

MATH 40005

PROBABILITY & STATISTICS

• sample space: the sample space Ω is defined as the set of all possible outcomes of an experiment. the elements of Ω are typically denoted by ω and called sample points.

• subsets of Ω are collections of elements of Ω and are called events.

• the complement $A^c = \Omega \setminus A$

• let I be the general index set, suppose $A_i \subseteq \Omega$ for all i and $B \subseteq \Omega$.

$$\left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} A_i^c \quad \star \quad \left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} A_i^c$$

$$B \cap \left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} (B \cap A_i), \quad B \cup \left(\bigcap_{i \in I} A_i\right) = \bigcap_{i \in I} (B \cup A_i)$$

proof of $\star$:

• we first prove $\left(\bigcap_{i \in I} A_i\right)^c \subseteq \bigcup_{i \in I} A_i^c$:

$$\text{assume } a \in \left(\bigcap_{i \in I} A_i\right)^c \Rightarrow a \notin \left(\bigcap_{i \in I} A_i\right)$$

$$\Rightarrow \exists i^* \in I, a \notin A_{i^*} \Rightarrow \exists i^* \in I, a \in A_{i^*}^c$$

$$\Rightarrow \text{since } A_{i^*}^c \subseteq \bigcup_{i \in I} A_i^c, a \in \bigcup_{i \in I} A_i^c$$

• we then prove $\left(\bigcap_{i \in I} A_i\right)^c \supseteq \bigcup_{i \in I} A_i^c$:

$$\text{assume } a \in \bigcup_{i \in I} A_i^c \Rightarrow \exists i^* \in I, a \in A_{i^*}^c$$

$$\Rightarrow \exists i^* \in I, a \notin A_{i^*}$$

$$\Rightarrow \text{since } \bigcap_{i \in I} A_i \subseteq A_{i^*}, a \notin \bigcap_{i \in I} A_i$$

$$\Rightarrow a \in \left(\bigcap_{i \in I} A_i\right)^c$$

cardinality: for any set A , define the cardinality of A (known as $\text{card}(A)$ or $|A|$) as the number of elements in A .

• two sets have the same cardinality if there exists a bijection between the two sets.

- (3) a set A is finite if it has a finite number of elements
 (4) a set A is countable infinite if there is a bijection between the elements of A and $\mathbb{N} = \{1, 2, \dots\}$
 (5) a set is countable (finite or countable infinite) or uncountable (uncountable infinite)
 注意: 判断是否 cardinality 相同, 要构造 $f(x)$ 后证明其 bijective

probability

(1) classical interpretation: 当所有事件发生概率率相同
 when $\text{card}(\Omega) < \infty$, or area of Ω is finite, $A \subseteq \Omega$:

$$P_{\text{naive}}(A) = \frac{\text{card}(A)}{\text{card}(\Omega)} \text{ or } \frac{\text{area of } A}{\text{area of } \Omega}$$

注意: 要写出 Ω 与 A

(2) limiting frequency: 通过实验判断

consider n_{total} replications of an experiment and n_A denotes the number of times event A occurs.

$$P(A) = \lim_{n_{\text{total}} \rightarrow \infty} \frac{n_A}{n_{\text{total}}}$$

(3) surjective: 主观判断

power set

a power set of a set A , denoted as $P(A)$, is defined as the set of all possible subsets of A including A and $\emptyset$.

unordered

with replacement let $k, n \in \mathbb{N}$, assume we have an urn with n balls with labels in $\{1, 2, \dots, n\}$, and we want to choose k balls one after the other with replacement. assume the order of the balls doesn't matter, then there are $\binom{n+k-1}{k}$ possibilities

proof:

我们将问题转化成: 我们有 k 个小球, 将它们扔入不同的 label 着 $\{1, 2, \dots, n\}$ 的箱子中, 根据结果抽球

因此, 有 k 个球, $(n-1)$ 个挡板, 答案为 $\binom{n+k-1}{n-1}$

注意: 不要套用公式! 有时隔板法时, 板之间可为空 $\binom{n+k-1}{n-1}$

有时板之间不为空 $\binom{k-1}{n-1}$

(1) a collection of subsets of Ω denoted by $\mathcal{F}$ is called an algebra (or a field) on Ω if:

① $\emptyset \in \mathcal{F}$

② $\forall A \in \mathcal{F}, A^c \in \mathcal{F}$,

③ $\forall A_1, A_2 \in \mathcal{F}, A_1 \cup A_2 \in \mathcal{F}$

(2) a collection of subsets of Ω denoted by $\mathcal{F}$ is called an σ -algebra (or an σ -field) on Ω if

① $\emptyset \in \mathcal{F}$

② $\forall A \in \mathcal{F}, A^c \in \mathcal{F}$

③ $\forall$ countable index set $\mathcal{I}$, we have $A_i \in \mathcal{F}, \bigcup_{i \in \mathcal{I}} A_i \in \mathcal{F}$

注意: • any algebra is closed under finite unions and finite intersections.

• any σ -algebra is closed under countable intersection since for any $A_i \in \mathcal{F}, \bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c$ and each $A_i^c \in \mathcal{F}$

• any algebra contains $\Omega = \emptyset^c$

• for any sample space Ω , then the so-called trivial σ -algebra is defined as $\mathcal{F}_{\text{trivial}} = \{\emptyset, \Omega\}$, and the total or power σ -algebra is defined as $\mathcal{F} = \mathcal{P}(\Omega)$

a mapping $P: \mathcal{F} \rightarrow \mathbb{R}$ is called a probability measure on $(\Omega, \mathcal{F})$ if it satisfies three conditions:

probability

(i) $P(A) \geq 0$ for all events $A \in \mathcal{F}$

(ii) $P(\Omega) = 1$

(iii) for any countable sequence of disjoint events $(A_i)_{i \in \mathbb{I}}$ with $A_i \in \mathcal{F}$, for all $i \in \mathbb{I}$, we have

$$P\left(\bigcup_{i \in \mathbb{I}} A_i\right) = \sum_{i \in \mathbb{I}} P(A_i)$$

用这条性质时
要说明是 disjoint
即 $A_i \cap A_j = \emptyset$

注意:

- we define a probability space as the triplet $(\Omega, \mathcal{F}, P)$.
- 在证明某些结论时, 相加计算时 events 要先转化成 disjoint!

① $P(A^c) = 1 - P(A)$

② if $A \subseteq B$, $P(A) \leq P(B)$

③ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

• $(P(A) = 0) \nRightarrow (A = \emptyset)$

conditional probability (1) $P(A_1 \cap \dots \cap A_n) = P(A_1 | A_2 \cap \dots \cap A_n) P(A_2 | A_3 \cap \dots \cap A_n) \dots P(A_{n-1} | A_n) P(A_n)$

(2) if $P(A) > 0$, $P(B) > 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

(3) let $\{B_i | i \in \mathbb{I}\}$ denote a partition of Ω , with $P(B_i) > 0$ for all $i \in \mathbb{I}$; $P(A) > 0$.

$$P(B_i | A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{k \in \mathbb{I}} P(A|B_k)P(B_k)}$$

(4) if $P(A \cap E) > 0$, $P(B \cap E) > 0$

$$P(A|B \cap E) = \frac{P(B|A \cap E)P(A \cap E)}{P(B \cap E)}$$

(5) let $\{B_i | i \in \mathbb{I}\}$ denote a partition of Ω with $P(B_i \cap E) > 0$ for all $i \in \mathbb{I}$, $P(E) > 0$.

$$P(A|E) = \sum_{i \in \mathbb{I}} \frac{P(A \cap B_i | E)}{P(E)} = \sum_{i \in \mathbb{I}} P(A|B_i \cap E) P(B_i | E)$$

3

if events A and B are independent, then the same is true for independence of each of pairs A^c and B , A and B^c , A^c and B^c

Proof: 仅证明第一组

$$P(A^c)P(B) = (1 - P(A))P(B) = P(B) - P(A)P(B) \\ = (P(B \cap A) + P(B \cap A^c)) - P(A \cap B) = P(A^c \cap B)$$

注意: 证明 independence 两个方法: $P(A)P(B) = P(A \cap B)$ 或 $(P(A|B) = P(A), P(B|A) = P(B))$

a countable or uncountable infinite collection of events is defined to be independent if each finite subcollection is independent

a finite collection of events $A_1, A_2, \dots, A_n$ is defined to be independent if

$$P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \dots P(A_{i_k})$$

for every subcollection $\{i_1, i_2, \dots, i_k\}$ of $\{1, 2, \dots, n\}$, $k = 1, 2, \dots, n$

注意: pairwise independence of (A_i, A_j) is in general not sufficient to conclude the independence of $(A_1, \dots, A_n)$

for events $A, B, C \in \mathcal{F}$ and $P(C) > 0$, then A and B are conditionally independent given C if

$$P(A \cap B | C) = P(A | C)P(B | C)$$

if $P(B \cap C) > 0$, then $P(A | B \cap C) = P(A | C)$

Proof:

$$P(A | B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)} = \frac{P(A \cap B | C)P(C)}{P(B \cap C)} = \frac{P(A | C)P(B | C)P(C)}{P(B \cap C)} = P(A | C)$$

注意: 常先证明 A, B conditionally independent, 再用 $P(A | B \cap C) = P(A | C)$
any countable union can be written as a countable union of disjoint sets.

Proof:

let $A, A_1, \dots \in \mathcal{F}$, define $D_1 = A_1, D_2 = A_2 \setminus A_1, D_3 = A_3 \setminus (A_1 \cup A_2) \dots$

(5) a sequence of sets $(A_i)_{i=1}^{\infty}$ is said to increase to A (expressed as $A_i \uparrow A$), if $A_1 \subseteq A_2 \subseteq \dots$ and $\bigcup_{i=1}^{\infty} A_i = A$. similarly, a sequence of sets $(A_i)_{i=1}^{\infty}$ is said to decrease to A (expressed as $A_i \downarrow A$) if $A_1 \supseteq A_2 \supseteq \dots$ and $\bigcap_{i=1}^{\infty} A_i = A$.

(6) $(A_i \uparrow A) \Leftrightarrow (A_i^c \downarrow A^c)$

(7) if $A_1, A_2, \dots \in \mathcal{F}$ and $A_i \uparrow A$ or $A_i \downarrow A$ then

$$\lim_{i \rightarrow \infty} P(A_i) = P(A)$$

proof: 仅证明 $A_i \uparrow A$ 的情况

$A = \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} D_i$, where $D_i = A_i \setminus (\bigcup_{k=1}^{i-1} A_k)$ is disjoint with others

$$P(A) = \sum_{i=1}^{\infty} P(D_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n P(D_i) = \lim_{n \rightarrow \infty} P(A_n)$$

$A_i \downarrow A$ 的情况可用 $1 - P(A)$ 解决

注意: for increasing or decreasing sets, $\lim_{i \rightarrow \infty} P(A_i) = P(\lim_{i \rightarrow \infty} A_i)$

(8) if $A_1, A_2, \dots$ is a countably infinite set of independent events, then $P(\bigcap_{i=1}^{\infty} A_i) = \prod_{i=1}^{\infty} P(A_i)$

proof:

let $B_n = \bigcap_{i=1}^n A_i$, then $B_n \downarrow B = \bigcap_{i=1}^{\infty} A_i$, so

$$\begin{aligned} P(\bigcap_{i=1}^{\infty} A_i) &= P(B) = \lim_{n \rightarrow \infty} P(B_n) = \lim_{n \rightarrow \infty} P(\bigcap_{i=1}^n A_i) \\ &= \lim_{n \rightarrow \infty} \prod_{i=1}^n P(A_i) = \prod_{i=1}^{\infty} P(A_i) \end{aligned}$$

discrete random variables

(1) consider a function with domain X and co-domain Y , i.e. $f: X \rightarrow Y$

- for any subset $A \subseteq X$, we define the image of A under f as:

$$f(A) = \{y \in Y \mid \exists x \in A, f(x) = y\} = \{f(x) \mid x \in A\}$$

if $A = X$, we call $f(X) = \text{Im} f$ the image of f

- for any subset $B \subseteq Y$, we define the pre-image of B under f as:

$$f^{-1}(B) = \{x \in X \mid f(x) \in B\}$$

$f: \Omega \rightarrow \mathbb{R}$

4

注意: $(x \in f^{-1}(B)) \iff (f(x) \in B)$

(2) for any collection of subsets $B_i \subseteq Y, i \in I$ where I denotes an arbitrary index set, we have that $f^{-1}(\bigcup_{i \in I} B_i) = \bigcup_{i \in I} f^{-1}(B_i)$

proof:

if $x \in f^{-1}(\bigcup_{i \in I} B_i)$, $f(x) \in \bigcup_{i \in I} B_i$

then $\exists j \in I, f(x) \in B_j, x \in f^{-1}(B_j), x \in \bigcup_{i \in I} f^{-1}(B_i)$

(3) a discrete random variable on the probability space (Ω, F, P) is defined as a mapping $X: \Omega \rightarrow \mathbb{R}$ such that 这也是条件!

- the image/range of Ω under X is a countable subset of $\mathbb{R}$.

- $X^{-1}(x) = \{\omega \in \Omega \mid X(\omega) = x\} \in F$ for all $x \in \mathbb{R}$

注意: 第二条的 "for all $x \in \mathbb{R}$ ": 如果 $x \notin \text{Im } X$, $X^{-1}(x) = \emptyset \in F$.

(4) the probability mass function (pmf) of the discrete random variable X is defined as the function $p_X: \mathbb{R} \rightarrow [0, 1]$ given by $P_X(x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$

注意: $P_X(x) = 0$ for $x \notin \text{Im } X$

- $X^{-1}(x_1) \cap X^{-1}(x_2) = \{\omega \in \Omega \mid X(\omega) = x_1 \text{ and } X(\omega) = x_2\} = \emptyset$

证明 pmf: $\sum_{x \in \mathbb{R}} P_X(x) = \sum_{x \in \text{Im } X} P_X(x) = \sum_{x \in \text{Im } X} P(X^{-1}(x))$

① $P_X(x) \geq 0$
 $= P(\bigcup_{x \in \text{Im } X} X^{-1}(x)) = P(\bigcup_{x \in \text{Im } X} \{\omega \in \Omega \mid X(\omega) = x\})$

② $\sum P_X(x) = 1$
 $= P(\Omega) = 1$

③ ω 表示一个试验结果, x 表示一个事件的结果, $P_X(x)$ 表示事件发生概率. e.g. toss 2 coins, use 0 represents heads, 1 represents tails. X = number of heads. then. if $\omega_1 = (1, 1), x_1 = 2, P_X(x_1) = P_X(2) = \frac{1}{4}$.

(5) let I denote a countable index set. suppose that $S = \{s_i; i \in I\}$ is a countable set of distinct real number, and $\{x_i; i \in I\}$

is a collection of numbers satisfying

$$\pi_i \geq 0 \text{ for all } i \in \mathcal{I}, \text{ and } \sum_{i \in \mathcal{I}} \pi_i = 1$$

then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a discrete random variable X on that probability space such that its probability mass function is given by

$$P_X(s_i) = \pi_i \text{ for all } i \in \mathcal{I}$$

$$P_X(s) = 0 \text{ for all } s \notin \mathcal{S}$$

(6) 常见的分布.

- Bernoulli distribution. $X \sim \text{Bern}(p)$.

$$P(X=1) = p, P(X=0) = 1-p$$

- Hypergeometric distribution $X \sim \text{HGeom}(N, k, n)$

$$P(X=x) = \frac{\binom{k}{x} \binom{N-k}{n-x}}{\binom{N}{n}}$$

- Discrete uniform distribution $X \sim \text{DUnif}(C)$

$$P(X=x) = \frac{1}{\text{card}(C)} \text{ for } x \in C$$

- Poisson distribution $X \sim \text{Poi}(\lambda)$

$$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x \in \mathbb{N}$$

- Negative binomial distribution $X \sim \text{NBin}(r, p)$

$$P(X=x) = \binom{x+r-1}{r-1} p^r (1-p)^x \text{ for } x \in \mathbb{N}$$

• CDF

(1) a random variable on the probability space $(\Omega, \mathcal{F}, P)$ is defined as the mapping $X: \Omega \rightarrow \mathbb{R}$ which satisfies

$$X^{-1}((-\infty, x]) = \{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F} \text{ for all } x \in \mathbb{R}$$

(2) Suppose that X is a random variable on $(\Omega, \mathcal{F}, P)$, then the cumulative distribution function of X is defined as the mapping

$$F_X: \mathbb{R} \rightarrow [0, 1] \text{ given by } F_X = P\{\omega \in \Omega \mid X(\omega) \leq x\} = P(X^{-1}((-\infty, x])),$$

which is typically abbreviated to $F_X(x) = P(X \leq x)$

5

注意: F_X is right-continuous. if $x_n \downarrow x$ (which means there exists a sequence $x_1 \geq x_2 \geq \dots$ and convergent to x), then $\lim_{n \rightarrow \infty} F(x_n) = F(x)$

proof:

$$E_n := \{\omega \mid X(\omega) \leq x_n\} \downarrow \bigcap_{n=1}^{\infty} E_n = \{\omega \mid X(\omega) \leq x\} := E$$

by continuity of the probability measure: $\lim_{n \rightarrow \infty} P(E_n) = P(E)$

since $P(E_n) = F_X(x_n)$, $P(E) = F_X(x)$, $\lim_{n \rightarrow \infty} F_X(x_n) = F_X(x)$

"a random variable X is called continuous if its CDF continuous random variable can be written as $F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(u) du$ for $\forall x \in \mathbb{R}$

where the function $f_X: \mathbb{R} \rightarrow \mathbb{R}$ satisfies:

(i) $f_X(u) \geq 0$ for all $u \in \mathbb{R}$

(ii) $\int_{-\infty}^{\infty} f_X(u) du = 1$

we call f_X the probability density function of X

注意: $f_X(x)$ 本身不是 probability. $f_X(x) dx$ 才是

0' for a continuous random variable X with density f_X , $P(X=x)=0$ for all $x \in \mathbb{R}$.

proof:

consider any $x \in \mathbb{R}$ with a sequence $x_n \uparrow x$, and define event

$$E_n = \{\omega: x_n < X(\omega) \leq x\} \downarrow E = \{\omega: X(\omega) = x\}$$

then use the continuity of probability measure

$$\begin{aligned} P(X=x) &= \lim_{n \rightarrow \infty} P(E_n) = \lim_{n \rightarrow \infty} P(\{\omega: x_n < X(\omega) \leq x\}) \\ &= \lim_{n \rightarrow \infty} (F_X(x) - F_X(x_n)) = 0 \end{aligned}$$

(3) 常见分布

• uniform distribution $X \sim U(a, b)$, $a < b$.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases} \quad F_X(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{x-a}{b-a} & \text{if } a < x < b \\ 1 & \text{if } x \geq b \end{cases}$$

- exponential distribution $X \sim \text{Exp}(\lambda), \lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad F_X(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0 \end{cases}$$

- Gamma function

$$\text{for } t > 0 \quad \Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$$

$$\text{properties: } \circ \text{ for } t > 1: \Gamma(t) = (t-1)\Gamma(t-1)$$

$$\circ \text{ for } t \in \mathbb{N}, \Gamma(t) = (t-1)!$$

- Gamma distribution $X \sim \text{Gamma}(\alpha, \beta), \alpha, \beta > 0$

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Chi-squared distribution $X \sim \chi^2(n), n \in \mathbb{N}$

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} \left(\frac{x}{2}\right)^{\frac{n}{2}-1} e^{-\frac{x}{2}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

* note that $\chi^2(n) \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2})$

- F-distribution $X \sim F(d_1, d_2), d_1, d_2 \in \mathbb{N}$

$$f_X(x) = \begin{cases} \frac{\Gamma(\frac{d_1+d_2}{2}) \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1}}{\Gamma(\frac{d_1}{2}) \Gamma(\frac{d_2}{2}) \left(1 + \frac{d_1}{d_2} x\right)^{\frac{d_1+d_2}{2}}} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

* if $X_1 \sim \chi^2(n), X_2 \sim \chi^2(m)$ then $\frac{X_1/n}{X_2/m} \sim F(n, m)$

- Beta function

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \text{for } \alpha, \beta > 0$$

- Beta distribution $X \sim \text{Beta}(\alpha, \beta), \alpha, \beta > 0$

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- normal distribution $X \sim N(\mu, \sigma^2)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- cauchy distribution

6

$$f_X(x) = \frac{1}{\pi(1+x^2)}$$

$$F_X(x) = \frac{1}{\pi} \arctan(x) + \frac{1}{2}$$

* if $X, Y \sim N(0, 1)$, $Z = \frac{X}{Y}$ follows cauchy distribution

• student t-distribution, $\nu > 0$

$$f_X(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} (1 + \frac{x^2}{\nu})^{-\frac{\nu+1}{2}}$$

1. let X be a discrete random variable on $(\Omega, \mathcal{F}, P)$ and let transformation $g: \mathbb{R} \rightarrow \mathbb{R}$ denote a deterministic. then $Y = g(X)$ is a discrete random variable with probability mass function given by

$$P_Y(y) = \sum_{x \in \text{Im } X: g(x)=y} P(X=x)$$

proof:

我们要证明 Y 是 discrete random variable 的话要证明 3 个条件。
这里只证明其中一个。

$$\begin{aligned} Y^{-1}(\{y\}) &= \{\omega \in \Omega \mid Y(\omega) = y\} = \{\omega \in \Omega \mid g(X(\omega)) = y\} \\ &= \{\omega \in \Omega \mid X(\omega) \in \{x \in \text{Im } X \mid g(x) = y\}\} \\ &= \{x \in \mathbb{Q} \mid X(\omega) \in \bigcup_{x \in \text{Im } X, g(x)=y} \{x\}\} \\ &= X^{-1}\left(\bigcup_{x \in \text{Im } X, g(x)=y} \{x\}\right) = \bigcup_{x \in \text{Im } X, g(x)=y} X^{-1}(\{x\}) \in \mathcal{F} \end{aligned}$$

for continuous random variable X , $Y = g(X)$ is a continuous random variable if $Y^{-1}((-\infty, y]) = \{\omega \in \Omega \mid Y(\omega) \leq y\} \in \mathcal{F}$

注意①我们一开始转化出 Y 的 CDF (F_Y), 随后才能找到 Y 的 PDF (f_Y)

②注意 CDF 转化时, $g(x)$ 是 increasing 还是 decreasing. decreasing 的话要处理

③边缘元素对于 discrete variable CDF 很重要! $P(a \leq X \leq b) = F_X(b) - (F_X(a) -)$, 其中 $(F_X(a) -) = \lim_{x \rightarrow a^-} F_X(x)$. 如果是 continuous variable 则不用

• joint distribution function

(1) the joint distribution function of the random variables (X, Y) is defined as the mapping $F_{X,Y}: \mathbb{R}^2 \rightarrow [0, 1]$ given by $F_{X,Y}(x, y) = P(\{\omega \in \Omega \mid X(\omega) \leq x, Y(\omega) \leq y\})$ or $F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$

① $F_{X,Y}$ is non-decreasing in each variable

② $F_{X,Y}$ is continuous from right

注意: 对于 discrete random variable, 注意 $P(X < x)$ 的形式!

③ $\lim_{x,y \rightarrow +\infty} F_{X,Y}(x, y) = 1, \lim_{x,y \rightarrow -\infty} F_{X,Y}(x, y) = 0$

④ marginal distributions: $F_X(x) = \lim_{y \rightarrow \infty} F_{X,Y}(x, y)$
 $F_Y(y) = \lim_{x \rightarrow \infty} F_{X,Y}(x, y)$

注意:

$$P(X \leq x, Y > y) + P(X \leq x, Y \leq y) = P(X \leq x)$$

$$\text{so } P(X \leq x, Y > y) = F_{X,Y}(x, \infty) - F_{X,Y}(x, y)$$

(2) the random variables X and Y are independent if for all $x, y \in \mathbb{R}$, $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$, or

we call the random variables $X_1, \dots, X_n$ are independent if

$$P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) = P(X_1 \leq x_1)P(X_2 \leq x_2) \dots P(X_n \leq x_n)$$

(3) we call the random variables $X_1, \dots, X_n$ pairwise independent if $F_{X_i, X_j}(x_i, x_j) = F_{X_i}(x_i)F_{X_j}(x_j)$ for all $x_i, x_j \in \mathbb{R}$ whenever $i \neq j$

注意: pairwise independent $\neq$ independent

(4) for independent random variables $X_1, \dots, X_n$ and continuous functions $f_i: \mathbb{R} \rightarrow \mathbb{R}$ for $i=1, 2, \dots, n$, the transformed random variables $Y_i = f_i(X_i), \dots, Y_n = f_n(X_n)$ are still linearly independent

(5) let X, Y be discrete random variables on $(\Omega, \mathcal{F}, P)$, their joint probability mass function denoted by $p_{X,Y}$ is denoted as the function: $p_{X,Y}: \mathbb{R}^2 \rightarrow [0, 1]$ given by: $p_{X,Y} = P(\{\omega \in \Omega \mid X(\omega) = x, Y(\omega) = y\})$.

we have

$$\textcircled{1} p_{X,Y}(x, y) \geq 0, \textcircled{2} \sum_x \sum_y p_{X,Y}(x, y) = 1 \quad [\text{2个证明pmf条件}]$$

$$\textcircled{3} p_X(x) = \sum_y p_{X,Y}(x, y), p_Y(y) = \sum_x p_{X,Y}(x, y)$$

$$\textcircled{4} \text{independent if } p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

(6) we call random variables (X, Y) on $(\Omega, \mathcal{F}, P)$ continuous

independent
对比pdf/pmf
X cdf 为 F

if $F_{X,Y}(x,y) = \int_{u=-\infty}^x \int_{v=-\infty}^y f_{X,Y}(u,v) du dv$

for a function $f_{X,Y}: \mathbb{R}^2 \rightarrow \mathbb{R}$ satisfying:

① $f_{X,Y}(u,v) \geq 0$ for all $u,v \in \mathbb{R}$, ② $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v) du dv = 1$

③ $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$ if derivative exists.

④ $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$, $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx$

⑤ independent if $f_{X,Y}(x,y) = f_X(x) f_Y(y)$

(7) consider two (one-dimensional) random variables X and Y on the same sample space with expectations $\mu_X = E(X)$ and $\mu_Y = E(Y)$, the covariance of X and Y is defined as $\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)$ if the expectation takes a finite value, we define the correlation of X and Y as $\rho = \text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$

注意: ① $-1 \leq \rho \leq 1$

② 在这里, $E(X)$ 考虑 $\int_{-\infty}^{\infty} x f_X(x) dx$ 或 $\sum x p_X(x)$

$E(XY)$ 考虑 $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy$ 或 $\sum_x \sum_y xy p_{X,Y}(x,y)$

(8) let X, Y denote independent and jointly discrete/continuous random variables with finite expectation, then $E(XY) = E(X)E(Y)$, so $\text{Cov}(X,Y) = 0$.

proof $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dx dy$ 或 $\sum_x \sum_y xy p_{X,Y}(x,y)$
 $= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) dx dy$ 或 $\sum_x \sum_y xy p_X(x) p_Y(y)$
 $= \int_{-\infty}^{\infty} x f_X(x) dx \int_{-\infty}^{\infty} y f_Y(y) dy$ 或 $\sum_x x p_X(x) \sum_y y p_Y(y) = E(X)E(Y)$

注意: X, Y independent $\Rightarrow \text{Cov}(X,Y) = 0$, $\text{Cov}(X,Y) = 0 \nRightarrow X, Y$ independent

(9) let X, Y denote two jointly discrete/continuous random variables with finite expectation, then $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$

• probability generating functions:

(1) let X denote a discrete random variable with $\text{Im} X \leq \mathbb{N}$, we denote

$$S_X = \{s \in \mathbb{R} \mid \sum_{x=0}^{\infty} |s|^x P(X=x) < \infty\}$$

then the pgf of X is defined as the function $G_X: S_X \rightarrow \mathbb{R}$ given by

$$G_X(s) = E(s^X) = \sum_{x=0}^{\infty} s^x P(X=x)$$

注意: we observe pgf is well-defined for $|s| \leq 1$ since

$$\sum_{x=0}^{\infty} |s|^x P(X=x) \leq \sum_{x=0}^{\infty} P(X=x) = 1 < \infty$$

(2) let X, Y denote discrete random variables with $\text{Im} X, \text{Im} Y \leq \mathbb{N}$, their pg

are denoted by G_X and G_Y respectively, then $G_X(s) = G_Y(s)$, for all $s \in S_X \cap S_Y$ if and only if $P(X=x) = P(Y=x)$ for all $x=0,1,2,\dots$

Proof 只证明 if $G_X(s) = G_Y(s)$ for all possible s , then $P(X=x) = P(Y=x)$

$$G_X(0) = G_Y(0) \rightarrow P(X=0) = P(Y=0)$$

$$G'_X(0) = G'_Y(0) \rightarrow P(X=1) = P(Y=1)$$

$\vdots$

$$G_X^{(n)}(0) = G_Y^{(n)}(0) \rightarrow n!P(X=n) = n!P(Y=n) \rightarrow P(X=n) = P(Y=n)$$

$\vdots$

(3) let X, Y be independent discrete random variables with $\text{Im } X, \text{Im } Y \in \mathbb{N}$, then $G_{X+Y}(s) = G_X(s)G_Y(s)$ for all $s \in S_X \cap S_Y$

Proof

$$G_{X+Y}(s) = E(s^{X+Y}) = E(s^X s^Y) = E(s^X)E(s^Y) = G_X(s)G_Y(s)$$

(4) for $k \in \mathbb{N}$, we call $E(X^k)$ the k th moment of X .

(5) let X be a discrete random variable with $\text{Im } X \in \mathbb{N}$, let $k \in \mathbb{N}$, then the k th derivative of pgf is:

$$\frac{d^k}{ds^k} G_X(s) \Big|_{s=1} = E[X(X-1)\cdots(X-k+1)]$$

Proof

$$\frac{d^k}{ds^k} G_X(s) = E[X(X-1)\cdots(X-k+1)s^{X-k}]$$

注意: $\textcircled{1} E(X) = G'_X(1), \text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

$\textcircled{2} X \sim \text{Bem}(p), G_X(s) = 1-p+sp$

$X \sim \text{Bin}(n, p), G_X(s) = (sp+1-p)^n$

$X \sim \text{Poi}(\lambda), G_X(s) = e^{\lambda(s-1)}$

• moment generating function

(1) let X be a random variable, its mgf is defined as $M_X(t) = E(e^{tx})$, provided the expectation exists in some neighborhood of 0.

notice that $M_X(t) = E(e^{tx}) = \sum_{n=0}^{\infty} E(X^n) \frac{t^n}{n!}$ by expanding e

$$M_X(t) = \sum_{n=0}^{\infty} M_X^{(n)}(0) \frac{t^n}{n!}$$

by expanding $M_X(t)$ around $t=0$

so $M_X^{(n)}(0) = E(X^n)$

如果要求 $E(X^n)$, 凑格式即可!

(2) for $a, b \in \mathbb{R}$, $M_{X+b}(t) = e^{bt} M_X(at)$

(3) let $X_1, \dots, X_n$ denote a sequence of independent random variables with mgf $M_{X_1}, \dots, M_{X_n}$, then $M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t)$

(4) if the mgfs of the random variables X and Y exist and $M_X(t) = M_Y(t)$ in a neighborhood of zero, then $F_X(u) = F_Y(u)$ for all u .

• characteristic function (不考)

8

for a random variable X , its characteristic function is defined as

$$\phi_X(t) = E(e^{itX}) = E[\cos tX + i \sin tX]$$

it turns out that characteristic function exists for all distributions.

• Laplace transformation (1.7.3)

for a non-negative random variable X , its Laplace transformation is defined as $\{s(t) = E(e^{-tX})$ for all $t \geq 0$

• conditional distribution

(1) let Y denote a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$. consider an event $B \in \mathcal{F}$ such that $P(B) > 0$. the conditional expectation of Y given B is defined as $E(Y|B) = \sum_{y \in \mathcal{M}_Y} y P(Y=y|B)$

(2) consider a partition $\{B_i : i \in \mathcal{I}\}$ of Ω with $P(B_i) > 0$ for all $i \in \mathcal{I}$. let Y denote a discrete random variable with finite expectation. then

$$E(Y) = \sum_{i \in \mathcal{I}} E(Y|B_i) P(B_i)$$

Proof:

$$E(Y) = \sum_y y P(Y=y) = \sum_y y \sum_{i \in \mathcal{I}} P(Y=y|B_i) P(B_i)$$

$$= \sum_{i \in \mathcal{I}} \sum_y y P(Y=y|B_i) P(B_i) \quad \text{absolute convergence, allows reordering}$$

$$= \sum_{i \in \mathcal{I}} E(Y|B_i) P(B_i)$$

(3) suppose (X, Y) are jointly discrete random variables. consider the event $B = \{X=x\}$ for some $x \in \mathcal{R}$ such that $p_X(x) = P(X=x) > 0$. then the conditional distribution/probability mass function of Y given $X=x$ is:

$$P_{Y|X}(y|x) = P(Y=y|X=x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \quad \text{for } y \in \mathcal{R}$$

also the conditional expectation of Y given $X=x$ is:

$$E(Y|X=x) = \sum_y y P_{Y|X}(y|x) \quad \text{provided the sum is absolutely convergent}$$

(4) note that we can also formulate an independence condition in terms of conditional p.m.f.s: discrete X and Y are independent if and only if:

$$P(Y=y|X=x) = P(Y=y) \quad \text{for all } x, y \text{ such that } P(X=x) > 0$$

also, we get a Bayes' type result of the form:

$$P_{Y|X}(y|x) p_X(x) = p_{X,Y}(x,y) = p_{X|Y}(x|y) p_Y(y) \quad \text{for all } x, y \text{ which } p_X(x), p_Y(y) > 0$$

(4) suppose (X, Y) are jointly continuous:

$$P(Y \leq y | x \leq X \leq x+\epsilon) = \frac{P(Y \leq y, x \leq X \leq x+\epsilon)}{P(x \leq X \leq x+\epsilon)} = \frac{\int_x^{x+\epsilon} \int_{-\infty}^y f_{X,Y}(u,v) dv du}{\int_x^{x+\epsilon} f_X(u) du}$$

$$= \frac{\frac{1}{\epsilon} \int_x^{x+\epsilon} \int_{-\infty}^y f_{X,Y}(u,v) dv du}{\frac{1}{\epsilon} \int_x^{x+\epsilon} f_X(u) du}$$

$$\text{for } \epsilon \rightarrow 0: P(Y \leq y | x \leq X \leq x+\epsilon) = \frac{\int_{-\infty}^y f_{X,Y}(x,v) dv}{f_X(x)}$$

$$F_{Y|X=x}(y|x) = \int_{-\infty}^y \frac{f_{X,Y}(x,v)}{f_X(x)} dv$$

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- (6) the independence conditions in terms of conditional p.d.f.s:
 X and Y are independent if and only if $f_{Y|X}(y|x) = f_Y(y)$ for all x, y such that $f_X(x) > 0$

also notice that for two jointly continuous random variables X, Y , the conditional expectation of Y given $X=x$ is

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x,y)}{f_X(x)} dy$$

provided that $f_X(x) > 0$

- (5) for jointly continuous random variables X, Y with $E|Y| < \infty$, we have $E(Y) = \int_{\{x: f_X(x) > 0\}} E(Y|X=x) f_X(x) dx$

proof:

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dx dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right) f_X(x) dx = \int_{\{x: f_X(x) > 0\}} E(Y|X=x) f_X(x) dx \end{aligned}$$

注意: 对于 double integral, 上下限的判断十分重要! 一定贴合定义! 上下限可以是变数! 通过结果是常数还是函数判断上下限十分有用。

各种 pdf / cdf, 一定要标定义域!

Statistics

- (1) for any random variable X and constant $a \in \mathbb{R}$, $E[(X-a)^2] \geq E[(X-E(X))^2]$
 that is said, given a random variable X , over all $a \in \mathbb{R}$,

$$\min_{a \in \mathbb{R}} E[(X-a)^2] = E[(X-E(X))^2]$$

proof: $E[(X-a)^2] = E[(X-E(X) + E(X)-a)^2]$

- (2) suppose that the random variable X is known to only take values in the bounded range $[a, b]$, then $\text{Var}(X) \leq \frac{(b-a)^2}{4}$

- (3) the support of a function is the set of values in the domain that are not mapped to 0.

- (4) for each positive integer k , the k th moment of the random variable X is denoted μ'_k and is defined by $\mu'_k = E[X^k]$
 furthermore, the k th central moment of X , denoted μ_k , is defined by $\mu_k = E[(X-\mu)^k]$

- (5) suppose random variables $X_1, \dots, X_n$ are independent sampled from a distribution F_X with mean μ and finite variance σ^2 , then $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, $E(S^2) = E(\frac{1}{n-1} \sum (X_i - \bar{X})^2) = \sigma^2$

Proof: $E(\bar{X}) = E(\frac{1}{n} \sum X_i) = \frac{1}{n} E(\sum X_i) = \mu$

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum X_i\right) = \frac{1}{n^2} \text{Var}(\sum X_i) = \frac{\sigma^2}{n}$$

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1} \sum (X_i - \bar{X})^2\right) = \frac{1}{n-1} E\left(\sum (X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1} E\left(\sum X_i^2 - n\bar{X}^2\right) = \frac{1}{n-1} \left(\sum E(X_i^2) - n E(\bar{X}^2)\right) \\ &= \frac{1}{n-1} \left[\sum (\text{Var}(X_i) + (E(X_i))^2) - n (\text{Var}(\bar{X}) + (E(\bar{X}))^2)\right] \\ &= \frac{1}{n-1} \left(n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2 \end{aligned}$$

(6) if a random variable X can only take nonnegative values, then

$$P(X \geq \alpha) \leq \frac{E(X)}{\alpha} \quad \text{for all } \alpha > 0.$$

proof: suppose $\alpha \in \mathbb{R}$ is fixed. let Y_α be the discrete random variable

$$Y_\alpha = \begin{cases} \alpha & \text{if } X \geq \alpha \\ 0 & \text{if } X < \alpha \end{cases}$$

notice that $Y_\alpha \leq X$, meaning that $E(Y_\alpha) \leq E(X)$

$$E(Y_\alpha) = \alpha P(X \geq \alpha) \leq E(X) \quad P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

(7) if X is a random variable with mean μ and variance σ^2 , then for all $c > 0$,

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

proof: use Markov's inequality and let $Z = (X - \mu)^2$, $\alpha = c^2$, then

$$P((X - \mu)^2 \geq c^2) \leq \frac{E[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

substitute X by $|X - \mu|$

[注意] 换皮形式: if X is a random variable with mean μ and variance

$$\sigma^2, \text{ then for all } k > 0: P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

我们可以考虑 X 的有关形式以确定在精确度下所需的 n

(8) given a sample of random variables $X_1, X_2, \dots, X_n$ with parameter θ , and those are observed as $x_1, x_2, \dots, x_n$

then $\hat{\theta} = \hat{\theta}(X_1, X_2, \dots, X_n)$ is an estimator (function)

$\hat{\theta} = \hat{\theta}(x_1, x_2, \dots, x_n)$ is an estimate (value)

(9) the standard error of the estimator $\hat{\theta}$ is defined as the square root of its variance

the estimation error of the estimator $\hat{\theta}$ of a parameter θ is defined to be

$\hat{\theta} - \theta$. the bias of the estimation $\hat{\theta}$ of a parameter θ is noted by $b_\theta(\hat{\theta})$

$$b_\theta(\hat{\theta}) = E(\hat{\theta}) - \theta$$

the mean squared error of an estimator $\hat{\theta}$ of a parameter θ is:

$$E[(\hat{\theta} - \theta)^2] = (b_\theta(\hat{\theta}))^2 + \text{Var}(\hat{\theta})$$

(10) suppose we have:

- random variables $X_1, X_2, \dots, X_n \sim F_X$ with parameters θ

- write $\vec{X} = (X_1, X_2, \dots, X_n)$, and corresponding observations $\vec{x} = (x_1, x_2, \dots, x_n)$

an interval estimate of a parameter θ is any pair of functions $L(\vec{x})$ and

$U(\bar{x})$ of a sample $\bar{x}$ that satisfies $L(\bar{x}) \leq U(\bar{x})$ for all $\bar{x}$.
for an interval estimator $[L(\bar{X}), U(\bar{X})]$, the coverage probability is the probability that the random interval $[L(\bar{X}), U(\bar{X})]$ covers the true parameter θ

if $[L(\bar{X}), U(\bar{X})]$ is designed so that

$$P_{\theta}(L(\bar{X}) \leq \theta \leq U(\bar{X})) \geq 1 - \alpha$$

for every θ and some $\alpha \in (0, 1)$, then $[L(\bar{X}), U(\bar{X})]$ is a $1 - \alpha$ confidence interval

注意: 比如如果我们用 inequality 来解

$$P(|Z - E(Z)| \geq C) \leq \frac{\text{Var}(Z)}{C^2}$$

$$\text{if } Z = \bar{X}$$

$$P(|\bar{X} - E(\bar{X})| \geq C) \leq \frac{\text{Var}(\bar{X})}{C^2}$$

$$P(|\bar{X} - \theta| \geq C) \leq \frac{\sigma^2}{nC^2}$$

$$P(|\bar{X} - \theta| < C) \geq 1 - \frac{\sigma^2}{nC^2}$$

一步一步来, 别搞到 $E(\bar{X})$ $\text{Var}(\bar{X})$ 与 $E(Z)$ $\text{Var}(Z)$

从这题起!

• use statistical theory, we have derived

$$P(\bar{X} - k \frac{\sigma}{\sqrt{n}} < \theta < \bar{X} + k \frac{\sigma}{\sqrt{n}}) \geq 1 - \frac{1}{k^2}$$

(11) 一组数据中可能有多 mode

(12) median m 的定义: $P(X \leq m) \geq 0.5$ and $P(X \geq m) \geq 0.5$

一组数据中可能有多 median

(13) suppose that m is a median of X : $\min_a E(|X - a|) = E(|X - m|)$

suppose that μ is the mean of X : $\min_a E[(X - a)^2] = E[(X - \mu)^2]$

(14) given a sample of observations $x_1, \dots, x_n$, the sample median m is defined as

$$m = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) & \text{if } n \text{ even} \end{cases}$$

where $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\} = \{x_1, x_2, \dots, x_n\}$ and $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$

(15) let F_X the cumulative distribution function of X then $F_X^{-1}: (0, 1) \rightarrow \mathbb{R}$

is called quantile function for distribution of X

for all $p \in (0, 1)$, $F_X^{-1}(p)$ is defined to be the smallest value x such that $F_X(x) \geq p$

(16) lower quantile $q_{0.25} = F_X^{-1}(0.25)$

higher quantile $q_{0.75} = F_X^{-1}(0.75)$

(17) index for median: $i_m = \frac{n+1}{2}$

index for $q_{0.25}$: $i_{0.25} = \frac{1}{2}(\lfloor Lin \rfloor + 1)$

index for $q_{0.75}$: $i_{0.75} = n - i_{0.25} + 1$

找 median / Q1 / Q3

要用公式!

samples

(1) suppose that $X_1, X_2, \dots, X_n$ are independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$, where μ_i and σ_i^2 are finite. then, define $Y = \sum_{i=1}^n X_i$

$$Y \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$$

proof: $M_Y(t) = M_{\sum_{i=1}^n X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}}$
 $= e^{\sum_{i=1}^n \mu_i t + \sum_{i=1}^n \frac{\sigma_i^2 t^2}{2}}$

2) suppose $X_1, X_2, \dots, X_n$ are identical independent random variables distributed according to $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
 in such a way, known $\bar{x}$ and σ^2 , the α confidence interval for μ is

$$(\bar{x} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

[注意: 1. confidence interval 不表明 interval 有 α 的概率包含 μ ! 对于单个 interval, 其包含 μ 的概率是 0 或 1!

应翻译为 α 的 interval 中会包含 μ

2. 如果查表时 Z_α 不存在, 查找 Z_β , $\beta > \alpha$ 微微大一些]

3) suppose $Z_1, Z_2, \dots, Z_n \sim N(0, 1)$ are identical independent and $\vec{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$
 if A is an orthogonal $n \times n$ matrix, $\vec{Y} = A\vec{Z}$, with $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$

① then $Y_1, \dots, Y_n \sim N(0, 1)$ are identical independent as well.

$$\textcircled{2} \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$$

proof ②: $\sum_{i=1}^n Y_i^2 = \vec{Y}^T \vec{Y} = \vec{Z}^T A^T A \vec{Z} = \vec{Z}^T \vec{Z} = \sum_{i=1}^n Z_i^2$

①: 已知: support Z is a continuous random variable with density f_Z and $h: \mathbb{R} \rightarrow \mathbb{R}$ is strictly increasing/decreasing, with inverse function h^{-1} then $Y=h(Z)$ has density $f_Y(y) = f_Z(h^{-1}(y)) \left| \frac{d}{dy} h^{-1}(y) \right|$
 so in our case: $f_Y(\vec{y}) = f_Z(A^{-1}\vec{y}) |1| = f_Z(\vec{z}) =$

$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n z_i^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} \sum_{i=1}^n y_i^2}$$

so $\vec{y}$ and $\vec{z}$ has same probability distribution.

4) suppose $X_1, \dots, X_n$ are identical independent random variables distributed according to $N(\mu, \sigma^2)$, with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
 then $\bar{X}$ and S^2 are independent random variables, and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.

[注意: for $Z_1, Z_2 \sim N(0, 1)$, $Z_1^2 \sim \chi_1^2$, $Z_1^2 + Z_2^2 \sim \chi_2^2$]

proof: assume $X_i \sim N(\mu, \sigma^2)$ identical independent

then let $Z_i = \frac{X_i - \mu}{\sigma}$, $Z_i \sim N(0, 1)$ identical independent

we can show: $\bar{Z} = \frac{\bar{X} - \mu}{\sigma}$, $\sum_{i=1}^n (Z_i - \bar{Z})^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$

construct $n \times n$ orthogonal matrix A with $\begin{pmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{pmatrix}$ on the first row.

if let $\vec{Y} = A\vec{Z}$, then $Y_i \sim N(0, 1)$ identical independent

$$\bullet \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2$$

$$\bullet Y_1 = \sqrt{n} \bar{Z} = \sqrt{n} \frac{\bar{X} - \mu}{\sigma}, \quad \sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Z_i^2 - n \bar{Z}^2 = \sum_{i=1}^n (Z_i - \bar{Z})^2 = \frac{(n-1)S^2}{\sigma^2}$$

thus, $\sum_{i=1}^n Y_i^2$ is independent of Y_i , so S^2 and $\bar{X}$ are independent, and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

(5) student's t distribution

suppose $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ are independent and $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, then the random variable $T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ can be written

as $T = \frac{U}{\sqrt{\frac{V}{p}}}$, where $U \sim N(0, 1)$, $p = n-1$, $V \sim \chi_p^2$. U and V are independent

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

规范格式:

$$P(t_{n-1, \frac{\alpha}{2}} < T < t_{n-1, 1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(t_{n-1, \frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{n-1, 1-\frac{\alpha}{2}}) = 1-\alpha$$

$$P(t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}} + \bar{X} < \mu < t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} + \bar{X}) = 1-\alpha$$

$$P(\bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) = 1-\alpha$$
 因为t是symmetric

(6) confidence interval for population variance

suppose $X_1, \dots, X_n$ are identical independent random variables distributed according to $N(\mu, \sigma^2)$, with $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

we know that $\bar{X}$ and S^2 are independent and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$, so

$$P(\chi_{n-1, \frac{\alpha}{2}}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{n-1, 1-\frac{\alpha}{2}}^2) = 1-\alpha$$

$$P(\frac{1}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi_{n-1, \frac{\alpha}{2}}^2}) = 1-\alpha$$
 此时 χ^2 不是symmetric

$$P(\frac{(n-1)S^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1, \frac{\alpha}{2}}^2}) = 1-\alpha$$

(7) hypothesis testing.

- 在 H_0 与 H_1 的选择, 倾向于将严重后果的条件化为 H_1 (将你想要的结果化为 H_1)
- 结果不能证明 H_0 正确与否, 只能证明 success / fail to reject H_0
- let X be a random variable with continuous cumulative distribution function F_X , and $Y = F_X(X)$, then $Y \sim U(0, 1)$

Proof: $P(Y \leq 0) = P(F_X(X) \leq 0) = 0$

$$P(Y \geq 1) = P(F_X(X) \geq 1) = 0$$

$$\text{let } y \in (0, 1): P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = y$$

similarly, we can show, if T be a random variable with continuous cumulative distribution function F_T :

$$\text{when } P = F_T(T), P \sim U(0, 1)$$

$$\text{when } P = 1 - F_T(T), P \sim U(0, 1)$$

$$\text{when } P = 2 \min\{F_T(T), 1 - F_T(T)\}, P \sim U(0, 1)$$

} 用于 one-sided

} 用于 two-sided

so once T is observed as t , P is observed as p -value p

这样做是为了确保 $P(P < \alpha) = \alpha$. when $p < \alpha$, reject H_0 .



• type I error: H_0 正确拒绝了 H_1 , 出现概率为 α

type II error: H_1 正确拒绝了 H_0 , 我们定义其出现概率为 β .

我们定义 power of hypothesis test 为 $1-\beta$, 即 H_1 正确时选择 H_1 的概率.

[注意: 我们只能选择 α , 但不能选择 β . 但是 when α decreases, β increases]

(8) student's two-sample test.

we have two groups of data with unknown means and variance, and wish to test if the two groups have the same mean or not. if we assume observations are independently normally distributed, then we can use student's two-sample test

$H_0: \theta_1 = \theta_2$ $H_1: \theta_1 \neq \theta_2$, assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown.

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \text{ where } S_p^2 = \frac{1}{n+m-2} ((n-1)S_x^2 + (m-1)S_y^2)$$

(9) permutation test.

如果我们不知道 two group of data 的 distribution 情况下, 想要判断其 mean 是否相同. 此时我们使用 permutation test.

- 步骤:
- compute statistic t for sample $\bar{x}$ and $\bar{y}$, $t = \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$
 - combine observations into $\bar{Z} = (x_1, x_1, \dots, x_n, y_1, y_1, \dots, y_m)$
 - randomly choose n observations as sample 1 $\bar{x}^*$, and remaining m observations as sample 2 $\bar{y}^*$
 - compute statistic t^* for samples $\bar{x}^*$ and $\bar{y}^*$
 - repeat process B times. (to get all possible permutation test statistics). based on the empirical permutation distribution of statistics, we sort statistics and find upper and lower limit for $(1-\alpha)\%$.
 - if t falls outside limits, then reject H_0 .

	10	12	5	8	16	4	11	9	7
$\bar{g}$	x	x	x	x	x	y	y	y	y
$\bar{g}^*1$	x	y	y	x	y	x	x	y	x
$\bar{g}^*2$	y	y	x	x	x	y	x	y	x

• Covariance

(1) recall $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$, $\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$, $-1 \leq \rho_{XY} \leq 1$

$$\text{Cov}(X, X) = \text{Var}(X), \text{Cov}(X, Y) = \text{Cov}(Y, X)$$

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

(2) for any two random variables X, Y , $|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$

proof: define $f(t) = E[(X - \mu_X)t + (Y - \mu_Y)]^2 = \sigma_X^2 t^2 + 2t\text{Cov}(X, Y) + \sigma_Y^2 \geq 0$

$$\text{so } \Delta = 4(\text{Cov}(X, Y))^2 - 4\sigma_X^2 \sigma_Y^2 \leq 0, |\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$$

(2) suppose $X_1, X_2, \dots, X_n \sim F_X, Y_1, Y_2, \dots, Y_n \sim F_Y$, then the sample covariance is $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$

the sample correlation is $\rho_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$

(3) for any two random variables X, Y

• $|\rho_{XY}| = 1 \Leftrightarrow P(Y = \alpha X + b) = 1$ for some $\alpha \neq 0, b \in \mathbb{R}$

• if $\rho_{XY} = 1$, then $\alpha > 0$

• if $\rho_{XY} = -1$, then $\alpha < 0$

Proof lemma: suppose Z is a nonnegative random variable. then

$$E(Z) = 0 \Leftrightarrow P(Z = 0) = 1$$

consider $f(t) = E[(X - \mu_X)t + (Y - \mu_Y)]^2$

$$= \sigma_X^2 t^2 + 2\text{Cov}(X, Y)t + \sigma_Y^2$$

so $|\rho_{XY}| = 1 \Leftrightarrow (\text{Cov}(X, Y))^2 = \sigma_X^2 \sigma_Y^2 \Leftrightarrow$ for $f(t), \Delta = 0$

$\Leftrightarrow f(t) = 0$ has only one unique root

$$\text{when } t = -\frac{\text{Cov}(X, Y)}{\sigma_X^2} = -\left(\frac{\sigma_Y}{\sigma_X}\right)\rho_{XY}$$

$$E[(X - \mu_X)t + (Y - \mu_Y)]^2 = 0 \Leftrightarrow$$

$$P[(X - \mu_X)t + (Y - \mu_Y)]^2 = 0 \Leftrightarrow$$

$$P[(X - \mu_X)t + (Y - \mu_Y)] = 0 \Leftrightarrow$$

$$P(Y = \alpha X + b) = 1, \alpha = -t, b = \mu_X t + \mu_Y$$

notice that $\alpha = \left(\frac{\sigma_Y}{\sigma_X}\right)\rho_{XY}$, so α and ρ_{XY} share same sign

(4) suppose the pairs of measurements $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are observed, define the pairs $(u_1, v_1), (u_2, v_2), \dots, (u_n, v_n)$ by

$u_i = ax_i + b, v_i = cy_i + d$ for some $a, b, c, d \in \mathbb{R}$ with $a, c > 0$, then

$$\rho_{XY} = \rho_{UV}$$

correlation doesn't imply causation!

likelihood

(1) suppose we have random variables X , statistical model $\{P_\theta | \theta \in \Theta\}$ for X , each P_θ is specified by probability density/mass function f_θ , and observed data x . the likelihood function $L(\cdot | x): \Theta \rightarrow \mathbb{R}$ is defined by $L(\theta | x) = f_\theta(x) = f(x | \theta)$ for any $\theta \in \Theta$

[注意: x 是 fixed, θ 是 variable!]

(2) suppose we have a statistical model $\{P_\theta | \theta \in \Theta\}$ where each P_θ is discrete and specified by the probability mass function f_θ . then, given data x , one can interpret $f_\theta(x)$ as the probability of obtaining the data x given that the true value of the parameter is θ

[注意: $L(\theta | x)$ is not the probability that θ is true value!]

12

如果 $L(\theta_1|x) > L(\theta_2|x)$, we say θ_1 is a more plausible value than θ_2 for the true value of θ . 不要用 probable!

(3) 对于 continuous random variables, $L(\theta|x) = P_\theta(x) = 0$. 因此为了做比较我们使用 likelihood ratio

for discrete $\frac{L(\theta_1|x)}{L(\theta_2|x)} = \frac{P_{\theta_1}(X=x)}{P_{\theta_2}(X=x)}, \theta_1, \theta_2 \in \Theta$

for continuous: $\frac{L(\theta_1|x)}{L(\theta_2|x)} \approx \frac{P_{\theta_1}(x-\delta < X < x+\delta)}{P_{\theta_2}(x-\delta < X < x+\delta)}, \theta_1, \theta_2 \in \Theta$

(4) given the likelihood function $L(\cdot|x)$, any function $L'(\cdot|x) = cL(\cdot|x)$ for $c > 0$ is an equivalent likelihood function for the parameter θ

(5) suppose $L(\theta|x)$ is the likelihood of the parameter θ given the observed data x . then the parameter value $\hat{\theta}(x)$ at which $L(\theta|x)$ attains its maximum as a function of θ , with x held fixed, is called the maximum likelihood estimate of θ , and then the maximum likelihood estimator of the parameter θ based on the random sample X is $\hat{\theta}(X)$

[注意: 求导 + 考虑 boundary $\Rightarrow$ 找 maximum likelihood estimate

将 $\hat{\theta}(x)$ 换成 $\hat{\theta}(X) \Rightarrow$ 得到 maximum likelihood estimator]

(6) 如果求 MLE 过于困难, 可以最大化 $L(\theta|x)$ 的 transform
最出名的为 $\log L(\theta|x)$.

• linear regression

(1) assume the true model is $Y = \beta_0 + \beta_1 X$ and $y_i = \beta_0 + \beta_1 x_i + e_i$, where β_0, β_1 are unknown parameters and e_i is unknown error

so $\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is the residual, and we want to pick up the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimise residual sum of squares $RSS = \sum_{i=1}^n (\hat{e}_i)^2$.

by calculation, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$, where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$,
 $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$

(2) assume the simple linear regression model is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where

- x_i are fixed, known values, β_0, β_1 are fixed, unknown values
- $\epsilon_i \sim N(0, \sigma^2)$ for unknown σ^2 are independent random variables.

then $E(Y_i) = \beta_0 + \beta_1 x_i, \text{Var}(Y_i) = \sigma^2, Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

by calculating the MLE, $\ln L(\beta_0, \beta_1, \sigma^2 | x, y) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$

$\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}, \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \hat{\sigma}^2 = \frac{1}{n} \hat{RSS}_{xy} = \frac{1}{n} \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

and the statistic $R^2 = 1 - \frac{\hat{RSS}_{xy}}{S_{yy}}, \hat{RSS}_{xy} = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

[注意: $\hat{RSS}_{xy} = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}, R^2 = \frac{(S_{xy})^2}{S_{xx} S_{yy}} = r^2_{XY}$]

based on those, we get $\hat{\epsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

通过plot出residual, 我们可以判断我们初始假设是否合理

如果不合理可通过transformation x, y , 后重新建模, 或通过remove outlier后重新计算

(b) R^2 measures the 'difference between the sum of squares and the estimated residual sum of squares, normalised by sum of squares

$R^2 \approx 1$: the fitted line is close to data.

$R^2 \approx 0$: the fitted line is close to the average horizontal line

• bootstrap.

(1) given data $\vec{x} = (x_1, x_2, \dots, x_n)$, the empirical distribution $\hat{F}$ on the (multi)set $\vec{x}$ is defined as the discrete distribution with probability mass function

$$p_{\vec{x}}(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \vec{x} \\ 0 & \text{otherwise} \end{cases}$$

if all of data in $\vec{x}$ are distinct. if one data is not distinct, then $p_{\vec{x}}(x) = \frac{c_{\vec{x}}(x)}{n}$, where $c_{\vec{x}}(x)$ is the number of times that x occurs in $\vec{x}$.

(2) 假设你只有单次数据, 在不知underlying distribution的情况下, 想要知道某-statistic的confidence interval

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

data set



$\hat{F}$

empirical distribution

$\vec{x}^{*1}$

$\vec{x}^{*2}$

$\vec{x}^{*B}$

bootstrap samples of data

放回取样, 取与原始数据集相同个数的元素

$$\theta^{*1} = s(\vec{x}^{*1})$$

$$\theta^{*2} = s(\vec{x}^{*2})$$

$$\dots \theta^{*B} = s(\vec{x}^{*B})$$

bootstrap replication of θ

$$SE_{\theta} \text{ or } (\hat{\theta}_L, \hat{\theta}_U)$$

bootstrap estimate

$$SE_{\theta} = \frac{1}{B-1} \sum_{i=1}^B (\hat{\theta}^{*i} - \frac{1}{B} \sum_{j=1}^B \hat{\theta}^{*j})^2$$

$$(\hat{\theta}_L, \hat{\theta}_U) = (\hat{\theta}^{*}_{(LB \cdot \frac{\alpha}{2})}, \hat{\theta}^{*}_{(LB \cdot (1 - \frac{\alpha}{2}))})$$

• Bonferroni correction

when we compute multiple p-values simultaneously, the probability of at least one p-value being significant is $\leq n\alpha$

therefore, when there are n tests, and the nominal threshold is α , then one should use the adjusted significance threshold $\alpha' = \frac{\alpha}{n}$

• spurious correlations:

correlation doesn't imply causation

13

• anscombe's quartet

特征数据并不能完美描述图像 画图验证!

一切概率考虑 domain!

① 如何证明 sets $A=B$

- 直接拆开求等
- $A \subseteq B$ and $B \subseteq A$
- $(\cup A_i)^c = \cap (A_i^c)$, $(\cap A_i)^c = \cup (A_i^c)$, $A \cup (\cap B_i) = \cap (A \cup B_i)$, $A \cap (\cup B_i) = \cup (A \cap B_i)$

② 公式

- classical probability: $P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$
- Choose k balls from n different balls with replacement (unordered) $\binom{n+k-1}{k}$
- probability space, countable sequence of disjoint events $(A_i)_{i \in \mathbb{N}}$
 $P(\cup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} P(A_i)$ [用这条件时一定要说明 disjoint, $A_i \cap A_j = \emptyset$]
- conditional probability: $P(A|B \cap E) = \frac{P(B \cap A \cap E) P(A|E)}{P(B|E)}$

$\{B_i\}_{i \in \mathbb{N}}$ denote the partition of Ω :

$$P(A|E) = \sum_{i \in \mathbb{N}} \frac{P(A \cap B_i \cap E)}{P(E)} = \sum_{i \in \mathbb{N}} P(A|B_i \cap E) P(B_i|E)$$

[使用 $\{B_i\}$ partition 分这种办法去简便运算]

- increase/decrease to A : $(A_i \uparrow A) \Leftrightarrow (A_i^c \downarrow A^c)$, $\lim_{i \rightarrow \infty} P(A_i) = P(A)$ [套用此证明 $A_i \uparrow A, A_i \downarrow A$]
- for function $f: X \rightarrow Y$, $B_i \subseteq Y$, $f^{-1}(\cup_{i \in \mathbb{N}} B_i) = \cup_{i \in \mathbb{N}} f^{-1}(B_i)$
- CDF transformation: (1) 对于 continuous variables 通过 F_X 转化
 注意 transformation 本身是否 decreasing, 如是则需处理 *
 (2) 对于 discrete variables, $P(a \leq X \leq b) = F_X(b) - (F_X(a)^-)$!
- $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - E(X)E(Y)$, $\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$. $[(\text{Cov}(X, Y) = 0) \nRightarrow (X, Y \text{ independent})]$

③ 定义

(1) a collection of subsets of Ω defined by $\mathcal{F}$ is called...

- an algebra on Ω : ① $\emptyset \in \mathcal{F}$, ② $\forall A \in \mathcal{F}, A^c \in \mathcal{F}$, ③ $\forall A_1, A_2 \in \mathcal{F}, A_1 \cup A_2 \in \mathcal{F}$.
- an σ -algebra on Ω : ① $\emptyset \in \mathcal{F}$, ② $\forall A \in \mathcal{F}, A^c \in \mathcal{F}$, ③ $\forall$ countable index set $\mathbb{I}, A_i \in \mathcal{F}, \cup_{i \in \mathbb{I}} A_i \in \mathcal{F}$

(2) a mapping $P: \mathcal{F} \rightarrow \mathbb{R}$ is called a probability measure on $(\Omega, \mathcal{F})$.

① $\forall A \in \mathcal{F}, P(A) \geq 0$, ② $P(\Omega) = 1$

③ $\forall$ countable sequence of disjoint events $(A_i)_{i \in \mathbb{N}} \in \mathcal{F}$, $P(\cup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} P(A_i)$ [不用写出证明]

we then define probability space as $(\Omega, \mathcal{F}, P)$

(3) $A_i \uparrow A, A_i \downarrow A$

(4) CDF is right-continuous. if $x_n \downarrow x$, then $\lim_{n \rightarrow \infty} F_X(x_n) = F_X(x)$ since $\{w | X(w) \leq x_n\} \downarrow \{w | X(w) \leq x\}$

(5) PGF: $G_X(s) = E(S^X) = \sum_{x=0}^{\infty} S^x P(X=x)$, $\frac{d^k}{ds^k} G_X(s)|_{s=0} = k! P(X=k)$

• independent X, Y , $G_{X+Y}(s) = G_X(s)G_Y(s)$, $G_{aX+b}(s) = s^b G_X(s^a)$

• $\frac{d^k}{ds^k} G_X(s)|_{s=1} = E[X(X-1) \dots (X-k+1)]$, $E(X) = G'_X(1)$, $\text{Var}(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

(6) MGF: $M_X(t) = E(e^{tX}) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E(X^n) = \sum_{n=0}^{\infty} \frac{d^n}{dt^n} M_X(t)|_{t=0} \frac{t^n}{n!}$

• independent X, Y , $M_{X+Y}(t) = M_X(t)M_Y(t)$, $M_{aX+b}(t) = e^{bt} M_X(at)$

④ 证明 independence

(pairwise independent) $\nRightarrow$ (independent)

- $P(A)P(B) = P(A \cap B)$
- $P(A|B) = P(A)$ or $P(B|A) = P(B)$
- conditionally independence: $P(A \cap B|C) = P(A|C)P(B|C)$
 [常用其推导出 $P(A \cap B \cap C) = P(A|C)P(B|C)P(C)$]

8

⑤ conditional distribution.

(1) $E(Y) = \sum E(Y|X_i)P(X_i)$ 或 $\int E(Y|X)f_X dx$

(2) for discrete random variables X, Y

• $P_{Y|X}(y|x) = \frac{P_{X,Y}(x,y)}{P_X(x)}$, $P_{X,Y}(x,y) = P_{Y|X}(y|x)P_X(x) = P_{X|Y}(x|y)P_Y(y)$

• independent X, Y : $P_{Y|X}(y|x) = P_Y(y)$ or $P_{X|Y}(x|y) = P_X(x)$

(3) for continuous random variables X, Y

• $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$

• independent X, Y : $f_{Y|X}(y|x) = f_Y(y)$ or $f_{X|Y}(x|y) = f_X(x)$

double integral 注意上下限!

搞清楚是关于哪个variable的function!

④ 如何证明是 discrete random variable

• a mapping from Ω to $\mathbb{R}$

• the image of range of Ω under X is a countable subset of $\mathbb{R}$

• $X^{-1}(x) = \{\omega \in \Omega \mid X(\omega) = x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$

注意: $\{\omega \in \Omega \mid Y(\omega) = y\} = \{\omega \in \Omega \mid g(X(\omega)) = y\}$

$$= \{\omega \in \Omega \mid X(\omega) \in \{z \in \text{Im} X \mid g(z) = y\}\}$$

$$= \{\omega \in \Omega \mid X(\omega) \in \bigcup_{z \in \text{Im} X, g(z)=y} \{z\}\}$$

$$= X^{-1}\left(\bigcup_{z \in \text{Im} X, g(z)=y} \{z\}\right)$$

$$= \bigcup_{z \in \text{Im} X, g(z)=y} X^{-1}(\{z\}) \in \mathcal{F}$$

⑤ sample space Ω : the set of all possible outcome

sample point: ω : elements of Ω

events: collection of ω s

power set $\mathcal{P}(A)$ of A : all subsets of A including $\emptyset, A$

$$F_W(w) = \iint_{\{(x,y) : g(x,y) \leq w\}} f_{X,Y}(x,y) dx dy$$

⑥ Bayes' rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

law of total probability: $P(A) = \sum P(A|B_i)P(B_i)$, $\{B_i\}$ are partitions

LOTUS

$$E(g(x)) = \int g(x) f(x) dx$$

$$= \sum g(x) P(X=x)$$

① statistics.

- $\min_{\alpha} E[(X-\alpha)]^2 = E[(X-E(X))^2]$, $\min_{\alpha} E(|X-\alpha|) = E(|X-m|)$
- for $X \in [\alpha, b]$, $\text{Var}(X) \leq \frac{(b-\alpha)^2}{4}$
- for X takes nonnegative value, $P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$ for $\alpha > 0$
- $P(|X-E(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}$ 或 $P(|X-E(X)| \geq K\sqrt{\text{Var}(X)}) \leq \frac{1}{K^2}$ 注意 M, G 是相对于当前 X 估计 confidence interval 时类 注意假设
- standard error $SE_{\hat{\theta}} = \sqrt{\text{Var}(\hat{\theta})}$
- bias: $b_{\theta}(\hat{\theta}) = E(\hat{\theta}) - \theta$
- mean square errors: $E[(\hat{\theta} - \theta)^2] = [b_{\theta}(\hat{\theta})]^2 + \text{Var}(\hat{\theta})$
- median 的定义: $P(X \leq m) \geq 0.5$ and $P(X > m) \geq 0.5$
- index for median: $i_m = \frac{1+n}{2}$
- index for $q_{0.25}, i_{0.25} = \frac{1}{2}(\lfloor i_m \rfloor + 1)$
- index for $q_{0.75}, i_{0.75} = n - i_{0.25} + 1$

confidence interval 的 α 不是包含正确值的概率

② sample: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

- $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} (E(X^2) - (E(X))^2)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$
- $t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$, $t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$, where $S_p^2 = \frac{1}{n+m-2} ((n-1)S_X^2 + (m-1)S_Y^2)$

- P 值有三种情况.
- type I error: H_0 正确选了 H_1 .
- type II error: H_1 正确选了 H_0 .
- power of test: H_1 正确时选 H_1 的概率

critical threshold value $t_{n, \alpha}$
test statistics t
 D -values: D (判断结果与 α 关系)

③ likelihood.

- $L(\theta|X)$ 不是 θ 的正确概率.
- likelihood ratio: $\frac{L(\theta_1|X)}{L(\theta_2|X)}$
- 求导 + 考虑 boundary $\rightarrow$ 得到 maximum likelihood estimate
- 改 θ 为 $\hat{\theta} \rightarrow$ 得到 maximum likelihood estimator
- 可通过最大化 $L(\theta|X)$ 的 transformation 得到 MLE.

验证 global maximum!
(= 阶导 + boundary limit)

考虑极端特殊情况!

④ linear regression.

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, $\hat{\beta}_0 = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$, $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$, 其中 $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$
- $\hat{\sigma}^2$ (estimate variance for noise) = $\frac{1}{n} R^2 S_{xy}$, 其中 $R^2 S_{xy} = \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$
- $= S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$
- $R^2 = 1 - \frac{R^2 S_{xy}}{S_{yy}} = \frac{(S_{xy})^2}{S_{xx} S_{yy}} = r_{xy}^2$, $R^2 \approx 1$, then fitted line close to data

* sample covariance 的计算公式: $\frac{1}{n-1} \sum (X_n - \bar{x})(Y_n - \bar{y})$
对于 dependent X, Y : $\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y) + 2ab \text{Cov}(X, Y)$

① 三个方法

- bootstrap
- permutation test
- bonferroni correction

justification!

(which is just the pdf/cdf of ...)

(name of model: classical/bayes/LOTUS ...)

(since independent, $P(X|Y) = P(X)$)

assumption!

(independence, randomly selected)

domain!