Author name(s)

Leture notes for STAT2602

– Monograph –

October 14, 2019

Springer

Contents

Chapter 1
Basic concepts

1.1 Discrete distribution

Let x be a realization of a discrete random variable $X \in \mathscr{R}$. Then,

$$f(x) = \mathrm{P}(X = x)$$

is the probability density function (p.d.f.) of X.

Theorem 1.1. *(Discrete univariate probability density function) A discrete univariate probability density function has the following properties:*

(1) $f(x) > 0$ *for* $x \in S$*;*
(2) $\sum_{x \in S} f(x) = 1$*;*
(3) $\mathrm{P}(X \in A) = \sum_{x \in A} f(x)$*, where* $A \subset S$*.*

Based on the p.d.f. $f(x)$, we define the function $F(x)$ by

$$F(x) = \mathrm{P}(X \leq x) = \sum_{s \in S \text{ and } s \leq x} f(s).$$

The function $F(x)$ is called the cumulative distribution function (c.d.f.) of the discrete random variable X. Note that $F(x)$ is a step function on $\mathscr{R}$ and the height of a step at x, $x \in S$, equals the probability $f(x)$ (see Fig.1.1 for an illustration).

From Theorem 1.1, we can obtain the following theorem.

Theorem 1.2. *(Discrete cumulative distribution function) A discrete univariate cumulative distribution function has the following properties:*

(1) $0 \leq F(x) \leq 1$ *for* $x \in \mathscr{R}$*;*
(2) $F(x)$ *is a nondecreasing function of* x*;*
(3) $F(\infty) = 1$ *and* $F(-\infty) = 0$.

Remark 1.1. The p.d.f. $f(x)$ and the c.d.f. $F(x)$ are one-to-one corresponding. We can first define the c.d.f. $F(x)$, and then define the p.d.f. $f(x)$ by

$$f(x) = F(x) - F(x-) \quad \text{for } x \in S.$$

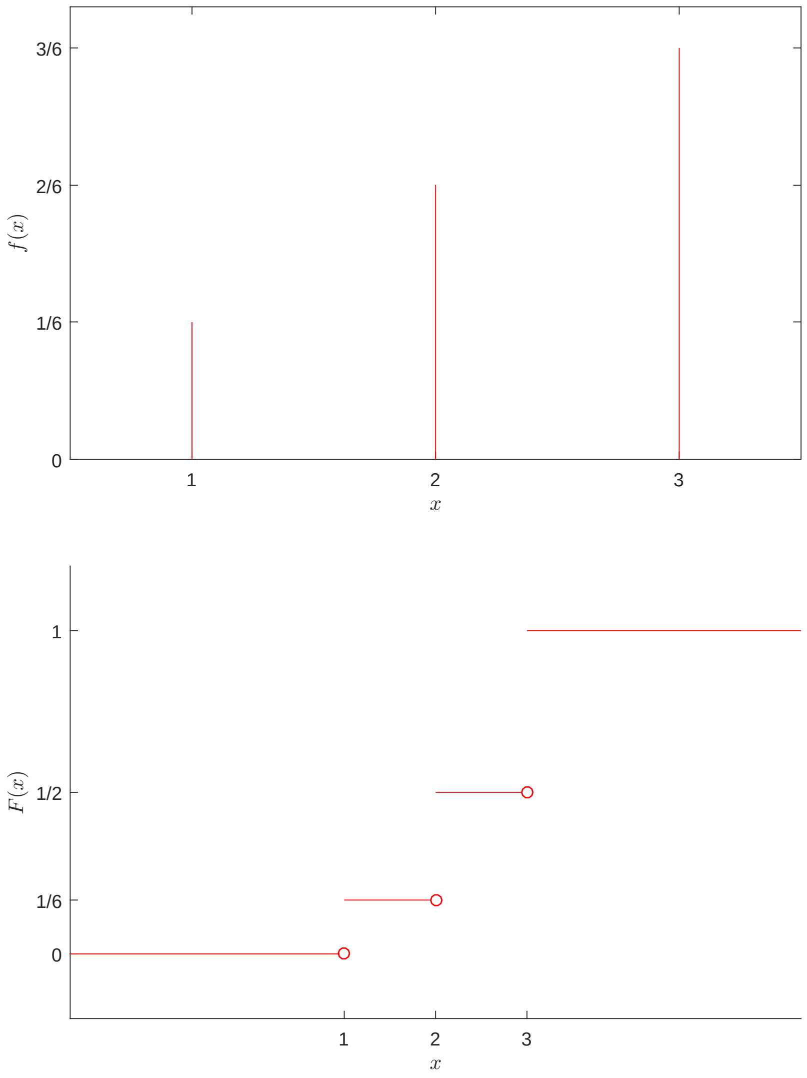

Fig. 1.1 The top panel is the p.d.f $F(x)$ of a discrete random variable X, where $f(x) = \mathrm{P}(X = x) = x/6$ for $x = 1,2,3$, and the bottom panel is the corresponding c.d.f. $F(x)$.

Property 1.1. *Two discrete random variables X and Y are independent if and only if $F(x,y) = F_X(x)F_Y(y)$ for all $(x,y) \in S$, where F is joint distribution of X and Y, and F_X (or F_Y) is the marginal distribution of X (or Y).*

Property 1.2. *Let X and Y be two independent discrete random variables. Then,*
(a) for arbitrary countable sets A and B,

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B);$$

(b) for any real functions $g(\cdot)$ and $h(\cdot)$, $g(X)$ and $h(Y)$ are independent.

1.2 Continuous distribution

Let $X \in \mathscr{R}$ be a continuous random variable. The probability of X lies in an interval $(a,b]$ is

$$\mathrm{P}(a < X \leq b) = \int_a^b f(x)dx$$

for some non-negative function $f(\cdot)$. We call $f(x)$ the p.d.f. of the continuous random variable X.

Theorem 1.3. *(Continuous univariate probability density function) A continuous univariate probability density function has the following properties:*

(1) $f(x) \geq 0$ *for* $x \in \mathscr{R}$*;*
(2) $\int_{\mathscr{R}} f(x)dx = 1$*;*
(3) $\mathrm{P}(X \in A) = \int_A f(x)dx$ *for* $A \subset \mathscr{R}$*.*

Based on the p.d.f. $f(x)$, the c.d.f of X is defined as

$$F(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f(s)ds,$$

which also satisfies Theorem 1.2. From the fundamental theorems of calculus, we have $F'(x) = f(x)$ if exists. Since there are no steps or jumps in a continuous c.d.f., it must be true that $\mathrm{P}(X = b) = 0$ for all real values of b.

As you can see, the definition for the p.d.f. (or c.d.f.) of a continuous random variable differs from the definition for the p.d.f. (or c.d.f.) of a discrete random variable by simply changing the summations that appeared in the discrete case to integrals in the continuous case.

Example 1.1. (Uniform distribution) A random variable X has a uniform distribution if

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Briefly, we say that $X \sim U(a,b)$.

Property 1.3. *If F is a continuous c.d.f. and $X \sim U(0,1)$, then $Y = F^{-1}(X) \sim F$.*

Proof.

$$\mathrm{P}(Y \leq x) = \mathrm{P}(F^{-1}(X) \leq x) = \mathrm{P}(X \leq F(x)) = F(x).$$

Note that this property helps us to generate a random variable from certain distribution. □

Example 1.2. (Normal distribution) A random variable X has a normal distribution if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \text{ for } x \in \mathscr{R},$$

where $\mu \in \mathscr{R}$ is the location parameter and $\sigma > 0$ is the scale parameter. Briefly, we say that $X \sim \mathrm{N}(\mu, \sigma^2)$. A simple illustration of $f(x)$ with different values of μ and σ is given in Fig.1.2.

Further, $Z = (X-\mu)/\sigma \sim \mathrm{N}(0,1)$ (the standard normal distribution), and the c.d.f. of Z is typically denoted by $\Phi(x)$, where

$$\Phi(x) = \mathrm{P}(Z \le x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{s^2}{2}\right] ds.$$

Numerical approximations for $\Phi(x)$ have been well tabulated in practice.

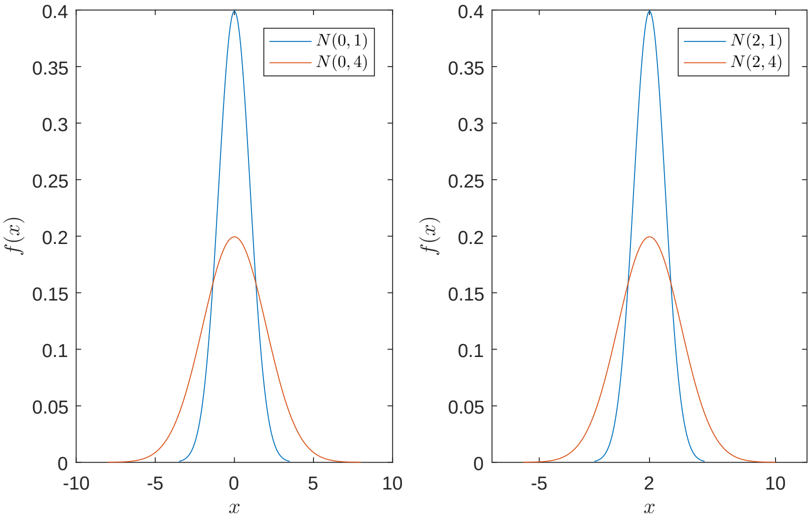

Fig. 1.2 The p.d.f $f(x)$ of $\mathrm{N}(\mu,\sigma^2)$.

Property 1.4. *If the p.d.f. of a continuous random variable X is $f_X(x)$ for $x \in \mathscr{R}$, the p.d.f. of $Y = aX + b$ for $a \neq 0$ is $f_Y(x) = \frac{1}{a} f_X(\frac{x-b}{a})$ for $x \in \mathscr{R}$.*

Proof. Let $F_X(x)$ be the c.d.f. of X. Then, the c.d.f. of Y is

$$F_Y(x) = \mathrm{P}(Y \le x) = \mathrm{P}(aX + b \le x) = \mathrm{P}\left(X \le \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right)$$

for $x \in \mathscr{R}$. Hence,

$$f_Y(x) = F_Y'(x) = \frac{1}{a} F_X'\left(\frac{x-b}{a}\right) = \frac{1}{a} f_X\left(\frac{x-b}{a}\right).$$

This completes the proof. □

Property 1.5. *Two continuous random variables X and Y are independent if and only if*

$$F(x,y) = F_X(x)F_Y(y) \quad \textit{for all } (x,y) \in R^2.$$

Property 1.6. *Let X and Y be two independent continuous random variables. Then,*
(a) for arbitrary intervals A and B,

$$\mathrm{P}(X \in A, Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B);$$

(b) for any real functions $g(\cdot)$ and $h(\cdot)$, $g(X)$ and $h(Y)$ are independent.

1.3 Empirical distribution

Suppose that $X \sim F(x)$ is a random variable resulting from a random experiment. Repeat this experiment n independent times, we get n random variables $X_1, \cdots, X_n$ associated with these outcomes. The collection of these random variables is called a sample from a distribution with c.d.f. $F(x)$ (or p.d.f. $f(x)$). The number n is called the sample size.

As all random variables in a sample follow the same c.d.f. as X, we expect that they can give us the information about the c.d.f of X. Next, we are going to show that the empirical distribution of $\{X_1, \cdots, X_n\}$ is close to $F(x)$ in some probability sense.

The empirical distribution of $\{X_1, \cdots, X_n\}$ is defined as

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathrm{I}(X_k \leq x)$$

for $x \in \mathscr{R}$, where $\mathrm{I}(A)$ is an indicator function such that $\mathrm{I}(A) = 1$ if A holds and $\mathrm{I}(A) = 0$ otherwise. Obviously, $F_n(x)$ assigns the probability $1/n$ to each X_k, and we can check that it satisfies Theorem 1.2 (please do it by yourself). Since $F_n(x)$ is the relative frequency of the event $X \leq x$, it is an approximation of the probability $\mathrm{P}(X \leq x) = F(x)$. Thus, the following result is expected.

Theorem 1.4. *As $n \to \infty$, $\sup_{x \in \mathscr{R}} |F_n(x) - F(x)| \to 0$ almost surely (a.s.).*

The proof of aforementioned theorem is omitted. Roughly speaking, the almost surely convergence in this theorem means that $F_n(x)$ provides an estimate of the c.d.f. $F(x)$ for each realization $\{x_1, \cdots, x_n\}$. To see it more clearly, Fig.1.3 plots the empirical distribution function $F_n(x)$ based on a realization $\{x_1, \cdots, x_n\}$ with $X_i \sim \mathrm{N}(0,1)$. As a comparison, the c.d.f. $\Phi(x)$ of $\mathrm{N}(0,1)$ is also included in Fig.1.3. From this figure, we can see that $F_n(x)$ is getting close to $\Phi(x)$ as the sample size n increases, and this is consistent to the conclusion in Theorem 1.4.

Example 1.3. Let X denote the number of observed heads when four coins are tosses independently and at random. Recall that the distribution of X is $\mathrm{B}(4, 1/2)$. One thousand repetitions of this experiment (actually simulated on the computer) yielded the following results:

Number of heads	Frequency
0	65
1	246
2	358
3	272
4	59

This information determines the following empirical distribution function:

x	$F_{1000}(x)$	x	$F_{1000}(x)$
$(-\infty, 0)$	0.000	$[2,3)$	0.669
$[0,1)$	0.065	$[3,4)$	0.941
$[1,2)$	0.311	$[4,\infty)$	1.000

The graph of the empirical distribution function $F_{1000}(x)$ and the theoretical distribution function $F(x)$ for the binomial distribution are very close (please check it by yourself).

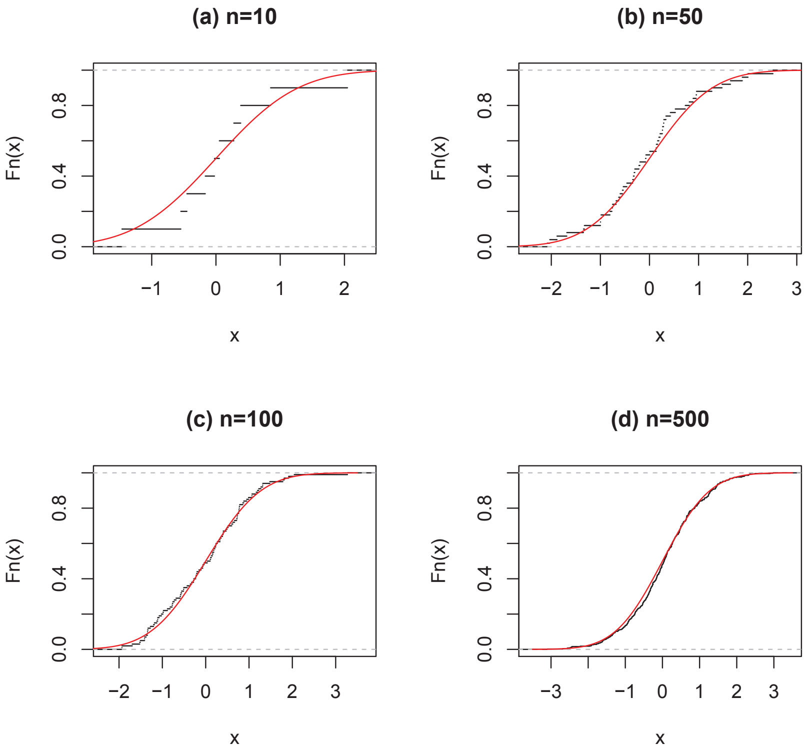

Fig. 1.3 The black step function is the empirical distribution function $F_n(x)$ based on a realization $\{x_1, \cdots, x_n\}$ with $X_i \sim \mathrm{N}(0,1)$. The red solid line is the c.d.f. $\Phi(x)$ of $\mathrm{N}(0,1)$.

Example 1.4. The following numbers are a random sample of size 10 from some distribution:

$$-0.49, 0.90, 0.76, -0.97, -0.73, 0.93, -0.88, -0.75, 0.88, 0.96.$$

(a) Write done the empirical distribution; (b) use the empirical distribution to estimate $\mathrm{P}(X \le -0.5)$ and $\mathrm{P}(-0.5 \le X \le 0.5)$.

Solution. Order the random sample:

$$-0.97, -0.88, -0.75, -0.73, -0.49, 0.76, 0.88, 0.90, 0.93, 0.96.$$

Then, the empirical distribution function $F_{10}(x)$ is as follows:

x	$F_{10}(x)$	x	$F_{10}(x)$
$(-\infty, -0.97)$	0.0	$[-0.49, 0.76)$	0.5
$[-0.97, -0.88)$	0.1	$[0.76, 0.88)$	0.6
$[-0.88, -0.75)$	0.2	$[0.88, 0.90)$	0.7
$[-0.75, -0.73)$	0.3	$[0.90, 0.93)$	0.8
$[-0.73, -0.49)$	0.4	$[0.93, 0.96)$	0.9
		$[0.96, \infty)$	1.0

Thus, $\mathrm{P}(X \le -0.5) = F(-0.5) \approx F_{10}(-0.5) = 0.4$ and $\mathrm{P}(-0.5 \le X \le 0.5) = F(0.5) - F(-0.5) \approx F_{10}(0.5) - F_{10}(-0.5) = 0.5 - 0.4 = 0.1$. □

The question now is how to estimate the p.d.f. $f(x)$? The answer is "relative frequency histogram".

For the discrete random variable X, we can estimate $f(x) = \mathrm{P}(X = x)$ by the relative frequency of occurrences of x. That is,

$$f(x) \approx f_n(x) = \frac{\sum_{k=1}^{n} \mathrm{I}(X_k = x)}{n}.$$

Example 1.3. (con't) The relative frequency of observing $x = 0, 1, 2, 3$ or 4 is listed in the second column, and it is close to the value of $f(x)$, which is the p.d.f of $\mathrm{B}(4, 1/2)$.

x	$f_{1000}(x)$	$f(x)$
0	0.065	0.0625
1	0.246	0.2500
2	0.358	0.3750
3	0.272	0.2500
4	0.059	0.0625

By increasing the value of n, the difference between $f_n(x)$ and $f(x)$ will become small. □

For the continuous random variable X, we first define the so-called class intervals. Choose an integer $l \geq 1$, and a sequence of real numbers $c_0, c_1, \cdots, c_l$ such that $c_0 < c_1 < \cdots < c_l$. The class intervals are

$$(c_0, c_1],\ (c_1, c_2], \cdots, (c_{l-1}, c_l].$$

Roughly speaking, the class intervals are a non-overlapped partition of the interval $[X_{\min}, X_{\max}]$. As $f(x) = F'(x)$, we expect that when c_{j-1} and c_j is close,

$$f(x) \approx \frac{F(c_j) - F(c_{j-1})}{c_j - c_{j-1}} \quad \text{for } x \in (c_{j-1}, c_j],\ j = 1, 2, \cdots, l.$$

Note that

$$F(c_j) - F(c_{j-1}) = \mathrm{P}(X \in (c_{j-1}, c_j]) \approx \frac{\sum_{k=1}^{n} \mathrm{I}(X_k \in (c_{j-1}, c_j])}{n}$$

is the relative frequency of occurrences of $X_k \in (c_{j-1}, c_j]$. Thus, we can approximate $f(x)$ by

$$f(x) \approx h_n(x) = \frac{\sum_{k=1}^{n} \mathrm{I}(X_k \in (c_{j-1}, c_j])}{n(c_j - c_{j-1})} \quad \text{for } x \in (c_{j-1}, c_j],\ j = 1, 2, \cdots, l.$$

We call $h_n(x)$ the relative frequency histogram. Clearly, the way that we define the class intervals is not unique, and hence the value of $h_n(x)$ is not unique. When the sample size n is large and the length of the class interval is small, $h_n(x)$ is expected to be a good estimate of $f(x)$.

The property of $h_n(x)$ is as follows:

(i) $h_n(x) \geq 0$ for all x;

(ii) The total area bounded by the x axis and below $h_n(x)$ equals one, i.e.,

$$\int_{c_0}^{c_l} h_n(x) dx = 1;$$

(iii) The probability for an event A, which is composed of a union of class intervals, can be estimated by the area above A bounded by $h_n(x)$, i.e.,

$$\mathrm{P}(A) \approx \int_A h_n(x)dx.$$

Example 1.5. A random sample of 50 college-bound high school seniors yielded the following high school cumulative grade point averages (GPA's).

3.77 2.78 3.40 2.20 3.26
3.00 2.85 2.65 3.08 2.92
3.69 2.83 2.75 3.97 2.74
2.90 3.38 2.38 2.71 3.31
3.92 3.29 4.00 3.50 2.80
3.57 2.84 3.18 3.66 2.86
2.81 3.10 2.84 2.89 2.59
2.95 2.77 3.90 2.82 3.89
2.83 2.28 3.20 2.47 3.00
3.78 3.48 3.52 3.20 3.30

(a) Construct a frequency table for these 50 GPA's using 10 intervals of equal length with $c_0 = 2.005$ and $c_{10} = 4.005$.
(b) Construct a relative frequency histogram for the grouped data.
(c) Estimate $f(3)$ and $f(4)$.

Solution. (a) and (b). The frequency and the relative frequency histogram based on the class intervals are given in the following table:

class interval	frequency	relative frequency histogram	class interval	frequency	relative frequency histogram
(2.005, 2.205]	1	0.1	(3.005, 3.205]	5	0.5
(2.205, 2.405]	2	0.2	(3.205, 3.405]	6	0.6
(2.405, 2.605]	2	0.2	(3.405, 3.605]	4	0.4
(2.605, 2.805]	7	0.7	(3.605, 3.805]	4	0.4
(2.805, 3.005]	14	1.4	(3.805, 4.005]	5	0.5

(c) As $3 \in (2.805, 3.005]$ and $4 \in (3.805, 4.005]$,

$$f(3) \approx h_{50}(3) = \frac{14}{50 \times (3.005 - 2.805)} = 1.4,$$

$$f(4) \approx h_{50}(4) = \frac{5}{50 \times (4.005 - 3.805)} = 0.5.$$

□

1.4 Expectation

Definition 1.1. *(Expectation of a discrete statistic) If $u(X)$ is a function of a discrete random variable X that has a p.d.f. $f(x)$, then*

$$\mathrm{E}[u(X)] = \sum_{x \in S} u(x) f(x),$$

where the summation is taken over all possible values of x. If $\mathrm{E}[u(X)]$ exists, it is called the mathematical expectation (or expected value) of $u(X)$.

Remark 1.2. $\mathrm{E}[u(X)]$ exists if

$$\sum_{x\in S}|u(x)|f(x) < \infty.$$

We say two random variables X_1 and X_2 are uncorrelated, if $\mathrm{Cov}(X_1,X_2) = 0$, where

$$\mathrm{Cov}(X_1,X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2)$$

is the covariance of X_1 and X_2.

Property 1.7. *Let X be a discrete random variable with finite mean $\mathrm{E}(X)$, and let a and b be constants. Then,*
(i) $\mathrm{E}(aX+b) = a\mathrm{E}(X)+b$;
(ii) if $\mathrm{P}(X=b) = 1$, then $\mathrm{E}(X) = b$;
(iii) if $\mathrm{P}(a < X \le b) = 1$, then $a < \mathrm{E}(X) \le b$;
(iv) if $g(X)$ and $h(X)$ have finite mean, then

$$\mathrm{E}(g(X)+h(X)) = \mathrm{E}(g(X))+\mathrm{E}(h(X)).$$

Property 1.8. *If $X \ge 0$ takes integer values, then $\mathrm{E}(X) = \sum_{x=1}^{\infty}\mathrm{P}(X \ge x) = \sum_{x=0}^{\infty}\mathrm{P}(X > x)$.*

Definition 1.2. *(Expectation of a continuous statistic) If $u(X)$ is a function of a continuous random variable that has a p.d.f. $f(x)$, then*

$$\mathrm{E}[u(X)] = \int_{\mathscr{R}} u(x)f(x)\mathrm{d}x.$$

If $\mathrm{E}[u(X)]$ exists, it is called the mathematical expectation (or expected value) of $u(X)$.

Remark 1.3. $\mathrm{E}[u(X)]$ exists if

$$\int_{\mathscr{R}}|u(x)|f(x)\mathrm{d}x < \infty.$$

Example 1.6. Let X have the $\mathrm{N}(\mu,\sigma^2)$ distribution. Then,

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty}\frac{x}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]dx\\ &= \int_{-\infty}^{\infty}\frac{s\sigma+\mu}{\sqrt{2\pi}\sigma}\exp\left[-\frac{s^2}{2}\right]d(s\sigma+\mu)\ \text{(letting } s=(x-\mu)/\sigma)\\ &= \int_{-\infty}^{\infty}\frac{s\sigma}{\sqrt{2\pi}}\exp\left[-\frac{s^2}{2}\right]ds+\mu\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{s^2}{2}\right]ds.\end{aligned}$$

The first integrand is an odd function, and so the integral over $\mathscr{R}$ is zero. The second integrand is one by some algebra. Hence, $\mathrm{E}(X) = \mu$.

Property 1.9. *Let X be a continuous random variable, a and b be constants, and g and h be functions. Then,*
(i) if $g(X)$ and $h(X)$ have finite mean then

$$\mathrm{E}(ag(X)+bh(X)) = a\mathrm{E}(g(X))+b\mathrm{E}(h(X));$$

(ii) if $\mathrm{P}(a \le X \le b) = 1$*, then* $a \le \mathrm{E}(X) \le b$*;*
(iii) if h is non-negative, then for $a > 0$*,* $\mathrm{P}(h(X) \ge a) \le \mathrm{E}(h(X)/a)$*;*
(iv) if g is convex, then $g(\mathrm{E}(X)) \le \mathrm{E}(g(X))$*.*

Property 1.10. *Let X be a non-negative random variable with c.d.f. F, p.d.f f, and finite expected value* $\mathrm{E}(X)$*. Then,*

$$\mathrm{E}(X) = \int_0^\infty (1-F(x))dx.$$

Property 1.11. *Let* a,b,c*, and d be constants. Then,*
(i) $E(X^2) = 0$ *if and only if* $\mathrm{P}(X=0)=1$*;*
(ii) $\mathrm{Cov}(aX+b, cY+d) = ac\,\mathrm{Cov}(X,Y)$*;*
(iii) $\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)$*;*
(iv) if X and Y are independent, $\mathrm{E}(h(X)g(Y)) = \mathrm{E}(h(X))\mathrm{E}(g(Y))$ *provided that* $\mathrm{E}(h(X)) < \infty$ *and* $\mathrm{E}(g(Y)) < \infty$*.*
(v) $-1 \le \rho(X,Y) \le 1$*;*
(vi) $|\rho(X,Y)| = 1$ *if and only if* $\mathrm{P}(X = aY+b) = 1$ *for some constants a and b;*
(vii) $\rho(aX+b, cY+d) = \mathrm{sgn}(ac)\rho(X,Y)$*, where* $\mathrm{sgn}(x)$ *denotes the sign of x;*
(viii) if X and Y are independent, $\rho(X,Y) = 0$*.*

Property 1.12. *(Cauchy-Schwarz inequality) If* $\mathrm{E}(X^2)\mathrm{E}(Y^2) < \infty$*, then*

$$\mathrm{E}(XY) \le \sqrt{\mathrm{E}(X^2)\mathrm{E}(Y^2)}.$$

Proof. Without loss generality, we assume that $\mathrm{E}(Y^2) > 0$. Note that

$$0 \le \mathrm{E}\left[(X\mathrm{E}(Y^2) - Y\mathrm{E}(XY))^2\right] = \mathrm{E}(Y^2)\left[\mathrm{E}(X^2)\mathrm{E}(Y^2) - (\mathrm{E}(XY))^2\right].$$

Hence, the conclusions holds. □

Chapter 2
Preliminary

2.1 Moment generating function

Let r be a positive integer. The r-th moment about the origin of a random variable X is defined as $\mu_r = \mathrm{E}(X^r)$. In order to calculate μ_r, we can make use of the moment generating function (m.g.f.).

Definition 2.1. *(Moment Generating Function) A moment generating function of X is a function of $t \in \mathscr{R}$ defined by $M_X(t) = \mathrm{E}(e^{tX})$ if exists.*

Property 2.1. *Suppose $M_X(t)$ exists. Then,*

(1) $M_X(t) = \sum_{r=0}^{\infty} \mu_r \left(\frac{t^r}{r!}\right)$;
(2) $\mu_r = M_X^{(r)}(0)$ for $r = 1, 2, \ldots$;
(3) For constants a and b, $M_{aX+b}(t) = e^{bt} M_X(at)$.

Proof. (1) For a discrete random variable X we have

$$M_X(t) = \sum_x e^{tx} \mathrm{P}(X = x) = \sum_x \sum_{r=0}^{\infty} \frac{(tx)^r}{r!} \mathrm{P}(X = x) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \sum_x x^r \mathrm{P}(X = x) = \sum_{r=0}^{\infty} \frac{t^r}{r!} \mu_r.$$

For a continuous random variable X, the proof is similar by using integrals instead of sums.

(2) Make use of (1).

(3) $M_{aX+b}(t) = \mathrm{E}\left[e^{(aX+b)t}\right] = e^{bt}\mathrm{E}(e^{atX}) = e^{bt} M_X(at)$. □

From parts (1)-(2) above, we have

$$M_X(t) = M_X(0) + M_X^{(1)}(0)\left(\frac{t}{1!}\right) + M_X^{(2)}(0)\left(\frac{t^2}{2!}\right) + M_X^{(3)}(0)\left(\frac{t^3}{3!}\right) + \cdots,$$

which is called the Maclaurin's series of $M_X(t)$ around $t = 0$. If the Maclaurin's series expansion of $M_X(t)$ can be found, the r-th moment μ_r is the coefficient of $t^r/r!$; or if $M_X(t)$ exists and the moments are given, we can frequently sum the Maclaurin's series to obtain the closed form of $M_X(t)$.

Property 2.2. *If $M_X(t)$ exists, there is a one-to-one correspondence between $M_X(t)$ and the p.d.f. $f(x)$ (or c.d.f. $F(x)$).*

Proof. The proof is omitted. □

The above property shows that we can decide the distribution of X by calculating its m.g.f.

Example 2.1. The m.g.f. of $\mathrm{N}(\mu, \sigma^2)$ is

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \int_{-\infty}^{+\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-2\sigma^2 tx + x^2 - 2\mu x + \mu^2}{-2\sigma^2} \right) \mathrm{d}x \\
&= \exp\left(\frac{-2\mu\sigma^2 t - \sigma^4 t^2}{-2\sigma^2} \right) \times \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu-\sigma^2 t)^2}{-2\sigma^2} \right) \mathrm{d}x \\
&= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2 \right),
\end{aligned}$$

because $\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(x-\mu-\sigma^2 t)^2}{-2\sigma^2} \right)$ is the density function of $\mathrm{N}(\mu + \sigma^2 t, \sigma^2)$. □

Example 2.2. Find the m.g.f. of a random variable X following a Poisson distribution with mean λ.

Solution.

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} \mathrm{P}(X = x) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}.$$

□

Example 2.3. Find the m.g.f. of a random variable which has a (probability) density function given by

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

and then use it to find μ_1, μ_2, and μ_3.

Solution.

$$\begin{aligned}
M_X(t) &= \mathrm{E}(e^{tX}) \\
&= \int_{-\infty}^{+\infty} e^{tx} f(x) \mathrm{d}x = \int_0^{+\infty} e^{tx} e^{-x} \mathrm{d}x = \begin{cases} \left. \dfrac{e^{(t-1)x}}{t-1} \right|_0^{+\infty} = \dfrac{1}{1-t}, & \text{for } t < 1; \\ \text{does not exist,} & \text{for } t \geq 1. \end{cases}
\end{aligned}$$

Then,

$$\begin{aligned}
&\mu_1 = M_X^{(1)}(0) = \left. \frac{1}{(1-t)^2} \right|_{t=0} = 1, \ \mu_2 = M_X^{(2)}(0) = \left. \frac{2}{(1-t)^3} \right|_{t=0} = 2, \\
&\mu_3 = M_X^{(3)}(0) = \left. \frac{2 \times 3}{(1-t)^4} \right|_{t=0} = 3!.
\end{aligned}$$

□

Property 2.3. *If $X_1, X_2, \ldots, X_n$ are independent random variables, $M_{X_i}(t)$ exists for $i = 1, 2, \cdots, n$, and $Y = X_1 + X_2 + \cdots + X_n$, then $M_Y(t)$ exists and*

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Proof. The proof is left as an excise. □

Example 2.4. Find the distribution of the sum of n independent random variables $X_1, X_2, \ldots, X_n$ following Poisson distributions with means $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively.

Solution. Let $Y = X_1 + X_2 + \cdots + X_n$. Then,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{(e^t-1)\sum_{i=1}^{n}\lambda_i},$$

which is the m.d.f. of Poisson random variable with mean $\sum_{i=1}^{n} \lambda_i$. Hence, by Example 2.2 and Property 2.3, $Y \sim$ Poisson distribution with mean $\sum_{i=1}^{n} \lambda_i$. □

Example 2.5. For positive numbers α and λ, find the moment generating function of a gamma distribution Gamma(α, λ) of which the density function is given by

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & \text{for } x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Solution.

$$\begin{aligned} M_X(t) = \mathrm{E}(e^{tX}) &= \int_0^{+\infty} e^{tx} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \mathrm{d}x \\ &= \int_0^{+\infty} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \mathrm{d}x \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha} \int_0^{+\infty} \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \mathrm{d}x \\ &= \begin{cases} \dfrac{\lambda^\alpha}{(\lambda-t)^\alpha}, & \text{for } t < \lambda; \\ \text{does not exist}, & \text{for } t \geq \lambda, \end{cases} \end{aligned}$$

where

$$\int_0^{+\infty} \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \mathrm{d}x = 1$$

is due to the fact that

$$\frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} \text{ for } x > 0$$

is the density function of a Gamma$(\alpha, \lambda - t)$ distribution. □

Example 2.6. Find the distribution of the sum of n independent random variables $X_1, X_2, \ldots, X_n$ where X_i follows Gamma(α_i, λ), $i = 1, 2, \ldots, n$, with the p.d.f. given by

$$f(x) = \begin{cases} \dfrac{\lambda^{\alpha_i} x^{\alpha_i - 1} e^{-\lambda x}}{\Gamma(\alpha_i)}, & \text{for } x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Solution. From the previous example, we know that the m.g.f. of X_i is

$$M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_i} \text{ for } t < \lambda,\ i = 1, 2, \ldots, n.$$

Hence, the m.g.f. of $X_1 + X_2 + \cdots + X_n$ is

$$\prod_{i=1}^{n} M_{X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1 + \alpha_2 + \cdots + \alpha_n} \quad \text{for } t < \lambda.$$

Therefore, $X_1 + X_2 + \cdots + X_n$ follows a $\text{Gamma}(\alpha_1 + \alpha_2 + \cdots + \alpha_n, \lambda)$ distribution. □

Example 2.7. Prove that the sum of n independent random variables $X_1, X_2, \ldots, X_n$ each following a Bernoulli distribution with parameter p follows $\text{B}(n, p)$, the binomial distribution with parameters n and p.

Proof. On one hand, for $i = 1, 2, \ldots, n$, $M_{X_i}(t) = e^{0t}\text{P}(X_i = 0) + e^{1t}\text{P}(X_i = 1) = (1 - p) + e^t p = 1 - p + pe^t$, and hence the m.g.f. of $X_1 + X_2 + \cdots + X_n$ is

$$\prod_{i=1}^{n} M_{X_i}(t) = (1 - p + pe^t)^n.$$

On the other hand, the m.g.f. of $\text{B}(n, p)$ is

$$\sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (pe^t + 1 - p)^n.$$

Therefore, $X_1 + X_2 + \cdots + X_n$ follows a $\text{B}(n, p)$ distribution. □

Example 2.8. Let $X_1, X_2, \ldots, X_n$ be independent $\text{N}(0, 1)$ random variables. Show that $Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$.

Solution. By the independence of $X_1, X_2, \ldots, X_n$ and Property 2.3,

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i^2}(t) = \prod_{i=1}^{n} \frac{1}{\sqrt{1 - 2t}} = (1 - 2t)^{-n/2},$$

for $t < 1/2$, where we have used the fact that

$$\begin{aligned} M_{X_i^2}(t) = \text{E}(e^{tX_i^2}) &= \int_{-\infty}^{\infty} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \,\text{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(1-2t)x^2}{2}\right) \text{d}x \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \text{d}\left(\frac{y}{\sqrt{1-2t}}\right) \quad \text{(by letting } y = \sqrt{1-2t}x\text{)} \\ &= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \text{d}y = \frac{1}{\sqrt{1-2t}}. \end{aligned}$$

Note that the m.g.f. of χ_n^2 is $(1-2t)^{-n/2}$ for $t < 1/2$. Hence, by Property 2.2, $Y \sim \chi_n^2$. □

2.2 Convergence

Functions of random variables are of interest in statistical applications. Usually, functions of random sample $\mathbf{X} = \{X_1, \cdots, X_n\}$ are called statistics. Two important statistics are the sample mean $\overline{X}$ and the sample variance S^2, where

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

Although in a particular sample, say $x_1, \cdots, x_n$, we observe definite values of these statistics, $\bar{x}$ and s^2, we should recognize that each value is only one observation of the respective random variables $\overline{X}$ and S^2. That is, each $\overline{X}$ or S^2 is also a random variable with its own distribution.

Suppose that the random sample $\mathbf{X}$ from a distribution $F(x)$ with mean $\mu = \mathrm{E}(X)$ and variance $\sigma^2 = \mathrm{Var}(X)$. When n is large, Theorem 1.4 shows that $F(x)$ can be well approximated by $F_n(x)$. Meanwhile, we can easily show that $\overline{X}$ and S^2 are the mean and variance of a random variable from a distribution $F_n(x)$. Therefore, it is expected that when n is large, μ and σ^2 can be well approximated by $\overline{X}$ and S^2, respectively.

Definition 2.2. *(Convergence in probability) Let $(Z_n; n \geq 1)$ be a sequence of random variables. We say the sequence Z_n converges in probability to Z if, for any $\varepsilon > 0$,*

$$\mathrm{P}(|Z_n - Z| > \varepsilon) \to 0 \quad \textit{as } n \to \infty.$$

For brevity, this is often written as $Z_n \to_p Z$.

Remark 2.1. 1. For a deterministic sequence $\{a_n\}$,

$$a_n \to a \text{ as } n \to \infty \Longleftrightarrow \text{for any } \varepsilon > 0, \text{ there exists an integer } N(\varepsilon) > 0$$
$$\text{such that when } n \geq N, |a_n - a| < \varepsilon \text{ (for sure!)}$$

2. For a random sequence $\{Z_n\}$,

$$Z_n \to_p Z \text{ as } n \to \infty \Longleftrightarrow \text{for any } \varepsilon > 0, \text{ there exists an integer } N(\varepsilon) > 0$$
$$\text{such that when } n \geq N, \mathrm{P}(|Z_n - Z| < \varepsilon) \text{ is very close}$$
$$\text{to one (but not for sure!)}$$

3. ε specifies the accuracy of the convergence, which can be achieved for large $n (\geq N)$.

Theorem 2.1. *(Weak law of large numbers (LLN)) Let $(X_i; i \geq 1)$ be a sequence of independent random variables having the same finite mean and variance, $\mu = \mathrm{E}(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1)$. Then, as $n \to \infty$,*

$$\overline{X} \to_p \mu.$$

It is customary to write $S_n = \sum_{i=1}^{n} X_i$ for the partial sums of the X_i.

Proof. It suffices to show that for any $\varepsilon > 0$,

$$\mathrm{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \varepsilon\right) \to 0 \text{ as } n \to \infty. \tag{2.1}$$

By Chebyshov's inequality below, we have

$$\begin{aligned}\mathrm{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \varepsilon\right) &\le \frac{1}{\varepsilon^2}\mathrm{E}\left[\left(\frac{1}{n}S_n - \mu\right)^2\right]\\ &= \frac{1}{\varepsilon^2}\mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right)^2\right]\\ &= \frac{1}{n^2\varepsilon^2}\mathrm{E}\left[\left(\sum_{i=1}^n (X_i - \mu)\right)^2\right].\end{aligned}$$

By the independence of $(X_n; n \ge 1)$, we can obtain that

$$\begin{aligned}\mathrm{E}\left[\left(\sum_{i=1}^n (X_i - \mu)\right)^2\right] &= \mathrm{E}\left[\sum_{i=1}^n\sum_{j=1}^n (X_i - \mu)(X_j - \mu)\right]\\ &= \mathrm{E}\left[\sum_{i=1}^n (X_i - \mu)^2\right] + \mathrm{E}\left[\sum_{i=1}^n \sum_{j\ne i, j=1}^n (X_i - \mu)(X_j - \mu)\right]\\ &= \mathrm{E}\left[\sum_{i=1}^n (X_i - \mu)^2\right] \quad \text{(by Property 1.11(iv))}\\ &= n\sigma^2.\end{aligned}$$

Thus, it follows that

$$\mathrm{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \varepsilon\right) \le \frac{n\sigma^2}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ as } n \to \infty,$$

which implies that (2.1) holds. This completes the proof. □

Property 2.4. *(Chebyshov's inequality) Suppose that* $\mathrm{E}(X^2) < \infty$*. Then, for any constant* $a > 0$*,*

$$\mathrm{P}(|X| \ge a) \le \frac{\mathrm{E}(X^2)}{a^2}.$$

Proof. This is left as an excise. □

Property 2.5. *If* $X_n \to_p \mu$ *and* $Y_n \to_p \nu$*, then (i)* $X_n + Y_n \to_p \mu + \nu$*; (ii)* $X_nY_n \to_p \mu\nu$*; (iii)* $X_n/Y_n \to \mu/\nu$ *if* $Y_n \ne 0$ *and* $\nu \ne 0$*; (iv)* $g(X_n) \to_p g(\mu)$ *for a continuous function* $g(\cdot)$*.*

Proof. The proof is omitted. □

Example 2.9. Let $(X_i; i \ge 1)$ be a sequence of independent random variables having the same finite mean $\mu = \mathrm{E}(X_1)$, finite variance $\sigma^2 = \mathrm{Var}(X_1)$, and finite fourth moment $\mu_4 = E(X_1^4)$. Show that

$$S^2 \to_p \mathrm{Var}(X_1).$$

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

Definition 2.3. *(Convergence in distribution) Let $(Z_n; n \geq 1)$ be a sequence of random variables. We say the sequence Z_n converges in distribution to Z if, as $n \to \infty$,*

$$G_n(x) \to G(x) \quad \textit{whereever } G(x) \textit{ is continuous.}$$

Here, $G_n(x)$ and $G(x)$ are the c.d.f. of Z_n and Z, respectively.

Theorem 2.2. *(Central limit theorem (CLT)) Let $(X_i; i \geq 1)$ be a sequence of independent random variables having the same finite mean and variance, $\mu = \mathrm{E}(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1)$. Then, as $n \to \infty$,*

$$\frac{\overline{X} - \mathrm{E}(\overline{X})}{\sqrt{\mathrm{Var}(\overline{X})}} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \to_d \mathrm{N}(0,1).$$

Central limit theorem shows that $\overline{X} \sim \mathrm{N}\big(\mathrm{E}(\overline{X}), \mathrm{Var}(\overline{X})\big)$, and hence it tells us the distribution of $\overline{X}$ when the sample size n is large. Next, we check the performance of central limit theorem by simulations:

(1) Generate a realization $\{x_1, x_2, \cdots, x_n\}$ of the independent random sample $\{X_1, X_2, \cdots, X_n\}$ from $\mathrm{N}(0,1)$;
(2) Calculate $z_n = \sqrt{n}(\bar{x} - \mu)/\sigma$ with $\mu = 0$ and $\sigma = 1$;
(3) Repeat (1)-(2) J times to get $\{z_n^{(1)}, z_n^{(2)}, \cdots, z_n^{(J)}\}$, which is a sequence of realizations of Z_n, where $Z_n = \sqrt{n}(\overline{X} - \mu)/\sigma$;
(4) Plot the (relative frequency) histogram of $\{z_n^{(1)}, z_n^{(2)}, \cdots, z_n^{(J)}\}$.

Fig. 2.1 plots the histogram of $\{z_n^{(1)}, z_n^{(2)}, \cdots, z_n^{(J)}\}$ for $n = 10, 50$, and 1000 with the repetition time $J = 10000$. From this figure, the histogram of $\{z_n^{(1)}, z_n^{(2)}, \cdots, z_n^{(J)}\}$ is very close to the density of $\mathrm{N}(0,1)$, especially for large n. Hence, this implies that the approximation in central limit theorem performs very well.

To prove the above central limit theorem, we need the following lemma:

Lemma 2.1. *If*

1. *$M_{Z_n}(t)$, the moment generating function of Z_n, exists, $n = 1, 2, \ldots$,*
2. *$\lim_{n\to\infty} M_{Z_n}(t)$ exists and equals the moment generating function of a random variable Z,*

then

$$\lim_{n\to\infty} G_{Z_n}(x) = G_Z(x) \qquad \textit{for all } x \textit{ at which } G_Z(x) \textit{ is continuous,}$$

where $G_{Z_n}(x)$ is the c.d.f. of Z_n, $n = 1, 2, \ldots$, and $G_Z(x)$ is the c.d.f. of Z.

Proof. As $(X_i; i \geq 1)$ is a sequence of independent random variables having the same finite mean $\mu = \mathrm{E}(X_1)$ and finite variance $\sigma^2 = \mathrm{Var}(X_1)$, simple algebra gives us that

$$\mathrm{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}(X_i) = \mu \text{ and } \mathrm{Var}(\overline{X}) = \mathrm{E}[(\overline{X})^2] - [\mathrm{E}(\overline{X})]^2 = \frac{\sigma^2}{n}.$$

Hence, it suffices to show that

$$Z_n = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \to_d \mathrm{N}(0,1). \tag{2.2}$$

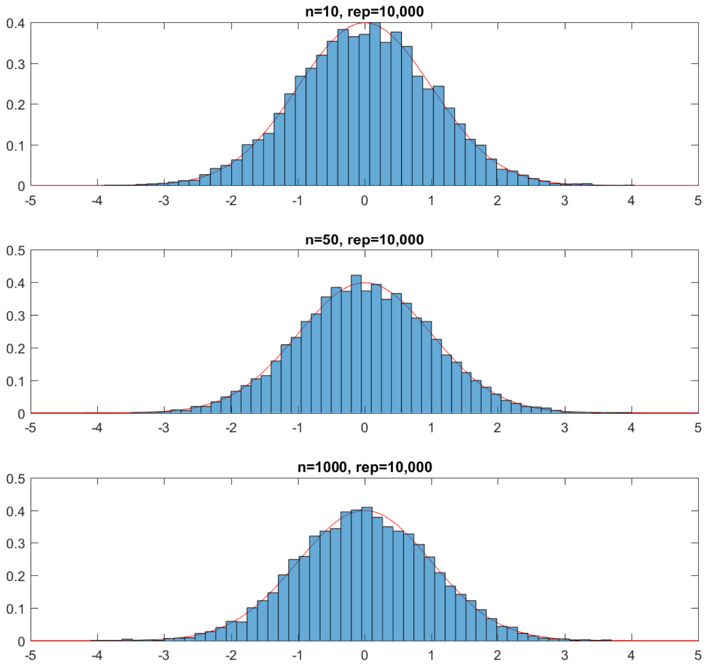

Fig. 2.1 The histogram of $\{z_n^{(1)}, z_n^{(2)}, \cdots, z_n^{(J)}\}$ with $J = 10000$ and $n = 10$ (top), $n = 50$ (middle) or $n = 1000$ (Bottom). The red line is the density of $\mathrm{N}(0,1)$.

A heuristic proof for (2.2) is based on Lemma 2.1. Let $Y_i = \dfrac{X_i - \mu}{\sigma}$, then $\mathrm{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = 1$. Suppose the moment generating function $M_{Y_i}(t)$ exists. A Taylor's expansion of $M_{Y_i}(t)$ around 0 gives:

$$M_{Y_i}(t) = M_{Y_i}(0) + tM_{Y_i}^{(1)}(0) + \frac{t^2}{2}M_{Y_i}^{(2)}(\varepsilon), \qquad \text{for some } 0 \le \varepsilon \le t.$$

Since $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i$, then the moment generating function of Z_n is thus given by

$$\begin{aligned}
M_{Z_n}(t) &= \prod_{i=1}^n M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) \\
&= \left[M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[M_{Y_i}(0) + \frac{t}{\sqrt{n}}M_{Y_i}^{(1)}(0) + \frac{(t/\sqrt{n})^2}{2}M_{Y_i}^{(2)}(\varepsilon)\right]^n \\
&= \left[1 + \frac{t}{\sqrt{n}}\mathrm{E}(Y_i) + \frac{t^2}{2n}M_{Y_i}^{(2)}(\varepsilon)\right]^n \\
&= \left[1 + \frac{t^2}{2n}M_{Y_i}^{(2)}(\varepsilon)\right]^n,
\end{aligned}$$

where $0 \le \varepsilon \le t/\sqrt{n}$. As $n \to \infty$, $\varepsilon \to 0$ and $M_{Y_i}^{(2)}(\varepsilon) \to M_{Y_i}^{(2)}(0) = \mathrm{E}(Y_i^2) = 1$. Hence,

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1+\frac{t^2}{2n}\right)^n = \exp\left(\frac{t^2}{2}\right) = \exp\left(0\times t + \frac{1}{2}\times 1 \times t^2\right),$$

which is the moment generating function of $\mathrm{N}(0,1)$ random variable. Hence, the conclusion follows directly from Lemma 2.1. □

Example 2.10. Suppose that $Y \sim \chi^2(50)$. Approximate $\mathrm{P}(40 < Y < 60)$.

Solution. By Example 2.8, $Y \sim \sum_{i=1}^{50} X_i^2$, where $X_1, X_2, \cdots, X_{50}$ are independent $\mathrm{N}(0,1)$ random variables. Let $\overline{X} = \frac{1}{50}\sum_{i=1}^{50} X_i^2$. Hence,

$$\begin{aligned}
\mathrm{P}(40 < Y < 60) &= \mathrm{P}(40 < 50\overline{X} < 60)\\
&= \mathrm{P}(4/5 < \overline{X} < 6/5)\\
&= \mathrm{P}\left(\frac{\sqrt{50}(4/5-\mu)}{\sigma} < \frac{\sqrt{50}(\overline{X}-\mu)}{\sigma} < \frac{\sqrt{50}(6/5-\mu)}{\sigma}\right)\\
&\approx \Phi(\frac{\sqrt{50}(6/5-\mu)}{\sigma}) - \Phi(\frac{\sqrt{50}(4/5-\mu)}{\sigma}) \ \text{(by CLT)}\\
&= \Phi(1) - \Phi(-1) \approx 0.68,
\end{aligned}$$

where $\mu = \mathrm{E}X_i^2 = 1$, $\sigma^2 = \mathrm{Var}X_i^2 = 2$, and $\Phi(\cdot)$ is the c.d.f. of $\mathrm{N}(0,1)$. □

2.3 Resampling

Suppose $\{X_1, \cdots, X_n\}$ be a random sample from one population with an unknown c.d.f. $F(\cdot)$. Let $\{x_1, \cdots, x_n\}$ be one realization of $\{X_1, \cdots, X_n\}$. Based on $\{x_1, \cdots, x_n\}$, we have a realization of the empirical distribution:

$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathrm{I}(x_k \le x).$$

By Theorem 1.4,

$$F(x) \approx F_n(x). \tag{2.3}$$

Since $F_n(x)$ is a discrete c.d.f, we can draw a random sample $\{X_1^*, X_2^*, \cdots, X_B^*\}$ from $F_n(x)$, and it is expected that the (relative frequency) histogram of $\{X_1^*, X_2^*, \cdots, X_B^*\}$ should be close to $f(x)$. Here, $X_i^* \sim X^* \sim F_n(x)$ is a discrete random variable such that

$$\mathrm{P}(X^* = x_j) = \frac{1}{n} \text{ for } j = 1, 2, \cdots, n.$$

Conventionally, $\{X_1^*, X_2^*, \cdots, X_B^*\}$ is called the bootstrap (resampling) random sample, and B is the bootstrap sample size.

Example 2.11. Let $\{x_i\}_{i=1}^{200}$ be a realization from $\mathrm{N}(0,1)$. Fig. 2.2 plots the histogram of original realization $\{x_i\}_{i=1}^{200}$ and bootstrapped realizations $\{x_i^*\}_{i=1}^{100}$, $\{x_i^*\}_{i=1}^{200}$, and $\{x_i^*\}_{i=1}^{500}$.

From this figure, we can see that the distribution of the bootstrapped realizations is very close to the distribution of $\mathrm{N}(0,1)$, especially for large B. □

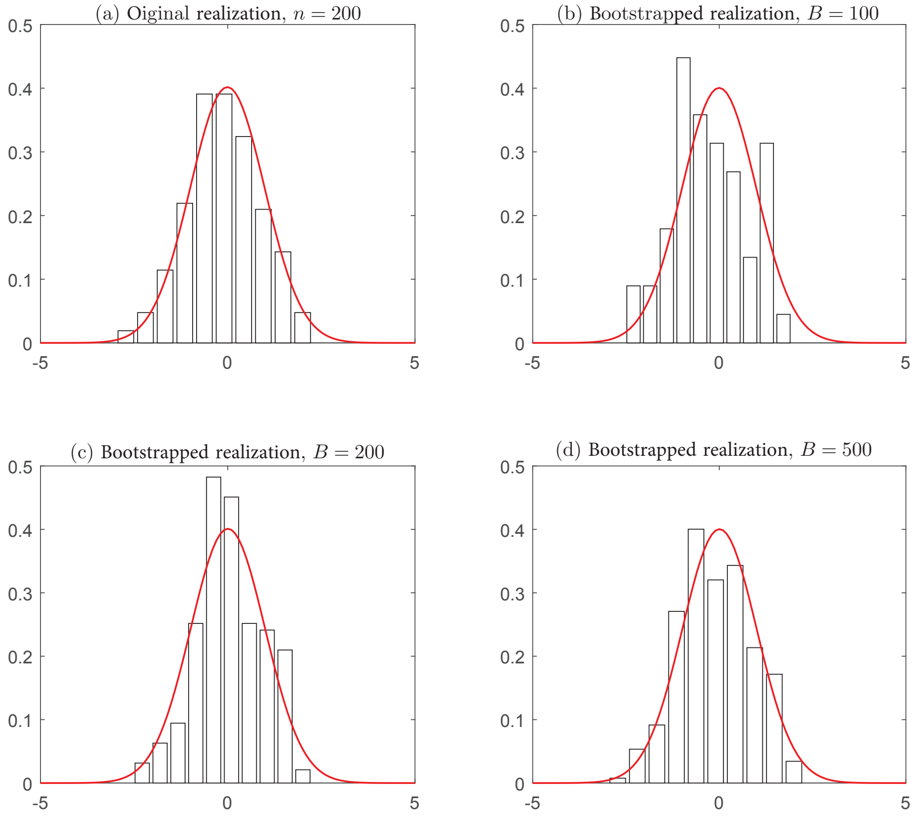

Fig. 2.2 The red line is the p.d.f of $\mathrm{N}(0,1)$. (a) The histogram of original realization $\{x_i\}_{i=1}^{200}$; (b) The histogram of one bootstrapped realization $\{x_i^*\}_{i=1}^{100}$; (c) The histogram of one bootstrapped realization $\{x_i^*\}_{i=1}^{200}$; (d) The histogram of one bootstrapped realization $\{x_i^*\}_{i=1}^{500}$.

Finally, we consider how to use bootstrap method to approximate the distribution of a statistic $T = g(X_1, X_2, \cdots, X_n)$, where $g(\cdot)$ is a given functional. In practice, a direct calculation of the distribution of T is usually infeasible. But in view of (2.3), the distribution of T is close to the distribution of T^*, where $T^* = g(X_1^*, X_2^*, \cdots, X_n^*)$. Hence, it motives us to use the following procedure to approximate the distribution of T:

(1) Generate a bootstrapped realization $\{x_1^*, x_2^*, \cdots, x_n^*\}$ from the distribution $F_n(\cdot)$;
(2) Calculate $t^* = g(x_1^*, x_2^*, \cdots, x_n^*)$, which is a realization of T^*;
(3) Repeat (1)-(2) J times to get $\{t^{*(1)}, t^{*(2)}, \cdots, t^{*(J)}\}$, which is a sequence of realizations of T^*;
(4) Plot the (relative frequency) histogram of $\{t^{*(1)}, t^{*(2)}, \cdots, t^{*(J)}\}$.

Since the histogram of $\{t^{*(1)}, t^{*(2)}, \cdots, t^{*(J)}\}$ is close to the p.d.f. of T^*, it is also close to the p.d.f. of T. Clearly, this bootstrap method provides us an easy way to calculate the percentile of the distribution of T, which is important in many applications.

Chapter 3
Point estimation

In many applications, a random variable X resulting from a random experiment is assumed to have a certain distribution with the p.d.f. $f(x;\theta)$, where $\theta \in \mathscr{R}^s$ is a unknown parameter taking value in a set Ω, and Ω is usually called the parameter space. For example, X is often assumed to follow a normal distribution $N(\mu,\sigma^2)$; in this case $\theta = (\mu,\sigma) \in \Omega$ is the unknown parameter, and the parameter space $\Omega = \{(\mu,\sigma) : \mu \in \mathscr{R}, \sigma > 0\}$. For the experimenter, the important question is how to find a "good" estimator for the unknown parameter θ. Heuristically, a random sample $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$, which forms an empirical distribution, can elicit information about the distribution of X. Hence, it is natural to expect that we can estimate the unknown parameter θ based on a random sample $\mathbf{X}$.

3.1 Maximum likelihood estimator

We first consider the maximum likelihood estimator, which is motivated by a simple example below.

Example 3.1. Suppose that X follows a Bernoulli distribution so that the p.d.f. of X is

$$f(x;p) = p^x(1-p)^{1-x}, \;\; x = 0, 1,$$

where the unknown parameter $p \in \Omega$ with $\Omega = \{p : p \in (0,1)\}$. Further, assume that we have a random sample $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ with the observable values $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, respectively. Then, the probability that $\mathbf{X} = \mathbf{x}$ is

$$\begin{aligned} L(x_1, \cdots, x_n; p) &= \mathrm{P}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}, \end{aligned}$$

which is the joint p.d.f. of $X_1, X_2, \cdots, X_n$ evaluated at the observed values. The joint p.d.f. is a function of p. Then, we want to find the value of p that maximizes this joint p.d.f., or equivalently, we want to find p_* such that

$$p_* = \arg\max_{p \in \Omega} L(x_1, \cdots, x_n; p).$$

The way to propose p_* is reasonable because p_* most likely has produced the sample values $x_1, \cdots, x_n$. We call p^* the maximum likelihood estimate, since "likelihood" is often used as a synonym for "probability" in informal contexts.

Conventionally, we denote $L(p) = L(x_1, \cdots, x_n; p)$, and p_* is easier to be computed by

$$p_* = \arg\max_{p \in \Omega} \log L(p).$$

[Note that p maximizes $\log L(p)$ also maximizes $L(p)$]. By simple algebra (see one example below), we can show that

$$p_* = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

which maximizes $\log L(p)$. The corresponding statistic, namely $n^{-1}\sum_{i=1}^{n} X_i$, is called the maximum likelihood estimator (MLE) of p; that is,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that $\mathrm{E}(\hat{p} - p)^2 \to 0$ as $n \to \infty$. Thus, $\hat{p}$ is a "good" estimator of p in some sense.

Definition 3.1. *(Likelihood Function) Let* $\mathbf{X}$ *be a random sample with a joint p.d.f.* $\mathbf{f}(x_1, \cdots, x_n; \theta)$*, where the parameter* θ *is within a certain parameter space* Ω*. Then, the likelihood function of this random sample is defined as*

$$L(\theta) = \mathbf{f}(\mathbf{X}; \theta) = \mathbf{f}(X_1, X_2, \cdots, X_n; \theta)$$

for $\theta \in \Omega$*. Moreover,* $\ell(\theta) = \log L(\theta)$ *is called the log-likelihood function.*

Definition 3.2. *(Maximum Likelihood Estimator) Given a likelihood function* $L(\theta)$ *for* $\theta \in \Omega$*, the maximum likelihood estimator (MLE) of* θ *is defined as*

$$\hat{\theta} = \arg\max_{\theta \in \Omega} L(\theta) = \arg\max_{\theta \in \Omega} \ell(\theta).$$

Definition 3.3. *(Maximum Likelihood Estimate) The observed value of* $\hat{\theta}$ *is called the maximum likelihood estimate.*

Example 3.2. Let $\mathbf{X}$ be an independent random sample from a Bernoulli distribution with parameter p with $0 < p < 1$. Find the maximum likelihood estimator of p.

Solution. For the random sample $\mathbf{X}$, the likelihood function is

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

Hence, the log-likelihood function is

$$\begin{aligned} \ell(p) = \log L(p) &= \sum_{i=1}^{n} [X_i \log p + (1 - X_i)\log(1-p)] \\ &= \log p \cdot \sum_{i=1}^{n} X_i + \log(1-p) \cdot \sum_{i=1}^{n} (1 - X_i). \end{aligned}$$

Note that

$$\begin{aligned}\frac{\mathrm{d}\ell(p)}{\mathrm{d}p} &= \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} X_i\right)\\ &= \frac{(1-p)\sum_{i=1}^{n} X_i - np + p\sum_{i=1}^{n} X_i}{p(1-p)}\\ &= \frac{n(\overline{X}-p)}{p(1-p)}.\end{aligned}$$

It is not hard to see that

$$p < \overline{X} \iff \frac{\mathrm{d}\ell(p)}{\mathrm{d}p} > 0$$

and

$$p > \overline{X} \iff \frac{\mathrm{d}\ell(p)}{\mathrm{d}p} < 0,$$

which imply that $\ell(p)$ attains its maximum at $p = \overline{X}$. Therefore, the maximum likelihood estimator of p is $\overline{X}$. □

Example 3.3. Let **X** be an independent random sample from a uniformly distribution over the interval $[0,\beta]$. Find the maximum likelihood estimator of β.

Solution. Note that a uniformly distribution over the interval $[0,\beta]$ has the p.d.f. given by

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

For the random sample **X**, the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta}\mathrm{I}(0 \le X_i \le \beta) = \frac{1}{\beta^n}\prod_{i=1}^{n}\mathrm{I}(0 \le X_i \le \beta).$$

In order that $L(\beta)$ attains its maximum, β must satisfy

$$0 \le X_i \le \beta,\ i = 1,2,\ldots,n.$$

Since $\frac{1}{\beta^n}$ increases as β decreases, we must select β to be as small as possible subject to the previous constraint. Therefore, the maximum of $L(\beta)$ should be selected to be the maximum of $X_1, X_2, \ldots, X_n$, that is, the maximum likelihood estimator $\hat{\beta} = X_{(n)} = \max_{1 \le i \le n} X_i$. □

Example 3.4. Let **X** be an independent random sample from $\mathrm{N}(\theta_1,\theta_2)$, where $(\theta_1,\theta_2) \in \Omega$ and $\Omega = \{(\theta_1,\theta_2) : \theta_1 \in \mathscr{R}, \theta_2 > 0\}$. Find the MLEs of θ_1 and θ_2. [Note: here we let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$].

Solution. Let $\theta = (\theta_1,\theta_2)$. For the random sample **X**, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\frac{(X_i-\theta_1)^2}{2\theta_2}\right].$$

Then, the log-likelihood function is

$$\ell(\theta) = \log L(\theta) = -\frac{n}{2}\log(2\pi\theta_2) - \frac{\sum_{i=1}^n (X_i - \theta_1)^2}{2\theta_2}.$$

As the MLE is the maximizer of $\ell(\theta)$, it should satisfy that

$$0 = \frac{\partial \ell(\theta)}{\partial \theta_1} = \frac{1}{\theta_2}\sum_{i=1}^n (X_i - \theta_1),$$

$$0 = \frac{\partial \ell(\theta)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^n (X_i - \theta_1)^2.$$

Solving the two equations above, we obtain that

$$\theta_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \theta_2 = S^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

By considering the usual condition on the second partial derivatives, these solutions do provide a maximum. Thus, the MLEs of θ_1 and θ_2 are

$$\hat{\theta}_1 = \overline{X} \quad \text{and} \quad \hat{\theta}_2 = S^2,$$

respectively. □

3.2 Method of moments estimator

The method of moments estimator is often used in practice, especially when we do not know the full information about X except for its certain moments. Recall that the r-th moment about the origin of X is defined as $\mu_r = \mathrm{E}X^r$. In many situations, μ_r contains the information about the unknown parameter θ. For example, if $X \sim \mathrm{N}(\mu, \sigma^2)$, we know that

$$\mu_1 = \mathrm{E}X = \mu \quad \text{and} \quad \mu_2 = \mathrm{E}X^2 = \mathrm{Var}(X) + [\mathrm{E}(X)]^2 = \sigma^2 + \mu^2.$$

or

$$\mu = \mu_1 \quad \text{and} \quad \sigma^2 = \mu_2 - \mu_1^2.$$

That is, the unknown parameters μ and σ^2 can be estimated if we find "good" estimators for μ_1 and μ_2. Note that by the weak law of large numbers (see Theorem 2.1),

$$m_1 = \frac{1}{n}\sum_{i=1}^n X_i \to_p \mathrm{E}(X) = \mu_1 \quad \text{and} \quad m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \to_p \mathrm{E}(X^2) = \mu_2.$$

Thus, it is reasonable to estimate μ and σ^2 by m_1 and $m_2 - m_1^2$, respectively, and these estimators are called the method of moments estimator.

To generalize the idea above, we assume that the unknown parameter $\theta \in \mathscr{R}^s$ can be expressed by

$$\theta = h(\mu_1, \mu_2, \cdots, \mu_k), \tag{3.1}$$

where $h : \mathscr{R}^k \to \mathscr{R}^s$. For the illustrating example above, $s = 2$, $k = 2$, $\theta = (\mu, \sigma^2)$ and $h = (h_1, h_2)$ with

$$h_1(\mu_1, \mu_2) = \mu_1 \text{ and } h_2(\mu_1, \mu_2) = \mu_2 - \mu_1^2.$$

Define the r-th sample moment of a random sample $\mathbf{X}$ by

$$m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r, \qquad r = 1, 2, \ldots.$$

Unlike μ_r, m_r always exists for any positive integer r. In view of (3.1), the **method of moments estimator** (MME) $\tilde{\theta}$ of θ is defined by

$$\tilde{\theta} = h(m_1, m_2, \cdots, m_k),$$

and the observed value of $\tilde{\theta}$ is called the method of moments estimate.

Example 3.5. Let $\mathbf{X}$ be an independent random sample from a gamma distribution with the p.d.f. given by

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & \text{for } x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Find a MME of (α, λ).

Solution. Some simple algebra shows that the first two moments are

$$\mu_1 = \frac{\alpha}{\lambda} \text{ and } \mu_2 = \frac{\alpha^2 + \alpha}{\lambda^2}.$$

[Note: μ_1 and μ_2 can be obtained from Example 2.5.] Substituting $\alpha = \lambda\mu_1$ in the second equation, we get

$$\mu_2 = \frac{(\lambda\mu_1)^2 + \lambda\mu_1}{\lambda^2} = \mu_1^2 + \frac{\mu_1}{\lambda} \text{ or } \lambda = \frac{\mu_1}{\mu_2 - \mu_1^2},$$

which implies that

$$\alpha = \lambda\mu_1 = \frac{(\mu_1)^2}{\mu_2 - (\mu_1)^2}.$$

Therefore, one MME of (α, λ) is $(\tilde{\alpha}, \tilde{\lambda})$, where

$$\tilde{\alpha} = \frac{m_1^2}{m_2 - m_1^2} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2} \text{ and } \tilde{\lambda} = \frac{m_1}{m_2 - m_1^2} = \frac{\overline{X}}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2}.$$

□

It is worth noting that the way to construct h in (3.1) is not unique. Usually, we use the lowest possible order moments to construct f, although this is may not be the optimal way. To consider the "optimal" MME, one may refer to the generalized method of moments estimator for a further reading.

3.3 Estimator properties

For the same unknown parameter θ, many different estimators may be obtained. Heuristically, some estimators are good and others bad. The question is how would we establish a criterion of goodness to compare one estimator with another? The particular properties of estimators that we will discuss below are unbiasedness, efficiency, and consistency.

3.3.1 Unbiasedness

Suppose that $\hat{\theta}$ is an estimator of θ. If $\hat{\theta}$ is a good estimator of θ, a fairly desirable property is that its mean be equal to θ, namely, $\mathrm{E}(\hat{\theta}) = \theta$. That is, in practice, we would want $\mathrm{E}(\hat{\theta})$ to be reasonably close to θ.

Definition 3.4. *(Unbiased estimator) The bias of an estimator $\hat{\theta}$ is defined as*

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

If $\mathrm{Bias}(\hat{\theta}) = 0$*, $\hat{\theta}$ is called an unbiased estimator of θ. Otherwise, it is said to be biased.*

Definition 3.5. *(Asymptotically unbiased estimator) $\hat{\theta}$ is an asymptotically unbiased estimator if*

$$\lim_{n\to\infty} \mathrm{Bias}(\hat{\theta}) = \lim_{n\to\infty} [\mathrm{E}(\hat{\theta}) - \theta] = 0,$$

where n is the sample size.

Example 3.3. (con't) (i) Show that $\hat{\beta} = X_{(n)}$ is an asymptotically unbiased estimator of β; (ii) modify this estimator of β to make it unbiased.

Solution. Let $Y = X_{(n)}$. For $0 \le y \le \beta$,

$$\mathrm{P}(Y \le y) = \prod_{i=1}^{n} \mathrm{P}(X_i \le y) = \left(\frac{y}{\beta}\right)^n.$$

Thus, by Property 1.10,

$$\begin{aligned}\mathrm{E}(Y) &= \int_0^\infty \mathrm{P}(Y > y)\mathrm{d}y = \int_0^\beta \left[1 - \left(\frac{y}{\beta}\right)^n\right] \mathrm{d}y = \beta - \frac{\beta^{n+1}}{(n+1)\beta^n} \\ &= \frac{n\beta}{n+1} \to \beta \text{ as } n \to \infty,\end{aligned}$$

which implies that Y is an asymptotically unbiased estimator of β.

Furthermore, since $\mathrm{E}\left(\frac{n+1}{n}Y\right) = \beta$, we know that $\frac{n+1}{n}Y$ is an unbiased estimator of β. □

Example 3.4. (con't) Show that $\overline{X}$ is an unbiased estimator of θ_1, and S^2 is an asymptotically unbiased estimator of θ_2.

Solution. As each $X_i \sim \mathrm{N}(\theta_1, \theta_2)$, we have

$$\mathrm{E}(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\theta_1 = \theta_1.$$

Hence, $\overline{X}$ is an unbiased estimator of θ_1.

Next, it is easy to see that

$$\mathrm{E}(S^2) = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\overline{X})^2\right].$$

Note that for each i, we have

$$\begin{aligned}
\mathrm{E}\left[(X_i-\overline{X})^2\right] &= \mathrm{E}\left[(X_1-\overline{X})^2\right]\\
&= \mathrm{Var}(X_1-\overline{X}) \qquad (\text{since } \mathrm{E}(X_1-\overline{X})=0)\\
&= \mathrm{Var}\left(X_1 - \frac{X_1+X_2+\cdots+X_n}{n}\right)\\
&= \mathrm{Var}\left(\frac{(n-1)X_1}{n} - \sum_{i=2}^{n}\frac{X_i}{n}\right)\\
&= \frac{(n-1)^2}{n^2}\theta_2 + \frac{(n-1)}{n^2}\theta_2 \qquad (\text{by the independence among } X_i\text{'s})\\
&= \frac{n-1}{n}\theta_2.
\end{aligned}$$

Hence, it follows that

$$\mathrm{E}(S^2) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\overline{X})^2\right] = \frac{n-1}{n}\theta_2.$$

As $\lim_{n\to\infty}\mathrm{E}(S^2) = \theta_2$, S^2 is an asymptotically unbiased estimator of θ_2. □

3.3.2 Efficiency

Suppose that we have two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$. The question is how to compare $\hat{\theta}$ and $\tilde{\theta}$ in terms of a certain criterion. To answer this question, we first introduce the so-called mean squared error of a given estimator $\hat{\theta}$.

Definition 3.6. *(Mean squared error) Suppose that $\hat{\theta}$ is an estimator of θ. The mean squared error of $\hat{\theta}$ is*

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

For a given estimator $\hat{\theta}$, MSE$(\hat{\theta})$ is the mean (expected) value of the square of the error (difference) $\hat{\theta}-\theta$. This criterion can be decomposed by two parts as shown below.

Property 3.1. *If Var$(\hat{\theta})$ exists, then the mean squared error of $\hat{\theta}$ is*

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2.$$

Proof.

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}\left[(\hat{\theta}-\theta)^2\right] \\
&= \mathrm{E}\left(\left\{[\hat{\theta}-\mathrm{E}(\hat{\theta})]+[\mathrm{E}(\hat{\theta})-\theta]\right\}^2\right) \\
&= \mathrm{E}\left([\hat{\theta}-\mathrm{E}(\hat{\theta})]^2+2[\hat{\theta}-\mathrm{E}(\hat{\theta})][\mathrm{E}(\hat{\theta})-\theta]+[\mathrm{E}(\hat{\theta})-\theta]^2\right) \\
&= \mathrm{Var}(\hat{\theta})+2\mathrm{E}[\hat{\theta}-\mathrm{E}(\hat{\theta})][\mathrm{E}(\hat{\theta})-\theta]+[\mathrm{Bias}(\hat{\theta})]^2 \\
&= \mathrm{Var}(\hat{\theta})+[\mathrm{Bias}(\hat{\theta})]^2.
\end{aligned}$$

□

Remark 3.1. The following result is straightforward:

$$\lim_{n\to\infty}\mathrm{MSE}(\hat{\theta})=0 \Longleftrightarrow \lim_{n\to\infty}\mathrm{Var}(\hat{\theta})=0 \text{ and } \lim_{n\to\infty}\mathrm{Bias}(\hat{\theta})=0.$$

Heuristically, one wants $\mathrm{MSE}(\hat{\theta})$ as small as possible. As we discussed above, the unbiasedness is the desirable property for a certain estimator $\hat{\theta}$. Thus, it is reasonable to restrict our attention to only the unbiased estimator $\hat{\theta}$. In this case,

$$\mathrm{MSE}(\hat{\theta})=\mathrm{Var}(\hat{\theta})$$

by Property 3.1. Now, for two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$, we only need to select the one with a smaller variance, and this motivates us to define the efficiency between $\hat{\theta}$ and $\tilde{\theta}$.

Definition 3.7. *(Efficiency) Suppose that $\hat{\theta}$ and $\tilde{\theta}$ are two unbiased estimators of θ. The efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined by*

$$Eff(\hat{\theta},\tilde{\theta})=\frac{Var(\tilde{\theta})}{Var(\hat{\theta})}.$$

If $Eff(\hat{\theta},\tilde{\theta})>1$, then we say that $\hat{\theta}$ is relatively more efficient than $\tilde{\theta}$.

Example 3.6. Let $(X_n;n\geq 1)$ be a sequence of independent random variables having the same finite mean and variance, $\mu=\mathrm{E}(X_1)$ and $\sigma^2=\mathrm{Var}(X_1)$. We can show that $\overline{X}$ is an unbiased estimator of μ, and $\mathrm{Var}(\overline{X})=\frac{\sigma^2}{n}$. Suppose that we now take two samples, one of size n_1 and one of size n_2, and denote the sample means as $\overline{X}^{(1)}$ and $\overline{X}^{(2)}$, respectively. Then,

$$\mathrm{Eff}\,(\overline{X}^{(1)},\overline{X}^{(2)})=\frac{\mathrm{Var}(\overline{X}^{(2)})}{\mathrm{Var}(\overline{X}^{(1)})}=\frac{n_1}{n_2}.$$

Therefore, the larger is the sample size, the more efficient is the sample mean for estimating μ.

Example 3.3. (con't) Note that $\dfrac{n+1}{n}X_{(n)}$ is an unbiased estimator of β. Show that
(i) $2\overline{X}$ is also an unbiased estimator of β;
(ii) Compare the efficiency of these two estimators of β.

Solution. (i) Since $\mathrm{E}(\overline{X})$ equals the population mean, which is $\beta/2$, $\mathrm{E}(2\overline{X})=\beta$. Thus, $2\overline{X}$ is an unbiased estimator of β.

(ii) First we must find the variance of the two estimators. Recall that $Y = X_{(n)}$. Before, we have already obtained

$$\mathrm{P}(Y \leq y) = \left(\frac{y}{\beta}\right)^n \qquad \text{for } 0 \leq y \leq \beta.$$

Therefore, for $Z = Y^2$, it is not hard to show that

$$\mathrm{P}(Z \leq z) = \mathrm{P}(Y \leq \sqrt{z}) = \left(\frac{\sqrt{z}}{\beta}\right)^n \qquad \text{for } 0 \leq z \leq \beta^2.$$

By Property 1.10, we have

$$\begin{aligned}
\mathrm{E}(Z) &= \int_0^{\beta^2} 1 - \left(\frac{\sqrt{z}}{\beta}\right)^n dz \\
&= 2\int_0^{\beta} y\left(1 - \frac{y^n}{\beta^n}\right)\mathrm{d}y \qquad \text{(by setting } y = \sqrt{z}) \\
&= 2\int_0^{\beta} \left(y - \frac{y^{n+1}}{\beta^n}\right)\mathrm{d}y \\
&= 2\left[\frac{\beta^2}{2} - \frac{\beta^{n+2}}{(n+2)\beta^n}\right] \\
&= \frac{n}{n+2}\beta^2.
\end{aligned}$$

Or by a direct calculation, we can show that

$$\mathrm{E}(Z) = \int_0^{\beta} y^2 \mathrm{d}\frac{y^n}{\beta^n} = \frac{n}{\beta^n}\int_0^{\beta} y^{n+1}\mathrm{d}y = \frac{n}{\beta^n}\cdot\frac{\beta^{n+2}}{n+2} = \frac{n}{n+2}\beta^2.$$

Hence,

$$\begin{aligned}
\mathrm{Var}\left(\frac{n+1}{n}Y\right) &= \mathrm{E}\left[\left(\frac{n+1}{n}Y\right)^2\right] - \left[\mathrm{E}\left(\frac{n+1}{n}Y\right)\right]^2 \\
&= \left(\frac{n+1}{n}\right)^2 \frac{n}{n+2}\beta^2 - \beta^2 \\
&= \left(\frac{n^2+2n+1}{n^2+2n} - 1\right)\beta^2 \\
&= \frac{\beta^2}{n(n+2)}.
\end{aligned}$$

Since $\mathrm{Var}(X_i) = \beta^2/12$ for each i, we have

$$\mathrm{Var}(2\overline{X}) = 4\mathrm{Var}(\overline{X}) = 4\cdot\frac{\beta^2}{12n} = \frac{\beta^2}{3n}.$$

Therefore,

$$\mathrm{Eff}\left(\frac{n+1}{n}Y, 2\overline{X}\right) = \frac{\mathrm{Var}(2\overline{X})}{\mathrm{Var}\left(\frac{n+1}{n}Y\right)} = \frac{n+2}{3}.$$

Thus, it can be seen that, for $n > 1$, $\frac{n+1}{n}Y$ is more efficient than $2\overline{X}$, and for $n = 1$, $\frac{n+1}{n}Y$ and $2\overline{X}$ have the same efficiency. □

Now, we face two important questions:

(i) For a given unbiased estimator $\tilde{\theta}$, could we find another unbiased estimator $\tilde{\theta}_*$, which has a smaller variance than $\tilde{\theta}$?

(ii) Among all unbiased estimators, could we find the **uniformly minimum variance unbiased estimator** (UMVUE), that is,

$$\text{UMVUE} = \underset{\tilde{\theta}\text{ is unbiased}}{\arg\min}\ \text{Var}(\tilde{\theta}).$$

The first question tells us how to find a more efficient unbiased estimator from an initial unbiased estimator; the second question tells us the UMVUE is relatively more efficient than any other unbiased estimators (in other words, the UMVUE is the best unbiased estimator).

To answer the first question, we need make use of the sufficient statistic.

Definition 3.8. *(Sufficient statistic) Suppose that the random sample* $\mathbf{X}$ *has a joint p.d.f.* $\mathbf{f}(x_1,\cdots,x_n;\theta)$*, where* θ *is the unknown parameter. The statistic* $T := T(\mathbf{X})$ *is sufficient for* θ *if and only if*

$$\mathbf{f}(x_1,\cdots,x_n;\theta) = g\big(T(x_1,\cdots,x_n);\theta\big)h(x_1,\cdots,x_n),$$

where g *depends on* $x_1,\cdots,x_n$ *only through* $T(x_1,\cdots,x_n)$*, and* h *does not depend on* θ*.*

Remark 3.2. (i) Definition 3.8 is also called Factorization Theorem; (ii) Sufficient statistic is not unique:

$$\begin{aligned} &T = T(\mathbf{X}) \text{ is a sufficent statistic for } \theta \\ \Longleftrightarrow\ & v(T) = v\big(T(\mathbf{X})\big) \text{ is also a sufficient statistic for } \theta, \end{aligned}$$

where $v(\cdot)$ is an invertible function. For example, if T is a sufficient statistic for θ, then T^3 is also a sufficient statistic for θ, while T^2 is not a sufficient statistic for θ.

Example 3.7. Suppose that $\mathbf{X}$ is an independent random sample from a uniform distribution $U(\alpha,\beta)$. Find a sufficient statistic for (α,β).

Solution. Let $\theta = (\alpha,\beta)$. The joint p.d.f. of $\mathbf{X}$ is

$$\begin{aligned} \mathbf{f}(x_1,\cdots,x_n;\theta) &= \prod_{i=1}^{n}\left(\frac{1}{\beta-\alpha}\right)\mathrm{I}(\alpha \le x_i \le \beta) = \left(\frac{1}{\beta-\alpha}\right)^n \mathrm{I}(\alpha \le x_i \le \beta, \forall i = 1,\ldots,n) \\ &= \left(\frac{1}{\beta-\alpha}\right)^n \mathrm{I}\left(\alpha \le \min_{1\le i\le n} X_i\right)\mathrm{I}\left(\max_{1\le i\le n} X_i \le \beta\right). \end{aligned}$$

Hence, by Definition 3.8, we know that $\left(\min_{1\le i\le n} X_i, \max_{1\le i\le n} X_i\right)$ is a sufficient statistic for (α,β). □

Example 3.8. Suppose that $\mathbf{X}$ is an independent random sample from a normal distribution $\mathrm{N}(\mu,\sigma^2)$. Find a sufficient statistic for (μ,σ^2).

Solution. Let $\theta = (\mu, \sigma^2)$. The joint p.d.f. of $\mathbf{X}$ is

$$\begin{aligned}
&\mathbf{f}(x_1, \cdots, x_n; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{((x_i-\bar{x})-(\mu-\bar{x}))^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(\mu-\bar{x})^2 - 2\sum_{i=1}^{n}(x_i-\bar{x})(\mu-\bar{x})\right)\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\mu-\bar{x})^2\right)\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right) \exp\left(-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{n}{2\sigma^2}s^2\right) \exp\left(-\frac{n}{2\sigma^2}(\mu-\bar{x})^2\right).
\end{aligned}$$

Hence, by Definition 3.8, we know that $(\overline{X}, S^2)$ is a sufficient statistic for (μ, σ^2). □

When the random sample $\mathbf{X}$ is from an exponential family, the sufficient statistic for θ can be easily found.

Property 3.2. *Let* $\mathbf{X}$ *be an i.i.d. random sample from a p.d.f. having the form:*

$$f(x;\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{s} p_i(\theta)t_i(x)\right) \quad \textit{(expontial family)},$$

where $\theta = (\theta_1, \theta_2, \cdots, \theta_s) \in \Theta \subset \mathscr{R}^s$. *Then,*

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), \cdots, \sum_{j=1}^{n} t_s(X_j)\right)$$

is a sufficient statistic for θ.

Remark 3.3. The exponential family includes many of the most common distributions, e.g., (1) normal; (2) exponential; (3) gamma; (4) chi squared; (5) beta; (6) Bernoulli; (7) Poisson; (8) geometric; ...

Example 3.8. (con't) Note that

$$\begin{aligned}
f(x;\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2+\mu^2-2\mu x}{2\sigma^2}\right) \\
&= \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right] \exp\left(\frac{n\mu}{\sigma^2}\frac{x}{n} - \frac{n}{2\sigma^2}\frac{x^2}{n}\right),
\end{aligned}$$

which implies that $h(x)=1$, $c(\theta)=\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\right]$, $p_1(\theta)=\frac{n\mu}{\sigma^2}$, $p_2(\theta)=-\frac{n}{2\sigma^2}$, $t_1(x)=\frac{x}{n}$, and $t_2(x)=\frac{x^2}{n}$. By Property 3.2, we know that $(\overline{X},\overline{X^2})$ is a sufficient statistic for θ.
(Note that we have found both $(\overline{X},S^2)$ and $(\overline{X},\overline{X^2})$ are sufficient statistics for θ, and these two sufficient statistics have the following link: $(\overline{X},S^2)=v(\overline{X},\overline{X^2})$, where the mapping v defined by

$$(z_1,z_2)\xrightarrow{v}(z_1,z_2-z_1^2)$$

is invertible. Here, we have used the fact $S^2=\overline{X^2}-(\overline{X})^2$.) □

Theorem 3.1. *(Rao-Blackwell Theorem) Let $\tilde{\theta}$ be an unbiased estimator of θ with $E(\tilde{\theta}^2)<\infty$, and $T:=T(\mathbf{X})$ be a sufficient statistic for θ. Let $w(t)=E(\tilde{\theta}|T=t)$. Then, $\tilde{\theta}_*=w(T)$ is an unbiased estimator of θ and $Var(\tilde{\theta}_*)\leq Var(\tilde{\theta})$.*

The previous theorem shows that by using the sufficient statistic T, we can always get a better unbiased estimator (in terms of efficiency) from an initial unbiased estimator. Moreover, this theorem implies that

the UMVUE is a function of the sufficient statistic T.

Next, we turn to the second question on how to find the UMVUE. The following property tells us that the UMVUE is unique.

Property 3.3. *If $\hat{\theta}$ is the UMVUE of θ, then $\hat{\theta}$ is unique.*

Intuitively, how to find the UMVUE is not an easy task. Below, we offer two approaches towards this goal.

3.3.2.1 UMVUE: complete and sufficient statistic method

The first approach to find the UMVUE is based on a complete and sufficient statistic. We have already introduced the sufficient statistic. Below, we introduce the complete statistic.

Definition 3.9. *(Complete statistic) For a given random sample $\mathbf{X}$, $T:=T(\mathbf{X})$ is a complete statistic of θ if*

$$E[z(T)]=0 \text{ for all } \theta \text{ implies } P(z(T)=0)=1 \text{ for all } \theta.$$

When the random sample $\mathbf{X}$ is from an exponential family, the complete statistic (as the sufficient statistic in Property 3.2) for θ can be easily found.

Property 3.4. *Let $\mathbf{X}$ be an i.i.d. random sample from a p.d.f. having the form:*

$$f(x;\theta)=h(x)c(\theta)\exp\left(\sum_{i=1}^{s}p_i(\theta)t_i(x)\right) \quad \text{(expontial family)},$$

where $\theta=(\theta_1,\theta_2,\cdots,\theta_s)\in\Theta\subset\mathscr{R}^s$. Then,

$$T(\mathbf{X})=\left(\sum_{j=1}^{n}t_1(X_j),\sum_{j=1}^{n}t_2(X_j),\cdots,\sum_{j=1}^{n}t_s(X_j)\right)$$

is a complete statistic for θ as long as the parameter Θ contains an open set in $\mathscr{R}^s$.

From Properties 3.2 and 3.4, we know that if $\mathbf{X}$ is from an exponential family and the related parameter Θ contains an open set in $\mathscr{R}^s$, then

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \sum_{j=1}^{n} t_2(X_j), \cdots, \sum_{j=1}^{n} t_s(X_j) \right)$$

is a complete and sufficient statistic for θ. If $\mathbf{X}$ is not from an exponential family, we need to check whether a statistic is complete or sufficient for θ by definition.

Example 3.8. (con't) It is easy to see that $(\overline{X}, \overline{X^2})$ is also a complete statistic for θ. □

The following theorem tells us the relationship between the complete and sufficient statistic and the UMVUE.

Theorem 3.2. *Let $T := T(\mathbf{X})$ be a complete and sufficient statistic for θ, and $\phi(T)$ be any estimator based only on T. Then, $\phi(T)$ is the unique UMVUE of its expected value $E[\phi(T)]$.*

From the preceding theorem, we know that if T is a complete and sufficient statistic for θ and $\mathrm{E}[\phi_0(T)] = \theta$ for some functional $\phi_0(\cdot)$, then $\phi_0(T)$ is the UMVUE of θ. In other words, the procedure to find the UMVUE is as follows:

$$(i) \text{ Find a complete and sufficient statistic } T \text{ for } \theta; \tag{3.2}$$

$$(ii) \text{ Find a functional } \phi_0(\cdot) \text{ such that } \mathrm{E}[\phi_0(T)] = \theta. \tag{3.3}$$

By Theorem 3.2, the unbiased estimator $\phi_0(T)$ in (3.3) is the UMVUE of θ.

Example 3.8. (con't) We have shown that $(\overline{X}, \overline{X^2})$ is a complete and sufficient statistic for θ. Next, since

$$\mathrm{E}(\overline{X}) = \mu, \ \mathrm{E}\left(\frac{n}{n-1}S^2\right) = \sigma^2 \ \text{ and } \ \frac{n}{n-1}S^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right),$$

by Theorem 3.2, we know that $\overline{X}$ is the UMVUE of μ, and $\frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the UMVUE of σ^2. □

Example 3.9. Let $\mathbf{X}$ be an i.i.d. random sample from Poisson distribution with parameter λ. Find the UMVUE of λ.

Solution. The p.d.f. of the Poisson distribution is

$$\frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-\lambda}}{x!}e^{\log(\lambda^x)} = \frac{e^{-\lambda}}{x!}e^{x\log(\lambda)} = \frac{e^{-\lambda}}{x!}e^{\frac{x}{n}n\log(\lambda)}. \tag{3.4}$$

Hence, by Properties 3.2 and 3.4, we know that

$$T(\mathbf{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a complete and sufficient statistic for λ. Moreover, since $\mathrm{E}(\overline{X}) = \lambda$, by Theorem 3.2, we know that $\overline{X}$ is the UMVUE of λ. □

3.3.2.2 UMVUE: CRLB method

The second approach to find the UMVUE is based on the following procedure:

(i) Find the lower bound of $\mathrm{Var}(\tilde{\theta})$ for all unbiased estimators; (3.5)

(ii) Find an unbiased estimator $\hat{\theta}$ whose variance achieves this lower bound. (3.6)

Clearly, $\hat{\theta}$ in (3.6) is the UMVUE of θ. It is worth noting that conditions (3.5)-(3.6) are not necessary for the UMVUE, since there are some cases that the UMVUE can not achieve the lower bound in (3.5).

To consider the lower bound of $\mathrm{Var}(\tilde{\theta})$, we need introduce the Fisher information.

Definition 3.10. *(Fisher information) The Fisher information about θ is defied as*

$$I_n(\theta) = \mathrm{E}\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right],$$

where $\ell(\theta) = \log L(\theta)$ is the log-likelihood function of the random sample.

Theorem 3.3. *Let* **X** *be an independent random sample from a population with the p.d.f. $f(x;\theta)$. Then, under certain regularity conditions, we have the following conclusions.*

(i) $I_n(\theta) = nI(\theta)$, where

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right],$$

and X has the same distribution as the population;

(ii) $I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right]$;

(iii) Cramer-Rao inequality:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)},$$

where $\hat{\theta}$ is an unbiased estimator of θ, and $\frac{1}{I_n(\theta)}$ is called the Cramer-Rao lower bound (CRLB).

Proof. (i) Let $\frac{\partial \log f(X;\theta)}{\partial \theta}$ be the score function. Under certain regularity conditions, it can be shown that the first moment of the score is

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial \log f(X;\theta)}{\partial \theta}\right] &= \mathrm{E}\left[\frac{\frac{\partial}{\partial \theta} f(X;\theta)}{f(X;\theta)}\right] \\
&= \int \frac{\frac{\partial}{\partial \theta} f(x;\theta)}{f(x;\theta)} f(x;\theta)\mathrm{d}x \\
&= \int \frac{\partial}{\partial \theta} f(x;\theta)\mathrm{d}x = \frac{\partial}{\partial \theta}\int f(x;\theta)\mathrm{d}x = \frac{\partial}{\partial \theta} 1 = 0.
\end{aligned}$$

Hence, by the independence of $X_1, \cdots, X_n$, it follows that

$$\begin{aligned}
I_n(\theta) &= \mathrm{E}\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial \theta}\right)^2\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right)^2\right] + \mathrm{E}\left[\sum_{i=1}^{n} \sum_{j=1 \text{ and } j\neq i}^{n} \frac{\partial \log f(X_i;\theta)}{\partial \theta}\frac{\partial \log f(X_j;\theta)}{\partial \theta}\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left(\frac{\partial \log f(X_i;\theta)}{\partial \theta}\right)^2\right] \\
&= nI(\theta).
\end{aligned}$$

(ii) Under certain regularity conditions, it can be shown that

$$\begin{aligned}
&\mathrm{E}\left[\frac{\partial^2}{\partial \theta^2}\log f(X;\theta)\right] \\
&= \int \left\{ \frac{\frac{\partial^2}{\partial \theta^2} f(x;\theta)}{f(x;\theta)} - \left(\frac{\frac{\partial}{\partial \theta} f(x;\theta)}{f(x;\theta)}\right)^2 \right\} f(x;\theta)\mathrm{d}x \\
&= \int \frac{\partial^2}{\partial \theta^2} f(x;\theta)dx - \int \left(\frac{\partial}{\partial \theta}\log f(x;\theta)\right)^2 f(x;\theta)dx \\
&= \frac{\partial^2}{\partial \theta^2}\int f(x;\theta)dx - I(\theta) \\
&= \frac{\partial^2}{\partial \theta^2} 1 - I(\theta) \\
&= -I(\theta).
\end{aligned}$$

(iii) Recall $\mathbf{f}(\theta) := \mathbf{f}(x_1,\cdots,x_n;\theta) = f(x_1;\theta)\cdots f(x_n;\theta)$ is the joint p.d.f. of $\mathbf{X}$. For any unbiased estimator $\hat{\theta}$, we can write $\hat{\theta} = g(\mathbf{X}) := g(X_1,\cdots,X_n)$ for some functional g. Then, we have

$$0 = \mathrm{E}(\hat{\theta} - \theta) = \int \cdots \int [g(x_1,\cdots,x_n) - \theta]\mathbf{f}(\theta)dx_1\cdots dx_n.$$

Differentiating both sides of the preceding equation, we obtain that

$$0 = \int \cdots \int -\mathbf{f}(\theta) + [g(x_1,\cdots,x_n) - \theta]\frac{\partial \mathbf{f}(\theta)}{\partial \theta}dx_1\cdots dx_n,$$

which implies that

$$1 = \int \cdots \int [g(x_1,\cdots,x_n) - \theta]\frac{\partial \mathbf{f}(\theta)}{\partial \theta}dx_1\cdots dx_n.$$

Using the Cauchy-Schwarz Inequality: $(\int s_1(x)s_2(x)dx)^2 \leq \int s_1^2(x)dx \int s_2^2(x)dx$, we have

$$\begin{aligned}
1 &= \left\{\int\cdots\int[g(x_1,\cdots,x_n)-\theta]\frac{\partial \mathbf{f}(\theta)}{\partial\theta}dx_1\cdots dx_n\right\}^2\\
&= \left\{\int\cdots\int[g(x_1,\cdots,x_n)-\theta]\sqrt{\mathbf{f}(\theta)}\left[\frac{1}{\sqrt{\mathbf{f}(\theta)}}\frac{\partial \mathbf{f}(\theta)}{\partial\theta}\right]dx_1\cdots dx_n\right\}^2\\
&= \left\{\int\cdots\int[g(x_1,\cdots,x_n)-\theta]\sqrt{\mathbf{f}(\theta)}\left[\sqrt{\mathbf{f}(\theta)}\frac{\partial \log\mathbf{f}(\theta)}{\partial\theta}\right]dx_1\cdots dx_n\right\}^2\\
&\le \int\cdots\int[g(x_1,\cdots,x_n)-\theta]^2\mathbf{f}(\theta)dx_1\cdots dx_n\times\int\cdots\int\mathbf{f}(\theta)\left(\frac{\partial\log\mathbf{f}(\theta)}{\partial\theta}\right)^2dx_1\cdots dx_n.
\end{aligned}$$

Hence, it gives us that $1\le \mathrm{Var}(\hat\theta)\times I_n(\theta)$, which implies that the Cramer-Rao inequality holds. □

Remark 3.4. From the proof above, we can find that "=" holds if and only if there exists a constant A such that

$$\begin{aligned}
&A[g(x_1,\cdots,x_n)-\theta]\sqrt{\mathbf{f}(\theta)}=\sqrt{\mathbf{f}(\theta)}\frac{\partial\log\mathbf{f}(\theta)}{\partial\theta}\ \text{ for all } x_1,x_2,\cdots,x_n\\
\Longleftrightarrow &A[g(X_1,\cdots,X_n)-\theta]=\frac{\partial\log\mathbf{f}(X_1,X_2,\cdots,X_n;\theta)}{\partial\theta}\ \text{ (with probability one)}\\
\Longleftrightarrow &A[\hat\theta-\theta]=\frac{\partial\log L(\theta)}{\partial\theta}\ \text{ (with probability one).}
\end{aligned}\tag{3.7}$$

Equation (3.7) is called the attainable condition for the CRLB. In other words, if $\hat\theta$ can achieve the lower bound, it must satisfy (3.7).

Corollary 3.1. *If $\hat\theta$ is an unbiased estimator of θ and* $\mathrm{Var}(\hat\theta)=\frac{1}{I_n(\theta)}$, *then $\hat\theta$ is the UMVUE of θ.*

Corollary 3.1 tells us how to find the UMVUE by the method of CRLB.

Example 3.10. Show that $\overline{X}$ is the UMVUE of the mean of a normal population $\mathrm{N}(\mu,\sigma^2)$.

Solution. We use the CRLB method to do this. For $-\infty<x<\infty$,

$$f(x;\mu)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right],$$

which implies that

$$\begin{aligned}
&\log f(x;\mu)=-\log(\sigma\sqrt{2\pi})-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2,\\
&\text{and }\frac{\partial\log f(x;\mu)}{\partial\mu}=\frac{x-\mu}{\sigma^2}.
\end{aligned}$$

Therefore,

$$I(\mu)=\mathrm{E}\left[\left(\frac{\partial\log f(X;\mu)}{\partial\mu}\right)^2\right]=\mathrm{E}\left[\frac{(X-\mu)^2}{\sigma^4}\right]=\frac{1}{\sigma^2}$$

or

$$I(\mu) = -\mathrm{E}\left[\frac{\partial^2 \log f(X;\mu)}{\partial \mu^2}\right] = -\mathrm{E}\left(\frac{-1}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

Hence,

$$CRLB = \frac{1}{I_n(\mu)} = \frac{1}{nI(\mu)} = \frac{\sigma^2}{n}.$$

Recall that $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$. Thus, $\overline{X}$ is the UMVUE of μ.
(A further question: what is the CRLB with respect to σ^2? Is this CRLB attainable?) □

Example 3.11. Show that $\overline{X}$ is the UMVUE of the parameter θ of a Bernoulli population.

Solution. We use the CRLB method to do this. For $x = 0$ or 1,

$$f(x;\theta) = \theta^x(1-\theta)^{1-x},$$

which implies that

$$\begin{aligned}\frac{\partial \log f(x;\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}[x\log\theta + (1-x)\log(1-\theta)]\\ &= \frac{x}{\theta} - \frac{1-x}{1-\theta}\\ &= \frac{x}{\theta(1-\theta)} - \frac{1}{1-\theta}.\end{aligned}$$

Noting that

$$\mathrm{E}\left[\frac{X}{\theta(1-\theta)}\right] = \frac{1}{1-\theta},$$

we have

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right] = \mathrm{Var}\left[\frac{X}{\theta(1-\theta)}\right] = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

Hence,

$$CRLB = \frac{1}{I_n(\theta)} = \frac{1}{nI(\theta)} = \frac{\theta(1-\theta)}{n}.$$

Since $\mathrm{E}(\overline{X}) = \theta$ and $\mathrm{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$, $\overline{X}$ is the UMVUE of θ. □

If we know the full information about the population distribution X, the following theorem tells us that the MLE tends to be the first choice asymptotically.

Theorem 3.4. *Suppose that $\hat{\theta}$ is the MLE of a parameter θ of a population distribution. Then, under certain regular conditions, as $n \to \infty$,*

$$\frac{\hat{\theta} - \theta}{\sqrt{1/I_n(\theta)}} \to_d \mathrm{N}(0,1).$$

If $\hat{\theta}$ is an unbiased estimator of θ, the above theorem implies that $\mathrm{Var}(\hat{\theta}) \approx \dfrac{1}{I_n(\theta)}$ when n is large. That is, the MLE $\hat{\theta}$ can achieve the CRLB asymptotically.

3.3.3 Consistency

In the previous discussions, we have restricted our attention to the unbiased estimator, and proposed a way to check whether an unbiased estimator is UMVUE. Now, we introduce another property of the estimator called the consistency.

Definition 3.11. *(Consistent estimator) $\hat{\theta}$ is a consistent estimator of θ, if*

$$\hat{\theta} \rightarrow_p \theta,$$

that is, for any $\varepsilon > 0$,

$$P\left(|\hat{\theta}-\theta| > \varepsilon\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Note that the definition of the convergence in probability is given in Definition 2.2.

Property 3.5. *If $\hat{\theta}$ is an unbiased estimator of a parameter θ and* $\mathrm{Var}(\hat{\theta}) \rightarrow 0$ *as $n \rightarrow \infty$, then $\hat{\theta}$ is a consistent estimator of θ.*

Property 3.6. *If $\hat{\theta}$ is an asymptotically unbiased estimator of a parameter θ and* $\mathrm{Var}(\hat{\theta}) \rightarrow 0$ *as $n \rightarrow \infty$, then $\hat{\theta}$ is a consistent estimator of θ.*

We shall mention that the unbiasedness along does not imply the consistency. A toy example is as follows. Suppose that

$$\hat{\theta} = \mathrm{I}(0 < X_1 < 1/2) - \mathrm{I}(1/2 < X_1 < 1),$$

where $X_1 \sim U(0,1)$. Then, $\mathrm{E}(\hat{\theta}) = 0$, i.e., $\hat{\theta}$ is an unbiased estimator of $\theta = 0$. But, as $\hat{\theta}$ takes value of either 1 or -1, $\hat{\theta} \not\rightarrow_p 0$.

Property 3.7. *If $\hat{\theta} \rightarrow_p \theta$ and $\tilde{\theta} \rightarrow_p \theta'$, then*
(i) $\hat{\theta} \pm \tilde{\theta} \rightarrow_p \theta \pm \theta'$;
(ii) $\hat{\theta} \cdot \tilde{\theta} \rightarrow_p \theta \cdot \theta'$;
(iii) $\hat{\theta}/\tilde{\theta} \rightarrow_p \theta/\theta'$ assuming that $\tilde{\theta} \neq 0$ and $\theta' \neq 0$;
(iv) if g is any real-valued function that is continuous at θ, $g(\hat{\theta}) \rightarrow_p g(\theta)$.

Example 3.12. Suppose that $\mathbf{X}$ is an independent random sample from a population with the finite mean $\mu = \mathrm{E}(X_1)$, finite variance $\sigma^2 = \mathrm{Var}(X_1)$, and finite fourth moment $\mu_4 = \mathrm{E}(X_1^4)$. Show that $\overline{X}$ is a consistent estimator of μ, and S^2 is a consistent estimator of σ^2.

Solution. Note that $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \rightarrow 0$ as $n \rightarrow \infty$. Hence, by Property 3.5, $\overline{X}$ is a consistent estimator of μ (This is just the weak law of large numbers in Theorem 2.1).

For S^2, we have $S^2 = \overline{X^2} - (\overline{X})^2$. By the weak law of large numbers, we have

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \rightarrow_p \mu_2 = \mathrm{E}(X_1^2).$$

As $\overline{X} \rightarrow_p \mu$, Property 3.7(iv) implies that

$$(\overline{X})^2 \rightarrow_p \mu^2.$$

Therefore, by Property 3.7(i), $S^2 \rightarrow_p \mu_2 - \mu^2 = \sigma^2$. □

Remark 3.5. If $\mathbf{X}$ is from a $\mathrm{N}(\mu, \sigma^2)$ population, we can show that

$$\frac{nS^2}{\sigma^2} = \chi^2_{n-1},$$

where χ^2_k is a chi-square distribution with k degrees of freedom. Therefore, $\mathrm{E}(\frac{nS^2}{\sigma^2}) = n-1$, which implies that $\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2 \to \sigma^2$ as $n \to \infty$. That is, S^2 is an asymptotically unbiased estimator of σ^2.

Moreover, since $\mathrm{Var}(\frac{nS^2}{\sigma^2}) = 2(n-1)$, we can obtain that

$$\mathrm{Var}(S^2) = \mathrm{Var}\left(\frac{\sigma^2}{n} \cdot \frac{nS^2}{\sigma^2}\right) = \frac{2\sigma^4(n-1)}{n^2} \to 0 \text{ as } n \to \infty.$$

Hence, by Property 3.6, S^2 is a consistent estimator of σ^2.

Example 3.13. Let $\mathbf{X}$ be an independent random sample from a population with a p.d.f.

$$f(x;\theta) = \frac{2x}{\theta^2}\mathrm{I}(0 < x \leq \theta).$$

(i) Find the UMVUE of θ;
(ii) Show that this UMVUE is a consistent estimator of θ;
(iii) Find the MLE of θ;
(iv) Find a MME of θ;
(v) Will the MLE be better than the MME in terms of efficiency?

Solution. (i) Since $f(x;\theta)$ is not continuous with respect to θ, the method of CRLB does not work. We use the method based on a complete and sufficient statistic. Note that $f(x;\theta)$ does not belong to the exponential family (Why?). Below, we look for a complete and sufficient statistic for θ by definition.

First, the joint p.d.f. of $\mathbf{X}$ is

$$\begin{aligned}
\mathbf{f}(x_1, x_2, \cdots, x_n; \theta) &= f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) \\
&= \prod_{i=1}^{n} \frac{2x_i}{\theta^2}\mathrm{I}(0 < x_i < \theta) \\
&= \left[\prod_{i=1}^{n} x_i \mathrm{I}(x_i > 0)\right]\left[\prod_{i=1}^{n} \frac{2}{\theta^2}\mathrm{I}(x_i \leq \theta)\right] \\
&= \left[\prod_{i=1}^{n} x_i \mathrm{I}(x_i > 0)\right]\left[\left(\frac{2}{\theta^2}\right)^n \prod_{i=1}^{n}\mathrm{I}(x_i \leq \theta)\right] \\
&= \left[\prod_{i=1}^{n} x_i \mathrm{I}(x_i > 0)\right]\left[\left(\frac{2}{\theta^2}\right)^n \mathrm{I}\left(\max_{1\leq i\leq n} x_i \leq \theta\right)\right].
\end{aligned}$$

By Definition 3.8, we know that $T := \max_{1\leq i\leq n} X_i$ is a sufficient statistic for θ.

Second, we show that T is also a complete statistic for θ. Let $z(\cdot)$ be a functional such that $\mathrm{E}(z(T)) = 0$ for all $\theta > 0$, i.e.,

$$0 = \mathrm{E}(z(T)) = \int_0^\theta z(t)\frac{2nt^{2n-1}}{\theta^{2n}}dt = \frac{2n}{\theta^{2n}}\int_0^\theta z(t)t^{2n-1}dt \tag{3.8}$$

for all $\theta > 0$, where we have used the fact that T has a p.d.f. $f(t) = \dfrac{2nt^{2n-1}}{\theta^{2n}}$ for $t \in (0, \theta)$ (Why?).

By (3.8), we can get that $\int_0^\theta z(t)t^{2n-1}dt = 0$ for all $\theta > 0$, which implies that $z(\theta)\theta^{2n-1} = 0$ and hence $z(\theta) = 0$. Therefore, by Definition 3.9, we know that T is a complete statistic for θ.

Third, we can show that

$$\mathrm{E}(T) = \int_0^\theta t\frac{2nt^{2n-1}}{\theta^{2n}}dt = \frac{2n}{\theta^{2n}}\int_0^\theta t^{2n}dt = \frac{2n}{2n+1}\theta,$$

which implies that $Y := \dfrac{2n+1}{2n}T$ is the UMVUE of θ by Theorem 3.2.

(ii) By simple calculation, we have

$$\mathrm{E}(Y^2) = \int_0^\theta \left(\frac{2n+1}{2n}t\right)^2 \frac{2nt^{2n-1}}{\theta^{2n}}dt = \frac{(2n+1)^2}{(2n)\theta^{2n}}\int_0^\theta t^{2n+1}dt = \frac{(2n+1)^2}{(2n)(2n+2)}\theta^2,$$

which implies that

$$\mathrm{Var}(Y) = \mathrm{E}(Y^2) - [\mathrm{E}(Y)]^2 = \frac{(2n+1)^2}{(2n)(2n+2)}\theta^2 - \theta^2 \to 0$$

as $n \to \infty$. Therefore, by Property 3.5, Y is a consistent estimator of θ.

(iii) The MLE of θ is T (Why?).

(iv) Note that

$$\mathrm{E}(X) = \int_0^\theta x\frac{2x}{\theta^2}dx = \frac{2}{3}\theta.$$

Hence, $\frac{3}{2}\overline{X}$ is a MME of θ.

(v) This is left as an exercise. □

Chapter 4
Interval estimation

4.1 Basic concepts

In the previous chapter, we have learned how to construct a point estimator for a unknown parameter θ, leading to the guess of a single value as the value of θ. However, a point estimator for θ does not provide much information about the accuracy of the estimator. It is desirable to generate a narrow interval that will cover the unknown parameter θ with a large probability (confidence). This motivates us to consider the interval estimator in this chapter.

Definition 4.1. *(Interval estimator) An interval estimator of θ is a random interval $[L(\mathbf{X}), U(\mathbf{X})]$, where $L(\mathbf{X}) := L(X_1, \cdots, X_n)$ and $U(\mathbf{X}) := U(X_1, \cdots, X_n)$ are two statistics such that $L(\mathbf{X}) \le U(\mathbf{X})$ with probability one.*

Definition 4.2. *(Interval estimate) If $\mathbf{X} = \mathbf{x}$ is observed, $[L(\mathbf{x}), U(\mathbf{x})]$ is the interval estimate of θ.*

Although the definition is based on a closed interval $[L(\mathbf{X}), U(\mathbf{X})]$, it will sometimes be more natural to use an open interval $(L(\mathbf{X}), U(\mathbf{X}))$, a half-open and half-closed interval $(L(\mathbf{X}), U(\mathbf{X})]$ (or $[L(\mathbf{X}), U(\mathbf{X}))$), or an one-sided interval $(\infty, U(\mathbf{X})]$ (or $[L(\mathbf{X}), \infty)$).

The next example shows that compared to the point estimator, the interval estimator can have some confidence (or guarantee) of capturing the parameter of interest, although it gives up some precision.

Example 4.1. For an independent random sample X_1, X_2, X_3, X_4 from $\mathrm{N}(\mu, 1)$, consider an interval estimator of μ by $[\overline{X} - 1, \overline{X} + 1]$. Then, the probability that μ is covered by the interval $[\overline{X} - 1, \overline{X} + 1]$ can be calculated by

$$\begin{aligned}
\mathrm{P}\left(\mu \in [\overline{X} - 1, \overline{X} + 1]\right) &= \mathrm{P}\left(\overline{X} - 1 \le \mu \le \overline{X} + 1\right) \\
&= \mathrm{P}\left(-1 \le \overline{X} - \mu \le 1\right) \\
&= \mathrm{P}\left(-2 \le \frac{\overline{X} - \mu}{\sqrt{1/4}} \le 2\right) \\
&= \mathrm{P}(-2 \le Z \le 2) \\
&\approx 0.9544,
\end{aligned}$$

where $Z \sim \mathrm{N}(0, 1)$ and we have used the fact that $\overline{X} \sim \mathrm{N}(\mu, 1/4)$. Thus, we have over a 95% change of covering the unknown parameter with our interval estimator. Note that for any point estimator $\hat{\mu}$ of μ, we have $\mathrm{P}(\hat{\mu} = \mu) = 0$. Sacrificing some precision in the interval estimator, in moving from a point to an interval, has resulted in increased confidence that our assertion about μ is correct. □

The certainty of the confidence (or guarantee) is quantified in the following definition.

Definition 4.3. *(Confidence coefficient) For an interval estimator* $[L(\mathbf{X}), U(\mathbf{X})]$ *of* θ*, the confidence coefficient of* $[L(\mathbf{X}), U(\mathbf{X})]$*, denoted by* $(1-\alpha)$*, is*

$$1-\alpha = \mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]),$$

where $\mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ *is the coverage probability of* $[L(\mathbf{X}), U(\mathbf{X})]$.

Remark 4.1. In some situations, the coverage probability $\mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})])$ may depend on θ, and then the the confidence coefficient is defined as

$$1-\alpha = \inf_{\theta} \mathrm{P}(\theta \in [L(\mathbf{X}), U(\mathbf{X})]).$$

Interval estimator, together with a measure of confidence (say, the confidence coefficient), is sometimes known as **confidence interval**. So, the terminologies of interval estimators and confidence intervals are interchangeable. A confidence interval with confidence coefficient equal to $1-\alpha$, is called a $1-\alpha$ confidence interval.

For example, a 95% (i.e., $\alpha = 0.05$) confidence interval means that if 100 confidence intervals were constructed based on 100 different samples from the same population, we would expect 95 of the intervals to contain θ.

Now, the question is how to construct the interval estimator. One important way to do it is using the pivotal quantity.

Definition 4.4. *(Pivotal Quantity) A random variable* $Q(\mathbf{X}, \theta) = Q(X_1, \cdots, X_n, \theta)$ *is a pivotal quantity if the distribution of* $Q(\mathbf{X}, \theta)$ *is free of* θ*. That is, regardless of the distribution of* $\mathbf{X}$*,* $Q(\mathbf{X}, \theta)$ *has the same distribution for all values of* θ*.*

Logically, when $Q(\mathbf{X}, \theta)$ is a pivotal quantity, we can easily construct a $1-\alpha$ confidence interval for $Q(\mathbf{X}, \theta)$ by

$$1-\alpha = \mathrm{P}\left(\widetilde{L}_\alpha \le Q(\mathbf{X}, \theta) \le \widetilde{U}_\alpha\right), \tag{4.1}$$

where $\widetilde{L}_\alpha$ and $\widetilde{U}_\alpha$ do not depend on θ. Suppose that the inequalities $\widetilde{L}_\alpha \le Q(\mathbf{X}, \theta) \le \widetilde{U}_\alpha$ in (4.1) are equivalent to the inequalities $L(\mathbf{X}) \le \theta \le U(\mathbf{X})$, Then, from (4.1), a $1-\alpha$ confidence interval of θ is $[L(\mathbf{X}), U(\mathbf{X})]$.

In the rest of this section, we will use the pivotal quantity method to construct our interval estimators.

4.2 Confidence intervals for means

4.2.1 One-sample case

Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be an independent random sample from the population $\mathrm{N}(\mu, \sigma^2)$. We first consider the interval estimator of μ when σ^2 is known. Note that

$$\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n). \tag{4.2}$$

Hence, when σ^2 is known, $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$ is a pivotal quantity involving μ. Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-z_{\alpha/2} \le Z \le z_{\alpha/2}\right) \\ &= \mathrm{P}\left(-z_{\alpha/2} \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) \\ &= \mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \end{aligned}$$

where z_α satisfies

$$\mathrm{P}(Z \ge z_\alpha) = \alpha$$

for $Z \sim \mathrm{N}(0,1)$. Usually, we call z_α the upper percentile of $\mathrm{N}(0,1)$ at the level α; see Fig. 4.1. So, when σ^2 is known, a $1-\alpha$ confidence interval of μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{4.3}$$

Given the observed value of $\overline{X} = \bar{x}$ and the value of $z_{\alpha/2}$, we can calculate the interval estimate of μ by

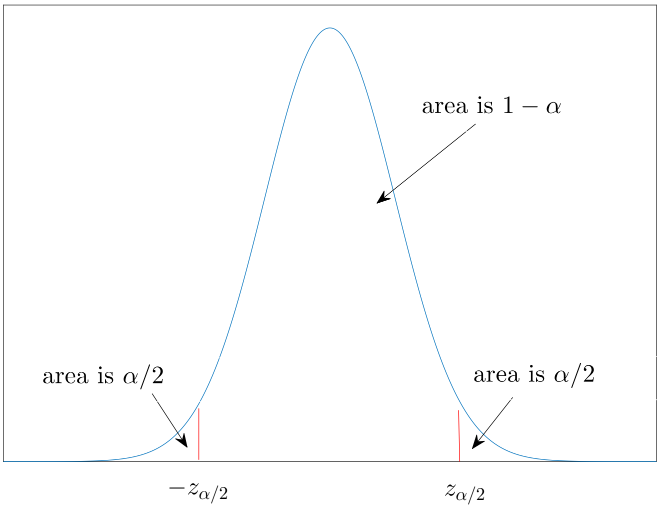

Fig. 4.1 Upper percentile of $\mathrm{N}(0,1)$ distribution

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

As the point estimator, the $1-\alpha$ confidence interval is also not unique. Ideally, we should choose it as narrow as possible in some sense, but in practice, we usually choose the equal-tail confidence interval as in (4.3) for convenience, since tables for selecting equal probabilities in the two tails are readily available.

Example 4.2. A publishing company has just published a new college textbook. Before the company decides the price of the book, it wants to know the average price of all such textbooks in the market. The research department at the company took a sample of 36 such textbooks and collected information on their prices. This information produced a mean price of \$48.40 for this sample. It is known that the standard deviation of the prices of all such textbooks is \$4.50. Construct a 90% confidence interval for the mean price of all such college textbooks assuming that the underlying population is normal.

Solution. From the given information, $n = 36$, $\bar{x} = 48.40$ and $\sigma = 4.50$. Now, $1-\alpha = 0.9$, i.e., $\alpha = 0.1$, and by (4.3), the 90% confidence interval for the mean price of all such college textbooks is given by

$$\begin{aligned}[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] &= [48.40 - z_{0.05}\frac{4.50}{\sqrt{36}}, 48.40 + z_{0.05}\frac{4.50}{\sqrt{36}}] \\ &\approx [47.1662, 49.6338].\end{aligned}$$

□

Example 4.3. Suppose the bureau of the census and statistics of a city wants to estimate the mean family annual income μ for all families in the city. It is known that the standard deviation σ for the family annual income is 60 thousand dollars. How large a sample should the bureau select so that it can assert with probability 0.99 that the sample mean will differ from μ by no more than 5 thousand dollars?

Solution. From the construction of a confidence interval, we have

$$1-\alpha = \mathrm{P}\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \mathrm{P}\left(|\overline{X} - \mu| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

where $1-\alpha = 0.99$ and $\sigma = 60$ thousand dollars. It suffices to have $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le 5$ or

$$n \ge \left(\frac{60 z_{\alpha/2}}{5}\right)^2 = \left(\frac{60 \times 2.576}{5}\right)^2 \approx 955.5517.$$

Thus, the sample size should be at least 956. (Note that we have to round 955.5517 up to the next higher integer. This is always the case when determining the sample size.) □

Next, we consider the interval estimator of μ when σ^2 is unknown. To find a pivotal quantity, we need to use the following result.

Property 4.1. *(i) $\overline{X}$ and S^2 are independent;*

(ii) $\frac{nS^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$ is χ^2_{n-1}, where χ^2_k is a chi-square distribution with k degrees of freedom;

(iii) $T = \frac{\overline{X} - \mu}{S/\sqrt{n-1}}$ is t_{n-1}, where t_k is a t distribution with k degrees of freedom.

Property 4.2. *(i) If $Z_1, \cdots, Z_k$ are k independent $N(0,1)$ random variables, then $\sum_{i=1}^{k} Z_i^2$ is χ^2_k;*

(ii) If Z is $N(0,1)$, U is χ^2_k, and Z and U are independent, then $T = \dfrac{Z}{\sqrt{U/k}}$ is t_k.

Proof of Property 4.1. (i) The proof of (i) is out of the scope of this course;

(ii) Note that

$$\begin{aligned} W = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 &= \sum_{i=1}^{n} \left[\frac{X_i - \overline{X}}{\sigma} + \frac{\overline{X} - \mu}{\sigma} \right]^2 \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 + \sum_{i=1}^{n} \left(\frac{\overline{X} - \mu}{\sigma} \right)^2 \\ &= \frac{nS^2}{\sigma^2} + Z^2, \end{aligned}$$

where we have used the fact that the cross-product term is equal to

$$2\sum_{i=1}^{n} \frac{(\overline{X} - \mu)(X_i - \overline{X})}{\sigma^2} = \frac{2(\overline{X} - \mu)}{\sigma^2} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) = 0.$$

Note that W is χ^2_n and Z^2 is χ^2_1 by Property 4.2(i). Since $\dfrac{nS^2}{\sigma^2}$ and Z^2 are independent by (i), we can show that the m.g.f. of $\dfrac{nS^2}{\sigma^2}$ is the same as the one of χ^2_{n-1}. Hence, (ii) holds.

(iii) Note that

$$T = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{nS^2/\sigma^2}\sqrt{1/(n-1)}}.$$

Hence, by Property 4.2(ii), T is t_{n-1}. □

From Property 4.1(iii), we know that T is a pivotal quantity of μ. Let

$$\begin{aligned} 1 - \alpha &= \mathrm{P}\left(-t_{\alpha/2,df=n-1} \le T \le t_{\alpha/2,df=n-1} \right) \\ &= \mathrm{P}\left(-t_{\alpha/2,df=n-1} \le \frac{\overline{X} - \mu}{S/\sqrt{n-1}} \le t_{\alpha/2,df=n-1} \right) \\ &= \mathrm{P}\left(\overline{X} - t_{\alpha/2,df=n-1} \frac{S}{\sqrt{n-1}} \le \mu \le \overline{X} + t_{\alpha/2,df=n-1} \frac{S}{\sqrt{n-1}} \right), \end{aligned}$$

where $t_{\alpha,df=k}$ satisfies

$$\mathrm{P}(T \ge t_{\alpha,df=k}) = \alpha$$

for a random variable $T \sim t_k$; see Fig. 4.2. So, when σ^2 is unknown, a $1 - \alpha$ confidence interval of μ is

$$\left[\overline{X} - t_{\alpha/2,df=n-1} \frac{S}{\sqrt{n-1}}, \overline{X} + t_{\alpha/2,df=n-1} \frac{S}{\sqrt{n-1}} \right]. \tag{4.4}$$

Given the observed value of $\overline{X} = \overline{x}$, $S = s$, and the value of $t_{\alpha/2,df=n-1}$, we can calculate the interval estimate of μ by

$$\left[\overline{x} - t_{\alpha/2,df=n-1} \frac{s}{\sqrt{n-1}}, \overline{x} + t_{\alpha/2,df=n-1} \frac{s}{\sqrt{n-1}} \right].$$

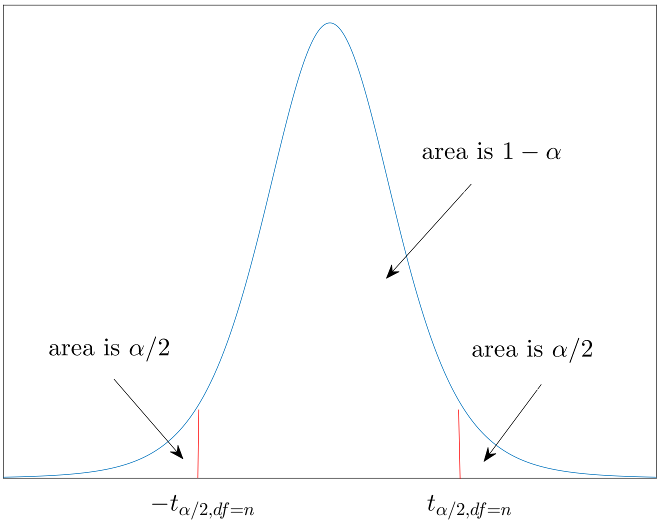

Fig. 4.2 Upper percentile of t_n distribution

Remark 4.2. Usually there is a row with ∞ degrees of freedom in a t-distribution table, which actually shows values of z_α. In fact, when $n \to \infty$, the distribution function of t_n tends to that of $\mathrm{N}(0,1)$; see Fig. 4.3. That is, in tests or exams, if n is so large that the value of $t_{\alpha,df=n}$ cannot be found, you may use z_α instead.

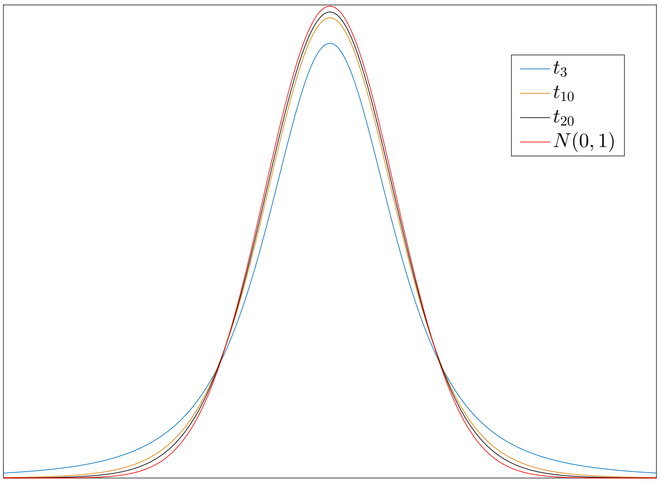

Fig. 4.3 Distributions of t_n and $N(0,1)$

Example 4.4. A paint manufacturer wants to determine the average drying time of a new brand of interior wall paint. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true population mean assuming normality.

Solution. As $n = 12$, $\bar{x} = 66.3$, $s = 8.4$, $\alpha = 1 - 0.95 = 0.05$ and $t_{\alpha/2,df=n-1} = t_{0.025,11} \approx 2.201$, the 95% confidence interval for μ is

$$\left[66.3 - 2.201 \times \frac{8.4}{\sqrt{12-1}},\ 66.3 + 2.201 \times \frac{8.4}{\sqrt{12-1}}\right],$$

that is, $[61.1722, 71.4278]$. □

Example 4.5. Construct a 95% confidence interval for the mean hourly wage of apprentice geologists employed by the top 5 oil companies. For a sample of 50 apprentice geologists, $\bar{x} = 14.75$ and $s = 3.0$ (in dollars).

Solution. As $n = 50$, $\alpha = 1 - 0.95 = 0.05$, $t_{0.025,df=49} \approx 2.010$, we have

$$t_{\alpha/2,df=n-1}\frac{s}{\sqrt{n}} \approx 2.010 \times \frac{3.0}{\sqrt{50-1}} = 0.8614.$$

Thus, the 95% confidence interval is $[14.75 - 0.86,\ 14.75 + 0.86]$, or $[13.89, 15.59]$. □

4.2.2 Tow-sample case

Besides the confidence interval for the mean of one single normal distribution, we shall also consider the problem of constructing confidence intervals for the difference of the means of two normal distributions when the variances are unknown.

Let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ and $\mathbf{Y} = \{Y_1, Y_2, \cdots, Y_m\}$ be random samples from independent distributions $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$, respectively. We are of interest to construct the confidence interval for $\mu_X - \mu_Y$ when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

First, we can show that

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2/n + \sigma^2/m}}$$

is $\mathrm{N}(0, 1)$. Also, by the independence of $\mathbf{X}$ and $\mathbf{Y}$, from Property 4.1(ii), we know that

$$U = \frac{nS_X^2}{\sigma^2} + \frac{mS_Y^2}{\sigma^2}$$

is χ^2_{n+m-2}. Moreover, by Property 4.1(i), Z and U are independent. Hence,

$$\begin{aligned} T &= \frac{Z}{\sqrt{U/(n+m-2)}} \\ &= \frac{[(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)]/\sqrt{\sigma^2/n + \sigma^2/m}}{\sqrt{(nS_X^2 + mS_Y^2)/[\sigma^2(n+m-2)]}} \\ &= \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{R} \end{aligned}$$

is t_{n+m-2}, where

$$R = \sqrt{\frac{nS_X^2 + mS_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}.$$

That is, T is a pivotal quantity of $\mu_X - \mu_Y$. Let

$$\begin{aligned}
1 - \alpha &= \mathrm{P}\left(-t_{\alpha/2,df=n+m-2} \le T \le t_{\alpha/2,df=n+m-2}\right) \\
&= \mathrm{P}\left(-t_{\alpha/2,df=n+m-2} \le \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{R} \le t_{\alpha/2,df=n+m-2}\right) \\
&= \mathrm{P}\left((\overline{X} - \overline{Y}) - t_{\alpha/2,df=n+m-2}R \le \mu_X - \mu_Y \le (\overline{X} - \overline{Y}) + t_{\alpha/2,df=n+m-2}R\right),
\end{aligned}$$

So, when $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ is unknown, a $1 - \alpha$ confidence interval of $\mu_X - \mu_Y$ is

$$\left[(\overline{X} - \overline{Y}) - t_{\alpha/2,df=n+m-2}R, (\overline{X} - \overline{Y}) + t_{\alpha/2,df=n+m-2}R\right]. \tag{4.5}$$

Given the observed value of $\overline{X} = \bar{x}, \overline{Y} = \bar{y}, S_X = s_X, S_Y = s_Y$, and the value of $t_{\alpha/2,df=n+m-2}$, we can calculate the interval estimate of $\mu_X - \mu_Y$ by

$$\left[(\bar{x} - \bar{y}) - t_{\alpha/2,df=n+m-2}r, (\bar{x} - \bar{y}) + t_{\alpha/2,df=n+m-2}r\right].$$

where

$$r = \sqrt{\frac{ns_X^2 + ms_Y^2}{n+m-2}\left(\frac{1}{n} + \frac{1}{m}\right)}.$$

Example 4.6. Suppose that scores on a standardized test in mathematics taken by students from large and small high schools are $\mathrm{N}(\mu_X, \sigma^2)$ and $\mathrm{N}(\mu_Y, \sigma^2)$, respectively, where σ^2 is unknown. If a random sample of $n = 9$ students from large high schools yielded $\bar{x} = 81.31$, $s_X^2 = 60.76$ and a random sample of $m = 15$ students from small high schools yielded $\bar{y} = 78.61$, $s_Y^2 = 48.24$, the endpoints for a 95% confidence interval for $\mu_X - \mu_Y$ are given by

$$81.31 - 78.61 \pm 2.074\sqrt{\frac{9 \times 60.76 + 15 \times 48.24}{22}\left(\frac{1}{9} + \frac{1}{15}\right)},$$

since $\mathrm{P}(T \le 2.074) = 0.975$. So, the 95% confidence interval is $[-3.95, 9.35]$. □

4.3 Confidence intervals for variances

4.3.1 One-sample case

First, we consider the one-sample case. By Property 4.1(ii),

$$\frac{nS^2}{\sigma^2} \sim \chi_{n-1}^2$$

is a pivotal quantity involving σ^2. Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(\chi^2_{1-\alpha/2,df=n-1} \le \frac{nS^2}{\sigma^2} \le \chi^2_{\alpha/2,df=n-1}\right) \\ &= \mathrm{P}\left(\frac{nS^2}{\chi^2_{\alpha/2,df=n-1}} \le \sigma^2 \le \frac{nS^2}{\chi^2_{1-\alpha/2,df=n-1}}\right), \end{aligned}$$

where $\chi^2_{\alpha,df=n}$ satisfies

$$\mathrm{P}(T \ge \chi^2_{\alpha,df=n}) = \alpha$$

for a random variable $T \sim \chi^2_n$; see Fig. 4.4. So, a $1-\alpha$ confidence interval of σ^2 is

$$\left[\frac{nS^2}{\chi^2_{\alpha/2,df=n-1}}, \frac{nS^2}{\chi^2_{1-\alpha/2,df=n-1}}\right]. \tag{4.6}$$

Given the observed value of $S = s$ and the values of $\chi^2_{\alpha/2,df=n-1}$ and $\chi^2_{1-\alpha/2,df=n-1}$, we can calculate the interval estimate of σ^2 by

$$\left[\frac{ns^2}{\chi^2_{\alpha/2,df=n-1}}, \frac{ns^2}{\chi^2_{1-\alpha/2,df=n-1}}\right].$$

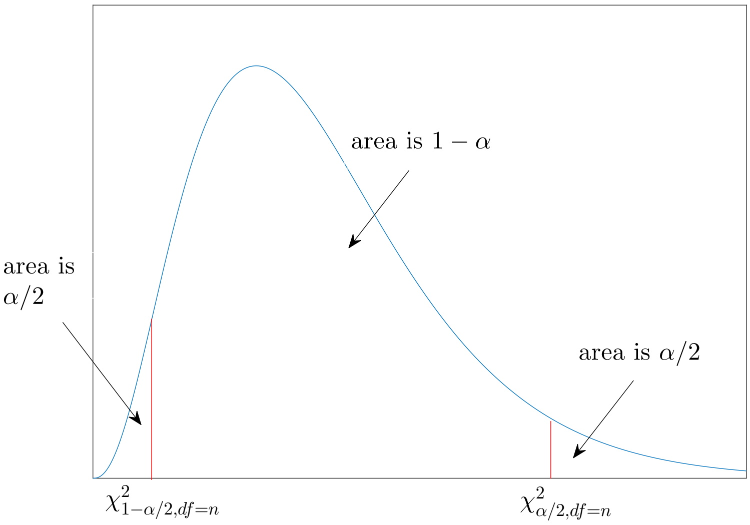

Fig. 4.4 Upper percentile of χ^2_n distribution

Example 4.7. A machine is set up to fill packages of cookies. A recently taken random sample of the weights of 25 packages from the production line gave a variance of 2.9 g^2. Construct a 95% confidence interval for the standard deviation of the weight of a randomly selected package from the production line.

Solution. As $n = 25$, $s^2 = 2.9$, $\alpha = 0.05$,

$$\frac{ns^2}{\chi^2_{\alpha/2,df=n-1}} = \frac{25(2.9)}{\chi^2_{0.025,df=24}} \approx \frac{25(2.9)}{39.36} \approx 1.8420,$$
$$\frac{ns^2}{\chi^2_{1-\alpha/2,df=n-1}} = \frac{25(2.9)}{\chi^2_{0.975,df=24}} \approx \frac{25(2.9)}{12.40} \approx 5.8468,$$

the 95% confidence interval for the population variance is $(1.8420,\ 5.8468)$. Taking positive square roots, we obtain the 95% confidence interval for the population standard deviation to be $(1.3572,\ 2.4180)$. □

4.3.2 Two-sample case

Next, we consider the two-sample case. Let

$$\mathbf{X} = \{X_1, X_2, \cdots, X_n\} \text{ and } \mathbf{Y} = \{Y_1, Y_2, \cdots, Y_m\}$$

be random samples from independent distributions $\mathrm{N}(\mu_X, \sigma_X^2)$ and $\mathrm{N}(\mu_Y, \sigma_Y^2)$, respectively. We are of interest to construct the confidence interval for σ_X^2/σ_Y^2.

Property 4.3. *Suppose that $U \sim \chi^2_{r_1}$ and $V \sim \chi^2_{r_2}$ are independent. Then,*

$$F_{r_1,r_2} = \frac{U/r_1}{V/r_2}$$

has an F_{r_1,r_2} distribution with r_1 and r_2 degrees of freedom.

By Property 4.1(ii),

$$\frac{nS_X^2}{\sigma_X^2} \sim \chi^2_{n-1} \text{ and } \frac{mS_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}.$$

Then, by Property 4.3, it follows that

$$\left[\frac{mS_Y^2}{\sigma_Y^2(m-1)}\right] \Big/ \left[\frac{nS_X^2}{\sigma_X^2(n-1)}\right] \sim F_{m-1,n-1},$$

which is a pivotal quantity involving σ_X^2/σ_Y^2. Let

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(F_{1-\alpha/2,df=(m-1,n-1)} \le \left[\frac{mS_Y^2}{\sigma_Y^2(m-1)}\right] \Big/ \left[\frac{nS_X^2}{\sigma_X^2(n-1)}\right] \le F_{\alpha/2,df=(m-1,n-1)}\right) \\ &= \mathrm{P}\left(\frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F_{1-\alpha/2,df=(m-1,n-1)} \le \frac{\sigma_X^2}{\sigma_Y^2} \le \frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F_{\alpha/2,df=(m-1,n-1)}\right), \end{aligned}$$

where $F_{\alpha,df=(m,n)}$ satisfies

$$\mathrm{P}(T \ge F_{\alpha,df=(m,n)}) = \alpha$$

for a random variable $T \sim F_{m,n}$; see Fig. 4.5. So, a $1-\alpha$ confidence interval of σ_X^2/σ_Y^2 is

$$\left[\frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F_{1-\alpha/2,df=(m-1,n-1)},\ \frac{n(m-1)S_X^2}{m(n-1)S_Y^2} F_{\alpha/2,df=(m-1,n-1)}\right]. \tag{4.7}$$

Given the observed value of $S_X = s_X$, $S_Y = s_Y$, and the values of $F_{\alpha/2,df=(m-1,n-1)}$ and $F_{1-\alpha/2,df=(m-1,n-1)}$, we can calculate the interval estimate of σ_X^2/σ_Y^2 by

$$\left[\frac{n(m-1)s_X^2}{m(n-1)s_Y^2}F_{1-\alpha/2,df=(m-1,n-1)},\ \frac{n(m-1)s_X^2}{m(n-1)s_Y^2}F_{\alpha/2,df=(m-1,n-1)}\right].$$

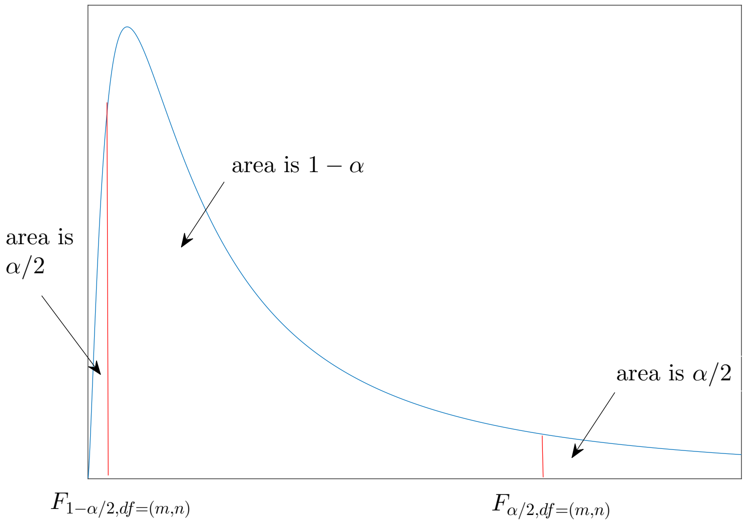

Fig. 4.5 Upper percentile of $F_{m,n}$ distribution

4.4 Confidence intervals: Large samples

In the previous sections, the confidence intervals are all constructed for the normal population, which allows us to deal with the case of a fixed sample size n. In practice, the normal assumption on the population is restricted. When the population is not normal, we can make use of the CLT to propose confidence intervals, which has an approximated confidence coefficient $1-\alpha$ for large n.

To elaborate the idea, we first introduce a useful theorem.

Theorem 4.1. *(Slutsky's theorem) If $X_n \to_d X$ and $Y_n \to_p C$ (a constant), then*
(i) $X_n + Y_n \to_d X + C$;
(ii) $X_n \cdot Y_n \to_d X \cdot C$;
(iii) $X_n/Y_n \to_d X/C$ provided that $C \neq 0$.

Note that Theorem 4.1 fails if C is not a constant.

Let $\mathbf{X}$ be an independent random sample from a population, which has the mean μ and the variance $\sigma^2 < \infty$. According to Theorem 4.1, CLT and the fact that $S \to_p \sigma$, we have

$$\frac{\sqrt{n}(\overline{X}-\mu)}{S} = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \cdot \frac{S}{\sigma} \rightarrow_d \mathrm{N}(0,1), \tag{4.8}$$

for large n. Hence, by (4.8), it follows that for large n,

$$\begin{aligned} 1-\alpha &\approx \mathrm{P}\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{S} \leq z_{\alpha/2}\right) \\ &= \mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right). \end{aligned}$$

So, an approximated $1-\alpha$ confidence interval of μ is

$$\left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]. \tag{4.9}$$

Given the observed value of $\bar{x}$ and s, we can calculate the interval estimate of μ by

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

Note that the confidence interval in (4.9) only requires a large n but not the normal population assumption. Clearly, the similar idea can be applied to the two-sample case.

To end this chapter, we consider the interval estimator for percentage p, where

$$p = \mathrm{P}(X \in (a,b)).$$

Define $\xi = \mathrm{I}(a < X < b)$. Then, $\mathrm{E}(\xi) = p$. This indicates that p is the theoretical mean of ξ. Hence, by (4.9), an approximated $1-\alpha$ confidence interval of p is

$$\left[\overline{\xi} - z_{\alpha/2}\frac{S_\xi}{\sqrt{n}},\ \overline{\xi} + z_{\alpha/2}\frac{S_\xi}{\sqrt{n}}\right], \tag{4.10}$$

where $\overline{\xi} = \frac{1}{n}\sum_{i=1}^n \xi_i$ and $S_\xi^2 = \frac{1}{n}\sum_{i=1}^n(\xi_i - \overline{\xi})^2 = \overline{\xi}(1-\overline{\xi})$ with $\xi_i = \mathrm{I}(a < X_i < b)$.

In general, we can treat the interval (a,b) as "success", $p = \mathrm{P}$("success"), and $\overline{\xi} =$ relative frequence of "success".

Example 4.8. In a certain political campaign, one candidate has a poll taken at random among the voting population. The results are $n = 112$ and $y = 59$ (for "Yes"). Should the candidate feel very confident of winning?

Solution. Let $p = \mathrm{P}$("the condidate wins the campaign"). Then, $\overline{\xi} = 59/112 \approx 0.527$. According to (4.10), since $z_{0.025} \approx 1.96$, an approximated 95% confident interval estimate for p is

$$\begin{aligned} &\left[0.527 - z_{0.025}\sqrt{\frac{0.527*(1-0.527)}{112}},\ 0.527 + z_{0.025}\sqrt{\frac{0.527*(1-0.527)}{112}}\right] \\ &\approx [0.435, 0.619]. \end{aligned}$$

There has certain possibility that p is less than 50%, and the candidate should take this into account in campaigning. □

Chapter 5
Hypothesis testing

In scientific activities, much attention is devoted to answering questions about the validity of theories or hypotheses concerning physical phenomena. For examples, (i) Is the new drug effective in combating a certain disease? (ii) Are females more talented in music than males? e.t.c. To answer these questions, we need use the hypothesis testing, which is a procedure used to determine (make a decision) whether a hypothesis should be rejected (declared false) or not.

5.1 Basic concepts

To study hypothesis, the following technical terms are needed:

Hypothesis a statement (or claim) about a population;
Hypothesis test a rule that leads to a decision to or not to reject a hypothesis;
Simple hypothesis a hypothesis that completely specifies the distribution of the population;
Composite hypothesis a hypothesis that does not completely specify the distribution of the population;
Null hypothesis (H_0) a hypothesis that is assumed to be true before it can be rejected;
Alternative hypothesis (H_1 or H_a) a hypothesis that will be accepted if the null hypothesis is rejected.

Often, a hypothesis has the special form "the unknown distributional parameter θ belongs to a set". There may be two competing hypotheses of this form:

$$H_0 : \theta \in \Omega_0 \text{ versus } H_1 : \theta \in \Omega_1,$$

where Ω_0 and Ω_1 are disjoint sets of possible values of the parameter θ.

Example 5.1. Suppose that the score of STAT2602 follows $\mathrm{N}(\theta, 10)$, and we want to know whether the theoretical mean $\theta = 80$. In this case,

$$H_0 : \theta = 80 \text{ versus } H_1 : \theta \neq 80.$$

Here, H_0 is a simple hypothesis, because θ is the only unknown parameter and Ω_0 consists of exactly one real number; and H_1 is a composite hypothesis, because it can not completely specify the distribution of the score. □

If both the two hypotheses are simple, the null hypothesis H_0 is usually chosen to be a kind of default hypothesis, which one tends to believe unless given strong evidence otherwise.

Example 5.2. Suppose that the score of STAT2602 follows $\mathrm{N}(\theta, 10)$, and we want to know whether the theoretical mean $\theta = 80$ or 70. In this case,

$$H_0 : \theta = 80 \text{ versus } H_1 : \theta = 70.$$

Here, both H_0 and H_1 are simple hypotheses. We tend to believe $H_0 : \theta = 80$ unless given strong evidence otherwise. □

In order to construct a rule to decide whether the hypothesis is rejected or not, we need to use the test statistic defined by

Test statistic the statistic upon which the statistical decision will be based.

Usually, the test statistic is a functional on the random sample $\mathbf{X} = \{X_1, \cdots, X_n\}$, and it is denoted by $W(\mathbf{X})$. Some important terms about the test statistic are as follows:

Rejection region or critical region the set of values of the test statistic for which the null hypothesis is rejected;
Acceptance region the set of values of the test statistic for which the null hypothesis is not rejected (is accepted);
Type I error rejection of the null hypothesis when it is true;
Type II error acceptance of the null hypothesis when it is false.

	Accept H_0	Reject H_0
H_0 is true	No error	Type I error
H_0 is false	Type II error	No error

Suppose that the rejection region is

$$\{W(\mathbf{X}) \in R\}.$$

Definition 5.1. *(***Power function***) The power function $\pi(\theta)$ is the probability of rejecting H_0 when the true value of the parameter is θ, i.e.,*

$$\pi(\theta) := \mathrm{P}_\theta(W(\mathbf{X}) \in R).$$

Let $\alpha(\theta)$ and $\beta(\theta)$ be probabilities of committing a type I and type II error respectively when the true value of the parameter is θ. That is,

$$\begin{aligned} \alpha(\theta) &= \mathrm{P}_\theta(W(\mathbf{X}) \in R) \;\; \text{for } \theta \in \Omega_0; \\ \beta(\theta) &= \mathrm{P}_\theta(W(\mathbf{X}) \in R^c) \;\; \text{for } \theta \in \Omega_1. \end{aligned}$$

From $\alpha(\theta)$ and $\beta(\theta)$, we know that

$$\pi(\theta) = \begin{cases} \alpha(\theta), & \text{for } \theta \in \Omega_0; \\ 1 - \beta(\theta), & \text{for } \theta \in \Omega_1. \end{cases}$$

Example 5.3. A manufacturer of drugs has to decide whether 90% of all patients given a new drug will recover from a certain disease. Suppose
(a) the alternative hypothesis is that 60% of all patients given the new drug will recover;
(b) the test statistic is W, the observed number of recoveries in 20 trials;
(c) he will accept the null hypothesis when $W > 14$ and reject it otherwise.
Find the power function of W.

Solution: Let $p = \mathrm{P}(\text{"recovery"})$. The hypotheses are

$$H_0 : p = 0.9 \text{ versus } H_1 : p = 0.6.$$

The test statistic W follows a binomial distribution $\mathrm{B}(n, p)$ with parameters $n = 20$ and p. The rejection region is $\{W \le 14\}$. Hence,

$$\begin{aligned}\pi(p) &= \mathrm{P}_p(W \le 14)\\ &= 1 - \mathrm{P}_p(W > 14)\\ &= 1 - \sum_{k=15}^{20} \binom{20}{k} p^k (1-p)^{20-k}\\ &\approx \begin{cases} 0.0113, & \text{for } p = 0.9;\\ 0.8744, & \text{for } p = 0.6. \end{cases}\end{aligned}$$

(This implies that the probability of committing a type I and type II error are 0.0113 and 0.1256, respectively.) □

Example 5.4. Let $\mathbf{X}$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. Consider a test statistic $W = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ for hypotheses $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$. Assume that the rejection region is $\{W \ge K\}$. Then, the power function is

$$\begin{aligned}\pi(\mu) &= \mathrm{P}_\mu(W \ge K)\\ &= \mathrm{P}_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ge K + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right)\\ &= \mathrm{P}\left(Z \ge K + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right),\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. It is easy to see that

$$\lim_{\mu\to-\infty} \pi(\mu) = 0, \ \lim_{\mu\to\infty} \pi(\mu) = 1, \ \text{ and } \ \pi(\mu_0) = \alpha \text{ if } \mathrm{P}(Z \ge K) = \alpha.$$

□

The ideal power function is 0 for $\theta \in \Omega_0$ and 1 for $\theta \in \Omega_1$. However, this ideal can not be attained in general. For a fixed sample size, it is usually impossible to make both types of error probability arbitrarily small. In searching for a good test, it is common to restrict consideration to tests that control the type I error probability at a specified level. Within this class of tests we then search for tests that have type II error probability that is as small as possible. The *size* defined below is used to control the type I error probability.

Definition 5.2. *(**Size**) For $\alpha \in [0,1]$, a test with power function $\pi(\theta)$ is a size α test if*

$$\max_{\theta\in\Omega_0} \pi(\theta) = \alpha.$$

Remark 5.1. α is also called the *level of significance* or *significance level*. If H_0 is a simple hypothesis $\theta = \theta_0$, then $\alpha = \pi(\theta_0)$.

Example 5.5. Suppose that we want to test the null hypothesis that the mean of a normal population with $\sigma^2 = 1$ is μ_0 against the alternative hypothesis that it is μ_1, where $\mu_1 > \mu_0$.
(a) Find the value of K such that $\{\overline{X} \geq K\}$ provides a rejection region with the level of significance $\alpha = 0.05$ for a random sample of size n.
(b) For the rejection region found in (a), if $\mu_0 = 10$, $\mu_1 = 11$ and we need the type II probability $\beta \leq 0.06$, what should n be?

Solution. (a) Note that $H_0 : \mu = \mu_0$, $H_1 : \mu = \mu_1$, and the rejection region is $\{\overline{X} \geq K\}$. By definition,

$$\begin{aligned}\alpha = \pi(\mu_0) &= \mathrm{P}_{\mu_0}(\overline{X} \geq K)\\ &= \mathrm{P}_{\mu_0}\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq \frac{K-\mu_0}{\sigma/\sqrt{n}}\right)\\ &= \mathrm{P}\left(Z \geq \frac{K-\mu_0}{1/\sqrt{n}}\right),\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. Hence, when $\alpha = 0.05$, we should have

$$0.05 = \mathrm{P}\left(Z \geq \frac{K-\mu_0}{1/\sqrt{n}}\right),$$

which is equivalent to

$$\frac{K-\mu_0}{1/\sqrt{n}} = z_{0.05} \approx 1.645 \quad \text{or} \quad K \approx \mu_0 + \frac{1.645}{\sqrt{n}}.$$

(b) By definition,

$$\begin{aligned}\beta = 1 - \pi(\mu_1) &= \mathrm{P}_{\mu_1}(\overline{X} < K)\\ &= \mathrm{P}_{\mu_1}\left(\frac{\overline{X}-\mu_1}{\sigma/\sqrt{n}} < \frac{K-\mu_1}{\sigma/\sqrt{n}}\right)\\ &= \mathrm{P}\left(Z < \frac{K-\mu_1}{\sigma/\sqrt{n}}\right)\\ &\approx \mathrm{P}\left(Z < \frac{\mu_0 + \frac{1.645}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}}\right).\end{aligned}$$

With $\mu_0 = 10$, $\mu_1 = 11$, $\sigma^2 = 1$, it follows that

$$\beta \approx \mathrm{P}\left(Z < \sqrt{n}(\mu_0 - \mu_1) + 1.645\right) = \mathrm{P}\left(Z < -\sqrt{n} + 1.645\right).$$

Hence,

$$\begin{aligned}\beta \leq 0.06 &\iff -\sqrt{n} + 1.645 \leq -z_{0.06} \approx -1.555\\ &\iff n \geq (1.645 + 1.555)^2 \approx 10.24,\end{aligned}$$

that is, $n \geq 11$. □

Remark 5.2. In the above example, the value of K in the rejection region $\{\overline{X} \geq K\}$ is determined by the significance level α. For the test statistic $\overline{X}$, the value of K uniquely decides whether the null hypothesis is rejected or not, and it is usually called a *critical value* of this test.

In summary, steps to perform a hypothesis test are as follows:
(1) State the null and alternative hypotheses and the level of significance α.
(2) Choose a test statistic.
(3) Determine the rejection region.
(4) Calculate the value of the test statistic according to the particular sample drawn.
(5) Make a decision: reject H_0 if and only if the value of the test statistic falls in the rejection region.

The key steps are (2) and (3), and they can be accomplished by using the likelihood ratio or generalized likelihood ratio. The test statistic, denoted by $W(\mathbf{X})$, is chosen case by case. The rejection region of $W(\mathbf{X})$ usually has the form of $\{W(\mathbf{X}) \leq K\}$, $\{W(\mathbf{X}) \geq K\}$, $\{|W(\mathbf{X})| \geq K\}$, or $\{W(\mathbf{X}) \leq K_1\} \cup \{W(\mathbf{X}) \geq K_2\}$, where the values of K, K_1 and K_2 are determined by the significance level α.

Instead of using step (5), we can also use *p-value* to make a decision.

Definition 5.3. *(***p-value***) Let $W(\mathbf{x})$ be the observed value of the test statistic $W(\mathbf{X})$.*
Case 1: The rejection region is $\{W(\mathbf{X}) \leq K\}$, then

$$p\text{-}value = \max_{\theta \in \Omega_0} \mathrm{P}_\theta(W(\mathbf{X}) \leq W(\mathbf{x}));$$

Case 2: The rejection region is $\{W(\mathbf{X}) \geq K\}$, then

$$p\text{-}value = \max_{\theta \in \Omega_0} \mathrm{P}_\theta(W(\mathbf{X}) \geq W(\mathbf{x}));$$

Case 3: The rejection region is $\{|W(\mathbf{X})| \geq K\}$, then

$$p\text{-}value = \max_{\theta \in \Omega_0} \mathrm{P}_\theta(|W(\mathbf{X})| \geq |W(\mathbf{x})|).$$

From the above example, we know that p-value does not depend on α, and it helps us to make a decision by comparing its value with α.

Property 5.1. *For a test statistic $W(\mathbf{X})$,*

$$H_0 \textit{ is rejected at the significance level } \alpha \iff p\text{-}value \leq \alpha.$$

Proof. We only prove it for Case 1. Note that

$$p\text{-value} = \max_{\theta \in \Omega_0} \mathrm{P}_\theta(W(\mathbf{X}) \leq W(\mathbf{x})) \text{ and } \alpha = \max_{\theta \in \Omega_0} \mathrm{P}_\theta(W(\mathbf{X}) \leq K).$$

By the monotonicity of the c.d.f., we have

$$\begin{aligned} p\text{-value} \leq \alpha &\iff W(\mathbf{x}) \leq K \\ &\iff \text{ the observed value of } W(\mathbf{X}) \text{ falls in the rejection region} \\ &\iff H_0 \text{ is rejected at the significance level } \alpha. \end{aligned}$$

□

Example 5.3. (con't) If the observed value of W is 12, then

$$\begin{aligned} p\text{-value} &= \mathrm{P}_p(W \le 12) \text{ for } p = 0.9 \\ &= \sum_{k=0}^{12} \binom{20}{k} (0.9)^k (0.1)^{20-k} \approx 0.0004. \end{aligned}$$

Hence, at the significance level $\alpha = 0.05$, the null hypothesis is rejected. □

Example 5.5. (con't) If the observed value of $\overline{X}$ is 10.417, then

$$\begin{aligned} p\text{-value} &= \mathrm{P}_\mu(\overline{X} \ge 10.417) \text{ for } \mu = 10 \\ &= \mathrm{P}_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ge \frac{10.417-\mu}{\sigma/\sqrt{n}}\right) \text{ for } \mu = 10 \\ &= \mathrm{P}\left(Z \ge \frac{10.417-10}{1/\sqrt{11}}\right) \\ &\approx 0.0833. \end{aligned}$$

Hence, at the significance level $\alpha = 0.05$, the null hypothesis $H_0 : \mu = 10$ is not rejected. □

5.2 Most powerful tests

Definition 5.4. (**Most powerful tests**) *A test concerning a simple null hypothesis $\theta = \theta_0$ against a simple alternative hypothesis $\theta = \theta_1$ is said to be most powerful if the power of the test at $\theta = \theta_1$ is a maximum.*

To construct a most powerful rejection region we refer to the likelihood function of a random sample $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ defined by

$$L(\theta) = \mathbf{f}(X_1, X_2, \ldots, X_n; \theta),$$

where $\mathbf{f}(x_1, x_2, \ldots, x_n; \theta)$ is the joint p.d.f. of the random variables $X_1, X_2, \ldots, X_n$ from a population with a parameter θ. Consider the **likelihood ratio** $\frac{L(\theta_0)}{L(\theta_1)}$. Intuitively speaking, the null hypothesis should be rejected when the likelihood ratio is small.

Theorem 5.1. *(Neyman-Pearson Lemma) Suppose $X_1, X_2, \ldots, X_n$ constitute a random sample of size n from a population with exactly one unknown parameter θ. Suppose that there is a positive constant k and a region C such that*

$$(i)\ \mathrm{P}_\theta\{(X_1, X_2, \ldots, X_n) \in C\} = \alpha \ \text{ for } \theta = \theta_0,$$

and

$$\begin{aligned} (ii)\ & \frac{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)}{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_1)} \le k \qquad \text{when } (x_1, x_2, \ldots, x_n) \in C, \\ (iii)\ & \frac{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)}{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_1)} \ge k \qquad \text{when } (x_1, x_2, \ldots, x_n) \notin C. \end{aligned}$$

Construct a test, called the ***likelihood ratio test****, which rejects $H_0 : \theta = \theta_0$ and accepts $H_1 : \theta = \theta_1$ if and only if $(X_1, X_2, \ldots, X_n) \in C$. Then any other test which has significance level $\alpha^* \le \alpha$ has power not more than that of this likelihood ratio test. In other words, the likelihood ratio test is most powerful among all tests having significance level $\alpha^* \le \alpha$.*

Proof. Suppose D is the rejection region of any other test which has significance level $\alpha^* \le \alpha$. We consider first the continuous case. Note that

$$\begin{aligned}\alpha &= \mathrm{P}_\theta\{(X_1, X_2, \ldots, X_n) \in C\} \quad \text{for } \theta = \theta_0 \text{ (by (i))}\\ &= \underset{C}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n;\\ \alpha^* &= \mathrm{P}_\theta\{(X_1, X_2, \ldots, X_n) \in D\} \quad \text{for } \theta = \theta_0\\ &= \underset{D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n.\end{aligned}$$

Since $\alpha \ge \alpha^*$, it follows that

$$\begin{aligned}&\underset{C}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n\\ \ge &\underset{D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n.\end{aligned}$$

Subtracting

$$\underset{C\cap D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n,$$

we get

$$\begin{aligned}&\underset{C\cap D'}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n\\ \ge &\underset{C'\cap D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n, \end{aligned} \tag{5.1}$$

where C' and D' are complements of C and D respectively. Hence,

$$\begin{aligned}&\underset{C\cap D'}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_1)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n\\ \ge &\underset{C\cap D'}{\int \cdots \int \int} \frac{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)}{k}\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n \text{ (by (ii))}\\ \ge &\underset{C'\cap D}{\int \cdots \int \int} \frac{\mathbf{f}(x_1, x_2, \ldots, x_n; \theta_0)}{k}\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n \text{ (by (5.1))}\\ \ge &\underset{C'\cap D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_1)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n. \text{ (by (iii))}\end{aligned}$$

Adding $\underset{C\cap D}{\int \cdots \int \int} \mathbf{f}(x_1, x_2, \ldots, x_n; \theta_1)\mathrm{d}x_1\mathrm{d}x_2 \cdots \mathrm{d}x_n$, we finally obtain

$$\begin{aligned}&\int\cdots\underset{C}{\int}\int \mathbf{f}(x_1,x_2,\ldots,x_n;\theta_1)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n\\ \geq &\int\cdots\underset{D}{\int}\int \mathbf{f}(x_1,x_2,\ldots,x_n;\theta_1)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n,\end{aligned}$$

or

$$\mathrm{P}_\theta\left\{(X_1,X_2,\ldots,X_n)\in C\right\}\geq \mathrm{P}_\theta\left\{(X_1,X_2,\ldots,X_n)\in D\right\}$$

for $\theta=\theta_1$. The last inequality states that the power of the likelihood ratio test at $\theta=\theta_1$ is at least as much as that corresponding to the rejection region D. The proof for discrete case is similar, with sums taking places of integrals. □

Neyman-Pearson Lemma says that to test $H_0:\theta=\theta_0$ versus $H_1:\theta=\theta_1$, the rejection region for the likelihood ratio test is

$$\frac{L(\theta_0)}{L(\theta_1)}\leq k\Longleftrightarrow (X_1,X_2,\cdots,X_n)\in C\Longleftrightarrow W(\mathbf{X})\in R,$$

where the interval R is chosen so that the test has the significance level α. Generally speaking, the likelihood ratio helps us to determine the test statistic and the form of its rejection region.

Example 5.6. A random sample $\{X_1,X_2,\ldots,X_n\}$ from a normal population $\mathrm{N}(\mu,\sigma^2)$, where $\sigma^2=\sigma_0^2$ is known, is to be used to test the null hypothesis $\mu=\mu_0$ against the alternative hypothesis $\mu=\mu_1$, where $\mu_1>\mu_0$. Use the Neyman-Pearson Lemma to construct the most powerful test.

Solution. The likelihood function of the sample is

$$L(\mu)=\left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n\exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n(X_i-\mu)^2\right].$$

The likelihood ratio test rejects the null hypothesis $\mu=\mu_0$ if and only if $\dfrac{L(\mu_0)}{L(\mu_1)}\leq k$, that is,

$$\begin{aligned}&\exp\left\{\frac{1}{2\sigma_0^2}\sum_{i=1}^n\left[(X_i-\mu_1)^2-(X_i-\mu_0)^2\right]\right\}\leq k\\ \Longleftrightarrow &\sum_{i=1}^n(-2\mu_1X_i+\mu_1^2+2\mu_0X_i-\mu_0^2)\leq 2\sigma_0^2\log k\\ \Longleftrightarrow &\, n(\mu_1^2-\mu_0^2)+2(\mu_0-\mu_1)\sum_{i=1}^nX_i\leq 2\sigma_0^2\log k\\ \Longleftrightarrow &\,\overline{X}\geq\frac{2\sigma_0^2\log k-n(\mu_1^2-\mu_0^2)}{2n(\mu_0-\mu_1)}\qquad(\text{since }\mu_1>\mu_0).\end{aligned}$$

Therefore, in order that the level of significance is α, we should choose a constant K such that $\mathrm{P}_\mu(\overline{X}\geq K)=\alpha$ for $\mu=\mu_0$, that is,

$$\mathrm{P}\left(\frac{\overline{X}-\mu_0}{\sigma_0/\sqrt{n}}\geq\frac{K-\mu_0}{\sigma_0/\sqrt{n}}\right)=\alpha\Longleftrightarrow\frac{K-\mu_0}{\sigma_0/\sqrt{n}}=z_\alpha\Longleftrightarrow K=\mu_0+\frac{\sigma_0z_\alpha}{\sqrt{n}}.$$

Therefore, the most powerful test having significance level $\alpha^* \leq \alpha$ is the one which has the rejection region

$$\left\{\overline{X} \geq \mu_0 + \frac{\sigma_0 z_\alpha}{\sqrt{n}}\right\} \quad \text{or} \quad \left\{\frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq z_\alpha\right\}.$$

(Note that the rejection region found does not depend on the value of μ_1). □

Example 5.7. Suppose $X_1, X_2, \ldots, X_n$ constitute a random sample of size n from a population given by a density

$$f(x) = \theta x^{\theta-1} \mathrm{I}(0 \leq x \leq 1).$$

If $0 \leq X_i \leq 1$ for $i = 1, 2, \ldots, n$, find the form of the most powerful test for testing

$$H_0 : \theta = 2 \text{ versus } H_1 : \theta = 1.$$

Solution. The likelihood function of the sample is

$$L(\theta) = \theta^n \left(\prod_{i=1}^n X_i\right)^{\theta-1} \prod_{i=1}^n \mathrm{I}(0 \leq X_i \leq 1) = \theta^n \left(\prod_{i=1}^n X_i\right)^{\theta-1}.$$

Hence, the likelihood ratio is

$$\frac{L(2)}{L(1)} = 2^n \prod_{i=1}^n X_i.$$

The likelihood ratio test rejects H_0 if and only if $2^n \prod_{i=1}^n X_i \leq K$ where K is a positive constant (or, equivalently, $\prod_{i=1}^n X_i \leq k$ where k is a positive constant). □

5.3 Generalized likelihood ratio tests: One-sample case

The Neyman-Pearson lemma provides a method of constructing most powerful rejection regions for testing a simple null hypothesis against a simple alternative hypothesis, but it does not always apply to composite hypotheses. We shall present a general method for constructing rejection regions for tests of composite hypotheses which in most cases have very satisfactory properties, although they are not necessarily uniformly most powerful.

Suppose that $\theta \in \Omega$, where Ω is the parametric space. Consider the following hypotheses:

$$H_0 : \theta \in \Omega_0 \quad \text{versus} \quad H_1 : \theta \in \Omega_1,$$

where Ω_1 is the complement of Ω_0 with respect to Ω (i.e., $\Omega_1 = \Omega/\Omega_0$). Let

$$L(\Omega_0) = \max_{\theta \in \Omega_0} L(\theta) \quad \text{and} \quad L(\Omega) = \max_{\theta \in \Omega} L(\theta).$$

The **generalized likelihood ratio** is defined as

$$\Lambda = \frac{L(\Omega_0)}{L(\Omega)}.$$

Since Ω_0 is a subset of Ω, it follows that $\Lambda \leq 1$. When the null hypothesis is false, we would expect Λ to be small. A generalized likelihood ratio test states, therefore, that the null hypothesis H_0 is rejected if and only if Λ falls in a rejection region of the form $\Lambda \leq k$, where $0 \leq k \leq 1$.

5.3.1 Testing for the mean: Variance is known

Example 5.8. Find the generalized likelihood ratio test for testing

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu \neq \mu_0$$

on the basis of a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where $\sigma^2 = \sigma_0^2$ is known.

Solution. Ω is the set of all real numbers (i.e., $\Omega = \mathscr{R}$) and $\Omega_0 = \{\mu_0\}$. On one hand, since Ω_0 contains only μ_0, it follows that

$$L(\Omega_0) = \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \mu_0)^2\right].$$

On the other hand, since the maximum likelihood estimator of μ is $\overline{X}$, it follows that

$$L(\Omega) = \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i - \overline{X})^2\right].$$

Hence,

$$\begin{aligned}\Lambda = \frac{L(\Omega_0)}{L(\Omega)} &= \exp\left\{-\frac{1}{2\sigma_0^2}\left[\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \overline{X})^2\right]\right\} \\ &= \exp\left[-\frac{n(\overline{X} - \mu_0)^2}{2\sigma_0^2}\right].\end{aligned}$$

Therefore, the rejection region is $\left\{|\overline{X} - \mu_0| \geq K\right\}$. In order that the level of significance is α, that is,

$$\mathrm{P}_\mu\left(|\overline{X} - \mu_0| \geq K\right) = \alpha \quad \text{for } \mu = \mu_0,$$

we should let $K = z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}$, so that

$$\begin{aligned}\mathrm{P}_\mu\left(|\overline{X} - \mu_0| \geq K\right) &= \mathrm{P}_\mu\left(|\overline{X} - \mu_0| \geq z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\ &= \mathrm{P}_\mu\left(\frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq z_{\alpha/2}\right) + \mathrm{P}_\mu\left(\frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \leq -z_{\alpha/2}\right) \\ &= \mathrm{P}\left(Z \geq z_{\alpha/2}\right) + \mathrm{P}\left(Z \leq -z_{\alpha/2}\right) \\ &= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha\end{aligned}$$

for $\mu = \mu_0$. So, the generalized likelihood ratio test has the rejection region

$$\left\{ \frac{|\overline{X} - \mu_0|}{\sigma_0/\sqrt{n}} \geq z_{\alpha/2} \right\}$$

at the significance level α. □

From aforemention example and the similar technique, we can have the following table:

Table 5.1 Testing for the mean when $\sigma^2 = \sigma_0^2$ is known

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$\left\{ \left| \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \right| \geq z_{\alpha/2} \right\}$	$\mathrm{P}\left(\lvert Z \rvert \geq \left| \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right| \right)$
Left-tailed	$\mu = \mu_0$ or $\mu \geq \mu_0$	$\mu < \mu_0$	$\left\{ \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \leq -z_\alpha \right\}$	$\mathrm{P}\left(Z \leq \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right)$
Right-tailed	$\mu = \mu_0$ or $\mu \leq \mu_0$	$\mu > \mu_0$	$\left\{ \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq z_\alpha \right\}$	$\mathrm{P}\left(Z \geq \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} \right)$

Example 5.9. The standard deviation of the annual incomes of government employees is \$1400. The mean is claimed to be \$35,000. Now a sample of 49 employees has been drawn and their average income is \$35,600. At the 5% significance level, can you conclude that the mean annual income of all government employees is not \$35,000?

Solution 1.

Step 1: “The mean ... is not 35,000” can be written as “$\mu \neq 35000$”, while “the mean ... is 35,000” can be written as “$\mu = 35000$”. Since the null hypothesis should include an equality, we consider hypothesis:

$$H_0 : \mu = 35000 \text{ versus } H_1 : \mu \neq 35000.$$

Step 2: The test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - 35000}{1400/\sqrt{49}} = \frac{\overline{X} - 35000}{200},$$

which follows $N(0,1)$ under H_0.

Step 3: At the significance level $\alpha = 5\%$, the rejection region is

$$\left\{ |Z| \geq z_{\alpha/2} \right\} \approx \{ |Z| \geq 1.960 \}.$$

Step 4: Since $\bar{x} = 35600$, the value of the test statistic is

$$\frac{35600 - 35000}{200} = 3.$$

Step 5: Since $|3| \geq 1.960$, we reject H_0 and accept H_1. Therefore, we conclude that the mean annual income of all government employees is not \$35,000 at the 5% level of significance.

Solution 2.

Step 1:

$$H_0 : \mu = 35000 \text{ versus } H_1 : \mu \neq 35000.$$

Step 2: The test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{X} - 35000}{1400/\sqrt{49}} = \frac{\overline{X} - 35000}{200},$$

which follows $N(0,1)$ under H_0.

Step 3: Since $\bar{x} = 35600$, the value of the test statistic is

$$\frac{35600 - 35000}{200} = 3.$$

Step 4:

$$p\text{-value} = \mathrm{P}(|Z| \geq |3|) = 2\mathrm{P}(Z \geq 3) \approx 2(0.5 - 0.4987) = 0.0026,$$

where Z follows $N(0,1)$.

Step 5: Since $0.0026 \leq 0.05 = \alpha$, we reject H_0 and accept H_1. Therefore we conclude that the mean annual income of all government employees is not \$35,000. □

Example 5.10. The chief financial officer in FedEx believes that including a stamped self-addressed envelope in the monthly invoice sent to customers will reduce the amount of time it takes for customers to pay their monthly bills. Currently, customers return their payments in 24 days on average, with a standard deviation of 6 days. It was calculated that an improvement of two days on average would cover the costs of the envelopes (because cheques can be deposited earlier). A random sample of 220 customers was selected and stamped self-addressed envelopes were included in their invoice packs. The amounts of time taken for these customers to pay their bills were recorded and their mean is 21.63 days. Assume that the corresponding population standard deviation is still 6 days. Can the chief financial officer conclude that the plan will be profitable at the 10% significance level?

Solution 1. The plan will be profitable when "$\mu < 22$", and not profitable when "$\mu \geq 22$". Since the null hypothesis should include an equality, we have

$$H_0 : \mu \geq 22 \text{ versus } H_1 : \mu < 22.$$

The value of the test statistic is

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{21.63 - 22}{6/\sqrt{220}} \approx -0.9147.$$

Since $-0.9147 > -1.282 \approx -z_{0.1}$, H_0 should not be rejected. The chief financial officer cannot conclude that the plan is profitable at the 10% significance level.

Solution 2: Consider

$$p\text{-value} \approx \mathrm{P}(Z \leq -0.9147) \approx 0.1814 > 0.1 = \alpha,$$

where Z follows $N(0,1)$. Therefore, H_0 should not be rejected. The chief financial officer cannot conclude that the plan is profitable at the 10% significance level. □

5.3.2 Testing for the mean: Variance is unknown

Example 5.11. Find the generalized likelihood ratio test for testing

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu > \mu_0$$

on the basis of a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$.

Solution. Now

$$\begin{aligned} \Omega &= \{(\mu,\sigma) : \mu \geq \mu_0, \sigma > 0\}, \\ \Omega_0 &= \{(\mu,\sigma) : \mu = \mu_0, \sigma > 0\}, \\ \Omega_1 &= \{(\mu,\sigma) : \mu > \mu_0, \sigma > 0\}. \end{aligned}$$

The likelihood function of the sample is

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

and hence,

$$\begin{aligned} \frac{\partial \ln L(\mu,\sigma)}{\partial \mu} &= \frac{n}{\sigma^2}(\overline{X}-\mu), \\ \frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n (X_i-\mu)^2. \end{aligned}$$

On Ω_0, the maximum value of $L(\mu,\sigma)$ is $L(\mu_0,\tilde{\sigma})$, where $\tilde{\sigma}$ satisfies

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

This is because $\tilde{\sigma}$ is the maximum value of $\ln L(\mu_0,\sigma)$, by noting that for all $\mu > 0$,

$$\begin{aligned} \sigma < \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2} &\iff \frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} > 0, \\ \sigma > \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2} &\iff \frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} < 0. \end{aligned}$$

Therefore,

$$L(\Omega_0) = L(\mu_0,\tilde{\sigma}) = \left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right).$$

On Ω, the maximum value of $L(\mu,\sigma)$ is $L(\hat{\mu},\hat{\sigma})$, where (*noting that $L(\mu,\sigma)$ decreases with respect to μ when $\mu > \overline{X}$ and increases with respect to μ when $\mu < \overline{X}$*)

$$\hat{\mu} = \begin{cases} \mu_0, & \text{if } \overline{X} \leq \mu_0; \\ \overline{X}, & \text{if } \overline{X} > \mu_0, \end{cases}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Therefore,

$$L(\Omega) = L(\hat{\mu}, \hat{\sigma}) = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right).$$

Thus, we have

$$\Lambda = \frac{L(\Omega_0)}{L(\Omega)} = \left(\frac{\hat{\sigma}}{\tilde{\sigma}}\right)^n = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{n/2} = \begin{cases} 1, & \text{if } \overline{X} \le \mu_0; \\ \left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\mu_0)^2}\right]^{n/2}, & \text{if } \overline{X} > \mu_0. \end{cases}$$

The rejection region is $\{\Lambda \le k\}$ for some nonnegative constant $k < 1$ (since we do not want α to be 1). Then $\{\Lambda \le k\} \subseteq \{\overline{X} > \mu_0\}$ and $\Lambda \le k$ is equivalent to

$$k^{2/n} \ge \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \mu_0)^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2} = \frac{1}{1 + \frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}},$$

that is,

$$\frac{(\overline{X} - \mu_0)^2}{S^2} = \frac{n(\overline{X} - \mu_0)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \ge k^{-2/n} - 1,$$

or (since $\overline{X} > \mu_0$)

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \ge c,$$

where c is the constant $\sqrt{(n-1)(k^{-2/n} - 1)}$. In order that the level of significance is α, that is,

$$\mathrm{P}_{(\mu,\sigma)}\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \ge c\right) = \alpha \quad \text{for } \mu = \mu_0,$$

we should let $c = t_{\alpha,n-1}$, since

$$\mathrm{P}_{(\mu,\sigma)}\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \ge t_{\alpha,n-1}\right) = \mathrm{P}\left(t_{n-1} \ge t_{\alpha,n-1}\right) \quad \text{for } \mu = \mu_0$$

by Property 4.1(iii). So, the generalized likelihood ratio test has the rejection region

$$\left\{\frac{\overline{X} - \mu_0}{S/\sqrt{n-1}} \ge t_{\alpha,n-1}\right\}$$

at the significance level α. □

Example 5.12. What will happen if we change H_0 in the previous example to be $\mu \le \mu_0$?

Solution. Note that

$$\begin{aligned}\Omega &= \{(\mu,\sigma): -\infty < \mu < \infty, \sigma > 0\},\\ \Omega_0 &= \{(\mu,\sigma): \mu \le \mu_0, \sigma > 0\},\\ \Omega_1 &= \{(\mu,\sigma): \mu > \mu_0, \sigma > 0\}.\end{aligned}$$

On Ω_0, the maximum value of $L(\mu,\sigma)$ is $L(\tilde{\mu},\tilde{\sigma})$ where

$$\tilde{\mu} = \begin{cases} \overline{X}, & \text{if } \overline{X} < \mu_0;\\ \mu_0, & \text{if } \overline{X} \ge \mu_0,\end{cases}$$

and

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{\mu})^2.$$

Therefore,

$$L(\Omega_0) = L(\tilde{\mu},\tilde{\sigma}) = \left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right).$$

On Ω, the maximum value of $L(\mu,\sigma)$ is $L(\hat{\mu},\hat{\sigma})$, where

$$\hat{\mu} = \overline{X} \qquad \text{and} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Therefore,

$$L(\Omega) = L(\hat{\mu},\hat{\sigma}) = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right).$$

Thus, we have

$$\Lambda = \frac{L(\Omega_0)}{L(\Omega)} = \left(\frac{\hat{\sigma}}{\tilde{\sigma}}\right)^n = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{n/2} = \begin{cases} 1, & \text{if } \overline{X} \le \mu_0;\\ \left[\dfrac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\mu_0)^2}\right]^{n/2}, & \text{if } \overline{X} > \mu_0.\end{cases}$$

So the generalized likelihood ratio test is the same as that in the previous example. □

From aforemention two examples and the similar technique, we can have the following table:

Table 5.2 Testing for the mean when σ^2 is unknown

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\mu = \mu_0$	$\mu \ne \mu_0$	$\left\{\left\lvert\frac{\overline{X}-\mu_0}{S/\sqrt{n-1}}\right\rvert \ge t_{\alpha/2,n-1}\right\}$	$\mathrm{P}\left(\lvert t_{n-1}\rvert \ge \left\lvert\frac{\bar{x}-\mu_0}{s/\sqrt{n-1}}\right\rvert\right)$
Left-tailed	$\mu = \mu_0$ or $\mu \ge \mu_0$	$\mu < \mu_0$	$\left\{\frac{\overline{X}-\mu_0}{S/\sqrt{n-1}} \le -t_{\alpha,n-1}\right\}$	$\mathrm{P}\left(t_{n-1} \le \frac{\bar{x}-\mu_0}{s/\sqrt{n-1}}\right)$
Right-tailed	$\mu = \mu_0$ or $\mu \le \mu_0$	$\mu > \mu_0$	$\left\{\frac{\overline{X}-\mu_0}{S/\sqrt{n-1}} \ge t_{\alpha,n-1}\right\}$	$\mathrm{P}\left(t_{n-1} \ge \frac{\bar{x}-\mu_0}{s/\sqrt{n-1}}\right)$

Example 5.13. According to the last census in a city, the mean family annual income was 316 thousand dollars. A random sample of 900 families taken this year produced a mean family annual income of 313 thousand dollars and a standard deviation of 70 thousand dollars. At the 2.5% significance level, can we conclude that the mean family annual income has declined since the last census?

Solution. Consider hypothesis

$$H_0 : \mu \geq 316 \ \text{ versus } \ H_1 : \mu < 316.$$

The value of the test statistic is

$$\frac{\bar{x}-\mu_0}{s/\sqrt{n-1}} = \frac{313-316}{70/\sqrt{900-1}} \approx -1.286.$$

Since $-1.286 > -1.963 = -t_{0.025,899}$, we do not reject H_0. Thus we cannot conclude that the mean family annual income has declined since the last census at the 2.5% level of significance. □

5.3.3 Testing for the variance

Example 5.14. Given a random sample of size n from a normal population with unknown mean and variance, find the generalized likelihood ratio test for testing the null hypothesis $\sigma = \sigma_0$ ($\sigma_0 > 0$) against the alternative hypothesis $\sigma \neq \sigma_0$.

Solution. Note that

$$\begin{aligned} \Omega &= \{(\mu,\sigma) : -\infty < \mu < \infty, \sigma > 0\}, \\ \Omega_0 &= \{(\mu,\sigma) : -\infty < \mu < \infty, \sigma = \sigma_0\}, \\ \Omega_1 &= \{(\mu,\sigma) : -\infty < \mu < \infty, \sigma > 0, \sigma \neq \sigma_0\}, \end{aligned}$$

and the likelihood function of the sample $\{X_1, \cdots, X_n\}$ is

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

On Ω_0, the maximum value of $L(\mu,\sigma)$ is $L(\tilde{\mu},\sigma_0)$ where $\tilde{\mu} = \overline{X}$. Therefore,

$$\begin{aligned} L(\Omega_0) &= L(\tilde{\mu},\sigma_0) \\ &= \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (X_i-\overline{X})^2\right]. \end{aligned}$$

On Ω, the maximum value of $L(\mu,\sigma)$ is $L(\hat{\mu},\hat{\sigma})$ where $\hat{\mu} = \overline{X}$ and $\hat{\sigma}^2 = S^2$. Therefore,

$$L(\Omega) = L(\hat{\mu},\hat{\sigma}) = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n \exp\left(-\frac{n}{2}\right).$$

Thus, we have

$$\begin{aligned}\Lambda = \frac{L(\Omega_0)}{L(\Omega)} &= \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{2\sigma_0^2}+\frac{n}{2}\right] \\ &= \left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n\sigma_0^2}\right]^{n/2} \exp\left[-\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{2\sigma_0^2}+\frac{n}{2}\right].\end{aligned}$$

The rejection region is $\{\Lambda \le k\}$ for some positive constant $k<1$ (since we do not want α to be 1). Letting $Y = \frac{1}{n\sigma_0^2}\sum_{i=1}^{n}(X_i-\overline{X})^2$,

$$\begin{aligned}\Lambda \le k &\Longleftrightarrow Y^{n/2}\exp\left(-\frac{nY}{2}+\frac{n}{2}\right) \le k, \\ &\Longleftrightarrow Y\exp(-Y+1) \le k^{2/n}, \\ &\Longleftrightarrow Y\exp(-Y) \le \frac{k^{2/n}}{e}.\end{aligned}$$

For $y>0$ define a function $g(y) = ye^{-y}$. Then,

$$\frac{\mathrm{d}g(y)}{\mathrm{d}y} = e^{-y} - ye^{-y} = (1-y)e^{-y}.$$

Since

$$y<1 \Longleftrightarrow \frac{\mathrm{d}g(y)}{\mathrm{d}y} > 0 \text{ and } y>1 \Longleftrightarrow \frac{\mathrm{d}g(y)}{\mathrm{d}y} < 0,$$

$g(y)$ will be small when y is close to zero or very large. Thus we reject the null hypothesis $\sigma = \sigma_0$ when the value of Y (or nY) is large or small, that is, the rejection region of our generalized likelihood ratio test has the rejection region:

$$\{nY \le K_1\} \cup \{nY \ge K_2\}.$$

Note that $nY = \frac{nS^2}{\sigma_0^2}$. In order that the level of significance is α, that is,

$$\mathrm{P}_{(\mu,\sigma)}\left(\frac{nS^2}{\sigma_0^2} \le K_1\right) + \mathrm{P}\left(\frac{nS^2}{\sigma_0^2} \ge K_2\right) = \alpha \quad \text{for } \sigma = \sigma_0,$$

we should let $K_1 = \chi^2_{1-\alpha/2,n-1}$ and $K_2 = \chi^2_{\alpha/2,n-1}$, since

$$\mathrm{P}_{(\mu,\sigma)}\left(\frac{nS^2}{\sigma_0^2} \le K_1\right) = \mathrm{P}\left(\chi^2_{n-1} \le \chi^2_{1-\alpha/2,n-1}\right) = \frac{\alpha}{2}$$

and

$$\mathrm{P}_{(\mu,\sigma)}\left(\frac{nS^2}{\sigma_0^2} \ge K_2\right) = \mathrm{P}\left(\chi^2_{n-1} \ge \chi^2_{\alpha/2,n-1}\right) = \frac{\alpha}{2}$$

for $\sigma = \sigma_0$ by using the fact that $nY \sim \chi^2_{n-1}$ from Property 4.1(ii). □

From the aforemention example and the similar technique, we can have the following table:

Table 5.3 Testing for the variance

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$\left\{\frac{nS^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2,n-1}\right\}$ $\cup\left\{\frac{nS^2}{\sigma_0^2} \geq \chi^2_{\alpha/2,n-1}\right\}$	$2\min\Big(\mathrm{P}\left(\chi^2_{n-1} \leq \frac{ns^2}{\sigma_0^2}\right),$ $\mathrm{P}\left(\chi^2_{n-1} \geq \frac{ns^2}{\sigma_0^2}\right)\Big)$
Left-tailed	$\sigma = \sigma_0$ or $\sigma \geq \sigma_0$	$\sigma < \sigma_0$	$\left\{\frac{nS^2}{\sigma_0^2} \leq \chi^2_{1-\alpha,n-1}\right\}$	$\mathrm{P}\left(\chi^2_{n-1} \leq \frac{ns^2}{\sigma_0^2}\right)$
Right-tailed	$\sigma = \sigma_0$ or $\sigma \leq \sigma_0$	$\sigma > \sigma_0$	$\left\{\frac{nS^2}{\sigma_0^2} \geq \chi^2_{\alpha,n-1}\right\}$	$\mathrm{P}\left(\chi^2_{n-1} \geq \frac{ns^2}{\sigma_0^2}\right)$

Example 5.15. One important factor in inventory control is the variance of the daily demand for the product. A manager has developed the optimal order quantity and reorder point, assuming that the variance is equal to 250. Recently, the company has experienced some inventory problems, which induced the operations manager to doubt the assumption. To examine the problem, the manager took a sample of 25 daily demands and found that $s^2 = 270.58$. Do these data provide sufficient evidence at the 5% significance level to infer that the management scientist's assumption about the variance is wrong?

Solution. Consider hypothesis

$$H_0 : \sigma^2 = 250 \ \text{ vesus } \ H_1 : \sigma^2 \neq 250.$$

The value of test statistic is

$$\frac{ns^2}{\sigma_0^2} = \frac{25 \times 270.58}{250} \approx 25.976.$$

Since $\chi^2_{1-0.05/2,25-1} \approx 12.401 \leq 25.976 \leq 39.364 \approx \chi^2_{0.05/2,25-1}$, we do not reject H_0. There is not sufficient evidence at the 5% significance level to infer that the management scientists assumption about the variance is wrong. □

5.3.4 Test and interval estimation

We can obtain the interval estimation by using the two-tailed hypothesis testing. For example, consider hypotheses

$$H_0 : \mu = \mu_0 \ \text{ versus } \ \mu \neq \mu_0.$$

If the variance is known, the acceptance region is

$$\left\{\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right|<z_{\alpha/2}\right\} \Longleftrightarrow \overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}<\mu_0<\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

at the significance level α. As H_0 is accepted, $\mu=\mu_0$ hence

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

That is, $\left[\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ is the $1-\alpha$ confidence interval of μ.

Similarly, we can find the confidence interval of μ when the variance is unknown, and σ^2 by using the two-tailed hypothesis testing.

5.4 Generalized likelihood ratio tests: Two-sample case

In this section, we assume that there are two populations following $\mathrm{N}(\mu_1,\sigma_1^2)$ and $\mathrm{N}(\mu_2,\sigma_2^2)$ respectively. A sample $\{X_i, i=1,2,\ldots,n_1\}$ is taken from the population $\mathrm{N}(\mu_1,\sigma_1^2)$ and a sample $\{Y_j, j=1,2,\ldots,n_2\}$ is taken from the population $\mathrm{N}(\mu_2,\sigma_2^2)$. Assume that these two samples are independent (that is, $X_1,X_2,\ldots,X_{n_1}, Y_1,Y_2,\ldots,Y_{n_2}$ are independent).

5.4.1 Testing for the mean: Variance is known

We first consider the hypothesis testing for $\mu_1-\mu_2$ when σ_1 and σ_2 are known.

Example 5.16. Assume that σ_1 and σ_2 are known. Find the generalized likelihood ratio for testing

$$H_0:\mu_1-\mu_2=\delta \qquad \text{versus} \qquad H_1:\mu_1-\mu_2\neq\delta.$$

Solution. Note that

$$\begin{aligned}
\Omega_0 &= \{(\mu_1,\mu_2):\mu_1-\mu_2=\delta\},\\
\Omega_1 &= \{(\mu_1,\mu_2):\mu_1-\mu_2\neq\delta\},\\
\Omega &= \Omega_0\cup\Omega_1=\{(\mu_1,\mu_2):-\infty<\mu_1<\infty,-\infty<\mu_2<\infty\}.
\end{aligned}$$

The likelihood function of the two samples is

$$\begin{aligned}
L(\mu_1,\mu_2) &= \left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^{n_1}\exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)^2\right]\\
&\quad\times\left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^{n_2}\exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(Y_j-\mu_2)^2\right].
\end{aligned}$$

On Ω_0, we have

$$\ln L(\mu_1,\mu_2)=\ln L(\mu_1,\mu_1-\delta)=C-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)^2-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(Y_j-\mu_1+\delta)^2,$$

where C depends on neither μ_1 nor μ_2. By direct calculation,

$$\begin{aligned}\frac{\partial}{\partial\mu_1}\ln L(\mu_1,\mu_1-\delta) &= \frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)+\frac{1}{\sigma_2^2}\sum_{j=1}^{n_2}(Y_j-\mu_1+\delta)\\ &= \frac{n_1(\overline{X}-\mu_1)}{\sigma_1^2}+\frac{n_2(\overline{Y}-\mu_1+\delta)}{\sigma_2^2}\\ &= \frac{n_1\overline{X}}{\sigma_1^2}+\frac{n_2(\overline{Y}+\delta)}{\sigma_2^2}-\left(\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}\right)\mu_1.\end{aligned}$$

Therefore, the maximum likelihood estimator of μ_1 is

$$\tilde{\mu}_1=\frac{\dfrac{n_1\overline{X}}{\sigma_1^2}+\dfrac{n_2(\overline{Y}+\delta)}{\sigma_2^2}}{\dfrac{n_1}{\sigma_1^2}+\dfrac{n_2}{\sigma_2^2}},$$

since

$$\begin{aligned}\mu_1<\tilde{\mu}_1 &\Longleftrightarrow \frac{\partial}{\partial\mu_1}\ln L(\mu_1,\mu_1-\delta)>0,\\ \mu_1>\tilde{\mu}_1 &\Longleftrightarrow \frac{\partial}{\partial\mu_1}\ln L(\mu_1,\mu_1-\delta)<0.\end{aligned}$$

On Ω, it is easy to see that the maximum likelihood estimator of μ_1 is $\overline{X}$ and that of μ_2 is $\overline{Y}$, since $\mu_1=\overline{X}$ maximizes

$$\left(\frac{1}{\sigma_1\sqrt{2\pi}}\right)^{n_1}\exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}(X_i-\mu_1)^2\right],$$

and $\mu_2=\overline{Y}$ maximizes

$$\left(\frac{1}{\sigma_2\sqrt{2\pi}}\right)^{n_2}\exp\left[-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}(Y_j-\mu_2)^2\right].$$

Thus, the generalized likelihood ratio is

$$\begin{aligned}\Lambda &= \frac{L(\Omega_0)}{L(\Omega)}\\ &= \exp\left[-\frac{1}{2\sigma_1^2}\sum_{i=1}^{n_1}\left[(X_i-\tilde{\mu}_1)^2-(X_i-\overline{X})^2\right]-\frac{1}{2\sigma_2^2}\sum_{j=1}^{n_2}\left[(Y_j-\tilde{\mu}_1+\delta)^2-(Y_j-\overline{Y})^2\right]\right]\\ &= \exp\left[-\frac{n_1(\overline{X}-\tilde{\mu}_1)^2}{2\sigma_1^2}-\frac{n_2(\overline{Y}-\tilde{\mu}_1+\delta)^2}{2\sigma_2^2}\right]\\ &= \exp\left[C'(\overline{X}-\overline{Y}-\delta)^2\right],\end{aligned}$$

where C' is negative and does not depend on the samples, because

$$\overline{X}-\tilde{\mu}_1=\frac{\frac{n_2}{\sigma_2^2}(\overline{X}-\overline{Y}-\delta)}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}} \text{ and } \overline{Y}-\tilde{\mu}_1+\delta=\frac{\frac{n_1}{\sigma_1^2}(\overline{Y}+\delta-\overline{X})}{\frac{n_1}{\sigma_1^2}+\frac{n_2}{\sigma_2^2}}.$$

Therefore the rejection region should be $\left\{|\overline{X}-\overline{Y}-\delta|\geq K\right\}$.

Under H_0, we have

$$\begin{cases}\overline{X} \text{ follows N}\left(\mu_1,\frac{\sigma_1^2}{n_1}\right),\\ \overline{Y} \text{ follows N}\left(\mu_1-\delta,\frac{\sigma_2^2}{n_2}\right),\end{cases}$$

and thus (by the independence between the two samples)

$$\overline{X}-\overline{Y} \text{ follows N}\left(\delta,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right).$$

Therefore, the rejection region is

$$\left\{\frac{|\overline{X}-\overline{Y}-\delta|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\geq z_{\alpha/2}\right\},$$

where the test statistic is $\dfrac{\overline{X}-\overline{Y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$. □

From the aforemention example and the similar technique, we can have the following table:

Table 5.4 Testing for the mean when variances σ_1^2 and σ_2^2 are known

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	$\left\{\frac{\lvert\overline{X}-\overline{Y}-\delta\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\geq z_{\alpha/2}\right\}$	$\mathrm{P}\left(\lvert Z\rvert\geq\frac{\lvert\bar{x}-\bar{y}-\delta\rvert}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$
Left-tailed	$\mu_1-\mu_2=\delta$ or $\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$\left\{\frac{\overline{X}-\overline{Y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\leq -z_{\alpha}\right\}$	$\mathrm{P}\left(Z\leq\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$
Right-tailed	$\mu_1-\mu_2=\delta$ or $\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$\left\{\frac{\overline{X}-\overline{Y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\geq z_{\alpha}\right\}$	$\mathrm{P}\left(Z\geq\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right)$

5.4.2 Testing for the mean: Variance is unknown

We second consider the hypothesis testing for $\mu_1 - \mu_2$ when σ_1 and σ_2 are unknown but equal.

Example 5.17. Assume that σ_1 and σ_2 are unknown but equal to σ. Find the generalized likelihood ratio for testing

$$H_0 : \mu_1 - \mu_2 = \delta \qquad \text{versus} \qquad H_1 : \mu_1 - \mu_2 \neq \delta.$$

Solution. Note that

$$\begin{aligned}
\Omega_0 &= \{(\mu_1, \mu_2, \sigma) : \mu_1 - \mu_2 = \delta, \sigma > 0\},\\
\Omega_1 &= \{(\mu_1, \mu_2, \sigma) : \mu_1 - \mu_2 \neq \delta, \sigma > 0\},\\
\Omega &= \Omega_0 \cup \Omega_1 = \{(\mu_1, \mu_2, \sigma) : \sigma > 0\}.
\end{aligned}$$

The likelihood function of the two samples is

$$L(\mu_1, \mu_2, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n_1+n_2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}(X_i - \mu_1)^2 + \sum_{j=1}^{n_2}(Y_j - \mu_2)^2\right]\right\}.$$

On Ω_0, we have

$$\begin{aligned}
\ln L(\mu_1, \mu_2, \sigma) &= \ln L(\mu_1, \mu_1 - \delta, \sigma)\\
&= C - (n_1 + n_2)\ln\sigma - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}(X_i - \mu_1)^2 + \sum_{j=1}^{n_2}(Y_j - \mu_1 + \delta)^2\right],
\end{aligned}$$

where C is a constant. Then,

$$\begin{aligned}
\frac{\partial}{\partial\mu_1}\ln L(\mu_1, \mu_1 - \delta, \sigma) &= \frac{n_1(\overline{X} - \mu_1) + n_2(\overline{Y} - \mu_1 + \delta)}{\sigma^2}\\
&= \frac{n_1\overline{X} + n_2(\overline{Y} + \delta)}{\sigma^2} - \frac{n_1 + n_2}{\sigma^2}\mu_1.
\end{aligned}$$

This implies that the maximum likelihood estimator of μ_1 is

$$\tilde{\mu}_1 = \frac{n_1\overline{X} + n_2(\overline{Y} + \delta)}{n_1 + n_2},$$

which does not depend on σ, since

$$\begin{aligned}
\mu_1 < \tilde{\mu}_1 &\Longleftrightarrow \frac{\partial}{\partial\mu_1}\ln L(\mu_1, \mu_1 - \delta, \sigma) > 0,\\
\mu_1 > \tilde{\mu}_1 &\Longleftrightarrow \frac{\partial}{\partial\mu_1}\ln L(\mu_1, \mu_1 - \delta, \sigma) < 0.
\end{aligned}$$

Therefore, it is now sufficient to consider $L(\tilde{\mu}_1, \tilde{\mu}_1 - \delta, \sigma)$ for finding the maximum likelihood estimator of σ. By direct calculation,

$$\begin{aligned}\frac{\partial}{\partial\sigma}\ln L(\tilde{\mu}_1,\tilde{\mu}_1-\delta,\sigma) &= -\frac{n_1+n_2}{\sigma}+\frac{1}{\sigma^3}\left[\sum_{i=1}^{n_1}(X_i-\tilde{\mu}_1)^2+\sum_{j=1}^{n_2}(Y_j-\tilde{\mu}_1+\delta)^2\right]\\ &= \frac{1}{\sigma^3}\left[-(n_1+n_2)\sigma^2+\sum_{i=1}^{n_1}(X_i-\tilde{\mu}_1)^2+\sum_{j=1}^{n_2}(Y_j-\tilde{\mu}_1+\delta)^2\right]\end{aligned}$$

and the maximum likelihood estimator of σ is

$$\tilde{\sigma}=\sqrt{\frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(X_i-\tilde{\mu}_1)^2+\sum_{j=1}^{n_2}(Y_j-\tilde{\mu}_1+\delta)^2\right]},$$

since

$$\sigma>\tilde{\sigma}\iff\frac{\partial}{\partial\sigma}\ln L(\tilde{\mu}_1,\tilde{\mu}_1-\delta,\sigma)<0.$$

Therefore,

$$\ln L(\Omega_0)=C-(n_1+n_2)\ln\tilde{\sigma}-\frac{n_1+n_2}{2}.$$

On Ω, we have

$$\ln L(\mu_1,\mu_2,\sigma)=C-(n_1+n_2)\ln\sigma-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}(X_i-\mu_1)^2+\sum_{j=1}^{n_2}(Y_j-\mu_2)^2\right],$$

where C is a constant. Then, by direct calculation,

$$\begin{aligned}\frac{\partial}{\partial\mu_1}\ln L(\mu_1,\mu_2,\sigma) &= \frac{n_1(\overline{X}-\mu_1)}{\sigma^2},\\ \frac{\partial}{\partial\mu_2}\ln L(\mu_1,\mu_2,\sigma) &= \frac{n_2(\overline{Y}-\mu_2)}{\sigma^2},\\ \frac{\partial}{\partial\sigma}\ln L(\mu_1,\mu_2,\sigma) &= -\frac{n_1+n_2}{\sigma}+\frac{1}{\sigma^3}\left[\sum_{i=1}^{n_1}(X_i-\mu_1)^2+\sum_{j=1}^{n_2}(Y_j-\mu_2)^2\right].\end{aligned}$$

Hence, by following the same routine as before, we can show that the maximum likelihood estimators are

$$\begin{aligned}\hat{\mu}_1 &= \overline{X},\\ \hat{\mu}_2 &= \overline{Y},\\ \hat{\sigma}^2 &= \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2\right].\end{aligned}$$

Therefore,

$$\ln L(\Omega)=C-(n_1+n_2)\ln\hat{\sigma}-\frac{n_1+n_2}{2}.$$

Now, the generalized likelihood ratio is

$$\Lambda=\frac{L(\Omega_0)}{L(\Omega)}=\frac{\tilde{\sigma}^{-(n_1+n_2)}}{\hat{\sigma}^{-(n_1+n_2)}}=\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{-(n_1+n_2)/2}.$$

Note that

$$\begin{aligned}
\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} &= \frac{\sum_{i=1}^{n_1}(X_i-\tilde{\mu}_1)^2+\sum_{j=1}^{n_2}(Y_j-\tilde{\mu}_1+\delta)^2}{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}\\
&= \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+n_1(\overline{X}-\tilde{\mu}_1)^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2+n_2(\overline{Y}-\tilde{\mu}_1+\delta)^2}{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}\\
&= 1+\frac{n_1(\overline{X}-\tilde{\mu}_1)^2+n_2(\overline{Y}-\tilde{\mu}_1+\delta)^2}{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}\\
&= 1+\frac{n_1\left[\frac{n_2(\overline{X}-\overline{Y}-\delta)}{n_1+n_2}\right]^2+n_2\left[\frac{n_1(\overline{Y}+\delta-\overline{X})}{n_1+n_2}\right]^2}{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}\\
&= 1+\frac{\frac{n_1n_2}{n_1+n_2}(\overline{X}-\overline{Y}-\delta)^2}{n_1S_1^2+n_2S_2^2}\\
&= 1+\frac{(\overline{X}-\overline{Y}-\delta)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left[n_1S_1^2+n_2S_2^2\right]},
\end{aligned}$$

where S_1^2 and S_2^2 are the sample variances of $\{X_i,\, i=1,2,\dots,n_1\}$ and $\{Y_j,\, j=1,2,\dots,n_2\}$ respectively, Therefore H_0 should be rejected when

$$\frac{|\overline{X}-\overline{Y}-\delta|}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{n_1S_1^2+n_2S_2^2}} \text{ is large.}$$

Under H_0, $\overline{X}$ follows $\mathrm{N}\left(\mu_1,\frac{\sigma^2}{n_1}\right)$ and $\overline{Y}$ follows $\mathrm{N}\left(\mu_1-\delta,\frac{\sigma^2}{n_2}\right)$, and thus $\overline{X}-\overline{Y}$ follows $\mathrm{N}\left(\delta,\frac{\sigma^2}{n_1}+\frac{\sigma^2}{n_2}\right)$, which implies that

$$\frac{\overline{X}-\overline{Y}-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ follows } \mathrm{N}(0,1).$$

Besides, the fact that the two independent random variables $\frac{n_1S_1^2}{\sigma^2}$ and $\frac{n_2S_2^2}{\sigma^2}$ follows $\chi^2_{n_1-1}$ and $\chi^2_{n_2-1}$ respectively implies that $\frac{n_1S_1^2+n_2S_2^2}{\sigma^2}$ follows $\chi^2_{n_1+n_2-2}$. Therefore,

$$W = \frac{\dfrac{\overline{X}-\overline{Y}-\delta}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{\sigma^2}\Big/(n_1+n_2-2)}} = \frac{\overline{X}-\overline{Y}-\delta}{\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}\sqrt{\dfrac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2}}}$$

follows $t_{n_1+n_2-2}$ by Property 4.2(ii). Letting

$$S_p^2 = \frac{n_1S_1^2+n_2S_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2+\sum_{j=1}^{n_2}(Y_j-\overline{Y})^2}{n_1+n_2-2},$$

then the rejection region is $\left\{|W| \geq t_{\alpha/2,n_1+n_2-2}\right\}$, where the test statistic is $W = \dfrac{\overline{X}-\overline{Y}-\delta}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$.

□

From the aforemention example and the similar technique, we can have the following table:

Table 5.5 Testing for the mean when variances $\sigma_1^2 = \sigma_2^2$ are unknown

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\mu_1-\mu_2=\delta$	$\mu_1-\mu_2\neq\delta$	$\left\{\left\lvert\dfrac{\overline{X}-\overline{Y}-\delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right\rvert \geq t_{\alpha/2,n_1+n_2-2}\right\}$	$\mathrm{P}\left(\lvert t_{n_1+n_2-2}\rvert \geq \left\lvert\dfrac{\overline{x}-\overline{y}-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right\rvert\right)$
Left-tailed	$\mu_1-\mu_2=\delta$ or $\mu_1-\mu_2\geq\delta$	$\mu_1-\mu_2<\delta$	$\left\{\dfrac{\overline{X}-\overline{Y}-\delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leq -t_{\alpha,n_1+n_2-2}\right\}$	$\mathrm{P}\left(t_{n_1+n_2-2} \leq \dfrac{\overline{x}-\overline{y}-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$
Right-tailed	$\mu_1-\mu_2=\delta$ or $\mu_1-\mu_2\leq\delta$	$\mu_1-\mu_2>\delta$	$\left\{\dfrac{\overline{X}-\overline{Y}-\delta}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geq t_{\alpha,n_1+n_2-2}\right\}$	$\mathrm{P}\left(t_{n_1+n_2-2} \geq \dfrac{\overline{x}-\overline{y}-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$

Remark 5.3. S_p is called the pooled sample variance, which is an unbiased estimator of σ^2 under H_0.

Example 5.18. A consumer agency wanted to estimate the difference in the mean amounts of caffeine in two brands of coffee. The agency took a sample of 15 500-gramme jars of Brand I coffee that showed the mean amount of caffeine in these jars to be 80 mg per jar and the standard deviation to be 5 mg. Another sample of 12 500-gramme jars of Brand II coffee gave a mean amount of caffeine equal to 77 mg per jar and a standard deviation of 6 mg. Assuming that the two populations are normally distributed with equal variances, check at the 5% significance level whether the mean amount of caffeine in 500-gramme jars is greater for Brand 1 than for Brand 2.

Solution. Let the amounts of caffeine in jars of Brand I be referred to as population 1 and those of Brand II be referred to as population 2.

We consider the hypotheses:

$$H_0 : \mu_1 \leq \mu_2 \text{ versus } H_1 : \mu_1 > \mu_2,$$

where μ_1 and μ_2 are the mean of population 1 and population 2, respectively.

Note that

$$n_1 = 15,\ \bar{x}_1 = 80,\ s_1 = 5,$$

and

$$n_2 = 12,\ \bar{x}_2 = 77,\ s_2 = 6,\ \alpha = 0.05.$$

Hence, $\bar{x}_1 - \bar{x}_2 = 80 - 77 = 3$, $t_{\alpha,n_1+n_2-2} = t_{0.05,25} \approx 1.708$, $\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} \approx 0.3873$, and

$$s_p = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{15 * 5^2 + 12 * 6^2}{15 + 12 - 2}} \approx 5.4626.$$

Therefore, the observed value of the test statistic is

$$w = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{3}{5.4626 * 0.3873} \approx 1.42.$$

As $1.42 < 1.708$, we can not reject H_0. Thus, we conclude that the mean amount of caffeine in 500-gramme jars is not greater for Brand 1 than for Brand 2 at the 5% significance level. □

5.4.3 Testing for the variance

In the above subsection, the assumption $\sigma_1 = \sigma_2$ is needed. Hence, it is interesting to perform tests comparing σ_1 and σ_2.

Example 5.19. Find the generalized likelihood ratio test for hypotheses

$$H_0 : \sigma_1 = \sigma_2 \ \text{ versus } \ H_1 : \sigma_1 \neq \sigma_2.$$

Solution. It can be proved (details omitted) that the generalized likelihood ratio is

$$\frac{C\left(\dfrac{S_1^2}{S_2^2}\right)^{n_1/2}}{\left[n_1 \dfrac{S_1^2}{S_2^2} + n_2\right]^{(n_1+n_2)/2}},$$

where C is a constant.

For $w > 0$ define the function

$$G(w) = \frac{w^{n_1/2}}{[n_1 w + n_2]^{(n_1+n_2)/2}}.$$

Then,

$$\begin{aligned} \ln G(w) &= \frac{n_1}{2}\ln w - \frac{n_1+n_2}{2}\ln[n_1 w + n_2], \\ \frac{\mathrm{d}}{\mathrm{d}w}\ln G(w) &= \frac{n_1}{2w} - \frac{n_1+n_2}{2} \times \frac{n_1}{n_1 w + n_2} \\ &= \frac{n_1 n_2 (1-w)}{2w[n_1 w + n_2]}, \end{aligned}$$

which is negative when $w > 1$ and is positive when $w < 1$. Therefore, the value of $G(w)$ will be small when w is very large or very small. Therefore H_0 should be rejected when $\frac{S_1^2}{S_2^2}$ is large or small.

When H_0 is true, $\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} = \frac{\frac{n_1 S_1^2}{(n_1-1)\sigma_1^2}}{\frac{n_2 S_2^2}{(n_2-1)\sigma_2^2}}$ follows F_{n_1-1,n_2-1} by Property 4.3. Thus, we let the test statistic be $W = \frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2}$, and the rejection region is

$$\{W \le F_{1-\alpha/2,n_1-1,n_2-1}\} \cup \{W \ge F_{\alpha/2,n_1-1,n_2-1}\}.$$

□

From the aforemention example and the similar technique, we can have the following table:

Table 5.6 Testing for the variance

Test	H_0	H_1	Rejection region	p-value
Two-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$\left\{\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \ge F_{\alpha/2,n_1-1,n_2-1}\right\}$ $\cup\left\{\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \le F_{1-\alpha/2,n_1-1,n_2-1}\right\}$	$2\min\left(\mathrm{P}\left(F_{n_1-1,n_2-1} \ge \frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2}\right),\right.$ $\left.\mathrm{P}\left(F_{n_1-1,n_2-1} \le \frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2}\right)\right)$
Left-tailed	$\sigma_1 = \sigma_2$ or $\sigma_1 \ge \sigma_2$	$\sigma_1 < \sigma_2$	$\left\{\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \le F_{1-\alpha,n_1-1,n_2-1}\right\}$	$\mathrm{P}\left(F_{n_1-1,n_2-1} \le \frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2}\right)$
Right-tailed	$\sigma_1 = \sigma_2$ or $\sigma_1 \le \sigma_2$	$\sigma_1 > \sigma_2$	$\left\{\frac{n_1(n_2-1)S_1^2}{n_2(n_1-1)S_2^2} \ge F_{\alpha,n_1-1,n_2-1}\right\}$	$\mathrm{P}\left(F_{n_1-1,n_2-1} \ge \frac{n_1(n_2-1)s_1^2}{n_2(n_1-1)s_2^2}\right)$

Remark 5.4. Recall $F_{\alpha,m,n}$ as the positive real number such that $\mathrm{P}(X \ge F_{\alpha,m,n}) = \alpha$ where X follows $F_{m,n}$. Suppose X follows $F_{m,n}$. Then $1/X$ follows $F_{n,m}$ and

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

because

$$\begin{aligned}
& 1-\alpha = \mathrm{P}(F_{1-\alpha,m,n} < X) = \mathrm{P}\left(\frac{1}{F_{1-\alpha,m,n}} > \frac{1}{X}\right) \\
\Longrightarrow \quad & \alpha = \mathrm{P}\left(\frac{1}{F_{1-\alpha,m,n}} \leq \frac{1}{X}\right) \\
\Longrightarrow \quad & \frac{1}{F_{1-\alpha,m,n}} = F_{\alpha,n,m}.
\end{aligned}$$

Example 5.20. A study involves the number of absences per year among union and non-union workers. A sample of 16 union workers has a sample standard deviation of 3.0 days. A sample of 10 non-union workers has a sample standard deviation of 2.5 days. At the 10% significance level, can we conclude that the variance of the number of days absent for union workers is different from that for nonunion workers?

Solution. Let all union workers be referred to as population 1 and all non-union workers be referred to as population 2.

We consider the hypotheses:

$$H_0 : \sigma_1 = \sigma_2 \text{ versus } H_1 : \sigma_1 \neq \sigma_2,$$

where σ_1^2 and σ_2^2 are the variance of population 1 and population 2, respectively.

Note that $n_1 = 16$, $s_1 = 3$, $n_2 = 10$, and $s_2 = 2.5$. Hence, the value of the test statistic is

$$\frac{n_1(n_2-1)}{n_2(n_1-1)}\frac{s_1^2}{s_2^2} = 0.96 * \frac{3.0^2}{2.5^2} = 1.3824.$$

Since

$$\frac{1}{f_{0.05,9,15}} < 1 < 1.3824 < 3.006 \approx f_{0.05,15,9},$$

we cannot reject H_0. Thus we conclude that the data do not indicate that the variance of the number of days absent for union workers is different from that for non-union workers at the 10% significance level. □

Remark 5.5. Since H_0 should not be rejected, we may test

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

using the pooled sample variance.

5.5 Generalized likelihood ratio tests: Large samples

Unfortunately, the likelihood ratio method does not always produce a test statistic with a known probability distribution. Nevertheless, if the sample size is large, we can obtain an approximation to the distribution of a generalized likelihood ratio.

Theorem 5.2. *Suppose that we are testing*

$$H_0 : \theta_i = \theta_{i,0} \textit{ for all } i = 1,2,\ldots,d$$

versus

$$H_1 : \theta_i \neq \theta_{i,0} \textit{ for at least one } i = 1,2,\ldots,d$$

and that Λ is the generalized likelihood ratio. Then, under very general conditions, when H_0 is true,

$$-2\ln\Lambda \rightarrow_d \chi_d^2 \textit{ as } n \rightarrow \infty.$$

Remark 5.6. For more general cases, d can be determined by

$$\begin{aligned} d = &\ [\text{number of parameters to estimate when determining } L(\Omega)] \\ &- [\text{number of parameters to estimate when determining } L(\Omega_0)]\,. \end{aligned}$$

For example, if we test

$$H_0 : \theta_i = \theta_{i,0} \text{ for all } i = 1,2,\ldots,m$$

against

$$H_1 : \theta_i \neq \theta_{i,0} \text{ for at least one } i = 1,2,\ldots,m,$$

then $d = m$.

5.5.1 Goodness-of-fit tests

Suppose m is an integer greater than 1, and there is a population X such that

$$\mathrm{P}(X = a_i) = p_i, \qquad i = 1,2,\ldots,m,$$

where $i \neq j$ implies $a_i \neq a_j$ for $i, j = 1,2,\ldots,m$, $p_i > 0$ for $i = 1,2,\ldots,m$ and

$$p_1 + p_2 + \cdots + p_m = 1.$$

Now suppose $X_1, X_2, \ldots, X_n$ constitute a random sample of size n from the population. For $i = 1,2,\ldots,m$, let Y_i be the number of k such that $X_k = a_i$. Then

$$\mathrm{P}(Y_i = y_i, i = 1,2,\ldots,m) = \frac{n!}{y_1!y_2!\cdots y_m!} p_1^{y_1} p_2^{y_2} \cdots p_m^{y_m}$$

for non-negative integers y_i, $i = 1,2,\ldots,m$, such that $y_1 + y_2 + \cdots + y_m = n$.

We want to test

$$H_0 : \ p_i = p_{i,0} \text{ for all } i = 1,2,\ldots,m$$

versus

$$H_1 : \ p_i \neq p_{i,0} \text{ for at least one } i = 1,2,\ldots,m,$$

where $p_{i,0} > 0$ for $i = 1,2,\ldots,m$ and

$$p_{1,0} + p_{2,0} + \cdots + p_{m,0} = 1.$$

This is equivalent to testing

$$H_0 : \ p_i = p_{i,0} \text{ for all } i = 1,2,\ldots,m-1$$

versus

$$H_1: \ p_i \neq p_{i,0} \text{ for at least one } i = 1,2,\ldots,m-1,$$

where $p_{i,0} > 0$ for $i = 1,2,\ldots,m-1$ and

$$p_{1,0} + p_{2,0} + \cdots + p_{m-1,0} < 1.$$

If we let (O for observed frequency and E for expected frequency when H_0 is true)

$$\begin{aligned} O_i &= Y_i \qquad \text{for } i = 1,2,\ldots,m-1, \\ O_m &= n - \sum_{i=1}^{m-1} Y_i, \\ E_i &= np_{i,0} \qquad \text{for } i = 1,2,\ldots,m-1, \\ E_m &= n\left(1 - \sum_{i=1}^{m-1} p_{i,0}\right), \end{aligned}$$

then, it can be proved that

$$\begin{aligned} -2\ln\Lambda &\approx \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m} \frac{O_i^2}{E_i} - 2\sum_{i=1}^{m} O_i + \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{O_i^2}{E_i} - 2n + n \\ &= \sum_{i=1}^{m} \frac{O_i^2}{E_i} - n. \end{aligned}$$

Note that $\{-2\ln\Lambda \geq K\}$ is the rejection region, and

$$\left\{ \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \geq \chi^2_{\alpha,m-1} \right\}$$

can serve as an approximate rejection region. Since this is only an approximate result, it is suggested that all expected frequencies should be no less than 5, so that the sample is large enough. To meet this rule, some categories may be combined when to do so is logical.

Example 5.21. A journal reported that, in a bag of m&m's chocolate peanut candies, there are 30% brown, 30% yellow, 10% blue, 10% red, 10% green and 10% orange candies. Suppose you purchase a bag of m&m's chocolate peanut candies at a nearby store and find 17 brown, 20 yellow, 13 blue, 7 red, 6 green and 9 orange candies, for a total of 72 candies. At the 0.1 level of significance, does the bag purchased agree with the distribution suggested by the journal?

Solution. Consider hypotheses:

$$H_0: \ \text{the bag purchased agrees with the distribution suggested by the journal,}$$

versus

$$H_1: \ \text{the bag purchased does not agree with the distribution suggested by the journal.}$$

Then we have the table below, in which all expected frequencies are at least 5.

Colour	O_i	E_i	$O_i - E_i$
Brown	17	$72 \times 30\% = 21.6$	-4.6
Yellow	20	$72 \times 30\% = 21.6$	-1.6
Blue	13	$72 \times 10\% = 7.2$	5.8
Red	7	$72 \times 10\% = 7.2$	-0.2
Green	6	$72 \times 10\% = 7.2$	-1.2
Orange	9	$72 \times 10\% = 7.2$	1.8
Total	72	72	0

Therefore, as the sample is large enough,

$$\begin{aligned} -2\ln\Lambda &\approx \sum_{i=1}^{6} \frac{O_i^2}{E_i} - n \\ &= \frac{17^2+20^2}{21.6} + \frac{13^2+7^2+6^2+9^2}{7.2} - 72 \\ &\approx 6.426 < 9.236 \approx \chi^2_{0.1,6-1}. \end{aligned}$$

Alternatively,

$$\begin{aligned} -2\ln\Lambda &\approx \sum_{i=1}^{6} \frac{(O_i-E_i)^2}{E_i} \\ &= \frac{(-4.6)^2}{21.6} + \frac{(-1.6)^2}{21.6} + \frac{5.8^2}{7.2} + \frac{(-0.2)^2}{7.2} + \frac{(-1.2)^2}{7.2} + \frac{1.8^2}{7.2} \\ &\approx 6.426 < 9.236 \approx \chi^2_{0.1,6-1}. \end{aligned}$$

Hence we should not reject H_0. At the significance level 10%, we cannot conclude that the bag purchased does not agree with the distribution suggested by the journal. □

Example 5.22. A traffic engineer wishes to study whether drivers have a preference for certain tollbooths at a bridge during non-rush hours. The number of automobiles passing through each tollbooth lane was counted during a randomly selected 15-minute interval. The sample information is as follows.

Tollbooth Lane	1	2	3	4	5	Total
Number of Cars observed	171	224	211	180	214	100

Can we conclude that there are differences in the numbers of cars selecting respectively each of the lanes? Test at the 5% significance level.

Solution. Consider hypotheses:

$$H_0: \text{ there is no preference among the five lanes,}$$

versus

$$H_1: \text{ there is a preference among the five lanes.}$$

All the five expected frequencies equal $1000 \div 5 = 200$, which is not less than 5. Therefore, as the sample is large enough,

$$\begin{aligned}
-2\ln\Lambda &\approx \sum_{i=1}^{5}\frac{O_i^2}{E_i}-n \\
&= \frac{171^2+224^2+211^2+180^2+214^2}{200}-1000 \\
&\approx 10.67 \geq 9.488 \approx \chi^2_{0.05,5-1}.
\end{aligned}$$

Hence, H_0 should be rejected. At the significance level 5%, we can conclude that there are differences in the numbers of cars selecting respectively each of the lanes. □

When testing goodness of fit to help select an appropriate population model, we usually are interested in testing whether some family of distributions seems appropriate and are not interested in the lack of fit due to the wrong parameter values. Suppose we want to test

H_0 : the population follows a particular distribution with k unknown parameters.

For calculating E_i's, we have to use the maximum likelihood estimate of the unknown parameters. Then the rejection region is $\{-2\ln\Lambda \geq K\}$ or, approximately,

$$\left\{\sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i} \geq \chi^2_{\alpha,m-1-k}\right\}.$$

5.5.2 *Pearson Chi-squared test of independence*

Consider the following joint distribution of two discrete random variables X and Y:

Probability		Value of Y					Row sum
		b_1	$\cdots$	b_j	$\cdots$	b_c	
Value of X	a_1	$p_{1,1}$	$\cdots$	$p_{1,j}$	$\cdots$	$p_{1,c}$	$p_{1.}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_i	$p_{i,1}$	$\cdots$	$p_{i,j}$	$\cdots$	$p_{i,c}$	$p_{i.}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_r	$p_{r,1}$	$\cdots$	$p_{r,j}$	$\cdots$	$p_{r,c}$	$p_{r.}$
Column sum		$p_{.1}$	$\cdots$	$p_{.j}$	$\cdots$	$p_{.c}$	1

We want to test

$$H_0 : X \text{ and } Y \text{ are independent}$$

versus

$$H_1 : X \text{ and } Y \text{ are not independent.}$$

That is, we want to test

$$H_0 : p_{i,j} = p_{i\cdot}p_{\cdot j} \quad \text{for } i = 1,2,\ldots,r-1 \text{ and } j = 1,2,\ldots,c-1$$

versus

$$\begin{aligned}
H_1 : p_{i,j} \neq p_{i\cdot}p_{\cdot j} \quad &\text{for at least one } i \text{ and } j, \\
&\text{where } i = 1,2,\ldots,r-1 \text{ and } j = 1,2,\ldots,c-1.
\end{aligned}$$

A random sample of size n taken from this distribution is a set of n independent vectors, or ordered pairs of random variables, (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n) each following this distribution. From such a sample we obtain the following table, where $O_{i,j}$ (called the observed frequency of the (i,j)-th cell) is the number of k such that $X_k = a_i$ and $Y_k = b_j$, $i = 1,2,\dots,r$, $j = 1,2,\dots,c$. A box containing an observed frequency is called a cell. Such a two-way classification table is also called a contingency table or cross-tabulation. Ours is an $r \times c$ contingency table.

Observed frequency		Value of Y					Row sum
		b_1	$\cdots$	b_j	$\cdots$	b_c	
Value of X	a_1	$O_{1,1}$	$\cdots$	$O_{1,j}$	$\cdots$	$O_{1,c}$	$n_{1.}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_i	$O_{i,1}$	$\cdots$	$O_{i,j}$	$\cdots$	$O_{i,c}$	$n_{i.}$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	a_r	$O_{r,1}$	$\cdots$	$O_{r,j}$	$\cdots$	$O_{r,c}$	$n_{r.}$
Column sum		$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.c}$	n

Let Λ be the generalized likelihood ratio. Then it can be proved that

$$-2ln\Lambda \approx \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{i,j}^2}{E_{i,j}} - n = n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{i,j}^2}{n_{i\cdot}n_{\cdot j}} - 1\right)$$

where $E_{i,j} = \dfrac{n_{i\cdot}n_{\cdot j}}{n}$ is the expected frequency corresponding to $O_{i,j}$ when H_0 is true, $i = 1,2,\dots,r$ and $j = 1,2,\dots,c$. The rejection region is approximately

$$\left\{\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \geq \chi^2_{\alpha,(r-1)(c-1)}\right\},$$

where

$$\begin{aligned}
&\text{the number of degrees of freedom}\\
=\ &[\text{number of parameters to estimate when determining } L(\Omega)]\\
&\quad -[\text{number of parameters to estimate when determining } L(\Omega_0)]\\
=\ &(rc-1)-[(r-1)+(c-1)]\\
=\ &rc-r-c+1\\
=\ &(r-1)(c-1).
\end{aligned}$$

This test is called the Pearson Chi-squared test of independence. As in previous sections, we require each expected frequency to be at least 5.

Example 5.23. Suppose we draw a sample of 360 students and obtain the following information. At the 0.01 level of significance, test whether a student's ability in mathematics is independent of the student's interest in statistics.

		Ability in Math			sum
		Low	Average	High	
Interest in Statistics	Low	63	42	15	120
	Average	58	61	31	150
	High	14	47	29	90
Sum		135	150	75	360

Solution. Consider hypotheses:

$$H_0: \text{ ability in mathematics and interest in statistics are independent,}$$

versus

$$H_1: \text{ ability in mathematics and interest in statistics are not independent (are related).}$$

The table below shows the expected frequencies (where, for example, $45 = 120 \times 135 \div 360$ and $50 = 120 \times 150 \div 360$).

		Ability in Math			sum
		Low	Average	High	
Interest in Statistics	Low	45	50	25	120
	Average	56.25	62.5	31.25	150
	High	33.75	37.5	18.75	90
Sum		135	150	75	360

All expected frequencies are at least 5. Therefore, as the sample is large enough,

$$n\left(\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{i,j}^2}{n_{i}.n._{j}}-1\right) = 360\left(\frac{63^2}{120\times 135}+\frac{42^2}{120\times 150}+\cdots+\frac{29^2}{90\times 75}-1\right)$$
$$\approx 32.140 \geq 13.277 \approx \chi^2_{0.01,(3-1)(3-1)}.$$

Hence, at the significance level 1%, we reject H_0 and conclude that there is a relationship between a student's ability in mathematics and the student's interest in statistics.

Alternatively, the value of the test statistic equals

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} = \frac{(63-45)^2}{45}+\frac{(42-50)^2}{50}+\cdots+\frac{(29-18.75)^2}{18.75}$$
$$\approx 32.140 \geq 13.277 \approx \chi^2_{0.01,(3-1)(3-1)}.$$

□