STA3007 Nonparametric Statistics

Lecture Notes

Xian Zhou

The Chinese University of Hong Kong, Shenzhen

Overview:

This course introduces methods of statistical inference without assumptions on the mathematical forms of underlying probability distributions. Topics include:

- Nonparametric statistical inference
- One-sample location and matched pair problems
- Two-sample location problem
- Two-sample dispersion and other problems
- One-way layout problems
- Two-way layout problems
- Independence problem
- Regression problems

1. Basics of Statistics

Sample space and events

- An experiment that produces *uncertain* outcomes is said to be a *statistical experiment*. The outcome is not known before it is *observed*.
- Each possible outcome of a statistical experiment is referred to as a *sample point*, often denoted by ω.
- The set of all possible outcomes is called the *sample space*, denoted by Ω.
- Before the experiment is carried out, it is only known that its outcome is in the sample space, expressed by $\omega \in \Omega$, but not which one in Ω.
- For example, tossing a coin is a statistical experiment with two possible outcomes of the upward face: head or tail. Hence $\Omega = \{\text{head}, \text{tail}\}$.
- Another simple example is to count the number of cars passing a bridge in a certain time period. Then every nonnegative integer is a sample point, and the sample space is $\Omega = \{0,1,2,\ldots\}$.

- To find out how much time it will take to complete a certain task is also a statistical experiment, where every nonnegative number $\omega \geq 0$ is a sample point, and the sample space is $\Omega = [0, \infty)$.
- A sample space Ω is said to be *discrete* if it contains a finite or countable number of points. For example, $\Omega = \{\text{head}, \text{tail}\}$ and $\Omega = \{0, 1, 2, \ldots\}$ are discrete sample spaces.
- A sample space Ω is said to be *continuous* if it is an interval of positive length. $\Omega = [0, \infty)$ is an example of a continuous sample space.
- While a set may be neither discrete nor continuous, it is generally sufficient to consider discrete and continuous sample spaces in this course.
- If Ω is a discrete sample space, then every subset E of Ω, written $E \subset \Omega$, is called an *event*.
- For a continuous sample space Ω, $E \subset \Omega$ is an event if and only if E is generated by subintervals of Ω through finite or countable set operations (union, intersection and complement).

- If the outcome of a statistical experiment is in event E, then we say that E *occurs*. In particular, the empty set ϕ is an event that never occurs.
- For example, in counting the number cars passing a bridge between 5–6am, $E = \{0,1,\ldots,100\} \subset \Omega = \{0,1,2,\ldots\}$ is an event, which occurs if and only if no more than 100 cars passed the bridge between 5–6am.
- In $\Omega = [0,\infty)$ for the time (in hours) to complete a task, $[8,\infty)$ is an event that occurs if and only if the require time is at least 8 hours. Other examples of events include $[0,1]$, $(2,4) \cup (8,10]$, $[0,1] \cup [2,3] \cup [4,5] \cup \cdots$, etc.
- If $E \subset \Omega$ is an event, so is its complement $E^c \subset \Omega$, which occurs if and only if E does not occur. In mathematical notation, $\omega \in E^c \Leftrightarrow \omega \notin E$ $(\omega \in \Omega)$.
- If $E_1, E_2, \ldots$ are events, so is $E_1 \cup E_2 \cup \cdots$, which occurs if and only if at least one of $E_1, E_2, \ldots$ occurs.
- As a result, if a subset E of the sample space Ω is generated from finite or countable set operations of events, then E is an event.

Random variable

- A *random variable* is a real-value function $X = X(\omega)$ defined on a sample space Ω such that $\{X \le x\} = \{\omega \in \Omega : X(\omega) \le x\} \subset \Omega$ is an event for every $x \in \mathbb{R}$, where $\mathbb{R}$ denotes the set of all real numbers.
- The value of $X = X(\omega)$ depends on the uncertain outcome ω of a statistical experiment, which explains the terminology of "random variable".
- If $\Omega \subset \mathbb{R} = (-\infty, \infty)$, $X = X(\omega) = \omega$ defines a random variable that is equal to the outcome of a statistical experiment. In general, however, random variables can be quite different from the outcomes (sample points).
- For instance, on $\Omega = \{\text{head}, \text{tail}\}$, if we define $X(\text{head}) = 1$ and $X(\text{tail}) = 0$, then X is a random variable on Ω. In this case, $\omega \in \{\text{head}, \text{tail}\}$ is qualitative (non-numerical), but the value of X is always quantitative (a real number).
- Even if $\Omega = \mathbb{R}$, there can be many different random variables defined on Ω, such as $X(\omega) = 3\omega + 1$, $X(\omega) = \omega^2$, and $X(\omega) = \log(\omega + 1)$, etc.
- $X(\omega_1) = X(\omega_2)$ is possible for $\omega_1 \ne \omega_2$, such as $X(\omega) = \omega^2$ on $\Omega = \mathbb{R}$.

- Let E be any event in a sample space Ω. Define the *indicator* I_E of E by

$$I_E = I_E(\omega) = \begin{cases} 1 & \text{if } E \text{ occurs } (\omega \in E) \\ 0 & \text{otherwise } (\omega \notin E) \end{cases}, \quad \omega \in \Omega.$$

  Then I_E is a random variable that takes on two values only: 1 and 0, which is referred to as a *binary* or *dichotomous* random variable.

- For example, let $\Omega = [0, \infty)$ and $E = [0,1] \subset \Omega$. Then $X = I_E = I_{\{0 \le \omega \le 1\}}$ is a binary random variable with $X = X(\omega) = 1$ if $0 \le \omega \le 1$ and $X = 0$ if $\omega > 1$.

- We will consider events defined by a random variable X in the form of $\{X \le x\}$ for $x \in \mathbb{R}$, or generated by such events through finite or countable set operations, such as $\{X \le 2\} \bigcap \{X \le 1\}^c = \{X \le 2\} \bigcap \{X > 1\} = \{1 < X \le 2\}$ and $\bigcup_{n=1}^{\infty} \{X \le 1 - 1/n\} = \{X < 1\}$.

- The notation ω of the outcome will be often omitted for simplicity. In some cases, however, it will be useful to count the number of outcomes that give the same value of a random variable X.

Probability and distribution

- Given a sample space Ω, a *probability measure* $\Pr$, or simply *probability*, is a function of events in Ω. In other words, $\Pr$ assigns a value, denoted by $\Pr(E)$, to each event $E \subset \Omega$.

- A probability $\Pr$ is assumed to satisfy three axioms:

  (i) $\Pr(E) \geq 0$ for every event $E \subset \Omega$;

  (ii) $\Pr(\Omega) = 1$;

  (iii) If events $E_1, E_2, \ldots$ satisfy $E_i \cap E_j = \phi$ for all $i \neq j$, then

  $$\Pr(E_1 \cup E_2 \cup \cdots) = \Pr(E_1) + \Pr(E_2) + \cdots$$

- Given a random variable X defined on a sample space Ω, a *probability distribution*, or simply a *distribution*, of X is a *rule* that determines $\Pr(E)$ for every event E defined by X.

- A distribution of X can be specified by its *cumulative distribution function* (cdf): $F(x) = F_X(x) = \Pr(X \leq x),\ x \in \mathbb{R}$.

- A random variable X and its distribution are said to be *discrete* if the range of X (the set of its values) is finite or countable.
- A random variable X and its distribution are said to be *continuous* if the range of X is an interval or a union of intervals.
- There exist random variables and distributions that are neither discrete nor continuous, but they will not be needed in this course.
- The distribution of a discrete random variable X can also be given by its *probability mass function* (pmf):

$$f(x) = f_X(x) = \Pr(X = x), \quad x \in \mathbb{R},$$

with $f(x) = 0$ if x is not in the range of X.

- The distribution of a continuous random variable X can also be given by its *probability density function* (pdf), or just *density*:

$$f(x) = f_X(x) = F'(x) = \frac{d}{dx} F_X(x) = \frac{d}{dx} \Pr(X \le x), \quad x \in \mathbb{R}.$$

- For convenience, we may call both pmf and pdf as *probability function* (pf).
- The cdf $F(x)$ of a random variable X can be obtained from its pf $f(x)$ by

$$F(x) = \Pr(X \le x) = \sum_{y \le x} \Pr(X = y) = \sum_{y \le x} f(y) \quad \text{if } X \text{ is discrete;}$$

$$F(x) = \Pr(X \le x) = \int_{-\infty}^{x} f(y)dy \quad \text{if } X \text{ is continuous.}$$

- For example, the following two particular distributions will be useful:
  - *Binomial* distribution with parameters $N = 1, 2, \ldots$ and $0 < p < 1$, denoted by $Bin(N, p)$. Its pmf is

$$f(x) = \Pr(X = x) = \frac{N!}{x!(N-x)!} p^x (1-p)^{N-x}, \; x = 0, 1, \ldots, N.$$

  - *Normal* distribution $N(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and pdf:

$$f(x) = \frac{d}{dx} \Pr(X \le x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}.$$

Symmetric distribution

- A random variable X with cdf $F(x) = \Pr(X \le x)$, or its distribution, is said to be *symmetric* about a real number $a \in \mathbb{R}$ if

$$\Pr(X \le a - x) = \Pr(X \ge a + x) \quad \text{for all } x \ge 0 \tag{1.1}$$

- Write $F(x-) = \Pr(X < x)$. Then by (1.1),

$$F(a-x) = \Pr(X \le a - x) = \Pr(X \ge a + x) = 1 - \Pr(X < a + x) = 1 - F(a + x-)$$

Thus (1.1) is equivalent to

$$F(a-x) + F(a+x-) = 1 \quad \text{for all } x \ge 0 \tag{1.2}$$

- Let $f(x)$ be the probability function (pf) of X (the density of a continuous X, or $f(x) = \Pr(X = x)$ of a discrete X).

- Then (1.1) is equivalent to

$$f(a-x) = f(a+x) \quad \text{for all } x \ge 0 \tag{1.3}$$

- Replace $a-x$ by x and so $a+x=2a-(a-x)$ by $2a-x$. Then the symmetry in (1.1) and (1.3) can be equivalently expressed respectively by

$$\Pr(X \le x) = \Pr(X \ge 2a-x) \quad \text{and} \quad f(x) = f(2a-x) \text{ for all } x \ge 0 \qquad (1.4)$$

- The normal distribution $N(\mu, \sigma^2)$ is obviously symmetric about μ because

$$f(\mu - x) = f(\mu + x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \text{ for all } x \ge 0.$$

- The binomial distribution $Bin(N, p=0.5)$ is symmetric about $a = 0.5N$. Since

$$f(N-x) = \frac{N!}{x!(N-x)!} 0.5^x \cdot 0.5^{N-x} = f(x), \; x = 0,1,\ldots,N,$$

and $f(x) = 0$ for $x \notin \{0,1,\ldots,N\}$, it follows that

$$f(a-x) = f(N-(a-x)) = f(N-0.5N+x) = f(0.5N+x) = f(a+x)$$

(Note that $Bin(N, p)$ is not symmetric if $p \ne 0.5$).

Mean, median and variance

- Let X be a random variable with cdf $F(x)$, pf $f(x)$ and range R. The *mean* (or *expectation* or *expected value*) of X is denoted and defined by

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x dF(x) = \begin{cases} \sum_{x \in R} x f(x) & \text{if } X \text{ is discrete;} \\ \int_{-\infty}^{\infty} x f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

- The mean of a function $g(X)$ of X can be calculated by

$$\mathrm{E}[g(X)] = \begin{cases} \sum_{x \in R} g(x) f(x) & \text{if } X \text{ is discrete;} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

- A real number $m = m_X$ is called a *median* of X if

$$\Pr(X \le m) \ge 0.5 \quad \text{and} \quad \Pr(X \ge m) \ge 0.5$$

- In terms of the cdf, a median m satisfies

$$F(m) \geq 0.5 \text{ and } F(m-) \leq 0.5 \text{ (as } F(m-) = \Pr(X < m) = 1 - \Pr(X \geq m))$$

- If $\Pr(X = m) = 0$, then $F(m) = F(m-) = 0.5$. Hence for a continuous X, a median m is a solution to equation $F(m) = 0.5$.
- If X has a symmetric distribution about a, then $\mathrm{E}[X] = a$ is a median of X.
- As examples, $m = \mathrm{E}[X] = \mu$ for $X \sim N(\mu, \sigma^2)$ (where "~" stands for "has a distribution"), and $m = \mathrm{E}[X] = 0.5N$ for $X \sim Bin(N, 0.5)$.
- If $F(x) = 0.5$ for $a \leq x < b$ with $a < b$, then the median is not unique: any $x \in [a, b)$ is a median of X.
- If the distribution of X is symmetric, however, we will take $m = \mathrm{E}[X]$ as the median whether the median is unique or not.
- For example, if $X \sim Bin(3, 0.5)$, then $F(x) = F(1) = 0.5$ for $1 \leq x < 2$. Hence any $x \in [1, 2)$ is a median of X. But we take $m = 1.5 = 0.5N = \mathrm{E}[X]$.

- Mean and median represent the *location* of a distribution.
- The *variance* of a random variable X is denoted and defined by

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2$$

- $\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of X.
- It is well known that
  - $\mathrm{Var}(X) = \sigma^2$ for $X \sim N(\mu, \sigma^2)$; and
  - $\mathrm{Var}(X) = Np(1-p)$ for $X \sim Bin(N, p)$.
- The variance measures the *dispersion* (or *spread*) of a distribution: a larger variance indicates that the random variable X is more likely to take values far away from the mean; whereas a small variance points to a great chance for X to be close to its mean.
- We will tackle the location problems based on the median and the dispersion problems based on the variance.

Independence

- Two events E and F in a common sample space are said to be (statistically) *independent* if $\Pr(E \bigcap F) = \Pr(E)\Pr(F)$; and *dependent* otherwise.

- Let X and Y be two random variables defined on a (common) sample space Ω. The joint cdf of (X,Y) is denoted and defined by

$$F(x,y) = F_{X,Y}(x,y) = \Pr(X \le x, Y \le y) = \Pr(\{X \le x\} \bigcap \{Y \le y\})$$

- X and Y are said to be *independent* if

$$F(x,y) = \Pr(X \le x)\Pr(Y \le y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

$F_X(x) = \Pr(X \le x)$ is called the *marginal* cdf of X, similarly for $F_Y(y)$.

- More generally, *n* random variables $X_1, \ldots, X_n$ are independent if

$$F(x_1, \ldots, x_n) = \Pr(X_1 \le x_1, \ldots, X_n \le x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

- Equivalently, $X_1,\ldots,X_n$ are independent if and only if

$$f(x_1,\ldots,x_n) = f_1(x_1)\cdots f_n(x_n) \text{ for all } x_1,\ldots,x_n \in \mathbb{R},$$

where $f(x_1,\ldots,x_n)$ is the joint pf of $X_1,\ldots,X_n$ given by

$$f(x_1,\ldots,x_n) = \begin{cases} \Pr(X_1 = x_1,\ldots,X_n = x_n) & \text{if } X_1,\ldots,X_n \text{ are discrete,} \\ \dfrac{\partial^n F(x_1,\ldots,x_n)}{\partial x_1 \cdots \partial x_n} & \text{if } X_1,\ldots,X_n \text{ are continuous,} \end{cases}$$

and $f_i(x)$ is the (marginal) pf of X_i, $i = 1,\ldots,n$.

- In practical data analyses, independence is interpreted as "the value of each variable does not affect the probabilities of events from other variables", and is commonly judged or assumed based on the relationships of the variables under consideration.
- For example, the lifetimes of unrelated people can be reasonably assumed to be independent, but not of people from the same family or bloodline.

Equally likely outcomes

- Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$ be a finite sample space with M equally likely outcomes, that is, $\Pr(\omega_i) = \Pr(\{\omega_i\}) = 1/M$, $i = 1, 2, \ldots, M$.

- Then for any event $E \subset \Omega$,

$$\Pr(E) = \frac{\text{Number of points } \omega_i \in E}{M} \tag{1.5}$$

- The distribution for any random variable X defined on Ω is given by

$$\Pr(X = x) = \frac{\text{Number of } \omega_i \in \Omega \text{ such that } X(\omega_i) = x}{M} \tag{1.6}$$

- The probabilities in (1.5) – (1.6) can be calculated by counting the numbers of outcomes in the relevant events using the following results.

- The number of ways (*permutations*) to order N items is

$$N! = N(N-1)\cdots 2 \cdot 1 \quad (\text{with } 0! = 1 \text{ by definition})$$

- There are N^n ways to choose n items from N items *with replacement*.

- The number of ways to choose ordered n items (e.g., $(1,2) \neq (2,1)$ considered as different ways) from N ($n \leq N$) items *without replacement* is

$$N(N-1)\cdots(N-n+1) = \frac{N!}{(N-n)!}, \quad n = 0,1,\ldots,N.$$

- If the choice is unordered (e.g., $(1,2) = (2,1)$ considered as the same choice), then the number of choices (*combinations*) without replacement is

$$\binom{N}{n} = \frac{N(N-1)\cdots(N-n+1)}{n!} = \frac{N!}{n!(N-n)!}, \quad n = 0,1,\ldots,N.$$

- Consider a statistical experiment to choose an ordered $B = (b_1, b_2, \ldots, b_n)$ with replacement from numbers $a_1, a_2, \ldots, a_N$ (not necessarily distinct) such that $\Pr(b_i = a_j) = 1/N$ for $i = 1,\ldots,n$ and $j = 1,\ldots,N$ (each b_i is equally likely to be any a_j). Then each choice of B is an outcome of the experiment.

- The sample space of all such outcomes consists of $M = N^n$ equally likely sample points (some of them may be identical), each with probability N^{-n}.

- For example, if choose ordered (b_1, b_2) from (a_1, a_2, a_3) with replacement, then there are 9 outcomes each with probability $N^{-n} = 3^{-2} = 1/9$:

  (a_1, a_1), (a_1, a_2), (a_1, a_3), (a_2, a_1), (a_2, a_2), (a_2, a_3), (a_3, a_1), (a_3, a_2), (a_3, a_3)

- If $(b_1, \ldots, b_n)$ is unordered, however, then it is no longer guaranteed to have equal probability.

- For example, if choose unordered (b_1, b_2) from (a_1, a_2, a_3) with replacement, then there are 6 outcomes (a_1, a_1), (a_1, a_2), (a_1, a_3), (a_2, a_2), (a_2, a_3), (a_3, a_3) who probabilities are not all equal:

$$\Pr(a_1, a_1) = \Pr(a_2, a_2) = \Pr(a_3, a_3) = 3^{-2} = 1/9$$

$$\Pr(a_1, a_2) = \Pr(a_1, a_3) = \Pr(a_2, a_3) = 2/9$$

- Define a random variable $X = X(B) = b_1 + \cdots + b_n$ for ordered $(b_1, \ldots, b_n)$ taken from numbers $(a_1, a_2, \ldots, a_N)$ with replacement. Then

$$\Pr(X = x) = \frac{\text{Number of } (b_1, \ldots, b_n) \text{ such that } b_1 + \cdots + b_n = x}{N^n} \qquad (1.7)$$

- If choose $(b_1,\ldots,b_n)$ from $(a_1,\ldots,a_N)$ without replacement, and the random variable $X = X(b_1,\ldots,b_n)$ does not vary with the order of $(b_1,\ldots,b_n)$, such as $X = b_1 + \cdots + b_n$, then it suffices to consider unordered $(b_1,\ldots,b_n)$, or a fixed order for all $(b_1,\ldots,b_n)$, such as $b_1 < \cdots < b_n$, for the distribution of X.

- In such a case, the total number M of outcomes may differ between ordered and unordered $(b_1,\ldots,b_n)$, but the distribution $\Pr(X = x)$ is the same whether $(b_1,\ldots,b_n)$ is ordered, unordered or fixed-ordered.

- For a random variable $X = X(b_1,\ldots,b_n)$, we will consider:

  (i) ordered $(b_1,\ldots,b_n)$ in the case with replacement;

  (ii) ordered $(b_1,\ldots,b_n)$ in the case without replacement if X varies with the order of $(b_1,\ldots,b_n)$; and

  (iii) unordered or fixed-ordered $(b_1,\ldots,b_n)$ in the case without replacement if X is invariant with the order of $(b_1,\ldots,b_n)$.

Example 1.1 Consider choosing $n = 2$ items from $\{1,2,3,4\}$ with replacement. Then $M = 4^2 = 16$ and $B = (b_1, b_2)$ with ordered $b_1, b_2 \in \{1,2,3,4\}$.

Let $X = X(b_1, b_2) = b_1 + b_2$. The range of X is $\{2,3,4,5,6,7,8\}$, with

$$X(1,1) = 1+1 = 2, \quad X(1,2) = 1+2 = 3 = X(2,1), \text{ and so on.}$$

By (1.7), the distribution of X is given in the following table:

x	$B = (b_1, b_2)$	$\Pr(X = x)$
2	(1,1)	1/16
3	(1,2), (2,1)	2/16
4	(1,3), (2,2), (3,1)	3/16
5	(1,4), (2,3), (4,1), (3,2)	4/16
6	(2,4), (3,3), (4,2)	3/16
7	(3,4), (4,3)	2/16
8	(4,4)	1/16

Example 1.2 Choose $n = 2$ items from $\{1,2,3,4\}$ without replacement. Then the range of $X = b_1 + b_2$ is $\{3,4,5,6,7\}$.

If $B = (b_1, b_2)$ is ordered, then $M = 4 \cdot 3 = 12$;

If $B = (b_1, b_2)$ is unordered (or with a fixed order $b_1 < b_2$), then $M = \binom{4}{2} = 6$.

The distribution of X in these two cases is given by

	Ordered; $M = 12$		Unordered; $M = 6$	
x	$B = (b_1, b_2)$	$\Pr(X = x)$	$B = (b_1, b_2)$	$\Pr(X = x)$
3	(1,2), (2,1)	$2/12 = 1/6$	(1,2)	$1/6$
4	(1,3), (3,1)	$2/12 = 1/6$	(1,3)	$1/6$
5	(1,4), (2,3), (4,1), (3,2)	$4/12 = 2/6$	(1,4), (2,3)	$2/6$
6	(2,4), (4,2)	$2/12 = 1/6$	(2,4)	$1/6$
7	(3,4), (4,3)	$2/12 = 1/6$	(3,4)	$1/6$

It shows that X has the same distribution whether B is ordered or unordered.

Statistical inference

- The idea of *statistical inference* is to gain meaningful information (referred to as *inference*) on uncertain matters based on the *samples* (data) drawn from the *population* in which the matters of interest arise.
- The topics include *estimation* of unknown quantities, *prediction* of future outcomes, and *hypothesis testing* to make a decision. While these topics are interrelated, this course will focus primarily on hypothesis testing, but some related estimation problems will be discussed as well.
- The problem of hypothesis testing defines a *null hypothesis*, denoted by H_0, for testing against an *alternative hypothesis* H_1. The purpose is to decide whether to accept or reject H_0 in favour of H_1 based on the available data.
- Rejection of a correct H_0 is referred to as the *Type I* error; and acceptance of a wrong H_0 is the *Type II* error. The principle of hypothesis testing is to achieve $\Pr(\text{Type I error}) \leq \alpha$ – which is known as the *level of significance*, commonly set at 5%, but 1% and 10% are often used as well.

- Rejection of H_0 at the 5% level (of significance) can be interpreted as "there is sufficient evidence against H_0", since the chance of error is only 5%.
- Accepting H_0, however, should not be interpreted as "sufficient evidence" in support of H_0, since the probability of Type II error is not under control. Instead, the right interpretation is "insufficient evidence against H_0".
- The decision is made according to intuition and the nature of H_0 and H_1 based on observed sample data to ensure $\Pr(\text{Type I error}) = \alpha$ under H_0.
- If $\Pr(\text{Type I error}) = \alpha$ is not achievable, we will aim to find the decision rule such that $\Pr(\text{Type I error})$ is the largest probability below α, which is referred to as the *achievable* level of significance.
- A function of the sample random variables used to make a decision for the test, or its observed value, is called a *test statistic*.
- The probability under H_0 that the test statistic is more extreme against H_0 than observed is called the *p-value*. H_0 is rejected $\Leftrightarrow$ p-value $\leq \alpha$.

Example 1.3 Consider a simple example with $H_0 : \mu = 0$ against $H_1 : \mu > 0$ at the 5% level of significance. The data $x_1, \ldots, x_n$ are drawn from independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n \sim N(\mu, 1)$.

Let $\bar{X} = (X_1 + \cdots + X_n)/n$ be the sample mean and $\bar{x}$ its observed value. Then an intuitive rule is to reject H_0 if $\bar{x} \geq c$ for some c such that $\Pr(\text{Type I error}) = 0.05$. By the properties of the normal distribution, $Z = \sqrt{n}\bar{X} \sim N(0,1)$ under H_0, hence

$$\Pr(\text{Type I error}) = \Pr(\bar{X} \geq c) = \Pr(Z \geq \sqrt{n}c) = 0.05 \;\Leftrightarrow\; \sqrt{n}c = z_{0.05} = 1.645,$$

where z_α satisfies $\Pr(Z \geq z_\alpha) = \alpha$. Thus the rejection rule for testing H_0 against H_1 is $\bar{x} \geq c = 1.645/\sqrt{n}$. This c is called the *critical point* of the test.

Furthermore, the test statistic is $\bar{X}$ in this case. Hence

$$p\text{-value} = \Pr(\bar{X} \geq \bar{x}) = \Pr(Z \geq \sqrt{n}\bar{x}) \leq 0.05 \;\Leftrightarrow\; \bar{x} \geq c = 1.645/\sqrt{n}$$

This confirms that H_0 is rejected if and only if p-value $\leq \alpha = 0.05$.

Parametric and nonparametric methods

- If a probability distribution is assumed of a known mathematical form, but with one or more unknown parameters, such as $Bin(N, p)$ and $N(\mu, \sigma^2)$, then it is said to be a *parametric* distribution.
- If the mathematical form of a probability distribution is unknown, then the distribution is said to be *nonparametric*.
- If the distribution of the sample is parametric, then a statistical method that utilizes its known mathematical form to draw its inference is parametric.
- One example is the moment method, which relies on the parametric form of the distribution to set the equation between the sample and population means.
- Another one is the maximum likelihood method, which requires parametric probabilities or densities to obtain likelihood functions.
- If the distribution of the sample $X_1, \ldots, X_n$ is nonparametric, then a statistical method to draw its inference is nonparametric (or *distribution-free*).

- A common and effective nonparametric method to estimate the cdf $F(x)$ of a sample $X_1,\ldots,X_n$ is the *empirical distribution function* (edf):

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}} = \frac{\text{Number of } X_i \le x}{n}$$

- This is equivalent to assign an equal probability $1/n$ to each observed value x_i of X_i: Let $\Pr_n$ denote the probability determined by the edf $F_n(x)$. Then

$$\Pr_n(X = x_i) = F_n(x_i) - F_n(x_i-) = \frac{1}{n}, \quad i = 1,\ldots,n.$$

- Note that a nonparametric method may still use a parametric distribution to make inference, but that is not the distribution of the sample.

- For example, $nF_n(x) \sim Bin(n, F(x))$ for the edf $F_n(x)$. Hence

$$\mathrm{E}\big[F_n(x)\big] = F(x) \quad \text{and} \quad \mathrm{Var}\big(F_n(x)\big) = \frac{1}{n}F(x)[1-F(x)]$$

- A function $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ of a sample of random variables $X_1, \ldots, X_n$ to estimate a quantity θ is called an *estimator* of θ, and its observed value $\hat{\theta}(x_1, \ldots, x_n)$ is a *point estimate* of θ. They often share the same notation $\hat{\theta}$ for simplicity.
- Parametric and nonparametric methods may produce the same estimator. For example, the sample mean $\overline{X}$ is a parametric estimator of the population mean μ from the maximum likelihood method for $N(\mu, \sigma^2)$.
- On the other hand, $\overline{X}$ is also a nonparametric estimator of $\mu = \mathrm{E}[X]$ from the edf $F_n(x)$ if we estimate μ by the mean of $F_n(x)$:

$$\mathrm{E}_n[X] = \sum_{i=1}^{n} x_i \Pr_n(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

- Both parametric and nonparametric methods have their merits and drawbacks, advantages and disadvantages. Some of them are discussed below.

- If a parametric distribution is known to be correct or closely approximate the true distribution, then the parametric method is preferred, as it can model the data more concisely and extrapolate to enable predictions.
- But in practice, it is often difficult or impossible to know what parametric distribution is appropriate, or even if one exists.
- Nonparametric methods, on the other hand, do not rely on any form of the distribution, hence they are usable without the need to assume or find one, and their results are more "robust" than parametric methods. Chapter 1 of the textbook discusses more advantages of nonparametric methods.
- These methods, however, may be more computation intensive, and are not good for extrapolations beyond the range of available data, hence they are generally unsuitable for predictions.
- The primary focus in this course is on nonparametric methods for hypothesis testing. But related estimation problems on parameters associated with tests will be discussed as well.