概率论与数理统计（缪柏其）

1 事件及其概率

1.1 概率论简史

1.2 随机实验和随机事件

随机试验：随机现象是自然界和社会中一类结果不可预先确定的客观现象. 当人们观测它时，所得结果不能预先确定，仅仅是多种可能结果中的一种. 对随机现象的实现或对它的某个特征的观测过程即称为随机试验，简称试验.

样本空间与事件：随机试验的每一个可能出现的结果称为基本事件，它是随机试验结果的最小单位，不能再分拆. 而由若干个基本事件组成的一个结果称为随机事件(简称事件). 事件通常用英文大写字母 A，B… 来表示. 随机试验中所有基本事件所构成的集合称为样本空间,通常用Ω或 S 来表示.样本空间的元素称为样本点，通常用ω表示.显然，每个样本点即为一个基本事件.

根据样本空间Ω的大小，可以将样本空间分为三类：有限样本空间（仅含有有限个样本点）、可数无穷样本空间（含有无穷且可数个样本点）和不可数样本空间（含有无穷且不可数个样本点）.

必然事件和不可能事件：样本空间本身称为必然事件，因其在随机试验中必然会发生，通常用Ω或 S 来表示.空集称为不可能事件，由于其不包含任何样本点，故在随机试验中不可能发生，通常用∅表示。

事件的和：事件 A 和事件 B 中至少有一个发生,记为 AUB,称为 A 与 B 的和.事件的和运算等同于集合运算“并”。两个事件和的运算可以推广到 n 个（或可数个）事件$A_1, A_2, \cdots, A_n$的和$A_1 \cup A_2 \cup \cdots \cup A_n$，简记为$\bigcup_{i=1}^{n} A_i$，表示这 n 个事件中至少有一个事件发生.

事件的差：事件 A 发生而事件 B 不发生，记为 A-B 或 $A\overline{B}$，称为 A 与 B 的差.

事件的积：事件 A 和事件 B 同时发生，记为 $A \cap B$,$A \cdot B$ 或 AB,称为 A 与 B 的积.事件的积运算等同于集合运算“交".两个事件积的运算可以推广到 n 个（或可数个）事件$A_1, A_2, \cdots, A_n$的积$A_1 \cap A_2 \cap \cdots \cap A_n$,简记为$\bigcap_{i=1}^{n} A_i$或$\prod_{i=1}^{n} A_i$,表示这 n 个事件同时发生。

不相容事件：事件 A 和事件 B 不能同时发生，即 $A \cap B = \emptyset$,称为事件 A 和事件 B 不相容或互斥.若一列事件中任意两个事件不相容，则称其为两两不相容事件列.

当事件$A_1, A_2, \cdots, A_n$两两不相容时，可以把“并”运算符改写成加号，$\bigcup_{i=1}^{n} A_i = \sum_{i=1}^{n} A_i$。

对立事件：｛事件 A 不发生｝这一事件称为 A 的对立事件(或余事件)，记为$\overline{A}$或A^c.

事件的运算公式：

$$A = A \cup A = A \cap A（等幂律）$$

$$A \cup BC = （A \cup B）\cap （A \cup C）$$

$$A \cap （B \cup C）= AB \cup AC \quad （分配律）$$

$$（A \cup B）（C \cup D）= AC \cup BC \cup AD \cup BD$$

$$\left(\bigcup_{k=1}^{n} A_k\right)^c = \bigcap_{k=1}^{n} A_k^c$$

$$\left(\bigcap_{k=1}^{n} A_k\right)^c = \bigcup_{k=1}^{n} A_k^c \quad （德摩根对偶法则）$$

$$A \cup B = B \cup A（交换律）$$

$$A \cup B \cup C = （A \cup B）\cup C = A \cup （B \cup C）$$

$$A \cap B \cap C = （A \cap B）\cap C = A \cap （B \cap C）\quad （结合律）$$

$$\begin{matrix}(\mathrm{A}\cup\mathrm{B})\cap\mathrm{A}=\mathrm{A}\\(\mathrm{A}\cap\mathrm{B})\cup\mathrm{A}=\mathrm{A}\end{matrix}\ （吸收律）$$

$$\begin{matrix}\mathrm{A}\cup\emptyset=\mathrm{A},\mathrm{A}\cap\emptyset=\emptyset\\\mathrm{A}\cap\mathrm{U}=\mathrm{A},\mathrm{A}\cup\mathrm{U}=\mathrm{U}\end{matrix}\ （0-1\ 律）$$

$$\mathrm{A}\cup\overline{\mathrm{A}}=\mathrm{U},\mathrm{A}\cap\overline{\mathrm{A}}=\emptyset（互补律）$$

$$\begin{matrix}(\overline{\mathrm{A}}\cup\mathrm{B})\cap\mathrm{A}=\mathrm{A}\cap\mathrm{B}\\(\overline{\mathrm{A}}\cap\mathrm{B})\cup\mathrm{A}=\mathrm{A}\cup\mathrm{B}\end{matrix}\ （重叠律）$$

$$\bigcup_{\mathrm{k}=1}^{\mathrm{n}}\mathrm{A_k}=\sum_{\mathrm{k}=1}^{\mathrm{n}}\mathrm{A_k}\bigcap_{\mathrm{j}=1}^{\mathrm{k}-1}\overline{\mathrm{A}}_{\mathrm{j}}$$

1.3 概率的定义和性质

概率是随机事件发生可能性大小的数字表征，取值于区间[0,1].概率是事件的函数.因为事件与集合有一一对应关系，所以概率可以视为集合的函数.

1.3.1 古典概型

假设样本空间Ω中试验结果只有有限个，记为$|\Omega|=N$(样本点个数)，且每个基本事件发生的可能性相同.若事件 A 中的基本事件个数$|A|=M$(称为事件 A 的有利场合数，因为这些基本事件的发生对事件 A 的发生“有利”)，则事件 A 的概率定义为 $\mathrm{P(A)}=\frac{|A|}{|\Omega|}=\frac{M}{N}$。在有限性和等可能性下定义概率的模型称为古典概型.

1.计数原理

加法原理：完成一件事情有 n 类办法，在第一类办法中有m_1种不同的方法，在第二类办法中有m_2种不同的方法，……，在第九类办法中有m_n种不同的方法,那么完成这件事共有 $\mathrm{N}=m_1+m_2+\cdots+m_n$种不同的方法.

乘法原理：完成一件事情需要分成 n 个步骤，做第一步有m_1种不同的方法，做第二步有m_2种不同的方法，……，做第八步有m_n种不同的方法，那么完成这件事有 $\mathrm{N}=m_1m_2\cdots m_n$种不同的方法.

（1）从 n 个不同的元素中，有放回地取出 r 个元素组成的可重复排列的不

同方式有n^r种.从 n 个不同的元素中，不放回地取出 r 个元素组成的不重复排列的不同方式有 $\mathrm{n(n-1)\cdots(n-r+1)=P_n^r}$种，这种排列称为选排列.特别 r=n 时,称为全排列.

（2）从 n 个不同的元素中，不放回地取 r 个元素组成的组合，不同方式个数为$\binom{n}{r}=\frac{n!}{r!(n-r)!}=C_n^r$.

（3）从 n 个不同的元素中，有放回地取 r 个元素组成的组合(不考虑顺序)，不同方式个数为$\binom{n+r-1}{r}$，这个数称为重复组合数.

2.盒子模型

把 r 个球随机放到不同编号的 n 个盒子中去：

（1）球可辨，每个盒子中不限球的个数.此为重复排列，不同的放法个数为n^r。

（2）球可辨，每个盒子中至多放一个球.此为选排列，不同的放法个数为$\frac{n!}{(n-r)!}$。

（3）球不可辨，每个盒中不限球的个数.不同的排法个数为$\binom{n+r-1}{r}$。

（4）球不可辨，每个盒子中至多放一个球.此为一般的组合，不同的放法个数为$\binom{n}{r}$。

3.多组组合

把n个不同的元素分为有序的k个部分，第i部分有r_i个元素，$\mathrm{i}=1,2,\cdots,\mathrm{k},r_1+r_2+\cdots+r_k$，则不同的分法个数为$\frac{n!}{r_1!r_2!\cdots r_k!}$，称为多项式系数，它是$(a_1+a_2+\cdots+a_k)^n$展开后$a_1^{r_1}a_2^{r_2}\cdots a_k^{r_k}$前面的系数.

4.不尽相异元素的排列

有 n 个元素，属 k 个不同类，同类不可辨认，各类元素分别有$n_1,n_2,\cdots,n_k$个，$n_1+n_2+\cdots+n_k=n$，排成一列，共有$\frac{n!}{n_1!n_2!\cdots n_k!}$种排法。

一个小概率事件在一次试验中是几乎不可能发生的，但在多次重复试验中几

乎是必然发生的，数学上称之为小概率原理.设 $P(A)=\alpha$,若α小于于某个给定的很小的数 c,则该事件称为小概率事件.在概率统计中常用的 c 有两个,c=0.01 和 c=0.05,我们可以分别称为严标准和宽标准。

1.3.2 概率的统计定义

古典概型中有两个限制，如果去掉有限性，保留基本事件的等可能性，就称为几何概型.

概率的统计定义：设事件 A 发生的概率为 p，为了确定他，把该实验在相同的条件下独立重复做 n 几次，用n_A表示事件 A 出现的频数，n_A/n称为事件 A 的频率.当 $n\to\infty$时，频率n_A/n会在某个数 p 附近波动，且慢慢稳定下来，我们就称该数为事件 A 发生的概率，记为 $P(A)=p$。

1.3.3 主观概率的定义

主观概率：一个事件发生的概率规定为某人在主观上相信该事件会发生程度的数字衡量.

1.3.4 概率的公理化定义

设 $P(\cdot)$是定义在样本空间Ω中事件上的一个实函数，它取值在 0 和 1 之间，满足：若 A 是样本空间上的一个事件，则（1）（非负性）$0\le P(A)\le 1$;（2）（规范性）设Ω为必然事件，则 $P(\Omega)=1$；（3）（可数可加性）对Ω中两两不相容事件列$A_1,A_2,\cdots,A_k,\cdots$，即$A_i\cap A_j=\emptyset,\forall i\neq j$，有 $P(\bigcup_{k=1}^{\infty}A_k)=\sum_{k=1}^{\infty}P(A_k)$，则 $P(\cdot)$称为一个概率函数，简称概率。

（1）$P(\emptyset)=0$。

（2）（有限可加性）若$A_1,A_2,\cdots,A_n$两两不相容，则 $P(\bigcup_{k=1}^{n}A_k)=\sum_{k=1}^{n}P(A_k)$。

（3）（可减性）若 $A\subset B$，则 $P(B-A)=P(B)-P(A)$.

（4）（单调性）若 $A \subset B$，则 $P(B) \geq P(A)$.

（5）$P(\overline{A}) = 1 - P(A)$.

（6）（加法定理/容斥定理）对任意事件$A_1, A_2, \cdots, A_n$，有 $P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k) - \sum_{1 \leq i < j \leq n} P(A_i A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) - \cdots + (-1)^{n-1} P(A_1 A_2 \cdots A_n)$。

（7）（次可加性）对任意事件列$A_1, A_2, \cdots, A_n, \cdots$，有 $P(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} P(A_n)$。

（8）（下连续性）若事件列满足$A_n \subset A_{n+1}$，则 $P(\bigcup_{n=1}^{\infty} A_n) = \lim_{n \to \infty} P(A_n)$。

（9）（上连续性）若事件列满足$A_n \supset A_{n+1}$，则则 $P(\bigcap_{n=1}^{\infty} A_n) = \lim_{n \to \infty} P(A_n)$。

1.4 条件概率

1.4.1 条件概率的定义

条件概率：设 A,B 为样本空间Ω中的两个事件，P(B)>0，称 $P(A|B) = \frac{P(AB)}{P(B)}$，为已知事件 B 发生的情况下，事件 A 发生的条件概率。

乘法公式：$P(AB) = P(A)P(B|A)$，那么若 $P(\bigcap_{k=1}^{n-1} A_k) > 0$，$P(\bigcap_{k=1}^{n} A_k) = P(A_1)P(A_2|A_1) \cdots P\left(A_n \middle| \bigcap_{k=1}^{n-1} A_k\right)$

$P(\cdot|B)$是一个概率函数. 概率函数具有的性质，对条件概率函数同样成立.

1.4.2 全概率公式

完备事件群：设$B_1, B_2, \cdots, B_n$是样本空间Ω中的一组概率大于 0 的事件，满足$B_i B_j = \emptyset, \sum_{i=1}^{n} B_i = \Omega$，则称$B_1, B_2, \cdots, B_n$是样本空间Ω的一个完备事件群（也称为Ω的一个划分）。

全概率公式：设$B_1, B_2, \cdots, B_n$是样本空间Ω中的一个完备事件群，A 为Ω中任一事件，则 $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$

1.4.3 贝叶斯公式

设$B_1, B_2, \cdots, B_n$是样本空间Ω中的一个完备事件群，A 为Ω中任一事件，则$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P\left(A\middle|B_j\right)P(B_j)}$

假设某个过程具有$B_1, B_2, \cdots, B_n$这样几个可能的前提(原因)，而$\{P(B_k)\}$是人们对这 n 个可能前提(原因)的可能性大小的一种事前(即还没有进行当前试验)估计，称之为先验概率.当这个过程有了一个结果 A 之后 (即当前试验完成后)，人们便会通过条件概率$\{P(B_k|A)\}$来对这 n 个可能前提的可能性大小作出一种新的认识，因此，将这些条件概率称为后验概率.

1.5 独立性

若 A,B 是样本空间Ω中的两个事件，满足 $\mathrm{P(AB) = P(A)P(B)}$，则称事件 A,B 相互独立。如果事件 A 与事件 B 的发生互不影响，那么两事件是相互独立的.

若 A,B 是样本空间Ω中的两个事件，若 $\mathrm{P(A) > 0}$,则事件 A,B 相互独立的充要条件为 $\mathrm{P(B|A) = P(B)}$。

A,B 是样本空间Ω中的两个事件，则下面陈述等价：（1）A 与 B 独立（2）A 与$\overline{B}$独立（3）$\overline{A}$与 B 独立（4）$\overline{A}$与$\overline{B}$独立。

设$A_1, A_2, \cdots, A_n$是样本空间Ω中的 n 个事件，若对∀k 及$\forall 1 \le i_1 \le i_2 \le \cdots \le i_k \le n$满足 $\mathrm{P}(A_{i_1}A_{i_2}\cdots A_{i_k}) = P(A_{i_1})P(A_{i_2})\cdots P(A_{i_k})$，则称事件$A_1, A_2, \cdots, A_n$相互独立。

设$A_1, A_2, \cdots, A_n$是样本空间Ω中的 n 个事件，记$\widetilde{A}_i = A_i$或$\overline{A}_i$，若 $\mathrm{P}(\widetilde{A}_1\widetilde{A}_2\cdots\widetilde{A}_n) = P(\widetilde{A}_1)P(\widetilde{A}_2)\cdots P(\widetilde{A}_n)$。

设$A_1, A_2, \cdots, A_n$是样本空间Ω中的 n 个事件，如果其中任意两个事件相互独立，那么称事件$A_1, A_2, \cdots, A_n$两两独立。

如果事件列$\{A_n: n = 1,2,\cdots,\infty\}$中任意有限个事件相互独立，那么称其为独立

事件列.如果其中任意两个事件相互独立，那么称其为两两独立事件列.

1.6 扩展进阶：求概率的一些方法

1.6.1 选择合适的样本空间

1.6.2 递推法（条件化）

1.6.3 利用概率性质求解

1.7 扩展阅读 1：贝叶斯公式和垃圾邮件识别

1.8 扩展阅读 2：三门问题

2 随机变量及其分布

2.1 随机变量的概念

随机变量 X 是一个映射，X：$\Omega \to \mathbb{R}$，对每个（可测集）$A \subset \mathbb{R}$，$\{X \in A\}$是一个（可定义概率的）随机事件，且 $P(X \in A) = P(\{\omega \in \Omega: X(\omega) \in A\})$。随机变量（r.v.）是取值随实验结果而定且有一定概率分布的变量。

2.2 离散型随机变量的分布

如果随机变量 X 只取有限多个或可数多个值，那么称 X 为离散型随机变量。设 X 取的一切可能值为 $x_1, x_2, \cdots, x_n, \cdots$，则 $P(X = x_k) = p_k$，其中 $p_k \geq 0$，$\sum_{k=1}^{\infty} p_k = 1$，称为离散型随机变量 X 的分布律或概率质量函数（pmf）。

2.2.1 0-1 分布

若随机变量 X 的分布律为 $\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$，其中 $0 < p < 1$，则称 X 服从 0-1 分布或者伯努利分布或两点分布。

2.2.2 离散均匀分布

若随机变量 X 的分布律为 $P(X = x_k) = \frac{1}{n}$，其中 x_k 为 n 个不同的实数，则称随

机变量 X 服从离散均匀分布，常用的$x_k = k$。

参数为(N, M, n)，$M \ll N$的超几何分布$P(X = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$。

2.2.3 二项分布

只有两种可能结果的实验称为伯努利实验。

设离散型随机变量 X 所有可能取值为$\{0,1,\cdots,n\}, 0 < p < 1$，如果其分布律为$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$，那么称 X 服从二项分布，记作$X \sim B(n,p)$，而$P(X = k)$常记作$b(n,p,k)$。

2.2.4 负二项分布

设随机变量X_r取正整数值，其分布律为$P(X_r = k) = \binom{k-1}{r-1} p^r q^{k-r}$，其中 r 为正整数，$0 < p < 1$，则称 X 服从参数 r,p 的负二项分布或者帕斯卡分布，记为$X \sim NB(r,p)$,$P(X_r = k)$记为$nb(r,p,k)$。

若 r=1，分布称为几何分布，记为$X_1 \sim Ge(p)$，$P(X_1 = k) = pq^{k-1}$.

以所有正整数为取值集合的随机变量 X 服从几何分布$Ge(p)$，当且仅当对任何正整数 m 和 n，都有 $\mathrm{P(X > m + n | X > m) = P(X > n)}$，该性质称为几何分布的无记忆性。

2.2.5 泊松分布

若随机变量 X 的分布律为$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$，$\lambda > 0$，则称 X 服从参数λ的泊松分布，记为 X~P(λ)。

泊松逼近定理：设一族随机变量$X_n \sim B(n, p_n)$，若当 $\mathrm{n} \to \infty$，$np_n \to \lambda > 0$，则$\lim_{n\to\infty} P(X_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, 一般 $\mathrm{n} \geq 30, np_n \leq 5$ 时可使用。当 $\mathrm{r} \to \infty, r(1-p) \to \lambda > 0$，那么$p \approx 1 - \frac{\lambda}{r}$，令$k - r = x$为固定数，$\lim_{r\to\infty} P(X_r = k) = e^{-\lambda}\frac{\lambda^x}{x!}$。

2.3 连续型随机变量的分布

2.3.1 随机变量的分布函数

设 X 为随机变量，x 为任一实数，称$F(x)=P(X\leq x)$为随机变量 X 的（累积）分布函数（cdf）。

设离散型随机变量 X 的可能值为$x_1,x_2,\cdots,x_n,\cdots$，其分布律$\{p_k\}$和分布函数$F(x)$互相确定，等价描述了 X 的概率分布情况。

$F(x)$为非减函数，且只存在第一类间断点；$\forall x\in\mathbb{R},0\leq F(x)\leq 1$，且$F(\infty)=1,F(-\infty)=0$；$F(x)$为右连续函数，$F(x+0)=F(x)$。

2.3.2 概率密度函数

设随机变量 X 的分布函数$F(x)$，若存在非负函数$f(x)\geq 0$,使得$\forall x\in\mathbb{R},F(x)=\int_{-\infty}^{x}f(t)dt$，则称 X 为连续型随机变量，$f(x)$称为分布函数的概率密度函数（pdf），简称密度函数，记为$X\sim f(x)$。

概率密度函数具有性质：

（1）$f(x)\geq 0$，$\forall x\in\mathbb{R}$.

（2）$\int_{-\infty}^{\infty}f(x)dx=1$

（3）$\forall x_1<x_2$,有 $P(x_1<X\leq x_2)=F(x_2)-F(x_1)=\int_{x_1}^{x_2}f(x)dx$。

（4）若$f(x)$在x_0连续，则$F'(x_0)=f(x_0)$

（5）$\forall x\in\mathbb{R}$，$P(X=x)=0$

2.3.3 几种重要的连续型分布

1.均匀分布

随机变量 X 在有限区间(a,b)内取值，且概率密度函数为$f(x)=\frac{1}{b-a}I_{(a,b)}(x)$，则称 X 服从区间$(a,b)$上的均匀分布，记为 $X\sim U(a,b)$。

2.指数分布

若随机变量 X 的密度函数为$f(x)=\lambda e^{-\lambda x}I_{(0,\infty)}(x)$，其中$\lambda>0$ 为参数，则称 X 服从参数为λ的指数分布，记为 X~Exp(λ)。

设 X~Exp(λ)，则对任意$s,t>0$，有 P(X > s + t|X > t) = P(X > s)，即无记忆性。

如果$\lambda(t)=\frac{k}{\delta}(\frac{t}{\delta})^{k-1}$，则$f(t)=\frac{k}{\delta}(\frac{t}{\delta})^{k-1}exp\left\{-(\frac{t}{\delta})^{k}\right\}I_{(0,\infty)}(t)$，称为参数为$\delta,k$的韦布尔分布。

3.正态分布

随机变量 X 的概率密度为$f(x)=\frac{1}{\sqrt{2\pi}\sigma}exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$，则称 X 服从参数为$\mu$，$\sigma^2$的正态分布，记为 X~N($\mu$，$\sigma^2$).

μ称为位置参数，表示图形的对称位置。σ称为尺度参数，表示图形高低的改变。当$\mu=0,\sigma=1$ 时，$f(x)$称为标准正态分布密度函数，记为$\varphi(x)$，对应的分布函数称为标准正态分布函数，记为$\Phi(x)$。

设 X~N(μ，σ^2)，则$F(x)=\Phi(\frac{x-\mu}{\sigma})$，即$\frac{x-\mu}{\sigma}$~N(0,1)，$\frac{x-\mu}{\sigma}$称为标准化变换。

2.4 随机变量函数的分布

2.4.1 离散型随机变量函数的分布

设随机变量$Y=g(X)$，X有分布律$P(X=x_k)=p_k$，则$P(Y=g(x_k))=\sum_{i:g(x_i)=g(x_k)}p_i$，如果 g 恒为常数，那么Y取该常数的概率为 1，称为 Y 的分布退化。

2.4.2 连续型随机变量函数的分布

对变换$Y=g(X)$，X~$f(x)$，则 Y 的分布函数为$F_1(y)=P(Y\le y)=P(g(X)\le y)=\int_{g(X)\le y}f(x)dx$。

若$g(x)$是严格单调的且反函数可导时，则随机变量Y仍为连续型随机变量，

且有概率密度函数$f_1(y)=\begin{cases}f(h(y))|h'(y)|,\alpha<y<\beta\\0,\text{其他}\end{cases}$，$h(y)$为$g(x)$反函数。

当$g(x)$不是全区间上单调而是逐段单调时，密度变换公式有下面形式：设随机变量X的密度函数为$f(x),a<x<b$。如果$(a,b)=\sum_j I_j$，使得$y=g(x)$在每个子区间有唯一的反函数$h_j(y)$，且导函数$h'_j(y)$连续，那么$Y=g(X)$是连续型随机变量，$f_1(y)=\sum_j f(h_j(y))\left|h'_j(y)\right|I_j$。

2.5 扩展阅读：正态分布的由来

棣莫弗-拉普拉斯中心极限定理：设随机变量X_n服从参数为 n 和 p 的二项分布，其中 0<p<1 为一给定的常数，那么，对任意的实数 x，有$\lim_{n\to\infty}P(\frac{X_n-np}{\sqrt{np(1-p)}}\le x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$。

随机误差分布亦是正态分布。

3 多维随机变量及其分布

3.1 多维随机变量及其分布

3.1.1 多维随机变量

设$X_1(\omega),\cdots,X_n(\omega)$为同一样本空间上的随机变量，则称$\boldsymbol{X}(\omega)=(X_1(\omega),\cdots,X_n(\omega))$为 n 维随机变量，或称 n 为随机向量，简记为$(X_1,\cdots,X_n)$或$\boldsymbol{X}$。

设(X,Y)是二维随机变量，$(x,y)\in\mathbb{R}^2$，称二元函数$F(x,y)=P(X\le x,Y\le y)=P(\{\omega:X(\omega)\le x,Y(\omega)\le y\})=P(\{X\le x\}\cap\{Y\le y\})$为$(X,Y)$的分布函数，或$(X,Y)$的联合分布函数。

二维随机变量的联合分布函数$F(x,y)$具有性质：

（1）$F(x,y)$分别对 x,y 单调不减；

（2）$\forall(x,y)\in\mathbb{R}^2$，$0\le F(x,y)\le 1$，$F(\infty,\infty)=1$，$F(-\infty,y)=F(x,-\infty)=F(-\infty,-\infty)=0$；

（3）$F(x,y)$分别关于x,y右连续；

（ 4 ） $\forall x_1<x_2,y_1<y_2,P(x_1<X<x_2,y_1<Y<y_2)=F(x_2,y_2)+F(x_1,y_1)-F(x_1,y_2)-F(x_2,y_1)$

设X,Y的可能取值为$\{(x_i,y_j),i=1,2,\cdots,j=1,2,\cdots\}$，记$P(X=x_i,Y=y_j)=p_{ij}$，称为二维离散型随机变量的联合概率质量函数或联合分布律。

称$\boldsymbol{X}=(X_1,\cdots,X_n)$为一 n 维离散型随机变量，若每一个$X_i$都是一个离散型随机变量，并设$X_i$的所有可能取值（有限个或可数个）为$\{a_{i1},a_{i2},\cdots\}$，则称$p(j_1,\cdots,j_n)=P(X_1=a_{1j_1},\cdots,a_{nj_n})$为 n 维随机变量$\boldsymbol{X}$的联合概率质量函数或联合分布律。

多项分布$\boldsymbol{X}\sim M(N;p_1,\cdots,p_n)$，$P(k_1,\cdots,k_n)=\frac{N!}{k_1!\cdots k_n!}{p_1}^{k_1}\cdots{p_n}^{k_n}$。

3.1.2 连续型多维随机变量的联合密度函数

设$(X,Y)\sim F(x,y)$，若存在可积的非负函数$f(x,y)$，使得对于$\forall(x,y)\in\mathbb{R}^2$，有$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(u,v)dudv$，则称$(X,Y)$为二维连续型随机变量，$F(x,y)$称为其联合分布函数，称$f(x,y)$为其联合概率密度函数，简称联合密度函数。

设(X,Y)的概率密度函数为$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-a)^2}{\sigma_1^2}-2\rho\frac{(x-a)(y-b)}{\sigma_1\sigma_2}+\frac{(y-b)^2}{\sigma_2^2}\right]\right\}$，称$(X,Y)\sim N(a,b,\sigma_1^2,\sigma_2^2,\rho)$，即二元正态分布。

设$(X,Y)\sim f(x,y)=\frac{1}{|G|}I_G(x,y)$，称$(X,Y)$在 G 上服从均匀分布。

设$\boldsymbol{X}=(X_1,\cdots,X_n)$，$\boldsymbol{x}=(x_1,\cdots,x_n)\in\mathbb{R}^n$，称$F(\boldsymbol{x})=P(X_1\le x_1,\cdots,X_n\le x_n)\equiv P(\boldsymbol{X}\le\boldsymbol{x})$为 n 维随机变量$\boldsymbol{X}$的联合分布函数。若存在非负 n 元函数$f(x_1,\cdots,x_n)$使得$F(\boldsymbol{x})=\int_{-\infty}^{\boldsymbol{x}}f(\boldsymbol{u})d\boldsymbol{u}$，则$f(\boldsymbol{x})$称为$\boldsymbol{X}$的联合密度函数，$\boldsymbol{X}$称为连续型 n 维随机变量。

3.2 边缘（际）分布

设(X,Y)的联合分布函数为$F(x,y)$，则其分量X,Y的分布函数$F_1(x)$和$F_2(y)$称为(X,Y)或F的边缘（际）分布，$F_1(x)=F(x,\infty)$。

3.2.1 二维离散型随机变量的边缘（际）分布

$$P(X=x_i)=\sum_j P(X=x_i,Y=y_j)=\sum_j p_{ij}=p_{i\cdot}$$

3.2.2 二维连续型随机变量的边缘分布

设二维连续型随机变量$(X,Y)\sim f(x,y)$，则$f_1(x)=\int f(x,y)dy$，称为(X,Y)或$f(x,y)$的边缘概率密度函数，简称边缘密度函数。

二元正态分布的概率密度函数可表示为$f(\boldsymbol{x})=(2\pi)^{-1}|\boldsymbol{A}|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{A}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$，$\boldsymbol{A}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$，它的随机变量服从正态分布，但与$\rho$无关。

多维随机变量的边缘分布和边缘密度函数：

设$\boldsymbol{X}=(X_1,\cdots,X_n)$，任取 k 个变量，不妨为前 k 个，记$\boldsymbol{U}=(X_1,\cdots,X_k)$，余下$\boldsymbol{V}=(X_{k+1},\cdots,X_n)$，则$\boldsymbol{U}$的分布即为$\boldsymbol{X}$的边缘分布，记为$F_{\boldsymbol{U}}(\boldsymbol{u})=P(\boldsymbol{U}\le\boldsymbol{u},\boldsymbol{V}<\infty)=F(\boldsymbol{u},\infty)$，$f_{\boldsymbol{U}}(\boldsymbol{u})=\int f(\boldsymbol{u},\boldsymbol{v})d\boldsymbol{v}$，n 维随机变量有$2^n-2$个边缘分布。

3.3 条件分布

$p_{i|j}=\frac{p_{ij}}{p_{\cdot j}}$，$F_{X|Y}(x|y)=\int_{-\infty}^{x}f_{X|Y}(u|y)du=\int_{-\infty}^{x}\frac{f(u,y)}{f_2(y)}du$。

如果$f_2(y)>0$，那么称$f_{X|Y}(x|y)=\frac{f(x,y)}{f_2(y)}$为给定$Y=y$下随机变量 X 的条件概率密度函数，常表达为$X|Y=x\sim f_{X|Y}(x|y)$。

$$f(x,y)=f_{X|Y}(x|y)f_2(y)$$

$$f_{X|Y}(x|y)=\frac{f_{Y|X}(y|x)f_1(x)}{f_2(y)}$$

多维随机变量的条件密度函数：

$$f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = \frac{f(\boldsymbol{x},\boldsymbol{y})}{f_Y(\boldsymbol{y})}$$

3.4 相互独立的随机变量

设随机变量X,Y的联合分布函数为$F(x,y)$，边缘分布为$F_1(x),F_2(y)$。若$\forall(x,y)\in\mathbb{R}^2$，都有$F(x,y)=F_1(x)F_2(y)$，则称随机变量$X,Y$相互独立。

对于离散型随机变量，$p_{ij}=p_{i\cdot}p_{\cdot j}$。

对于连续型随机变量，$f(x,y)=f_1(x)f_2(y)$，等价于$f(x,y)$可分离变量，即$f(x,y)=g_1(x)g_2(y)$而不必是概率密度函数。

两个正态分布随机变量独立的充要条件是$\rho=0$。

多维随机向量的相互独立性：

设$\boldsymbol{X}\sim F(\boldsymbol{x})$，记$X_i\sim F_i(x_i)$，若对任意的$\boldsymbol{x}$，有$F(\boldsymbol{x})=F_1(x_1)\cdots F_n(x_n)$，则称$X_1,\cdots,X_n$相互独立。

（1）对于离散型随机变量，若对任意的$\boldsymbol{x}$，有$P(\boldsymbol{X}=\boldsymbol{x})=\prod P(X_i=x_{ik_i})$，则称$X_1,\cdots,X_n$相互独立。

（2）对于连续型随机变量，若对任意的$\boldsymbol{x}$，有$f(\boldsymbol{x})=\prod f_i(x_i)$，则称$X_1,\cdots,X_n$相互独立。

（3）若n个随机变量$X_1,\cdots,X_n$相互独立，则随机变量$(X_1,\cdots,X_k)$和$(X_{k+1},\cdots,X_n)$相互独立，随机向量的函数$Y_1=g_1(X_1,\cdots,X_k)$和$Y_2=g_2(X_{k+1},\cdots,X_n)$也相互独立。

3.5 随机向量函数的分布

设$\boldsymbol{X}\sim F(\boldsymbol{x})$，$Z=g(\boldsymbol{X})$为一维随机变量，若$Z_i=g_i(\boldsymbol{X})$分别为一维随机变量，$\mathrm{A}\subset\mathbb{R}^{\mathrm{n}}$，则$P(\boldsymbol{Z}\in A)=\int f(\boldsymbol{x})d\boldsymbol{x}$，$F_{\boldsymbol{Z}}(\boldsymbol{z})=P(\boldsymbol{Z}\leq\boldsymbol{z})$。如果$\boldsymbol{g}$的逆映射$\boldsymbol{g}^{-1}$存在一阶连续偏导数，那么$p(\boldsymbol{z})=f(\boldsymbol{g}^{-1}(\boldsymbol{z}))|\boldsymbol{J}|I_D(\boldsymbol{z}),\boldsymbol{J}=\left|\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{x}}\right|$。

独立随机变量的和的概率密度函数为卷积：$f_1 * f_2(z) = \int_{-\infty}^{\infty} f_1(x) f_2(z-x)dx$。

如果$X_1, \cdots, X_n$相互独立且服从相同指数分布，那么 $\mathrm{Z} = X_1 + \cdots + X_n$的概率密度函数为$f_n(z) = \frac{\lambda^n}{n!} z^{n-1} exp\{-\lambda z\} I_{[0,\infty)}(z)$，称为参数是$n$，$\lambda$的$\Gamma$分布，记作$\mathrm{Z}\sim\mathrm{Ga}(n, \lambda)$。如果互相独立的两个同类型随机变量之和仍服从同一类型的分布，那么称此分布类型具有再生性。Γ分布对参数 n 有再生性。$Z = X - Y$的概率密度函数为$f_Z(z) = \frac{\lambda}{2} exp\{-\lambda|z|\}$，称为拉普拉斯分布，即独立指数分布的差是拉普拉斯分布。

正态分布$N(\mu, \sigma^2)$对参数μ, σ^2具有再生性。

独立随机变量商的分布$f_Z(z) = \int_{-\infty}^{\infty} |v| f_1(zv) f_2(v) dv$。

独立标准正态分布随机变量的商服从柯西分布$f_Z(z) = \frac{1}{\pi(1+z^2)}$。

$X_1, \cdots, X_n$相互独立，则$F_{max}(z) = F_1 \cdots F_n$，$F_{min}(z) = 1 - (1 - F_1)(1 - F_n)$。

3.6 扩展阅读：辛普森悖论

若有$\frac{a_1}{b_1} > \frac{a_2}{b_2}, \frac{c_1}{d_1} > \frac{c_2}{d_2}$，不能保证$\frac{a_1+c_1}{b_1+d_1} > \frac{a_2+c_2}{b_2+d_2}$。

4 随机变量的数字特征和极限定理

4.1 数学期望和中位数

4.1.1 数学期望和中位数

设X为离散型随机变量，如果$\sum |x_k| p_k < \infty$，那么称$E(X) = \sum x_k p_k$为随机变量X的数学期望，简称期望。

设$X\sim f(x)$，如果$\int |x| f(x)dx < \infty$，即 $\mathrm{E}(|x|) < \infty$，那么称 $\mathrm{E(X)} = \int x f(x)dx$ 为连续型随机变量的数学期望。

相同分布且相互独立，称为 i.i.d.。

泊松分布的数学期望为参数λ。

正态分布的数学期望为参数μ。

柯西分布不存在数学期望。

二项分布的数学期望为 np。

负二项分布数学期望为$\frac{r}{p}$。

数学期望的性质：

（1）线性性：$\mathrm{E}(X_1+\cdots+X_n)=\sum \mathrm{E}(X_i)$。

（2）若$X_1,\cdots,X_n$相互独立，$\mathrm{E}(X_1\cdots X_n)=\prod \mathrm{E}(X_k)$。

（3）设$\boldsymbol{X}\sim F_{\boldsymbol{X}}(\boldsymbol{x})$，$\boldsymbol{Y}=g(\boldsymbol{X})$为 m 维随机变量，有分布函数$F_{\boldsymbol{Y}}(\boldsymbol{y})$，则 $\mathrm{E}(\mathbf{Y})=\begin{cases}\sum g(\boldsymbol{x})P(\boldsymbol{X}=\boldsymbol{x}),\text{离散型}\\ \int g(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x},\text{连续型}\end{cases}$。

（4）若 $\mathrm{X}\geq \mathrm{Y}$，则$E(X)\geq E(Y)$。

4.1.2 条件数学期望（条件期望）

$E(Y|X=x)=\begin{cases}\sum y_i P(Y=y_i|X=x),\text{离散型}\\ \int yf_{Y|X}(y|x)dy,\text{连续型}\end{cases}$，$E(Y|x)$是 x 的函数，记为$h(x)$，如果不固定 X，则$E(Y|X)=h(X)$，取其期望，$E(E(Y|X))=E(Y)$，称为条件期望的平滑公式或全期望公式。

4.1.3 中位数和众数

设随机变量$X\sim F(x)$，若存在常数 m，满足$F(m)\geq 0.5\geq F(m-0)$，则称 m 为 X 的中位数。中位数不唯一，除非$f(m)>0$。

对于离散型随机变量，概率质量函数最大值对应的随机变量取为众数，记作m_d；对于连续型随机变量，概率密度函数最大的称为众数，记作m_d。众数不唯一，如果密度函数有唯一的极大值点，称该密度函数为单峰的。

设$0 < p < 1$，称Q_p是随机变量X的p分位数，即$P(X \le Q_p) \ge p, P(X \ge Q_p) \le 1-p$。密度函数大于零的连续型随机变量的分布分位数唯一，有$P(X \le Q_p) = p$。当p取0.25,0.5,0.75时称Q_p为四分位数，$IQR(X) = Q_{0.75} - Q_{0.25}$为内四分位距。当p恰好取百分比例，分位数称作百分位数。

正态分布的内四分位距约为1.35σ。

2 方差和矩

2.1 方差和标准差

随机变量在位置参数附近散布程度的数字特征中最重要的是方差。

设随机变量X平方可积，则$\sigma^2 = Var(X) = E(X-\mu)^2, \sigma = \sqrt{Var(X)}$分别称为X的方差和标准差。

方差的性质：

（1）$Var(X) = E(X^2) - \mu^2$，$E(X^2) \ge [E(X)]^2$

（2）$Var(c) = 0$

（3）$Var(cX) = c^2 Var(X), Var(X+c) = Var(X)$

（4）独立随机变量和的方差等于随机变量方差的和。

（5）$Var(X) = 0$当且仅当$P(X = c) = 1$，其中$c = E(X)$，此时称X退化到常数c。对任何常数c，由$Var(X) \le E(X-c)^2$，当且仅当$c = E(X)$等号成立。

随机变量的标准化：$Y = \frac{X-\mu}{\sigma}$，用来消除由于计算单位不同带来的影响。

泊松分布方差为λ。

二项分布方差为$np(1-p)$。

正态分布方差为σ^2。

马儿可夫不等式：若随机变量$Y \ge 0$，任取$\varepsilon > 0$，有$P(Y \ge \varepsilon) \le \frac{E(Y)}{\varepsilon}$。

切比雪夫不等式：任取$\varepsilon > 0$，有$P(|X - \mu| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon}$。

4.2.2 矩

设 X 为随机变量，满足$E(|X|^k) < \infty, k$为正整数，则$E(X-c)^k$称为 X 关于 c 的 k 阶矩。称α_k为$E(X^k)$为随机变量 X 的 k 阶原点矩，称$\mu_k = E(X - E(X))^k$为 X 的 k 阶中心距。

一阶原点矩α_1即为单位质量物体的质心，$Var(X)$是二阶中心距，为单位质量物体绕质心旋转的转动惯量。

设 *X~f(x)*，若 *f* 关于直线 *x=a* 对称，即$f(a+x) = f(a-x)$，那么$E(X) = a$，且$\mu_3 = 0$。如果$\mu_3 > 0$，f的图形最高点偏右，称右偏或正偏，否则称负偏或左偏。μ_3是数据分布与正态分布偏离程度的标志，标准化为$\beta_1 = \frac{\mu_3}{\sigma^3}$，称为随机变量 X 的偏度系数。

μ_4可以衡量密度函数在期望附近的陡峭程度，越集中越小，标准化$\beta_2 = \frac{\mu_4}{\sigma^4}$为随机变量 X 的峰度系数。

正态分布的$\beta_1 = 0, \beta_2 = 3$，有时也将$\frac{\mu_4}{\sigma^4} - 3$ 定义为峰度系数。

随机变量 X 的矩母函数或者矩生成函数(MGF)$M_X(s)$定义为$M_X(s) = E[e^{sX}]$，如果存在正常数 *a*，使得$M_X(s)$对所有的*[-a,a]*有限，那么称 X 的矩母函数$M_X(s)$存在。

$M_X(s) = E[e^{sX}] = \sum_0^\infty E(X^k)\frac{s^k}{k!}$，即$E(X^k) = \frac{d^k}{ds^k}M_X(s)\Big|_{s=0}$。

正态分布$M_X(s) = e^{\mu s + \frac{1}{2}\sigma^2 s^2}$。

假设存在正常数 c 使得随机变量 X 和 Y 的矩母函数对所有的*[-a,a]*均有限且相等，则它们的分布相同，即$F_X(t) = F_Y(t)$。

指数分布的矩母函数$M_X(s) = \frac{\lambda}{\lambda - s}, s < \lambda$。

如果$X_1,\cdots,X_n$相互独立，则$M_{X_1+\cdots+X_n}(s)=M_{X_1}(s)\cdots M_{X_n}(s)$。

4.2.3 协方差和相关系数

设随机变量 X 和 Y 均平方可积，则称$Cov(X,Y)=E[(X-E(X))(Y-E(Y))]$为随机变量 X 和 Y 的协方差。

协方差有性质：

（1）$Cov(X,Y)=Cov(Y,X)$。

（2）$Cov(X,Y)=E(XY)-E(X)E(Y)$。

（3）对任意实数a,b,c,d有$Cov(aX+b,cY+d)=acCov(X,Y),Cov(aX+bY,cX+dY)=acVar(X)+(ad+bc)Cov(X,Y)+bdVar(Y)$。随机变量的方差是协方差的特例，$Cov(aX+bY,cX+dY)=(a,b)\begin{pmatrix}Var(X) & Cov(X,Y)\\ Cov(X,Y) & Var(Y)\end{pmatrix}\binom{c}{d}$，即协方差是一个双线性函数。

（4）（a）若X,Y独立，则$Cov(X,Y)=0$；（b）$[Cov(X,Y)]^2\le Var(X)Var(Y)$，当且仅当$X,Y$有严格线性关系时等号成立，称为随机变量场合的柯西-施瓦茨不等式。

设随机变量 X 和 Y 均平方可积，则称$\rho_{X,Y}=Cov\left(\frac{X-E(X)}{\sqrt{Var(X)}},\frac{Y-E(Y)}{\sqrt{Var(Y)}}\right)=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$为$X,Y$的相关系数，不混淆情况下简记为$\rho$。

$|\rho|\le 1$，当且仅当随机变量有严格的线性关系时等号成立。如果$\rho<0$，称X,Y负相关；当$\rho>0$，称正相关；$\rho=0$，称为不相关。相关系数刻画了随机变量之间的线性相关关系，并不能反映有无函数关系。

引入线性空间$\mathcal{L}=\{r.v.X:E(X^2)<\infty\}$，在$\mathcal{L}$上，对任意$X,Y$，定义内积$\langle X,Y\rangle=Cov(X,Y)$，可以定义两个期望为零的$r.v.X,Y$的模$|\cdot|$和夹角余弦$cos\theta$：$|X|^2=\langle X,X\rangle=Var(X),cos\theta=\frac{\langle X,Y\rangle}{\sqrt{\langle X,X\rangle\langle Y,Y\rangle}}=\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$，相关系数就是空间$\mathcal{L}$中的余弦，

所以相关系数只能反映性关系，并不能反映函数关系。

相关系数只能刻画两个随机变量之间的线性关系程度。

如果$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，那么$Cov(X,Y)=\rho\sigma_1\sigma_2,\rho_{X,Y}=\rho$。

如果X,Y相互独立，则$\rho=0$，反之不必成立。

若$(X,Y)\sim U(|r|=1)$，那么X,Y不相关，也不相互独立。

对于任何非退化的$r.v.X,Y$存在方差，那么下面命题等价：（1）X,Y不相关；（2）$Cov(X,Y)=0$；（3）$E(XY)=E(X)E(Y)$；（4）$Var(X+Y)=Var(X)+Var(Y)$。

4.3 熵的基本概念

熵是$r.v.$最重要的数字特征之一，度量了随机变量中所含有的信息量的大小，它体现的是随机变量的不确定度程度，熵越大不确定度越大。

设X为离散型$r.v.$，分布律为$P(X=x_k)=p_k$，则熵定义为$H(X)=-\sum_{k=1}^{\infty}p_k log_2(p_k)$，如果$X$为连续性$r.v.$，有概率密度函数$f_X(x)$，那么有$H(X)=-\int_{-\infty}^{\infty}f_X(x)lnf_X(x)dx$。

熵只和随机变量取值的相对散布状况而非绝对状况有关，$H(X+c)=H(X)$。

如果X是0-1分布$r.v.$，$H(X)=-[plog_2(p)+(1-p)log_2(1-p)]$。

离散型$r.v.$的熵有性质：

（1）$H(X)\geq 0$；

（2）如果取有限个值的$r.v.X$的概率分布，那么有$H(X)\leq log_2(n)$，当且仅当X为离散均匀分布$r.v.$时等号成立。

设连续型$r.v.X$在实数集取值，$\max_X H(X),s.t.E(X)=\mu,Var(X)=\sigma^2$，那么有$X\sim N(\mu,\sigma^2)$。

设连续型$r.v.X$在正实数集取值，$\max_X H(X),s.t.E(X)=\frac{1}{\lambda}$，那么有$X\sim Exp(\lambda)$。

设连续型$r.v.X$在$[a,b]$取值，$\max_X H(X)$，那么有$X\sim U(a,b)$。

4.4 大数定律和中心极限定理

一般情况下，$r.v.$和的极限就是正态分布。习惯上把$r.v.$和的分布收敛于正态分布的那一类定理称为中心极限定理。

依概率收敛：设$X_1,\cdots,X_n,\cdots$是一随机变量序列，X为$r.v.$，如果$\forall\varepsilon>0$，有$\lim_{n\to\infty}P(|X_n-X|\geq\varepsilon)=0$，那么称随机变量序列$\{X_n\}$依概率收敛于随机变量$X$，记为$X_n\to X,in\ P$，或$X_n\xrightarrow{P}X$。

设$X_1,\cdots,X_n,\cdots$是一 *i.i.d* 的$r.v.$序列，记它们的期望和方差为μ,σ^2。记$S_n=X_1+\cdots+X_n$，那么$\forall\varepsilon>0$，$\lim_{n\to\infty}P(\left|\frac{S_n}{n}-\mu\right|\geq\varepsilon)=0$。

（伯努利大数定律）设$\{X_k\}$为独立的 0-1 分布$r.v.$序列，$P(X_k=1)=p$，那么$\frac{1}{n}\sum_{k=1}^{n}X_k\xrightarrow{P}p$。

强大数定律：随机变量序列的前 n 项和的平均几乎在每个样本点上收敛。

设$X_1,\cdots,X_n,\cdots$是一列实值$r.v.$，X为$r.v.$，F_n和F分别为随机变量X_n和X的分布函数，如果对F的所有连续点有$\lim_{n\to\infty}F_n(x)=F(x)$，那么称$\{F_n\}$弱收敛于$F$，也称$\{X_n\}$依分布收敛于$X$，记作$X_n\xrightarrow{\mathcal{L}}X$。

设$X_1,\cdots,X_n,\cdots$是一列实值$r.v.$，X为$r.v.$，那么：（1）若$X_n\xrightarrow{P}X$，则$X_n\xrightarrow{\mathcal{L}}X$；（2）若$X_n\xrightarrow{\mathcal{L}}c$，则$X_n\xrightarrow{P}c$。

林德伯格-莱维中心极限定理：设$X_1,\cdots,X_n,\cdots$是一 *i.i.d* 的$r.v.$序列，记它们的期望和方差为μ,σ^2。记$S_n=\sum_{i=1}^{n}X_i$，那么$\forall x$为实值，有$\lim_{n\to\infty}P\left(\frac{\sqrt{n}(S_n/n-\mu)}{\sigma}\leq x\right)=\Phi(x)$，即标准正态分布，即$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\xrightarrow{\mathcal{L}}N(0,1)$，或表示为$\lim_{n\to\infty}P\left(\frac{(S_n-E(S_n))}{\sqrt{Var(S_n)}}\leq x\right)=\Phi(x)$。

棣莫弗-拉普拉斯中心极限定理：设$X_1,\cdots,X_n,\cdots$是一 *i.i.d* 的$r.v.$序列，$S_n=$

$\sum_{i=1}^{n} X_i$，$X_i \sim B(1,p)$，那么$\forall x$为实值，有$\lim_{n\to\infty} P\left(\frac{(S_n-np)}{\sqrt{np(1-p)}} \le x\right) = \Phi(x)$。

注意到$S_n \sim B(n,p)$，当np较小时可用泊松分布逼近二项分布，当np较大时可用正态分布逼近二项分布，如果区间在两头须对连续性进行修正。

一般而言，当$n \ge 30$，中心极限定理的近似程度能满足需求。它指出大数定律$\overline{X}$依概率收敛于$E(x)$的速度为$1/\sqrt{n}$。

4.5 扩展阅读：数学期望的计算

利用期望的可加性求期望：将随机变量分解为若干期望容易计算的随机变量的和。

利用条件期望求期望。

5 统计学基本概念

5.1 统计学发展简史

5.2 基本概念

统计分析一般分为四个分支：描述性统计，对数据进行汇总，有助于以简洁的方式解释数据，不会丢失太多信息；探索性分析，侧重于使用图形方法深入研究数据并确定数据集中不同变量之间存在的关系，类似于数据可视化；预测分析，旨在基于一个或几个自变量来预测因变量，包含线性回归或分类；推断性统计，使用从总体中抽取的随机数据样本进行推断，得到关于总体的结论，主要工具是置信区间和假设检验。

5.2.1 总体

研究对象某个指标取值的全体以及取这些值的概率分布，称为统计总体，简称总体。

5.2.2 样本

从总计中按一定的方式抽取 n 个个体$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$，称为样本量为 n 的一个样本。

从总体中按一定方式抽取样本的行为称为抽样，目的是通过取得的样本对总体分布中某些未知要素做出推断。常用的抽样方法有不放回抽样和放回抽样。最常用的一种抽样方法叫做简单随机抽样，它满足要求：（1）代表性。总体中每一个个体都有同等机会被抽入样本。（2）独立性。样本中每一个个体取什么值不影响其他个体取什么值。由简单随机抽样获得的样本$(X_1,X_2,\cdots,X_n)$称为简单随机样本，也称为简单样本，也可简称样本。

设$X_1,X_2,\cdots,X_n$是从总体F中抽取的样本量为 n 的简单随机样本，则$X_1,X_2,\cdots,X_n$相互独立且有相同分布F。$(X_1,X_2,\cdots,X_n)$的联合分布函数为$\prod F(x_i)$，联合概率密度函数为$\prod f(x_i)$。

对样本来说，（1）不放回抽样具有代表性，但不具有独立性。实际抽样还有分层抽样、整体抽样和等距抽样等方式。（2）抽样方案实施前，样本视为随机变量，实施后，是一组数，称为样本的一个实现，也称样本值，该特点称为样本的“二重性”。

样本具有一定的概率分布，称为样本分布。

5.2.3 统计量

完全由样本$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)$决定的量称为统计量。

常见的统计量有：

样本均值$\overline{X}=\frac{1}{n}\sum X_i$，它反映了总体均值的信息。

样本方差$S^2=\frac{1}{n-1}\sum(X_i-\overline{X})^2$，它反映总体方差的信息，而$S$称为样本标准差，反映了总体标准差的信息。

样本矩$a_k=\frac{1}{n}\sum X_i^k$称为样本 k 阶原点矩，$m_k=\frac{1}{n}\sum(X_i-\overline{X})^k$称为样本 k 阶中心距，当样本为简单随机样本，样本矩依概率收敛到相应总体矩。

样本偏度系数$\widehat{\beta}_1=\frac{m_3}{m_2^{3/2}}$反映了总体偏度。

样本峰度性质$\widehat{\beta}_2=\frac{m_4}{m_2^2}-3$反映了总体峰度。

样本相关系数：设$(\boldsymbol{X},\boldsymbol{Y})=\big((X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)\big)$为从二维总体$F(X,Y)$中抽取的样本，称$\rho_n=\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum(X_i-\overline{X})^2}\sqrt{\sum(X_i-\overline{X})^2}}$为样本相关系数，也称皮尔逊相关系数，反映了总体相关系数信息。

次序统计量及其有关统计量：把样本按大小排列为$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)}$，则称$\big(X_{(1)},X_{(2)},\cdots,X_{(n)}\big)$称为次序统计量，它的任一部分也是次序统计量。样本中位数：$m_n=\begin{cases}X_{(\frac{n+1}{2})},n\text{为奇数}\\\frac{1}{2}[X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}],n\text{为偶数}\end{cases}$，它反映总体中位数的信息；极值：$X_{(1)}$和$X_{(n)}$称为样本的极小值和极大值，$X_{(n)}-X_{(1)}$称为极差；样本 p 分位数：$X_{([(n+1)p])}$，常见的是样本四分位数，以及样本四分位矩$X_{([3(n+1)/4])}-X_{([(n+1)/4])}$。

经验分布函数：$F_n(x)=\frac{\{X_1,X_2,\cdots,X_n\text{中}\ll x\text{的个数}\}}{n}=\frac{1}{n}\sum I_{(-\infty,x]}(X_i)$称为样本$X_1,X_2,\cdots,X_n$的经验分布函数，$F_n(x)\xrightarrow{P}F(x)$。记$Y_i=I_{(-\infty,x]}(X_i)$，那么$P(Y_i=1)=F(x),P(Y_i=0)=1-F(x)$，且$Y_1,Y_2,\cdots,Y_n i.i.d\sim B(1,F(x))$，故$nF(x)=\sum Y_i\sim b(n,F(x))$，因此有$P(F_n(x)=\frac{k}{n})=P(\sum Y_i=k)=\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$。

5.3 抽样分布

设$(X_1,X_2,\cdots,X_n)$为一个样本，统计量$T=T(X_1,X_2,\cdots,X_n)$的分布称为抽样分布。

样本均值和样本方差的分布与下面“统计三大分布”有密切联系。

1.χ^2分布

设样本$(X_1, X_2, \cdots, X_n)$为来自标准正态总体的一个简单随机样本，称$X = X_1^2 + \cdots + X_n^2$服从自由度为 n 的$\chi^2$分布，记为$X \sim \chi_n^2$。

χ_n^2分布的概率密度函数为$k_n(x) = \frac{1}{\Gamma(n/2)2^{n/2}} e^{-x/2} x^{(n-2)/2} I_{(0,\infty)}(x)$。当 n=1,2 时曲线单调下降，从 n=3 开始曲线有单峰。

如果$X \sim \chi_n^2$，记$P(X > c) = \alpha$，则$c = \chi_n^2(\alpha)$称为自由度为 n 的χ_n^2分布的上α分位数。

χ^2分布具有性质：（1）若$X \sim \chi_n^2$，则$E(X) = n, Var(X) = 2n$；（2）若$X \sim \chi_m^2, Y \sim \chi_n^2$，且$X + Y$独立，则$Z = X + Y \sim \chi_{m+n}^2$。

若记$Ga(\alpha, \lambda)$的概率密度函数为$p(x; \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I_{(0,\infty)}(x)$，那么自由度为 n 的$\chi^2$分布与$\Gamma$分布关系为$\xi = \sum_{i=1}^n X_i^2 \sim Ga(n/2, 1/2)$。也可以用此给出$\chi^2$分布定义，即若随机变量$\xi \sim Ga(n/2, 1/2)$，则称$\xi$为服从自由度为 n 的$\chi^2$分布，另一方面，若$Y \sim Ga(\alpha, \lambda)$，则$Z = 2\lambda Y \sim \chi_{2n}^2$。

2.t 分布

设$X \sim N(0,1), Y \sim \chi_n^2$，且$X, Y$相互独立，称$T = \frac{X}{\sqrt{Y/n}}$服从自由度为 n 的 t 分布，记为$T \sim t_n$。

t_n的概率密度函数$f_n(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$。

t 分布具有性质：（1）当 n=1 时，t_1分布就是柯西分布，$f_1(t) = \frac{1}{\pi(1+t^2)}$；（2）若$T \sim t_n$，则当 n ≥ 2 时，由对称性$E(T) = 0$，当 n ≥ 3 时，$Var(T) = \frac{n}{n-2}$；（3）当 n → ∞时，$t_n$分布趋于标准正态分布，即$\lim_{n\to\infty} f_n(t) = \varphi(t)$。

若$T \sim t_n$，记$P(T > c) = \alpha$，则$c = t_n(\alpha)$称为自由度为 n 的 t 分布的上α分位数。

3.F 分布

设$X \sim \chi_m^2, Y \sim \chi_n^2$且$X, Y$相互独立，称$F = \frac{X/m}{Y/n}$服从自由度为 m,n 的 F 分布，记为

$F \sim F_{m,n}$。

$f_{m,n}(x) = m^{m/2}n^{n/2}\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}x^{m/2-1}(mx+n)^{-(m+n)/2}I_{(0,\infty)}(x)$。

F 分布有性质：（1）若$Z \sim F_{m,n}$，则$1/Z \sim F_{n,m}$；（2）若$T \sim t_n$，则$T^2 \sim F_{1,n}$；（3）$F_{m,n}(1-\alpha) = 1/F_{n,m}(\alpha)$。

设随机变量$X_1, X_2, \cdots, X_n i.i.d \sim N(\mu, \sigma^2), c_1, c_2, \cdots, c_n$是不全为零的常数，则

（1）独立的正态随机变量线性组合服从正态分布，即$T = \sum c_k X_k \sim N\left(\mu\sum c_k, \sigma^2\sum c_k^2\right)$，特别，当$c_1 = c_2 = \cdots = c_n = \frac{1}{n}$时，有$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$。

（2）$S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$为样本方差，则$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$。

（3）$\overline{X}$与S^2相互独立。

（4）$\frac{\sqrt{n}(\overline{X}-\mu)}{S} \sim t_{n-1}$。

设$X_1, X_2, \cdots, X_m i.i.d \sim N(\mu_1, \sigma_1^2), Y_1, Y_2, \cdots, Y_n i.i.d \sim N(\mu_2, \sigma_2^2)$，且假定$\sigma_1^2 = \sigma_2^2 = \sigma^2$，样本$X_1, X_2, \cdots, X_m$与$Y_1, Y_2, \cdots, Y_n$相互独立，则$T = \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{S_T}\sqrt{\frac{mn}{n+m}} \sim t_{n+m-2}$，其中$(n+m-2)S_T^2 = (m-1)S_X^2 + (n-1)S_Y^2$。

设$X_1, X_2, \cdots, X_m i.i.d \sim N(\mu_1, \sigma_1^2), Y_1, Y_2, \cdots, Y_n i.i.d \sim N(\mu_2, \sigma_2^2)$，且合样本$X_1, X_2, \cdots, X_m$与$Y_1, Y_2, \cdots, Y_n$相互独立，则$F = \frac{S_X^2}{S_Y^2}\frac{\sigma_2^2}{\sigma_1^2} \sim F_{m-1,n-1}$。

设$X_1, X_2, \cdots, X_m i.i.d$服从指数分布$f(x,\lambda) = \lambda e^{-\lambda x}I_{(0,+\infty)}(x)$，则有$2\lambda n\overline{X} = 2\lambda\sum X_i \sim \chi^2_{2n}$。

5.4 扩展阅读 1：民意调查

统计理论上可以证明，抽样调查结果的可靠性不在于样本数量大小。

1.比例估计与样本量

在实际情况中，通常民意调查面对总体庞大，可以近似为无限总体，即总体可视为 0-1 分布，民众对某个具体问题的看法比例为p，它客观存在但未知。

以简单随机抽样的方式，得到样本量为n的样本（通常$n \ll N$），其中表示支持的人数为n_0，那么$p_0 = n_0/n$是p的一个良好的点估计（可以证明p_0既是矩估计，又是最大似然估计）。我们通过置信区间的方式刻画估计误差。当n比较大时，$\frac{\sqrt{n}(p_0-p)}{\sqrt{p(1-p)}}$服从标准正态分布，在置信水平 $1-\alpha$下，参数p的置信区间近似为$[p_0 - \frac{u_{\alpha/2}}{\sqrt{n}}\sqrt{p_0(1-p_0)}, p_0 + \frac{u_{\alpha/2}}{\sqrt{n}}\sqrt{p_0(1-p_0)}]$，其中$u_{\alpha/2}$表示标准正态分布的$\alpha/2$ 分位数。

在调查中，需要考虑误差$d = |p - p_0|$，给定误差边界d_0时，样本量$n \approx \frac{u_{\alpha/2}^2}{d_0^2}p_0(1-p_0)$，特别地，如果设定$d = 3\%$，置信水平为 95%，那么$n \approx 1067$。$n$只与置信水平、绝对误差有关，与总体数量无关。

5.5 扩展阅读 2：双盲对照试验

双盲对照试验是指在实验过程中，测试者和被测试者都不知道被测试者组别，旨在消除参与者有意识或无意识的个人偏差。这种试验方法可以防止研究结果被安慰剂效应或观测者偏差影响。

6 参数点估计

6.1 参数点估计的概念

一般地，设有一个统计总体，记为$f(x;\theta_1,\cdots,\theta_k)$，当总体分布为连续型分布时，$f$为概率密度函数，总体分布为离散型分布时，$f$为概率质量函数，同意约定$f(x;\theta_1,\cdots,\theta_k)$为总体分布，它包含$k$个未知参数。

参数估计问题一般指，在有了从总体抽取的样本$\boldsymbol{X} = (X_1,\cdots,X_k)$后，要用样本$\boldsymbol{X}$对参数$\theta_1,\cdots,\theta_k$或其函数$g(\theta) = g(\theta_1,\cdots,\theta_k)$进行估计。为了特定目的构造的统计量$[\hat{g}(\theta)](\boldsymbol{X})$称为$g(\theta)$的估计量，$[\hat{g}(\theta)](\boldsymbol{x})$称为$g(\theta)$的估计值。这种估计等于用一个点取估计另一个点，称为点估计。

点估计常用的构造方法有矩估计和最大似然估计（MLE）。

6.2 矩估计法

连续型总体分布的j阶原点矩和中心距分别为$\alpha_j = E(X^j) = \int x^j f(x;\theta_1,\cdots,\theta_k)dx$，$\mu_j = E(X-\alpha_1)^j = \int (x-\alpha_1)^j f(x;\theta_1,\cdots,\theta_k)dx$，离散型总体分布的$j$阶原点矩和中心距分别为$\alpha_j = \sum x_i^j f(x;\theta_1,\cdots,\theta_k)$，$\mu_j = \sum (x_i-\alpha_1)^j f(x;\theta_1,\cdots,\theta_k)$。这些矩依赖于参数$\theta_1,\cdots,\theta_k$。另一方面，由大数定律，样本矩依概率收敛到总体矩，所以$\alpha_j = \alpha_m(\theta_1,\cdots,\theta_k) \approx a_j = \frac{1}{n}\sum X_i^j$，$\mu_j = \mu_m(\theta_1,\cdots,\theta_k) \approx m_j = \frac{1}{n}\sum (X_i-\overline{X})^j$。将近似改为等式，选择适当的$k$个样本原点矩或样本中心距，就可以得到解$\hat{\theta}_i(X_1,\cdots,X_k)$，将$\hat{\theta}_i$作为$\theta_i$的估计。若要估计$g(\theta_1,\cdots,\theta_k)$，则用$\hat{g}(\theta_1,\cdots,\theta_k) = g(\hat{\theta}_1,\cdots,\hat{\theta}_k)$来估计。这样得到的估计称为矩估计。为了区别其他估计量，有时记作$\hat{\theta}_M$。矩估计量具有二重性。

正态总体$X\sim N(\mu,\sigma^2)$的参数矩估计$\hat{\mu} = \overline{X}, \hat{\sigma}^2 = m_2$，应用中一般用样本方差$S^2$来估计$\sigma^2$，即$\hat{\sigma}^2 = S^2$，$\hat{\sigma} = S$。

指数总体$X\sim Exp(\lambda)$的参数据估计$\hat{\lambda}_M = \overline{X}^{-1}$。

均匀总体$X\sim U(\theta_1,\theta_2)$的参数据估计$\hat{\theta}_1 = \overline{X} - \sqrt{3}S, \hat{\theta}_2 = \overline{X} + \sqrt{3}S$。

二项总体的参数矩估计$\hat{p} = \overline{X}/N$。

泊松总体的参数矩估计$\hat{\lambda} = \overline{X}$。

在合理的优劣准则下，低阶矩的估计优于高阶矩。

6.3 最大似然估计

设样本$\boldsymbol{X} = (X_1,\cdots,X_n)$有联合概率密度函数或联合概率质量函数$f(\boldsymbol{x};\theta) = f(\boldsymbol{x};\theta_1,\cdots,\theta_k)$，当固定$\boldsymbol{x}$时把$f(\boldsymbol{x};\theta)$看成$\theta$的函数，称为似然函数，记为$L(\theta;\boldsymbol{x})$或$L(\theta)$。

似然函数表示了观测到样本值$\boldsymbol{x}$的概率大小。反过来说，如果已经观测到样本值$\boldsymbol{x}$，那么$L(\theta;\boldsymbol{x})$应达到或接近最大值。

用似然程度最大的那个点$\theta^*=(\theta_1^*,\cdots,\theta_k^*)$，即满足条件$L(\theta_1^*,\cdots,\theta_k^*;x_1,\cdots,x_n)=\max\limits_{(\theta_1,\cdots,\theta_k)\epsilon\Theta}L(\theta_1,\cdots,\theta_k;x_1,\cdots,x_n)$的$(\theta_1^*,\cdots,\theta_k^*)$作为$(\theta_1,\cdots,\theta_k)$的估计值，这样的估计称为最大似然估计。若要估计$g(\theta_1,\cdots,\theta_k)$，则$g(\theta_1^*,\cdots,\theta_k^*)$就是它的最大似然估计。有时把最大似然估计量记作$\hat{\theta}_L$。

若似然函数是严格单调的，则似然函数的最大值在边界取到，从而得到最大似然估计。若似然函数是光滑的，且样本是简单随机样本，则似然函数是n个因子的乘积。可以先取自然对数，再求极值。若$\ell(\theta_1,\cdots,\theta_k)=ln\,L(\theta_1,\cdots,\theta_k;x_1,\cdots,x_n)$关于$(\theta_1,\cdots,\theta_k)$可微，那么通过似然方程$\frac{\partial\ell}{\partial\theta_i}=0$求驻点，进一步验证是否为最大值点。需要注意最大值点可能在边界取到，故需与边界比较。最大似然估计具有二重性。除少数情形，似然估计的显示表示一般不存在，需要用数值优化方法。

正态总体的最大似然估计$\hat{\mu}_L=\overline{X},\hat{\sigma}_L^2=m_2$。

指数总体的最大似然估计$\hat{\lambda}_L=\overline{X}^{-1}$。

均匀总体$U(0,\theta)$的最大似然估计$\hat{\theta}_L=maxX_i$。

二项总体的最大似然估计$\hat{p}_L=\overline{X}/N$。

柯西总体$f(x;\theta)=\frac{1}{\pi[1+(x-\theta)^2]}$的最大似然估计$\hat{\theta}_L$不好判断，由于$\theta$是总体分布的中位数，那么样本中位数可以作为它的估计。

一般地，对称中心都可以用样本中位数估计。

6.4 优良性准则

6.4.1 点估计的无偏性

设$\hat{g}(X_1,\cdots,X_n)$是$g(X_1,\cdots,X_n)$的一个估计量，称$E_\theta(\hat{g}(X_1,\cdots,X_n))-g(X_1,\cdots,X_n)$为估计量$\hat{g}$的偏差。若对任意可能的$(\theta_1,\cdots,\theta_k)\epsilon\Theta$，都有$E_\theta(\hat{g}(X_1,\cdots,X_n))=g(X_1,\cdots,X_n)$，则称$\hat{g}$是一个无偏估计量。

无论总体什么分布，只要期望存在，$\sum\omega_iX_i$都是期望的无偏估计。样本方差是σ^2的无偏估计。

一般而言，二阶以上的样本矩估计都不是总体矩的无偏估计，都要作修正。

6.4.2 最小方差无偏估计

优良性标准有均方误差$MSE_\theta(\hat{\theta})=E_\theta(\hat{\theta}(X_1,\cdots,X_n)-\theta)^2$，它兼顾了偏差和波动，也可以用平均绝对误差$MAD_\theta(\hat{\theta})=E_\theta(\left|\hat{\theta}(X_1,\cdots,X_n)-\theta\right|)$。

设$\hat{\theta}_1,\hat{\theta}_2$都是总体参数$\theta$的无偏估计，方差存在，若$Var_\theta(\hat{\theta}_1)\leq Var_\theta(\hat{\theta}_2)$，且至少存在一个$\theta$使等式成立，则称$\hat{\theta}_1$比$\hat{\theta}_2$更有效。

设$\hat{\theta}$是θ的一个无偏估计，若对θ的任一无偏估计$\hat{\theta}'$，都有$Var_\theta(\hat{\theta})\leq Var_\theta(\hat{\theta}')$，则称$\hat{\theta}$是$\theta$的一个最小方差无偏估计（MVUE）。

6.4.3 克拉默-拉奥方差下界

对$g(\theta)$的任一无偏估计$\hat{g}(\boldsymbol{X})$，在正则条件下有$Var_\theta(\hat{g}(\boldsymbol{X}))\geq(g'(\theta))^2[nI(\theta)]^{-1}$，其中$I(\theta)=\int_{-\infty}^{\infty}\left[\frac{\partial f(x;\theta)}{\partial\theta}\middle/f(x;\theta)\right]^2f(x;\theta)dx$称为费希尔信息函数，$g'(\theta)=\frac{\partial}{\partial\theta}\int\hat{g}(\boldsymbol{X})f(\boldsymbol{x}|\theta)d\boldsymbol{x}$。

6.5 点估计量的大样本理论

当样本量$n\to\infty$时点估计量的性质称为大样本性质。估计量在样本量固定时的性质称为小样本性质。它们的区别在于样本量是固定量还是动态量。

设$\hat{\theta}(X_1,\cdots,X_n)$是参数$\theta$的一个点估计，若当样本量$n\to\infty$时有$\hat{\theta}(X_1,\cdots,X_n)\xrightarrow{P}\theta$，则称$\hat{\theta}(X_1,\cdots,X_n)$是$\theta$的一个（弱）相合估计量。

直观上，相合性就是当样本量趋于无穷时，估计量只能在θ附近作越来越小的摆动，这称为相合性，也称为一致性。相合性是对一个估计量的基本要求。

设$\hat{\theta}(X_1,\cdots,X_n)$是参数$\theta$的一个点估计，设它的方差存在，记$Var_\theta(\hat{\theta}(X_1,\cdots,X_n))=\sigma_n^2(\theta)$，若当样本量$n\to\infty$时有$\lim\limits_{n\to\infty}P\left(\frac{\hat{\theta}(X_1,\cdots,X_n)}{\sigma_n(\theta)}\le x\right)=\Phi(x)$，则称估计量$\hat{\theta}(X_1,\cdots,X_n)$具有渐进正态性，常记为$\frac{\hat{\theta}(X_1,\cdots,X_n)}{\sigma_n(\theta)}\xrightarrow{\mathcal{L}}N(0,1)$。

估计量$\hat{\theta}(X_1,\cdots,X_n)$是否具有渐进正态性是其优良性的一个重要标志。渐进正态性提供了估计量$\hat{\theta}$的一个近似分布，利用它可以完成相关的统计推断。

在一般情况下，据估计和最大似然估计都有渐进正态性。

6.6 扩展阅读：德军坦克问题

7 区间估计

7.1 基本概念

设$(X_1,\cdots,X_n)$是从总体中抽取的一个简单随机样本，$\theta\in\Theta$为未知参数。$\hat{\theta}_1(X_1,\cdots,X_n)$和$\hat{\theta}_2(X_1,\cdots,X_n)$为两个统计量，给定一个小的正数$\alpha\in(0,1)$，若$P_\theta(\hat{\theta}_1(X_1,\cdots,X_n)\le\theta\le\hat{\theta}_2(X_1,\cdots,X_n))=1-\alpha$，则称区间$[\hat{\theta}_1,\hat{\theta}_2]$为参数$\theta$的置信区间估计，置信系数为$1-\alpha$。

如果知道P_θ不会小于$1-\alpha$，那么称$1-\alpha$为$[\hat{\theta}_1,\hat{\theta}_2]$的置信水平。置信系数是置信水平中的最大者。

设w_α是总体的上α分位数，v_α是下α分位数，那么$w_\alpha=v_{1-\alpha}$。

7.2 枢轴变量法

找区间估计的一般方法，称为枢轴变量法。设参数θ，找一个θ的良好点估计$T(\boldsymbol{X})$，一般为θ的最大似然估计；构造一个函数$S(T,\theta)$，不包含其他未知参数，称为枢轴变量，使得它的分布F已知；枢轴变量必须满足：$\forall a<b, a\le S(T,\theta)\le b$

能改写成$A \le \theta \le B$，A和B只能与$T(\boldsymbol{X})$，a和b有关，与θ无关；取分布F的上$\alpha/2$分位数$w_{\alpha/2}$和上$1-\alpha/2$分位数$w_{1-\alpha/2}$，那么有$P(w_{1-\alpha/2} \le S(T,\theta) \le w_{\alpha/2}) = 1-\alpha$。

正态总体均值μ的置信区间为$\bar{x} \pm d$，其中误差界限$d = \begin{cases} \frac{\sigma}{\sqrt{n}} u_{\alpha/2}, \sigma\text{已知} \\ \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2), \sigma\text{未知} \\ \frac{\hat{\sigma}}{\sqrt{n}} u_{\alpha/2}, n > 30, \sigma\text{未知，总体不必正态} \end{cases}$。

对于两个正态总体均值差的估计，如果方差之比在 1 附近，$\mu_2 - \mu_1 \in \tilde{y} - \tilde{x} \pm \sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}} u_{\alpha/2}$。两个正态总体方差比的区间估计为，$\frac{\sigma_1^2}{\sigma_2^2} \in \frac{s_1^2}{s_2^2}[F_{n-1,m-1}(1-\alpha/2), F_{n-1,m-1}(\alpha/2)]$。

7.3 大样本方法

7.3.1 比例 p 的区间估计

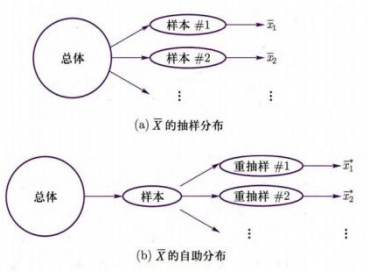
设事件 A 在每次实验中发生的概率为 p，作 n 次独立实验，以Y_n记事件 A 发生的次数，那么$P_p(-u_{\alpha/2} \le \frac{Y_n - np}{\sqrt{np(1-p)}} \le u_{\alpha/2}) \approx 1-\alpha$，于是可以得到 p 的置信区间，称为得分区间。

对于两个比例 p 的差，当 n 足够大时，$\frac{X_1/n_1 - X_2/n_2}{\sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)}} \sim N(0,1)$。

7.3.2 一般总体均值μ的置信区间

设$(X_1, \cdots, X_n)$是从总体X中抽取的一个简单随机样本，根据大数定律，$\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim N(0,1)$，所以$\mu$的置信系数近似为$1-\alpha$的置信区间为$\mu \in \bar{x} \pm \frac{s}{\sqrt{n}} u_{\alpha/2}$。

7.4 自助法置信区间



得到自助分布的步骤是，从$x_1,\cdots,x_n$中有放回地抽取一组样本量同样为 n 的样本，记为$x_1^*,\cdots,x_n^*$称为一个自助样本，基于它得到统计量的一个自助版本。重复操作 B 次，得到 B 个自助版本，它们用来近似统计量的自助分布。通常取 B=1000。

设参数θ的点估计为$\hat{\theta}$，记 B 个自助版本统计量$\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$的样本$\alpha/2$和$1-\alpha/2$分位数分别为$\hat{\theta}^*_{([(B+1)_{\alpha/2}])},\hat{\theta}^*_{([(B+1)_{1-\alpha/2}])}$，称$\left[2\hat{\theta}-\hat{\theta}^*_{([(B+1)(1-\alpha/2)])},2\hat{\theta}-\hat{\theta}^*_{([(B+1)\alpha/2])}\right]$为$\theta$的$1-\alpha$基本自助置信区间。

称$\left[\hat{\theta}-t^*_{1-\alpha/2}\widehat{se}^*_B(\hat{\theta}),\hat{\theta}-t^*_{\alpha/2}\widehat{se}^*_B(\hat{\theta})\right]$为$\theta$的$1-\alpha$自助 t 置信区间。它的算法是：由样本值得到$\hat{\theta}$；对样本值再抽样，得到$se(\hat{\theta})$的估计值$\widehat{se}(\hat{\theta})$以及 B 个自助版本统计量$\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$；对每个再抽样样本再抽样，得到$se(\hat{\theta}_b^*)$的估计值$\widehat{se}(\hat{\theta}_b^*)$，计算统计量$t^{(b)}=\frac{\hat{\theta}_b^*-\hat{\theta}}{se(\hat{\theta}_b^*)}$；找出$t^{(1)},\cdots,t^{(B)}$分布的样本分位数，得到$\theta$的$1-\alpha$自助 t 置信区间。

7.5 置信限

设$(X_1,\cdots,X_n)$是从总体中抽取的一个简单随机样本，$\bar{\theta}=\bar{\theta}(X_1,\cdots,X_n),\underline{\theta}=\underline{\theta}(X_1,\cdots,X_n)$为两个统计量。

（1）若对θ的一切可能取值，有$P_\theta(\bar{\theta}(X_1,\cdots,X_n)\geq\theta)=1-\alpha$，则称$\bar{\theta}$为$\theta$的一个置信系数为$1-\alpha$的置信上限。

（2）若对θ的一切可能取值，有$P_\theta(\bar{\theta}(X_1,\cdots,X_n)\leq\theta)=1-\alpha$，则称$\bar{\theta}$为$\theta$的一个置信系数为$1-\alpha$的置信下限。

7.6 扩展阅读：“足球赛会杀人”的真假

8 假设检验

8.1 问题的提法和基本概念

8.1.1 例子和问题解法

对统计总体（即总体分布）的性质所作的假设称为统计假设.使用样本对所作出的统计假设进行检查的方法和过程称为假设检验.如果总体分布的类型是已知的，要检验的假设是有关总体参数的某个取值范围，就称为参数假设检验问题.如果总体分布类型完全未知，我们称之为非参数假设检验问题.

8.1.2 假设检验中的几个基本概念

1.原假设和备择假设

在统计学中，我们把关于总体分布的某个特征的假设命题称为一个“假设”或“统计假设”，这个命题是否成立还需要通过样本来检验。一般我们把认为是正确的命题称为原假设H_0，换言之，将错误拒绝会带来很大后果的事情作为原假设。

当原假设不成立时的结论需要提前明确规定，称为备择假设，记为H_1或H_α，即拒绝原假设后可供选择的假设。

原假设也称为零假设，即没有变化，备择假设有时也称为“对立假设”，就是与原假设对立的意思. 这个词既可以指全体，也可以指一个或一些特殊情况.一般地，记Θ_0和Θ_1是参数空间Θ的两个不相交的非空子集，一个统计假设可以表示为$H_0: \theta \in \Theta_0 \leftrightarrow H_1: \theta \in \Theta_1$。

2.简单假设和复合假设

如果假设只有一个参数值，就称为简单假设，否则称为复合假设。如果参数为$\theta \in \Theta$，那么常见的假设有分为两点假设、双侧假设或双边假设、单侧假设或单边假设。

3.检验统计量、接受域、拒绝域和临界值

在检验一个假设时用到的统计量称为检验统计量.使原假设得到接受的样本

所在区域 A 称为该检验的接受域，而使原假设被拒绝的样本所在区域 D 称为拒绝域或否定域。在常见的假设检验中接受域和拒绝域通常可以简化为检验统计量$T(\boldsymbol{X})$所处区域。检验统计量区域变化时值为临界值，它可能不止一个。上述可以表示为$\Psi(X_1,\cdots,X_n)=\begin{cases}1,(X_1,\cdots,X_n)\in D\\0,(X_1,\cdots,X_n)\in A\end{cases}$，它称为对$H_0$和$H_1$的一个检验函数或法则。$\Psi=1$ 表示拒绝H_0，$\Psi=0$ 表示不能拒绝H_0。

8.1.3 功效函数

设总体为$F(x,\theta)$，H_0是关于参数θ的一个原假设。设Ψ是根据样本$(X_1,\cdots,X_n)$对假设所作的一个检验，称$\beta_\psi(\theta)=P_\theta$(在检验Ψ下$\mathrm{H}_0$被否定)为检验Ψ的功效函数。

设Ψ是假设的一个检验，$\beta_\psi(\theta)$为其功效函数，若$\beta_\psi(\theta)\le\alpha,\forall\theta\in H_0$，则称$\Psi$是$H_0$的一个水平$\alpha$的检验，或者说，检验$\Psi$的水平为$\alpha$。检验的水平是检验$\Psi$错误拒绝$H_0$所允许的最大概率。

8.1.4 两类错误

Ψ必犯两类错误之一：（1）当H_0成立时，检验法则Ψ拒绝了H_0，称为犯了第一类错误，也称“弃真错误”，记为$\alpha_{1\psi}(\theta)$，简记为α；（2）当H_0不成立时，检验法则Ψ没有拒绝H_0，称为犯了第二类错误，也称“存伪错误”，记为$\alpha_{2\psi}(\theta)$。

我们只能犯两种错误之一，它们与功效函数关系为：$\alpha_{1\psi}(\theta)=\begin{cases}\beta_\psi(\theta),\theta\in H_0\\0,\theta\in H_1\end{cases},\alpha_{2\psi}(\theta)=\begin{cases}0,\theta\in H_0\\1-\beta_\psi(\theta),\theta\in H_1\end{cases}$。奈曼提出先保证犯第一类错误的概率不超过某个给定的很小的数α,在此基础上使犯第二类错误的概率尽量小.如果仅仅考虑控制犯第一类错误的概率，而不涉及犯第二类错误概率所得到的检验，我们称为显著性检验，α也称显著性水平，此时原假设受到保护，接受原假设的结论是没有保障的，更恰当的结论应该是不能拒绝原假设。

显著性检验方法可以从直观出发来构造合理的检验法则。设定显著性水平为α，一般步骤如下：（1）求出未知参数θ的点估计$\hat{\theta}$；（2）寻找一个检验统计量$\mathrm{T}=\mathrm{T}(\hat{\theta},\theta_0)$，使之分布已知；（3）根据备择假设实际意义，寻找拒绝域；（4）根据犯弃真错误概率的最大值，即显著性水平，给出临界值方程，确定检验的拒绝域；（5）根据样本计算检验统计量的值，若落在拒绝域中，则可拒绝原假设,否则不能拒绝原假设.

8.2 正态总体参数检验

8.2.1 单个正态总体均值的检验

关于单个正态总体均值μ的假设检验问题，也称为一样本均值检验问题，常见的假设形式有左侧假设、右侧假设、双侧假设，称呼原因在于各自拒绝域的形式。

设$(X_1,\cdots,X_n)$是从该正态总体$N(\mu,\sigma^2)$中抽取的一个简单样本。

1.σ^2已知

对于左侧假设，合理检验是，Ψ:当$Z=\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma}<C$时拒绝原假设$H_0$，否则不能拒绝$H_0$。

要确定常数C，使检验Ψ有给定的水平α，为此$\beta_\Psi(\mu)=P_\mu(Z<C)=P_\mu(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}<C+\frac{\sqrt{n}(\mu_0-\mu)}{\sigma})$，即$\beta_\Psi(\mu)=\Phi(C+\frac{\sqrt{n}(\mu_0-\mu)}{\sigma})$，它需要满足$\beta_\Psi(\mu)\leq\alpha.\forall\mu>\mu_0$，故$C=u_{1-\alpha}=-u_\alpha$，于是$\beta_\Psi(\mu)=\Phi(\frac{\sqrt{n}(\mu_0-\mu)}{\sigma}-u_\alpha)$。

如果一个检验在控制犯弃真错误不超过α时还要求犯存伪错误概率小于β，即$\beta_\Psi(\mu)\geq1-\beta,\forall\mu<\mu_0$。放松要求后，$\beta_\Psi(\mu)\geq1-\beta,\forall\mu\leq\mu_1<\mu_0$，所以$n\geq\sigma^2\frac{(u_\alpha+u_\beta)^2}{(\mu_0-\mu_1)^2}$。

2.σ^2未知

检验统计量 $\mathrm{T}=\frac{\sqrt{n}(\overline{X}-\mu_0)}{S},\frac{\sqrt{n}(\overline{X}-\mu)}{S}\sim t_{n-1}$。

设立原假设和备择假设的两条原则：把已有的经过考验的结论或事实作为原假设H_0；把希望得到的结论放在备择假设H_1，希望能通过拒绝原假设得到结论。

8.2.2 两个正态总体均值差的检验

1.成组比较

设$(X_1,\cdots,X_m)$是从正态总体$N(\mu_1,\sigma^2)$中抽取的一个简单样本，$(Y_1,\cdots,Y_m)$是从正态总体$N(\mu_2,\sigma^2)$中抽取的一个简单样本，两组样本相互独立。假设时$\mu_1-\mu_2$与某常数δ的比较。

当σ^2已知时，检验统计量取$Z=\frac{\overline{X}-\overline{Y}-\delta}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}},\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}\sim N(0,1)$。当$\sigma^2$未知时，取$S_T=\sqrt{\frac{(m+1)S_1^2+(n-1)S_2^2}{m+n-2}}$为$\sigma$的优良点估计，检验统计量$\mathrm{T}=\sqrt{\frac{mn}{m+n}}\frac{\overline{X}-\overline{Y}-\delta}{S_T},\sqrt{\frac{mn}{m+n}}\frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_T}\sim t_{m+n-2}$。

2.成对比较

有一类实验中数据成对出现，称为成对比较。样本$\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$满足数据对之间通常可以认为是独立的，数据对内部的两个样本值通常不独立。对于两组数据均值是否有差异的检验问题，通常对数据对内部样本取差，构成虚拟总体及样本。假设时μ_z与某一常数 C 比较。若数据是连续数据，可以假设虚拟总体是正态分布，假设检验转为一样本正态检验问题。在大样本场合，由中心极限定理构造标准正态检验统计量也能得到拒绝域。这类问题的检验统称为成对 t 检验。

8.2.3 正态总体方差的检验

对于单个正态总体$N(\mu,\sigma^2)$的检验，检验统计量$\chi^2=\frac{(n-1)S^2}{\sigma_0^2},\frac{(n-1)S^2}{\sigma^2}\sim\chi_{n-1}^2$。

对于两个正态总体方差比的检验，设$(X_1,\cdots,X_m)$和$(Y_1,\cdots,Y_m)$分别是从$N(\mu_1,\sigma_1^2)$和$N(\mu_2,\sigma_2^2)$抽取的简单样本，两组样本之间相互独立。检验统计量$F=$

$S_1^2/bS_2^2, S_1^2/(\sigma_1^2-\sigma_2^2)S_2^2 \sim F_{m-1,n-1}$。

8.3 比例 p 的检验

设$(X_1,\cdots,X_n)$是 0-1 分布总体$B(1,p)$的一个样本，关于p仍有三类常见假设。

对于右侧假设，Ψ:当$\overline{X}>C$时拒绝H_0，否则不能拒绝H_0。由于$X=\sum_{i=1}^{n}X_i \sim B(n,p)$，因此检验等价于$\psi$:当$X>C$时拒绝$H_0$，否则不能拒绝$H_0$。记$X$的分布函数为$F_p(x)$，限制$C$为整数，对于给定的$\alpha$，功效函数$\beta_\psi(p)=1-F_p(C)$，求解$1-F_{p_0}(C)=\alpha$，一般没有精确解，会得到两个检验$C=C_0$或$C=C_0+1$，前者检验水平超过$\alpha$而后者不足。常采用的随机化检验是，当$X=C_0+1$时，从[0,1]中任取一个随机数$U$，若$U>\frac{1-\alpha-F_{p_0}(C_0)}{F_{p_0}(C_0+1)-F_{p_0}(C_0)}$则拒绝原假设，否则不能拒绝。

如果样本量 n 比较大，那么根据中心极限定理，$F_{p_0}(C)\approx\Phi(\frac{C-np_0}{\sqrt{np_0(1-p_0)}}), C\approx np_0+u_\alpha\sqrt{np_0(1-p_0)}$。

8.4 似然比检验

设样本$\boldsymbol{X}$有联合密度函数或联合概率质量函数$f(\boldsymbol{x};\theta)$，称统计量$LR(\boldsymbol{x})=\sup_{\theta\in\Theta}f(\boldsymbol{x};\theta)/\sup_{\theta\in\Theta_0}f(\boldsymbol{x};\theta)$为检验问题的似然比。$\phi$:当$LR(\boldsymbol{x})>c$时拒绝原假设$H_0$，否则不能拒绝$H_0$，称为检验的一个似然比检验，$c$由检验水平决定。

若$X_1,\cdots,X_n$时简单随机样本，在原假设成立下，当$n\to\infty$时，似然比有一个简单的极限分布。设Θ的维数为k，Θ_0的维数为s，若$k-s=t>0$，则对于检验问题在原假设成立下，当$n\to\infty$时，有$P(2ln\,LR(\boldsymbol{X})\le x)\to F_{\chi_t^2}(x)$，记为$2ln\,LR(\boldsymbol{X})\xrightarrow{\ell}\chi_t^2$。

8.5 p 值

p 值=P（得到当前样本下检验统计量的值或更极端值|原假设下）。取检验水平α，如果 p 值小于α，那么检验统计量的值落入拒绝域，即拒绝原假设，由此

可以得到关于 p 值的水平α检验法则。在检验统计量T在原假设下的分布难以得到时，可以用自助法计算 p 值。

置信区间和假设检验有明显关系。参数θ的双侧假设$H_0: \theta = \theta_0 \leftrightarrow H_1: \theta \neq \theta_0$的检验，接受域就是参数$\theta$的$1-\alpha$置信区间。类似的，置信系数为$1-\alpha$的单侧置信区间与显著性水平为$\alpha$的单侧检验问题的接受域也有对应关系。

8.6 扩展阅读：多重假设检验

在多重假设检验问题中，一个检验的显著性水平或犯第一类错误的概率不再表示所有检验综合在一起时候的显著性水平或犯第一类错误的概率.多重假设检验是一类综合校正多个假设下错误率的方法.

对一组原假设，至少错误拒绝一个原假设的概率称为总 I 型错误率。为了使得总 I 型错误率不超过α，可以使用帮费罗尼矫正方法：拒绝第i个假设，如果其p值满足$p_i \leq \alpha/m$。这种方法十分保守。

记R为拒绝的总数目，V为错误拒绝的原假设数目。本杰明尼-霍赫伯格过程（$BH(q)$）控制错误发现比例的期望，称为错误发现率$FDR = E\left(\frac{V}{R}|R > 0\right), P(R > 0) < q \leq \alpha$，其中$q$为允许的最大错误发现率上界。

$BH(q)$过程如下：（1）将每个假设检验的 p 值从小到大排序；（2）选择$R = max\left\{i: p_{(i)} \leq \frac{i}{m}q\right\}$；（3）拒绝 p 值小于$p_{(R)}$的假设。

$BH(q)$过程仅在个假设之间相互独立时能够控制FDR。当真实的原假设个数$m_0 = m$时，$FDR = P(V \geq 1) = FWER$，当$m_0 < m$时，$FWER \geq FDR$，所以控制$FWER$的过程也控制了FDR。

9 非参数假设检验

9.1 拟合优度检验

非参数假设检验的是样本所在总体是不是已知的理论分布，它用到的检验统计量一般近似为服从χ^2分布，也称皮尔逊χ^2检验。

9.1.1 理论分布完全已知且只取有限个值

假设一个以有限集$\{a_1,\cdots,a_k\}$为值域的总体X，从总体中抽取一组样本量为n的简单样本，其中有n_i次取值a_i。给定一个总体分布律（理论分布）为$P(X=a_i)=p_i$。检验问题：$H_0: P(X=a_i)=p_i \leftrightarrow H_1: \exists j\ s.t.\ P(X=a_j)\neq p_j$。

当n充分大，按照大数定律，$n_i/n\approx p_i$，称np_i为a_i这个类的理论值或者期望值，即在原假设成立时的期望值E，而将n_i称为观测值O。

统计量$Z=\sum\frac{(O-E)^2}{E}=\sum_{i=1}^{k}\frac{n_i^2}{np_i}-n$。如果原假设成立，当$n\to\infty$时，$Z$的分布趋于$\chi^2_{k-1}$。

当$Z>C$时拒绝H_0，否则接收H_0。C的选取根据给定的水平α，若近似认为Z的分布为χ^2_{k-1}，那么$C=\chi^2_{k-1}(\alpha)$。

假定根据一组数据算出$Z=Z_0$，定义$p(Z_0)$为原假设成立时，出现Z_0或更极端情况的概率，它近似为$p(Z_0)=1-F_{\chi^2_{k-1}}(Z_0)$，称为拟合优度。当$p(Z_0)<\alpha$时，拒绝$H_0$。

9.1.2 理论分布类型已知但含有有限个未知参数

若总体X取有限个值$\{a_1,\cdots,a_k\}$，但分布中含有r个未知参数$\theta_1,\cdots,\theta_r$，即原假设为$H_0: P(X=a_i)=p_i(\theta_1,\cdots,\theta_r)$。

记样本为$(X_1,\cdots,X_n)$，在原假设下，参数$\theta_1,\cdots,\theta_r$的最大似然估计为$\hat{\theta}_1,\cdots,\hat{\theta}_r$，从而$p_i$的最大似然估计为$\hat{p}_i$。用$n_i$表示样本中取$a_i$的个数，构建统计量$Z=\sum\frac{(O-\hat{E})^2}{\hat{E}}=\sum_{i=1}^{k}\frac{n_i^2}{n\hat{p}_i}-n$。若原假设成立，当$n\to\infty$时，$Z$的分布趋于$\chi^2_{k-r-1}$。

当总体X取无穷多个值，但分布中仅有有限个未知参数，此时原假设

$H_0: X \sim F_\theta(x)$。可以将总体的取值划分为k段，定义离散型随机变量$Y = a_i, x_{i-1} < X \leq x_i$，那么当原假设成立时，$Y$有分布$P(Y = a_i) = p_i(\theta_1, \cdots, \theta_r) = F_\theta(x_i) - F_\theta(x_{i-1})$，这样将无穷个值总体问题转换为有限个值总体问题。对总体的取值划分应使理论频数$np_i \geq 5$，否则将相邻子区间合并，此外划分点不能依赖于样本，必须事先规定。

9.1.3 列联表检验

理论分布类型已知，但有若干参数未知的检验常用于列联表检验。

列联表是一种按两个属性作双向分类的表。设 X:属性 A 的水平，Y:属性 B 的水平，记$p_{ij} = P(X = i, Y = j)$。$H_0: p_{ij} = p_{i\cdot} p_{\cdot j}$，共有$a + b - 2$ 个参数。取统计量$Z = \sum\sum \frac{(nn_{ij} - n_{i\cdot} n_{\cdot j})^2}{nn_{i\cdot} n_{\cdot j}}$，当$n \to \infty$时，$Z$的分布趋于$\chi^2_{(a-1)(b-1)}$。

齐一性检验指检验某一属性 A 各个水平对应的另一个属性 B 的分布是否全部相同，即$H_0: P(B = j | A = i)$关于i同分布。不同于χ^2检验，在齐一性检验问题中$n_{i\cdot}$提前给定，但可以证明当$n \to \infty$时，Z的分布趋于$\chi^2_{(a-1)(b-1)}$，即方法与χ^2检验相同。

9.2 威尔克科森秩和检验

如果两个总体分布分别为$F(x)$和$F(x - \theta)$，考虑$H_0: \theta = 0 \leftrightarrow H_0: \theta > 0$，$\theta$称为位置参数。在正态分布场合，就是两个总体均值差的检验问题。设从两个总体中抽取了样本$(X_1, \cdots, X_{n_1})$和$(Y_1, \cdots, Y_{n_2})$。

非参数检验中常用的是把数据按从小到大的顺序排列，根据秩的大小来检验，称为秩检验。将两个样本合在一起，指标重排为$(Z_1, \cdots, Z_n)$，设$Y_j = Z_{R_j}$，R_j称为Y_j在合样本中的秩。称$W_Y = \sum R_j$为Y样本在合样本中的秩和。

若$\theta > 0$，那么，拒绝域$W_Y > c_\alpha = inf\{c: P\{W_Y \geq c\} \leq \alpha\}$，或者$W_X < d_{1-\alpha} =$

$sup\{d: P\{W_X \leq d\} \leq \alpha\}$，这两种检验方法等价。

当两组样本的容量不等时，应取容量小的那一组样本对应的秩和作为检验统计量。

当样本量充分大时，$[W_Y - \frac{n_2(n+1)}{2}]/\sqrt{\frac{(n+1)n_1n_2}{12}}$近似为标准正态分布。

9.3 符号检验

符号检验是根据一对数据差的正负号来检验两个总体位置参数之间是否有差异.要处理的问题也可以说是一种成对比较问题。可以用成对比较的前提是一对数据的差有一个明确的值，第二是数据差构成的虚拟总体服从正态分布.

符号检验方法的优点在于一是没有任何分布的假定，二是对数据值的精确度要求不高.

对于n对数据，将结果表示为n个符号，其中m个不为零。记S_m^+为+号个数，则$S_m^+ \sim B(m, 1/2)$，通常不能给出临界值。常用办法是计算p值。

9.4 其他非参数检验概述

9.4.1 科尔莫戈罗夫检验

设$(X_1, \cdots, X_n)$是从总体F中抽取的一个样本，由此构造的经验分布函数记为$F_n(x)$。$H_0: F(x) = F_0(x)$，$F_0(x)$是一个完全已知的分布。取检验统计量$D_n = \sup\limits_{-\infty<x<\infty} |F_n(x) - F_0(x)|$，由于经验分布函数$F_n(x) \xrightarrow{P} F(x)$，所以原假设成立时$D_n$很小。

如果理论分布$F_0(x)$连续，那么在原假设H_0成立时，$\lim\limits_{n\to\infty} P(D_n \leq \frac{x}{\sqrt{n}}) = K(x) = \begin{cases} \sum (-1)^k exp\{-2k^2x^2\}, x > 0 \\ 0, x \leq 0 \end{cases}$。

在总体X为一维且理论分布完全已知的连续分布时，科尔莫戈罗夫检验优于χ^2检验。χ^2统计量的值依赖于区间划分，D_n没有依赖性，一般来说科尔莫戈罗

夫检验鉴别能力强。当总体X为多维时，χ^2检验处理方法与一维一样，极限分布的形式与维数无关，对于包含未知参数的理论分布，χ^2检验容易处理。

9.4.2 斯米尔诺夫检验

对于两个总体X和Y的连续分布是否相同的问题，设$(X_1,\cdots,X_m)$和$(Y_1,\cdots,Y_n)$分别是从两个总体X和Y中抽取的样本，F_{1m}和F_{2n}为对应的经验分布。定义统计量$D_{mn}^{+}=\sup\limits_{-\infty<x<\infty}(F_{1m}(x)-F_{2n}(x))$，$D_{mn}=\sup\limits_{-\infty<x<\infty}|F_{1m}(x)-F_{2n}(x)|$。

在原假设成立时，$\lim\limits_{\substack{n\to\infty\\m\to\infty}}P(\sqrt{\frac{mn}{m+n}}D_{mn}^{+}\le x)=(1-exp\{-2x^2\})I_{(x>0)}$，$\lim\limits_{\substack{n\to\infty\\m\to\infty}}P(\sqrt{\frac{mn}{m+n}}D_{mn}\le x)=K(x)$。

当检验问题$H_0:F_1(x)=F_2(x)$，取D_{mn}。当检验问题$H_0:F_1(x)\le F_2(x)$，取D_{mn}^{+}。

9.5 扩展阅读：正态性检验

正态性检验问题表述为：H_0:总体服从正态分布↔H_1:总体不服从正态分布。

9.5.1 定性检验法

1.直方图

如果样本数据来自一正态总体且样本容量不太小，那么它应该满足“中间大、两头小”的特点，直方图是单峰且对称的。此外还可以计算样本均值和方差，从而显示正态分布图像来与直方图轮廓进行比较。

2.Q-Q 图

Q-Q 图中的 Q 指分位数，原理是：如果样本数据来自一正态总体，那么嘉定的正态总体的分位数与样本数据的经验分位数基本一致。Q-Q 图实际上是一个散点图，纵坐标为实际样本数据的分位数，横坐标为假定正态总体的分位数。判断正态性原则就是：如果图中的点大致呈左下至右上的直线，那么可以认为是正态

的。

9.5.2 定量检验法

1.偏度和峰度

总体X的偏度和峰度即其标准化后的三阶和四阶中心距，表示为$\beta_1 = E\left[\frac{X-E(X)}{\sqrt{Var(X)}}\right]^3, \beta_2 = E\left[\frac{X-E(X)}{\sqrt{Var(X)}}\right]^4$。偏度反映分布形状是否对称；峰度反映分布形状的尖锐程度，正态分布$\beta_1 = 0, \beta_2 = 3$。

设$x_1, \cdots, x_n$为来自某总体X的一组简单随机样本，样本k阶中心距$m_k = \frac{1}{n}\sum(x_i - \bar{x})^k$，则$\widehat{\beta}_1 = \frac{m_3}{m_2^{3/2}}, \widehat{\beta}_2 = \frac{m_4}{m_2^2}$。原假设成立时，$\sqrt{\frac{n}{6}}\widehat{\beta}_1, \sqrt{\frac{n}{24}}(\widehat{\beta}_2 - 3)$的极限分布服从标准正态分布。构造检验统计量$\widehat{\beta} = \frac{n}{6}[\widehat{\beta}_1^2 + \frac{1}{4}(\widehat{\beta}_2 - 3)^2]$，当$n \to \infty$时，服从$\chi_2^2$。

2.拟合优度

SW 检验的思想基于次序统计量，将家样本按从小到大排列以构造统计量$W = \frac{(\sum a_i x_{(i)})^2}{\sum(x_i - \bar{x})^2}$，其中$(a_1, \cdots, a_n) = \frac{\boldsymbol{m^T V^{-1}}}{\|\boldsymbol{V^{-1} m^T}\|}$，$\boldsymbol{m}$为$n$个 i.i.d 标准正态分布随机变量的次序期望，$\boldsymbol{V}$是它们的协方差矩阵。这种检验方法适合样本比较小的情形。

10 相关分析和回归分析

研究相关关系，一种方法是相关分析，致力于寻找一些数量指标以刻画相关程度，一种是回归分析，着重寻求变量之间近似的函数关系，此外还有方差分析，它考虑一个或一些变量对某一特定变量的影响大小。

10.1 相关分析

按取值满足的运算性质将数据分为定性数据和定量数据。定性数据又称分类数据，用数字表示类别属性，定性数据仅能计数。定量数据有大小的区别，有序数据有序但数据的差没有意义，间隔数据可以用数值反映量的差别，比例数据有

绝对的零点，可以进行所有的算术运算。

10.1.1 比例数据的相关系数

1.皮尔逊相关系数

对变量X,Y观测到n对数据$(x_1,y_1),\cdots,(x_n,y_n)$，则变量$X,Y$的皮尔逊相关系数$r=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{[n\sum x_i^2-(\sum x_i)^2][n\sum y_i^2-(\sum y_i)^2]}}$，记$S_{xx}=\sum(x_i-\bar{x})^2,S_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y}),S_{yy}=\sum(y_i-\bar{y})^2$，那么$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$。

如果把$(x_1,\cdots,x_n)$和$(y_1,\cdots,y_n)$视为两个向量，那么r就是这两个向量中心化后夹角的余弦。r不能反映因果关系，它是X,Y相关系数ρ的估计。

r^2称为决定系数或判刑系数，它反映了由于相关关系，y的变化可以由x的变化来解释的百分比。

2.可以转换为线性关系的非线性相关的刻画

10.1.2 有序数据的相关系数：肯德尔相关系数

设变量X,Y为有序数据，取值$1,\cdots,I$和$1,\cdots,J$。肯德尔相关系数$r_\tau(X,Y)=\binom{n}{2}^{-1}\sum_{i<j}sgn[(X_i-X_j)(Y_i-Y_j)]$。定义$X,Y$相关系数$\rho_\tau(X,Y)=P((X_i-X_j)(Y_i-Y_j)>0)-P((X_i-X_j)(Y_i-Y_j)<0)$，$r_\tau$为其估计。

分母改为$\sum_{i<j}\left|sgn[(X_i-X_j)(Y_i-Y_j)]\right|$，称为修正的肯德尔相关系数，或集中系数。

肯德尔相关系数也可以用来描述两个比例数据的相关性，但精度不如皮尔逊相关系数。

10.2 回归分析

10.2.1 一元线性回归模型

考虑的自变量在一定程度上影响了因变量的取值，在模型中，将其余自变量

的影响看作随机误差，因此因变量被视为随机变量，自变量可以是随机变量，也可以是确定的量，这取决于是否能够控制自变量的取值。

设想因变量 Y 由两部分组成，一部分由自变量$x_1,\cdots,x_p$的影响所致，表示为$f(x_1,\cdots,x_p)$，另一部分是随机误差，记为e，那么建立模型$Y=f(x_1,\cdots,x_p)+e$，要求$E(e)=0$，在给定自变量$(x_1,\cdots,x_p)$下，$E(Y|x_1,\cdots,x_p)=f(x_1,\cdots,x_p)$，称$f(x_1,\cdots,x_p)$为 Y 关于$x_1,\cdots,x_p$的回归方程。函数$f$最简单的形式是$x_1,\cdots,x_p$的线性函数，即$f(x_1,\cdots,x_p)=a+\boldsymbol{\beta}^T\boldsymbol{x},\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^T$。

对于只有一个自变量的情况，$Y=a+\beta_1x+e,E(Y|x)=a+\beta_1x$，此时$y=a+\beta_1x$称为$y$关于$x$的回归直线。设独立观测出 n 对数据$(x_i,y_i)$，作它的散点图，如果$y$关于$x$的回归曲线是直线，那么观测数据应有$y_i=a+\beta_1x_i+e_i,E(e_i)=0$。

拟合值与观测值之间距离的平方最小时拟合最好，即求a,β_1，使得$g(a,\beta_1)=min\sum(y_i-a-\beta_1x_i)^2$，求取极值，得$\widehat{\beta}_1=\frac{S_{xy}}{S_{xx}},\hat{a}=\bar{y}-\widehat{\beta}_1\bar{x}$，称为$a,\beta_1$的最小二乘估计。如果误差$e_1,\cdots,e_pi.i.d{\sim}N(0,\sigma^2)$，那么上式微最大似然估计。

10.2.2 回归系数最小二乘估计的几何意义

记$\boldsymbol{y}=(y_1,\cdots,y_n)^T,\boldsymbol{1}=(1,\cdots,1)^T,\boldsymbol{x}=(x_1,\cdots,x_n)^T,\boldsymbol{e}=(e_1,\cdots,e_n)^T$，那么$\boldsymbol{y}=a\boldsymbol{1}+\beta_1\boldsymbol{x}+\boldsymbol{e}=(\boldsymbol{1},\boldsymbol{x})\begin{pmatrix}a\\\beta_1\end{pmatrix}+\boldsymbol{e}=\boldsymbol{X}\eta+\boldsymbol{e}$，即两个 n 维向量的线性组合生成了一个线性空间，记为$\mathcal{S}_2$，$g(a,\beta_1)=min\|\boldsymbol{y}-a\boldsymbol{1}-\beta_1\boldsymbol{x}\|^2$，最小值时为$\boldsymbol{y}$在$\mathcal{S}_2$上的投影。

10.2.3 误差方差的估计

记$\hat{y}_i=\widehat{\beta}_0+\widehat{\beta}_1x_i^*,\widehat{\beta}_0=\hat{a}+\widehat{\beta}_1\bar{x}=\bar{y},x_i^*=x_i-\bar{x},\hat{\boldsymbol{y}}=(\hat{y}_1,\cdots,\hat{y}_n)^T$，那么$\hat{\boldsymbol{y}}$称为$\boldsymbol{y}$的回归值，残差向量$\hat{\boldsymbol{e}}=(\hat{e}_1,\cdots,\hat{e}_n)^T$，$\hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}}$的自由度为 n-2，误差方差的估计$\widehat{\sigma}^2=\frac{1}{n-2}\sum\hat{e}_i^2$。称$TSS=S_{yy}$为总平方和，$ESS=\sum(y_i-\hat{y}_i)^2$为残差平方和，$RSS=$

$\sum(\hat{y}_i-\bar{y})^2$称为回归平方和，那么$TSS=RSS+ESS$，称为平方和分解公式。在线性回归关系下，决定系数$r^2=\frac{RSS}{TSS}$，在计算机中记为$R^2$。注意到$s_y^2=TSS/(n-1)$，所以$\hat{\sigma}=\sqrt{\frac{n-1}{n-2}(1-r^2)s_y}$。

10.2.4 回归系数最小二乘法估计的性质

在回归方程中，通常对误差有两种假定：（1）高斯-马儿可夫假定，$E(e_i)=0$，$e_1,\cdots,e_n$不相关且$Var(e_i)=\sigma^2$；（2）正态假定：$e_1,\cdots,e_n$不相关且$e_i\sim N(0,\sigma^2)$。

在高斯-马儿可夫假定下，$E(\hat{\beta}_0)=\beta_0,E(\hat{\beta}_1)=\beta_1,Var(\hat{\beta}_0)=\sigma^2/n,Var(\hat{\beta}_1)=\sigma^2/S_{xx},Cov(\hat{\beta}_0,\hat{\beta}_1)=0,E(\hat{\sigma}^2)=\sigma^2,(\hat{\beta}_0,\hat{\beta}_1)\perp\hat{e}$。

在正态假定下，$\hat{\beta}_0\sim N(\beta_0,\sigma^2/n),\hat{\beta}_1\sim N(\beta_1,\sigma^2/S_{xx}),(n-2)\hat{\sigma}^2\sim\sigma^2\chi^2_{n-2},\frac{\sqrt{S_{xx}}}{\hat{\sigma}}(\hat{\beta}_1-\beta_1)\sim t_{n-2},\frac{\hat{a}_0-a_0}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}}\sim t_{n-2}$。

10.2.5 回归系数的检验和区间估计

1.回归系数β_1的检验和区间估计

$H_0:\beta_1=c\leftrightarrow H_1:\beta_1\neq c$，在误差服从独立正态分布的假定下，$\frac{\sqrt{S_{xx}}}{\hat{\sigma}}(\hat{\beta}_1-\beta_1)\sim t_{n-2}$。

2.回归函数和因变量 Y 的预测

因变量y平均值$m(x)=a+\beta_1x$称为回归函数。可以用$\hat{y}=\hat{a}+\hat{\beta}_1x$作为$m(x)$的点估计。在正态假定下，$\hat{y}\sim N(a+\beta_1x,(1/n+(x-\bar{x})^2/S_{xx})\sigma^2)$，而$y\sim N(a+\beta_1x,(1+1/n+(x-\bar{x})^2/S_{xx})\sigma^2)$。

10.2.6 可化为线性函数的非线性回归

10.2.7 多元线性回归模型

$y=a_0+\beta_1x_1+\cdots+\beta_px_p+e,E(y|x_1,\cdots,x_p)=a_0+\beta_1x_1+\cdots+\beta_px_p$，这个模型称为 p 元线性回归模型。当$f$是多元可微函数时可以作泰勒展开，在一定范

围内用多项式近似，视为多元线性回归模型。

$$\widehat{\beta}_0 = \bar{y}, \widehat{\beta} = (\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{y}$$

在高斯-马儿可夫假定下，$E(\widehat{\beta}_0) = \beta_0, E(\widehat{\beta}_i) = \beta_i, Var(\widehat{\beta}_0) = \sigma^2/n, Var(\widehat{\beta}) = \sigma^2/\boldsymbol{S_{xx}}, Cov(\widehat{\beta}_0, \widehat{\beta}_i) = 0, E(\widehat{\sigma}^2) = \sigma^2, (\widehat{\beta}_0, \widehat{\beta}) \perp \hat{e}$。

在正态假定下，$\widehat{\beta}_0 \sim N(\beta_0, \sigma^2/n), \widehat{\beta} \sim N_p(\beta_1, \sigma^2\boldsymbol{S_{xx}^{-1}})$，若记$d(i)$为方阵$\boldsymbol{S_{xx}^{-1}}$的对角元，那么$\widehat{\beta}_i \sim N_p(\beta_1, d(i)\sigma^2), (n-p-1)\widehat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}, \frac{1}{\widehat{\sigma}\sqrt{d(i)}}(\widehat{\beta}_1 - \beta_1) \sim t_{n-p-1}, (\frac{1}{n} + \bar{\boldsymbol{x}}^T\boldsymbol{S_{xx}^{-1}}\bar{\boldsymbol{x}})^{-1/2}\frac{\hat{a}_0 - a_0}{\widehat{\sigma}} \sim t_{n-p-1}, \widehat{m} \sim N(m(\boldsymbol{x}), (1/n + (\boldsymbol{x} - \bar{\boldsymbol{x}})^T\boldsymbol{S_{xx}^{-1}}(\boldsymbol{x} - \bar{\boldsymbol{x}}))\sigma^2)$。

10.3 多元回归中自变量的选择和模型诊断简述

10.3.1 变量选择的准则和方法

建立多元线性回归方程后，首先要检查回归方程是否有效. 有效反映在两个方面，一是自变量对因变量的影响是否可用自变量的线性组合来表达，即回归是线性的还是非线性的.统计上称为回归模型的非线性性检验。二是指因变量能否用选定的这些变量的线性组合来表达。

记$R^2 = \frac{\hat{\boldsymbol{y}}^{*T}\hat{\boldsymbol{y}}^{*}}{\boldsymbol{y}^{*T}\boldsymbol{y}^{*}} = \frac{RSS}{TSS}$，$R$称为复相关系数，$R^2$反映了线性回归好坏的程度。

另一种方法是在正态假定下作回归方程的显著性检验，即要检验$H_0: \beta_i = 0 \leftrightarrow H_1: \beta_i \neq 0$，$\hat{\boldsymbol{y}}^{*T}\hat{\boldsymbol{y}}^{*} \sim \sigma^2\chi^2_p, \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} \sim \sigma^2\chi^2_{n-p-1}$，$F = \frac{RSS/p}{ESS/(n-p-1)} \sim F_{p,n-p-1}$。

对进入模型的自变量要进行筛选，剔除对因变量影响不大的自变量，即抓主要矛盾.这称为自变量的选择，我们称包含 p 个自变量的模型为全模型，剔除一些自变量后的模型称为选模型。剔除自变量的准则，大多数是从残差平方和出发建立的。

（1）RMS_k（修正的残差平方和准则）。记$RMS_k = \frac{ESS_k}{n-k}$，选择$k$，使上式最

小。

（2）C_p准则。令$C_p = \frac{ESS_k}{TSS} - (n - 2k)$最小。

（3）AIC 准则。令$AIC = nln(\frac{ESS_k}{n}) + 2k$最小。

（4）BIC 准则。令$BIC = nln(\frac{ESS_k}{n}) + kln(n)$最小。

在计算机上进行变量选择时，需要计算ESS_k和RSS_k，有最优子集回归法、逐步回归法、向前选择法、向后选择法。

10.3.2 回归诊断

如果残差估计的散点图中，散点没有趋势，大体分布在以 x 轴为中心的两条水平线之间，说明回归时有效的。

10.4 扩展阅读：相关与因果

10.5 附录