_1, \cdots, X_N$ are random samples from Gamma(α,β). By using Exercise 2.6.22, one sees that

$$\mathbb{E}X_i = \alpha\beta \quad \text{and} \quad \mathbb{E}X_i^2 = \beta^2(\alpha+1)\alpha.$$

We are now want to estimate the parameters $\theta_1 = \alpha$ and $\theta_2 = \beta$. A little straightforward algebra gives

$$\alpha = \frac{(\mathbb{E}X_i)^2}{\mathbb{E}X_i^2 - (\mathbb{E}X_i)^2} \quad \text{and} \quad \beta = \frac{\mathbb{E}X_i^2 - (\mathbb{E}X_i)^2}{\mathbb{E}X_i}.$$

By substituting $\mathbb{E}X_i \to \overline{X} = \frac{1}{N}\sum_{i=1}^N X_i$ and $\mathbb{E}X_i^2 \to \frac{1}{N}\sum_{i=1}^N X_i^2$ in the above equation, we see that the m.m.e. of α and β are (3.1.4)

$$\hat{\alpha} = \frac{\overline{X}^2}{\frac{1}{N}\sum_{i=1}^N X_i^2 - \overline{X}^2} \overset{(3.1.4)}{=} \frac{\overline{X}^2}{S_b^2} \quad \text{and} \quad \hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^N X_i^2 - \overline{X}^2}{\overline{X}} \overset{(3.1.4)}{=} \frac{S_b^2}{\overline{X}}$$

where

$$S_b^2 = \frac{1}{N}\sum_{i=1}^N (X_i - \overline{X})^2 \overset{(3.1.4)}{=} \frac{1}{N}\sum_{i=1}^N X_i^2 - \overline{X}^2$$

is the biased variance estimator in Example 3.1.13. By using the strong law of large number (Theorem 2.3.24), one sees that $\hat{\alpha}$ (resp. $\hat{\beta}$) is a consistent estimator of α (resp. β).

EXAMPLE 3.2.7. Let $X_1, X_2, \cdots, X_N$ are random sample from uniform distribution on $(0,\theta)$, i.e.

$$\mathbb{P}(X_i \le y) = \begin{cases} 0 & ,y \le 0, \\ y/\theta & ,0 \le y \le \theta, \\ 1 & ,y \ge \theta. \end{cases}$$

One sees that $\mathbb{E}X_i = \frac{\theta}{2}$. By substituting $\mathbb{E}X_i$ using $\overline{X} = \frac{1}{N}\sum_{i=1}^N X_i$ in the equation $\theta = 2\mathbb{E}X_i$, we see that the m.m.e. of θ is

$$\hat{\theta} = 2\overline{X},$$

which is an unbiased estimator of θ.

EXAMPLE 3.2.8. Let $X_1, X_2, \cdots, X_N$ are random sample from uniform distribution on $(\theta, \theta+1)$, i.e.

$$\mathbb{P}(X_i \le y) = \begin{cases} 0 & ,y \le \theta, \\ y-\theta & ,\theta \le y \le \theta+1, \\ 1 & ,y \ge \theta+1. \end{cases}$$

One sees that $\mathbb{E}X_i = \theta + \frac{1}{2}$. By substituting $\mathbb{E}X_i$ using $\overline{X} = \frac{1}{N}\sum_{i=1}^N X_i$ in the equation $\theta = \mathbb{E}X_i - \frac{1}{2}$, we see that the m.m.e. of θ is

$$\hat{\theta} = \overline{X} - \frac{1}{2},$$

which is an unbiased estimator of θ.

3.3. The method of maximum likelihood

The method of maximum likelihood was first introduced by R. A. Fisher in 1922 [Fis22]. Most statisticians recommend this method, at least when the sample size is large, since the resulting estimators have certain desirable efficiency properties.

DEFINITION 3.3.1. Let $X_1, \cdots, X_N$ be absolute continuous random variables (resp. discrete random variables) with a joint p.d.f. (resp. joint p.m.f.) that depends on a parameter θ whose value is unknown. This joint distribution, regarded as a function of θ, is called the *likelihood function*, and is denoted by $L(\theta)$. The natural logarithm of the likelihood function $\ell(\theta) := \ln(L(\theta))$ is commonly referred to as the *log-likelihood function*. The *maximum likelihood estimator* (later we use the abbreviation "m.l.e.") $\hat{\theta}$ is the value of θ that maximizes the likelihood function (or log-likelihood function).

Maximizing the likelihood gives the parameter value for which the observed sample is most likely to have been generated, that is, the parameter value that "agrees most closely" with the observed data.

EXAMPLE 3.3.2. Let $X_1, \cdots, X_N$ be random samples from $\mathscr{E}(\lambda)$. Since $X_1, \cdots, X_N$ are i.i.d., then the likelihood (i.e. their joint p.d.f.) is

$$L(\lambda) = \prod_{i=1}^{N} (\lambda e^{-\lambda X_i}) = \lambda^N e^{-\lambda \sum_{i=1}^{N} X_i} = \lambda^N e^{-\lambda N \overline{X}},$$

then the log-likelihood is

$$\ell(\lambda) := \ln L(\lambda) = N \ln \lambda - \lambda N \overline{X}.$$

By solving the equation $\ell'(\lambda) = 0$, one sees that $\hat{\lambda} = 1/\overline{X}$ maximizes ℓ, that is, $\ell(\hat{\lambda}) \geq \ell(\lambda)$ for all $\ell > 0$. In other words, $\hat{\lambda} = 1/\overline{X}$ is the m.l.e. of the parameter λ. Note that such $\hat{\lambda} = 1/\overline{X}$ is exactly the m.m.e. that we found in Example 3.2.4, which is a consistent but biased estimator.

EXAMPLE 3.3.3. Let $X_1, \cdots, X_N$ are random samples from $\mathscr{N}(\mu, \sigma^2)$. Since $X_1, \cdots, X_N$ are i.i.d., then the likelihood (i.e. their joint p.d.f.) is

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} \right) = (2\pi\sigma^2)^{-\frac{N}{2}} e^{-\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{2\sigma^2}}$$

then the log-likelihood is

$$\ell(\mu, \sigma^2) := \ln L(\mu, \sigma^2) = -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2} (\sigma^2)^{-1} \sum_{i=1}^{N} (X_i - \mu)^2.$$

Note that its first partial derivatives are

$$\begin{aligned}\partial_\mu \ell(\mu,\sigma^2) &= -(\sigma^2)^{-1}\sum_{i=1}^N(\mu - X_i) = -(\sigma^2)^{-1}N(\mu-\overline{X}),\\ \partial_{\sigma^2}\ell(\mu,\sigma^2) &= -\frac{N}{2}(\sigma^2)^{-1}+\frac{1}{2}(\sigma^2)^{-2}\sum_{i=1}^N(X_i-\mu)^2 = -\frac{N}{2}(\sigma^2)^{-2}\left(\sigma^2-\frac{1}{N}\sum_{i=1}^N(X_i-\mu)^2\right).\end{aligned}$$

By solving the equation $\partial_\mu \ell(\mu,\sigma^2) = \partial_{\sigma^2}\ell(\mu,\sigma^2) = 0$, we see that

$$(\hat{\mu},\hat{\sigma}^2) = \left(\overline{X},\frac{1}{N}\sum_{i=1}^N(X_i-\overline{X})^2\right) \equiv (\overline{X},S_b) \tag{3.3.1}$$

is the only critical point (Definition 1.5.1) of $\ell : \mathbb{R}\times\mathbb{R}_{>0}\to\mathbb{R}$. We also see that

$$\begin{aligned}\partial_\mu^2\ell(\mu,\sigma^2) &= -(\sigma^2)^{-1}N,\\ \partial_\mu\partial_{\sigma^2}\ell(\mu,\sigma^2) = \partial_{\sigma^2}\partial_\mu\ell(\mu,\sigma^2) &= (\sigma^2)^{-2}N(\mu-\overline{X}),\\ \partial_{\sigma^2}^2\ell(\mu,\sigma^2) &= \frac{N}{2}(\sigma^2)^{-2}-(\sigma^2)^{-3}\sum_{i=1}^N(X_i-\mu)^2 = \frac{N}{2}(\sigma^2)^{-3}\left(\sigma^2-\frac{2}{N}\sum_{i=1}^N(X_i-\mu)^2\right),\end{aligned}$$

then

$$\begin{aligned}\partial_\mu^2\ell(\hat{\mu},\hat{\sigma}^2) &= -(\hat{\sigma}^2)^{-1}N,\\ \partial_\mu\partial_{\sigma^2}\ell(\hat{\mu},\hat{\sigma}^2) &= 0,\\ \partial_{\sigma^2}^2\ell(\mu,\sigma^2) &= \frac{N}{2}(\hat{\sigma}^2)^{-3}\left(\hat{\sigma}^2-2\hat{\sigma}^2\right) = -\frac{N}{2}(\hat{\sigma}^2)^{-2},\end{aligned}$$

which shows that

$$\nabla^{\otimes 2}_{(\mu,\sigma^2)}\ell(\hat{\mu},\hat{\sigma}^2) = \begin{pmatrix} -(\hat{\sigma}^2)^{-1}N & 0 \\ 0 & -\frac{N}{2}(\hat{\sigma}^2)^{-2}\end{pmatrix} \prec 0,$$

therefore the second derivative test (Theorem 1.5.6) shows that (3.3.1) is the local maximizer of $\ell : \mathbb{R}\times\mathbb{R}_{>0}\to\mathbb{R}$. Since

$$\lim_{\sigma^2\to 0_+}\ell(\mu,\sigma^2) = \ln\left(\lim_{\sigma^2\to 0_+}L(\mu,\sigma^2)\right) = -\infty,$$

this shows that the local maximizer (3.3.1) is indeed global, and we conclude that (3.3.1) is the m.l.e. of parameters (μ,σ^2), which is also the m.m.e. of (μ,σ^2), see Example 3.2.5. We point out that $\hat{\mu} = \overline{X}$ is the m.v.u.e. of μ (Theorem 3.1.20), but however $\hat{\sigma}^2 = S_b$ is a biased estimator (Example 3.1.13).

EXAMPLE 3.3.4. Let $X_1, X_2, \cdots, X_N$ are random sample from uniform distribution on $(0,\theta)$, i.e.

$$\mathbb{P}(X_i \leq y) = \begin{cases} 0 & ,y \leq 0, \\ y/\theta & ,0 \leq y \leq \theta, \\ 1 & ,y \geq \theta. \end{cases}$$

One can easily see that the p.d.f. of each X_i is

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

Since $X_1, \cdots, X_N$ are i.i.d., then the likelihood (i.e. their joint p.d.f.) is

$$L(\theta) = \begin{cases} \theta^{-N} & ,0 \leq X_i \leq \theta \text{ for all } i = 1, \cdots, N, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} \theta^{-N} & \theta \geq \max_{i=1,\cdots,N} X_i, \\ 0 & \theta < \max_{i=1,\cdots,N} X_i. \end{cases}$$

Since $L(\theta) > 0$ if and only if $\theta \geq \max_{i=1,\cdots,N} X_i$, then we see that its global maximizer $\hat{\theta}$ (if exists) must satisfies $\hat{\theta} \geq \max_{i=1,\cdots,N} X_i$. Now we see that θ^{-N} is a monotone decreasing function, then we see that

$$\hat{\theta} = \max_{i=1,\cdots,N} X_i$$

is the m.l.e. of the parameter θ, which is exactly the biased estimator $\hat{\theta}_b$ mentioned in Example 3.1.9. In this case, the m.l.e. is different to the m.m.e. (Example 3.2.7).

The following example demonstrates that m.l.e. may not unique.

EXAMPLE 3.3.5. Let $X_1, X_2, \cdots, X_N$ are random sample from uniform distribution on $(\theta, \theta+1)$, i.e.

$$\mathbb{P}(X_i \leq y) = \begin{cases} 0 & ,y \leq \theta, \\ y-\theta & ,\theta \leq y \leq \theta+1, \\ 1 & ,y \geq \theta+1. \end{cases}$$

One can easily see that the p.d.f. if each X_i is

$$f(x) = \begin{cases} 1 & \theta \leq x \leq \theta+1, \\ 0 & \text{otherwise.} \end{cases}$$

Since $X_1, \cdots, X_N$ are i.i.d., then the likelihood (i.e. their joint p.d.f.) is

$$L(\theta) = \begin{cases} 1 & ,\theta \leq X_i \leq \theta+1 \text{ for all } i = 1, \cdots, N, \\ 0 & \text{otherwise,} \end{cases} = \begin{cases} 1 & \max_{i=1,\cdots,N} X_i - 1 \leq \theta \leq \min_{i=1,\cdots,N} X_i, \\ 0 & \text{otherwise.} \end{cases}$$

Then we see any $\hat{\theta}$ with

$$\max_{i=1,\cdots,N} X_i - 1 \leq \hat{\theta} \leq \min_{i=1,\cdots,N} X_i$$

maximizes $L(\theta)$. In other words, for each statistic $0 \leq \alpha = \alpha(X_1, \cdots, X_N) \leq 1$, we see that the convex combination

$$\hat{\theta}_\alpha := \alpha \left(\max_{i=1,\cdots,N} X_i - 1 \right) + (1-\alpha) \min_{i=1,\cdots,N} X_i$$

maximizes $L(\theta)$. In other words, any such of the estimator $\hat{\theta}$ is a m.l.e. of the parameter θ.

3.4. Sufficiency

Given random samples (i.i.d. random variables) $X_1, \cdots, X_N$ and our goal is to make an inference about some parameter θ. As a first point, we note that a statistic $T = t(X_1, \cdots, X_N)$ will not be useful for drawing conclusions about θ unless the distribution of T depends on θ. For example, we consider random samples X_1, X_2 of size $N = 2$ from a normal distribution with mean μ and variance σ^2, and we consider the statistic $T = X_1 - X_2$, which is a normal distribution with mean 0 and variance $2\sigma^2$. Since this statistic does not depends on μ, then it cannot be used as a basis for drawing any conclusion about μ.

We now want to make an inference about some parameter θ based on one or more statistics T, which depend(s) on such parameter θ. Among these statistics, we may expect that some of them contain more information about θ than will others. The main theme of this section is to decide which statics are most informative for making inferences.

DEFINITION 3.4.1. Suppose the joint distribution of $X_1, \cdots, X_N$ involves an unknown parameter θ. A statistic $T = t(X_1, \cdots, X_N)$ is said to be *sufficient* for making inference about a parameter θ if the joint distribution of $X_1, \cdots, X_N$ given that $T = t$ does not depend upon θ, for every possible value t of the statistic T. Such statistic T is called a *sufficient statistic*.

We now exhibit some examples in [**LM21**] to explain the definition.

EXAMPLE 3.4.2 (An estimator that is sufficient). Let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) be taken from the Bernuolli distribution, with p.m.f.

$$p(k;\theta) = \theta^k (1-\theta)^{1-k}, \quad k = 0, 1,$$

where θ is an unknown parameter, which is the mean of X_i. We now consider the statistic $T = \overline{X} := \frac{1}{n} \sum_{i=1}^N X_i$, which is both m.m.e. and m.l.e. of θ. To show that T is a sufficient estimator for θ, it is suffice to compute the conditional probability of $X_1 = k_1, \cdots, X_N = k_N$ given that $\overline{X} = \overline{k} := \frac{1}{N} \sum_{i=1}^N k_i$, because

$$\mathbb{P}(X_1 = k_1, \cdots, X_N = k_N | T = k) = 0 \quad \text{for all } k \neq \overline{k}.$$

First of all, we see that

$$\begin{aligned} &\mathbb{P}\left(X_1 = k_1, \cdots, X_N = k_N | T = \overline{k}\right) \\ &\quad = \frac{\mathbb{P}\left((X_1 = k_1, \cdots, X_N = k_N) \cap (T = \overline{k})\right)}{\mathbb{P}(T = \overline{k})} = \frac{\mathbb{P}(X_1 = k_1, \cdots, X_N = k_N)}{\mathbb{P}(T = \overline{k})}. \end{aligned}$$

We see that

$$\begin{aligned}\mathbb{P}(X_1 = k_1, \cdots, X_N = k_N) &= \prod_{i=1}^{N} \mathbb{P}(X_i = k_i) = \prod_{i=1}^{N} \theta^{k_i}(1-\theta)^{1-k_i} \\ &= \theta^{\sum_{i=1}^{N} k_i}(1-\theta)^{\sum_{i=1}^{N}(1-k_i)} = \theta^{N\bar{k}}(1-\theta)^{N(1-\bar{k})},\end{aligned}$$

and

$$\mathbb{P}(T = \bar{k}) = \mathbb{P}\left(\sum_{i=1}^{N} X_i = N\bar{k}\right) = \binom{N}{N\bar{k}} \theta^{N\bar{k}}(1-\theta)^{N(1-\bar{k})}$$

since $\sum_{i=1}^{N} X_i$ has binomial distribution with parameters N and p (Example 2.5.8). Therefore, we conclude that

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_N = k_N | T = \bar{k}\right) = \binom{N}{N\bar{k}}^{-1},$$

which is *independent* of θ. Therefore, we conclude that the statistic T is *sufficient* for making inference about the unknown parameter θ, in the sense of Definition 3.4.1.

EXAMPLE 3.4.3 (An estimator that is not sufficient). Let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) given in Example 3.4.2. We now consider the statistic $T = \frac{1}{N-1}\sum_{i=1}^{N-1} X_i$. To show that T is a sufficient estimator for θ, it is suffice to compute the conditional probability of $X_1 = k_1, \cdots, X_N = k_N$ given that $T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i$, because

$$\mathbb{P}(X_1 = k_1, \cdots, X_N = k_N | T = k) = 0 \quad \text{for all } k \neq \frac{1}{N-1}\sum_{i=1}^{N-1} k_i.$$

First of all, we compute that

$$\begin{aligned}&\mathbb{P}\left(X_1 = k_1, \cdots, X_N = k_N | T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i\right) \\ &= \frac{\mathbb{P}\left((X_1 = k_1, \cdots, X_N = k_N) \cap (T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i)\right)}{\mathbb{P}(T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i)} = \frac{\mathbb{P}(X_1 = k_1, \cdots, X_N = k_N)}{\mathbb{P}(T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i)}.\end{aligned}$$

We see that

$$\mathbb{P}\left(T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i\right) = \mathbb{P}\left(\sum_{i=1}^{N-1} X_i = \sum_{i=1}^{N-1} k_i\right) = \binom{N-1}{\sum_{i=1}^{N-1} k_i} \theta^{\sum_{i=1}^{N-1} k_i}(1-\theta)^{(N-1)-\sum_{i=1}^{N-1} k_i},$$

which gives

$$\mathbb{P}\left(X_1 = k_1, \cdots, X_N = k_N | T = \frac{1}{N-1}\sum_{i=1}^{N-1} k_i\right) = \binom{N-1}{\sum_{i=1}^{N-1} k_i}^{-1} \theta^{k_N}(1-\theta)^{1-k_N},$$

which *depends* on θ. Therefore, we conclude that the statistic T is *not sufficient* for making inference about the unknown parameter θ, in the sense of Definition 3.4.1.

However, the above arguments only works for random samples taken from discrete probability distributions. For random samples taken from continuous probability distribution, we need the following factorization theorem.

THEOREM 3.4.4 (Fisher-Neyman factorization theorem). *Let $f(x_1,\cdots,x_N;\theta)$ denote the joint p.m.f. or p.d.f. of random samples $X_1,\cdots,X_N$. Then $T = t(X_1,\cdots,X_N)$ is a sufficient statistic for θ if and only if there exist nonnegative functions g and h such that*

$$f(x_1,\cdots,x_N;\theta) = g\left(t(x_1,\cdots,x_N);\theta\right) h(x_1,\cdots,x_N).$$

The proof of the case when $X_1,\cdots,X_N$ are discrete is exactly same as the procedure in Example 3.4.2 and Example 3.4.3, here we shall not repeat the details here. A general proof when $X_1,\cdots,X_N$ are continuous is fraught with technical details that are beyond the level of our text, so lets skip those details here.

REMARK 3.4.5. Let $x_1,\cdots,x_N$ and $y_1,\cdots,y_N$ be any two sets of observations, and let $T = t(X_1,\cdots,X_N)$ is a sufficient statistic for θ. If $t(x_1,\cdots,x_N) = t(y_1,\cdots,y_N)$, then using the Fisher-Neyman factorization theorem (Theorem 3.4.4) we see that the likelihood ratio

$$\frac{f(x_1,\cdots,x_N;\theta)}{f(y_1,\cdots,y_N;\theta)} = \frac{h(x_1,\cdots,x_N)}{h(y_1,\cdots,y_N)}$$

does not depend on θ.

We now introduce the following concept:

DEFINITION 3.4.6. Suppose the joint distribution of $X_1,\cdots,X_N$ involves an unknown parameter θ. A sufficient statistic is said to be *minimal sufficient* for θ if it can be represented as a function of any other sufficient statistic for θ. In other words, the statistic $T_{\min} = t_{\min}(X_1,\cdots,X_N)$ is *minimal sufficient* for θ if and only if the following two conditions hold:

(1) $T_{\min}$ is sufficient for θ, and

(2) if $T = t(X_1,\cdots,X_N)$ is sufficient for θ, then there exists a function f such that $t_{\min} = f \circ t$, which means that

$$t(x_1,\cdots,x_N) = t(y_1,\cdots,y_N) \implies t_{\min}(x_1,\cdots,x_N) = t_{\min}(y_1,\cdots,y_N).$$

Intuitively, a minimal sufficient statistic most effectively captures all possible information about the parameter θ. Here is a result that allows for easy identification of a minimal sufficient statistic:

THEOREM 3.4.7 ([**DBC21**, Exercise 74 in Chapter 7]). *Let $f(x_1,\cdots,x_N;\theta)$ denote the joint p.m.f. or p.d.f. of random variables $X_1,\cdots,X_N$. Suppose that there is a function $t(x_1,\cdots,x_N)$ such that the following holds: for any two sets of observations $x_1,\cdots,x_N$ and $y_1,\cdots,y_N$ the likelihood ratio*

$$\frac{f(x_1,\cdots,x_N;\theta)}{f(y_1,\cdots,y_N;\theta)} \text{ does not depend on } \theta \iff t(x_1,\cdots,x_N) = t(y_1,\cdots,y_N),$$

then $T = t(X_1, \cdots, X_N)$ is a minimal sufficient statistic for θ.

PROOF. We first show that $T = t(X_1, \cdots, X_N)$ is a sufficient statistic for θ. Define an equivalence relation $\sim$ by setting

$$(x_1, \cdots, x_N) \sim (y_1, \cdots, y_N) \iff t(x_1, \cdots, x_N) = t(y_1, \cdots, y_N).$$

Let $\tau \in \operatorname{range}(f)$ and let $(x_1, \cdots, x_N)$ be a set of observation, and suppose that $t(x_1, \cdots, x_N) = \tau$. Then $(x_1, \cdots, x_N)$ is in the equivalence class $\{(y_1, \cdots, y_N) : t(y_1, \cdots, y_N) = \tau\}$, which has a representative

$$\boldsymbol{x}_{t(x_1,\cdots,x_N)} \equiv \boldsymbol{x}_\tau := [x_1, \cdots, x_N].$$

By the hypothesis, the ratio

$$\frac{f(x_1, \cdots, x_N; \theta)}{f(\boldsymbol{x}_{t(x_1,\cdots,x_N)}; \theta)} \text{ does not depend on } \theta,$$

and we define

$$h(x_1, \cdots, x_N) := \frac{f(x_1, \cdots, x_N; \theta)}{f(\boldsymbol{x}_{t(x_1,\cdots,x_N)}; \theta)}.$$

Let $g(t, \theta) := f(\boldsymbol{x}_t; \theta)$, then

$$f(x_1, \cdots, x_N; \theta) = f(\boldsymbol{x}_{t(x_1,\cdots,x_N)}; \theta) h(x_1, \cdots, x_N) = g\left(t(x_1, \cdots, x_N)\right) h(x_1, \cdots, x_N),$$

and by using the Fisher-Neyman factorization theorem (Theorem 3.4.4) so we conclude that $T = t(X_1, \cdots, X_N)$ is a sufficient statistic for θ.

Next, we aim to show that T is minimal sufficient. Suppose that $S = s(X_1, \cdots, X_N)$ is also sufficient for θ, so that, by using the Fisher-Neyman factorization theorem (Theorem 3.4.4), there exist functions g_s and h_s such that

$$f(x_1, \cdots, x_N; \theta) = g_s(s(x_1, \cdots, x_N)) h_s(x_1, \cdots, x_N).$$

Suppose that $s(x_1, \cdots, x_N) = s(y_1, \cdots, y_N)$, then

$$\frac{f(x_1, \cdots, x_N; \theta)}{f(y_1, \cdots, y_N; \theta)} = \frac{h_s(x_1, \cdots, x_N)}{h_s(y_1, \cdots, y_N)} \text{ does not depend on } \theta,$$

and this implies that $t(x_1, \cdots, x_N) = t(y_1, \cdots, y_N)$ by hypothesis. In other words, we showed that

$$s(x_1, \cdots, x_N) = s(y_1, \cdots, y_N) \implies t(x_1, \cdots, x_N) = t(y_1, \cdots, y_N),$$

which means that t is a function of s. Hence we conclude that $T = t(X_1, \cdots, X_N)$ is a minimal sufficient statistic for θ. □

The following example demonstrates the standard argument to argue a statistic which is sufficient by using the Fisher-Neyman factorization theorem (Theorem 3.4.4).

EXAMPLE 3.4.8 (A m.l.e. that is minimal sufficient). Let $X_1,\cdots,X_N$ be random samples (i.i.d. random variables) drawn from the uniform distribution on $[0,\theta]$, i.e. the p.d.f.

$$f(x;\theta)=\frac{1}{\theta},\quad 0\le x\le\theta,$$

where θ is an unknown parameter. We have showed in Example 3.3.4 that the m.l.e. of θ is the statistic $T_{\min}\equiv t_{\min}(X_1,\cdots,X_N)=\max\{X_1,\cdots,X_N\}$. The joint p.d.f. of $X_1,\cdots,X_N$ is given by

$$f(x_1,\cdots,x_N;\theta)=\frac{1}{\theta^N}\quad ,0\le x_1\le\theta,\cdots,0\le x_N\le\theta.$$

To obtain the desired factorization, we introduce notation for an indicator function: $I(A)=1$ if the statement A is true, and $I(A)=0$ otherwise. By using this notations, we write

$$\begin{aligned}f(x_1,\cdots,x_N;\theta)&=\frac{1}{\theta^N}I\left(0\le x_1\le\theta,\cdots,0\le x_N\le\theta\right)\\&=\frac{1}{\theta^N}I\left(0\le\min\{x_1,\cdots,x_N\}\text{ and }\max\{x_1,\cdots,x_N\}\le\theta\right)\\&=\overbrace{\left(\frac{1}{\theta^N}I\left(\max\{x_1,\cdots,x_N\}\le\theta\right)\right)}^{=:g(t(x_1,\cdots,x_N))}\overbrace{I\left(0\le\min\{x_1,\cdots,x_N\}\right)}^{=:h(x_1,\cdots,x_N)}.\end{aligned}$$

By using the Fisher-Neyman factorization theorem (Theorem 3.4.4), we conclude that the statistic T is *sufficient* for making inference about the unknown parameter θ, in the sense of Definition 3.4.1. Finally, by using Theorem 3.4.7, we see that T is also minimal sufficient for θ.

The following example demonstrates that “natural” estimators may not be sufficient. However, it is not easy to study a statistic, which is not sufficient, directly using the Fisher-Neyman factorization theorem (Theorem 3.4.4). In many cases, the concept of minimal sufficient statistic (Definition 3.4.6) is very helpful, as showed in the following example.

EXAMPLE 3.4.9 (A m.m.e. that is not sufficient). Let $X_1,\cdots,X_N$ be random samples (i.i.d. random variables) be given in Example 3.4.8. We have showed in Example 3.2.7 that the m.m.e. of θ is the statistic

$$T=2\overline{X}:=\frac{2}{N}\sum_{i=1}^{N}X_i\equiv t(X_1,\cdots,X_N),$$

which is an unbiased estimator of θ. Let $t_{\min}$ be the function given in Example 3.4.8. We now want to show that T is not sufficient for making inference about the unknown parameter θ. Otherwise, since

$$t\left(\frac{\theta}{2},\frac{\theta}{2},0,\cdots,0\right)=\frac{2\theta}{N}=t\left(\theta,0,0,\cdots,0\right),$$

but

$$t_{\min}\left(\frac{\theta}{2},\frac{\theta}{2},0,\cdots,0\right)=\frac{\theta}{2}\neq\theta=t_{\min}\left(\theta,0,0,\cdots,0\right),$$

which contradicts with the fact that $T_{\min} \equiv t_{\min}(X_1, \cdots, X_N) = \max\{X_1, \cdots, X_N\}$ is a minimal sufficient statistic for θ.

EXAMPLE 3.4.10 (A m.l.e. that is not sufficient). Let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) drawn from the uniform distribution on $[\theta, \theta+1]$, i.e. the p.d.f.

$$f(x;\theta) = 1, \quad \theta \leq x \leq \theta+1,$$

where θ is an unknown parameter. We have showed in Example 3.3.5 *sufficient* that, each statistic $0 \leq \alpha = \alpha(X_1, \cdots, X_N) \leq 1$, the statistic

$$T_\alpha := \alpha \left(\max_{i=1,\cdots,N} X_i - 1 \right) + (1-\alpha) \min_{i=1,\cdots,N} X_i$$

is a m.l.e. of θ. The joint p.d.f. of $X_1, \cdots, X_N$ is given by

$$f(x_1, \cdots, x_N; \theta) = 1 \quad , \theta \leq x_1 \leq \theta+1, \cdots, \theta \leq x_N \leq \theta+1,$$

that is,

$$\begin{aligned} f(x_1, \cdots, x_N; \theta) &= I(\theta \leq x_1 \leq \theta+1, \cdots, \theta \leq x_N \leq \theta+1) \\ &= I(\theta \leq \min\{x_1, \cdots, x_N\} \text{ and } \max\{x_1, \cdots, x_N\} \leq \theta+1) \\ &= I(\theta \leq \min\{x_1, \cdots, x_N\}) I(\max\{x_1, \cdots, x_N\} \leq \theta+1), \end{aligned} \tag{3.4.1}$$

which is not possible to be expressed only through t_α, which shows that any of the statistics T_α above is not sufficient.

It is interesting to mention the following theorem regarding the sufficiency of the m.l.e., which still strongly suggests us to consider m.l.e. in practical applications.

THEOREM 3.4.11 ([Moo71]). *If there exists a* unique *m.l.e. $\hat{\theta}$ of θ, then $\hat{\theta}$ is a minimal sufficient statistic for θ.*

In view of the p.d.f. in (3.4.1), we see that the parameter θ can be inferred, given *both* of the following statistics:

$$\begin{aligned} T_1 &= t_1(X_1, \cdots, X_N) = \min\{X_1, \cdots, X_N\}, \\ T_2 &= t_2(X_1, \cdots, X_N) = \max\{X_1, \cdots, X_N\}. \end{aligned}$$

This strongly suggests the following definition.

DEFINITION 3.4.12. Suppose the joint distribution of $X_1, \cdots, X_N$ involves m unknown parameters $\theta_1, \cdots, \theta_m$. The k-dimensional statistic

$$\mathbf{T} = (T_1, \cdots, T_k) = (t_1(X_1, \cdots, X_N), \cdots, t_k(X_1, \cdots, X_N)) = \mathbf{t}(X_1, \cdots, X_N)$$

is said to be (jointly) *sufficient* for making inference about the parameters $\theta_1, \cdots, \theta_m$ if the conditional distribution of the X_i's given that $\mathbf{T} = (t_1, \cdots, t_k)$ does not depend on any of the unknown parameters, and this is true for all possible values $t_1, \cdots, t_k$ of the statistic.

THEOREM 3.4.13. *The Fisher-Neyman factorization theorem (Theorem 3.4.4) also holds true for k-dimensional statistic.*

EXAMPLE 3.4.14. Let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) drawn from the uniform distribution on $[\theta, \theta + 1]$ as in Example 3.4.10. The Fisher-Neyman factorization theorem (Theorem 3.4.13) shows that the 2-dimensional statistic

$$\mathbf{T} = (\min\{X_1, \cdots, X_N\}, \max\{X_1, \cdots, X_N\})$$

is sufficient for making inference about the parameter θ.

EXERCISE 3.4.15. Let $X_1, \cdots, X_N$ be random sample from normal distribution $\mathscr{N}(\mu, \sigma^2)$. Show that the 2-dimensional statistic

$$\mathbf{T} = \left(\sum_{i=1}^{N} X_i, \sum_{i=1}^{N} X_i^2 \right)$$

is sufficient for making inference about the parameters μ and σ^2.

An estimator (or any function) of a parameter θ should depend on the data only through the sufficient statistic. A general result due to C.R. Rao and D. Blackwell shows that how to use an unbiased statistic (this is not sufficient) to create an estimator that is both unbiased and sufficient.

THEOREM 3.4.16 (Rao-Blackwell Theorem). *Suppose that the joint distribution of random variables $X_1, \cdots, X_N$ depends on some unknown parameter θ, and that T is sufficient for θ. If U is an unbiased estimator of $h(\theta)$, where h is a given function, then the estimator $U^* := \mathbb{E}(U|T)$ is also an unbiased estimator of $h(\theta)$ and has variance no greater than the original unbiased estimator U.*

PROOF. By the conditional expectation formula (2.4.1), we have

$$\mathbb{E}U = \mathbb{E}\left(\mathbb{E}(U|T)\right) = \mathbb{E}U^*,$$

which shows that $U^* := \mathbb{E}(U|T)$ is also an unbiased estimator of $h(\theta)$. In the other hand, by using the conditional variance formula (2.4.2), we have

$$\mathrm{var}\,(U) = \mathbb{E}\left(\mathrm{var}\,(U|T)\right) + \mathrm{var}\,\left(\mathbb{E}(U|T)\right) = \mathbb{E}\left(\mathrm{var}\,(U|T)\right) + \mathrm{var}\,(U^*) \geq \mathrm{var}\,(U^*),$$

which conclude our theorem. □

REMARK 3.4.17. Since T is sufficient for θ, the joint distribution of $X_1, \cdots, X_N$ given $T = t$ does not depend on θ, hence U^* does not depend on the unknown parameter θ, and so is a bona fide estimator. If U is already a function of T, then $U = U^*$.

EXAMPLE 3.4.18. Let $X_1, \cdots, X_N$ be random samples (i.i.d. random variables) be given in Example 3.3.4. We have showed in Example 3.4.8 that the m.l.e. of θ is the statistic $T_{\min} \equiv t_{\min}(X_1, \cdots, X_N) = \max\{X_1, \cdots, X_N\}$, which is a minimal sufficient statistic for θ. We have showed in Example 3.4.9 that the m.m.e. of θ is the statistic

$$\hat{\theta} = 2\overline{X} := \frac{2}{N}\sum_{i=1}^{N} X_i,$$

which is an unbiased estimator of θ. We also showed in Example 3.4.9 that $\hat{\theta}$ is not a sufficient statistic for θ, and thus it is not a function of $T_{\min}$. By using the Rao-Backwell theorem (Theorem 3.4.16), we see that $\hat{\theta}^* = \mathbb{E}(\hat{\theta}|T)$ is also an unbiased estimator of θ. We compute that

$$\begin{aligned}\mathbb{E}(\hat{\theta}|T=t) &= \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left(X_i|\max\{X_1,\cdots,X_N\}=t\right) \\ &= \frac{2}{N}\sum_{i=1}^{N}\left(\overbrace{\mathbb{E}\left(X_i|X_i=t\right)}^{=t}\overbrace{\mathbb{P}\left(X_i=\max\{X_1,\cdots,X_N\}\right)}^{=\frac{1}{N}}\right. \\ &\qquad \left. +\overbrace{\mathbb{E}\left(X_i|X_i<t\right)}^{=t/2}\overbrace{\mathbb{P}\left(X_i<\max\{X_1,\cdots,X_N\}\right)}^{=\frac{N-1}{N}}\right) \\ &= \frac{N+1}{N}t \quad \text{for all } t\in(0,\theta),\end{aligned}$$

that is,

$$\hat{\theta}^* = \mathbb{E}(\hat{\theta}|T) = \frac{N+1}{N}T = \frac{N+1}{N}\max\{X_1,\cdots,X_N\},$$

which is an unbiased estimator of θ given in Example 3.1.19, and most important, it is independent of the unknown parameter θ. The probability $\mathbb{P}\left(X_i = \max\{X_1, \cdots, X_N\}\right)$ means that the probability that the i^{th} sample is the largest among all others, therefore it has probability $\frac{1}{N}$. It is not difficult to verify that $\mathbb{E}\left(X_i|X_i < t\right)$ is just the expectation of the uniform distribution on $(0,t)$. It is not easy to compute the variance of $\hat{\theta}^*$, but the Rao-Backwell theorem (Theorem 3.4.16) guarantees that it is bounded from above by

$$\operatorname{var}(\hat{\theta}) = \frac{4}{N^2}\sum_{i=1}^{N}\operatorname{var}(X_i) = \frac{1}{3N}\theta^2.$$

CHAPTER 4

Statistical Intervals based on a single set of samples

Since each point estimator is just a single number, it provides no information about the precision and reliability of estimation. For example, the sample mean $\overline{X}$ is a point estimator of the mean μ, which says nothing about how close it might be to μ. An alternative to report a single sensible value for an unknown parameter θ being estimated is to calculate and report an entire interval of plausible values, or a *confidence interval*. A confidence interval is calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval. The higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval. Information about the precision of an interval estimate is conveyed by the width of the interval:

- If the confidence level is high and the resulting interval is quite narrow, our knowledge of the value of the parameter is reasonably precise.
- A very wide confidence interval, however, gives the message that there is a great deal of uncertainty concerning the value of what we are estimating.

4.1. Basic concept of confidence intervals

We first introduce the concept of confidence intervals by first focusing on a simple but unrealistic problem situation: Suppose that the parameter of interest is a population mean μ and that

(1) the population distribution is normal; and
(2) the value of the population standard deviation σ is known.

In some cases, the population normality is still reasonable, which can be checked by examined a normal probability plot of the sample data $X_1, \cdots, X_N$ of large sample size N. However, if the value of μ is unknown, it is unlikely that the value of σ would be available. However, since we assumed that the samples are dawned from normal distribution, then σ can be estimated by its MVUE (Theorem 3.1.21):

$$\hat{\sigma}(X_1, \cdots, X_N) := K_N S_u = K_N \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})^2}, \tag{4.1.1}$$

where

$$K_N = \sqrt{\frac{N-1}{2}}\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N}{2})} = \overbrace{\sqrt{\frac{N-1}{2}}\exp\left(\ln\Gamma\left(\frac{N-1}{2}\right) - \ln\Gamma\left(\frac{N}{2}\right)\right)}^{\text{more numerically stable for large } N}.$$

For simplicity, we can simply estimate σ by using the *sample standard deviation*

$$s(X_1,\cdots,X_N) := \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}, \tag{4.1.2}$$

which is slightly biased but consistent as the sample size N increases, if the bias is not of primary concern. In later sections, we will introduce some methods based on less restrictive assumptions (especially for samples drawn from distribution which is not normal).

Let $X_1,\cdots,X_N \overset{\text{iid}}{\sim} \mathscr{N}(\mu,\sigma^2)$, we have showed in Exercise 2.6.13 that $\overline{X} \sim \mathscr{N}(\mu,\frac{\sigma^2}{N})$ and

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}} \sim \mathscr{N}(0,1). \tag{4.1.3}$$

Let $\alpha \in (0,1)$ be a given parameter (usually small), let $z_{\alpha/2}$ be the unique number (known as the *two-sided z-critical value*) such that

$$\mathbb{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right) = \mathbb{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha. \tag{4.1.4}$$

EXERCISE 4.1.1. Show that $z_{\alpha/2} := \Phi^{-1}(1 - \frac{\alpha}{2})$, where Φ is the c.d.f. (Definition 2.3.14) of $\mathscr{N}(0,1)$.

This strongly suggests the following definition:

DEFINITION 4.1.2. Let $x_1,\cdots,x_N$ be actual sample observations drawn from i.i.d. normal distribution $\mathscr{N}(\mu,\sigma^2)$ and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ *confidence interval* for the mean μ of a normal population *when the value of* σ *is known* is given by

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right). \tag{4.1.5}$$

The (two-sided) z critical values for the most commonly used confidence levels are displayed in Table 1.

Confidence level (%)	α	$\alpha/2$	approximation of $z_{\alpha/2}$
90	0.10	0.050	1.645
95	0.05	0.025	1.960
99	0.01	0.005	2.576

TABLE 1. Some approximations of $z_{\alpha/2} := \Phi^{-1}(1 - \frac{\alpha}{2})$ for 90, 95 and 99% confidence

In the case when σ is unknown, in view of Theorem 3.1.21, it is make sense to replace it by

$$\hat{\sigma}(x_1,\cdots,x_N) := K_N\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2},$$

or simply

$$s(x_1,\cdots,x_N) := \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2},$$

which is slightly biased, if the bias is not of primary concern. In view of (4.1.3), we now consider the random variable

$$T = \frac{\overline{X}-\mu}{s(X_1,\cdots,X_N)/\sqrt{N}} \sim t_{N-1}, \quad \text{(Gosset's theorem, Example 2.6.15)} \tag{4.1.6}$$

where s is the sample standard deviation. Let $\alpha \in (0,1)$ be a given parameter (usually small), similar to (4.1.4), let $t_{\alpha/2,N-1}$ be the unique number (known as the *two-sided t-critical value*) such that

$$\begin{aligned}&\mathbb{P}\left(\overline{X}-t_{\alpha/2,N-1}\frac{s(X_1,\cdots,X_N)}{\sqrt{N}} < \mu < \overline{X}+t_{\alpha/2,N-1}\frac{s(X_1,\cdots,X_N)}{\sqrt{N}}\right)\\ &= \mathbb{P}\left(-t_{\alpha/2,N-1} < T < t_{\alpha/2,N-1}\right) = 1-\alpha.\end{aligned}$$

EXERCISE 4.1.3. Show that $t_{\alpha/2,N-1} := F^{-1}(1-\frac{\alpha}{2})$, where F is the c.d.f. (Definition 2.3.14) of t_{N-1}.

We then introduce a definition similar to Definition 4.1.2:

DEFINITION 4.1.4. Let $x_1,\cdots,x_N$ be actual sample observations drawn from i.i.d. normal distribution with mean μ and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ *t-confidence interval* for the mean μ of a normal population is given by

$$\left(\bar{x}-t_{\alpha/2,N-1}\frac{s(x_1,\cdots,x_N)}{\sqrt{N}}, \bar{x}+t_{\alpha/2,N-1}\frac{s(x_1,\cdots,x_N)}{\sqrt{N}}\right).$$

It is interesting to mention that the t-distribution with $N-1$ degree of freedom t_{N-1} has a symmetric, bell-shaped density curve centered at 0 that is wider than a standard normal $\mathscr{N}(0,1)$ that is wider than a standard normal curve but converges to the standard normal curve as $N\to\infty$ (so that the standard normal $\mathscr{N}(0,1)$ may be formally understood as a "t-distribution with ∞ degree of freedom"). There is a nice GeoGebra project (`https://www.geogebra.org/m/y3UPKHuH`), not only plot the t-distribution with $N-1$ degree of freedom t_{N-1} and the standard normal $\mathscr{N}(0,1)$, it also compute both two-sided z-critical value $z_{\alpha/2}$ and two-sided t-critical value $t_{\alpha/2,N-1}$.

However, the higher the desired degree of confidence, the wider the resulting interval, in other words, the gain in reliability entails a loss in precision: In fact, the only 100% confidence interval for μ is $(-\infty,+\infty)$, which is not informative since we knew that this interval covers μ even before sampling. Not only the "balance" between reliability and precision, another question is how many samples which we need? We first specify both the desired confidence level (equivalently, the z-critical value $z_{\alpha/2}$) and interval width at most w, then from (4.1.5) we have $w \geq 2z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$, that is,

$$N \geq \left(\frac{z_{\alpha/2}\sigma}{w/2}\right)^2. \tag{4.1.7}$$

This suggests us that an appealing strategy is to specify both the desired confidence level (equivalently, the z-critical value $z_{\alpha/2}$) and interval width w, and then determine the necessary sample size based on the formula (4.1.7). When σ is not known, it might seem like the natural update to the formula (4.1.7) is

$$N \geq \left(\frac{t_{\alpha/2,N-1}s(x_1,\cdots,x_N)}{w/2}\right)^2. \tag{4.1.8}$$

However, this formula presents two practical problems. First, sample size determination typically occurs before a study is carried out, in which case the researcher does not yet have a value for $s(x_1,\cdots,x_N)$. Second, N now appears on both sides of (4.1.8): we need to know N before finding the two-sided t-critical value, which then determines the sample size N on the left-hand side of (4.1.8). Therefore, it is not easy to find the minimal sample size.

Before summarize the ideas, we now introduce the following concept:

DEFINITION 4.1.5. Let $X_1,\cdots,X_N$ are random samples drawn from a (continuous) probability distribution depends on an unknown parameter θ. Suppose that there exists a random variable Z, which is a function of $X_1,\cdots,X_N,\theta$, such that its distribution is independent of θ (but may depend on other unknown parameters), then such a random variable is called a *pivotal quantity*.

EXAMPLE 4.1.6. The random variable Z in (4.1.3) is an example of pivotal quantity: it is a function of $X_1,\cdots,X_N,\mu,\sigma$, but its distribution is $\mathcal{N}(0,1)$, which is independent of the target parameter μ. The random variable T in (4.1.6) is also an example of pivotal quantity: it is a function of $X_1,\cdots,X_N,\mu$, but its distribution is t_{N-1}, which is independent of the target parameter μ.

Let $h(X_1,\cdots,X_N,\theta)$ denote a general pivotal quantity. For any $\alpha \in (0,1)$. constants a and b can be found (but the pair may not unique) to satisfy

$$\mathbb{P}\left(a < h(X_1,\cdots,X_N,\theta) < b\right) = 1-\alpha. \tag{4.1.9}$$

Since the distribution of $h(X_1,\cdots,X_N,\theta)$ does not depend on θ, then the choice of a and b is independent of θ.

EXAMPLE 4.1.7. In the normal example above, $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$.

Suppose that we can write (4.1.9) as

$$\mathbb{P}\left(\ell(X_1,\cdots,X_N) < \theta < u(X_1,\cdots,X_N)\right) = 1-\alpha.$$

EXAMPLE 4.1.8. In the normal example above,

$$\ell(X_1,\cdots,X_N) = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \quad u(X_1,\cdots,X_N) = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \quad \text{(if } \sigma \text{ is known)},$$

$$\begin{cases} \ell(X_1,\cdots,X_N) = \overline{X} - t_{\alpha/2,N-1}\dfrac{s(X_1,\cdots,X_N)}{\sqrt{N}} \\ u(X_1,\cdots,X_N) = \overline{X} + t_{\alpha/2,N-1}\dfrac{s(X_1,\cdots,X_N)}{\sqrt{N}} \end{cases} \quad \text{(if } \sigma \text{ is not known)}.$$

We now ready summarize the basic concept which we want to deliver.

DEFINITION 4.1.9. Let $x_1,\cdots,x_N$ be actual sample observations, called the *realizations* of random samples $X_1,\cdots,X_N$, and let $\alpha \in (0,1)$. Then the interval

$$(\ell(x_1,\cdots,x_N), u(x_1,\cdots,x_N))$$

is called a $100(1-\alpha)\%$ *confidence interval* for the unknown parameter θ.

In view of the central limit theorem (Theorem 2.6.12), we may consider the case when $X_1,\cdots,X_N$ be random samples from any population having a mean μ and standard deviation σ (square root of variance), provided that the second moment is finite, because

$$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$$

has approximately a normal distribution $\mathscr{N}(0,1)$, when the sample size N is extremely large (but itself may not a pivotal quantity in the sense of Definition 4.1.5). Here we remind the readers that the assumption of extremely large sample size is not realistic in practical application. Let $x_1,\cdots,x_N$ are realizations of random samples $X_1,\cdots,X_N$, and let $\alpha \in (0,1)$. An argument parallel with that given earlier in this section yields (4.1.5) as a large-sample confidence interval for μ with a confidence level of approximately $100(1-\alpha)\%$. In this case, we also can approximate σ by its MVUE $\hat{\sigma}(x_1,\cdots,x_N)$ or sample standard deviation $s(x_1,\cdots,x_N)$.

The confidence intervals discussed thus far give both a lower confidence bound and an upper confidence bound for the parameter being estimated. In some circumsta\sigma^{2}nces, and investigator will want only one of these two types of bounds. In general an upper confidence bound for a parameter θ with confidence level $100(1-\alpha)\%$ based on a random sample $X_1,\cdots,X_N$ is a quantity $u(X_!,\cdots,X_N)$ such that

$$\mathbb{P}(\theta < u(X_1,\cdots,X_N)) = 1-\alpha, \tag{4.1.10}$$

which corresponding to the choice $a = -\infty$ in (4.1.9). In the normal example, by using Exercise 4.1.1,

$$u(X_1,\cdots,X_N) = \overline{X} + z_\alpha \frac{\sigma}{\sqrt{N}}.$$

Similar, a lower confidence bound $\ell(X_1,\cdots,X_N)$ satisfies

$$\mathbb{P}\left(\ell(X_1,\cdots,X_N) < \theta\right) = 1-\alpha, \tag{4.1.11}$$

which corresponding to the choice $b = +\infty$ in (4.1.9). In the normal example, by using Exercise 4.1.1,

$$\ell(X_1,\cdots,X_N) = \overline{X} + z_\alpha \frac{\sigma}{\sqrt{N}}.$$

We summarize the above in the following definition:

DEFINITION 4.1.10 (see also Definition 4.1.2). Let $x_1,\cdots,x_N$ be actual sample observations drawn from i.i.d. normal distribution $\mathscr{N}(\mu,\sigma^2)$ and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ *lower and upper confidence interval* for the mean μ of a normal population *when the value of σ is known* is given by

$$\left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{N}}, \infty\right) \quad \text{and} \quad \left(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{N}}\right)$$

respectively.

If the standard deviation σ is unknown, similarly, by replacing Exercise 4.1.1 by Exercise 4.1.3, the following definition is also natural:

DEFINITION 4.1.11 (see also Definition 4.1.4). Let $x_1,\cdots,x_N$ be actual sample observations drawn from i.i.d. normal distribution $\mathscr{N}(\mu,\sigma^2)$ and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ *lower and upper t-confidence interval* for the mean μ of a normal population *when the value of σ is known* is given by

$$\left(\bar{x} - t_{\alpha,N-1} \frac{s(x_1,\cdots,x_N)}{\sqrt{N}}, \infty\right). \quad \text{and} \quad \left(-\infty, \bar{x} + t_{\alpha,N-1} \frac{s(x_1,\cdots,x_N)}{\sqrt{N}}\right)$$

respectively.

4.2. Basic concept of prediction intervals

We wish to predict a single value of variable to be observed at some future time. For example, we have available a set of random samples $X_1,X_2,\cdots,X_N$ from a normal population distribution $\mathscr{N}(\mu,\sigma^2)$, says, and we wish to predict the value of X_{N+1}, a single future observation. A natural way to do this is to consider the sample mean $\overline{X} = \frac{1}{N}\sum_{i=1}^N X_i$ as a point predictor of X_{N+1}, and the resulting prediction error is $\overline{X} - X_{N+1}$, with expectation

$$\mathbb{E}(\overline{X} - X_{N+1}) = \mathbb{E}\overline{X} - \mathbb{E}X_{N+1} = \mu - \mu = 0.$$

Since X_{N+1} is independent of $X_1,\cdots,X_N$, then it is independent of $\overline{X}$, so using Exercise 2.3.30 we see that the variance of the prediction error is

$$\operatorname{var}(\overline{X} - X_{N+1}) = \operatorname{var}(\overline{X}) + \operatorname{var}(X_{N+1}) = \frac{\sigma^2}{N} + \sigma^2 = \sigma^2\left(1 + \frac{1}{N}\right).$$

The prediction error is a linear combination of independent normally distributed random variables, so it self is normally distributed. Thus

$$Z = \frac{\overline{X} - X_{N+1} - \mathbb{E}(\overline{X} - X_{N+1})}{\sqrt{\operatorname{var}(\overline{X} - X_{N+1})}} = \frac{\overline{X} - X_{N+1}}{\sigma\sqrt{(1+\frac{1}{N})}} \sim \mathcal{N}(0,1).$$

In general, the parameter σ is unknown, as above, it can be estimated by its MVUE (4.1.1), or simply the sample standard deviation (4.1.2) if the bias is not of primary concern:

$$T = \frac{\overline{X} - X_{N+1}}{s\sqrt{1+\frac{1}{N}}}. \tag{4.2.1}$$

EXERCISE 4.2.1. Show that the random variable T given in (4.2.1) obeys the distribution t_{N-1}. (**Hint.** Modifying the ideas in (4.1.6))

This leads the following definition:

DEFINITION 4.2.2. Let $x_1, \cdots, x_N$ be actual sample observations drawn from i.i.d. normal distribution with mean μ and let $\alpha \in (0,1)$. A $100(1-\alpha)\%$ *prediction interval* for a single observation to be selected from a normal population distribution is

$$\left(\overline{x} - t_{\alpha/2,N-1}s(x_1,\cdots,x_N)\sqrt{1+\frac{1}{N}}, \overline{x} + t_{\alpha/2,N-1}s(x_1,\cdots,x_N)\sqrt{1+\frac{1}{N}}\right).$$

It is worth contrasting the behavior of the t-confidence interval (Definition 4.1.4) with the prediction interval (Definition 4.2.2). The prediction interval (Definition 4.2.2) is wider than the t-confidence interval (Definition 4.1.4). It is also interesting to see that , as N gets arbitrarily large, the t-confidence interval (Definition 4.1.4) shrinks to the single value μ:

$$\bigcap_{N\in\mathbb{N}} \left(\overline{x} - t_{\alpha/2,N-1}\frac{s(x_1,\cdots,x_N)}{\sqrt{N}}, \overline{x} + t_{\alpha/2,N-1}\frac{s(x_1,\cdots,x_N)}{\sqrt{N}}\right) = \{\mu\},$$

while the prediction interval (Definition 4.2.2) approaches to an interval with nonempty interior:

$$\begin{aligned}&\bigcap_{N\in\mathbb{N}} \left(\overline{x} - t_{\alpha/2,N-1}s(x_1,\cdots,x_N)\sqrt{1+\frac{1}{N}}, \overline{x} + t_{\alpha/2,N-1}s(x_1,\cdots,x_N)\sqrt{1+\frac{1}{N}}\right)\\ &\quad = \left[\overline{x} - z_{\alpha/2}\sigma, \overline{x} + z_{\alpha/2}\sigma\right],\end{aligned}$$

which covers the middle $100(1-\alpha)\%$ of a normal distribution $\mathcal{N}(\mu,\sigma^2)$. This demonstrates that there is uncertainty about a single future value even when there is no need to estimate any parameters. Here we also introduce a branch of statistics, called the *uncertainty quantification*, which is the science of quantitative characterization and estimation of uncertainties in both computational and real world applications. It tries to determine how likely certain outcomes are if some aspects of the system are not exactly known.

The t-confidence interval (Definition 4.1.4) for mean μ is robust to small sample size N (or even moderate departures from normality). However, if the sample size N is small and the population distribution is highly non-normal, they the *actual confidence level* may be considerably different from the one we think we get from using a particular t critical value. We will later discuss the bootstrap technique, which has been found to be quite sucessful at estimating parameters in a wide variety of non-normal situations.

In contrast to the confidence interval, the validity of the prediction interval (Definition 4.2.2) is closely tied to the normality assumption. The prediction interval (Definition 4.2.2) should not be used in the absence of compelling evidence for normality.

4.3. Bootstrap confidence intervals

As mentioned in the very beginning of Chapter 3, in practice we always expect that the sample size N is small. We now interested to construct/estimate a confidence interval for a statistic (for example, mean) if the population distribution is unknown (which is not normal in general) and the sample size N is small. The bootstrap, developed by Bradley Efron in the later 1970s [**Efr79**], facilitates calculating estimates in situation where statistical theory does not produce a formula for a confidence interval. In this section we are concerned with the case of an unknown distribution, for which the *nonparametric bootstrap* is appropriate.

Traditional inference (e.g. the presentations in Section 4.1 and Section 4.2) relies on the sampling distribution of a statistic. In contrast, the (nonparametric) bootstrap method considers what would happen if we were to draw repeatedly from the sample at hand. According to [**Efr79**], the (basic) bootstrap method for a set of sample is extremely simple, at least in principle:

Algorithm 1 Basic bootstrap method (see also Figure 4.3.1)

```
Require: Observed samples $x_1, x_2, \cdots, x_N$, which are realizations of random samples (random variables) $X_1, X_2, \cdots, X_N$.
for $b = 1, \cdots, B$ do
    Construct the sample probability distribution $\hat{U}$ by putting mass $1/N$ at each point $x_1, x_2, \cdots, x_N$.
    With $\hat{U}$ fixed, draw a random sample of size $N$ from $\hat{F}$, say $x_1^*, x_2^*, \cdots, x_N^*$.    % resample
    Compute the value of the statistic from the bootstrap sample $x_1^*, x_2^*, \cdots, x_N^*$, and label the resulting value $\hat{\theta}_b^*$
end for
return The values $\mu_1^*, \mu_2^*, \cdots, \mu_B^*$, which approximate the bootstrap distribution of $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_N)$.
```

EXAMPLE 4.3.1. In the case when we consider the sample mean $\hat{\theta} = \overline{X}$, the number $\hat{\theta}_b^*$ is simply the average of the bootstrap sample $x_1^*, x_2^*, \cdots, x_N^*$.

REMARK. In each iteration, we are not getting a permutation distribution distribution since the values of x_j^* are selected *with replacement* from the set $\{x_1, x_2, \cdots, x_N\}$. Obviously, for that to make sense, bootstrap sampling must occur with replacement, otherwise, we would get the same sample over and over again.

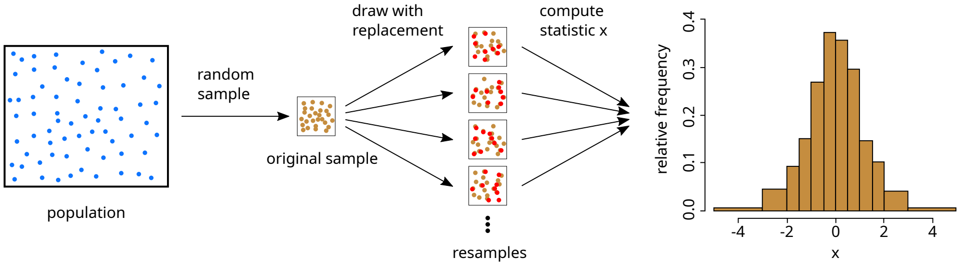

FIGURE 4.3.1. Graphical explaination of Algorithm 1 (Biggerj1, Marsupilami, CC BY-SA 4.0, via Wikimedia Commons)

In practice, $B = 1000$ is often used. In view of sample standard deviation (4.1.2), we define the *bootstrap standard error* to be the sample standard deviation of $\hat{\theta}_b^*$:

$$s_{\text{boot}} := \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \overline{\theta}^*)},$$

where $\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$ is the average of the bootstrap values of the statistic $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_N)$. Once we have the bootstrap distribution of a statistic, several different methods can be used to obtain a confidence interval for the corresponding parameter. For example, in view of the t-confidence interval (Definition 4.1.4):

DEFINITION 4.3.2. Let $x_1, \cdots, x_N$ be actual sample observations drawn from i.i.d. normal distribution with mean μ and let $\alpha \in (0,1)$. A bootstrap $100(1-\alpha)\%$ *t-confidence interval* for the mean μ is given by

$$\left(\overline{x} - t_{\alpha/2,N-1}s_{\text{boot}}, \overline{x} + t_{\alpha/2,N-1}s_{\text{boot}}\right).$$

The bootstrap t confidence interval is appropriate when the bootstrap distribution of the statistic is approximately normal and the bias of the bootstrap distribution is small.

A great advantage of bootstrap is its simplicity. Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality [**DE96**]. Bootstrapping is also a convenient method that avoids the cost of repeating the experiment to get other groups of sample data. According to the original developer of the bootstrapping method [**ERT01**], even setting the number of samples at 50 is likely to lead to fairly good standard error estimates.

However, bootstrapping depends heavily on the estimator used and, though simple, naive use of bootstrapping will not always yield asymptotically valid results and can lead to inconsistency [**Hin94**]. Although bootstrapping is (under some conditions) asymptotically consistent, it does not provide general finite-sample guarantees. The result may depend on the representative sample. Therefore it is recommended to obtain more bootstrap samples as available computing power has increased. Some recommendation on the bootstrap procedure also can be found in [**AMH08**]. However, Athreya has shown in [**Ath87**] that if one performs a naive bootstrap on the sample mean when the underlying population lacks a finite variance (for example, a power law distribution), then the bootstrap distribution will not converge to the same limit as the sample mean. As a result, confidence intervals on the basis of a Monte Carlo simulation of the bootstrap could be misleading. Athreya states that "Unless one is reasonably sure that the underlying distribution is not heavy tailed, one should hesitate to use the naive bootstrap".

Finally, we end this section by referring the monograph [**ET93**] for more details about this topic.

4.4. Understanding the concept of Tests of Hypotheses

A parameter can be estimated from sample data either by a single number (i.e. a point estimator in Chapter 3) or an entire interval of plausible values (i.e. a confidence interval mentioned above). In this section, rather than estimate a parameter, we are now interested in the problem to decide which of two contradictory claims about the parameter is correct, called the *hypothesis testing*. We will discuss some of the basic concepts in hypothesis testing and then introduce some decision procedures for the most frequently encountered testing problems.

DEFINITION 4.4.1. A (statistical) *hypothesis* is a claim of assertion either about the value of a single/multiple parameter(s), or about the form of an entire probability distribution.

In this section, we will only concentrate on hypotheses about a single/multiple parameter(s). In any hypothesis testing problem, there are two contradictory hypotheses under consideration. The objective is to decide, based on sample information, which of the two hypotheses is correct. In statistics, hypothesis testing problems are formulated so that one of the claims is initially assumed to be true. This initial claim will not be rejected in favor of the alternative claim unless sample evidence provides strong evidence for the latter.

DEFINITION 4.4.2. The *null hypothesis* H_0 is the claim that is initially assumed to be true (the "prior belief" claim). The alternative hypothesis, denoted by H_1, is the assertion that is contradictory to H_0.

The null hypothesis will be rejected only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the plausibility of the null hypothesis. The two possible conclusions from a hypothesis testing analysis are then[1]

$$\text{reject } H_0 \quad \text{or} \quad \textit{fail to reject } H_0. \tag{4.4.1}$$

The word "null" means "of no value, effect, or consequence", which suggests that H_0 should be identified with the hypothesis of no change/no difference (from current opinion). In statistical hypothesis testing there are two potential errors whose consequences must be considered when reaching a conclusion:

- A *type I error* consists of rejecting the null hypothesis H_0 when it is true.
- A *type II error* involves not rejection H_0 when it is false (i.e. H_1 is true).

EXAMPLE 4.4.3. The *presumption of innocence* is a legal principle that every person accused of any crime is considered innocent until proven guilty. It is also an international human right under the UN's Universal Declaration of Human Rights, Article 11. Under the presumption of innocence, the legal burden of proof is thus on the prosecution, which must present compelling evidence to the trier of fact (a judge or a jury). If the prosecution does not prove the charges true, then the person is acquitted of the charges. The prosecution must prove that the accused is guilty *beyond* a reasonable doubt, otherwise the accused must be acquitted (i.e. let the accused go free) despite the presence of reasonable doubt. This can be formulated in terms of hypothesis testing problems (from the trier of fact point of view):

$$H_0 : \text{the accused is innocence}, \quad H_1 : \text{the accused in guilty}. \tag{4.4.2}$$

In this case,

- a type I error is convicting an innocent person, while
- a type II error is a false acquittal (i.e. letting a guilty person go free).

The opposite system is a presumption of guilt (contradict to UN's Universal Declaration of Human Rights, Article 11), which can be formulated in terms of hypothesis testing problems:

$$H_0 : \text{the accused in guilty}, \quad H_1 : \text{the accused is innocence}. \tag{4.4.3}$$

This example also illustrate that one should not interchanging H_0 and H_1, which may lead very different outcomes.

We first consider the case when H_0 is stated as an equality claim. If θ denotes the parameter of interest, the null hypothesis will have the form

$$H_0 : \theta \le \theta_0, \quad H_0 : \theta \ge \theta_0 \quad \text{or} \quad H_0 : \theta = \theta_0, \tag{4.4.4}$$

[1] Some authors use the term "accept" rather than "fail to reject" while stating a hypothesis testing problem (4.4.1), which is misleading in my opinion.

where θ_0 is a specified number called the *null value* of parameter (i.e. the value claimed for θ by the null hypothesis). The alternative to the null hypothesis (4.4.4) will look like:

$$H_1 : \theta > \theta_1, \quad H_1 : \theta < \theta_0 \quad \text{or} \quad H_1 : \theta \neq \theta_0,$$

respectively. A test procedure is a rule, based on sample data, for deciding whether to reject H_0 or not:

DEFINITION 4.4.4. A test procedure is specified by the following:

(1) A *test statistic*, a function of the sample data on which the decision is to be based
(2) A *rejection region*, the set of all test statistic values for which H_0 will be rejected.

The null hypothesis will then be rejected if and only if the observed or computed test statistic value falls in the rejection region.

EXAMPLE 4.4.5. Let's take a look on the samples in Figure 4.4.1.

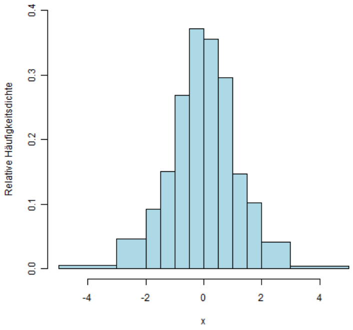

FIGURE 4.4.1. Histogram of 1000 samples (MM-Stat, CC BY-SA 3.0, via Wikimedia Commons)

Do you believe that the random samples are collected according to a certain distribution with mean zero, even though the average of the samples are not zero? If yes, this means that we set the null hypothesis (the "prior belief" claim)

$$H_0 : \mu = 0.$$

In this case, the alternative is $H_1 : \mu \neq 0$. In this case, the test statistic is the average of the samples, and the rejection region is $\mathbb{R} \setminus (-\varepsilon, \varepsilon)$ for some pre-chosen $\varepsilon > 0$.

Suppose a study and a sample size are fixed and a test statistic is chosen. We write

$$\alpha = \mathbb{P}(\text{type I error}) = \mathbb{P}(H_0 \text{ is rejected when it is true}),$$

and

$$\begin{aligned}\beta(\theta') &= \mathbb{P}(\text{type II error occur when the `true' value is } = \theta') \\ &= \mathbb{P}(H_0 \text{ is not rejected when the `true' value is } = \theta')\end{aligned}$$

Note that we consider the same definition for β for all three cases in (4.4.4). We see that decreasing the size of the rejection region to obtain a smaller value of α results in a larger value of β for any particular parameter value consistent with H_1, and vice versa. This says that once the test statistic and the sample size N are fixed, there is no rejection region that will simultaneously make both α and all β's small. A region must be chosen to effect a compromise between α and β. The approach adhered to by most statistical practitioners is to specify the largest value of α that can be tolerated and find a rejection region having that value of α. This makes β as small as possible subject to the bound on α.

DEFINITION 4.4.6. The resulting value of α is often referred to as the *significance level* of the test. The corresponding test procedure is called a *level α test*. A test with significance level α is one for which the type I error probability is controlled at the specified level.

4.5. Tests about a population mean

Similar as in the introduction of confidence intervals for a population mean μ (Section 4.1 and Section 4.2), we first consider the unrealistic scenario when:

(1) the population distribution is normal, and
(2) the value of the population standard deviation σ is known.

Let $X_1, \cdots, X_N$ represent a set of random samples of size N from the normal population with standard deviation $\sigma_{\overline{X}} = \sigma$, then the sample mean $\overline{X}$ has a normal distribution with standard deviation $\sigma_{\overline{X}} = \sigma/\sqrt{N}$. The null hypothesis is

$$H_0 : \mu \leq \mu_0, \quad H_0 : \mu \geq \mu_0 \quad \text{or} \quad H_0 : \mu = \mu_0 \quad (\text{so } \mu_0 \text{ is the null value of the parameter}),$$

which means that, we prior believe that the mean of the population is μ_0. In practice, we usually choose μ_0 which is "somehow close" to the average of the samples $x_1, \cdots, x_N$ (i.e. realizations of $X_1, \cdots, X_N$). Consider now the statistic Z obtained by standardizing $\overline{X}$ under the assumption that H_0 is true:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{N}}.$$

We first consider the hypothesis testing problem

$$H_0 : \mu \leq \mu_0, \quad H_1 : \mu > \mu_0.$$

An $\bar{x}$ value less than μ_0, which corresponds to a negative value of z, certainly does not provide support for H_1. If an $\bar{x}$ value exceeds μ_0, then we need to divide the discuss into two cases:

(1) An $\bar{x}$ value that exceeds μ_0 by only a small amount, in the sense that the corresponding z value is positive but small, does not suggest that H_0 should be rejected.
(2) The rejection of H_0 is appropriate only when $\bar{x}$ considerably exceeds μ_0, in the sense that the corresponding z value is positive and large.

In summary, the appropriate rejection region has the form $z \geq c$ for some relatively large positive constant c. As discussed in previous section, the cutoff value c should be chosen to control the probability of a type I error at the desired level α. This can be easily done in this case because the statistic Z when H_0 is true is the standard normal distribution:

$$\begin{aligned}\alpha &= \mathbb{P}(\text{type I error}) = \mathbb{P}(H_0 \text{ is rejected when it is true})\\ &= \mathbb{P}(Z \geq c \text{ when } Z \sim \mathscr{N}(0,1)) = 1 - \Phi(c),\end{aligned}$$

where Φ is the c.d.f. (Definition 2.3.14) of $\mathscr{N}(0,1)$. This shows that $c = \Phi^{-1}(1-\alpha) = z_\alpha$ (recall Exercise 4.1.1), that is, if a level α test is desired, then H_0 should be rejected if $z \geq z_\alpha$. For example:

- If a level .01 test is desired, then H_0 should be rejected if $z \geq z_{.01} \approx 2.33$.
- If a level .10 test is desired, then H_0 should be rejected if $z \geq z_{.10} \approx 1.28$.

This test procedure is *upper-tailed* because the rejection region consists only large values of the test statistic. In this case, we reject H_0 if and only if μ_0 falls outside a $100(1-\alpha)\%$ lower confidence interval (Definition 4.1.10) for μ.

Analogous reasoning for the hypothesis testing problem

$$H_0 : \mu \geq \mu_0, \quad H_1 : \mu < \mu_0.$$

If a level α test is desired, then H_0 should be rejected if $z \leq z_\alpha^{\text{low}} \equiv -z_\alpha^{\text{up}}$. This is a *lower tailed* test. In this case, we reject H_0 if and only if μ_0 falls outside a $100(1-\alpha)\%$ upper confidence interval (Definition 4.1.10) for μ.

We now consider the hypothesis testing problem

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0.$$

In this case, H_0 should be rejected if the sample mean $\bar{x}$ is too far to either side of μ_0. This is equivalent to rejecting H_0 if $|z| \geq c$:

$$\begin{aligned}\alpha &= \mathbb{P}(\text{type I error}) = \mathbb{P}(H_0 \text{ is rejected when it is true})\\ &= \mathbb{P}(|Z| \geq c \text{ when } Z \sim \mathscr{N}(0,1)) = 2(1 - \Phi(c)).\end{aligned}$$

Solving the equation yields $c = \Phi^{-1}(1-\frac{\alpha}{2}) = z_{\frac{\alpha}{2}}$ (recall Exercise 4.1.1), that is if a level α test is desired, then H_0 should be rejected if $|z| \geq z_{\frac{\alpha}{2}}$. Some approximation of $z_{\frac{\alpha}{2}}$ is shown in Table 1 above. In this case, we reject H_0 if and only if μ_0 falls outside a $100(1-\alpha)\%$ two tailed confidence interval (Definition 4.1.2) for μ. We summarize the above discussions in the following table:

testing	Test statistic value	Rejection region for level α test
$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$	$z \geq z_\alpha$ (upper tailed test)
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$	$z \leq -z_\alpha$ (lower tailed test)
$H_0 : \mu = \mu_0$ $H_1 : \mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{N}}$	$\lvert z \rvert \geq z_{\frac{\alpha}{2}}$ (two tailed test)

TABLE 2. z-test based on a set of random samples

In the z-test based on a set of random samples, there are simple formulas available for the probability β of a type II error.

EXERCISE 4.5.1. Proof $\Phi^{-1}(\beta) = -z_\beta$ for all $0 < \beta < 1$.

Consider first the upper tailed test with rejection region $z \geq z_\alpha$, i.e. $H_0 : \mu \leq \mu_0$. In this case,

$$H_0 \text{ will not be rejected } \iff \bar{x} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{N}}.$$

Now let μ' be any particular value of the parameter μ that exceeds the null value μ_0 (i.e. the value which we belief in prior). Then

$$\begin{aligned}
\beta(\mu') &= \mathbb{P}(H_0 \text{ is not rejected when the `true' value is } = \mu') \\
&= \mathbb{P}\left(\overline{X} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{N}} \text{ when the `true' value is } = \mu'\right) \\
&= \mathbb{P}\left(\frac{\overline{X} - \mu'}{\sigma/\sqrt{N}} < z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}} \text{ when the `true' value is } = \mu'\right) = \Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right).
\end{aligned}$$

If we consider two restrictions $\mathbb{P}(\text{type I error}) = \alpha$ and $\beta(\mu') \leq \beta$ for specified parameters α, μ' and β, then the sample size N should be chosen to satisfy

$$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) \leq \beta \iff z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}} \leq \Phi^{-1}(\beta) = -z_\beta \text{ (Exercise 4.5.1)}$$

which can be guaranteed by

$$N \geq \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu' - \mu_0}\right)^2. \tag{4.5.1}$$

The *power* (defined as the probability that the test procedure will reject null hypothesis H_0) of the upper tailed z-test based on a set of random samples is then $1 - \beta(\mu')$. As the 'true' value μ'

increases, $\mu_0 - \mu'$ becomes more negative, so $\beta(\mu')$ will be small and power will be large when μ' greatly exceeds μ_0.

Consider next the lower tailed test with rejection region $z \leq -z_\alpha$, i.e. $H_0 : \mu \geq \mu_0$. In this case,

$$H_0 \text{ will not be rejected } \iff \bar{x} > \mu_0 - z_\alpha \frac{\sigma}{\sqrt{N}}.$$

Now let μ' be any particular value of the parameter μ that less the null value μ_0 (i.e. the value which we belief in prior). Then

$$\begin{aligned}\beta(\mu') &= \mathbb{P}(H_0 \text{ is not rejected when the ‘true’ value is } = \mu') \\ &= \mathbb{P}\left(\overline{X} > \mu_0 - z_\alpha \frac{\sigma}{\sqrt{N}} \text{ when the ‘true’ value is } = \mu'\right) \\ &= 1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right).\end{aligned}$$

If we consider two restrictions $\mathbb{P}(\text{type I error}) = \alpha$ and $\beta(\mu') \leq \beta$ for specified parameters α, μ' and β, then the sample size N should be chosen to satisfy

$$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) \leq \beta \iff -z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}} \geq \Phi^{-1}(1-\beta) = z_\beta \text{ (Exercise 4.1.1)}$$

which can be guaranteed by (4.5.1). The *power* (defined as the probability that the test procedure will reject null hypothesis H_0) of the upper tailed z-test based on a set of random samples is then $1 - \beta(\mu')$. As the ‘true’ value μ' decreasing, $\mu_0 - \mu'$ becomes more positive, so $\beta(\mu')$ will be small and power will be large when μ' greatly lesser that μ_0.

Consider also the two tailed test with rejection region $|z| \geq z_{\frac{\alpha}{2}}$, i.e. $H_0 : \mu = \mu_0$. In this case,

$$H_0 \text{ will not be rejected } \iff |\bar{x} - \mu_0| < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}}.$$

Now let μ' be any particular value of the parameter μ that not equal to the null value μ_0 (i.e. the value which we belief in prior). Then

$$\begin{aligned}\beta(\mu') &= \mathbb{P}(H_0 \text{ is not rejected when the ‘true’ value is } = \mu') \\ &= \mathbb{P}\left(|\overline{X} - \mu_0| < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \text{ when the ‘true’ value is } = \mu'\right) \\ &= \mathbb{P}\left(\overline{X} - \mu_0 < z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \text{ when the ‘true’ value is } = \mu'\right) \\ &\quad - \mathbb{P}\left(\overline{X} - \mu_0 \leq -z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{N}} \text{ when the ‘true’ value is } = \mu'\right) \\ &= \Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) - \Phi\left(-z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right).\end{aligned}$$

If we consider two restrictions $\mathbb{P}(\text{type I error}) = \alpha$ and $\beta(\mu') \leq \beta$ for specified parameters α, μ' and β, then the sample size N should be chosen to satisfy

$$\Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) - \Phi\left(-z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) \leq \beta.$$

For simplicity, we consider a slightly stronger condition:

$$\Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) \leq \beta, \text{ which can be guaranteed by } N \geq \left(\frac{\sigma(z_{\frac{\alpha}{2}} + z_\beta)}{\mu' - \mu_0}\right)^2.$$

We now summarize the above in the following table:

testing	$\beta(\mu')$	Sufficient for $\beta(\mu') \leq \beta$
$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right)$	$N \geq \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu' - \mu_0}\right)^2$
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right)$	$N \geq \left(\frac{\sigma(z_\alpha + z_\beta)}{\mu' - \mu_0}\right)^2$
$H_0 : \mu = \mu_0$ $H_1 : \mu = \mu_0$	$\Phi\left(z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right) - \Phi\left(-z_{\frac{\alpha}{2}} + \frac{\mu_0 - \mu'}{\sigma/\sqrt{N}}\right)$	$N \geq \left(\frac{\sigma(z_{\frac{\alpha}{2}} + z_\beta)}{\mu' - \mu_0}\right)^2$

TABLE 3. Type II error probability $\beta(\mu')$ for a level α z-test based on a set of random samples

We now modify the z-test based on a set of random samples to accommodate the more realistic situation when σ is unknown, following a path similar to what was outlined in Section 4.1. Consider the test statistic obtained by replacing σ by the sample standard deviation $s(x_1, \cdots, x_N)$ similar as in (4.1.6):

$$T = \frac{\overline{X} - \mu_0}{s(X_1, \cdots, X_N)/\sqrt{N}} \sim t_{N-1}. \quad \text{(Gosset's theorem, Example 2.6.15)}$$

The rejection region for the t test differs from that of the z test only in that a t-critical value $t_{\alpha,N-1}$ replaces the z-critical value z_α:

testing	Test statistic value	Rejection region for level α test
$H_0 : \mu \leq \mu_0$ $H_1 : \mu > \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s(x_1, \cdots, x_N)/\sqrt{N}}$	$t \geq t_{\alpha, N-1}$ (upper tailed test)
$H_0 : \mu \geq \mu_0$ $H_1 : \mu < \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s(x_1, \cdots, x_N)/\sqrt{N}}$	$t \leq -t_{\alpha, N-1}$ (lower tailed test)
$H_0 : \mu = \mu_0$ $H_1 : \mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s(x_1, \cdots, x_N)/\sqrt{N}}$	$\lvert t \rvert \geq t_{\frac{\alpha}{2}, N-1}$ (two tailed test)

TABLE 4. t-test based on a set of random samples (see Table 2)

When the sample size is large, power and sample size calculations (as in Table 3) for the t-test based on a set of random samples can be approximated by the formulas provided in Table 3. In this case, a plausible value of σ must be specified; the sample standard deviation $s(x_1, \cdots, x_N)$ may be used for this purpose. However, exact calculations of power and the type II error probability $\beta(\mu')$ are much less straightforward. This is because the test statistic

$$T = \frac{\overline{X} - \mu_0}{s(X_1, \cdots, X_N)/\sqrt{N}} \tag{4.5.2}$$

does not have a t distribution when H_0 is false. This says that, when the true value of μ is anything other than μ_0, T has a much more complicated distribution, related to the following definition.

DEFINITION 4.5.2. Let $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_\nu$ (Exercise 2.6.14) be independent random variables. For any real number δ, the random variable

$$\frac{Z + \delta}{\sqrt{Y/\nu}}$$

is said to have a *noncentral t-distribution* with ν degrees of freedom and *noncentrality parameter* δ. Note that when $\delta = 0$, this random variable has a t_ν-distribution (Example 2.6.15).

The p.d.f. for the noncentral t-distribution with ν degrees of freedom and noncentrality parameter δ can be expressed as [**Sch91**, page 177]:

$$f(x; \nu, \delta) = \frac{\nu^{\frac{\nu}{2}} \exp(-\frac{\nu\delta^2}{2(x^2+\nu)})}{\sqrt{\pi}\Gamma(\frac{\nu}{2})2^{\frac{\nu-1}{2}}(x^2+\nu)^{\frac{\nu+1}{2}}} \int_0^\infty y^\nu \exp\left(-\frac{1}{2}\left(y - \frac{\delta x}{\sqrt{x^2+\nu}}\right)^2\right) \mathrm{d}y.$$

EXERCISE 4.5.3. Show that the test statistic T has a noncentral t-distribution with $N - 1$ degree of freedom and noncentrality parameter

$$\delta = \frac{\mu' - \mu_0}{\sigma/N}$$

when $\mathbb{E}X_i = \mu'$.

Let $F(x;\nu,\delta)$ denote the c.d.f. of the noncentral t-distribution mentioned above. The power (defined as the probability that the test procedure will reject null hypothesis H_0) of the lower tailed t-test based on a set of random samples is

$$\begin{aligned}&\mathbb{P}(T \le -t_{\alpha,N-1} \text{ when the `true' value is } \mu = \mu')\\ &\quad = \mathbb{P}(T \le -t_{\alpha,N-1} \text{ when } T \text{ has the distribution given in Exercise 4.5.3})\\ &\quad = F\left(-t_{\alpha,N-1}; N-1, \frac{\mu'-\mu_0}{\sigma/N}\right).\end{aligned}$$

Similarly, the power of the upper tailed t-test based on a set of random samples is

$$\begin{aligned}&\mathbb{P}(T \ge t_{\alpha,N-1} \text{ when the `true' value is } \mu = \mu')\\ &\quad = \mathbb{P}(T \ge t_{\alpha,N-1} \text{ when } T \text{ has the distribution given in Exercise 4.5.3})\\ &\quad = 1 - F\left(t_{\alpha,N-1}; N-1, \frac{\mu'-\mu_0}{\sigma/N}\right),\end{aligned}$$

and the power of the two tailed t-test based on a set of random samples is

$$\begin{aligned}&\mathbb{P}(|T| \ge t_{\frac{\alpha}{2},N-1} \text{ when the `true' value is } \mu = \mu')\\ &\quad = \mathbb{P}(|T| \ge t_{\frac{\alpha}{2},N-1} \text{ when } T \text{ has the distribution given in Exercise 4.5.3})\\ &\quad = 1 - F\left(t_{\frac{\alpha}{2},N-1}; N-1, \frac{\mu'-\mu_0}{\sigma/N}\right) + F\left(-t_{\frac{\alpha}{2},N-1}; N-1, \frac{\mu'-\mu_0}{\sigma/N}\right).\end{aligned}$$

Here we remind the readers that the noncentral t-distribution is *not symmetric*.

As we noted in Section 4.1, the t distributions are “robust” against violations of normality when the sample size N is reasonably large. That is, when using data from a large sample, the results of applying the t-test based on a set of random samples should be reasonably accurate even if the underlying population is not normal. We have also seen that, for large sample size N, the z and t_{N-1} distributions are quite similar, so that using a z distribution to determine rejection region cutoffs gives very similar results to the t-test procedure. In current practice, researchers typically use the t test even for large samples.

The one situation in which inferences for μ cannot be based on a t procedure is when the sample size is small and the data strongly suggests a non-normal population. In this case, it is possible to use the bootstrap technique as in Section 4.3 for testing hypotheses about an unknown parameter (here, the mean μ). The fundamental bootstrap concepts (Algorithm 1) may carry over to the hypothesis testing situation:

(1) First, a sample of data $x_1, \cdots, x_N$ is obtained.

(2) To approximate the sampling distribution of a statistic (here, $\overline{X}$), many resamples of size N are randomly selected with replacement from $x_1, \cdots, x_N$, and the statistic of interest (here, the average) is calculated for each resample.

(3) We bootstrap (i.e. repeat the above procedure) for B times, and we obtain B values of the statistic of interest (here, the resample means $\overline{x}_1^*, \overline{x}_2^*, \cdots, \overline{x}_B^*$).

(4) We then approximate the distribution of $\overline{X}$ by the bootstrap distribution (here, the distribution of the $\overline{x}_1^*, \overline{x}_2^*, \cdots, \overline{x}_B^*$), and inferences about the population mean μ can then be made.

However, this approach has a minor shortcoming: the mean of the original samples $x_1, \cdots, x_N$ do not equal μ_0. This indicates that the algorithm fails to satisfy a fundamental principle of hypothesis testing: making decisions based on the distribution of $\overline{X}$ *under the assumption that the null hypothesis* $H_0 : \mu = \mu_0$ *is true*. To address this issue, the bootstrap procedure must be adjusted as follows:

Algorithm 2 Adjusted bootstrap method

```
Require: Observed samples $x_1, x_2, \cdots, x_N$, which are realizations of random samples (random
    variables) $X_1, X_2, \cdots, X_N$.
Require: Specify a null value $\mu_0$                    % We first guess the population mean is $\mu_0$
Compute $\{w_i\}_{i=1}^N$ according to the formula $w_i = x_i - \overline{x} + \mu_0$, where $\overline{x}$ is the average of
    $x_1, x_2, \cdots, x_N$.
for $b = 1, \cdots, B$ do
    Construct the sample probability distribution $\hat{U}$ by putting mass $1/N$ at each point
        $w_1, w_2, \cdots, w_N$.
    With $\hat{U}$ fixed, draw a random sample of size $N$ from $\hat{F}$, say $w_1^*, w_2^*, \cdots, w_N^*$.    %
        resample
    Compute the average of the bootstrap sample $w_1^*, w_2^*, \cdots, w_N^*$, and label the resulting value
        $\overline{w}_b^*$.
end for
return The values $\overline{w}_1^*, \overline{w}_2^*, \cdots, \overline{w}_B^*$, which approximate the bootstrap distribution of $\overline{X}$.
```

Now the average of the original samples $w_1, \cdots, w_N$ is μ_0. This guarantees that the sample probability distribution $\hat{U}$ has expectation μ_0 and then the resulting resample means $\overline{w}_1^*, \overline{w}_2^*, \cdots, \overline{w}_B^*$ provide a semblance of what the distribution of $\overline{X}$ would look like if the null hypothesis $H_0 : \mu = \mu_0$ is true.

4.6. p-value

Using the rejection region method to test hypothesis entails first selecting a significance level α. Then after computing the value of the test statistic, the null hypothesis H_0 is rejected if the value falls in the rejection region and is otherwise not rejected. We now consider another way of reaching a conclusion in a hypothesis-testing analysis. This alternative approach is based on calculation of a certain probability called a p-value. The p-value is the probability of obtaining "extreme" observed sample result assuming the null hypothesis ("prior belief") is true. A small p-value suggests that such an extreme outcome is unlikely under the null hypothesis. If such an outcome does occur, then according to the principle of rare events, we may believe that it is reasonable to reject the null hypothesis. In other words, p-value provides an intuitive measure of the strength of evidence in the data against the null hypothesis. Then it is natural to consider the following method:

- Select a significance level α (as before, the desired type I error probability), then reject H_0 if p-value $\leq \alpha$ (otherwise, do not reject H_0 if p-value $> \alpha$).

However, unlike optimistic statements made by many textbooks (including the textbook [**DBC21**]), according to the statement made by American Statistical Association (ASA) [**WL16**], while the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced. Before further explaining this, let us first exhibit some examples.

EXAMPLE 4.6.1. Let Φ be the c.d.f. (Definition 2.3.14) of $\mathscr{N}(0,1)$. The p-value for an upper tailed z-test (i.e. reject H_0 if and only if $z \geq z_\alpha$) is just the area to the right of the computed value z under the standard normal curve:

$$\text{p-value} = 1 - \Phi(z).$$

The p-value for an lower tailed z-test (i.e. reject H_0 if and only if $z \leq -z_\alpha$) is just the area to the left of the computed value z under the standard normal curve:

$$\text{p-value} = \Phi(z).$$

More care must be exercised in the case of a two tailed test. Suppose first z is positive. We know to reject H_0 if and only if $z \geq z_{\frac{\alpha}{2}}$, which occurs precisely when $1 - \Phi(z) \leq \frac{\alpha}{2}$, i.e. $2(1 - \Phi(z)) \leq \alpha$. Comparing this to the earlier decision rule, we infer that the

$$\text{p-value} = 2(1 - \Phi(z)) = 2(1 - \Phi(|z|)).$$

If z is negative, a similar argument leads to

$$\text{p-value} = 2(1 - \Phi(-z)) = 2(1 - \Phi(|z|)).$$

Therefore, we conclude that:

$$\text{p-value} = \begin{cases} 1 - \Phi(z) & \text{for an upper tailed test (i.e. reject } H_0 \text{ if and only if } z \geq z_\alpha), \\ \Phi(z) & \text{for an lower tailed test (i.e. reject } H_0 \text{ if and only if } z \leq -z_\alpha), \\ 2(1 - \Phi(|z|)) & \text{for an two tailed test (i.e. reject } H_0 \text{ if and only if } |z| \geq z_{\frac{\alpha}{2}}). \end{cases}$$

EXAMPLE 4.6.2. Just a the p-value for a z-test is a z curve area, the p-value for a t-test will be a t curve area:

$$\text{p-value} = \begin{cases} 1 - F(z) & \text{for an upper tailed test (i.e. reject } H_0 \text{ if and only if } t \geq t_{\alpha,N-1}), \\ F(z) & \text{for an lower tailed test (i.e. reject } H_0 \text{ if and only if } t \leq -t_{\alpha,N-1}), \\ 2(1 - F(|z|)) & \text{for an two tailed test (i.e. reject } H_0 \text{ if and only if } |t| \geq t_{\frac{\alpha}{2},N-1}). \end{cases}$$

where F is the c.d.f. (Definition 2.3.14) of t_{N-1}.

EXAMPLE 4.6.3. From the bootstrap distribution of $\overline{w}_i^*$ obtained by Algorithm 2, a *bootstrap p-value* can be obtained by determining what proportion of bootstrap means $\overline{w}_1^*, \overline{w}_2^*, \cdots, \overline{w}_B^*$ are at least as contradictory to H_0 as the observed value of the test statistic $\overline{x}$:

- If $H_0 : \mu \geq \mu_0$ (i.e. we believe that the population mean is $\geq \mu_0$), then the bootstrap p-value is the proportion of values among $\overline{w}_1^*, \overline{w}_2^*, \cdots, \overline{w}_B^*$ that are $< \overline{x}$.
- If $H_0 : \mu \leq \mu_0$ (i.e. we believe that the population mean is $\leq \mu_0$), then the bootstrap p-value is the proportion of values among $\overline{w}_1^*, \overline{w}_2^*, \cdots, \overline{w}_B^*$ that are $> \overline{x}$.

Even though reporting p-values of statistical tests is common practice in academic publications of many quantitative fields, misinterpretation and misuse of p-values is widespread and has been a major topic in mathematics and metascience. Lets us borrow some words from the statement made by American Statistical Association (ASA) [**WL16**] in 2016: In February 2014, George Cobb[2] posed these questions to an American Statistical Association (ASA) forum:

- **Q.** Why do so many colleges and graduate students teach p-value $= 0.05$?
- **A.** Because that's still what the scientific community and journal editors use.
- **Q.** Why do so many people still use p-value $= 0.05$?
- **A.** Because that's what they were taught in college or graduate school.

This concern was brought to the attention of the American Statistical Association (ASA) board. Here are principles stated in [**WL16**]:

(1) p-values can indicate how incompatible the data are with a specified statistical model.
(2) p-values do not measure the probability that studied hypothesis is true, or the probability that the data were produced by random chance alone.
(3) Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
(4) Proper inference requires full reporting and transparency (otherwise, it may lead to false positives in published studies which should be strictly avoided).
(5) A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

[2] Professor Emeritus of Mathematics and Statistics at Mount Holyoke College

(6) By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

The statement made by American Statistical Association (ASA) [**WL16**] also provide a brief p-values and statistical significance reference list.

4.7. What to remember

We finally end this chapter by citing some words from the statement made by American Statistical Association (ASA) [**WL16**]: Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.

“No single index should substitute for scientific reasoning”

Bibliography

[AMH08] H. J. Adèr, G. J. Mellenbergh, and D. J. Hand. *Advising on research methods: a consultant's companion*. Huizen: Van Kessel, 2008. `https://hdl.handle.net/11245/1.293909`.

[Apo74] T. M. Apostol. *Mathematical analysis*. Addison-Wesley Publishing Co., second edition, 1974. MR0344384, Zbl:0309.26002.

[Ath87] K. B. Athreya. Bootstrap of the mean in the infinite variance case. *Ann. Statist.*, 15(2):724–731, 1987. MR0888436, Zbl:0628.62042, doi:10.1214/aos/1176350371.

[DBC21] J. L. Devore, K. N. Berk, and M. A. Carlton. *Modern mathematical statistics with applications*. Springer Texts Statist. Springer, Cham, third edition, 2021. MR4260283, Zbl:1459.62001, doi:10.1007/978-3-030-55156-8.

[DE96] T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statist. Sci.*, 11(3):189–228, 1996. MR1436647, Zbl:0955.62574, doi:10.1214/ss/1032280214.

[Dur19] R. Durrett. *Probability – theory and examples*, volume 49 of *Camb. Ser. Stat. Probab. Math.* Cambridge University Press, Cambridge, fifth edition, 2019. MR3930614, Zbl:1440.60001, doi:10.1017/9781108591034.

[Efr79] B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, (1):1–26, 1979. MR0515681, Zbl:0406.62024, doi:10.1214/aos/1176344552.

[ERT01] B. Efron, D. Rogosa, and R. Tibshirani. Resampling Methods of Estimation. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 13216–13220. 2001. doi:10.1016/B0-08-043076-7/00494-0.

[ET93] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monogr. Statist. Appl. Probab.* Chapman and Hall, New York, 1993. MR1270903, Zbl:0835.62038, doi:10.1007/978-1-4899-4541-9.

[Ete81] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrsch. Verw. Gebiete*, 55(1):119–122, 1981. MR0606010, Zbl:0438.60027, doi:10.1007/BF01013465.

[Fis22] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Lond. Phil. Trans. (A)*, 222:309–368, 1922. JFM:48.1280.02, doi:10.1098/rsta.1922.0009.

[Goo60] L. A. Goodman. On the exact variance of products. *J. Amer. Statist. Assoc.*, 55:708–713, 1960. MR0117809, Zbl:0099.13603, doi:10.2307/2281592.

[Gos08] W. S. ("Student") Gosset. The probable error of a mean. *Biometrika*, 6:1–25, 1908. Zbl:1469.62201, doi:10.1093/biomet/6.1.1.

[HSPW14] W. K. Härdle, V. Spokoiny, V. Panov, and W. Wang. *Basics of modern mathematical statistics – Exercises and solutions*. Springer Texts Statist. Springer, Heidelberg, 2014. MR3135149, Zbl:1286.00001, doi:10.1007/978-3-642-36850-9.

[Hin94] D. Hinkley. [Bootstrap: More than a Stab in the Dark?]: Comment. *Statist. Sci.*, 9(3):400–403, 1994. doi:10.1214/ss/1177010387.

[HPS71] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to probability theory*. The Houghton Mifflin Series in Statistics. Houghton Mifflin Co., Boston, MA, 1971. MR0358880, Zbl:0258.60002.

[Kow23] P.-Z. Kow. *Complex Analysis*. National Chengchi University, Taipei, 2023. https://puzhaokow1993.github.io/homepage.

[Kow24] P.-Z. Kow. *Calculus*. National Chengchi University, Taipei, 2024. https://puzhaokow1993.github.io/homepage.

[LM21] R. J. Larsen and M. L. Marx. *An introduction to mathematical statistics and its applications*. Pearson, sixth edition, 2021. https://www.pearson.com.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Texts Statist. Springer-Verlag, New York, second edition, 1998. MR1639875, Zbl:0916.62017, doi:10.1007/b98854.

[Moo71] D. S. Moore. Classroom notes: maximum likelihood and sufficient statistics. *Amer. Math. Monthly*, 78(1):50–52, 1971. MR1536181, Zbl:0205.45902, doi:10.2307/2317488.

[Rud87] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987. MR0924157, Zbl:0925.00005.

[Sch91] L. L. Scharf. *Statistical signal processing: Detection, estimation, and time series analysis*. Addison-Wesley Series in Electrical and Computer Engineering: Digital Signal Processing. Addison-Wesley, 1991. Zbl:1130.62303.

[SD15] V. Spokoiny and T. Dickhaus. *Basics of modern mathematical statistics*. Springer Texts Statist. Springer, Heidelberg, 2015. MR3289985, Zbl:1401.62006, doi:10.1007/978-3-642-39909-1.

[vB98] L. von Bortkewitsch. *Das Gesetz der kleinen Zahlen (German)*. Teubner, Leipzig, 1898. JFM:29.0188.03, EuDML:204250.

[WMS07] D. Wackerly, W. Mendenhall, and R. L. Scheaffer. *Mathematical statistics with applications*. Cengage, seventh edition, 2007. https://www.cengage.ca.

[WL16] R. L. Wasserstein and N. A. Lazar. The ASA's statement on p-values: context, process, and purpose. *Amer. Statist.*, 70(2):129–133, 2016. MR3511040, Zbl:7665862, doi:10.1080/00031305.2016.1154108.

[WZ15] R. L. Wheeden and A. Zygmund. *Measure and integral. An introduction to real analysis*. Pure Appl. Math. CRC Press, Boca Raton, FL, second edition, 2015. MR3381284, Zbl:1326.26007.