

概率论与数理统计

Probability theory and mathematical statistics

尹超

中国科学院大学，北京 100049

Carter Yin

University of Chinese Academy of Sciences, Beijing 100049, China

2024.8 – 2025.1

序言

本笔记为概率论与数理统计的笔记整理

讲课教师:

- 李雷, 国科大教授, 中科院数学院南楼 511
- Office Hours: TBD
- telephone: 82541585
- email: lilei@ucas.ac.cn
- Website: <https://mooc.ucas.edu.cn/portal>

助教:

- 王钊, 博士研究生, 中科院数学院-email: wangke22@mails.ucas.ac.cn

目录

序言	I
目录	III
1 事件的概率	1
1.1 Space	1
1.2 Algebra	1
1.3 Definition of Probability	2
1.4 Deductions from the Axioms	3
1.5 Conditional Probability	4
1.6 Total probability and Bayes' Theorem	4
1.7 Independence	5
2 Random Variables and their Distributions	6
2.1 Random Variables	6
2.2 Continuous Random Variables	9
2.3 随机变量的独立性	18
2.4 随机变量的函数的概率分布	18
2.5 Vector Case	20
2.6 两个随机变量	20
2.7 正态分布变量的和的再生性、封闭性	21
2.8 极小值和极大值的分布	23
2.9 Extreme Value Distribution	23
3	25
4	37
4.1 总体 (Population)	37
4.1.1 定义	37
4.1.2 类型	37
4.1.3 说明	37
4.2 样本 (Sample)	37
4.2.1 定义	37
4.2.2 示例	37
4.3 I.I.D. Sample (简单随机样本)	37
4.3.1 定义	37
4.3.2 性质	38
4.4 统计推断	38
4.5 统计量 (Statistic)	38
4.5.1 定义	38

4.5.2 常用统计量	38
4.6 次序统计量 (Order Statistics) 及其有关统计量	38
4.7 三大分布	39
4.8 分布	39
4.8.1 χ^2 分布	39
4.8.2 Gamma 分布	39
4.8.3 Quantiles, Percentiles, 分位点	40
4.8.4 t 分布 (Student distribution)	40
4.8.5 F 分布	40
4.9 正态总体样本均值和样本方差的分布	40
4.10 Sample Mean and Sample Variance	41
4.10.1 Sample Mean and Sample Variance, continued	41

第1章 事件的概率

§ 1.1 Space

Probability Space:

- 概率模型的三个要素, (Ω, Σ, P)
- Samples space, event sets, probability measure
- Σ : the set of subsets

Samples Space (样本空间), Event:

- 样本空间: 随机试验中所有基本事件所构成的集合, Ω 或 S 表示
- 例 1.2.1. 掷一枚骰子观察出现的点数则 $\Omega = 1, 2, 3, 4, 5, 6$
- 把样本空间中的基本事件与空间中的点相对应, 则事件与集合相对应, 因此事件运算与集合运算可以建立一一对应关系

§ 1.2 Algebra

Algebra:

- Commutative Law
- $A \cup B = B \cup A, A \cap B = B \cap A.$
- Associative Law
- $(A \cup B) \cup C = A \cup (B \cup C), (A \cap B) \cap C = A \cap (B \cap C).$
- $a + b = b + a, a \times b = b \times a$
- $(a + b) + c = a + (b + c), (a \times b) \times c = a \times (b \times c).$
- Distributive laws
- $(A \cup B) \cap C = (A \cap C) \cup (B \cap C);$
- $(A \cap B) \cup C = (A \cup C) \cap (B \cup C).$

De Morgan's Law:

- $\neg(A \wedge B) = \neg A \vee \neg B$
- $\neg(A \vee B) = \neg A \wedge \neg B$

De Morgan's Law (Generalized):

- $\neg(\bigwedge_{i=1}^n A_i) = \bigvee_{i=1}^n \neg A_i$
- $\neg(\bigvee_{i=1}^n A_i) = \bigwedge_{i=1}^n \neg A_i$

More:

- Difference. The set $A \setminus B$ is the set of points that belong to A and (but) not to B. In symbols:
-

$$A \setminus B = A \cap B^c = \{\omega \mid \omega \in A \text{ and } \omega \notin B\}$$

§ 1.3 Definition of Probability

Probability:

- 古典概型：有两个条件，
 1. (有限性) 试验结果只有有限个 (记为 n),
 2. (等可能性) 每个基本事件发生的可能性相同。
- 为计算事件 A 的概率，设 A 中包含 m 个基本事件，则定义事件 A 的概率为

$$P(A) = \frac{m}{n}$$

- 为方便起见，以 $\#(B)$ 记事件 B 中基本事件的个数，因此，

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

Counting:

- 排列 (Permutation): 从 n 个相异的物件中选 r ($1 \leq r \leq n$) 的不同排列数目

$$P(n, r) = \frac{n!}{(n-r)!}$$

- 组合 (Combination): 从 n 个相异的物件中选 r ($1 \leq r \leq n$) 的不同组合数目

$$C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

- 分组: n 个相异的物件分成 k ($1 \leq k \leq n$) 堆，每堆物件个数为 r

$$\text{分组数目} = \frac{n!}{(r!)^k}$$

概率的统计定义:

- 古典概型的两个条件往往不能满足，此时如何定义概率？

常用的一种方法是把含有事件 A 的随机试验独立重复做 n 次 (Bernoulli 试验)，设事件 A 发生了 n_A 次，称比值

$$\frac{n_A}{n}$$

为事件 A 发生的频率。当 n 越来越大时，频率会在某个值 $P(A)$ 附近波动，且波动越来越小，这个值 $P(A)$ 就定义为事件 A 的概率。

- 为什么不能写为

$$\lim_{n \rightarrow \infty} \frac{n_A}{n} = P(A)$$

主观概率:

• 关于概率的统计定义，我们可能会想到，如果试验不能在相同的条件下独立重复很多次时该怎么办？还有人们常谈论种种事件出现机会的大小，如某人有 80% 的可能性办成某事 • 如某人有 80% 的可能性办成某事拐一人则认为仅有 50% 的可能性。即我们常常会拿一个数字去估计这类事件发生的可能性，而心目中并不把它与频率挂钩，这种概率称为主观概率，这类概率有相当的生活基础。在金融和管理等方面有大量的应用，这一学派称为 Bayes 学派，近来得到越来越多的认可。但是当前用频率来定义概率的频率派仍是数理统计的主流，焦点是频率派认为概率是客观存在，不可能因人而异。

概率的公理化定义:

- 对概率运算规定一些简单的基本法则:

1. 设 A 是随机事件, 则 $0 \leq P(A) \leq 1$
2. 设 B 为必然事件, 则 $P(B) = 1$
3. 若事件 A 和 B 不相容, 则 $P(A \cup B) = P(A) + P(B)$
4. 为了对可数无穷个事件仍能成立, 我们要把上面公式中的两个事件推广到可数无穷个两两不相容的事件序列 $\{A_i\}_{i=1}^{\infty}$, 则

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Classical Enunciation of Probability:

- The probability of an event is the ratio of the number of cases favorable to that event to the total number of cases, provided that all these are equally likely.

- To translate this into our language: the sample space is a finite set of possible cases: $\{\omega_1, \omega_2, \dots, \omega_m\}$, each ω_i being a “case.” An event A is a subset $\{\omega_{i1}, \omega_{i2}, \dots, \omega_{in}\}$, each ω_{ij} being a “favorable case.” The probability of A is then the ratio

$$P(A) = \frac{|A|}{|\Omega|} = \frac{n}{m}$$

where $|A|$ is the number of favorable cases in event A , and $|\Omega|$ is the total number of cases in the sample space.

Definition of Probability:

- Definition. A probability measure on the sample space Ω is a function of subsets of Ω satisfying three axioms:

1. For every set $A \subset \Omega$, the value of the function is a nonnegative number: $P(A) \geq 0$.
2. For any two disjoint sets A and B , the value of the function for their union $A \cup B$ is equal to the sum of its value for A and its value for B : $P(A \cup B) = P(A) + P(B)$ provided $A \cap B = \emptyset$.
3. The value of the function for Ω (as a subset) is equal to 1: $P(\Omega) = 1$.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{n}{m}$$

§ 1.4 Deductions from the Axioms**Deductions from the Axioms:**

- For any set A , we have $P(A) \leq 1$.
- Proof:

$$A + A^c = \Omega$$

$$P(A) + P(A^c) = P(\Omega) = 1$$

$$P(A) = 1 - P(A^c) \leq 1$$

- For any two sets such that $A \subset B$, we have $P(A) \leq P(B)$, and $P(B - A) = P(B) - P(A)$.
- Proof:

$$B = A + (B - A)$$

$$P(B) = P(A) + P(B - A) \geq P(A)$$

$$\bullet P(A \cup B) \leq P(A) + P(B)$$

$$A \cup B = A + A^c B$$

$$P(A \cup B) = P(A) + P(A^c B)$$

General but Finite Cases:

- For any finite number of disjoint sets A_1, \dots, A_n , we have

$$P(A_1 + \dots + A_n) = P(A_1) + \dots + P(A_n)$$

- For any finite number of arbitrary sets A_1, \dots, A_n , we have

$$P(A_1 \cup \dots \cup A_n) \leq P(A_1) + \dots + P(A_n)$$

Proof: induction on n .

§ 1.5 Conditional Probability

Conditional Probability:

- 定义: 设事件 A 和 B 是随机试验 Ω 中的两个事件, 且 $P(B) > 0$, 称条件概率 $P(A | B)$ 为

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

事件 B 发生条件下事件 A 发生的条件概率

Multiplication Rule:

- 乘法公式: 设 A 和 B 是随机试验 Ω 中的两个事件, 且 $P(A) > 0$, 则 $P(A \cap B) = P(A)P(B | A)$
- 证明: $P(A \cap B) = P(A)P(B | A)$

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

- For a sequence of events A_1, A_2, \dots, A_n , we have

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

§ 1.6 Total probability and Bayes' Theorem

Total Probability:

• Proposition 2. Suppose that $\{A_1, A_2, \dots, A_n\}$ is a partition of the sample space into disjoint sets. Then for any set B we have

$$P(B) = \sum_{i=1}^n P(B | A_i)P(A_i)$$

- Proof:

$$B = B \cap \Omega = B \cap (A_1 \cup A_2 \cup \cdots \cup A_n) = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n) = P\left(\sum_{i=1}^n B A_i\right)$$

Bayes' Theorem:

• Theorem 1. Suppose that $\{A_1, A_2, \dots, A_n\}$ is a partition of the sample space into disjoint sets. Then for any set B such that $P(B) > 0$, we have

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^n P(B | A_j)P(A_j)}$$

• Proof:

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

$$P(B) = \sum_{j=1}^n P(B | A_j)P(A_j)$$

Bayesian Calculations:

a test positive

$$\frac{P(D | A)}{P(D^c | A)} = \frac{P(D)}{P(D^c)} \frac{P(A | D)}{P(A | D^c)}$$

§ 1.7 Independence

Independence:

• 定义: 设 A 和 B 是随机试验中的两个事件, 若满足

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

则称事件 A 和 B 相互独立。

• 定义: 设 A_1, A_2, \dots, A_n 是随机试验中的 n 个事件, 若对于任意的子集 $\{i_1, i_2, \dots, i_k\} \subset \{1, 2, \dots, n\}$, 满足

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

则称事件 A_1, A_2, \dots, A_n 相互独立。

上面有 $2^n - n - 1$ 个独立性条件。

上面有 2^n 个等式, 减去 n 个 $P(A_i) = P(A_i)$ 和一个 $P(\Omega) = 1$ 。

第2章 Random Variables and their Distributions

§ 2.1 Random Variables

Random Variables:

• 随机变量: 设 Ω 是随机试验的样本空间, 如果对于每一个实数 x , 事件 $\{\omega \mid X(\omega) \leq x\}$ 是 Ω 的事件, 则称 $X(\omega)$ 为随机变量。

- 随机变量的取值范围称为随机变量的值域, 记为 R_X 。
- 随机变量的分布函数 $F(x)$ 定义为

$$F(x) = P(X \leq x)$$

Discrete Random Variables:

- 离散型随机变量: 若随机变量 X 的值域 R_X 是有限集或可数集, 则称 X 为离散型随机变量。
- 离散型随机变量的分布函数 $F(x)$ 为

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

The Indicator of An Event:

- Indicator function: For an event A , the indicator function I_A is defined as

$$I_A(\omega) = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{if } A \text{ does not occur} \end{cases}$$

$$I_{A \cap B}(\omega) = I_A(\omega) \wedge I_B(\omega) = I_A(\omega) \cdot I_B(\omega)$$

$$I_{A \cup B}(\omega) = I_A(\omega) \vee I_B(\omega)$$

where $a \vee b$ is the maximum of a and b , and $a \wedge b$ is the minimum of a and b .

$I_A(\omega)$ is a random variable.

$I_A(\omega)$ is a Bernoulli random variable.

Bernoulli Trial:

- 定义: 设一个随机试验只有两个可能结果 A 和 A^c , 则称此试验为一 Bernoulli 试验。
- 定义: 设将一个可能结果为 A 和 A^c 的 Bernoulli 试验独立地重复 n 次, 使得事件且每次出现的概率相同, 则称此试验为 n 重 Bernoulli 试验。
- Bernoulli 试验: 一次随机试验称为 Bernoulli 试验, 如果试验只有两个可能结果, 记为 A 和 A^c , 且 $P(A) = p$, $P(A^c) = 1 - p$, 且两次试验之间相互独立。

$$P(A) = p, \quad P(A^c) = 1 - p$$

- Bernoulli 试验的特点: 二项分布、几何分布、超几何分布、泊松分布等。

• 0-1 分布: 设随机变量 X 只取 0 和 1 两值, $P(X = 1) = p$, $P(X = 0) = 1 - p$, 则称 X 服从 0-1 分布或 Bernoulli 分布, 记为 $X \sim B(1, p)$ 。

• 二项分布: 设某事件 A 在一次试验中发生的概率为 p , 现把试验独立地重复 n 次。以 X 记 A 在这次试验中发生的次数, 则 X 取值 $0, 1, \dots, n$, 且有

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

称 X 服从二项分布, 记为 $X \sim \text{Bin}(n, p)$ 或 $B(n, p)$ 。

二项分布: 构造性归纳法:

Assume it is correct for $k-1$. Then

$$P(X = k) = P(X = k-1) \frac{n-k+1}{k} \frac{p}{1-p}$$

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Random number and pseudo-random number:

- 随机数: 一般指在一定范围内的均匀分布的随机数。
- 伪随机数: 由计算机生成的数列, 其数值是按照一定规律生成的, 但是在一定范围内表现出随机性。
- 伪随机数的特点: 1. 周期性; 2. 重现性; 3. 均匀性; 4. 独立性。
- 伪随机数的生成方法: 1. 线性同余法; 2. 递推法; 3. 逆变换法; 4. 拒绝法; 5. 接受-拒绝法; 6. 分布抽样法; 7. 蒙特卡洛法。

Poisson Distribution:

- 泊松分布: 设随机变量 X 取值为 $0, 1, 2, \dots$, 且有

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

称 X 服从参数为 λ 的泊松分布, 记为 $X \sim \text{Poisson}(\lambda)$ 。

- 泊松分布的性质:

1. $E(X) = \lambda$;
2. $D(X) = \lambda$;
3. 泊松分布的和仍然服从泊松分布;
4. 泊松分布的和服从泊松分布。

Remember the Taylor expansion of e^x :

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Derivation of Poisson Distribution from "Introduction to Statistics" by Mood and Graybill:

• 假定体积为 V 的液体包含有一个大数目 N 的微生物, 再假定微生物没有群居的本能, 它们能够在液体的任何部分出现, 且在体积相等的部分出现的机会相同。现在我们取体积为 D 的微量液体在显微镜下观察, 问在这微量液体中将发现 x 个微生物的概率是什么? 我们假定 V 远远大于 D , 由于假定了这些微生物是以一致的概率在液体中到处散布, 因此任何一个微生物在 D 中出现的概率都是 D/V 。再由于假定了微生物没有群居的本能, 所以一个微生物在 D 中的出现, 不会影响另一个微生物在 D 中的出现与否。因此微生物中有 x 个在 D 中出现的概率就是

$$\binom{N}{x} \left(\frac{D}{V}\right)^x \left(1 - \frac{D}{V}\right)^{N-x}$$

- 令 V 和 N 趋向于无穷, 且微生物的密度 $N/V=d$ 保持常数 • 将上式改写成如下形式:

$$\lim_{N \rightarrow \infty, V \rightarrow \infty} \binom{N}{x} \left(\frac{d}{V}\right)^x \left(1 - \frac{d}{V}\right)^{N-x} = \frac{N(N-1)(N-2) \dots (N-x+1)}{x! N^x} \left(\frac{ND}{V}\right)^x \left(1 - \frac{ND}{NV}\right)^{N-x}$$

- 当 N 变成无限时其极限为

$$\lim_{N \rightarrow \infty, V \rightarrow \infty} \binom{N}{x} \left(\frac{d}{V}\right)^x \left(1 - \frac{d}{V}\right)^{N-x} = \frac{(Dd)^x}{x!} e^{-Dd}$$

• 这一推导过程还证明了 λ 是 x 的平均数，因为所考察的一部分体积 D 乘以整个的密度 d 就给出了在 D 中所预计的平均数目。

Poisson distribution as a limit of Binomial Distribution:

• 当 N 很大且 p 很小且趋于一个极限时，Poisson 分布是二项分布的一个很好的近似。而在 N 未知时，Poisson 分布更显得有用。

我们有下面的定理：

• 定理：在 n 重 Bernoulli 试验中，以 p 为成功概率，若 $n \rightarrow \infty$ 且 $np \rightarrow \lambda$ 保持常数，则二项分布 $B(n, p)$ 的极限为 Poisson 分布 $P(\lambda)$ 。具体来说，

$$\lim_{n \rightarrow \infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

其中 $\lambda = np$ 。

Geometric Random Variable:

• Suppose that independent trials, each having a probability p , $0 < p < 1$, of being a success, are performed until a success occurs. If we let X equal the number of trials required, then

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$

• Generalization: Negative binomial distribution.

Geometric Random Variable:

• Suppose that independent trials, each having a probability p , $0 < p < 1$, of being a success, are performed until a success occurs. If we let X equal the number of trials required, then

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, \dots$$

• Generalization: Negative binomial distribution.

The Hypergeometric Random Variable:

• Suppose that a sample of size n is to be chosen randomly (without replacement) from an urn containing N balls, of which m are white and $N - m$ are black. If we let X denote the number of white balls selected, then

$$P(X = i) = \frac{\binom{m}{i} \binom{N-m}{n-i}}{\binom{N}{n}}, \quad i = 0, 1, 2, \dots, \min(m, n)$$

• Sampling from a finite population:

• Suppose that a sample of size n is to be chosen randomly from an urn containing N balls, of which m are white and $N - m$ are black. If we let X denote the number of white balls selected:

– Sample without replacement: hypergeometric distribution

– Sample with replacement: binomial $X \sim \text{Binomial}(n, \frac{m}{N})$

– Sample without replacement: $N \rightarrow \infty, \frac{m}{N} \rightarrow p$, binomial $X \sim \text{Binomial}(n, p)$

§ 2.2 Continuous Random Variables

连续型随机变量:

- 离散随机变量只取有限个或可数无限个值
- 而连续型随机变量取不可数个值, 这就决定了不能用描述离散型随机变量的办法来刻画连续型随机变量。

Continuous Random Variables:

- 连续型随机变量: 若随机变量 X 的值域 R_X 是连续集, 则称 X 为连续型随机变量。
- 连续型随机变量的分布函数 $F(x)$ 为

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

- 连续型随机变量的密度函数 $f(x)$ 为

$$f(x) = \frac{dF(x)}{dx}$$

概率密度函数:

- X 为连续型随机变量, 如果存在一个函数 f , 叫做 X 的概率密度函数, 它满足下面的条件:
- 对所有的 $-\infty < x < \infty$, $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x) dx = 1$
- $-\infty < a < b < \infty$, $P(a \leq X \leq b) = \int_a^b f(x) dx$
- Note: $P(X = x) = 0 = \int_x^x f(u) du$

分布函数 (Distribution Function):

- X 为一连续型随机变量, 则

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

是 X 的分布函数, 且

$$f(x) = \frac{dF(x)}{dx}$$

是 X 的概率密度函数。

- 分布函数 $F(x)$ 具有下列性质:
- $0 \leq F(x) \leq 1$
- $F(x)$ 是非减函数
- $F(x)$ 是右连续的
- $F(x)$ 在正无穷处趋于 1, 在负无穷处趋于 0
- $P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$

正态分布 (Normal distribution):

- 如果一个随机变量 x 具有概率密度函数

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

其中 $-\infty < x < \infty, \sigma > 0, -\infty < \mu < \infty$, 则称 X 服从参数为 μ 和 σ^2 的正态分布, 记为 $X \sim N(\mu, \sigma^2)$ 。

- 正态分布的性质:

1. $E(X) = \mu$;
2. $D(X) = \sigma^2$;
3. 正态分布的和仍然服从正态分布;
4. 正态分布的和服从正态分布。

- 正态分布的标准化:

1. 若 $X \sim N(\mu, \sigma^2)$, 则 $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$ 。
2. 若 $X \sim N(0, 1)$, 则 $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$ 。

- 正态分布的性质:

1. 若 $X \sim N(\mu, \sigma^2)$, 则 $aX + b \sim N(a\mu + b, a^2\sigma^2)$ 。
2. 若 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, 且 X 与 Y 相互独立, 则 $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 。
3. 若 $X \sim N(\mu, \sigma^2)$, 则 $X^2 \sim \chi^2(1)$ 。

• 具有参数 $\mu = 0, \sigma = 1$ 的正态分布称为标准正态分布。用 $\phi(x)$ 和 $\Phi(x)$ 表示标准正态分布 $N(0, 1)$ 的密度函数和分布函数。

- 以 $F(x)$ 记正态分布 $N(\mu, \sigma^2)$ 的概率分布函数, 则恒有

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

所以任一正态分布的概率分布函数都可通过标准正态分布的分布函数计算出来。

density function: $\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$

- 中心极限定理 (Central limit theorem):

当 n 足够大时, 独立同分布的随机变量的和近似服从正态分布。

If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

Uniform distribution:

- If a random variable X has a density function

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

$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x \leq b \\ 1, & x > b \end{cases}$$

then X is said to have a uniform distribution on the interval (a, b) , denoted by $X \sim U(a, b)$.

- The mean and variance of $X \sim U(a, b)$ are

$$E(X) = \frac{a+b}{2}, \quad D(X) = \frac{(b-a)^2}{12}$$

指数分布 (Exponential distribution):

- If a random variable X has a density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

then X is said to have an exponential distribution with parameter λ , denoted by $X \sim \text{Exp}(\lambda)$ (其中 $\lambda > 0$ 为常数, 则称 X 服从参数为 λ 的指数分布。)

$$E(X) = \frac{1}{\lambda}, \quad D(X) = \frac{1}{\lambda^2}$$

指数分布的分布函数为

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Memoryless distribution:

- A continuous random variable X is said to be memoryless if

$$P(X > s + t \mid X > s) = P(X > t)$$

for all $s, t > 0$.

- Exponential distribution is memoryless.

多维分布:

- 把多个随机变量放在一起组成向量, 称为多维随机变量或者随机向量。
- Example: 从一副扑克牌中抽牌时, 可以用纸牌的花色和数字来说明其特征。
- Example: 考虑一个打靶的试验, 在靶面上取定一个直角坐标系。则命中的位置可由其坐标 X, Y 来刻画。 X 和 Y 都是随机变量。
- 定义: 设 $X = (X_1, \dots, X_n)$, 如果每个 X_i 都是一个随机变量, 称 X 为 n 维随机变量或者随机向量。

Representation of Dice by Indicators:

- Example: Let X be the sum of the numbers on two dice. We can represent the sum as

$$X = I_1 + I_2 + \dots + I_6$$

where I_i is the indicator of the event that the number on the i -th die is face up.

Marginal distribution (边缘分布):

• Definition: Let X and Y be two random variables with joint density function $f(x, y)$. The marginal density functions of X and Y are

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

- Example: Let X and Y be two random variables with joint density function

$$f(x, y) = \begin{cases} 2, & 0 < x < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Then the marginal density functions of X and Y are

$$f_X(x) = \int_x^1 2 dy = 2(1-x)$$

$$f_Y(y) = \int_0^y 2 dx = 2y$$

- $X = (X_1, \dots, X_n)$ 为 n 维离散随机变量, 有概率分布 $F(X)$ 。任意一个子集的分布是 m 维边缘分布。
- 二维离散随机变量 (X, Y) 的边缘概率质量函数为

$$P_X(x_i) = P(X = x_i) = \sum_{j=1}^m P(X = x_i, Y = y_j) = \sum_{j=1}^m p_{ij} = p_i$$

$$P_Y(y_j) = P(Y = y_j) = \sum_{i=1}^n P(X = x_i, Y = y_j) = \sum_{i=1}^n p_{ij} = p_j$$

Conditional Distribution:

一个随机变量 (或向量) 的条件概率分布, 就是在给定 (或已知) 某种条件 (某种信息) 下该随机变量 (向量) 的概率分布

离散型随机变量的条件分布

设 X 和 Y 是两个离散型随机变量, 它们的联合分布律为

$$P(X = x_i, Y = y_j) = p_{ij}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m$$

则在 $Y = y_j$ 的条件下, X 的条件分布律为

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_j}$$

$$P(Y = y_j | X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_i}$$

多维连续型随机向量: 概率密度函数:

- n 维连续型随机向量 $X = (X_1, X_2, \dots, X_n)$ 的概率密度函数 $f(x_1, x_2, \dots, x_n)$ 是满足下列条件的函数:

- 对所有的 (x_1, x_2, \dots, x_n) , $f(x_1, x_2, \dots, x_n) \geq 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n = 1$
- 对任意的 n 维区域 G , 有

$$P((X_1, X_2, \dots, X_n) \in G) = \int \int \dots \int_G f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

- 连续型随机向量 $X = (X_1, \dots, X_n)$, 如果存在 \mathbb{R}^n 上的非负函数 $f(x_1, x_2, \dots, x_n)$, 使得对任意的 $-\infty < a_1 \leq b_1 < \infty, \dots, -\infty < a_n \leq b_n < \infty$,

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2, \dots, a_n \leq X_n \leq b_n) = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \dots \int_{a_1}^{b_1} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

则称 f 为 X 的概率密度函数, X 为连续型随机向量。

多维连续型随机向量: 概率分布函数:

- n 维连续型随机向量 $X = (X_1, X_2, \dots, X_n)$ 的概率分布函数 $F(x_1, x_2, \dots, x_n)$ 是满足下列条件的函数:

- 对所有的 (x_1, x_2, \dots, x_n) , $0 \leq F(x_1, x_2, \dots, x_n) \leq 1$
- 对任意的 n 维区域 G , 有

$$P((X_1, X_2, \dots, X_n) \in G) = \int \int \dots \int_G f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

- 概率分布函数: $F(x) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$
- 1. $F(x_1, x_2, \dots, x_n)$ 对每个单元单调非降
- 2. 对任意的 $1 \leq j \leq n$,

$$\lim_{x_j \rightarrow -\infty} F(x_1, x_2, \dots, x_n) = 0,$$

- 3.

$$\lim_{x_1 \rightarrow -\infty, x_2 \rightarrow -\infty, \dots, x_n \rightarrow -\infty} F(x_1, x_2, \dots, x_n) = 0, \quad \lim_{x_1 \rightarrow +\infty, x_2 \rightarrow +\infty, \dots, x_n \rightarrow +\infty} F(x_1, x_2, \dots, x_n) = 1$$

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_n} \int_{-\infty}^{x_{n-1}} \dots \int_{-\infty}^{x_1} f(t_1, t_2, \dots, t_n) dt_1 dt_2 \dots dt_n$$

2-D Uniform distribution:

- $X = (X_1, X_2)$ 的概率密度函数

$$f(x_1, x_2) = \begin{cases} \frac{1}{S}, & (x_1, x_2) \in D \\ 0, & \text{otherwise} \end{cases}$$

$$S = \text{Area}(D) = \iint_D dx_1 dx_2$$

- 二维均匀分布的边缘分布:

1. $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 = \frac{1}{S} \int_{-\infty}^{\infty} dx_2 = \frac{1}{S}$
2. $f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 = \frac{1}{S} \int_{-\infty}^{\infty} dx_1 = \frac{1}{S}$

- 二维均匀分布的条件分布:

1. $f_{X_1|X_2}(x_1 | x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)} = \frac{1/S}{1/S} = 1$
2. $f_{X_2|X_1}(x_2 | x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)} = \frac{1/S}{1/S} = 1$

Bivariate normal distribution:

- If a random vector $X = (X_1, X_2)$ has a joint density function

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right] \right\}$$

then X is said to have a bivariate normal distribution with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$

$$(X_1, X_2) \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right)$$

- The marginal density functions of X_1 and X_2 are

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left\{ -\frac{(x_1 - \mu_1)^2}{2\sigma_1^2} \right\}$$

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp \left\{ -\frac{(x_2 - \mu_2)^2}{2\sigma_2^2} \right\}$$

- The conditional density functions of X_1 and X_2 given $X_2 = x_2$ and $X_1 = x_1$ are

$$f_{X_1|X_2}(x_1 | x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)}$$

$$f_{X_2|X_1}(x_2 | x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}$$

$$X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$$

Bivariate normal distribution

$$(X_1, X_2) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}$$

Where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $|\Sigma|$ 是 Σ 的行列式. In details,

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right] \right\}}$$

图 2.1: Bivariate normal distribution

二维连续型随机向量：Marginal distribution

设 (X, Y) 的概率密度函数 $f(x, y)$

- $P(x_1 \leq X \leq x_2) = P(x_1 \leq X \leq x_2, -\infty < Y < \infty) =$
 $\iint_{x_1 \leq X \leq x_2, -\infty < Y < \infty} f(u, v) \, du \, dv = (\text{Fubini Theorem})$
 $= \int_{x_1}^{x_2} \left[\int_{-\infty}^{\infty} f(u, v) \, dv \right] du = \int_{x_1}^{x_2} f_X(u) \, du$
- $f_X(u) = \int_{-\infty}^{\infty} f(u, v) \, dv$

图 2.2: Marginal Distribution1

多维连续型随机向量：Marginal distribution

- Generally, 连续型随机向量 $X = (X_1, X_2 \cdots, X_n)$, 如果存在 \mathbb{R}^n 上的非负函数 $f(x_1, x_2 \cdots, x_n)$, $(X_{i_1}, X_{i_2} \cdots X_{i_m})$ 是任意一个子集, 它的分布是 m 维边缘分布
- $f(x_{i_1}, x_{i_2} \cdots x_{i_m}) = \int \cdots \int f(x_1, x_2 \cdots, x_n) dx$
 其中积分是对除 $(X_{i_1}, X_{i_2} \cdots X_{i_m})$ 之外的所有变量来求积

图 2.3: Marginal Distribution2

多维连续型随机向量： conditional distribution

- 先考虑二维的情形. 设 (X, Y) 的概率密度函数 $f(x, y)$, consider the conditional distribution of X given $y \leq Y \leq y + \epsilon$,

$$P(X \leq x | y \leq Y \leq y + \epsilon) = \frac{P(X \leq x, y \leq Y \leq y + \epsilon)}{P(y \leq Y \leq y + \epsilon)}$$

$$= \frac{\int_{-\infty}^x \int_y^{y+\epsilon} f(u, v) du dv}{\int_y^{y+\epsilon} f_Y(v) dv} = \int_{-\infty}^x \frac{\int_y^{y+\epsilon} f(u, v) dv}{\int_y^{y+\epsilon} f_Y(v) dv} du$$

Take derivative w.r.t. x , and let $\epsilon \rightarrow 0$, $f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$

This is the conditional density function.

图 2.4: conditional distribution1

Bivariate normal distribution

$$(X_1, X_2) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} e^{\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}}$$

Conditional distribution:

$$X_1 | X_2 = x_2 \sim N(\mu_1 + \rho\sigma_1\sigma_2^{-1}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)),$$

图 2.5: conditional distribution2

多维连续型随机向量：条件密度函数（一般情形）

- $(X_1, X_2 \dots, X_n) \sim f(x_1, x_2 \dots, x_n)$, density function

- 记 $X = (X_1, X_2 \dots, X_k)$, $Y = (X_{k+1}, X_{k+2} \dots, X_n)$

$$(X, Y) \sim f(x, y),$$

where $x = (x_1, x_2 \dots, x_k)$, $y = (x_{k+1}, x_{k+2} \dots, x_n)$

$$\begin{aligned} f_{X|Y}(x|y) &= f_{X|Y}(x_{k+1}, x_{k+2} \dots, x_n | x_1, x_2 \dots, x_k) = \frac{f(x, y)}{f_Y(y)} \\ &= \frac{f(x_1, x_2 \dots, x_n)}{f_Y(x_{k+1}, x_{k+2} \dots, x_n)} \end{aligned}$$

Where $f_Y(x_{k+1}, x_{k+2} \dots, x_n) = \int \dots \int f(x_1, x_2 \dots, x_n) dx_1, dx_2 \dots, dx_k$

图 2.6: conditional distribution3

§ 2.3 随机变量的独立性

随机变量的独立性:

- 离散型随机变量 X_1, \dots, X_n 相互独立, 若它们的联合分布律等于各自的边缘分布律的乘积, 即

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

- 连续型随机变量 X_1, \dots, X_n 相互独立, 若它们的联合密度等于各自的边缘密度的乘积, 即

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdot \dots \cdot f_{X_n}(x_n)$$

- 最一般的情况, 随机变量 X_1, \dots, X_n 相互独立, 若它们的联合分布函数等于各自边缘分布函数的乘积, 即

$$F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \dots \cdot F_{X_n}(x_n)$$

对于二维正态分布 (Bivariate normal distribution)

X_1 和 X_2 相互独立, 当且仅当 $\rho = 0$

边缘分布: $X_1 \sim N(\mu_1, \sigma_1^2)$

条件分布: $X_1 | X_2 = x_2 \sim N\left(\mu_1 + \rho \frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right)$

§ 2.4 随机变量的函数的概率分布

随机变量的函数的概率分布:

- 设 X 是一个随机变量, $Y = g(X)$ 是 X 的函数, $g(x)$ 是连续函数, X 的概率密度函数为 $f_X(x)$, 则 Y 的概率密度函数为

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

其中 $x = g^{-1}(y)$ 。

- Example:** 设 $X \sim U(0, 1)$, 求 $Y = -\ln X$ 的概率密度函数。

$$1. Y = -\ln X \Rightarrow X = e^{-Y}$$

$$2. f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = 1 \cdot e^{-y} = e^{-y}$$

$$3. Y \sim \text{Exp}(1)$$

- Example:** 设 $X \sim N(0, 1)$, 求 $Y = X^2$ 的概率密度函数。

$$1. Y = X^2 \Rightarrow X = \pm\sqrt{y}$$

$$2. f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right| = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left| \frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \left| -\frac{1}{2\sqrt{y}} \right| = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$$

$$3. Y \sim \chi^2(1)$$

Random vector case (discrete):

- 特别当 ξ 和 η 是相互独立的非负整值随机变量, 各有分布律:

$$P(\xi = k) = a_k, \quad k = 0, 1, 2, \dots$$

$$P(\eta = l) = b_l, \quad l = 0, 1, 2, \dots$$

则它们的联合分布律为:

$$P(\xi = k, \eta = l) = P(\xi = k) \cdot P(\eta = l) = a_k \cdot b_l$$

if $\xi + \eta = n$ then

$$P(\xi + \eta = n) = \sum_{k=0}^n a_k \cdot b_{n-k}$$

称此公式为离散卷积公式。

Example:

设 $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, 且 X 与 Y 相互独立, 则 $X + Y \sim \text{Bin}(n + m, p)$ 。

这种性质称为再生性。Proof:

1. 离散卷积公式:

$$\begin{aligned} P(X + Y = k) &= \sum_{i=0}^k P(X = i)P(Y = k - i) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \binom{m}{k-i} p^{k-i} (1-p)^{m-(k-i)} \\ &= \sum_{i=0}^k \binom{n}{i} \binom{m}{k-i} p^k (1-p)^{n+m-k} \\ &= \binom{n+m}{k} p^k (1-p)^{n+m-k} \end{aligned}$$

2. 由此可知 $X + Y \sim \text{Bin}(n + m, p)$ 。

3. 概率推理, 构造独立同分布的 $Z_i \sim \text{Bernoulli}(p)$, 则 $\tilde{X} = \sum_{i=1}^n Z_i$, $\tilde{Y} = \sum_{i=1}^m Z_{n+i}$, 那么 $\tilde{X} + \tilde{Y} = \sum_{i=1}^{n+m} Z_i \sim \text{Bin}(n + m, p)$ 。

\tilde{X} 和 \tilde{Y} 的分布与 X 和 Y 相同, 所以 $X + Y$ 与 $\tilde{X} + \tilde{Y}$ 的分布相同。

Generalization:

- 可推广至多项和, 设 $X_i \sim \text{Bin}(n_i, p)$, $i = 1, 2, \dots, m$, 且 X_1, X_2, \dots, X_m 相互独立, 则 $\sum_{i=1}^m X_i \sim \text{Bin}(\sum_{i=1}^m n_i, p)$ 。

- 可推广至泊松分布, 设 $X_i \sim P(\lambda_i)$, $i = 1, 2, \dots, m$, 且 X_1, X_2, \dots, X_m 相互独立, 则 $\sum_{i=1}^m X_i \sim P(\sum_{i=1}^m \lambda_i)$ 。

- 可推广至正态分布, 设 $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, m$, 且 X_1, X_2, \dots, X_m 相互独立, 则 $\sum_{i=1}^m X_i \sim N(\sum_{i=1}^m \mu_i, \sum_{i=1}^m \sigma_i^2)$ 。

特别地, 若 X_1, \dots, X_n 为独立同分布, 且 $X_i \sim \text{Bin}(1, p)$, $i = 1, \dots, m$, 则 $\sum_{i=1}^m X_i \sim \text{Bin}(m, p)$, 此结论揭示了二项分布与 0-1 分布之间的密切关系。

连续型随机变量密度变换公式:

- 定理 (密度变换公式) 设随机变量 X 有概率密度函数 $f(x)$, $x \in (a, b)$ (a, b 可以是 ∞), 而 $y = g(x)$ 在 $x \in (a, b)$ 上是严格单调的连续函数, 存在唯一的反函数 $x = h(y)$, $y \in (\alpha, \beta)$ 并且 $h'(y)$ 存在且连续, 那么 $Y = g(X)$ 也是连续型随机变量且有概率密度函数

$$p(y) = f(h(y)) |h'(y)|, \quad y \in (\alpha, \beta)$$

注意: $f(x)dx = f(h(y))d(h(y)) = f(h(y)) |h'(y)| dy$

注意: $P(x < X \leq x + \delta x) \approx f(x)\delta x$, $P(y < Y \leq y + \delta y) \approx p(y)\delta y$

- A Generalization

- Note: 当 g 不是在全区间上单调而是逐段单调时, 密度变换公式为下面的形式: • 设随机变量 X 的密度函数为 $p_X(x)$, $x \in (a, b)$, 如果可以把 (a, b) 分割为一些 (有限个或可列个) 互不重叠的子区间的和

$(a, b) = \bigcup_j I_j$, 使得函数 $y = g(x)$ 在每个子区间上有唯一的反函数 $x = h_j(y)$, 并且 $h'_j(y)$ 存在且连续, 那么 $Y = g(x)$ 也是连续型随机变量且有概率密度函数

$$p_Y(y) = \sum_j p_X(h_j(y)) |h'_j(y)|, \quad y \in (\alpha, \beta)$$

§ 2.5 Vector Case

Vector Case: $n = 2$:

• 设 (X_1, X_2) 是二维连续型随机向量, 具有联合密度函数 $p(x_1, x_2)$, 设 $Y_j = f_j(X_1, X_2)$, $j = 1, 2$ 。若 (X_1, X_2) 和 (Y_1, Y_2) 一一对应, 逆映射 $X_j = h_j(Y_1, Y_2)$, $j = 1, 2$ 。

假定每个 $h_j(Y_1, Y_2)$ 有一阶连续偏导数, 则 (Y_1, Y_2) 也是连续型随机向量, 且有联合密度函数

$$q(y_1, y_2) = \begin{cases} p(h_1(y_1, y_2), h_2(y_1, y_2)) |J|, & (y_1, y_2) \in D \\ 0, & \text{otherwise} \end{cases}$$

其中 D 为随机向量 (X_1, X_2) 所有可能取值的集合。 J 为雅可比行列式, 即

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} \end{vmatrix}$$

General Case:

• 设 (X_1, X_2, \dots, X_n) 是 n 维连续型随机向量, 具有联合密度函数 $p(x_1, x_2, \dots, x_n)$, 设 $Y_j = f_j(X_1, X_2, \dots, X_n)$, $j = 1, 2, \dots, m$ 。若 (X_1, X_2, \dots, X_n) 和 (Y_1, Y_2, \dots, Y_m) 一一对应, 逆映射 $X_j = h_j(Y_1, Y_2, \dots, Y_m)$, $j = 1, 2, \dots, n$ 。

假定每个 $h_j(Y_1, Y_2, \dots, Y_m)$ 有一阶连续偏导数, 则 (Y_1, Y_2, \dots, Y_m) 也是连续型随机向量, 且有联合密度函数

$$q(y_1, y_2, \dots, y_m) = \begin{cases} p(h_1(y_1, y_2, \dots, y_m), h_2(y_1, y_2, \dots, y_m), \dots, h_n(y_1, y_2, \dots, y_m)) |J|, & (y_1, y_2, \dots, y_m) \in D \\ 0, & \text{otherwise} \end{cases}$$

其中 D 为随机向量 (X_1, X_2, \dots, X_n) 所有可能取值的集合。 J 为雅可比行列式, 即

$$J = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \frac{\partial h_1}{\partial y_2} & \cdots & \frac{\partial h_1}{\partial y_m} \\ \frac{\partial h_2}{\partial y_1} & \frac{\partial h_2}{\partial y_2} & \cdots & \frac{\partial h_2}{\partial y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \frac{\partial h_n}{\partial y_2} & \cdots & \frac{\partial h_n}{\partial y_m} \end{vmatrix}$$

§ 2.6 两个随机变量

两个随机变量之和的概率分布:

• 设 (X, Y) 的联合概率密度为 $f(x, y)$, 则 $X + Y$ 的概率密度 $p(z)$ 为

$$p(z) = \int_{-\infty}^{\infty} f(x, z-x) dx = \int_{-\infty}^{\infty} f(z-y, y) dy$$

• 例: 设 X, Y 独立, $X \sim N(0, 1)$, $Y \sim N(0, 1)$, 求 $X + Y$ 的概率密度。

$$1. f(x, y) = f(x)f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

$$2. p(z) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{x^2 + (z-x)^2}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

特别，当 X 与 Y 独立时，分别记 X 和 Y 的概率密度为 $f_1(x)$ 和 $f_2(y)$ ，则 $X + Y$ 的概率密度为

$$p(z) = \int_{-\infty}^{\infty} f_1(x)f_2(z-x)dx = \int_{-\infty}^{\infty} f_1(z-y)f_2(y)dy \triangleq f_1 * f_2(z) = f_2 * f_1(z)$$

称此公式为卷积公式。

• 例：设 X 服从期望为 2 的指数分布， $Y \sim U(0,1)$ ，且 X 和 Y 相互独立。求 $X - Y$ 的概率密度和 $P(X \leq Y)$ 。

$$1. f_1(x) = \frac{1}{2}e^{-\frac{x}{2}}, f_2(y) = 1, f_1 * f_2(z) = \int_0^1 \frac{1}{2}e^{-\frac{z+y}{2}} dy = \frac{1}{2}e^{-\frac{z}{2}}(1 - e^{-\frac{1}{2}})$$

$$2. P(X \leq Y) = \int_0^1 \int_0^y \frac{1}{2}e^{-\frac{x}{2}} dx dy = \int_0^1 (1 - e^{-\frac{y}{2}}) dy = 1 - \frac{1}{2}(1 - e^{-1}) = \frac{1}{2}(1 + e^{-1})$$

§ 2.7 正态分布变量的和的再生性、封闭性

正态分布变量的和的再生性、封闭性：

• 设 $X_1 \sim N(\mu_1, \sigma_1^2)$ ， $X_2 \sim N(\mu_2, \sigma_2^2)$ ，且 X_1 与 X_2 相互独立，则 $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ 。

• 推广至 n 个独立正态分布变量之和：设 $X_i \sim N(\mu_i, \sigma_i^2)$ ， $i = 1, 2, \dots, n$ ，且 X_1, X_2, \dots, X_n 相互独立，则 $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ 。

• 推广至 n 个独立正态分布变量之和的封闭性：设 $X_i \sim N(\mu_i, \sigma_i^2)$ ， $i = 1, 2, \dots, n$ ，且 X_1, X_2, \dots, X_n 相互独立，则 $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ 。

• 推广至 n 个独立正态分布变量之和的再生性：设 $X_i \sim N(\mu_i, \sigma_i^2)$ ， $i = 1, 2, \dots, n$ ，且 X_1, X_2, \dots, X_n 相互独立，则 $\sum_{i=1}^n X_i \sim N(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$ 。

Note: if the density of a r.v. $f(x) \propto e^{-cx^2}$, then its distribution is normal.

Generally, $X_i \sim N(\mu_i, \sigma_i^2)$, X_1, \dots, X_n are jointly independent, a_1, \dots, a_n, b are $n+1$ arbitrary real numbers, Let $X = b + \sum_{i=1}^n a_i X_i$, then $X \sim N(\mu, \sigma^2)$, where $\mu = b + \sum_{i=1}^n a_i \mu_i$, $\sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2$

图 2.7: 7

- $X_1 + X_2 = (X_1 - \mu_1) + (X_2 - \mu_2) + (\mu_1 + \mu_2)$
- WLOG, $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim N(0, \sigma_2^2)$, $Z = X_1 + X_2$

Proof

$$\begin{aligned}
 f_Z(z) &= \int_{-\infty}^{\infty} \frac{e^{-\frac{(z-t)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1} \frac{e^{-\frac{t^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2} dt \propto \int_{-\infty}^{\infty} e^{-\left\{\frac{(z-t)^2}{2\sigma_1^2} + \frac{t^2}{2\sigma_2^2}\right\}} dt = \int_{-\infty}^{\infty} e^{-\frac{\sigma_2^2(z-t)^2 + \sigma_1^2 t^2}{2\sigma_1^2 \sigma_2^2}} dt \\
 &= \int_{-\infty}^{\infty} e^{-\frac{(\sigma_1^2 + \sigma_2^2)t^2 - 2\sigma_2^2 z t + \sigma_2^2 z^2}{2\sigma_1^2 \sigma_2^2}} dt = \int_{-\infty}^{\infty} e^{-\frac{(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2 \sigma_2^2} \left\{t^2 - \frac{2\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2} t + \frac{\sigma_2^2 z^2}{\sigma_1^2 + \sigma_2^2}\right\}} dt = \\
 &= \int_{-\infty}^{\infty} e^{-\frac{(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2 \sigma_2^2} \left\{\left(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2}\right)^2 - \frac{\sigma_2^4 z^2}{(\sigma_1^2 + \sigma_2^2)^2} + \frac{\sigma_2^2 z^2}{\sigma_1^2 + \sigma_2^2}\right\}} dt \\
 &= e^{-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}} \int_{-\infty}^{\infty} e^{-\frac{(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2 \sigma_2^2} \left(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2}\right)^2} dt = e^{-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}} \int_{-\infty}^{\infty} e^{-\frac{(\sigma_1^2 + \sigma_2^2)}{2\sigma_1^2 \sigma_2^2} t^2} dt \\
 &\propto e^{-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}}
 \end{aligned}$$

NOTE: we do not carry out real integration except algebra

图 2.8: 7

随机变量之商的概率密度:

如果 (X, Y) 是二维连续型随机变量, 它们的联合密度函数为 $f(x, y)$, 则它们的商 X/Y 是连续型随机变量, 且有密度函数

$$\begin{aligned}
 p_{\frac{X}{Y}}(z) &= \int_{-\infty}^{\infty} |t| f(z t, t) dt, \forall z \in \mathbb{R} \\
 p_{\frac{Y}{X}}(z) &= \int_{-\infty}^{\infty} |t| f(t, z t) dt, \forall z \in \mathbb{R}
 \end{aligned}$$

Example: 设随机变量 ϵ 和 η 相互独立, 同服从参数 $\lambda = 1$ 的指数分布, 求 ϵ/η 的概率密度。

1. $f(x, y) = e^{-x} e^{-y} = e^{-(x+y)}$
2. $p_{\frac{\epsilon}{\eta}}(z) = \int_0^{\infty} |t| e^{-zt} e^{-t} dt = \int_0^{\infty} t e^{-(z+1)t} dt = \frac{1}{(z+1)^2}, z > 0$
3. $p_{\frac{\eta}{\epsilon}}(z) = \int_0^{\infty} |t| e^{-t} e^{-zt} dt = \int_0^{\infty} t e^{-(1+z)t} dt = \frac{1}{(z+1)^2}, z > 0$
4. $p_{\frac{\epsilon}{\eta}}(z) = p_{\frac{\eta}{\epsilon}}(z), z > 0$
5. ϵ/η 服从参数 $\lambda = 1$ 的指数分布。

注意: 当 X 和 Y 相互独立时, 分别记 X 和 Y 的概率密度为 $f_1(x)$ 和 $f_2(y)$, 则 X/Y 的概率密度为

$$p_{\frac{X}{Y}}(z) = \int_{-\infty}^{\infty} |t| f_1(z t) f_2(t) dt = \int_{-\infty}^{\infty} |t| f_1(t) f_2(z t) dt \triangleq f_1 * f_2(z) = f_2 * f_1(z)$$

称此公式为卷积公式。

t distribution:

- 设 $X_1 \sim N(\mu, \sigma^2)$, $X_2 \sim \chi_n^2$, 且 X_1 与 X_2 相互独立。令 $Y = \frac{X_1}{\left(\frac{X_2}{n}\right)^{0.5}}$, 则 Y 服从自由度为 n 的 t 分布。即

$$Y \sim t_n$$

F distribution:

- $X_1 \sim \chi_m^2$, $X_2 \sim \chi_n^2$, 且 X_1 与 X_2 相互独立。令 $Y = \frac{\frac{X_1}{m}}{\frac{X_2}{n}}$, 则 Y 服从自由度为 m 和 n 的 F 分布。即

$$Y \sim F_{m,n}$$

§ 2.8 极小值和极大值的分布

极小值和极大值的分布

- Suppose we have n random variables, $\xi_1, \xi_2, \dots, \xi_n$, consider their maximum and minimum

$$\eta_1 = \max(\xi_1, \dots, \xi_n)$$

$$\eta_2 = \min(\xi_1, \dots, \xi_n)$$

η_1, η_2 are random variables, if ξ_1, \dots, ξ_n are independent, we can find the distribution function of η_1, η_2

$$F_{\eta_1} = P(\max(\xi_1, \dots, \xi_n) \leq x) = P\left(\bigcap_{k=1}^n \{\xi_k \leq x\}\right) = \prod_{k=1}^n P(\xi_k \leq x) = \prod_{k=1}^n F_k(x)$$

$$\begin{aligned} F_{\eta_2} &= P(\min(\xi_1, \dots, \xi_n) \leq x) = 1 - P(\min(\xi_1, \dots, \xi_n) > x) \\ &= 1 - P(\bigcap_{k=1}^n \{\xi_k > x\}) = 1 - \prod_{k=1}^n P(\xi_k > x) = 1 - \prod_{k=1}^n (1 - F_k(x)) \end{aligned}$$

When n goes large, the distribution converges to only a few prototypes, one of which is Weibull distribution. It is widely used in Reliability analysis.

图 2.9: 9

§ 2.9 Extreme Value Distribution

Extreme Value Distribution

- $G(z) \propto e^{-(1+\xi z)^{-1/\xi}}$
- Weibull, $\xi < 0$: $G(z) = \begin{cases} e^{-(-(\frac{z-b}{a}))^\alpha}, & z < b \\ 1, & z \geq b \end{cases}$
reverse Weibull
- Gumbel, $\xi = 0$: $G(z) = e^{-e^{-\frac{z-b}{a}}}$
- Frechet, $\xi > 0$: $G(z) = \begin{cases} 0, & z \leq b \\ e^{-(\frac{z-b}{a})^{-\alpha}}, & z > b \end{cases}$
- Location-scale family

图 2.10: 10

第3章

Definition of Expectation: discrete random variables • 设 X 为一离散型随机变量, 其分布律为 $P(X = x_i) = p_i$, 则期望值 $E(X)$ 定义为

$$E(X) = \sum_i x_i p_i$$

Definition of Expectation: continuous random variables • 如果连续型随机变量 X 具有密度函数 $f(x)$, 则期望值 $E(X)$ 定义为

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Probability as Expectation • 给定事件 A , 我们定义其指示函数为随机变量 I_A , 其中

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c \end{cases}$$

则事件 A 的概率可以表示为期望值

$$P(A) = E[I_A(\omega)]$$

Examples

- Binomial: $X \sim B(n, p)$, $E(X) = np$
- Poisson: $X \sim P(\lambda)$, $E(X) = \lambda$
- Normal: $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$
- Uniform distribution: $X \sim U(a, b)$, $E(X) = \frac{a+b}{2}$
- Exponential Distribution: $X \sim \text{Exp}(\lambda)$, $E(X) = \frac{1}{\lambda}$

随机变量函数的期望 • 设随机变量 X 为离散型, 有分布 $P(X = a_i) = p_i$, 则

$$E[g(X)] = \sum_{i=1}^{\infty} g(a_i) p_i, \quad \sum_{i=1}^{\infty} |g(a_i)| p_i < \infty$$

或者 X 为连续型, 有概率密度函数 $f(x)$, 则

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx, \quad \int_{-\infty}^{\infty} |g(x)| f(x) dx < \infty$$

这是一个定理, 而不是定义。

- The theorem is valid when X is a random vector.
- This theorem can be used to prove the previous results of expectation of sums and products.

Proof:

- 设 X 的分布律: $P(X = x_i) = p_i$

• $Y = g(X)$, 则 Y 的分布律为

$$P(Y = y_j) = P(g(X) = y_j) = \sum_{x_i: g(x_i)=y_j} P(X = x_i) = \sum_{i: g(x_i)=y_j} p_i$$

Proof:

$$\sum_i g(x_i)p_i = \sum_j \sum_{i: g(x_i)=y_j} g(x_i)p_i = \sum_j \sum_{i: g(x_i)=y_j} y_j p_i = \sum_j y_j \sum_{i: g(x_i)=y_j} p_i = \sum_j y_j P[g(X) = y_j] = E(Y) = E[g(X)]$$

Properties of Expectation

• **Linearity:** 若干个随机变量线性组合的期望, 等于各变量期望的线性组合。假设 c_1, c_2, \dots, c_n 为常数, 则有

$$E\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i E(X_i)$$

这里假定各变量的期望都存在。

Independence • 若干个独立随机变量之积的期望, 等于各变量的期望之积, 即

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$$

这里假定各变量相互独立且期望都存在。

Example • 飞机场载客汽车上有 20 位乘客离开机场后共有 10 个车站可以下车, 若某个车站没有人下车则该车站不停车。设乘客在每个车站下车的可能性相等, 而且 20 位乘客下车的选择都相互独立。以 X 表示停车的次数, 求 $E(X)$ 。

$$Y_i = \begin{cases} 1 & \text{if some passenger(s) gets off at the } i\text{-th stop} \\ 0 & \text{if no passenger gets off at the } i\text{-th stop} \end{cases}$$

$$X = \sum_{i=1}^{10} Y_i$$

$$E[X] = \sum_{i=1}^{10} E[Y_i] = \sum_{i=1}^{10} P(Y_i = 1) = \sum_{i=1}^{10} (1 - P(Y_i = 0)) = \sum_{i=1}^{10} \left[1 - P\left(\bigcap_{j=1}^{20} (Z_j \neq i)\right)\right]$$

Conditional Expectation

• 设 X 和 Y 为随机变量, 若 (X, Y) 为离散型, 且在给定 $X = x$ 之下, Y 有条件分布 $P(Y = a_i | X = x) = p_i$, $i = 1, 2, \dots$, 或者 (X, Y) 为连续型, 且在给定 $X = x$ 之下 Y 的条件密度函数 $f(y|x)$, 则

$$E(Y|X = x) = \begin{cases} \int_{-\infty}^{\infty} y f(y|x) dy & (X, Y \text{ 为连续型}) \\ \sum_i a_i P(Y = a_i | X = x) & (X, Y \text{ 为离散型}) \end{cases}$$

$E(Y|X)$ 是随机变量/向量 X 的函数。

全期望公式

• 设 X 和 Y 为两个随机变量, 则有

$$E(X) = E[E(X|Y)]$$

Example • 一窃贼被关在有 3 个门的地牢里, 其中第一个门通向自由, 出这门走 3 个小时便可以回到地面, 第 2 个门通向另一个地道, 走 5 个小时将返回到地牢; 第 3 个门通向更长的地道, 走 7 个小时也回到地牢。若窃贼每次选择 3 个门的可能性总相同, 求他为获得自由而奔走的平均时间。假设窃贼没有记忆。

Median

- 称 μ 为连续型随机变量 X 的中位数, 如果

$$P(X \leq \mu) = \frac{1}{2}, \quad P(X \geq \mu) = \frac{1}{2}$$

- 对于离散型随机变量, 中位数可能不是唯一的。
- 中位数具有鲁棒性, 不易受极端值影响。

方差、标准差和矩

Variance, Standard Deviation and Moment

- 刻画随机变量在其中心位置附近散布程度的数字特征, 其中最重要的是方差。在实际应用中, 方差不仅是信息度量的标准, 也是风险度量的标准。
- 定义设 X 为随机变量, 分布为 F , 则称

$$\text{Var}(X) = E[(X - E[X])^2] = \sigma^2$$

为 X 或分布 F 的方差, 其平方根

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sigma$$

称为 X 或分布 F 的标准差。

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

Properties • 设 c 为常数, 则有

$$0 \leq \text{Var}(X) = E[X^2] - (E[X])^2, \quad \text{Var}(X) \leq E[X^2]$$

$$\text{Var}(cX) = c^2 \text{Var}(X)$$

$$\text{Var}(X) = 0 \text{ 当且仅当 } P(X = c) = 1, \text{ 其中 } c = E[X]$$

- 对任何常数 c 有

$$\text{Var}(X) \leq E[(X - c)^2] \text{ 其中等号成立当且仅当 } c = E[X]$$

- 如果随机变量 X 和 Y 相互独立, a, b 为常数, 则

$$\text{Var}(aX + bY) = a^2 \text{Var}(X) + b^2 \text{Var}(Y)$$

Examples

- Binomial: $X \sim B(n, p)$, $\text{Var}(X) = np(1 - p)$
- Poisson: $X \sim P(\lambda)$, $\text{Var}(X) = \lambda$
- Uniform distribution: $X \sim U(a, b)$, $\text{Var}(X) = \frac{(b-a)^2}{12}$
- Exponential Distribution: $X \sim \text{Exp}(\lambda)$, $\text{Var}(X) = \frac{1}{\lambda^2}$
- Normal: $X \sim N(\mu, \sigma^2)$, $\text{Var}(X) = \sigma^2$

标准化随机变量

- $X^* = \frac{X - E[X]}{\sqrt{\text{Var}(X)}}$ 为 X 的标准化随机变量。易见 $E(X^*) = 0$, $\text{Var}(X^*) = 1$ 。
- 我们引入标准化随机变量是为了消除由于计量单位的不同而给随机变量带来的影响。
- 标准正态分布

- Location-scale family

矩 (Moment)

- 定义: 设 X 为随机变量, c 为常数, r 为正整数, 则称 $E[(X - c)^r]$ 为 X 关于 c 点的 r 阶矩。
- 当 $c = 0$ 时, $\alpha_r = E[X^r]$ 称为 X 的 r 阶原点矩。
- 当 $c = E[X]$ 时, $\mu_r = E[(X - E[X])^r]$ 称为 X 的 r 阶中心矩。
- 容易看出, 一阶原点矩就是期望, 二阶中心矩就是 X 的方差。

协方差 (Covariance)

- 定义: 称 $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$ 为 X 和 Y 的协方差。
- Properties:
 - $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
 - $\text{Cov}(X, X) = \text{Var}(X)$
 - $\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$
 - 若 X 和 Y 相互独立, 则 $\text{Cov}(X, Y) = 0$
 - $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$
 - $\text{Cov}(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) = \sum_{i=1}^2 \sum_{j=1}^2 a_i b_j \text{Cov}(X_i, Y_j)$

相关系数 (Correlation)

- 定义: 设随机变量 X, Y , 称 $\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ 为 X 和 Y 的相关系数。当 $\rho_{X,Y} = 0$ 时, 则称 X 和 Y 不相关。
- 由定义容易看出, 若令 $X^* = \frac{X - E[X]}{\sqrt{\text{Var}(X)}}$ 和 $Y^* = \frac{Y - E[Y]}{\sqrt{\text{Var}(Y)}}$ 分别为 X 和 Y 相应的标准化随机变量, 则 $\rho_{X,Y} = \text{Cov}(X^*, Y^*)$ 。因此, 形式上可以把相关系数视为“标准尺度下的协方差”, 从这个角度上说, 相关系数可以更好地反映两个随机变量间的关系, 而不受它们各自所用度量单位的影响。

Example

- 设随机变量 $(X, Y) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$, 则 $\rho_{X,Y} = \rho$ 。
- 为了简化计算, 我们可以考虑标准化后的 X^* 和 Y^* 。

解答

1. 首先, 标准化 X 和 Y :

$$X^* = \frac{X - \mu_1}{\sigma_1}, \quad Y^* = \frac{Y - \mu_2}{\sigma_2}$$

2. 计算标准化后的协方差:

$$\text{Cov}(X^*, Y^*) = \text{Cov}\left(\frac{X - \mu_1}{\sigma_1}, \frac{Y - \mu_2}{\sigma_2}\right) = \frac{1}{\sigma_1\sigma_2} \text{Cov}(X, Y)$$

3. 由于 (X, Y) 服从二维正态分布, 协方差 $\text{Cov}(X, Y) = \rho\sigma_1\sigma_2$, 因此:

$$\text{Cov}(X^*, Y^*) = \frac{1}{\sigma_1\sigma_2} \cdot \rho\sigma_1\sigma_2 = \rho$$

4. 计算标准化后的方差:

$$\text{Var}(X^*) = \text{Var}\left(\frac{X - \mu_1}{\sigma_1}\right) = \frac{1}{\sigma_1^2} \text{Var}(X) = \frac{1}{\sigma_1^2} \sigma_1^2 = 1$$

$$\text{Var}(Y^*) = \text{Var}\left(\frac{Y - \mu_2}{\sigma_2}\right) = \frac{1}{\sigma_2^2} \text{Var}(Y) = \frac{1}{\sigma_2^2} \sigma_2^2 = 1$$

5. 计算相关系数:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\rho\sigma_1\sigma_2}{\sqrt{\sigma_1^2\sigma_2^2}} = \rho$$

因此, 标准化后的 X^* 和 Y^* 的相关系数仍然是 ρ 。

Properties of Correlation Coefficients

- 若 X 和 Y 相互独立, 则 $\rho_{X,Y} = 0$ 。
- $|\rho_{X,Y}| \leq 1$, 等号成立当且仅当 X 和 Y 之间存在严格的线性关系, 即
 - 当 $\rho_{X,Y} = 1$ 时, 存在 $a > 0, b \in \mathbb{R}$, 使得 $X = aY + b$ (正相关)
 - 当 $\rho_{X,Y} = -1$ 时, 存在 $a < 0, b \in \mathbb{R}$, 使得 $X = aY + b$ (负相关)
- Note: $\rho_{X,Y}$ 也常称作 X 和 Y 的线性相关系数, 只能刻画 X 和 Y 间的线性相依程度, $|\rho_{X,Y}|$ 越接近 1, 就表示 X 和 Y 间的线性相关程度越高; $|\rho_{X,Y}| = 0$ 时, 只是表示 X 和 Y 间不存在线性相关, 但可以存在非线性的函数关系。

线性相关和非线性相关

- 设 $X \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$, 而 $Y = \cos(X)$, 则

$$\text{Cov}(X,Y) = E[XY] = \int_{-\frac{1}{2}}^{\frac{1}{2}} x \cos(x) dx = 0$$

- 所以 X 和 Y 不相关, 但是 X 和 Y 之间存在着非线性的函数关系。

不相关与独立性之间的关系

• 定理: 对随机变量 X 和 Y , 如果 X 和 Y 相互独立, 那么它们一定不相关; 但是如果它们不相关却未必相互独立。

- Note: 只在正态情形下不相关与独立等价。我们举二维正态的例子来说明, 不妨设

$$(X,Y) \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

, 则 X 和 Y 独立等价于 $\rho_{X,Y} = \rho = 0$, 从而等价于 X 和 Y 不相关。

其他一些数字特征与相关函数

- 平均绝对差 $E[|X - E[X]|]$
- 矩母函数 $E[e^{tX}]$, 其中 $t \in \mathbb{R}$, moment generating function
- 特征函数 $E[e^{itX}]$, 其中 $t \in \mathbb{R}$, i 为虚数, Characteristic function

- 定义: 如果 X 为一离散型随机变量, 其分布律为 $P(X = a_k) = p_k$, $k \in \mathbb{N}$, 那么

$$E[e^{itX}] = \sum_{k=1}^{\infty} p_k e^{ita_k}$$

- 如果连续型随机变量 X 具有密度函数 $f(x)$, 那么

$$E[e^{itX}] = \int_{-\infty}^{\infty} f(x) e^{itx} dx$$

- 概率质量函数/密度函数可以通过其特征函数唯一确定, 反之亦然, 通过逆傅里叶变换。

矩母函数、特征函数用于计算

- 矩母函数: $M(t) = E[e^{tX}]$

$$M^{(k)}(t) = E[X^k e^{tX}], \quad M^{(k)}(0) = E[X^k]$$

- $H(t) = \log M(t)$, $H'(0) = E[X]$, $H''(0) = \text{Var}(X)$

- 特征函数: $G(t) = E[e^{it}]$

$$G^{(k)}(t) = E[i^k X^k e^{it}], \quad G^{(k)}(0) = i^k E[X^k]$$

- 积分运算变为微分运算

- Example: exponential distribution

Characteristic function of Poisson Random Variable

$$E[e^{itX}] = \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

- Note we use the Taylor expansion of the complex function e^z .

Sum of Random variables, Convolution,

Characteristic function

- 设 X 与 Y 独立, 分别记 X 和 Y 的概率密度为 $f_1(x)$ 和 $f_2(y)$, 则 $Z = X + Y$ 的概率密度为

$$p(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x) dx = \int_{-\infty}^{\infty} f_1(z-y) f_2(y) dy \triangleq f_1 * f_2(z) = f_2 * f_1(z)$$

- $f_1(x)$ 和 $f_2(y)$ 的特征函数分别是 $G_1(t)$ 和 $G_2(t)$, 则 Z 的特征函数是 $G_1(t) \cdot G_2(t)$ 。

大数定律

- 大数定律, 什么意义下的极限

依概率收敛, weak convergence

定义: 如果对任何 $\epsilon > 0$, 都有

$$\lim_{n \rightarrow \infty} P(|\xi_n - \xi| > \epsilon) = 0, \quad \lim_{n \rightarrow \infty} P(|\xi_n - \xi| < \epsilon) = 1$$

那么我们就称随机变量序列 $\{\xi_n, n \in \mathbb{N}\}$ 依概率收敛到随机变量 ξ , 记为 $\xi_n \xrightarrow{p} \xi$ 。

- 设 $\{X_n\}$ 是一列独立同分布 (i.i.d.) 的随机变量序列, 具有公共的数学期望 μ 和方差 σ^2 , 则

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \xrightarrow{p} \mu$$

即 $\{X_n\}$ 服从弱大数定律。

Example

- 如果以 $\{\zeta_n\}$ 表示 n 重 Bernoulli(p) 试验中的成功次数, 则有

$$\frac{\zeta_n}{n} \xrightarrow{P} p$$

- 如果用 $f_n = \frac{\zeta_n}{n}$ 表示成功出现的频率, 则上例说明了 $f_n \xrightarrow{P} p$, 即频率 (依概率) 收敛到概率。

Proof of Weak Law of Large Numbers:

Chebyshev 不等式

- 设随机变量 X 的方差存在, 则

$$P(|X - E[X]| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2}, \forall \epsilon > 0$$

- More General, Markov 不等式: 若 Y 为只取非负值的随机变量, 则对任何 $\epsilon > 0$, 有

$$P(Y \geq \epsilon) \leq \frac{E[Y]}{\epsilon}$$

A version of Strong Law of Large Numbers

- Let X_n be i.i.d. random variables, $E[|f(X_n)|] < \infty$, Then we have

$$\frac{f(X_1) + f(X_2) + \cdots + f(X_n)}{n} \rightarrow E[f(X_1)]$$

almost surely. Based on this, we can compute expectation, or integrals by simulations, or Monte Carlo.

Law of Large Numbers • Let X_n be i.i.d. random variables, $E[|f(X_n)|] < \infty$, Then we have

$$\frac{f(X_1) + f(X_2) + \cdots + f(X_n)}{n} \rightarrow E[f(X_1)]$$

almost surely.

- Perspectives:
 - Monte Carlo simulations
 - The probabilistic meaning of faith
 - Average over space and average over time
 - n is sufficiently large

中心极限定理 (Central Limit Theorem)

- 定理: 设 $\{X_n\}$ 为 i.i.d. 的随机变量序列, 具有公共的数学期望 μ 和方差 σ^2 . 则 $X_1 + X_2 + \cdots + X_n$ 的标准化形式

$$\frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + \cdots + X_n - n\mu)$$

满足中心极限定理。即对任意 $x \in \mathbb{R}$, 有

$$\lim_{n \rightarrow \infty} F_n(x) = \Phi(x)$$

其中 $F_n(x)$ 为 $\frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + \cdots + X_n - n\mu)$ 的分布函数, 而 $\Phi(x)$ 为标准正态分布 $N(0, 1)$ 的分布函数, 记为

$$\frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + \cdots + X_n - n\mu) \xrightarrow{d} N(0, 1)$$

de Moivre-Laplace Theorem

- 设 $X_n, n \in \mathbb{N}$ 为 Bernoulli 试验, 相互独立且具有相同的分布

$$P(X_1 = 1) = 1 - P(X_1 = 0) = p, \quad 0 < p < 1$$

则有

$$\frac{1}{\sqrt{np(1-p)}}(X_1 + X_2 + \cdots + X_n - np) \xrightarrow{d} N(0, 1)$$

即

$$\lim_{n \rightarrow \infty} P\left(\frac{X_1 + X_2 + \cdots + X_n - np}{\sqrt{np(1-p)}} \leq x\right) = \Phi(x), \quad \forall x \in \mathbb{R}$$

Central Limit Theorem

- Summation version:

$$P\left(\frac{1}{\sqrt{n}\sigma}(X_1 + X_2 + \cdots + X_n - n\mu) \leq x\right) \rightarrow \Phi(x)$$

- Average version:

$$P\left(\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \leq x\right) \rightarrow \Phi(x)$$

Example • 设一考生参加 100 道题的英语标准化考试, 每道题均为有四个备选答案的选择题, 有且仅有一个答案是正确的, 每道题他都随机地选择一个答案, 假设评分标准为: 选对得一分, 选错或不选不得分。试给出该考生最终得分大于等于 25 的概率。

解答

1. 设 X_i 表示第 i 道题的得分, X_i 是一个 Bernoulli 随机变量, $P(X_i = 1) = \frac{1}{4}$, $P(X_i = 0) = \frac{3}{4}$ 。

2. 设 $S = \sum_{i=1}^{100} X_i$ 表示考生的总得分。 S 服从参数为 $n = 100$ 和 $p = \frac{1}{4}$ 的二项分布, 即 $S \sim \text{Binomial}(100, \frac{1}{4})$ 。

3. 我们需要计算 $P(S \geq 25)$ 。

利用中心极限定理进行近似计算:

1. 计算期望和方差:

$$E[S] = np = 100 \times \frac{1}{4} = 25$$

$$\text{Var}(S) = np(1-p) = 100 \times \frac{1}{4} \times \frac{3}{4} = 18.75$$

2. 标准化:

$$Z = \frac{S - E[S]}{\sqrt{\text{Var}(S)}} = \frac{S - 25}{\sqrt{18.75}}$$

3. 计算 $P(S \geq 25)$:

$$P(S \geq 25) = P\left(\frac{S - 25}{\sqrt{18.75}} \geq \frac{25 - 25}{\sqrt{18.75}}\right) = P(Z \geq 0)$$

4. 根据标准正态分布的性质, $P(Z \geq 0) = 0.5$ 。

因此, 该考生最终得分大于等于 25 的概率为 0.5。

Example

• 每天有 1000 个旅客需要乘坐火车从芝加哥到洛杉矶, 这两个城市之间有两条竞争的铁路它们的火车同时开出同时到达并且具有同样的设备。设这 1000 个人乘坐那一条铁路的火车是相互独立而且又是任意的, 于是每列火车的乘客数目可视为概率为 $\frac{1}{2}$ 的 1000 重 Bernoulli 试验中成功的次数。如果一列火车设置 $s < n$ 个座位, 那么一旦有多于 s 个旅客来乘车就容纳不下了。令这个事件发生的概率为 f_s 。利用中心极限定理, 有

$$f_s \approx 1 - \Phi\left(\frac{2s - 1000}{\sqrt{1000}}\right)$$

• 要求 s 使得 $f_s < 0.01$, 即在 100 次中有 99 次是有足够的座位的, 查表容易求出 $s = 537$ 。这样, 两列火车所有的座位数为 1074, 其中只有 74 个空位, 可见由于竞争而带来的损失是很小的。

Comparison of Limit theorems

• Law of large numbers

$$\bar{X} - \mu = o_p(1)$$

• Central limit theorem

$$\bar{X} - \mu = O_p\left(\frac{1}{\sqrt{n}}\right)$$

• Law of the iterated logarithm

$$\limsup_{n \rightarrow \infty} \frac{\bar{X} - \mu}{\sqrt{2 \log \log n}} = 1$$

Wright-Fisher Model

- In a diploid population, N individuals, $2N$ copies of each gene.
- A gene with two alleles, A or B. An individual can have two copies of the same allele or two different alleles.
- Assumption: generations do not overlap; each copy of the gene found in the new generation is drawn independently at random from all copies of the gene in the old generation.
- Last generation: allele frequency of A is p .
- The probability of obtaining k copies of the allele A in the next generation is

$$P(k) = \binom{2N}{k} p^k (1-p)^{2N-k}$$

- Can be formulated as a Markov Chain.
- An allele can disappear due to genetic drift.

Problems in population genetics

- Hardy–Weinberg principle: for sufficiently large populations, the allele frequencies remain constant from one generation to the next.

- Neutral or not.
- Selection.
- Linkage.

Genetic Drift

- Last generation: allele frequency of A is p .
- The probability of obtaining k copies of the allele A is

$$P(k) = \binom{2N}{k} p^k (1-p)^{2N-k}$$

- Neutral.
- Selection, Linkage.
- Hardy–Weinberg principle: for sufficiently large populations, the allele frequencies remain constant from one generation to the next.

Coalescence

- Setup: haploid, population size $2N_e$, and consider one gene locus.
- Assumption: generations do not overlap; each copy of the gene found in the new generation is drawn independently at random from all copies of the gene in the old generation.
- Problem: two random samples from one generation, trace backwards in time to the generation when they coalesce in their most recent common ancestor (MRCA).
- Under the random setting, trace backward one generation, the probability that they share the same parent is $\frac{1}{2N_e}$. Let this coalescence time be T_c , then T_c follows a geometric distribution.

$$P(T_c = t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \frac{1}{2N_e} \rightarrow \frac{1}{2N_e} e^{-\frac{t-1}{2N_e}}$$

- $E[T_c] \approx 2N_e$, $SD[T_c] \approx 2N_e$.

Data Compression

- Lossless coding: important files.
- Lossy coding: images, audio files.

Lossless Coding

- Prefix coding: A code is a prefix code if no target bit string in the mapping is a prefix of the target bit string of a different source symbol in the same mapping.
 - Note: if we represent a prefix code in a binary tree, they are all on the leafs.
- Instantaneous code, or context-free code: symbols can be decoded instantaneously after their entire code word is received.

• Kraft inequality: Suppose we have n symbols $\{a_1, a_2, \dots, a_K\}$. A binary prefix code assigns n_i bits to a_i , then

$$\sum_{i=1}^K \frac{1}{2^{n_i}} \leq 1$$

Lossless coding and probabilistic models

- Lossless coding: 无损失编码, 文件压缩, zip software。
- 最优编码 (optimal coding)。
- Shannon, 1948, A Mathematical Theory of Communication。

Shannon's lossless coding theorem

- Let a discrete random variable X take on the values $\{a_1, a_2, \dots, a_K\}$ with respective probabilities $\{p(a_1), p(a_2), \dots, p(a_K)\}$

Then, for any coding of X that assigns n_i bits to a_i , the average coding length,

$$\sum_{i=1}^K n_i p(a_i) \geq - \sum_{i=1}^K \log p(a_i) p(a_i) = H(X)$$

$H(X)$ is called the entropy of the r.v. X .

- $\{-\log_2 p(a_i), i = 1, \dots, K\}$, the optimal coding length of $\{a_1, a_2, \dots, a_K\}$ based on the model of X 。
- Proof: Let $q_i = \frac{1}{2^{n_i}} \sum_{j=1}^K \frac{1}{2^{n_j}}$, $i = 1, \dots, K$, then

$$- \sum_{i=1}^K p_i \log_2 p_i \leq - \sum_{i=1}^K p_i \log_2 q_i$$

That is, the Kullback-Leibler divergence $D(p||q)$ is nonnegative

$$- \sum_{i=1}^K p_i \log_2 \frac{q_i}{p_i} = D(p||q) \geq 0$$

In addition, we need the Kraft inequality.

Note: $\log x \leq x - 1$ 。

Where does the probability model come from?

- Suppose $n = 3$, we can code the 1-symbol or 1-letter words, $\{a_1, a_2, a_3\}$, or we can code 2-letter words (9 words), etc.
- The word frequencies change correspondingly。
- The code lengths of words generally are not constant, namely, variable-length。
- Given a model, quite some (sub)-optimal coding are available now such as Huffman coding。

信息的单位: bit(比特)

- Toss a fair coin, Bernoulli trial (p), $p = \frac{1}{2}$ 。
- Entropy $= -0.5 \times \log_2 0.5 - 0.5 \times \log_2 0.5 = 1$ bit。

Lempel-Ziv Algorithm

- Example, binary sequence

001101100011010101001001001101000001010010110010110

- Parse into “word”

0, 01, 1, 011, 00, 0110, 10, 101, 001, 0010, 01101, 000, 00101, 001011, 0010110

- Optimality: code lengths of all words have the same length. Notice that each word occurs only once, chance $\frac{1}{N}$.
 - When $N \rightarrow \text{large}$, word lengths $\rightarrow \text{large}$ while code length of each word remain unchanged.
 - The coding is data-dependent.

第4章

§ 4.1 总体 (Population)

4.1.1 定义

• **Example:** 要研究某大学学生的学习情况，则该校的全体学生构成问题的总体，每一个学生则是该总体中的一个个体。

4.1.2 类型

- 有限总体
- 无限总体：用概率分布 $F(x)$ 来描述。

4.1.3 说明

• 总体总是有限的，有限总体相应的分布只能是离散的，其具体形式将与个体总数有关且缺乏一个简洁的数学形式，这会使有力的概率方法无法使用。引进无限总体的概念，在概率论上相当于用一个连续分布去逼近离散分布，当总体所含个体极多时，这种逼近所带来的误差，从应用的观点看已可以忽略不计。更好的是，事实证明：几种常见且在概率论上较易处理的分布，如指数分布和正态分布等，尤其是正态分布，对许多实用问题的总体分布给出了足够好的近似，而围绕着这些分布建立了深入而有效的统计方法。

§ 4.2 样本 (Sample)

4.2.1 定义

• 样本是按一定的规定从总体中抽出的一部分个体，所谓“按一定的规定”，就是指总体中的每一个个体有同等的被抽出的机会，以及在这个基础上设立的某种附加条件。

4.2.2 示例

• 设一批产品包含 N 个，内有废品 M 个， M 未知。因而废品率也未知，现从其中抽出 n 个逐一检查它们是否为废品，据此去估计 p 。

- Sample with replacement, Binomial (n, p)
- Sample without replacement, hypergeometric distribution

§ 4.3 I.I.D. Sample (简单随机样本)

4.3.1 定义

• 为了使抽取的样本能很好地反映总体的信息，必须考虑抽样方法。最常用的一种抽样方法叫作“简单随机抽样”，它要求满足下列两条：

1. 代表性: 总体中的每一个体都有同等机会被抽入样本, 这意味着样本中每个个体与所考察的总体具有相同分布。因此, 任一样本中的个体都具有代表性。

2. 独立性: 样本中每一个体取什么值并不影响其它个体取什么值。这意味着, 样本中各个体 X_1, X_2, \dots, X_n 是相互独立的随机变量。

- 由简单随机抽样获得的样本 (X_1, X_2, \dots, X_n) 称为简单随机样本。

4.3.2 性质

• 定义: 设有一总体 $F(x)$, (X_1, X_2, \dots, X_n) 为从 F 中抽取的容量为 n 的样本, 若 X_1, X_2, \dots, X_n 相互独立, 且有相同分布, 即同有分布 F :

$$F(x_1, \dots, x_n) = \prod_{i=1}^n F(x_i)$$

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

§ 4.4 统计推断

- 从总体中抽取一定大小的样本去推断总体的概率分布的方法称为统计推断。
- 当样本的分布形式已知, 但含有未知参数时, 有关其参数的推断, 称为参数统计推断。

§ 4.5 统计量 (Statistic)

4.5.1 定义

• 完全由样本所决定的量, 叫做统计量。这里要注意的是“完全”这两个字, 它表明: 统计量只依赖于样本, 而不能依赖于任何其他未知的量。特别是, 它不能依赖于总体分布中所包含的未知参数。

4.5.2 常用统计量

- X_1, X_2, \dots, X_n 为从总体 $F(t)$ 中抽取的容量为 n 的样本:
- 样本均值: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- 样本方差: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- 样本矩: $a_k = \frac{1}{n} \sum_{i=1}^n X_i^k$, $m_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$
- 经验分布: $F_n(t) = \frac{1}{n} \sum_{i=1}^n I(t \geq X_i)$, This is a r.v.

§ 4.6 次序统计量 (Order Statistics) 及其有关统计量

- X_1, X_2, \dots, X_n 为从总体 $F(t)$ 中抽取的容量为 n 的样本:
- Sorting: $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$, 次序统计量
- 样本中位数: $m_{1/2} = \begin{cases} X_{(\frac{n+1}{2})}, & n \text{ is odd} \\ \frac{1}{2}[X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}], & n \text{ is even} \end{cases}$
- 极值: $X_{(1)}, X_{(n)}$

§ 4.7 三大分布

- χ^2

- t

- F 分布

All rooted in normal distribution $N(0, 1)$

§ 4.8 分布

4.8.1 χ^2 分布

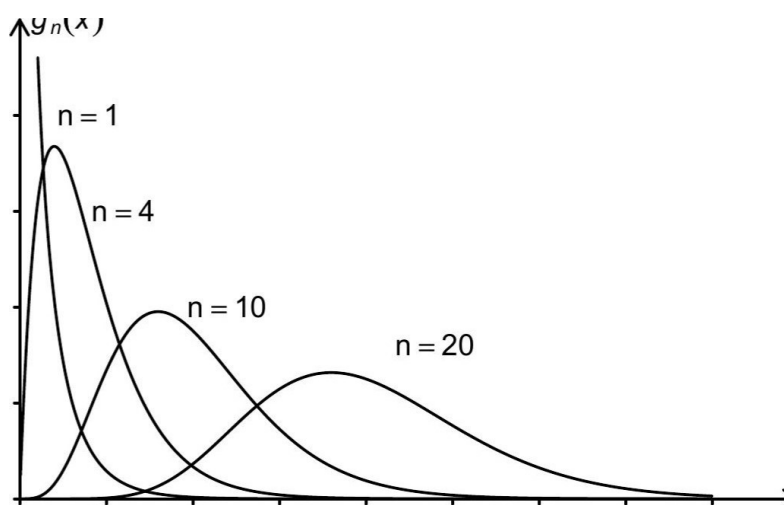
- 设 X_1, X_2, \dots, X_n 为独立同分布的随机变量，且 $X_i \sim N(0, 1)$ ，则

$$X = \sum_{i=1}^n X_i^2$$

服从自由度为 n 的 χ^2 分布，记为 χ_n^2 。

- $X \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$ 。

- 当自由度 n 越大，密度曲线越趋于对称； n 越小，曲线越不对称。当 $n = 1, 2$ 时，曲线是单调下降趋于 0；当 $n \geq 3$ 时，曲线有单峰，从 0 开始先单调上升，在一定位置达到峰值，然后单调下降趋向于 0。



4.8.2 Gamma 分布

- Gamma 分布 $\Gamma(\alpha, \lambda)$ 的概率密度函数为

$$f(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

- 设 Y_1, Y_2, \dots, Y_n 为独立同分布的随机变量，且 $Y_i \sim \text{Exp}(\lambda)$ ，则

$$Y = \sum_{i=1}^n Y_i \sim \Gamma(n, \lambda)$$

- Gamma 函数: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \lambda^\alpha e^{-\lambda x} dx$

4.8.3 Quantiles, Percentiles, 分位点

- 设 $X \sim F(x)$, 若 $P(X > c) = \alpha$, 则 c 成为此分布的上侧 (upper) α 分位点。

4.8.4 t 分布 (Student distribution)

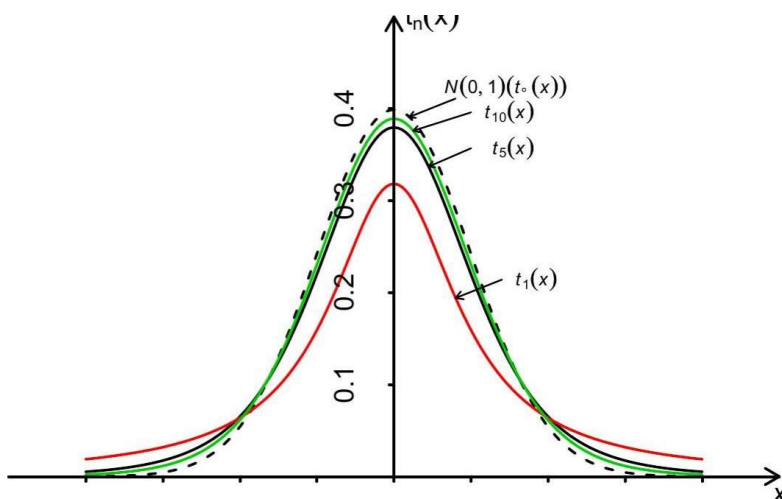
- 设 $X \sim N(0, 1)$, $Y \sim \chi_n^2$, 且 X 与 Y 独立。令

$$T = \frac{X}{\sqrt{Y/n}}$$

则 $T \sim t_n$, 自由度为 n , 其密度函数为

$$t_n(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$$

- $E[T] = 0$, $\text{Var}(T) = \frac{n}{n-2}$ 。 • 当 $n \rightarrow \infty$ 时, t_n 分布趋近于标准正态分布 $N(0, 1)$ 。



4.8.5 F 分布

- 设 $X \sim \chi_m^2$, $Y \sim \chi_n^2$, 且 X 与 Y 独立。令

$$F = \frac{(X/m)}{(Y/n)}$$

则 $F \sim F_{m,n}$, 自由度为 m 和 n , 其密度函数为

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (n + mx)^{-\frac{m+n}{2}}$$

- 若 $Z \sim F_{m,n}$, 则 $1/Z \sim F_{n,m}$ 。
- 若 $T \sim t_n$, 则 $T^2 \sim F_{1,n}$ 。
- $F_{m,n}(1 - \alpha) = 1/F_{n,m}(\alpha)$ 。

§ 4.9 正态总体样本均值和样本方差的分布

- 设 X_1, X_2, \dots, X_n 为独立同分布的随机变量, 且 $X_i \sim N(\mu, \sigma^2)$, 则

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

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.
- \bar{X} 和 S^2 独立。
- 定理: $T = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}$ 。

§ 4.10 Sample Mean and Sample Variance

- Sample Mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Sample Variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
- 设 X_1, X_2, \dots, X_n 为独立同分布的随机变量, 且 $X_i \sim N(\mu, \sigma^2)$, 则有以下结论:
 - 定理: 随机变量 \bar{X} 和随机向量 $(X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X})$ 独立。
 - \bar{X} 是正态随机变量。
 - $X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_n - \bar{X}$ 是正态随机向量。
 - $\text{Cov}(\bar{X}, X_1 - \bar{X}) = 0$ 。

4.10.1 Sample Mean and Sample Variance, continued

- \bar{X} 和 S^2 独立分布。• $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$ 。

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2 \\ \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2 \end{aligned}$$

- 由于正交性, 上述等式成立。
- 第一项和第二项是独立的, $\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2 \sim \chi_1^2$, 因此第二项是 χ_{n-1}^2 。