第一节 · 概率

司继春

上海对外经贸大学统计与信息学院

首先我们回顾一下**概率空间**（**probability space**）的定义。我们知道，在概率论中，概率空间为一个三元组：$(\Omega, \mathscr{F}, \mathscr{P})$，其中 Ω 为样本空间，$\mathscr{F}$ 为所有事件的集合，$\mathscr{P}$ 为概率测度。下面我们分别探讨三元组的每个元素。

1 样本空间

样本空间（**sample space**）Ω 为我们关心的随机试验的所有结果的集合，而 Ω 中的元素 $\omega \in \Omega$ 称之为**样本点**（**sample point**）。例如：

1. 随机从一堆扑克牌中抽取一张扑克，其花色的样本空间为：

$$\Omega_1 = \{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}$$

   而样本点为 $\heartsuit$、$\spadesuit$、$\clubsuit$ 以及 $\diamondsuit$。

2. 某银行一天所接待的所有客户数，其样本空间为 $\Omega_2 = \mathbb{N}$，样本点为自然数。

3. 随机从人群中抽取一个人，其身高的样本空间为 $\Omega_3 = \mathbb{R}^+$，而样本点为正实数。

注意在上面的三个例子中，样本空间有细微差别。Ω_1 的元素个数为有限个，而 Ω_2 和 Ω_3 的元素个数有无穷多个。其中，Ω_1 与 Ω_2 都可以与自然数 $\mathbb{Z}$ 或者 $\mathbb{Z}$ 的子集建立起一一对应的关系，我们称之为**可数集**（**countable set**），而像 Ω_3 这样不能与自然数 $\mathbb{Z}$ 或者其子集建立起一一对应关系的，我们称之为**不可数集**（**uncountable set**）。

在概率论中，显示的定义样本空间是非常重要的，同一个问题，如果定义的样本空间不同，可能会得到完全不同的结果。

例 1.（**贝特朗悖论**）考虑一个内接于圆的等边三角形。若随机选方圆上的个弦，则此弦的长度比三角形的边较长的概率是多少?

根据不同的假设，有三种解法：

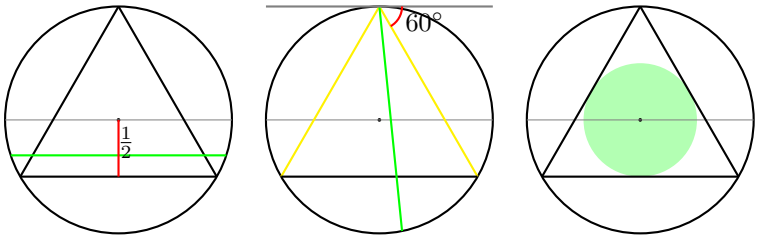


图 1: 贝特朗悖论

解法一，如图所示，在垂直于三角形任意一边的直径上随机取一个点，并通过该点做一条垂直于该直径的弦，在该点位于半径中点的时候弦长度等于三角形的边长度，所以概率为 $\frac{1}{2}$。

解法二，如图所示，通过三角形任意一个顶点做圆的切线，因为等边三角形内角为 60°，所以左边右边的角都是 60°。由该顶点做一条弦，弦的另一端在圆上任意一点。由图可知弦与切线成 60° 角和 120° 角之间的时候弦长度大于三角形边长，所以概率为 $\frac{1}{3}$。

解法三，如图所示，当弦的中点在阴影标记的圆内时，弦的长度大于三角形的边长，而大圆的弦中点一定在圆内，大圆的面积是 πr^2，小圆的面积是 $\frac{1}{4}\pi r^2$，所以概率为 $\frac{1}{4}$。

同一个问题为什么会得到三种不同的答案呢？原因在于，圆内“取弦”时规定尚不够具体，不同的“等可能性假定”导致了不同的样本空间：第一种解法中，假设弦的中点在直径上均匀分布；第二种解法中，假设弦的另一端在圆周上均匀分布；第三种解法中，假设弦的中点在大圆内均匀分布。

因而在定义概率时，第一步必须明确地指出样本空间是什么。

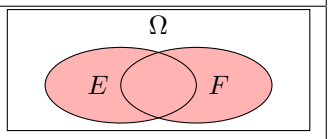
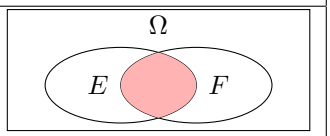
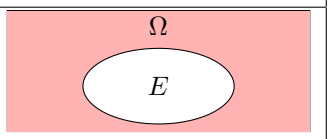
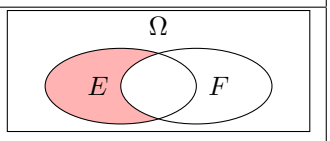
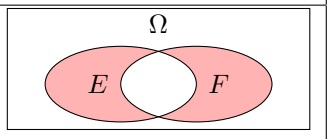
2 事件

我们称样本空间 Ω 的子集（包含 Ω 本身）为**事件**（**event**）。如果 A 为 Ω 的一个子集，如果随机试验的结果为 A 中的一个样本点，我们称之为发生了事件 A。在通常情况下，当我们称概率时，指的是事件发生的概率。

我们首先回忆集合的运算与性质。表1列出了常见集合运算的定义。对于表1中列出的集合运算，我们有以下的运算法则：

1. 交换律：$A \cup B = B \cup A$，$A \cap B = B \cap A$
2. 结合律：$A \cup (B \cup C) = (A \cup B) \cup C$，$A \cap (B \cap C) = (A \cap B) \cap C$
3. 分配率：$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$，$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
4. 德摩根律：$(A \cup B)^c = A^c \cap B^c$，$(A \cap B)^c = A^c \cup B^c$

表 1: 集合的运算

运算	符号	定义	图示
并	$E \cup F$	$\{\omega \in E \text{ OR} \omega \in F\}$	
交	$E \cap F$	$\{\omega \in E \text{ AND} \omega \in F\}$	
补集	E^c	$\{\omega \in \Omega, \omega \notin E\}$	
差	$E \backslash F$	$E \cap F^c$	
对称差	$E \triangle F$	$(E \backslash F) \cup (F \backslash E)$	

此外，以上运算法则容易推广至可数个集合的运算，比如德摩根律：

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

以及：

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$$

如果两个事件 A 和 B 满足 $A \cap B = \emptyset$，我们称之为**互斥事件**（disjoint or exclusive）。如果对于一系列事件 $A_1, A_2 \ldots$，对于任意 i 和 j，有 $A_i \cap A_j = \emptyset$，则称之为**两两互斥事件**。如果 $A_1, A_2 \ldots$ 为两两互斥事件，且 $\cup_{i=0}^{\infty} A_i = \Omega$，则 $A_1, A_2 \ldots$ 为样本空间的一个**划分**（**partition**）。

对于一个一般的样本空间，有数量巨大的子集。我们希望挑出那些我们需要研究的子集，同时剔除那些性质不是十分良好的子集，这就诞生了 σ-**代数**的概念。

定义 1. （σ-**代数**）如果样本空间 Ω 的一系列子集的集合 $\mathscr{F}$ 满足：

1. $\emptyset \in \mathscr{F}$
2. 若 $A \in \mathscr{F}$，则 $A^c \in \mathscr{F}$

3. 若 $A_1, A_2 \ldots \in \mathscr{F}$，则 $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$

我们称 $\mathscr{F}$ 为一个 **σ-代数**，或者 **σ-域**。

注意 σ-代数 $\mathscr{F}$ 为一个集合，其成员为样本空间 Ω 的子集，即事件，所以 $\mathscr{F}$ 为事件的集合。以上定义中 3 要求如果可数个集合在 $\mathscr{F}$ 中，那么这可数个集合的并集也要求在 $\mathscr{F}$ 中。而同时，结合 2 中关于补集的要求，根据德摩根律：

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$$

因而 2 和 3 结合起来要求可数个集合的交集也需要在 $\mathscr{F}$ 中。

例 2. 对于上述定义的 Ω_1，如果我们关心单个样本点：$\{\heartsuit\}, \{\clubsuit\}, \{\spadesuit\}, \{\diamondsuit\}$，现构建对应的 σ-代数。根据要求，首先集合中应该包含：

$$\{\emptyset, \{\heartsuit\}, \{\clubsuit\}, \{\spadesuit\}, \{\diamondsuit\}, \Omega_1\} \triangleq F_1$$

同时，F_1 中的元素的并集也需要包含在其中，因而：

$$\left\{ \begin{array}{cccccc} \emptyset, & \{\heartsuit\}, & \{\clubsuit\}, & \{\spadesuit\}, & \{\diamondsuit\}, & \Omega_1, \\ \{\heartsuit,\clubsuit\}, & \{\heartsuit,\spadesuit\}, & \{\heartsuit,\diamondsuit\}, & \{\clubsuit,\spadesuit\}, & \{\clubsuit,\diamondsuit\}, & \{\spadesuit,\diamondsuit\}, \end{array} \right\} \triangleq F_2$$

进而，F_1 中的元素的补集也需要包含在其中，因而：

$$\left\{ \begin{array}{cccccc} \emptyset, & \{\heartsuit\}, & \{\clubsuit\}, & \{\spadesuit\}, & \{\diamondsuit\}, & \Omega_1, \\ \{\heartsuit,\clubsuit\}, & \{\heartsuit,\spadesuit\}, & \{\heartsuit,\diamondsuit\}, & \{\clubsuit,\spadesuit\}, & \{\clubsuit,\diamondsuit\}, & \{\spadesuit,\diamondsuit\}, \\ \{\heartsuit,\clubsuit,\spadesuit\}, & \{\heartsuit,\clubsuit,\diamondsuit\}, & \{\heartsuit,\spadesuit,\diamondsuit\}, & \{\clubsuit,\spadesuit,\diamondsuit\} & & \end{array} \right\} \triangleq F_3$$

仔细观察，发现 F_3 已经满足定义中的要求，因而 $\mathscr{F} = F_3$ 即我们要构建的 σ-代数。

若我们关心的样本空间为 $\Omega = \mathbb{R}$，我们令 $\mathscr{I}$ 为所有开区间的集合：$\mathscr{I} = \{(a,b) \,|\, -\infty < a < b < +\infty\}$，那么包含 $\mathscr{I}$ 的最小 σ-代数我们记为 $\mathscr{B}$，并称之为 **Borelσ-代数**或 **Borel 域**，而 $\mathscr{B}$ 中的元素成为 **Borel 集**（**Borel set**）。注意由于：

$$(a,b] = \bigcap_{i=1}^{\infty} \left(a, b + \frac{1}{i}\right)$$

因而所有的左开右闭区间也都是 Borel 集。同理可证所有的左闭右开区间 $[a,b)$、闭区间 $[a,b]$ 及其可数并、交都为 Borel 集。

3 概率

现在，经过以上准备之后，我们可以定义概率了。

定义 2.（**Kolmogorov axioms**）给定一个样本空间 Ω 以及相应的 σ-代数 $\mathscr{F}$，函数 $\mathscr{P}:\mathscr{F}\to[0,1]$ 若满足：

1. 对于所有的事件 $A\in\mathscr{F}$，$\mathscr{P}(A)\geq 0$
2. $\mathscr{P}(\Omega)=1$
3. 若 $A_1,A_2,\ldots\in\mathscr{F}$ 为两两互斥事件，则 $\mathscr{P}\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mathscr{P}(A_i)$（**可数可加性或可列可加性**）

则我们称 $\mathscr{P}$ 为**概率函数**或**概率测度**。

以上概率的定义通常称之为概率的公理化定义（Axioms of Probability），或者柯尔莫哥洛夫公理（Kolmogorov Axioms）。注意以上的定义并没有限定概率函数的形式，只要满足以上三个条件的函数 $\mathscr{P}$ 都可以被定义为概率函数。

例 3.（抛硬币的概率）现在我们进行一项抛硬币的随机试验。记正面为 H，反面为 T，那么我们关心的样本空间为 $\Omega=\{H,T\}$。假设硬币质地均匀，即正面和反面的概率相等，那么

$$\mathscr{P}(\{H\})=\mathscr{P}(\{T\})$$

包含 $\{H\},\{T\}$ 的最小 σ-代数为 $\mathscr{F}=\{\emptyset,\Omega,\{H\},\{T\}\}$。根据概率的定义，$\mathscr{P}(\Omega)=\mathscr{P}(\{H\}\cup\{T\})=\mathscr{P}(\{H\})+\mathscr{P}(\{T\})=1$，因而 $\mathscr{P}(\{H\})=\mathscr{P}(\{T\})=0.5$。因而 $(\Omega,\mathscr{F},\mathscr{P})$ 即组成了概率空间。

如果假设硬币是不均匀的，且获得正面的概率为 0.1，那么同样根据概率定义，$\mathscr{P}'(\{H\})=0.1,\mathscr{P}'(\{T\})=\mathscr{P}'(\Omega)-\mathscr{P}'(\{H\})=1-0.1=0.9$，从而 $(\Omega,\mathscr{F},\mathscr{P}')$ 组成了新的概率空间。

以上我们使用抛硬币的例子构建了样本空间有限时的概率空间，然而当样本点逐渐增多时，$\mathscr{F}$ 的元素个数也会相应增加，逐个检验 $\mathscr{F}$ 中元素是否满足概率定义的三个条件变的更为复杂。而对于一般的有限或者可数个样本点的情形（或统称为离散的样本空间），我们可以使用如下的方法定义概率函数。

定理 1. 令 $\Omega=\{s_1,s_2,\ldots,s_n\}$ 为有限集，令 $\mathscr{F}$ 为 S 子集的任意 σ-代数。令 $p_1,p_2,\ldots p_n$ 为非负实数且 $\sum_{i=1}^{n}p_i=1$。对于任意集合 $A\in\mathscr{F}$，定义

$$\mathscr{P}(A)=\sum_{\{s_i,i\in A\}}p_i \tag{1}$$

则 $\mathscr{P}$ 为 $\mathscr{F}$ 上的概率函数。对于可数集 $\Omega=\{s_1,s_2,\ldots\}$ 可类似构建概率函数。

Proof. 定义 (2) 中第 1 条显然满足，对于第 2 条，根据上述定义，$\mathscr{P}(\Omega)=\sum_{i=1}^{n}p_i=1$，满足。对于第三条，对于两两互斥事件 $A_1,\ldots A_k$：

$$\mathscr{P}\left(\bigcup_{i=1}^{k}A_i\right)=\sum_{\left\{s_i,i\in\bigcup_{i=1}^{k}A_i\right\}}p_i=\sum_{i=1}^{k}\sum_{\{s_j,j\in A_i\}}p_j=\sum_{i=1}^{k}\mathscr{P}(A_i)$$

因而上述定义的概率函数 $\mathscr{P}$ 满足 Kolmogorov 公理。 □

例 4. 如果我们重复、独立的进行抛硬币试验（**伯努利试验**）N 次，且 $\mathscr{P}(\{H\}) = p$，那么 N 次实验中得到正面的次数的集合为：$\Omega = \{0, 1, \ldots N\}$，对应的概率为：$\mathscr{P}(\{k\}) = \begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k}$，$\mathscr{P}(\{k\}) \geqslant 0$ 且 $\sum_{k=0}^{N} \mathscr{P}(\{k\}) = 1$，因而根据定理 (1)，使用 (1) 式定义的 $\mathscr{P}$ 即定义了样本空间 Ω 上任意 σ-代数的所有子集的概率函数。

我们关心在一个小时之内到达某银行的客户数，客户数为可数集，样本空间为 $\Omega = \{0, 1, 2 \ldots\}$。取一个非常大的自然数 n，我们可以把一个小时分解为等长的 n 段，即 $(0, \frac{1}{n}], (\frac{1}{n}, \frac{2}{n}] \ldots (\frac{n-1}{n}, 1]$，当 n 很大时，一个区间段内有两个客户到达的概率几乎可以忽略不计。假设每段时间客户到达的概率相等，且反比于 n，不妨假设为 $\frac{\lambda}{n}$，那么一小时内总的人数：

$$\mathscr{P}^*(\{k\}) = \begin{pmatrix} n \\ k \end{pmatrix} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

令 $n \to \infty$，则 $\dfrac{\begin{pmatrix} n \\ k \end{pmatrix}}{n^k} = \frac{n!}{k! \cdot (n-k)! \cdot n^k} \to \frac{1}{k!}$，$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}$，因而

$$\mathscr{P}(\{k\}) = \frac{\lambda^k}{k!} e^{-\lambda} > 0$$

且 $\sum_{k=0}^{\infty} \mathscr{P}(\{k\}) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1$（$e^{\lambda x}$ 在 $x = 0$ 处泰勒展开可得），因而根据定理 (1)，使用 (1) 式定义的概率函数 $\mathscr{P}$ 即定义了样本空间 Ω 上任意 σ-代数的概率函数。

3.1 分布函数与概率定义

尽管在样本空间为可数的情况下定义概率函数相对简单，然而当我们考虑的样本空间为不可列时，概率函数的定义变得尤为困难。

例 5. （勒贝格不可测集）如果我们选取样本空间 $\Omega = [0, 1]$，令 Ω 中的所有有理数集合为 Q'，由于有理数为可数集合，因而可以写成 $Q' = \{q_1, q_2, \ldots\}$。对于 $(0, 1)$ 之间的任意实数 a，定义集合

$$S_a = \left\{ \begin{array}{ccc} a+q & if & a+q < 1 \\ a+q-1 & if & a+q \geqslant 1 \end{array} \forall q \in Q \right\}$$

那么可知 $\bigcup_{a \in (0,1)} S_a = [0, 1]$。由于 S_a 也是可数集，因而可以将其写为：

$$S_a = \{s_{a1}, s_{a2}, \ldots\}$$

令 T_1 为所有 S_a 中的 s_{a1}，T_2 为所有 S_a 中的 s_{a2}，因而我们有可数个 T_k，$\bigcup_{k=1}^{\infty} T_k = [0,1]$，且 T_k 两两不相交。每个 T_k 地位相等因而 $\mathscr{P}(T_k) = \mathscr{P}(T_{k'})$。若 $\mathscr{P}(T_k) > 0$，则：

$$1 = \mathscr{P}([0,1]) = \mathscr{P}\left(\bigcup_{k=1}^{\infty} T_k\right) = \sum_{k=1}^{\infty} \mathscr{P}(T_k) = \infty$$

如果 $\mathscr{P}(T_k) = 0$，则：

$$1 = \mathscr{P}([0,1]) = \mathscr{P}\left(\bigcup_{k=1}^{\infty} T_k\right) = \sum_{k=1}^{\infty} \mathscr{P}(T_k) = 0$$

无论如何都会得到矛盾。

因而在概率论中，在仅仅给定样本空间的情况下，并非任意集合都可以确定其概率。我们一般将上述性质不够良好的集合称之为（勒贝格）不可测集，而概率空间中 $\mathscr{F}$ 应该排除这些性质不够良好的不可测集。

定义 3. （**分布函数**）如果函数 $F : \mathbb{R} \to \mathbb{R}$ 满足：

1. 单调性：$F(a) \leq F(b), a \leq b$
2. 右连续：$\lim_{x \to x_0^+} F(x) = F(x_0)$
3. $F(-\infty) = 0, F(\infty) = 1$

则称 F 为分布函数（**distribution function**, d.f.）。特别的，令

$$\delta_t(x) = \begin{cases} 0 & x < t \\ 1 & x \geq t \end{cases}$$

若 $\{a_j\}$ 为可数集，$b_j > 0, \sum_j b_j = 1$，则 $F(x) = \sum_j b_j \delta_{a_j}(x)$ 为分布函数，我们称之为**离散型分布函数**（discrete d.f.）；处处连续的分布函数成为**连续型分布函数**（countinuous d.f.）。

例 6. 令 $a_1 = 0, a_2 = 1, b_1 = \frac{1}{3}, b_2 = \frac{2}{3}$，则 $F_d(x) = \sum_{j=1}^{2} b_j \delta_{a_j}(x)$ 为离散型分布函数，如图 (2.a) 所示；$F_c(x) = \frac{e^x}{1+e^x}$（Logistic 分布）为连续型分布函数，如图 (2.b) 所示，而 $F(x) = \frac{1}{3} F_d(x) + \frac{2}{3} F_c(x)$ 也为分布函数，如图 (2.c) 所示。

定理 2. 每个分布函数都可以写为一个离散型分布函数和一个连续型分布函数的凸组合，且该分解唯一。

在有了分布函数之后，我们可以使用分布函数定义 $\mathbb{R}$ 上的概率函数。例 (2) 中我们通过开区间定义了 Borel 域，如果我们定义：

$$P((-\infty, x]) = F(x) \tag{2}$$

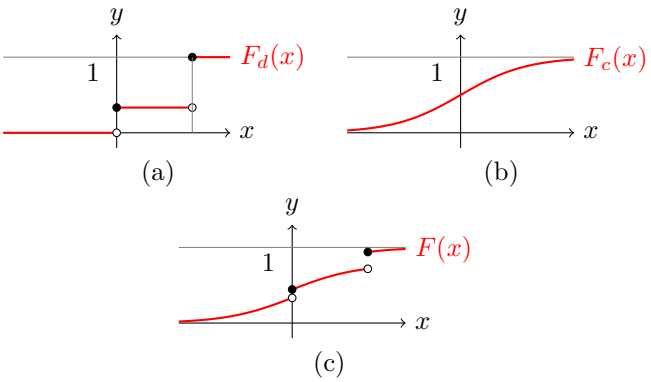


图 2: 分布函数

则对于任意的 $-\infty < a < b < +\infty$，有[1]：

$$
\begin{aligned}
P((a,b]) &= F(b) - F(a) \\
P((a,b)) &= F(b-) - F(a) \\
P([a,b)) &= F(b-) - F(a-) \\
P([a,b]) &= F(b) - F(a-)
\end{aligned}
$$

定理 3. 给定任意的分布函数 F，式 *(2)* 定义了 *Borel* 域 $\mathscr{B}$ 上的概率测度。

在此，我们仅仅概括性的了解一下整个定义过程，至于具体的理论推导，感兴趣可以参考 Ash (2000) Ch.1.3-1.4。现在考虑一个集合 $S \subset \mathbb{R}$，如果 S 可以写成可数个不相交的左开右闭区间的并集，即：

$$
S = \bigcup_{i=1}^{\infty} (a_i, b_i]
$$

那么 $P(S) = \sum_{i=1}^{\infty} P((a_i, b_i]) = \sum_{i=1}^{\infty} (F(a_i) - F(b_i))$。而类似的，对于任意的一个开集 $U \subset \mathbb{R}$，都可以写为可数个开区间的并集，即：

$$
U = \bigcup_{i=1}^{\infty} (c_i, d_i)
$$

类似地，$P(U) = \sum_{i=1}^{\infty} [F(d_i-) - F(c_i)]$，故对于所有的开集，我们定义了其概率。进而，由于闭集是开集的补集，我们可以使用开集的概率来定义闭集的概率。

在定义了开集和闭集的概率之后，对于任意一个集合 $S \subset \mathbb{R}$，我们可以定

[1]这也就是为什么开区间、闭区间、半开半闭区间在形成 Borel 域以及定义概率函数上都是等价的，然而一般用左开右闭区间的原因，因为 $P((a,b]) = F(b) - F(a)$，不像其他三个定义需要使用极限，表达更方便。

义其**外测度**（**outer measure**）：

$$P^*(S) = \inf_{\text{开集}U, S\subset U} P(U)$$

和**内测度**（**inner measure**）：

$$P_*(S) = \sup_{\text{闭集}C, C\subset S} P(C)$$

易知 $P_*(S) \leq P^*(S)$。一般来说，等号不一定成立，然而如果等号成立，我们就定义 $P(S) = P^*(S) = P_*(S)$，并称 S 为**可测集**（**meansurable set**）。可以证明，所有的可测集是一个 σ 代数，且包含了所有的开区间。由于 Borel 域 $\mathscr{B}$ 是包含所有开区间的最小 σ 代数，因而这个概率定义了 Borel 域 $\mathscr{B}$ 上的概率函数。

至此，我们就定义了实数集 $\mathbb{R}$ 上的概率空间：$(\mathbb{R}, \mathscr{B}, P)$[2]。

有一些命题，尽管其并非对于每个 $\omega \in \Omega$ 都成立，但是 $\mathscr{P}(\{\omega : \text{命题成立}\}) = 1$，即这个命题成立的概率为 1，那么我们称这个命题是几乎处处（**almost everywhere**）成立的，或者几乎必然（**almost sure**）成立的，简记为 $a.e.$ 或者 $a.s.$。例如：

例 7. 取 $\Omega = \mathbb{R}$，定义

$$X_n(\omega) = \begin{cases} 0 & if\ |\omega| > \frac{1}{n} \\ n & if\ |\omega| \leq \frac{1}{n} \end{cases}$$

因而除了在一个点 $\omega = 0$ 处之外，$\lim_{n\to\infty} X_n(\omega) = 0$。如果分布函数连续，单点集 $P(\{0\}) = 0$，因而 $\lim_{n\to\infty} X_n(\omega) = 0$ 以概率 1 成立，我们称 $\lim_{n\to\infty} X_n(\omega) = 0$ 几乎必然成立，简记为：$\lim_{n\to\infty} X_n(\omega) = 0\ a.e.$ 或者 $\lim_{n\to\infty} X_n(\omega) = 0\ a.s.$。

3.2 概率函数的性质

定理 4. 对于概率函数 $\mathscr{P}$，有以下性质：

1. $\mathscr{P}(A) \leq 1$
2. $\mathscr{P}(A^c) = 1 - P(A)$
3. $\mathscr{P}(\emptyset) = 0$
4. $\mathscr{P}(A \cup B) + \mathscr{P}(A \cap B) = \mathscr{P}(A) + \mathscr{P}(B)$
5. $A \subset B \Rightarrow \mathscr{P}(A) = \mathscr{P}(B) - \mathscr{P}(B \backslash A) \leq \mathscr{P}(B)$

[2]注意在此讲义中，对于一般的概率空间，我们使用 $\mathscr{P}$ 作为概率函数的标记，而对于 Borel 集上的概率函数，我们使用 P 作为特殊标记。

6. $\mathscr{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathscr{P}\left(A_i\right)$

7. 如果 $C_1, C_2, \ldots$ 为样本空间 Ω 的一个划分，那么 $\mathscr{P}(A)=\sum_{i=1}^{\infty} \mathscr{P}\left(A \cap C_i\right)$。

例 8. （**Bonferroni's Inequality**）根据定理 (4)，我们有：

$$\mathscr{P}(A \cap B)=\mathscr{P}(A)+\mathscr{P}(B)-\mathscr{P}(A \cup B) \geq \mathscr{P}(A)+\mathscr{P}(B)-1$$

该不等式确定了事件 A 和 B 同时发生的概率的下界。比如，天气预报预计明天上海下雨（A）的概率为 0.90，北京下雨（B）的概率为 0.8。如果假设两地是否下雨是独立事件，那么两地同时下雨的概率为 $\mathscr{P}(A \cap B)=\mathscr{P}(A) \cdot \mathscr{P}(B)=0.72$。然而现实情况是，我们并不知道两地下雨是否独立，但是根据上面的不等式，我们仍然可以得到两地同时下雨的概率的一个下界：$\mathscr{P}(A \cap B) \geqslant 0.8+0.9-1=0.7$，即两地同时下雨的概率至少有 0.7。

4 条件概率与独立

在此之前我们讨论的都是无条件概率。然而现实的应用中，我们经常碰到用已知信息推断未知信息的问题，这就涉及到条件概率的概念。条件概率即指，给定事件 B 发生的情况下，事件 A 发生的概率是多少。无条件概率和有条件概率有着不同的应用，比如如果我们想要标记山体滑坡的危险路段，那么我们可以通过统计路段长时间以来发生山体滑坡的概率，即无条件概率，进行标记；而当我们进行灾害预警时，我们知道给定天气是阴雨天的情况下，山体滑坡的概率会变的异常高，这个时候我们就是在使用条件概率了，即 $\mathscr{P}$(发生山体滑坡 | 阴雨天) 可能是灾害预警所关注的，而不是 $\mathscr{P}$(发生山体滑坡)。

定义 4. （**条件概率**）如果 A 和 B 为 Ω 中的两个时间，且 $\mathscr{P}(B)>0$，那么给定 B，事件 A 发生的条件概率为：

$$\mathscr{P}(A|B)=\frac{\mathscr{P}(A \cap B)}{\mathscr{P}(B)} \tag{3}$$

注意只有当事件 B 有正概率发生时，条件概率的定义才有意义。实际上，条件概率可以理解为我们把原始的样本空间 Ω 限定在新的样本空间 B 中，并相应对原概率函数使用式 (3) 对概率函数进行了重新定义，因而概率的性质在条件概率下依然成立。

定理 5.（全概率公式）如果 $C_1, C_2, \ldots$ 为样本空间 Ω 的一个划分，那么 $\mathscr{P}(A)=\sum_{i=1}^{\infty} \mathscr{P}(A|C_i) \cdot P\left(C_i\right)$。特别的，对于任意事件 B，有 $\mathscr{P}(A)=\mathscr{P}(A|B) \cdot \mathscr{P}(B)+\mathscr{P}(A|B^c) \cdot \mathscr{P}(B^c)$。

Proof. 根据条件概率定义，$\sum_{i=1}^{\infty} \mathscr{P}(A|C_i) \cdot \mathscr{P}(C_i)=\sum_{i=1}^{\infty} \mathscr{P}(A \cap C_i)$，根据定理 (4.7) 可证。 □

如果事件 A 和 B 都有正的概率，那么我们可以同时定义 $\mathscr{P}(A|B)$ 以及 $\mathscr{P}(B|A)$，根据定义，

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

以上关系式我们称之为**贝叶斯法则**（**Bayes' Rule**）。更进一步，根据全概率公式：

$$\mathscr{P}(A|B)=\frac{\mathscr{P}(B|A)\cdot P(A)}{\mathscr{P}(B)}=\frac{\mathscr{P}(B|A)\cdot\mathscr{P}(A)}{\mathscr{P}(B|A)\cdot\mathscr{P}(A)+\mathscr{P}(B|A^c)\cdot\mathscr{P}(A^c)}$$

贝叶斯法则被广泛应用在统计学中，因为贝叶斯法则体现了我们认识世界的一般规律，很好的将理论与现实、已知与未知、主观与客观联系在一起。在贝叶斯公式中，条件概率 $\mathscr{P}(B|A)$ 是我们已知的部分，或者我们的理论，而条件概率 $\mathscr{P}(A|B)$ 则是我们未知的需要判断的部分，而 $P(A)$ 则是主观的先验部分。贝叶斯公式的优雅之处在于，它告诉我们，如何使用我们的理论（$\mathscr{P}(B|A)$）和我们的主观先验（$\mathscr{P}(A)$）结合起来，使用更多的信息（B），推断我们所未知的知识 $\mathscr{P}(A|B)$。

例 9. 小明正在追求一个女孩小红，但是小明不知道小红的心意。假设 A 代表小红喜欢小明，A^c 代表小红不喜欢小明，由于小明不确定小红的心思，因而主观先验概率 $\mathscr{P}(A)=\mathscr{P}(A^c)=0.5$。为了探明小红的心意，4 月 1 日小明发短信跟小红表白，小红回短信表示愿意接受。小明的理论认为，如果小红喜欢自己，有 99% 的可能性会表示接受；而如果小红不喜欢自己，有一半的可能性小红是在开愚人节玩笑。如果令 B 代表小红接受的行为，那么小红的理论可以归结为：$\mathscr{P}(B|A)=0.99,\mathscr{P}(B|A^c)=0.5$。那么现在，根据贝叶斯法则，小明可以对小红喜欢自己的概率做出判断：

$$\begin{aligned}\mathscr{P}(A|B) &= \frac{\mathscr{P}(B|A)\cdot\mathscr{P}(A)}{\mathscr{P}(B|A)\cdot\mathscr{P}(A)+\mathscr{P}(B|A^c)\cdot\mathscr{P}(A^c)}\\ &= \frac{0.99\times0.5}{0.99\times0.5+0.5\times0.5}\approx66\%\end{aligned}$$

贝叶斯法则可以扩展到更一般的形式：

定理 6. （贝叶斯法则）如果 $B_1,B_2,...$ 为样本空间 Ω 的一个划分，令 $A\in\mathscr{B}$，那么：

$$\mathscr{P}(B_i|A)=\frac{\mathscr{P}(A|B_i)\mathscr{P}(B_i)}{\sum_{j=1}^{\infty}\mathscr{P}(A|B_j)\mathscr{P}(B_j)}$$

例 10. 有三扇门，其中一扇门里有奖品，三选一，你选择其中一扇门之后，主持人先不揭晓答案，而是从另外两扇门中排除掉一个没有奖品的门，现在主持人问你，要不要换个门，请问你换还是不换？这里假设第 i 扇门里面有奖品的事件为 A_i，没有奖品为 A_i^c，因而总共三种情况，即 $\{A_1A_2^cA_3^c,A_1^cA_2A_3^c,A_1^cA_2^cA_3\}$，

三种情况是等可能的。不失一般性，假设你选择了第一扇门，而主持人打开了第三扇门。如果根据第三扇门内没有奖品这一信息，可以得出：$\mathscr{P}\left(A_1A_2^c|A_3^c\right) = \frac{\mathscr{P}(A_1A_2^cA_3^c)}{\mathscr{P}\left(A_3^c\right)} = \mathscr{P}\left(A_1^cA_2|A_3^c\right) = \frac{\mathscr{P}(A_1^cA_2A_3^c)}{\mathscr{P}\left(A_3^c\right)} = \frac{1}{2}$，即是否换门都可以。

然而，主持人选择哪一扇门这一动作本身就可以的带来信息。记 S_i 为主持人翻开某扇门的概率。如果奖品藏在第一扇门内，那么 $\mathscr{P}\left(S_3|A_1A_2^cA_3^c\right) = \frac{1}{2}$，而 $\mathscr{P}\left(S_3|A_1^cA_2A_3^c\right) = 1$，也就是说，当你选择了第一扇门，而奖品在第二扇门里面，那么主持人一定会选择打开第二扇门。根据贝叶斯公式：

$$\begin{aligned}\mathscr{P}\left(A_1A_2^c|S_3A_3^c\right) &= \frac{\mathscr{P}\left(S_3|A_1A_2^cA_3^c\right)\cdot\mathscr{P}\left(A_1A_2^c|A_3^c\right)}{\mathscr{P}\left(S_3|A_1A_2^cA_3^c\right)\cdot\mathscr{P}\left(A_1A_2^c|A_3^c\right)+\mathscr{P}\left(S_3|A_1^cA_2A_3^c\right)\cdot\mathscr{P}\left(A_1^cA_2|A_3^c\right)} \\ &= \frac{\frac{1}{2}\cdot\frac{1}{2}}{\frac{1}{2}\cdot\frac{1}{2}+1\cdot\frac{1}{2}} = \frac{1}{3}\end{aligned}$$

同理 $\mathscr{P}\left(A_1^cA_2A_3^c|S_3\right) = \frac{2}{3}$，因而考虑到主持人选择打开第三扇门这一行为，换第二扇门是最好的选择。

然而很多时候，两个事件发生与否并没有什么关联，因而使用条件概率获并不能获得更多的信息。比如尽管从个人经验来看，我每次看国足国足都会输球，然而实际上我是否看国足的比赛与国足比赛是否能赢球，并没有任何关联，因而不能使用我是否看过去比赛去预测国足比赛是否能赢球，否则我可以通过足彩立即实现财务自由。在统计上，我们经常需要假设两类事件没有什么关联，才能简化并进行分析。比如当我们做抛硬币实验时，我们就潜在假设了每次抛出硬币与其他次抛出硬币的结果是没有任何关联的。这就引出了「统计独立性」的概念。

定义 5.（**统计独立性**）如果两个事件 A 和 B 满足：

$$\mathscr{P}\left(A\cap B\right) = \mathscr{P}\left(A\right)\cdot\mathscr{P}\left(B\right)$$

那么我们称事件 A 和 B 为独立事件。

定理 7. 如果 A 和 B 为独立事件，那么以下事件对也为独立事件：

1. A 和 B^c
2. A^c 和 B
3. A^c 和 B^c。

注意当我们考虑多于两个事件时，以上定义并不能进行扩展。

例 11. 现在抛掷两枚骰子，我们关心如下三个事件：

$$\begin{gathered}A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),\} \\ B = \{\text{两枚骰子之和介于 7 和 10 之间}\} \\ C = \{\text{两枚骰子之和为 2,7 或者 8}\}\end{gathered}$$

可以计算得到：$\mathscr{P}(A)=\frac{1}{6}$，$\mathscr{P}(B)=\frac{1}{2}$，$\mathscr{P}(C)=\frac{1}{3}$，而：

$$\begin{aligned}\mathscr{P}(A\cap B\cap C) &= \mathscr{P}(\{(4,4)\})\\ &= \frac{1}{36}\\ &= \mathscr{P}(A)\cdot\mathscr{P}(B)\cdot\mathscr{P}(C)\end{aligned}$$

然而：

$$\mathscr{P}(B\cap C)=\frac{11}{36}\neq\mathscr{P}(B)\cdot\mathscr{P}(C)$$

因而事件 B 和 C 并不独立。

为了更好的定义多于两个事件时的独立，我们使用如下定义：

定义 6. 我们称一系列事件 $A_1,A_2,...A_n$ 为**相互独立**的（**mutually independent or jointly independent**），如果对于任意的子列 $A_{i_1},A_{i_2},...A_{i_k}$，有：

$$\mathscr{P}\left(\bigcap_{j=1}^{k}A_{i_j}\right)=\prod_{j=1}^{k}\mathscr{P}\left(A_{i_j}\right)$$

习题

练习 1. 整数集 $\mathbb{Z}$ 是可数集还是不可数集？有理数集 $\mathbb{Q}$ 是可数集还是不可数集？开区间 $(0,1)$ 是可数集还是不可数集？

练习 2. 思考题：标准篮球的直径为 $24.6cm$，而标准篮筐的直径为 $45cm$，如果篮球垂直入框，中心落点均匀的落在篮筐内，请问投出空心球的概率有多大？如果篮球以 60° 角入框呢？篮球以多大的角度入框则不可能投出空心球？

练习 3. 试用交、并、补三个运算表示 $(E\triangle F)^c$。

练习 4.

1. 上例中，包含 $\{\clubsuit\}$ 的最小 σ-代数是？
2. 现在抛一枚骰子，则结果的样本空间为 $\Omega_4=\{1,2,3,4,5,6\}$，那么包含所有单个样本点的 σ-代数 $\mathscr{F}_4$ 中有多少个事件？

练习 5. 百年一遇的自然灾害（即每年发生的概率为 $p=1\%$）10 年期间至少发生 1 次的概率是多少？（假设这种自然灾害每年发生与否是独立的）

练习 6. 七个人玩桌游「炸碉堡」，经过第一轮投票，已知第一轮三个行动人中有一个是坏人，剩下的四个人中也有一个是坏人。如果在第二轮中由你选择三个人作为行动人，你的目标是尽可能的选出三个好人。你有如下三个策略：

1. 从三个人中选一个，另外四个人中选两个

2. 从三个人中选两个，另外四个人中选一个

3. 完全从四个人中选出新的三个

请问以上三个策略中，哪一个策略选出三个好人的概率最高?

练习 7. 给定任意一个连续的分布函数 F 及由其定义的概率函数 P，$\mathbb{R}$ 上的单点集 $\{a, a \in \mathbb{R}\}$ 的概率 $P(\{a, a \in \mathbb{R}\})$ 是多少? 根据概率函数的性质，$\mathbb{R}$ 上的任意可数集的概率 $P(\{a_i, i = 1, 2, \ldots, a_i \in \mathbb{R}\})$ 是多少? 所以概率为 0 的事件一定是不可能事件么?

练习 8. 试证明定理 (4)。

练习 9. 请给出两个事件至少有一个发生的概率，$\mathscr{P}(A \cup B)$ 的一个上界和一个下界。

练习 10. 试证明以上定理。

参考文献

[1] Ash, R.B., Doleans-Dade, C., 2000. Probability and measure theory. Academic Press.

[2] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[3] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[4] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

第二节 · 随机变量

司继春

上海对外经贸大学统计与信息学院

1 一元随机变量

上面介绍了一般的概率空间构建所需要的步骤，即一个概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$，我们需要一个样本空间 Ω，一个性质良好的集合族 $\mathscr{F}$ 以及定义在这个集合族上的概率函数 P。特别地，当我们选取样本空间 $\Omega = \mathbb{R}$ 时，我们有 Borelσ-代数 $\mathscr{B}$ 以及由分布函数 F 定义的概率函数 P 组成的概率空间 $(\mathbb{R}, \mathscr{B}, P)$。

尽管对于一般的样本空间 Ω，我们都可以使用以上的方法构建概率空间，然而很多时候，直接对原始的样本空间 Ω 和集合族 $\mathscr{F}$ 进行分析并不是非常的方便。比如，Ω 作为样本空间，我们并没有限制 Ω 具有代数结构，因而一般我们不能对样本点进行加减等运算。再比如如果我们的研究对象为抛 1000 次硬币，那么我们的样本空间有 2^{1000} 个元素，而这些元素不能相加相减（比如没有正面 + 正面 这样的运算）。为了方便分析，我们一般会把原始的概率空间 Ω 映射到实轴 $\mathbb{R}$ 上进行分析，于是就有了随机变量的概念。

定义 1.（**随机变量**）对于概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$，映射 $X : \Omega \to \mathbb{R}^*$ 满足：对于任意的 $B \in \mathscr{B}$，有：

$$X^{-1}(B) \triangleq \{\omega : X(\omega) \in B\} \in \mathscr{F}$$

那么我们称 X 为随机变量（random variable, r.v.）。

例 1. 对于抛硬币的实验，$\Omega = \{H, T\}$，我们可以定义一个随机变量 X 如下：

$$\begin{cases} X(H) = 0 \\ X(T) = 1 \end{cases}$$

对于 $\mathscr{F} = \{\emptyset, \{H\}, \{T\}, \Omega\}$，我们有 $X^{-1}(0) = \{H\}, X^{-1}(1) = \{T\}$，对于其他任何 Borel 集 B，如果 $1 \in B$ 则 $T \in X^{-1}(B)$，如果 $0 \in B$ 则 $H \in X^{-1}(B)$。如此我们便定义了一个 $\Omega \to \mathbb{R}^*$ 的随机变量 X。

例 2. 对于泊松分布，其 $\Omega = \{0, 1, 2 \ldots\} = \mathbb{Z}$，定义 $X(\omega) = \omega, \omega \in \mathbb{Z}$，同上，我们定义了从自然数集合 $\mathbb{Z} \to \mathbb{R}$ 的随机变量 X。

例 3. 电灯泡的寿命的样本空间为 $\Omega = (0, +\infty)$，我们可以定义 $X(\omega) = \omega, \omega \in \Omega$，如此我们定义了从正实数集 $\mathbb{R}^+ \to \mathbb{R}$ 的随机变量 X。

在此之前我们定义了离散型的样本空间和离散型的分布函数，以上例 (1) 和例 (2) 都属于这种情况。下面我们来定义离散型的随机变量：

定义 2.（**离散型随机变量**）如果存在一个可数集 $B \in \mathscr{B}$，满足 $P(X \in B) = 1$，则随机变量 X 成为离散型随机变量。

在得到随机变量的定义之后，我们还需要定义在 $(\mathbb{R}^*, \mathscr{B})$ 上的概率函数才能完成随机变量的概率空间的定义。由于随机变量是定义在一个一般的样本空间 Ω 及其对应的概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上，因而自然的想法是使用原概率空间中的概率函数 $\mathscr{P}$ 来定义 $(\mathbb{R}^*, \mathscr{B})$ 上的新的概率函数 P。

定理 1. 对于一个随机变量 $X : \Omega \to \mathbb{R}$，定义

$$P_X(B) = \mathscr{P}\left(X^{-1}(B)\right) = \mathscr{P}(\{\omega : X(\omega) \in B\})$$

则 P_X 为概率函数，$(\mathbb{R}, \mathscr{B}, P_X)$ 为概率空间，我们称 $(\mathbb{R}, \mathscr{B}, P_X)$ 为 $(\Omega, \mathscr{F}, \mathscr{P})$ 导出的概率空间。

例 4. 对于例 (1) 中的随机变量 X，如果原概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中定义 $\mathscr{P}(H) = \mathscr{P}(T) = 0.5$，则 $P_X(\{1\}) = \mathscr{P}\left(X^{-1}(1)\right) = \mathscr{P}(T) = 0.5$，同理可定义 $P(\{0\})$，从而完成概率空间 $(\mathbb{R}, \mathscr{B}, P_X)$ 的定义。

在此之前，我们定义了分布函数，而对于每一个随机变量，由于其值域均为 $\mathbb{R}$，因而其总对应一个分布函数。

定义 3.（累积分布函数）对于一个随机变量 X，函数

$$F_X(x) = P_X((-\infty, x]) = \mathscr{P}\left(X^{-1}((-\infty, x])\right), \forall x \in \mathbb{R}$$

为一个分布函数（满足分布函数定义的要求），我们称其为**累积分布函数**（**cumulative distribution function, c.d.f.**）。

对于随机变量来说，累积分布函数包含了所有概率函数 P_X 的信息，因而使用 P_X 和使用累积分布函数 F_X 是等价的。因而我们通常使用标记 $X \sim F_X(x)$ 表示随机变量 X **服从** F_X **分布**。此外，如果随机变量 X 和 Y 具有同样的分布，则记为 $X \sim Y$。

例 5.（累积分布函数）泊松分布和 Logistic 分布函数如图 (1.1) 和 (1.2) 所示：

定义 4. 如果两个随机变量的累积分布函数 $F_X(x) = F_Y(x)$，则我们称两个随机变量**同分布**。

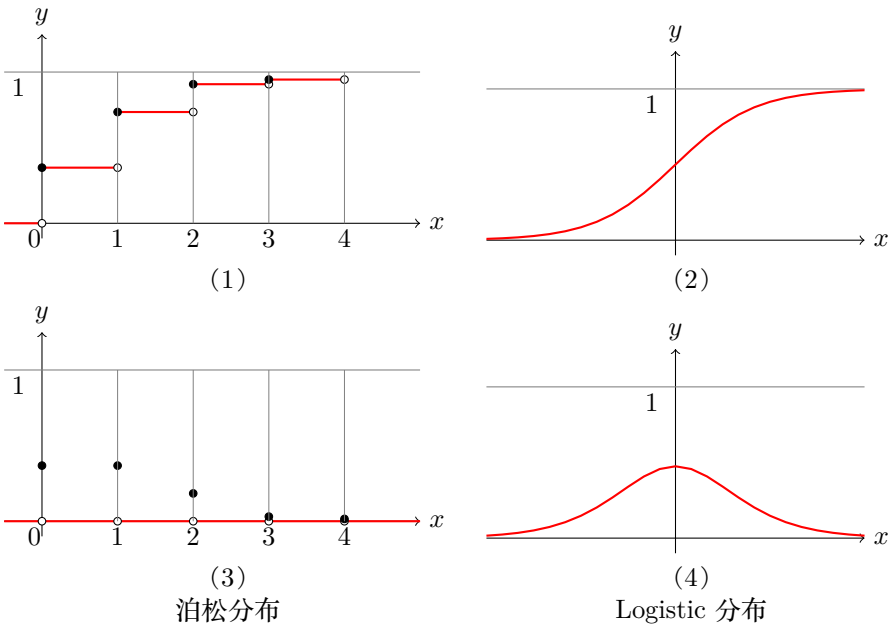


图 1: 累积分布函数与概率密度函数

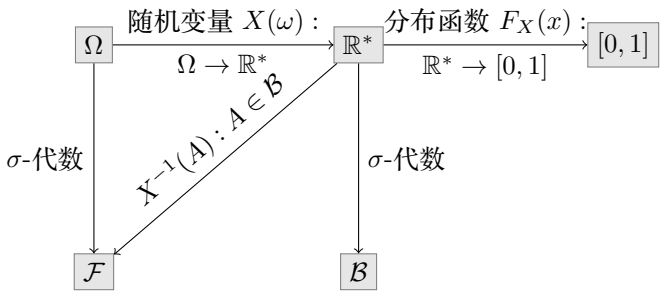


图 2: 随机变量

总结一下，图 (2) 回顾了从原概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 定义随机变量 X 并定义新的概率空间 $(\mathbb{R}, \mathscr{B}, P_X)$，并在此基础上定义分布函数的整个过程。由于随机变量极大的简化了分析，因此下面的分析中主要以随机变量为主要研究手段研究概率与统计问题。

虽然累计分布函数描述了随机变量的所有特征，然而很多时候，使用概率密度函数可以使分析和计算更为便利。

定义 5. 对于离散随机变量，**概率质量函数**（**probability mass function**, p.m.f）定义为：

$$f_X(x) = P(X = x), \text{for all } x$$

定义 6. （概率密度函数）对于连续型随机变量，**概率密度函数**（**probability density function**, p.d.f），$f_X(x)$ 定义为：

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \text{for all } x$$

例 6. （概率密度函数）泊松分布的 p.m.f 如图 (1.3) 所示，Logistic 分布的 p.d.f 如图 (1.2) 所示。

2 期望及其性质

2.1 期望的定义与性质

数学期望（**mathematical expectation**）的概念在概率论和统计学中有着最广泛的应用。在一个一般的概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中，数学期望实际上就是在这个概率空间中的积分。下面我们将给出一个在一般的概率空间中期望（积分）的定义。

定义 7. （离散型随机变量的期望）在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中，对于正的离散型随机变量 X：

$$X(\omega) = b_j \; if \; \omega \in \Lambda_j$$

其中 Λ_j 为样本空间 Ω 的划分，$b_j \geq 0$。期望定义为：

$$\mathbb{E}(X) = \sum_j b_j \mathscr{P}(\Lambda_j)$$

注意在上面的定义中，我们要求随机变量为离散型随机变量，而对于样本空间是否离散并没有做假定。

例 7. 对于例 (2) 中定义的随机变量 $X(\omega) = \omega, \omega \in \mathbb{Z}$，其期望为：

$$\mathbb{E}(X) = \sum_{j=0}^{\infty} j \cdot \mathscr{P}(\{j\}) = \sum_{j=0}^{\infty} j \cdot \frac{\lambda^j}{j!} e^{-\lambda} = \lambda \sum_{j=1}^{\infty} \frac{\lambda^{j-1}}{(j-1)!} e^{-\lambda} = \lambda$$

特别的，如果令 $X(\omega)=1_A(\omega)$，其中 1_A 为指示函数，即：

$$1_A(\omega)=\begin{cases}1 & if\ \omega\in A\\ 0 & otherwise\end{cases}$$

那么其期望：

$$\mathbb{E}(1_A)=1\cdot\mathscr{P}(A)+0=\mathscr{P}(A)$$

下面我们使用离散型随机变量的数学期望继续定义连续型随机变量的数学期望。令 X 为定义在 $(\Omega,\mathscr{F},\mathscr{P})$ 上任意的正的随机变量（$X(\omega)\geq 0$），定义集合：

$$\Lambda_{mn}=\left\{\omega:\frac{n}{2^m}\leq X(\omega)<\frac{n+1}{2^m}\right\}\subset\mathscr{F}$$

如此，对于任意一个 m，我们可以定义一个离散随机变量 X_m：

$$X_m(\omega)=\frac{n}{2^m}\ if\ \omega\in\Lambda_{mn}$$

注意对于任意 m，有：

$$X_m(\omega)\leq X_{m+1}(\omega);0\leq X(\omega)-X_m(\omega)<\frac{1}{2^m},\forall\omega\in\Omega$$

即 X_m 为单调递增的随机变量，从而

$$\lim_{m\to\infty}X_m(\omega)=X(\omega),\forall\omega\in\Omega$$

根据以上离散随机变量的定义，X_m 的期望可以定义为：

$$\mathbb{E}(X_m)=\sum_{n=0}^{\infty}\frac{n}{2^m}\mathscr{P}\left(\left\{\frac{n}{2^m}\leq X(\omega)<\frac{n+1}{2^m}\right\}\right)$$

如果存在一个 m 使得 $E(X_m)=+\infty$，那么定义 $E(X)=+\infty$，否则，定义：

$$\mathbb{E}(X)=\lim_{m\to\infty}\mathbb{E}(X_m)$$

如果上述极限存在且小于 $+\infty$，我们称随机变量 X 是**可积**（integrable）的。

至此我们定义了任意正随机变量的期望。对于任意的随机变量 X，定义 $X^+(\omega)=\max\{X(\omega),0\}$，$X^-(\omega)=\max\{-X(\omega),0\}$，则 $X=X^+-X^-$，$|X|=X^++X^-$。X^+ 和 X^- 都是正的随机变量，如果 $|X|$ 是可积的，那么我们称随机变量 X 是**可积**的，此时可以定义随机变量 X 的期望为：

$$\mathbb{E}(X)=\mathbb{E}\left(X^+\right)-\mathbb{E}\left(X^-\right)$$

至此，任意随机变量的期望即定义完毕。由于该期望是由概率函数 $\mathscr{P}$ 定义的，我们一般记期望为：

$$\mathbb{E}(X)=\int_{\Omega}X(\omega)\mathscr{P}(d\omega)$$

更一般的，对于 $A\in\mathscr{F}$，定义随机变量 X 在集合 A 上的积分为：

$$\int_{A}X(\omega)\mathscr{P}(d\omega)=\mathbb{E}(X\cdot 1_{A})=\int_{\Omega}X(\omega)\cdot 1_{A}(\omega)\mathscr{P}(d\omega)$$

特别的，令 $X(\omega)=1$ 为退化的随机变量，根据以上积分的定义，同样有：

$$\mathscr{P}(A)=\mathbb{E}(1_{A})=\int_{A}\mathscr{P}(d\omega)$$

积分有以下性质：

定理 2. （积分的性质）

1. （线性性）$\int_{A}(aX(\omega)+bY(\omega))\mathscr{P}(d\omega)=a\int_{A}X(\omega)\mathscr{P}(d\omega)+b\int_{A}Y(\omega)\mathscr{P}(d\omega)$
2. （可加性）如果 A_n 不相交，则 $\int_{\bigcup_n A_n}X\mathscr{P}(d\omega)=\sum_n\int_{A_n}X\mathscr{P}(d\omega)$
3. 如果在 A 上，$X\geq 0\,a.s.$，则 $\int_{A}X\mathscr{P}(d\omega)\geq 0$
4. （单调性）如果在 A 上，$X_1\leq X\leq X_2 a.s.$，则 $\int_{A}X_1\mathscr{P}(d\omega)\leq\int_{A}X\mathscr{P}(d\omega)\leq\int_{A}X_2\mathscr{P}(d\omega)$
5. （均值定理）如果在 A 上有 $a\leq X\leq b a.s.$，那么 $a\mathscr{P}(A)\leq\int_{A}X\mathscr{P}(d\omega)\leq b\mathscr{P}(A)$
6. $\left|\int_{A}X\mathscr{P}(d\omega)\right|\leq\int_{A}|X|\mathscr{P}(d\omega)$
7. （有界收敛定理）如果 $\lim_{n\to\infty}X_n=X\ a.s.$，且存在一个常数 M 使得 $\forall n:|X_n|\leq M\ a.s.$，那么

$$\lim_{n\to\infty}\int_{A}X_n\mathscr{P}(d\omega)=\int_{A}X\mathscr{P}(d\omega)=\int_{A}\left(\lim_{n\to\infty}X_n\right)\mathscr{P}(d\omega)\qquad(1)$$

8. （单调收敛定理）如果 $X_n\geq 0$ 且 $X_n\uparrow X\ a.s.$，那么 *(1)* 式成立
9. （控制收敛定理）如果 $\lim_{n\to\infty}X_n=X\ a.s.$，且存在一个随机变量 Y 使得 $|X_n|\leq Y\ a.s.$，且 $\int_{A}Y\mathscr{P}(d\omega)<\infty$，则 *(1)* 式成立。
10. 如果 $\sum_n\int_{A}|X_n|\mathscr{P}(d\omega)<\infty$，那么 $\sum_n|X_n|<\infty\ a.s.\ on\ A$，从而

$$\int_{A}\sum_{n}X_n\mathscr{P}(d\omega)=\sum_{n}\int_{A}X_n\mathscr{P}(d\omega)$$

由于期望即定义为 Ω 上的积分，因而以上性质对于期望都成立，比如由积分的线性性可以得到期望的线性性：

$$\mathbb{E}(aX+bY)=a\mathbb{E}(X)+b\mathbb{E}(Y)$$

而定理 (2.7-2.9) 则解决了极限符号和积分符号互换的问题。注意如果不满足以上定理的条件，积分和极限符号不必然可以互换。

例 8. （积分与极限互换）如果令 $\Omega=\mathbb{R}$，分布函数为：

$$F(\omega)=\begin{cases}0 & \omega<-1\\ \frac{1}{2}\omega+\frac{1}{2} & -1\leq\omega\leq 1\\ 1 & \omega>1\end{cases}$$

即在 $[-1,1]$ 上的均匀分布，进而使用此分布函数构建 $\mathbb{R}$ 上的概率测度 P。随机变量

$$X_n(\omega)=\begin{cases}0 & if\ \ |\omega|>\frac{1}{n}\\ n & if\ \ |\omega|\leq\frac{1}{n}\end{cases}$$

因而除了在一个点 $\omega=0$ 处之外，$\lim_{n\to\infty}X_n(\omega)=0$，或者 $\lim_{n\to\infty}X_n(\omega)=0\ a.s.$，因而[1]

$$\int_\Omega \lim_{n\to\infty}X_nP(d\omega)=0$$

然而由于 $\int_\Omega X_nP(d\omega)=1$，因而 $\lim_{n\to\infty}\int_\Omega X_nP(d\omega)=1$。因而在这个例子里 $\lim_{n\to\infty}\int_\Omega X_nP(d\omega)\neq\int_\Omega\lim_{n\to\infty}X_nP(d\omega)$。

以上我们介绍了数学期望的定义，然而给定一个随机变量，使用概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 计算数学期望非常不方便，我们通常希望使用导出的概率空间 $(\mathbb{R},\mathscr{B},P)$ 计算数学期望，因此我们有以下定理：

定理 3. 如果概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 根据定理 *(1)* 导出了概率空间 $(\mathbb{R},\mathscr{B},P)$，令 g 为一个可测函数，则：

$$\mathbb{E}(g(X))=\int_\Omega g(X(\omega))\,\mathscr{P}(d\omega)=\int_{\mathbb{R}}g(x)\,P(dx)$$

如果等式两边积分都存在。

如果 F 为对应于概率函数 P 的分布函数，则在 $(a,b]$ 上的积分也可以写为：

$$\int_{(a,b]}g(x)\,P(dx)=\int_{(a,b]}g(x)\,dF(x)$$

[1]在定义积分时，当遇到 $0\cdot\infty$ 时，定义 $0\cdot\infty=0$。

因而随机变量 X 的数学期望可以写为：

$$\mathbb{E}(X)=\int_{\mathbb{R}} x dF(x)$$

此外，根据积分以及分布函数的定义：

$$\int_{(a,b]} dF(x)=P((a,b])=F(b)-F(a)$$

例 9. （Cauchy 分布的期望）Cauchy 的密度函数为：

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

如果一个随机变量服从 Cauchy 分布，则：

$$\mathbb{E}(|X|)=\int_{\mathbb{R}}\frac{|x|}{\pi}\frac{1}{1+x^2}dx=\frac{2}{\pi}\int_{[0,\infty)}\frac{x}{1+x^2}dx$$

对于任意正数 M：

$$\int_{[0,M]}\frac{x}{1+x^2}dx=\left.\frac{\log\left(1+x^2\right)}{2}\right|_0^M=\frac{\log\left(1+M^2\right)}{2}$$

因此：

$$\mathbb{E}(|X|)=\frac{1}{\pi}\lim_{M\to\infty}\log\left(1+M^2\right)=\infty$$

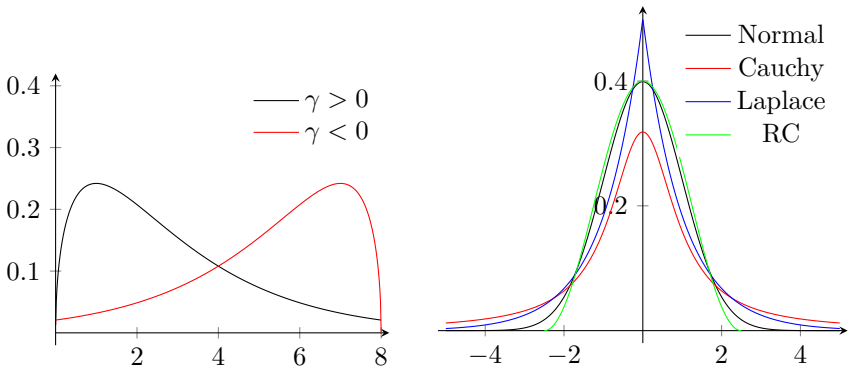
因而该随机变量是不可积的。

对于任意的正数 p，如果 $\mathbb{E}(|X|^p)<\infty$，则记 $X\in L^p=L^p(\Omega,\mathscr{F},\mathscr{P})$。对于整数 r，随机变量 X 的 r **阶矩**被定义为 $\mathbb{E}(X^r)$。一阶矩即为随机变量 X 的期望。此外，随机变量 X 的 r **阶中心矩**被定义为 $\mathbb{E}([X-E(X)]^r)$。特别的，当 $r=2$ 时，2 阶中心矩即为随机变量的**方差**（**variance**），记为 $\mathrm{Var}(X)$ 或者 $\sigma^2(X)$，**标准差**（**standard deviation**）定义为 $\sigma(X)=\sqrt{\mathrm{Var}(X)}$。$X$ 的方差可以使用一阶和二阶矩计算得到：

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left([X-\mathbb{E}(X)]^2\right)=\mathbb{E}\left(X^2-2\mathbb{E}(X)\cdot X+\mathbb{E}(X)^2\right)\\ &= \mathbb{E}\left(X^2\right)-2\mathbb{E}(X)^2+\mathbb{E}(X)^2=\mathbb{E}X^2-(\mathbb{E}X)^2\end{aligned}$$

注意 $\mathbb{E}X^2\geq(\mathbb{E}X)^2$。此外，根据方差定义，有 $\mathrm{Var}(aX+b)=a^2\mathrm{Var}(X)$。

除了前两阶矩之外，经常我们还会关心更高阶的矩。其中，**偏度**（**skewness**）和**峰度**（**kurtosis**）是最经常被关心的高阶矩。其中偏度为随机变量的三阶中心



(a) 左偏与右偏 (b) 尖峰厚尾

图 3: 偏度与峰度

矩，即定义：

$$\gamma = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^3\right)$$

如果 X 为对称分布，那么必然有 $\gamma = 0$。顾名思义，偏度与分布的不对称性有关，如图 (3.a) 所示，当 $\gamma > 0$ 时，分布函数右边的尾巴比较厚，我们称其分布为右偏，反之为左偏。

而峰度则是随机变量的四阶中心矩，即：

$$\operatorname{Kurt}(X) = \mathbb{E}\left(\left[\frac{X - \mathbb{E}(X)}{\sigma(X)}\right]^4\right) = \frac{\mathbb{E}\left([X - \mathbb{E}(X)]^4\right)}{[\operatorname{Var}(X)]^2}$$

尽管我们一般将其称为「峰度」，然而这一称呼并不准确，更准确的称呼应该为「尾厚度」。如图 (3.a) 所示，其中 RC 代表 Raised cosine 分布。如果比较正态分布，会发现正态分布的尾巴比 Raised cosine 分布要厚，而相应的正态分布的峰要「尖」一点，所以我们经常说「尖峰厚尾」。实际上，正态分布的峰度为 3，而图中 Raised cosine 分布的峰度约为 $2.40623 < 3$。

然而注意到，峰度大的并不一定代表峰更「尖」，而仅仅是尾巴更「厚」。如图中 Laplace 分布和 Cauthy 分布相比，Laplace 分布的峰更尖，但是其峰度小于 Cauthy 分布的峰度。实际上，Cauthy 分布的峰度为 $+\infty$。因而判断峰度时，不能顾名思义只看其峰的尖锐程度，而应该看尾巴的厚度。

我们一般会把峰度与正态分布的峰度相比，定义超额峰度（excess kurtosis）为峰度减 3，因而如果超额峰度 > 0，那么其尾巴比正态分布的尾巴要厚，而如果超额峰度 < 0，那么其尾巴要比正态分布的尾巴要薄。

2.2 积分与导数互换

在实际应用中，我们经常需要对积分进行求导，比如：

$$\frac{d}{d\theta}\int_{\mathbb{R}} g(x,\theta)\,dF(x)$$

然而积分常常难以计算，经常我们希望使用：

$$\int_{\mathbb{R}} \frac{dg(x,\theta)}{d\theta}dF(x)$$

来计算。然而这一计算方法是有条件的。

定理 4. （积分与微分互换）如果函数 $g(x,\theta)$ 在 $\theta=\theta_0$ 处可微，即对于任意 x，极限

$$\lim_{\delta\to 0}\frac{g(x,\theta_0+\delta)-g(x,\theta_0)}{\delta}=\frac{\partial}{\partial\theta}g(x,\theta)\bigg|_{\theta=\theta_0}$$

存在，并且存在一个函数 $h(x,\theta_0)$ 以及一个常数 δ_0，有

1. 对于任意的 x 和 $|\delta|\le\delta_0$，有：$\left|\frac{g(x,\theta_0+\delta)-g(x,\theta_0)}{\delta}\right|\le h(x,\theta_0)$
2. $\int_{\mathbb{R}} h(x,\theta_0)\,dF(x)<\infty$

那么：

$$\frac{d}{d\theta}\int_{\mathbb{R}} g(x,\theta)\,dF(x)\bigg|_{\theta=\theta_0}=\int_{\mathbb{R}}\left[\frac{dg(x,\theta)}{d\theta}\bigg|_{\theta=\theta_0}\right]dF(x)$$

该定理即控制收敛定理的应用。该定理表明，在一定的条件下，积分与微分操作可以交换顺序。在该条件成立下，进一步我们有莱布尼茨法则：

定理 5. （***Leibnitz* 法则**）如果 $g(x,\theta),a(\theta),b(\theta)$ 对 θ 可微，那么：

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x,\theta)\,dx=g(b(\theta),\theta)\frac{d}{d\theta}b(\theta)-g(a(\theta),\theta)\frac{d}{d\theta}a(\theta)+\int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}g(x,\theta)\,dx$$

2.3 常用不等式

下面我们介绍几个常用的不等式。

定理 6. （***Chebyshev* 不等式**）如果函数 ψ 满足：$\psi(u)=\psi(-u)\ge 0$，且在 $(0,\infty)$ 上单调递增，X 为随机变量，且 $\psi(X)<\infty$，则对于 $u>0$，有：

$$\mathscr{P}(|X|\ge u)\le\frac{\mathbb{E}[(\psi(X))]}{\psi(u)}$$

Proof. 根据均值定理：

$$\mathbb{E}[\psi(X)]=\int_{\Omega}\psi(X)\,\mathscr{P}(d\omega)\ge\int_{\{|X|\ge u\}}\psi(X)\,\mathscr{P}(d\omega)\ge\psi(u)\,\mathscr{P}(|X|\ge u)$$

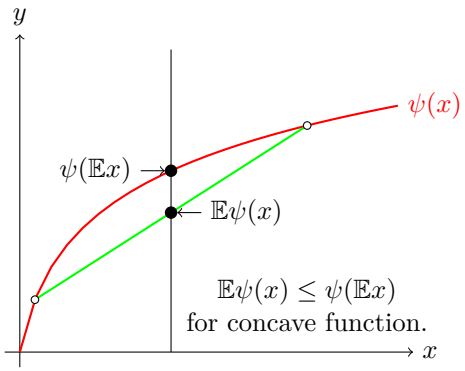


图 4: Jensen 不等式

□

特别的，令 $Y=\frac{[X-\mathbb{E}(X)]}{\sigma(X)}$，则 $\mathbb{E}(Y)=0$，$\mathbb{E}\left(Y^2\right)=1$。令 $\psi(x)=x^2$，有：

$$\mathscr{P}(|Y|\geq u)\leq\frac{1}{u^2}$$

定理 7. （***Jensen* 不等式**）如果 ψ 为凹函数，且随机变量 X 和 $\psi(X)$ 可积，则：

$$\psi(\mathbb{E}(X))\geq\mathbb{E}[\psi(X)]$$

Jensen 不等式表明，对于一般的非线性函数，期望的函数与函数的期望并不相等，如图 (4) 所示。

作为 Jensen 不等式的一个应用，考虑 $0<r<s$，并令 $p=\frac{s}{r}>1$，注意 $\psi(x)=|x|^p$ 为凸函数，因而根据 Jensen 不等式，有：

$$\mathbb{E}(|X|^{rp})=\mathbb{E}(|X|^s)\geq[\mathbb{E}(|X|^r)]^p=[\mathbb{E}(|X|^r)]^{\frac{s}{r}}$$

整理之后可以得到：

$$\mathbb{E}(|X|^r)\leq[\mathbb{E}(|X|^s)]^{\frac{r}{s}}$$

以上不等式被称为 **Liapounov 不等式**。

定理 8. （***Cauchy-Schwarz* 不等式**）对于两个随机变量 X,Y，若满足可积性，有：

$$|\mathbb{E}(XY)|\leq\mathbb{E}(|XY|)\leq\left[\mathbb{E}\left(|X|^2\right)\right]^{\frac{1}{2}}\left[\mathbb{E}\left(|Y|^2\right)\right]^{\frac{1}{2}}$$

3 随机变量的变换

对于一个可测函数 $g(\cdot)\mathbb{R}\to\mathbb{R}$，$Y=g(X)$ 也是一个随机变量。如果我们已知 X 的分布，如何获得新的随机变量 Y 的分布呢?

首先仿照随机变量的定义，我们定义函数 $g(\cdot)$ 对于一个集合的逆为：

$$g^{-1}(A)=\{x\in\mathbb{R}:g(x)\in A\}$$

因而对于单点集 $g^{-1}(\{y\})=\{x\in\mathbb{R}:g(x)=y\}$。

对于离散型的随机变量 X，$g(X)$ 也是离散型的，因而随机变量 Y 的 p.m.f. 为：

$$f_Y(y)=P(Y=y)=\sum_{x\in g^{-1}(\{y\})}P(X=x)=\sum_{x\in g^{-1}(\{y\})}f_X(x)$$

由此可以确定随机变量 $Y=g(X)$ 的概率质量函数。

对于一个一般的随机变量 X，随机变量 $Y=g(X)$ 的累积分布函数可以使用如下定义计算：

$$\begin{aligned}F_Y(y) &= P(Y\leq y)\\ &= P(g(X)\leq y)\\ &= P(\{x:g(x)\leq y\})\\ &= \int_{\{x:g(x)\leq y\}}dF_X\end{aligned}$$

特别的，如果 $g(\cdot)$ 严格单调递增，则：

$$F_Y(y)=F_X\left(g^{-1}(y)\right)$$

如果 $g(\cdot)$ 严格单调递减，则：

$$F_Y(y)=1-F_X\left(g^{-1}(y)\right)$$

现在，假设我们有一个随机变量 $U\sim Uniform(0,1)$，即 $(0,1)$ 区间上的均匀分布，而 $F(\cdot)$ 为一个分布函数，我们可以定义一个新的随机变量 $X=F^{-1}(U)$。注意如果分布函数 $F(\cdot)$ 存在「平台」，即 $F(x)=c,\ for\ a\leq x<b$，那么定义 $F^{-1}(u)=\inf\{x:F(x)\geq u\}$。注意对于一个存在「平台」的分布函数 F，$F(x)=c,\ for\ a\leq x<b$，$P(X\leq x)=P(X\leq a),\ for\ a\leq x<b$，也就是说如果某个随机变量的分布函数为 F，那么其取值在 $[a,b)$ 范围内的概率为 0。根据以上推理，随机变量 $X=F^{-1}(U)$ 的分布函数为：

$$\begin{aligned}F_X(x) &= \int_{F^{-1}(u)\leq x}dG(u)\\ &= \int_{F^{-1}(u)\leq x}du\\ &= \int_{u\leq F(x)}du\\ &= F(x)\end{aligned}$$

其中第二个等号由于均匀分布的分布函数 $G(u)=u, 0\leq u\leq 1$。这也就意味着，如果我们有分布函数 $F(\cdot)$，可以生成一个 $(0,1)$ 区间内的随机变量 U，进而生成 $X=F^{-1}(U)$，那么我们就生成了一个新的随机变量，其分布函数为 F。

例 10. （指数分布）若 $U\sim Uniform(0,1)$，令 $F(x)=1-e^{-x}, x>0$，即指数分布的分布函数。令 $X=F^{-1}(U)$，则随机变量 X 的分布函数为:

$$\begin{aligned} F_X(x) &= \int_{F^{-1}(u)\leq x} dG(u) \\ &= \int_{F^{-1}(u)\leq x} du \\ &= \int_{u\leq(1-e^{-x})} du \\ &= 1-e^{-x} \end{aligned}$$

因而为了生成指数分布的随机变量，只要生成均匀分布 U，并令 $X=-\ln u$ 即可。

在绝大多数计算机语言中，都有生成均匀分布的指令。比如在 C 语言中，可以使用标准库 <stdlib.h> 中的 rand() 函数生成均匀分布:

```
double x=(double)rand()/RAND_MAX;
```

在 Python 中，可以使用 Python 标准库的 random 包，或者 NumPy 的 random 包，比如:

```
import random as rd
x=rd.random()
```

或者:

```
import numpy.random as nprd
x=nprd.random()
```

即可以生成均匀分布。在 Excel 中，也可以使用 RAND() 生成均匀分布的随机数。

习题

练习 1. 在研究中，对收入等变量取对数是非常常见的处理手段。如果 $X>0$ 代表总体收入，那么 $\mathbb{E}(X)$ 和 $\exp[\mathbb{E}(\ln X)]$ 哪一个更大?

练习 2. 如果 r.v.$X\sim Binomial(n,p)$，求随机变量 $Y=n-X$ 的概率质量函数。

练习 3. 求对数正态分布（$Y = e^X, X \sim N(0,1)$）的概率密度函数。

练习 4. 证明：对于一个随机变量 $X \sim F_X$，随机变量 $Y = F_X(X) \sim Uniform(0,1)$。

练习 5. 使用任何编程语言，通过均匀分布生成 100 个 Logistic 分布（分布函数 $\frac{e^x}{1+e^x}$）的随机数，并将理论的分布函数及其经验分布函数画在一张图中。其中经验分布函数的定义为：

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(X_i \leq x)$$

即样本中小于 x 的样本的 s 比例。观察两者是否贴近？生成 1000 个数据，重复以上步骤，并比较两张图的差异。

练习 6. 使用任何编程语言，通过均匀分布生成 100 个服从泊松分布（$\lambda = 1$）的随机数，并计算均值。

参考文献

[1] Ash, R.B., Doleans-Dade, C., 2000. Probability and measure theory. Academic Press.

[2] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.

[3] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[4] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[5] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

第三节 · 常用分布

司继春

上海对外经贸大学统计与信息学院

1 离散分布

1.1 离散均匀分布

随机变量 X 服从离散均匀分布如果 $P(X=x)=\frac{1}{N}, x\in(a_1,a_2,...,a_N)$。

- 期望: $\mathbb{E}(X)=\frac{1}{N}\sum_{i=1}^{N}a_i$
- 方差: $\mathrm{Var}(X)=\frac{1}{N}\sum_{i=1}^{N}(a_i-\mathbb{E}X)^2$

1.2 伯努利分布

伯努利分布**即伯努利试验**（**Bernoulli trial**）的结果。伯努利试验即试验结果只有两种可能性（成功, 失败），独立重复 N 次的试验结果。一个随机变量 X 定义为 $X=1\,if$ 成功, $=0\,if$ 失败。如果成功的概率为 p，那么**伯努利分布**（**Bernoulli Distribution**）$X\sim Ber(p)$ 即:

$$X=\begin{cases}1 & with\ probability\ p\\ 0 & with\ probability\ 1-p\end{cases}\quad 0\le p\le 1$$

- 期望: $\mathbb{E}(X)=p$
- 方差: $\mathrm{Var}(X)=p(1-p)$

1.3 二项分布

二项分布即独立重复 N 次伯努利实验，成功次数 $Y=\sum_{i=1}^{N}X_i, X_i\sim Ber(p)$ 的分布。简单计算可以得到，随机变量 Y 的概率质量函数:

$$P(Y=y|N,p)=\left(\begin{array}{c}N\\ y\end{array}\right)p^y(1-p)^{N-y}, y=1,2,...,N$$

我们成 Y 服从**二项分布**（**Binomial Distribution**），记为 $Y\sim Bi(N,p)$。由于对于任意的实数 a,b 以及任意整数 $N\ge 0$, 有: $(a+b)^N=\sum_{i=0}^{N}\left(\begin{array}{c}N\\ i\end{array}\right)a^ib^{n-i}$,

因而：

$$\sum_{y=1}^{N} P\left(Y=y|N,p\right)=\sum_{y=0}^{N}\left(\begin{array}{c} N \\ y \end{array}\right)p^{y}\left(1-p\right)^{N-y}=\left(p+1-p\right)^{N}=1$$

- 期望：$\mathbb{E}\left(Y\right)=Np$
- 方差：$\mathrm{Var}\left(Y\right)=Np\left(1-p\right)$

1.4 负二项分布

二项分布关心在 N 次伯努利实验中成功的次数，而有时我们会关心为了达到 r 次成功所必须要做的试验的次数。记 Z 为为了达到 r 次成功所需的试验次数，那么事件 $Z=z$ 即 z 次试验达到了 r 次成功，这一事件可以进一步分解为在前 $z-1$ 次试验中得到了 $r-1$ 次成功，而第 r 次也成功了。所以其概率质量函数为：

$$\begin{aligned} P\left(Z=z|r,p\right) &= \left(\begin{array}{c} z-1 \\ r-1 \end{array}\right)p^{r-1}\left(1-p\right)^{z-r}\cdot p \\ &= \left(\begin{array}{c} z-1 \\ r-1 \end{array}\right)p^{r}\left(1-p\right)^{z-r},z=r,r+1,... \end{aligned}$$

我们称随机变量 Z 服从**负二项分布**（**Negative Binomial Distribution**），记为 $Z\sim NB\left(r,p\right)$。

下面计算 Z 的期望：

$$\begin{aligned} E\left(Z\right) &= \sum_{z=r}^{\infty} z\cdot\left(\begin{array}{c} z-1 \\ r-1 \end{array}\right)p^{r}\left(1-p\right)^{z-r} \\ &= \sum_{z=r}^{\infty} z\cdot\frac{\left(z-1\right)!}{\left(r-1\right)!\left(z-r\right)!}p^{r}\left(1-p\right)^{z-r} \\ &= \frac{r}{p}\sum_{z=r}^{\infty}\frac{z!}{r!\left(z-r\right)!}p^{r+1}\left(1-p\right)^{z-r} \\ &\overset{z'=z+1,r'=r+1}{=} \frac{r}{p}\sum_{z'=r+1}^{\infty}\frac{\left(z'-1\right)!}{\left(r'-1\right)!\left(z'-r'\right)!}p^{r'}\left(1-p\right)^{z'-r'} \\ &\overset{NB\left(r+1,p\right)\text{的质量函数和为 }1}{=} \frac{r}{p} \end{aligned}$$

同理可计算其方差。

- 期望：$\mathbb{E}\left(Z\right)=\frac{r}{p}$
- 方差：$\mathrm{Var}\left(Z\right)=\frac{r(1-p)}{p^{2}}$

1.5 几何分布

几何分布（**Geometric Distribution**）是最简单形式的负二项分布，如果一个随机变量 $V \sim NB(1,p)$，则随机变量 V 服从几何分布：

$$P(V=v|p) = p(1-p)^{v-1}$$

我们记 $V \sim G(p)$。几何分布具有无记忆性，即：

$$P(V>s|V>t) = P(V>s-t), s>t$$

即在给定我们为了等待成功已经等了 t 次的情况下，还需要等待的次数跟已经等待的次数是没有关系的。比如北京的车牌摇号，10 次没有中签的人跟 100 次没有中签的人，其需要继续等待时间的分布是一样的。因而此分布不适于建模带有生命长度的问题。

- 期望：$\mathbb{E}(V) = \frac{1}{p}$
- 方差：$\mathrm{Var}(V) = \frac{1-p}{p^2}$

1.6 超几何分布

即在一个 N 个试验组成的总体中，有已知 K 次成功，从这 N 个总体中抽取出 n 个样本，抽到的成功次数 $M=k$ 的概率。我们称随机变量 M 服从**超几何分布**（**Hypergeometric Distribution**），其概率质量函数为：

$$P(M=k|N,K,n) = \frac{\begin{pmatrix} K \\ k \end{pmatrix}\begin{pmatrix} N-K \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

- 期望：$\mathbb{E}(M) = \frac{n}{N} \cdot K$
- 方差：$\mathrm{Var}(M) = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$

1.7 泊松分布

在第一节中我们研究了到达次数问题，相应的，如果随机变量 X 的概率质量函数为：

$$P(X=x) = \frac{\lambda^x}{x!}e^{-\lambda}, x=0,1,2,...$$

那么我们称随机变量 X 服从**泊松分布**（**Poisson Distribution**），记为 $X \sim P(\lambda)$。泊松分布经常被用来建模次数问题。

- 期望：$\mathbb{E}(X) = \lambda$

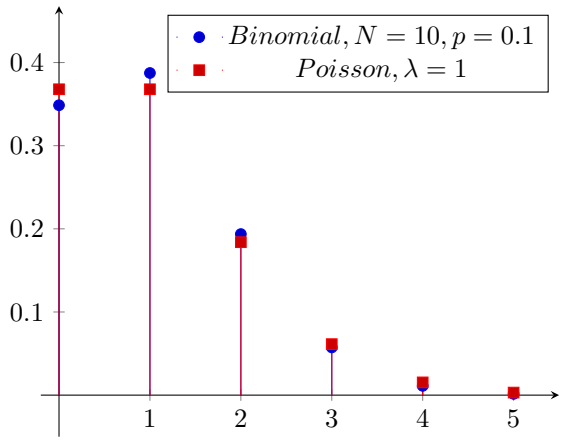


图 1: 泊松分布与二项分布

- 方差: $\mathrm{Var}(X) = \lambda$

根据第一节的例题中，我们已经知道如果一个随机变量 $Y \sim Bi(N, p)$，令 $\lambda = Np$，并令 $N \to \infty$，那么随机变量 Y 近似服从泊松分布，因而对于相对较大的 N，可以使用泊松分布去近似二项分布。

例 1. 如果一个人打字出错的概率为 $p = 0.001$，那么他写 1000 字出错的概率应该服从二项分布。记出错的字数为 Y，则 $Y \sim Bi(1000, 0.001)$。比如写 1000 字出错两个以内的概率为::

$$P(Y \leq 2) = \sum_{y=0}^{2} \begin{pmatrix} 1000 \\ y \end{pmatrix} 0.001^y 0.999^{1000-y} \approx 0.91979$$

以上计算比较繁琐，如果使用泊松逼近，令 $X \sim P(1)$:

$$P(Y \leq 2) \approx P(X \leq 2) = \sum_{y=0}^{2} \frac{1^y}{y!} e^{-1} = 0.91970$$

两者非常接近。

2 连续分布

2.1 均匀分布

如果随机变量 X 在区间 $[a, b]$ 内的概率密度函数为常数，即:

$$f(x|a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & otherwise \end{cases}$$

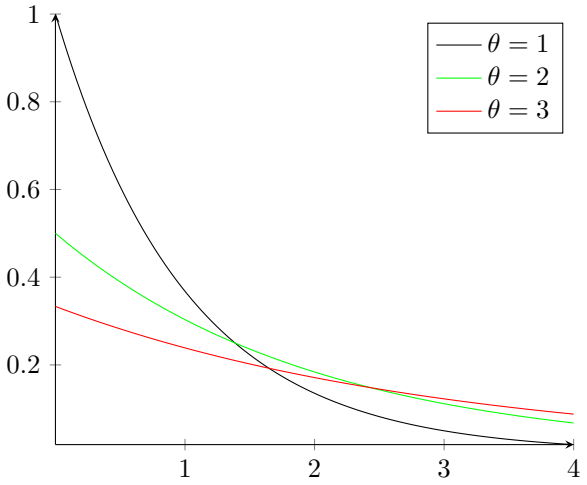


图 2: 指数分布密度函数

则称随机变量 X 服从区间 $[a,b]$ 上的**均匀分布**（**Uniform Distribution**），记为 $X \sim U(a,b)$。

- 分布函数：

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

- 期望：$\mathbb{E}(X) = \frac{a+b}{2}$
- 方差：$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

2.2 指数分布

若随机变量 X 的概率密度函数为：

$$f(x|a,\beta) = \frac{1}{\beta} e^{-\frac{x-a}{\beta}}, x > a$$

那么我们称 X 服从**指数分布**（**Exponential Distribution**），其密度函数如图 (2) 所示，记为 $X \sim E(a,\beta)$。

- 分布函数：$F(x) = 1 - e^{-\frac{x-a}{\beta}}, x > a$
- 期望：$\mathbb{E}(X) = \beta + a$
- 方差：$\mathrm{Var}(X) = \beta^2$

例 2.（指数分布）如果一个随机变量 $Y \sim P(\lambda)$，其中参数 λ 代表一段时间 T 内到达的平均次数，所以对于时间 t 内，平均到达的人数即 $\lambda\frac{t}{T}$，因而在 t 时间内只有 0 次到达的概率为 $e^{-\frac{\lambda t}{T}}$。如果在时间 t 内有 0 次到达，意味着等待时间大于 t，因而等待时间 X 大于 t 的概率 $P(X>t)=e^{-\frac{\lambda t}{T}}$，所以 $P(X\leq t)=1-e^{-\frac{\lambda t}{T}}$，因而等待时间 X 服从指数分布。

同样，指数分布也具有无记忆性，即如果 $X \sim E(0,\beta)$，那么

$$P(X>s|X>t)=P(X>s-t),s>t$$

以为着为了等待第一个人到达，如果已经等了 t 时间，那么继续等待的时间仍然服从指数分布：给定已经等了五分钟的情况下，继续等待五分钟有到达的概率与第一个五分钟有到达的概率是一样的。

现在考虑，如果一件产品的使用寿命为 T，其分布函数为 $F(t)$，那么寿命大于 t 的概率为 $S(t)=1-F(t)$，我们称之为**生存函数**（**Survival function**）。很多时候，我们关心一件产品已经使用了时间 t，在 $(t,t+dt)$ 一小段时间内死亡的概率，如果令 $dt\to 0$，即瞬时的死亡风险。可以计算：

$$\lambda(t)=\lim_{dt\to 0}\frac{P(t\leq T<t+dt)}{dt\cdot S(t)}=\frac{f(t)}{S(t)}=-\frac{S'(t)}{S(t)}=-\frac{d}{dt}\ln S(t)$$

我们称 $\lambda(t)$ 为**风险函数**（**hazard function**）。指数分布的风险函数为：

$$\lambda(t)=-\frac{d}{dt}\ln S(t)=-\frac{d}{dt}\ln\left(e^{-\frac{x}{\beta}}\right)=\frac{1}{\beta}$$

因而指数函数的风险函数为常数，即其死亡（到达）的可能性不随着时间的增加而增加（或者减少）。

2.3 伽马分布

在介绍伽马分布之前，我们需要先介绍伽马函数（Γ Function）。我们定义 Γ 函数为：

$$\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}dt$$

易知当 $\alpha>0$ 时，$\Gamma(\alpha)<\infty$，而当 $\alpha\leq 0$ 时，该积分不一定有限。图 (3) 展示了 Γ 函数的逆函数。

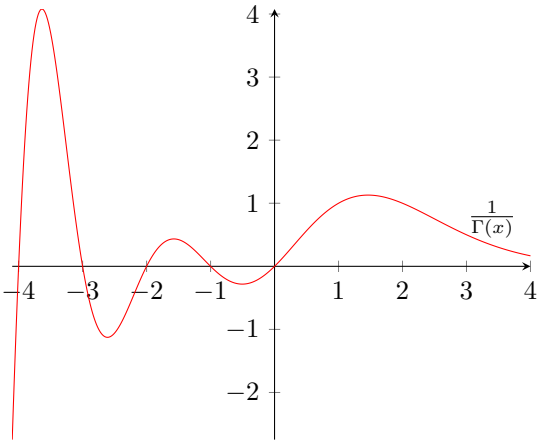


图 3: $\Gamma^{-1}(\alpha)$

根据 Γ 函数的定义，可知 $\Gamma(1)=1$，而当 $\alpha>0$ 时，有：

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty t^\alpha e^{-t}dt \\ &= -\int_0^\infty t^\alpha de^{-t} \\ &= -t^\alpha e^{-t}\big|_0^\infty + \int_0^\infty e^{-t}dt^\alpha \\ &= \alpha\int_0^\infty t^{\alpha-1}e^{-t}dt \\ &= \alpha\Gamma(\alpha)\end{aligned}$$

结合 $\Gamma(1)=1$，可以得到对于任意的正整数 n，$\Gamma(n)=(n-1)!$。此外，$\Gamma\left(\frac{1}{2}\right)=\int_0^\infty \frac{1}{\sqrt{t}}e^{-t}dt=\sqrt{\pi}$，因而：

$$\begin{aligned}\int_0^\infty \frac{1}{\sqrt{t}}e^{-t}dt &\overset{t=\frac{x^2}{2}}{=} \int_0^\infty \frac{\sqrt{2}}{x}e^{-\frac{x^2}{2}}d\frac{x^2}{2} \\ &= \sqrt{2}\int_0^\infty e^{-\frac{x^2}{2}}dx \\ &= \sqrt{\pi}\end{aligned}$$

因而

$$\int_0^\infty e^{-\frac{x^2}{2}}dx=\sqrt{\frac{\pi}{2}} \tag{1}$$

根据 Γ 函数的定义，函数：

$$f(t|\alpha)=\frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)},t>0$$

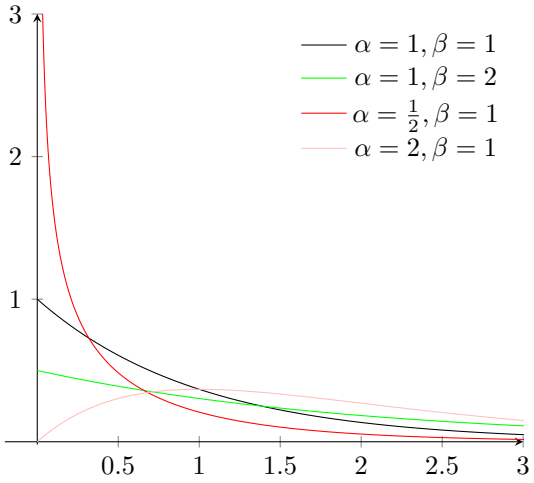


图 4: $\Gamma(\alpha, \beta)$ 分布的密度函数

为一个密度函数。如果令 $T \sim f(t|\alpha)$，那么对于任意的 $\beta > 0$，我们称 $X = \beta T$ 服从**伽马分布**（**Gamma Distribution**），记为 $X \sim \Gamma(\alpha, \beta)$，其密度函数为：

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, x > 0, \alpha > 0, \beta > 0$$

图 (4) 展示了不同参数下 $\Gamma(\alpha, \beta)$ 分布的密度函数。可知，指数分布为 $\alpha = 1$ 时 Γ 分布的特例。其期望：

$$\begin{aligned} E(X) &= \int_0^{\infty} x f(x|\alpha, \beta)\, dx \\ &= \int_0^{\infty} x \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \Gamma(\alpha+1)\beta^{\alpha+1} \\ &= \alpha\beta \end{aligned}$$

类似地，可以计算其方差：

- 期望：$\mathbb{E}(X) = \alpha\beta$
- 方差：$\mathrm{Var}(X) = \alpha\beta^2$

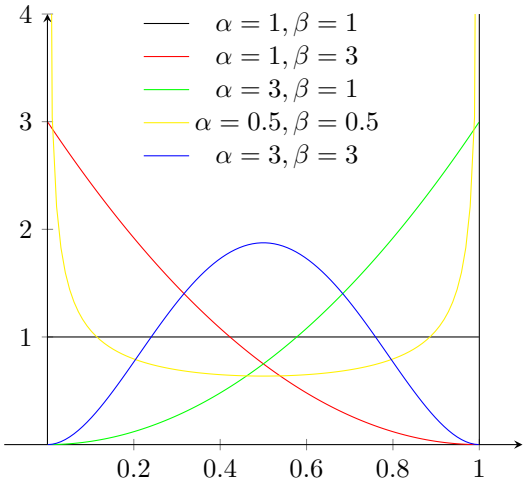


图 5: 指数分布密度函数

2.4 Beta 分布

如果一个随机变量 $X \in (0,1)$，其分布函数为：

$$f(x|a,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0$$

那么我们称 X 服从 **Beta 分布**（**Beta Distribution**）。其中

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

图 (5) 给出了 Beta 分布的密度函数。当 $\alpha > 1, \beta = 1$ 时，Beta 分布的密度函数为单调递增的；当 $\alpha = 1, \beta > 1$ 时，Beta 分布的密度函数为单调递减的；当 $\alpha < 1, \beta < 1$ 时，Beta 分布的密度函数为 U 型的；当 $\alpha > 1, \beta > 1$ 时，Beta 分布的密度函数为钟型的；当 $\alpha = 1, \beta = 1$ 时，Beta 分布的密度函数退化为均匀分布。由于 Beta 分布的取值范围在 $(0,1)$ 范围内，因而 Beta 分布经常被用于研究比率等问题，或者在贝叶斯分析中作为概率的先验。

Beta 分布的 r 阶矩可以计算为：

$$\begin{aligned}
E(X^r) &= \frac{1}{B(\alpha,\beta)} \int_0^1 x^r x^{\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{r+\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{B(r+\alpha,\beta)}{B(\alpha,\beta)} \frac{1}{B(r+\alpha,\beta)} \int_0^1 x^{r+\alpha-1}(1-x)^{\beta-1}\,dx \\
&= \frac{B(r+\alpha,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+r)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+r)\Gamma(\alpha)}
\end{aligned}$$

从而：

- 期望：$\mathbb{E}(X)=\frac{\alpha}{\beta+a}$
- 方差：$\mathrm{Var}(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

2.5 正态分布

如果一个随机变量 X 潜在受到非常多的独立因素的影响，即 $X=f_1+f_2+\cdots$，而每个 f_i 又不能单独对 X 有非常大的影响，那么一般来说 X 将会服从**正态分布**（**Normal Distribution**）或者**高斯分布**（**Gaussian Distribution**）。正态分布的密度函数为：

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

记为 $X\sim N\left(\mu,\sigma^2\right)$。

- 期望：$\mathbb{E}(X)=\mu$
- 方差：$\mathrm{Var}(X)=\sigma^2$

如果 $X\sim N\left(\mu,\sigma^2\right)$，如果令 $Z=\frac{X-\mu}{\sigma}$，则 $E(Z)=\frac{E(X)-\mu}{\sigma}=0$，$\mathrm{Var}(Z)=\frac{1}{\sigma^2}\cdot\mathrm{Var}(X)=1$，随机变量 Z 服从 $\mu=0,\sigma=1$ 的正态分布，即 $Z\sim N(0,1)$，我们称之为**标准正态分布**（**Standard Normal Distribution**）。标准正态分布的密度函数为：

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

根据公式 (1) 可知，其密度函数在 $\mathbb{R}$ 上的积分为 1。而其分布函数为：

$$\Phi(x)=\int_{-\infty}^{x}\phi(t)\,dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$$

由于 $\Phi(x)$ 没有显示解，因而一般我们用 $\Phi(z)$ 来代表标准正态的分布函数。

图 (6) 列出了标准正态分布的密度函数以及分布函数。由于正态分布为对称分布，即 $\phi(x)=\phi(-x)$，因而分布函数 $\Phi(x)=1-\Phi(-x)$。如果 $X\sim N\left(\mu,\sigma^2\right)$，那么 $Z=\frac{X-\mu}{\sigma}\sim N(0,1)$，根据标准正态的分布函数 $\Phi(x)$ 可以计算 X 在区间内取值的概率，比如：

$$P(|X-\mu|\le\sigma)=P(|Z|<1)=\Phi(1)-\Phi(-1)=2\Phi(1)-1=0.6827$$

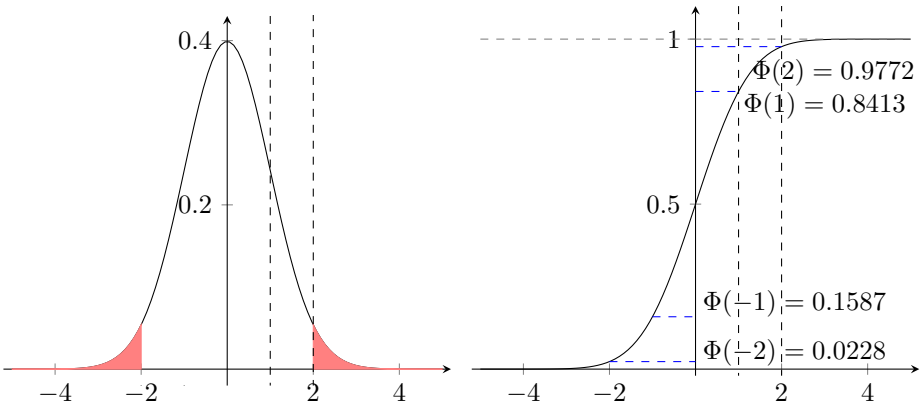


图 6: 标准正态分布密度函数与分布函数

同理，有：

$$\begin{aligned}
P\left(|X-\mu| \leq 1.65\sigma\right) &= P\left(|Z| < 1.65\right) \approx 0.90 \\
P\left(|X-\mu| \leq 1.96\sigma\right) &= P\left(|Z| < 1.96\right) \approx 0.95 \\
P\left(|X-\mu| \leq 2\sigma\right) &= P\left(|Z| < 2\right) = 0.9545 \\
P\left(|X-\mu| \leq 2.58\sigma\right) &= P\left(|Z| < 2.58\right) \approx 0.99 \\
P\left(|X-\mu| \leq 3\sigma\right) &= P\left(|Z| < 3\right) = 0.9973 \\
P\left(|X-\mu| \leq 5\sigma\right) &= P\left(|Z| < 5\right) \geq 1 - 10^{-6} \\
P\left(|X-\mu| \leq 6\sigma\right) &= P\left(|Z| < 6\right) \geq 1 - 10^{-8}
\end{aligned}$$

正态分布有很多优良的性质，因而在建模中是最经常被使用的一种分布。比如，如果两个随机变量 X, Y 独立，且 $X \sim N\left(\mu_1, \sigma_1^2\right), Y \sim N\left(\mu_2, \sigma_2^2\right)$，那么 $V = X + Y \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$，即两个独立的正态分布的和仍然是正态分布。在下一节多元随机变量中，我们还将介绍更多的正态分布独有的性质。

2.6 对数正态分布

若随机变量 $Y = e^X, X \sim N\left(\mu, \sigma^2\right)$，则我们称随机变量 Y 服从**对数正态分布**（**Lognormal Distribution**），记为 $Y \sim LN\left(\mu, \sigma^2\right)$。

- 期望：$\mathbb{E}\left(Y\right) = e^{\mu + \frac{\sigma^2}{2}}$
- 方差：$\mathrm{Var}\left(Y\right) = e^{2\left(\mu+\sigma^2\right)} - e^{2\mu+\sigma^2}$

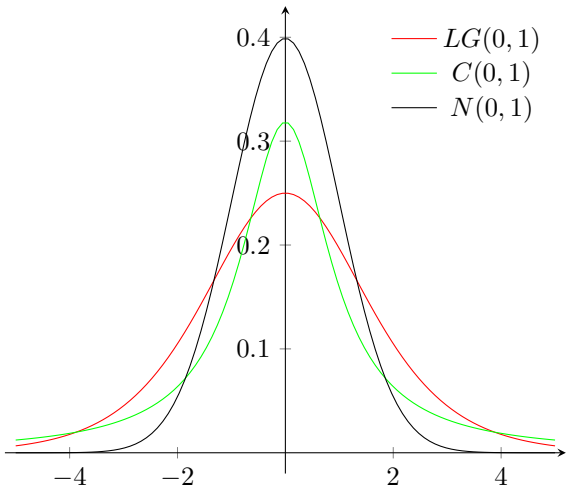


图 7: Logistic 分布、Cauchy 分布与正态分布密度函数

2.7 逻辑斯蒂分布

若随机变量 X 的分布函数为:

$$F(x|\mu,\sigma) = \frac{e^{\frac{x-\mu}{\sigma}}}{1+e^{\frac{x-\mu}{\sigma}}}$$

那么我们称 X 服从**逻辑斯蒂分布**（**Logistic Distribution**），其密度函数为:

$$f(x|\mu,\sigma) = \frac{1}{\sigma}\frac{e^{-\frac{x-\mu}{\sigma}}}{\left[1+e^{-\frac{x-\mu}{\sigma}}\right]^2}$$

记为: $X \sim LG(\mu,\sigma)$。Logistic 分布被广泛运用在**离散选择**（**Discrete Choice**）模型或者**分类器**（**Classifier**）中。

- 期望: $\mathbb{E}(X) = \mu$
- 方差: $\mathrm{Var}(X) = \sigma^2\frac{\pi^2}{3}$

2.8 柯西分布

若随机变量 X 的概率密度函数为:

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

那么我们称 X 服从**柯西分布**（**Cauchy Distribution**），记为 $X \sim C(\mu,\sigma)$。

- 期望: 不存在

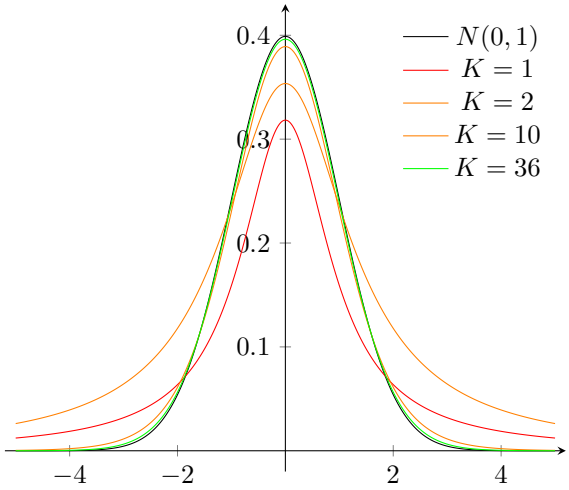


图 8: t 分布与正态分布密度函数

- 方差：不存在

由于柯西分布的一阶矩不存在，因而通常作为不可积的分布的例子。此外，可以得到，两个独立的标准正态分布 X, Y，其比例 $U = \frac{Y}{X}$ 刚好服从柯西分布。因而在实践中计算两个数的比例时需要特别小心。

2.9 χ^2 分布

K 个独立的正态分布的平方和的分布被称为**卡方分布**（**Chi-square Distribution**）或者 χ^2 **分布**。即，如果 $X_1, X_2, ..., X_K$ 为 K 个**独立**的**标准**正态分布，那么

$$X = \sum_{i=1}^{K} X_i^2 \sim \chi_K^2$$

其中参数 K 为卡方分布的**自由度**（**degrees of freedom**）。χ^2 分布的密度函数为：

$$f\left(x|K\right) = \frac{1}{\Gamma\left(K/2\right) 2^{K/2}} x^{K/2-1} e^{-x/2}, x > 0$$

因而，χ^2 分布实际是 Γ 分布的特殊形式（$\alpha = K/2, \beta = 2$）。

- 期望：$\mathbb{E}\left(X\right) = K$
- 方差：$\mathrm{Var}\left(X\right) = 2K$

χ^2 分布在假设检验中非常常用。

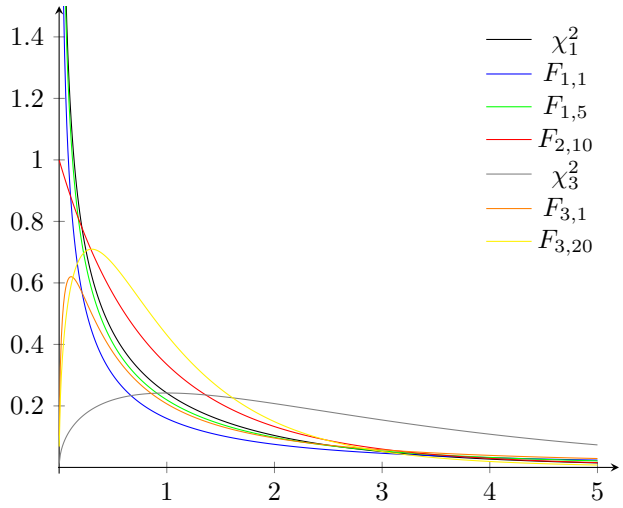


图 9: F 分布与 χ^2 分布密度函数

2.10 学生氏 t 分布

如果存在一个标准正态分布 $Z \sim N(0,1)$，以及一个 χ^2 分布 $X \sim \chi_K^2$，且 Z 和 X **独立**，那么随机变量 $T = \frac{Z}{\sqrt{X/K}}$ 即服从**学生氏 t 分布**（**Students't-distribution**）或者简称 t **分布**（**t-distribution**），记为 $T \sim t_K$，参数 K 为**自由度**。t 分布的密度函数为：

$$f(x|K) = \frac{\Gamma[(K+1)/2]}{\sqrt{K\pi}\Gamma(K/2)}\left(1+\frac{x^2}{K}\right)^{-\frac{K+1}{2}} = \frac{1}{\sqrt{K}B(1/2,K/2)}\left(1+\frac{x^2}{K}\right)^{-\frac{K+1}{2}}$$

当 $K=1$ 时，t 分布即退化为柯西分布，而当 $K \to \infty$ 时，t 分布趋向于正态分布。图 (8) 显示了不同自由度下的 t 分布与正态分布的比较。

- 期望：$\mathbb{E}(X) = 0, K > 1$
- 方差：$\mathrm{Var}(X) = \frac{K}{K-2}, K > 2$

t 分布在假设检验中也扮演着至关重要的地位。

2.11 F 分布

如果存在两个**独立**的 χ^2 分布 $X_1 \sim \chi_n^2, X_2 \sim \chi_m^2$，那么随机变量 $F = \frac{X_1/n}{X_2/m}$ 即服从 F **分布**（**F-distribution**），记为 $T \sim F_{n,m}$，参数 n, m 为**自由度**。F 分布的密度函数为：

$$f(x|K) = \frac{n^{n/2}m^{m/2}\Gamma[(n+m)/2]}{\Gamma(n/2)\Gamma(m/2)(m+nx)^{(n+m)/2}}x^{\frac{n}{2}-1}, x > 0$$

图 (9) 显示了不同自由度下的 χ^2 分布与 F 分布的密度函数。

- 期望：$\mathbb{E}(X)=\frac{m}{m-2}, m>2$
- 方差：$\operatorname{Var}(X)=\frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}, m>4$

F 分布与其他多种分布有关系：

- 如果 $X_k\sim\Gamma(\alpha_k,\beta_k)$ 且独立，那么 $\frac{\alpha_2\beta_1X_1}{\alpha_1\beta_2X_2}\sim F_{2\alpha_1,2\alpha_2}$
- 如果 $X\sim B(n/2,m/2)$，那么 $\frac{mX}{n(1-X)}\sim F_{n,m}$，反之亦成立
- 如果 $X\sim F_{n,m}$，那么 $\lim_{m\to\infty}nX\sim\chi^2_n$
- 如果 $T\sim t_K$，那么 $T^2\sim F_{1,K}$

3 分布族

给定任意的一个 $(\Omega,\mathcal{F})$，在这个空间里面我们可以定义各种不同的概率函数，统计学需要解决的问题就是使用样本数据去推断总体的概率函数 $\mathscr{P}$。然而一般情况下，概率函数 $\mathscr{P}$ 的可能性太多，因而我们经常把研究的重点放在一个概率函数的可能集合内，并在此集合内部对总体的概率函数做出推断。为此我们可以定义参数族的概念。

定义 1. 如果对于每一个已知的 $\theta\in\Theta,\Theta\subset\mathbb{R}^d$，$P_\theta$ 为在 $(\Omega,\mathcal{F})$ 上的一个已知的概率函数，那么 $\{P_\theta,\theta\in\Theta\}$ 即被称为**参数族**（**Parametric family**），其中 Θ 为**参数空间**（**Parametric space**），正整数 d 为参数空间的维数。

其中**位置尺度族**（**Location-scale families**）和**指数分布族**（**Exponent families**）是两类最为特殊且重要的参数族，我们这节将分别介绍这两类分布族。

3.1 位置尺度族

对于一个随机变量 X，我们令 $Y=\sigma X+\mu,\sigma>0$，那么其分布函数：

$$\begin{aligned}F_Y(y)&=P(Y\le y)\\&=P(\sigma X+\mu\le y)\\&=P\left(X\le\frac{y-\mu}{\sigma}\right)\\&=F_X\left(\frac{y-\mu}{\sigma}\right)\end{aligned}$$

所以其密度函数满足：

$$f_Y(y)=\frac{1}{\sigma}f_X\left(\frac{x-\mu}{\sigma}\right)$$

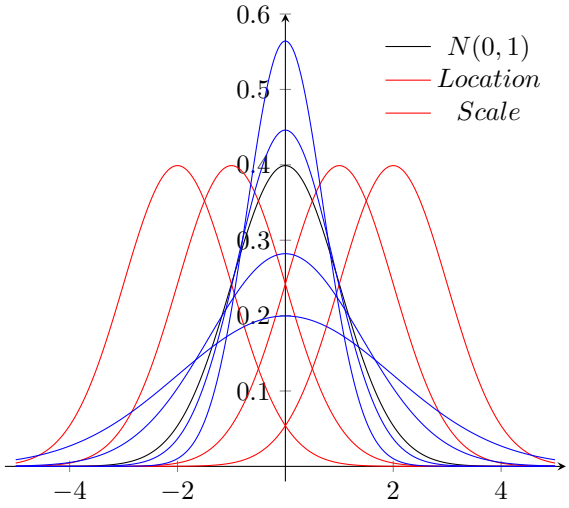


图 10: 正态分布的位置尺度族的密度函数

同时，其期望和方差满足：$\mathbb{E}(Y) = \sigma\mathbb{E}(X) + \mu$，$\mathrm{Var}(Y) = \sigma^2\mathrm{Var}(X)$。

令 $f(x)$ 为任意的密度函数，那么形如 $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ 的密度函数就形成了以 $\theta = (\mu, \sigma)$ 为参数的参数族，我们称之为位置尺度族。位置尺度族即对于任意的随机变量做位移和数乘所得到的随机变量的分布组成的分布族。其中参数 μ 一般称为**位置参数**（**Location parameter**），而 σ 一般称为**尺度参数**（**Scale parameter**）。

例 3. 标准正态分布的密度函数为

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

对于任意的 $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$，其位置尺度族的密度函数可以写为：

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

即正态分布的密度函数。因而正态分布族 $\{P_{\mu,\sigma}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$ 为一个位置尺度族。图 (10) 展示了正态分布的位置族和尺度族。

上面介绍的包括指数分布、Γ 分布等都是各种类型的位置尺度族。正如正态分布一样，很多时候我们会将一个随机变量变换为期望为 0、方差为 1 的随机变量，即令随机变量 $Z = \frac{X-\mu}{\sigma}$，易得 $\mathbb{E}(Z) = 0, \mathrm{Var}(Z) = 1$，我们称之为**标准化**（**standardized**）。由于 $P(X \le x) = P(Z \le \sigma x + \mu)$，因而应用中对于位置尺度族，经常将先研究标准化之后的随机变量，进而推广到其位置尺度族。

3.2 单参数指数分布族

指数分布族是非常常用的分布族，其包含了我们上面介绍的多数分布。在此我们首先讨论单参数的指数分布族。其定义如下：

定义 2. （指数分布族）对于一个参数族 $\{P_\theta, \theta \in \Theta\}$，如果其概率密度（质量）函数可以写成如下形式：

$$f(x|\theta) = h(x) \cdot \exp\{\eta(\theta) \cdot T(x) - B(\theta)\} \tag{2}$$

那么我们称 $\{P_\theta, \theta \in \Theta\}$ 为**单参数指数分布族**（**One-parameter exponential family**）。

由于很多随机变量的取值范围不是 $\mathbb{R}$，然而随机变量的值域为 $\mathbb{R}$，因而我们一般使用指示函数（indicator function）来表示随机变量的取值范围，即定义

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

例如 $1_{(-\infty,0]}(x)$ 即当 $x <= 0$ 时，$1_{(-\infty,0]}(x) = 1$；当 $x > 0$ 时，$1_{(-\infty,0]}(x) = 0$。

我们之前遇到的很多分布都属于指数分布族。例如：

例 4. 令 N 为已知的正整数，如果随机变量 $X \sim Bi(N,p)$，参数 $\theta = p \in \Theta = (0,1)$，其质量函数：

$$\begin{aligned} P(x|p) &= \binom{N}{x} p^x (1-p)^{N-x} \\ &= \binom{N}{x} \exp\{x\ln(p) + (N-x)\ln(1-p)\} \\ &= \binom{N}{x} \exp\{x[\ln(p) - \ln(1-p)] + N\ln(1-p)\} \\ &= \binom{N}{x} \exp\left\{x\left[\ln\left(\frac{p}{1-p}\right)\right] + N\ln(1-p)\right\} \end{aligned}$$

故令 $h(x) = \binom{N}{x}$，$\eta(\theta) = \ln\left(\frac{p}{1-p}\right)$，$T(x) = x$，$B(\theta) = -N\ln(1-p)$，所以二项分布属于指数分布族。

例 5. 泊松分布参数为 $\theta = \lambda \in \Theta = (0,\infty)$，其质量函数为：

$$\begin{aligned} P(x|\lambda) &= \frac{\lambda^x}{x!} e^{-\lambda} \\ &= \frac{1}{x!} \exp\{x\ln(\lambda) - \lambda\} \end{aligned}$$

故令 $h(x)=\frac{1}{x!}$，$\eta(\theta)=\ln(\lambda)$，$T(x)=x$，$B(\theta)=\lambda$，所以泊松分布属于指数分布族。

例 6. 指数分布的密度函数为：

$$\begin{aligned} f(x|\beta) &= \frac{1}{\beta}e^{-\frac{x}{\beta}}\cdot 1_{(0,\infty)}(x) \\ &= 1_{(0,\infty)}(x)\exp\left\{-x\cdot\frac{1}{\beta}-\ln\beta\right\} \end{aligned}$$

因而令 $h(x)=1_{(0,\infty)}(x)$，$\eta(\theta)=-\frac{1}{\beta}$，$T(x)=x$，$B(\theta)=\ln\beta$，故指数分布属于指数分布族。

然而并非所有带指数的密度函数都属于指数分布族，比如：

例 7. 若某一密度函数为：

$$f(x|\beta)=\frac{1}{\beta}\exp\left\{1-\frac{x}{\beta}\right\},x>\beta>0$$

可知 $f(x|\beta)$ 为密度函数。然而：

$$\begin{aligned} f(x|\beta) &= \frac{1}{\beta}\exp\left\{1-\frac{x}{\beta}\right\}\cdot 1_{(\beta,\infty)}(x) \\ &= 1_{(\beta,\infty)}(x)\cdot\exp\left\{-\frac{x}{\beta}-\ln\beta+1\right\} \end{aligned}$$

由于 $1_{(\beta,\infty)}(x)$ 不仅仅依赖于 x，而且依赖于 β，因而这一分布不属于指数分布族。

很多时候为了方便起见，我们会把密度函数重新参数化，即对于指数分布族

$$f(x|\theta)=h(x)\cdot\exp\{\eta(\theta)\cdot T(x)-B(\theta)\}$$

我们令 $\lambda=\eta(\theta)$，那么指数分布族可以写为：

$$f(x|\theta)=h(x)\cdot\exp\{\lambda\cdot T(x)-C(\lambda)\} \tag{3}$$

我们将指数分布族重新参数化为式 (3) 的形式，并将这种形式成为**规范形式**（**Canonical form**）。

例 8. 在例 (4) 中，如果令 $\lambda=\ln\left(\frac{p}{1-p}\right)$，那么

$$P(x|\lambda)=\begin{pmatrix} N \\ x \end{pmatrix}\exp\left\{\lambda x-N\ln\left(1+e^{\lambda}\right)\right\}$$

即为二项分布的质量函数的规范形式。

对于指数分布族的规范形式，我们有如下定理：

定理 1. 如果随机变量 X 的概率密度（质量）函数为式 *(3)* 的形式，那么 $\mathbb{E}\left[T\left(X\right)\right]=C'\left(\lambda\right)$，$Var\left[T\left(X\right)\right]=C''\left(\lambda\right)$。

例 9. 在例 (8) 中，$T\left(x\right)=x$，因而 $\mathbb{E}\left[T\left(X\right)\right]=\mathbb{E}\left(X\right)=C'\left(\lambda\right)=\frac{Ne^{\lambda}}{1+e^{\lambda}}=Np$，$\mathrm{Var}\left[T\left(X\right)\right]=\mathrm{Var}\left(X\right)=C''\left(\lambda\right)=\frac{Ne^{\lambda}}{(1+e^{\lambda})^{2}}=Np\left(1-p\right)$。

习题

练习 1. 求负二项分布的方差。

练习 2. 证明几何分布的无记忆性。

练习 3. 计算泊松分布的方差。

练习 4. 计算 $\Gamma\left(\alpha,\beta\right)$ 分布的方差。

练习 5. 求标准正态分布的偏度、峰度。

练习 6. Pareto 分布的密度函数为：

$$f\left(x|a,\beta\right)=\beta\cdot a^{\beta}x^{-(\beta+1)},x>a$$

那么 Pareto 分布是否属于指数分布族?

练习 7. 请用定理 (1) 计算指数分布、泊松分布的期望和方差。

参考文献

[1] Bickel, P.J., Doksum, K.A., 2001. Mathematical Statistics: Basic Ideas and Selected Topics. Prentice-Hall, Inc, New Jersey.

[2] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[3] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

[4] Schervish, M.J., 1995. Theory of Statistics. Springer-Verlag, New York.

第四节 · 多元随机变量

司继春

上海对外经贸大学统计与信息学院

在前两节中，我们讨论了一元随机变量的定义及其期望等概念。此外，我们还可以把随机变量的概念扩展到随机向量。在引入随机向量的定义之前，我们先回忆一些基础知识。

1 数学准备

对于两个集合 A, B，我们记 $A \times B = \{(a, b), \forall a \in A, b \in B\}$，即 $\times$ 运算定义了一个二元组的集合，我们称 $\times$ 为**笛卡尔乘积**（**Cartesian product**）。比如，如果我们选取 $A = \{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}, B = \{2, ..., 10, \mathrm{J}, \mathrm{Q}, \mathrm{K}, \mathrm{A}\}$，那么我们就得到了一副扑克牌共 52 张牌的集合。而如果选取 $A = \mathbb{R}, B = \mathbb{R}$，那么 $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ 为二维平面。

更一般的，我们可以记

$$\begin{aligned}\Omega_1 \times \Omega_2 \times \cdots \times \Omega_d &= \times_{i=1}^{n} \Omega_i \\ &= \{(\omega_1, \omega_2, ..., \omega_n), \omega_i \in \Omega_i, i = 1, ..., n\}\end{aligned}$$

特别的，令 $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ 为 n 维的**欧几里得空间**（**Euclidean space**），其中 $\omega = (\omega_1, ...\omega_n) \in \mathbb{R}^n$ 为向量。如果 $x \in \mathbb{R}^n, y \in \mathbb{R}^n$ 我们可以定义**内积**（**inner product**）为：

$$\langle x, y \rangle = x \cdot y = \sum_{i=1}^{n} x_i y_i$$

如果 $\langle x, y \rangle = 0$，我们称两个向量**正交**（**orthogonal**）。有了内积之后，可以使用内积定义（欧几里得）**范数**（**norm**）：

$$\|x\| = \sqrt{\langle x, x \rangle}$$

以及两个向量间的**距离**（**metric**）：

$$d(x, y) = \|x - y\| = \sqrt{\langle x - y, x - y \rangle} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

在本讲义中，我们一般把向量写成**列向量**的形式。

对于一个 n 维的欧几里得空间 $\mathbb{R}^n$，我们可以在这个空间上定义 Borel 域：

$$\mathscr{B}^n = \times_{i=1}^n \mathscr{B}_i = \sigma\left(\{\times_{i=1}^n A_i, A_i \in \mathscr{B}_i\}\right)$$

如果我们把 k 个 n 维向量按列摆放在一起，我们得到了一个 $k \times n$ 维的矩阵 $A_{k\times n} = [a_1, ..., a_n]$，其中 a_i 为 k 维向量。如果我们将矩阵 A 左乘一个 n 维向量 x，那么 $y = Ax$ 为一个 k 维向量。现在我们可以把矩阵左乘向量视为一个函数，即 $y = A(x) = Ax$，易知 $A(x_1 + x_2) = Ax_1 + Ax_2$，以及 $A(\alpha x) = \alpha Ax$，我们一般把符合如上两个性质的函数成为**线性映射**（**linear mapping**）。特别的，当 $k = n$，即 A 为 $n \times n$ 维方阵时，线性映射 A 将 $\mathbb{R}^n$ 上的一个向量 x 映射到 $\mathbb{R}^n$ 上的另外一个向量 y，此时我们称 A 为**线性变换**（**linear transformation**）。比如变换：

$$A = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$$

就讲一个二维空间 $\mathbb{R}^2$ 上的向量逆时针旋转 θ 度。取 $\theta = \frac{\pi}{2}$，那么：

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

取 $x = [1, 0]'$，那么 $y = Ax = [0, 1]'$，逆时针旋转了 90 度。

使用分块矩阵，如果令 $x = [x_1, x_2, ..., x_n]' \in \mathbb{R}^n$，$A_{k\times n} = [a_1, ..., a_n]$，其中 a_i 为 k 维列向量，那么：

$$y = Ax = [a_1, ..., a_n] \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^n x_i a_i$$

也就是说线性映射的结果 y 实际上是矩阵 A 的列向量 a_i 的一个线性组合。因而矩阵 A 的秩 $\text{rank}(A)$，即矩阵 A 的列向量的极大线性无关组，也就是对于所有 $x \in \mathbb{R}^n$，所有 $y = Ax$ 的极大线性无关组，或者所有向量 $\{y = Ax, \forall x \in \mathbb{R}^n\}$ 这个线性空间的维数。

实对称矩阵是我们接下来将要大量遇到的一类矩阵，任何的实对称矩阵 $A_{n\times n}$ 都可以被对角化为一个正交矩阵及其转置和一个对角矩阵的乘积：

$$A = \Gamma' \Lambda \Gamma$$

其中 Γ 为正交矩阵，即 $\Gamma\Gamma' = \Gamma'\Gamma = I$，$\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ 为**特征值**（**eigenvalue**）的对角阵。正交矩阵 $\Gamma'\Gamma = I$，因而矩阵 Γ 的列向量（**特征向量**，**eigen-**

vector）是两两正交的，且每个列向量的范数为 1。比如矩阵：

$$\Gamma = \begin{bmatrix} cos(\theta) & -sin(\theta) \\ sin(\theta) & cos(\theta) \end{bmatrix}$$

为正交矩阵，其每个列向量都是正交的且范数为 1。这类矩阵对应着等距变换（isometry），即两个点经过正交矩阵 Γ 的变换之后，$d(\Gamma x, \Gamma y) = \sqrt{x'\Gamma'\Gamma y} = \sqrt{x'y} = d(x, y)$。正交矩阵对应着旋转、翻转等变换，而相应的，对角矩阵 Λ 则对应着在不同的方向上的拉伸变换。

如果对于任何一个向量 $x \in \mathbb{R}^n$，$x'Ax > 0$，我们称矩阵 A 为**正定矩阵**（**Positive-definite matrix**）；如果满足 $x'Ax \geq 0$，则成为**半正定矩阵**（**Positive semi-definite matrix**）；负定矩阵和半负定矩阵可以类似定义。显然，如果一个实对称矩阵的所有特征值都 > 0（≥ 0），那么这个矩阵即为正定矩阵（半正定矩阵）。

此外，如果一个矩阵 A 可以被对角化，其特征值为 $\lambda_1, \ldots, \lambda_n$，那么 A 的行列式值 $|A| = \prod_{i=1}^n \lambda_i$。从这个角度看，对焦矩阵 Λ 对应着放缩变换，代表着在不同方向上的放缩，而等距变换则不会引起放缩。此外定义矩阵的**迹**（**trace**）为其对角元之和，即若 $A = [a_{ij}]_{n\times n}$，那么 $\mathrm{tr}(A) = \sum_{i=1}^n a_{ii}$。矩阵的迹有如下简单的性质：$\mathrm{tr}(AB) = \mathrm{tr}(BA)$。使用如上性质容易验证，如果矩阵 A 可以被对角化，那么 $\mathrm{tr}(A) = \sum_{i=1}^n \lambda_i$。

在实对称矩阵中，有一类矩阵是我们接下来非常频繁使用的，即**幂等矩阵**（**Idempotent matrix**）。如果一个方阵 P 满足 $P^2 = P$，那么我们称矩阵 P 为幂等矩阵。例如，矩阵：

$$P = \frac{1}{4} \cdot \begin{bmatrix} -4 & 6 & 2 \\ -12 & 13 & 3 \\ 20 & -15 & -1 \end{bmatrix}$$

为幂等矩阵，可以验证 $P^2 = P$。

特别的，当 P 为实对称矩阵时，我们称其为**投影矩阵**（**Projection matrix**）。由于所有实对称矩阵都可以被对角化，所以对于任意的投影矩阵，都可以写为：

$$P = \Gamma'\Lambda\Gamma$$

而由于 $P^2 = \Gamma'\Lambda\underbrace{\Gamma\Gamma'}_{I}\Lambda\Gamma = \Gamma'\Lambda^2\Gamma = \Gamma'\Lambda\Gamma$，且 Γ 为可逆矩阵，所以 $\Lambda^2 = \Lambda$。由于 Λ 为对角阵，所以 Λ 的对角元必为 0 或者 1。因而 $\mathrm{rank}(P) = \mathrm{rank}(\Lambda) = \mathrm{tr}(\Lambda)$。

投影矩阵顾名思义，与**投影**（**Projection**）的概念密不可分。如果把投影矩阵 P 视为线性变换，幂等矩阵的定义意味着一个向量经过 P 的变换以后，再

次经过 P 的变换仍然保持不变，即 $P(Px) = P^2x = Px$。比如矩阵：

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

即把一个 $x-y-z$ 三维坐标系中的一个向量 $x = (x_1, x_2, x_3)'$ 映射到 $x-y$ 二维平面上的点 Px，而一个本身就在 $x-y$ 二维平面的点，如 Px，再次经过 P 的映射，还是在 $x-y$ 二维平面上，且就是其本身。可以验证，$P^2 = P$。类似的，矩阵：

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

则把一个三维向量 $x = (x_1, x_2, x_3)'$ 映射到 $y = x$ 这条直线上，同样有 $P^2 = P$。

如果定义 $M = I - P$，那么 $M^2 = (I-P)(I-P) = I - P - P + P^2 = I - P = M$，即 $M = I - P$ 也为投影矩阵。注意 $(Mx)' Px = x'(I-P)Px = x'\left(P - P^2\right)x = 0$，因而 Px 与 Mx 是正交的。也就是说，幂等矩阵把一个向量 x 分解成了正交的两个部分：Px 和 Mx，$x = Px + Mx$ 且 $\langle Mx, Px\rangle = 0$。

例 1. 令 $\iota \in \mathbb{R}^n, \iota = (1, 1, ..., 1)'$，那么矩阵 $P_0 = \frac{1}{n}\iota\iota'$ 为投影矩阵，即 $P_0' = P_0$，且 $P_0^2 = \frac{1}{n^2}\iota \underbrace{\iota'\iota}_{n} \iota' = \frac{1}{n}\iota\iota' = P_0$。对于一个向量 x，

$$P_0 x = \frac{1}{n}\iota\iota' x = \frac{1}{n}\iota \cdot \sum_{i=1}^{n} x_i = \iota \cdot \bar{x} = \begin{pmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{pmatrix}$$

即 P_0 将一个向量投影变换为其均值向量。易知 $\operatorname{rank}(P_0) = \operatorname{tr}(P_0) = \operatorname{tr}\left(\frac{1}{n}\iota\iota'\right) = \frac{1}{n}\operatorname{tr}(\iota'\iota) = 1$。如果令 $M_0 = I - P_0$，根据上述结论，易知 M_0 也是幂等矩阵，且 $\operatorname{rank}(M_0) = \operatorname{tr}(M_0) = \operatorname{tr}(I - P_0) = \operatorname{tr}(I) - \operatorname{tr}(P_0) = n - 1$，且：

$$M_0 x = x - \frac{1}{n}\iota\iota' x = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix}$$

那么：

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (M_0 x)' M_0 x = x' M_0 x$$

为一个二次型的形式。

对于一个向量 $\theta = [\theta_1, \theta_2, \ldots, \theta_n]'$，其实值函数：$f(\theta) : \mathbb{R}^n \to \mathbb{R}$，我们可

以定义函数 $f(\cdot)$ 对向量 θ 的导数为:

$$\frac{\partial f}{\partial \theta}=\begin{bmatrix}\frac{\partial f}{\partial \theta_1}\\ \frac{\partial f}{\partial \theta_2}\\ \vdots\\ \frac{\partial f}{\partial \theta_n}\end{bmatrix}$$

同时定义其二阶导:

$$\frac{\partial^2 f}{\partial \theta \partial \theta'}=\begin{bmatrix}\frac{\partial^2 f}{\partial \theta_1^2} & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_1 \partial \theta_n}\\ \frac{\partial^2 f}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_2^2} & \cdots & \frac{\partial^2 f}{\partial \theta_2 \partial \theta_d}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial^2 f}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 f}{\partial \theta_n \partial \theta_2} & \cdots & \frac{\partial^2 f}{\partial \theta_n^2}\end{bmatrix}$$

比如，如果 $f(\theta)=\frac{\mu^2}{2\sigma^2}+\ln(\sigma)$，$\theta=(\mu,\sigma)'$ 那么:

$$\frac{\partial f(\theta)}{\partial \theta}=\begin{bmatrix}\frac{\mu}{\sigma^2}\\ \frac{1}{\sigma}-\frac{\mu^2}{\sigma^3}\end{bmatrix}$$

$$\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta'}=\begin{bmatrix}\frac{1}{\sigma^2} & -\frac{2\mu}{\sigma^3}\\ -\frac{2\mu}{\sigma^3} & \frac{3\mu^2}{\sigma^4}-\frac{1}{\sigma^2}\end{bmatrix}$$

我们知道对于一个实值函数，$\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}=\frac{\partial^2 f}{\partial \theta_j \partial \theta_i}$，因而 $\frac{\partial^2 f}{\partial \theta \partial \theta'}$ 是一个实对称矩阵。回忆极值原理，如果函数 f 可微，那么函数 f 在 θ_0 处为极值点的必要条件是 $\frac{\partial f(\theta_0)}{\partial \theta}=0$，如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 为正定矩阵（$f(\theta)$ 的所有特征值为正），那么 f 在 θ_0 处为极小值点，否则如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 为负定矩阵（$f(\theta)$ 的所有特征值为负），那么 f 在 θ_0 处为极大值点，如果 $\frac{\partial^2 f(\theta_0)}{\partial \theta \partial \theta'}$ 的特征值既有正值又有负值，那么 f 在 θ_0 处为鞍点（saddle point）。我们称 $\frac{\partial^2 f(\theta)}{\partial \theta \partial \theta'}$ 为**海塞矩阵**（**Hessian matrix**）。

2 多元随机变量

在有了以上准备之后，我们可以定义随机向量的概念。

定义 1.（随机向量）给定一个概率空间 $(\Omega,\mathscr{F},\mathscr{P})$，一个 k 维的随机向量 X 即从样本空间到 k 维欧几里得空间的函数，$X:\Omega\to\mathbb{R}^n$。

即，如果一个向量其每个分量都是随机变量，那么此向量被称为随机向量。

例 2.（随机向量）投两个均匀的四面骰子，则

$$\Omega=\{(1,1),(1,2),\ldots(4,4)\}$$

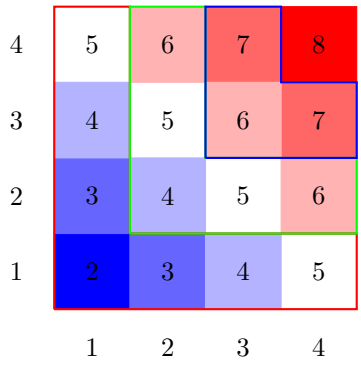


图 1: 四面骰子

定义随机变量 Y 为两个骰子的数值之和，定义 Z 为两个骰子中较小的骰子的数值，如图 (1) 所示。那么向量 $(Y,Z)' = X : \Omega \to \mathbb{R}^2$ 为一个随机向量，其可能的取值为 $\{(y,z), y \in \{2,...,8\}, z \in \{1,2,3,4\}\}$。例如，$X^{-1}(\{(5,3)\}) = \{(2,3),(3,2)\}$。

进而，我们可以使用 $(\Omega, \mathscr{F}, \mathscr{P})$ 和一个随机向量 X 的定义导出一个 $(\mathbb{R}^n, \mathscr{B}^n)$ 上的概率函数的定义。即定义：

$$P_X(A) = \mathscr{P}\left(X^{-1}(A)\right), \forall A \in \mathscr{B}^n$$

例 3. 在例 (2) 中，如果 $A = \{(5,2)\}$，那么：

$$P_X(A) = \mathscr{P}\left(X^{-1}(A)\right) = \mathscr{P}(\{(2,3),(3,2)\}) = \frac{2}{16}$$

同理，$P_X(\{(2,1)\}) = \frac{1}{16}$，$P_X(\{(5,a), a \in \{1,2,3,4\}\}) = \frac{4}{16}$ 等等。

给定一个随机向量 X，在得到了由原始概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 导出的概率空间 $(\mathbb{R}^n, \mathscr{B}^n, P)$ 后，仿照一元随机变量，我们还可以定义随机向量的**联合分布函数**（**joint cumulative distribution funtion**）：

定义 2. （联合分布函数）由 $(\Omega, \mathscr{F}, \mathscr{P})$ 导出的概率空间 $(\mathbb{R}^n, \mathscr{B}^n, P)$ 的**联合分布函数**（**joint c.d.f.**）定义为：

$$\begin{aligned} F(x) &= F(x_1, x_2, ..., x_n) \\ &= P((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n]) \\ &= \mathscr{P}\left(X^{-1}((-\infty, x_1] \times (-\infty, x_2] \times \cdots (-\infty, x_n])\right) \end{aligned}$$

$\forall x \in \mathbb{R}^n$。

易得，联合分布函数为单调递增且 $F(-\infty, -\infty, ..., -\infty) = 0, F(\infty, \infty, ..., \infty) = 1$。相应的，对于连续（离散）型的随机向量 X，我们还可以定义其联合概率密

度（质量）函数。

定义 3.（随机向量的联合密度函数与联合质量函数）

1. 如果随机向量 X 的每个分量都是离散型随机变量，那么可以定义联合概率质量函数 p.m.f 为：$f(x)=P(\{x\})=P(\{X_1=x_1,...,X_n=x_n\})$。

2. 如果随机变量 X 的联合分布函数连续，如果函数 $f(x)$ 满足：

$$P(X\in A)=\int_A f(x)\,dx, x\in\mathbb{R}^n, A\in\mathscr{B}^n$$

那么我们称 $f(x)$ 为其联合概率密度函数 p.d.f。特别的，如果联合分布函数 $F(x)$ 可微那么：

$$f(x)=\frac{\partial^n F(x)}{\partial x_1\partial x_2\cdots\partial x_n}$$

例 4.（概率质量函数）例 (2) 中的概率质量函数可以用下表描述：

$Z\backslash Y$	2	3	4	5	6	7	8
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0
4	0	0	0	0	0	0	$\frac{1}{16}$

例 5.（概率密度函数）如果随机向量 $X=(X_1,X_2)$ 的两个分量分别服从正态分布，且相互独立，那么其概率密度函数为：

$$f(x)=\frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}$$

现在，如果 $X=(X_1,...,X_n)$ 为随机向量，那么 $\widetilde{X}=(X_{i_1},X_{i_2},...,X_{i_k}),1\leq i_1<i_2<...<i_k\leq n$ 也是一个随机向量。$\widetilde{X}$ 的联合分布函数可以通过 $F(x)$ 来定义，即令 $F(x)$ 中满足 $j\notin\{i_1,...i_k\}$ 的分量为 ∞。如对于三维随机变量 $X=(X_1,X_2,X_3)$，则 $\widetilde{X}=(X_1,X_2)$ 的分布函数为：$F_{\widetilde{X}}(\widetilde{x})=F(\widetilde{x}_1,\widetilde{x}_2,\infty)$。

特别的，对于随机向量 X 的每个分量 X_i，我们可以定义其**边缘分布函数**（**marginal c.d.f.**）为：

$$F_{X_i}(x_i)=F(\infty,...,x_i,...,\infty)$$

注意边缘分布函数对应着一元随机变量 X_i 的分布函数：

$$\begin{aligned}F(\infty,...,x_i,...,\infty)&=P(\mathbb{R}\times\mathbb{R}\times\cdots\times(-\infty,x_i]\times\cdots\times\mathbb{R})\\&=\mathscr{P}\left(X^{-1}(\mathbb{R}\times\mathbb{R}\times\cdots\times(-\infty,x_i]\times\cdots\times\mathbb{R})\right)\\&=\mathscr{P}\left(X_i^{-1}((-\infty,x_i])\right)\end{aligned}$$

对于连续（离散）型的随机变量 X_i，其边缘概率密度（质量）函数可以相应定义。

例 6. （边缘质量函数）例 (2) 中，$X=(Y,Z)$，Y 和 Z 的边缘概率质量函数如下表所示：

$Z\backslash Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$ $\vert\vert$
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y=$	1

例 7. （边缘密度函数）例 (5) 中的联合正态分布函数，其边缘分布函数为：

$$\begin{aligned}
F_{X_1}(t) &= \int_{\mathbb{R}}\int_{-\infty}^{t} f(x_1,x_2)\,dx_1dx_2 \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\int_{\mathbb{R}}\int_{-\infty}^{t}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}dx_1dx_2 \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\int_{\mathbb{R}}\exp\left\{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right\}dx_2\int_{-\infty}^{t}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\}dx_1 \\
&= \frac{1}{\sqrt{2\pi}\sigma_1}\int_{-\infty}^{t}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\right\}dx_1
\end{aligned}$$

则其边缘密度函数为：

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

即例 (5) 中联合正态分布的边缘分布仍然是正态分布。

注意边缘分布函数由联合分布函数导出，然而如果只确定了边缘分布，联合分布并不能唯一确定。

例 8. （联合分布与边缘分布）以下两个联合质量函数具有相同的边缘分布，然而其联合质量函数并不相同：

$Z\backslash Y$	0	1	f_Z
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

$Z\backslash Y$	0	1	f_Z
0	$\frac{1}{12}$	$\frac{5}{12}$	$\frac{1}{2}$
1	$\frac{5}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
f_Y	$\frac{1}{2}$	$\frac{1}{2}$	1

例 9. 如果随机向量 (U,V) 的分布函数为:

$$F_{U,V}(u,v)=\min\{u,v\}$$

其边缘分布:

$$F_U(u)=F_{U,V}(u,\infty)=u$$
$$F_V(v)=F_{U,V}(\infty,v)=v$$

即其边缘分布为均匀分布。如果另一分布函数为:

$$F_{U,V}(U,V)=u\cdot v$$

其边缘分布也为均匀分布。因而如果只知道边缘分布，不能确定其联合分布。

3 多元随机变量的期望

与一元随机变量类似，对于随机向量 X 以及相应的从概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 导出的概率空间 $(\mathbb{R}^n,\mathscr{B}^n,P)$，对于实值可测函数 $g(X(\omega)):\Omega\to\mathbb{R}$，可以使用导出的概率空间计算数学期望:

$$\mathbb{E}(g(X))=\int_\Omega g(X(\omega))\,\mathscr{P}(d\omega)=\int_{\mathbb{R}^n}g(x)\,P(dx)$$

根据此定义，如果令 $g(X)=\iota_i'X=X_i$，其中 $\iota_i=(0,0,...,1,...0)$，那么:

$$\mathbb{E}(g(X))=\int_\Omega X_i(\omega)\,\mathscr{P}(d\omega)=\mathbb{E}(X_i)$$

即多元随机变量的分量的期望与一元随机变量的期望定义相同。因而我们经常把随机向量的期望写为:

$$\mathbb{E}(X)=\begin{bmatrix}\mathbb{E}X_1\\ \mathbb{E}X_2\\ \vdots\\ \mathbb{E}X_n\end{bmatrix}$$

如果我们令 $g(X)=\sum_{i=1}^n X_i=\iota'X$，其中 $\iota=(1,1,...,1)'$ 为全部由 1 构

成的向量，那么：

$$\begin{aligned}\mathbb{E}\left(\sum_{i=1}^{n}X_i\right)&=\int_{\mathbb{R}^n}\sum_{i=1}^{n}X_iP\left(dx\right)\\&=\sum_{i=1}^{n}\int_{\mathbb{R}^n}X_iP\left(dx\right)\\&=\sum_{i=1}^{n}\mathbb{E}\left(X_i\right)\end{aligned}$$

即期望的**线性性**。如果令 $\mu=\mathbb{E}\left(X\right)=\left[\mathbb{E}\left(X_1\right),\mathbb{E}\left(X_2\right),...,\mathbb{E}\left(X_d\right)\right]'$，令 $a\in\mathbb{R}^n$，那么我们有 $\mathbb{E}\left(\sum_{i=1}^{n}a_iX_i\right)=\mathbb{E}\left(a'X\right)=a'\mathbb{E}\left(X\right)=a'\mu$。

而对于一个实数矩阵 $A_{h\times n}=\left[a_1,a_2,...,a_h\right]'$，其乘积 $AX=\left[a_1'X,a_2'X,...,a_h'X\right]'$，其期望为：

$$\mathbb{E}\left(AX\right)=\mathbb{E}\begin{bmatrix}a_1'X\\a_2'X\\\vdots\\a_h'X\end{bmatrix}=\begin{bmatrix}\mathbb{E}\left(a_1'X\right)\\\mathbb{E}\left(a_2'X\right)\\\vdots\\\mathbb{E}\left(a_h'X\right)\end{bmatrix}=\begin{bmatrix}a_1'\mathbb{E}\left(X\right)\\a_2'\mathbb{E}\left(X\right)\\\vdots\\a_h'\mathbb{E}\left(X\right)\end{bmatrix}=A\mathbb{E}\left(X\right)$$

因而对于 $A_{h\times n}$ 以及 h 维向量 b，有：$\mathbb{E}\left(AX+b\right)=A\mathbb{E}\left(X\right)+b$。

此外，如果对于两个一元随机变量 Y,Z，如果 $\mathbb{E}\left|Y\right|^2<\infty,\mathbb{E}\left|Z\right|^2<\infty$，根据 Cauchy-Schwarz 不等式，$\mathbb{E}\left|YZ\right|\leq\sqrt{\mathbb{E}\left|Y\right|^2\mathbb{E}\left|Z\right|^2}<\infty$，即 YZ 可积，我们可以定义两个随机变量的**协方差**（**Covariance**）：

$$\begin{aligned}\mathrm{Cov}\left(Y,Z\right)&=\mathbb{E}\left[\left(Y-\mathbb{E}\left(Y\right)\right)\left(Z-\mathbb{E}\left(Z\right)\right)\right]\\&=\mathbb{E}\left[YZ-\mathbb{E}\left(Y\right)Z-Z\mathbb{E}\left(Y\right)+\mathbb{E}\left(Y\right)\mathbb{E}\left(Z\right)\right]\\&=\mathbb{E}\left(YZ\right)-2\mathbb{E}\left(Y\right)\mathbb{E}\left(Z\right)+\mathbb{E}\left(Y\right)\mathbb{E}\left(Z\right)\\&=\mathbb{E}\left(YZ\right)-\mathbb{E}\left(Y\right)\mathbb{E}\left(Z\right)\end{aligned}$$

当 $Y=Z$ 时，$\mathrm{Cov}\left(Y,Y\right)=\mathbb{E}\left(Y^2\right)-\left[\mathbb{E}\left(Y\right)\right]^2=\mathrm{Var}\left(Y\right)$。

进而可以使用协方差定义**相关系数**（**correlation coefficient**）：

$$\rho_{Y,Z}=\frac{\mathrm{Cov}\left(Y,Z\right)}{\sqrt{\mathrm{Var}\left(Y\right)\mathrm{Var}\left(Z\right)}}$$

由于:

$$\begin{aligned}\operatorname{Cov}(Y,Z) &= \mathbb{E}\left[(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))\right] \\ &\leq \mathbb{E}\left|(Y-\mathbb{E}(Y))(Z-\mathbb{E}(Z))\right| \\ &\leq \sqrt{\mathbb{E}\left|(Y-\mathbb{E}(Y))\right|^2 \mathbb{E}\left|Z-\mathbb{E}(Z)\right|^2} \\ &= \sqrt{\operatorname{Var}(Y)\operatorname{Var}(Z)}\end{aligned}$$

可知 $-1 \leq \rho_{Y,Z} \leq 1$。如果 $\rho_{Y,Z} = \pm 1$，那么 $P(Y = c_1 Z + c_2) = 1, c_1 \neq 0$；如果 $\rho_{Y,Z} > 0$，我们称随机变量 Y 和 Z 正相关，反之成为负相关，如果 $\rho_{Y,Z} = 0$，我们称随机变量 Y 和 Z 不相关（uncorrelated）。这里所谓的「相关系数」特指**皮尔森相关系数**（**Pearson correlation coefficient**），实际上只度量了随机变量之间的线性相关性。相关系数等于 0 并不意味着两个随机变量没有非线性的相关性。

例 10. 如果随机变量 $Y = Z^2$，$Z \sim N(0,1)$，那么:

$$\begin{aligned}\operatorname{Cov}(Z,Y) &= \mathbb{E}ZY - \mathbb{E}Z\mathbb{E}Y \\ &= \mathbb{E}Z^3 \\ &= 0\end{aligned}$$

两者相关系数为 0，然而显然两者存在着非线性的函数关系。

此外，如果 a, b 为任意实数，那么:

$$\begin{aligned}\operatorname{Var}(aY+bZ) &= \mathbb{E}(aY+bZ)^2 - \left[a\mathbb{E}(Y)+b\mathbb{E}(Z)\right]^2 \\ &= \mathbb{E}\left(a^2Y^2 + b^2Z^2 + 2abYZ\right) \\ &\quad - \left[a^2\left(\mathbb{E}(Y)\right)^2 + b^2\left(\mathbb{E}(Z)\right)^2 + 2ab\mathbb{E}(Y)\mathbb{E}(Z)\right] \\ &= a^2\operatorname{Var}(Y) + b^2\operatorname{Var}(Z) + 2ab\operatorname{Cov}(Y,Z)\end{aligned}$$

如果 Y, Z 不相关，那么 $\operatorname{Var}(aY+bZ) = a^2\operatorname{Var}(Y) + b^2\operatorname{Var}(Z)$。

对于一个随机向量 $X = (X_1, X_2, ..., X_n)'$，我们可以定义**方差协方差矩阵**（**variance-covariance matrix**），或者**协方差矩阵**为:

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}\left[(X-\mathbb{E}X)(X-\mathbb{E}X)'\right] \\ &= \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1,X_2) & \cdots & \operatorname{Cov}(X_1,X_n) \\ \operatorname{Cov}(X_2,X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n,X_1) & \operatorname{Cov}(X_n,X_2) & \cdots & \operatorname{Var}(X_n) \end{bmatrix}\end{aligned}$$

由于 $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$，因而协方差矩阵为实对称矩阵。根据定义，

有:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mathbb{E}X)(X-\mathbb{E}X)'\right] \\ &= \mathbb{E}\left[XX' - X\mathbb{E}(X') - \mathbb{E}(X)X' + \mathbb{E}(X)\mathbb{E}(X')\right] \\ &= \mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')\end{aligned}$$

此外，根据协方差矩阵的定义，对于任意的 n 维向量 c，我们有:

$$\begin{aligned}c'\mathrm{Var}(X)c &= c'\left[\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)'\right]c \\ &= \mathbb{E}\left[c'(X-\mathbb{E}X)(X-\mathbb{E}X)'c\right] \\ &= \mathbb{E}\left\{[c'(X-\mathbb{E}X)][c'(X-\mathbb{E}X)]'\right\} \\ &= \mathbb{E}\left[([c'(X-\mathbb{E}X)])^2\right] \\ &\geq 0\end{aligned}$$

因而协方差矩阵是一个半正定矩阵，通常我们记为 $\mathrm{Var}(X) \geq 0$。

由于 $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$，因而协方差矩阵为实对称矩阵。根据定义，对于实数矩阵 $A_{h\times n}$ 以及 h 维向量 b，我们有:

$$\begin{aligned}\mathrm{Var}(AX+b) &= \mathbb{E}\left[(AX+b-\mathbb{E}(AX+b))(AX+b-\mathbb{E}(AX+b))'\right] \\ &= \mathbb{E}\left[(AX-\mathbb{E}(AX))(AX-\mathbb{E}(AX))'\right] \\ &= \mathbb{E}\left[(AX-A\mathbb{E}(X))(X'A'-\mathbb{E}(X')A')\right] \\ &= \mathbb{E}\left[AXX'A' - AX\mathbb{E}(X')A' - A\mathbb{E}(X)X'A' + A\mathbb{E}(X)\mathbb{E}(X')A'\right] \\ &= A\left[\mathbb{E}(XX') - \mathbb{E}(X)\mathbb{E}(X')\right]A' \\ &= A\mathrm{Var}(X)A'\end{aligned}$$

4 多元随机变量的独立性

在概率一节中，我们学习了事件的独立性，现在我们讨论随机变量的独立性。

定义 4. 如果 $\{X_i, 1 \leq i \leq n\}$ 是定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量，如果对于任意的 Borel 集 $\{B_i, 1 \leq i \leq n\}$，有:

$$\mathscr{P}\left(\bigcap_{i=1}^{n}(X_i(\omega)\in B_i)\right) = \prod_{i=1}^{n}\mathscr{P}(X_i(\omega)\in B_i) \tag{1}$$

那么我们称随机变量 $\{X_i, 1 \leq i \leq n\}$ 相互独立。

根据以上定义，随机变量的相互独立意味着对于任意的 Borel 集 B_i，事件集 $\left\{X_i^{-1}(B_i), 1 \leq i \leq n\right\}$ 内的事件都是相互独立的。如果我们选取 $B_i = (-\infty, x_i]$，

那么：

$$\mathscr{P}\left(\bigcap_{i=1}^{n}\{X_i(\omega)\le x_i\}\right)=\prod_{i=1}^{n}\mathscr{P}(\{X_i(\omega)\le x_i\}) \tag{2}$$

实际上，(1) 式与 (2) 式是等价的。如果一系列随机变量 $(X_1,...,X_n)$ 是相互独立的，那么其联合分布函数：

$$\begin{aligned}F(x_1,...x_n)&=P(X_1\le x_1,...,X_n\le x_n)\\&=\mathscr{P}\left(\bigcap_{i=1}^{n}\{X_i(\omega)\le x_i\}\right)\\&=\prod_{i=1}^{n}\mathscr{P}(\{X_i(\omega)\le x_i\})\\&=\prod_{i=1}^{n}P(X_i\le x_i)\\&=\prod_{i=1}^{n}F_{X_i}(x_i)\end{aligned} \tag{3}$$

即独立随机向量的联合分布函数等于其边际分布函数的乘积。(2) 式与 (3) 式也是等价的，因而当我们说一系列随机变量 $\{X_i,1\le i\le n\}$ 相互独立时，等价于其联合分布函数可以写成边际分布相乘的形式。

如果密度（质量）函数存在，那么根据 (3) 式可得：

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

例 11. 在例 (6) 中，概率质量函数为：

$Z\backslash Y$	2	3	4	5	6	7	8	F_Z	f_Z
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{7}{16}$	$\frac{7}{16}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	$\frac{12}{16}$	$\frac{5}{16}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	$\frac{15}{16}$	$\frac{3}{16}$
4	0	0	0	0	0	0	$\frac{1}{16}$	$\frac{16}{16}$	$\frac{1}{16}$
F_Y	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{6}{16}$	$\frac{10}{16}$	$\frac{13}{16}$	$\frac{15}{16}$	$\frac{16}{16}$		$\sum f_Z$ ||
f_Y	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\sum f_Y=$	1

可见 $f_{Z,Y}\ne f_Z\cdot f_Y$，所以随机变量 (Y,Z) 不独立。

例 12. 例 (9) 中的两个联合分布函数：

$$\begin{aligned}F^1_{U,V}(u,v)&=\min\{u,v\}\\F^2_{U,V}(u,v)&=u\cdot v\end{aligned}$$

其边缘分布都为均匀分布，即 $F_U(u)=u, F_V(v)=v$，然而由于：

$$\begin{aligned} F^1_{U,V}(u,v) &= \min\{u,v\} && \neq F_U(u)\cdot F_C(v) \\ F^2_{U,V}(u,v) &= u\cdot v && = F_U(u)\cdot F_C(v) \end{aligned}$$

因而联合分布服从 $F^1_{U,V}$ 的随机变量不是相互独立的，而服从 $F^2_{U,V}$ 的随机变量是相互独立的。

定理 1. $\{X_j, 1\le j\le n\}$ 为一系列相互独立的随机变量，$1\le n_1\le n_2\le\cdots\le n_k=n$，那么对于 *Borel* 可测函数 $f_1, f_2, ...f_k$，那么：

$$\left\{f_1(X_1,...X_{n_1}), f_1(X_{n_1+1},...X_{n_2}), \cdots, f_k\left(X_{n_{k-1}+1},...X_{n_k}\right)\right\}$$

也为相互独立的随机变量

上述定理表明，任意独立的随机变量的函数仍然是相互独立的。此外，对于独立的随机变量的乘积，我们有如下结论：

定理 2. 如果概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的随机向量 $X=(Y,Z)'$，Y 和 Z 相互独立且可积，那么：

$$\mathbb{E}(YZ)=\mathbb{E}(Y)\mathbb{E}(Z)$$

因而，如果两个随机变量相互独立，那么其协方差 $\mathrm{Cov}(Y,Z)=\mathbb{E}(YZ)-\mathbb{E}(Y)\mathbb{E}(Z)=0$。然而反之并不成立，参见例 (10)。

5 条件期望

令 (Y,X) 为一个二元的随机向量。我们经常碰到的问题是，如何使用随机变量 X 预测随机变量 Y，在统计中，我们把这类问题成为**回归**（**Regression**）。如果我们观察到了随机变量 X 的值，那么 X 的何种函数形式可以更好的预测 Y 呢？为此比较常见的做法是最小化**均方误差**（**mean squared error**）：

$$\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[(Y-h(X))^2\right]\right\} \tag{4}$$

即选择一个函数 h 使得目标函数 $\mathbb{E}\left[(Y-h(X))^2\right]$ 最小，其中

$$\mathbb{H}=\left\{h|h:\mathbb{R}\to\mathbb{R}, \mathbb{E}\left[(h(X))^2\right]<\infty\right\}$$

注意到，如果

$$h_0(X)=\arg\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[(Y-h(X))^2\right]\right\}$$

那么我们可以定义误差项 $\epsilon=Y-h_0(X)$，我们有：$\mathbb{E}[\epsilon\cdot g(X)]=0$，其中 $g(X)$ 为随机变量 X 的任意函数。通过反证法证明，如果存在 $g(X)$ 使得 $\mathbb{E}[\epsilon\cdot g(X)]\neq 0$，

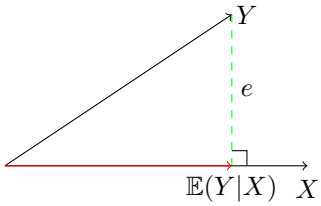


图 2: 条件期望图示

那么我们令

$$h(X)=h_0(X)+\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}g(X)$$

那么：

$$\begin{aligned}\mathbb{E}\left[(Y-h(X))^2\right]&=\mathbb{E}\left[\left(Y-h_0(X)-\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}g(X)\right)^2\right]\\&=\mathbb{E}\left[(Y-h_0(X))^2\right]+\mathbb{E}\left(\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}g(X)\right)^2\\&\quad-2\mathbb{E}\left(e(X)g(X)\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}\right)\\&=\mathbb{E}\left[(Y-h_0(X))^2\right]+\left(\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}\right)^2\mathbb{E}g^2(X)\\&\quad-2\mathbb{E}[\epsilon g(X)]\frac{\mathbb{E}[g(X)\epsilon]}{\mathbb{E}[g^2(X)]}\\&=\mathbb{E}\left[(Y-h_0(X))^2\right]-\frac{(\mathbb{E}[g(X)\epsilon])^2}{\mathbb{E}[g^2(X)]}\\&<\mathbb{E}\left[(Y-h_0(X))^2\right]\end{aligned}$$

因而如果 $h_0(X)$ 使得 (4) 式最小化，那么对于任意的函数 $g(X)$，我们一定有 $\mathbb{E}(g(X)[Y-h_0(X)])=0$。由于这个特性，我们一般称 $h(X)$ 为 Y 在 X 上的**正交投影**（**Orthogonal projection**）。直观上，我们可以把随机变量 X,Y 想象为两个向量，那么如图 (2) 所示，在 X 上距离 Y 最近的一点即 Y 点向 X 的方向上做垂线，而垂线与 X 是正交的。

如果令 $g(X)=1$，那么我们有 $\mathbb{E}[\epsilon\cdot g(X)]=\mathbb{E}[\epsilon]=\mathbb{E}[Y-h_0(X)]=0$，因而 $\mathbb{E}(Y)=\mathbb{E}(h_0(X))$。

我们知道，$\mathbb{E}(Y)=\arg\min_{c\in\mathbb{R}}\left\{\mathbb{E}(Y-c)^2\right\}$，仿照上式，我们可以定义随机变量 Y 给定 X 的**条件期望**（**Conditional expectation**）：

$$\mathbb{E}(Y|X)=h_0(X)=\arg\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[(Y-h(X))^2\right]\right\}$$

因而随机变量 Y 给定 X 的条件期望实际上是一个关于 X 的函数。对于条件期

望，我们有如下几个结论：

定理 3. （条件期望的性质）对于任意的可测函数 $g(X)$，条件期望有如下性质：

1. $\mathbb{E}[g(X)|X] = g(X)$；
2. $\mathbb{E}[(Y - \mathbb{E}(Y|X)) \cdot g(X)] = 0$；
3. $\mathbb{E}[\mathbb{E}(Y|X)] = \mathbb{E}(Y)$，$\mathbb{E}[Y - \mathbb{E}(Y|X)] = 0$；
4. $\mathbb{E}[(g(X) \cdot Y)|X] = g(X) \cdot \mathbb{E}(Y|X)$；
5. $\mathbb{E}(aY_1 + bY_2|X) = a\mathbb{E}(Y_1|X) + b\mathbb{E}(Y_2|X)$。

其中第一条性质可以由条件期望的定义得到；第二条性质与第三条性质上文已经说明，两者意味着 $\mathrm{Cov}(g(X), Y - \mathbb{E}(Y|X)) = 0$，即误差项 $\epsilon = Y - h_0(X)$ 与 X 的任意函数都不相关；第四条性质同样可以使用条件期望的定义证明；最后一条即条件期望的线性可加性。

相应的，我们还可以定义随机变量的条件方差 $\mathrm{Var}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right]$。根据条件期望的性质：

$$\begin{aligned}
\mathrm{Var}(Y|X) &= \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right] \\
&= \mathbb{E}\left\{\left[Y^2 + [\mathbb{E}(Y|X)]^2 - 2Y\mathbb{E}(Y|X)\right]|X\right\} \\
&= \mathbb{E}\left(Y^2|X\right) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}[Y\mathbb{E}(Y|X)|X] \\
&= \mathbb{E}\left(Y^2|X\right) + [\mathbb{E}(Y|X)]^2 - 2\mathbb{E}(Y|X)\mathbb{E}[Y|X] \\
&= \mathbb{E}\left(Y^2|X\right) - [\mathbb{E}(Y|X)]^2
\end{aligned}$$

其中第 4 个等号由于 $\mathbb{E}(Y|X)$ 也是 X 的函数，所以根据定理 (3.4)，可以从条件期望中提取出来。

例 13. 假设每天到达银行的人数服从泊松分布 $N \sim P(\lambda)$，而每个到达银行的人，办理外汇业务的概率为 p。那么每一天来银行办理外汇业务的人数 M 服从二项分布，即 $M|N \sim Bi(N, p), N \sim P(\lambda)$。那么每天来银行办理外汇业务的人数的期望：

$$\mathbb{E}(M) = \mathbb{E}[\mathbb{E}(M|N)] = \mathbb{E}(Np) = p\mathbb{E}(N) = p\lambda$$

如果对于随机变量 X, Y，我们取 $1_A(x) = 1\ if\ x \in X(A)$，这是一个随机变量 X 的函数，因而根据定理 (3.2)，有：

$$\mathbb{E}(Y \cdot 1_A(X)) = \mathbb{E}[\mathbb{E}(Y|X) \cdot 1_A(X)] = \mathbb{E}[h_0(X) \cdot 1_A(X)] \tag{5}$$

如果 X 是一个离散的随机变量，那么我们令 $A = \{X = x_i\}$，那么：

$$\mathbb{E}(Y \cdot 1\{X = x_i\}) = h_0(x_i) \cdot P(X = x_i)$$

从而：

$$\mathbb{E}\left(Y|X=x_i\right)=h_0\left(x_i\right)=\frac{\mathbb{E}\left(Y\cdot 1\left\{X=x_i\right\}\right)}{P\left(X=x_i\right)}=\frac{\sum_{k=0}^{\infty}\left[y_k\cdot P\left(Y=y_k,X=x_i\right)\right]}{P\left(X=x_i\right)}$$

而对于连续型随机变量，可以证明

$$\mathbb{E}\left(Y|X=x\right)=h_0\left(x\right)=\frac{\int yf\left(x,y\right)dy}{f_X\left(x\right)}$$

如果对于离散型随机变量，定义

$$f_{Y|X}\left(y|x\right)=\frac{P\left(Y=y,X=x\right)}{P\left(X=x\right)}$$

对于连续型随机变量，定义

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

那么条件期望可以写为 $\mathbb{E}\left(Y|X\right)=\int yf_{Y|X}\left(y|x\right)dy$，因而我们把 $f_{Y|X}\left(y|x\right)$ 定义为**条件密度函数**（**conditional denstiy function**）。根据定义，如果随机变量 X 和 Y 是独立的，那么：

$$f_{Y|X}\left(y|x\right)=\frac{f\left(x,y\right)}{f_X\left(x\right)}=\frac{f_X\left(x\right)\cdot f_Y\left(y\right)}{f_X\left(x\right)}=f_Y\left(y\right)$$

因而两个随机变量独立的充要条件是 $f_{Y|X}=f_Y$。

例 14. （条件密度函数）例 (2) 中，其条件概率密度函数如下表所示：

$Z\backslash Y$	2	3	4	5	6	7	8	$f_{Z\|Y}\left(z\|Y=2\right)$	$f_{Z\|Y}\left(z\|Y=4\right)$
1	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	1	$\frac{2}{3}$
2	0	0	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{2}{16}$	0	0	0	$\frac{1}{3}$
3	0	0	0	0	$\frac{1}{16}$	$\frac{2}{16}$	0	0	0
4	0	0	0	0	0	0	$\frac{1}{16}$	0	0
$f_{Y\|Z}\left(y\|Z=1\right)$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{2}{7}$	0	0			
$f_{Y\|Z}\left(y\|Z=2\right)$	0	0	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	0			

例 15. 对于联合正态密度函数：

$$f_{X,Y}\left(x,y\right)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{\left(1-\rho^2\right)}}$$
$$\cdot\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{\left(x-\mu_X\right)^2}{\sigma_X^2}+\frac{\left(y-\mu_Y\right)^2}{\sigma_Y^2}-\frac{2\rho\left(x-\mu_X\right)\left(y-\mu_Y\right)}{\sigma_X\sigma_Y}\right]\right\}$$

其中 $-1<\rho<1$，其边际密度函数为:

$$\begin{aligned}
f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \\
&\cdot \int_{\mathbb{R}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(1-\rho^2)(x-\mu_X)^2}{\sigma_X^2}+\left(\frac{y-\mu_Y}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right]\right\}dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \\
&\cdot \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{y-\mu_Y}{\sigma_Y}-\frac{\rho(x-\mu_X)}{\sigma_X}\right)^2\right\}dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \\
&\cdot \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}}\int_{\mathbb{R}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_X}\exp\left\{-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}
\end{aligned}$$

因而二元联合正态分布的边缘密度分布仍然是正态分布。其条件分布:

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}}\cdot\sqrt{2\pi}\sigma_X \\
&\cdot \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\} \\
&\cdot \exp\left\{\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\} \\
&= \frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}}\exp\left\{-\frac{1}{2}\left(\frac{y-\left[\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]}{\sigma_Y\sqrt{(1-\rho^2)}}\right)^2\right\}
\end{aligned}$$

因而 $Y|X\sim N\left(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\sigma_Y^2(1-\rho^2)\right)$，也是正态分布。进而，条件期望 $\mathbb{E}(Y|X=x)=\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$。

以上我们针对两个随机变量 Y 和 X 定义了条件期望 $\mathbb{E}(Y|X)$。条件期望

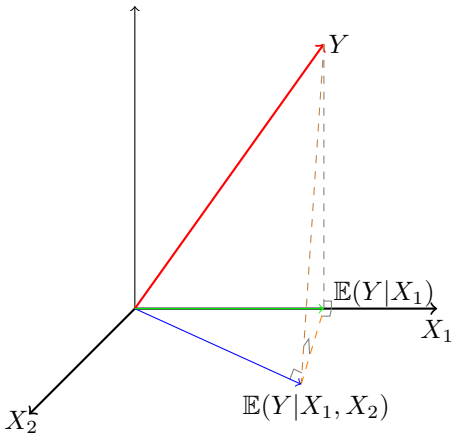


图 3: 迭代期望公式图示

可以很方便的扩充到多个 X 的情形，比如 $\mathbb{E}(Y|X_1,X_2)$ 可以定义为：

$$\mathbb{E}(Y|X_1,X_2)=h_0(X_1,X_2)=\arg\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[(Y-h(X_1,X_2))^2\right]\right\}$$

条件期望有如下性质：

$$\mathbb{E}\left[\mathbb{E}(Y|X_1,X_2)|X_1\right]=\mathbb{E}(Y|X_1)$$

即如果我们对随机变量 Y，先在大的空间上投影，再在这个大的空间上的一个小的子空间上进行投影，与直接在这个小的空间上进行投影是相等的。图 (3) 展示了一个线性投影的示例，注意条件期望是一个更加广义的非线性投影。以上公式我们称之为**迭代期望公式** (**Law of iterated expectation**)。定理 (3.4) 可以看成是令 X_1 为常数的特殊情形。

以上条件期望的概念还可以继续推广。首先我们引入一个随机变量生成的 $\sigma-$ 代数的概念。

定义 5. 令 X 为一个随机变量，令

$$\sigma\langle X\rangle=\sigma\left\langle X^{-1}(A):A\in\mathcal{B}\right\rangle$$

即包含 $\left\{X^{-1}(A):A\in\mathcal{B}\right\}$ 的最小 $\sigma-$ 代数。

例 16. 例 (2) 中，随机变量 Z 可能取值为：$\{1,2,3,4\}$，因而：

$$\begin{aligned}\sigma\langle X\rangle &= \sigma\left\langle Z^{-1}(A) : A \in \mathcal{B}\right\rangle \\ &= \sigma < \{(1,1),(2,1),(3,1),(4,1),(1,2),(1,3),(1,4)\}, \\ &\quad \{(2,2),(2,3),(2,4),(3,2),(4,2)\}, \\ &\quad \{(3,3),(3,4),(4,3)\}, \\ &\quad \{(4,4)\} >\end{aligned}$$

实际上，如果我们只知道 $Z=3$，我们知道实际发生的情况应该是 $\{(3,3),(3,4),(4,3)\}$ 中的某一种。因而如果给定 $Z=3$，我们把之前的 16 种情况降低到了 3 种情况。

在上例中，Z 总共有 4 种可能的取值，在每种 Z 的可能取值的情况下，都可以把 16 种情况降低为更少的情况，因而增大了信息量。而如果我们使用随机变量 Y，Y 共有 7 种可能的取值，给定 Y 也会增大我们的信息量。而如果给定 (X,Y) 两个随机变量，可以更加细分为 10 种情况，我们可以得到 $\sigma\langle X\rangle \subset \sigma\langle X,Y\rangle, \sigma\langle Y\rangle \subset \sigma\langle X,Y\rangle$，即两个随机变量提供了比单独一个随机变量更多的信息。

现在，如果给定 $Z=3$，那么我们可以把 $\mathbb{E}(Y|Z=3)$ 看成是 $\{(3,3),(3,4),(4,3)\}$ 中三种情况下 Y 的均值，即

$$\mathbb{E}(Y|Z=3) = \frac{1}{3}[(3+3)+(3+4)+(4+3)] = \frac{20}{3}$$

类似的，对于概率空间 $(\Omega,\mathscr{F},\mathscr{P})$，我们可以对 $\mathscr{F}$ 的一个子 $\sigma-$ 代数 $\mathscr{G}\subset\mathscr{F}$ 定义条件期望如下：

定义 6. 对于概率空间 $(\Omega,\mathscr{F},\mathscr{P})$，$\mathscr{G}\subset\mathscr{F}$ 为一个 $\sigma-$ 代数，如果对于任意的 $A\in\mathscr{G}$，随机变量 H 满足：

$$\mathbb{E}(Y\cdot 1_A) = \mathbb{E}(H\cdot 1_A)$$

那么我们称 H 为给定 $\mathscr{G}$ 随机变量 Y 的条件期望，记为 $\mathbb{E}(Y|\mathscr{G})$。令 $B\in\mathscr{F}$，定义 $\mathscr{P}(B|\mathscr{G}) = \mathbb{E}(1_B|\mathscr{G})$ 为条件概率。

注意以上定义与式 (5) 相同，所以以上定义的 $\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma\langle X\rangle)$。特别的，令 $\mathscr{G}=\{\emptyset,\Omega\}$，$\mathbb{E}(Y|\{\emptyset,\Omega\}) = \mathbb{E}(Y)$，即信息量最小的条件期望即为期望本身。而以上的迭代期望公式也可以相应推广，即如果 $\mathscr{G}_1\subset\mathscr{G}_2\subset\mathscr{F}$，那么：

$$\mathbb{E}(Y|\mathscr{G}_1) = \mathbb{E}\{[\mathbb{E}(Y|\mathscr{G}_2)]|\mathscr{G}_1\}$$

即先在大的信息集上做投影，再将其投影到小的信息集上，等价于直接投影在小的信息集上。

6 常用多元随机变量

6.1 多元随机变量的位置尺度族

对于一个 n 维随机向量 X，不失一般性，我们假设 $\mathbb{E}(X)=0$，我们记其协方差矩阵 $\mathrm{Var}(X)=\mathbb{E}(XX')=\Sigma$。根据定义，$\Sigma$ 为 $n\times n$ 维实对称矩阵，因而该矩阵一定可以被对角化为一个正交矩阵 Γ 和一个对角矩阵 $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_n)$：

$$\Sigma=\Gamma\Lambda\Gamma'$$

其中 Γ 为正交矩阵。此外，由于 $\mathrm{Var}(X)$ 是一个半正定矩阵，因而我们有特征值 $\lambda_i\geq 0, i=1,...,n$。现在我们定义对角矩阵的幂为：

$$\Lambda^p=\mathrm{diag}(\lambda_1^p,...,\lambda_n^p)$$

进而定义实对称矩阵的幂为：

$$\Sigma^p=\Gamma\Lambda^p\Gamma'$$

特别的，$\Sigma^{-1}=\Gamma\Lambda^{-1}\Gamma'=\Gamma\mathrm{diag}\left(\lambda_1^{-1},...,\lambda_n^{-1}\right)\Gamma'$，$\Sigma^{-\frac{1}{2}}=\Gamma\Lambda^{-\frac{1}{2}}\Gamma'=\Gamma\mathrm{diag}\left(\lambda_1^{-\frac{1}{2}},...,\lambda_n^{-\frac{1}{2}}\right)\Gamma'$。我们有：

$$\begin{aligned}
\Sigma^{-1}\Sigma&=\Gamma\Lambda^{-1}\underbrace{\Gamma'\Gamma}_{I}\Lambda\Gamma'=\Gamma\underbrace{\Lambda^{-1}\Lambda}_{I}\Gamma'=I\\
\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}&=\Gamma\Lambda^{-\frac{1}{2}}\underbrace{\Gamma'\Gamma}_{I}\Lambda\underbrace{\Gamma'\Gamma}_{I}\Lambda^{-\frac{1}{2}}\Gamma=\Gamma\underbrace{\Lambda^{-\frac{1}{2}}\Lambda\Lambda^{-\frac{1}{2}}}_{I}\Gamma'=I\\
\Sigma^{\frac{1}{2}}\Sigma^{\frac{1}{2}}&=\Gamma\Lambda^{\frac{1}{2}}\underbrace{\Gamma'\Gamma}_{I}\Lambda^{\frac{1}{2}}\Gamma=\Gamma\underbrace{\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}}_{\Lambda}\Gamma'=\Sigma
\end{aligned}$$

现在令 $Y=\Sigma^{-\frac{1}{2}}X$，那么：

$$\begin{aligned}
\mathrm{Var}(Y)&=\mathbb{E}\left(\Sigma^{-\frac{1}{2}}XX'\Sigma^{-\frac{1}{2}}\right)\\
&=\Sigma^{-\frac{1}{2}}\mathbb{E}(XX')\Sigma^{-\frac{1}{2}}\\
&=\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}\\
&=I
\end{aligned}$$

因而新生成的随机向量 Y 为方差为 1 且两两不相关的随机变量。

一般的，对于任意 $n\times n$ 维**实对称正定矩阵** M 以及 n 维向量 b，令 $Y=M^{\frac{1}{2}}X+b$，那么 $X=M^{-\frac{1}{2}}(Y-b)$，那么其分布函数为：

$$F_Y(y)=F_X\left(M^{-\frac{1}{2}}(y-b)\right)$$

相反，对于满足上式的一系列分布 $\{P_{b,M}:M\text{为实对称正定矩阵}\}$，我们称

之为多元随机变量的位置尺度族，这是对一元随机变量的自然推广。如果密度函数存在，那么其密度函数为：

$$f_Y(y) = \left|M^{-\frac{1}{2}}\right| f_X\left(M^{-\frac{1}{2}}(y-b)\right)$$

其中 $\left|M^{-\frac{1}{2}}\right|$ 为 $M^{-\frac{1}{2}}$ 的行列式值。

例 17. （多元正态分布）如果 $Z_1, ...Z_n$ 为独立的正态分布，那么随机向量 $(Z_1, ..., Z_n)$ 的联合密度函数为：

$$f(z) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\sum_{i=1}^n \frac{z_i^2}{2}\right\} = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{z'z}{2}\right), x = (z_1, ..., z_n)'$$

那么给定一个 $n \times n$ 维实对称正定矩阵 Σ 以及 n 维向量 μ，$X = \Sigma^{\frac{1}{2}} Z + \mu$，的密度函数为：

$$f_{\mu,\Sigma}(x) = (2\pi)^{-\frac{n}{2}} \left|\Sigma^{-\frac{1}{2}}\right| \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}, x = (x_1, ..., x_n)' \quad (6)$$

我们称满足以上密度函数的所有分布为**多元正态分布**（**Multivariate normal distribution**），如果随机向量 X 服从上述多元正态分布，我们简记为 $X \sim N(\mu, \Sigma)$。由于标准正态分布的期望为 0，协方差矩阵为单位阵，因而 $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \Sigma$。

6.2 多元正态分布

前面在例 (17) 中，我们定义了联合正态分布。由于接下来我们将大量使用联合正态分布，这里我们将详细讨论联合正态分布的一些性质。

由前所述，n 维多元正态分布实际上是 n 个独立的正态分布的联合分布生成的位置尺度族，如果 $X \sim N(\mu, \Sigma)$，那么 $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \Sigma$。现在，假设随机向量 X 的分量两两不相关，$\mathrm{Cov}(X_i, X_j) = 0$，那么 $\Sigma = \mathrm{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)$。带入式 (6) 中，得到：

$$\begin{aligned} f_{\mu,\Sigma}(x) &= (2\pi)^{-\frac{n}{2}} \left|\Sigma^{-\frac{1}{2}}\right| \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right\} \\ &= \prod_{i=1}^n f_{\mu_i,\sigma_i^2}(x_i) \end{aligned}$$

其中 $f_{\mu_i,\sigma_i^2}(x_i)$ 为一元正态分布的密度函数。因而如果 X 服从多元正态分布且其分量之间两两不相关，那么其分量之间也是独立的。尽管一般来说不相关得不到独立，但是如果随机变量服从联合正态分布，不相关可以得到独立。

在位置尺度族中我们限定矩阵必须为实对称矩阵，而实际上，任意给定一个矩阵 $M_{k\times n}$ 以及一个向量 $\zeta_{k\times 1}$，如果 $X\sim N(\mu,\Sigma)$，随机向量 $Y=MX+\zeta$ 仍然服从正态分布，即 $Y\sim N(M\mu+\zeta,M\Sigma M')$。特别的，令 $k=1$，即 M 为 $1\times n$ 维向量，那么 Y 为一个一元的随机变量，也服从正态分布。因而正态分布之和也为正态分布。

现在考虑分量之间两两不相关且方差相同的联合正态分布 $X\sim N\left(\mu,\sigma^2 I\right)$，如果我们有一个正交矩阵 $\Gamma_{n\times n}$，$\Gamma\Gamma'=I$，那么 $\mathbb{E}(\Gamma X)=\Gamma\mu$，$\mathrm{Var}(\Gamma X)=\Gamma\mathrm{Var}(X)\Gamma'=\sigma^2\Gamma\Gamma'=\sigma^2 I$，因而：

$$\Gamma X\sim N\left(\Gamma\mu,\sigma^2 I\right)$$

特别的，如果 $X\sim N(0,I)$，那么 $\Gamma X\sim N(0,I)$，即联合标准正态分布经过一个正交矩阵变换之后，仍然是联合标准正态分布。

此外，根据例 (15)，如果 $X\sim N(\mu,\Sigma)$，那么其边缘分布和条件分布都为正态分布。特别的，对于二维的联合正态分布随机变量 $X=(X_1,X_2)\sim N(\mu,\Sigma)$，其中 $\mu=(\mu_1,\mu_2)'$，

$$\Sigma=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$$

边缘分布 $X_1\sim N\left(\mu_1,\sigma_1^2\right)$，条件分布

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

相关系数 $\mathrm{Corr}(X_1,X_2)=\rho$，条件期望 $\mathbb{E}(X_1|X_2)=\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)$。现在定义

$$\epsilon=X_1-\mathbb{E}(X_1|X_2)=X_1-\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)$$

由于正态分布之和（差）仍为正态分布，因而随机变量 ϵ 也为正态分布，其期望 $\mathbb{E}(\epsilon)=\mathbb{E}(X_1-\mathbb{E}(X_1|X_2))=0$，方差 $\mathrm{Var}(\epsilon)=\mathrm{Var}\left(X_1-\rho\frac{\sigma_1}{\sigma_2}X_2\right)=\sigma_1^2+\rho^2\frac{\sigma_1^2}{\sigma_2^2}\sigma_2^2-2\cdot\rho\frac{\sigma_1}{\sigma_2}\cdot\rho\sigma_1\sigma_2=\left(1-\rho^2\right)\sigma_1^2$，因而 $\epsilon\sim N\left(0,\left(1-\rho^2\right)\sigma_1^2\right)$。将上式重写，有 $X_1=\mu_1+\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)+\epsilon=\mu_1-\rho\frac{\sigma_1}{\sigma_2}\mu_2+\rho\frac{\sigma_1}{\sigma_2}X_2+\epsilon$，上式对二维联合正态进行了分解，将其中的一个分量分解为另外一个分量和一个误差项 (ϵ) 的线性相加的形式。

这里需要提示的一点是，尽管多元正态分布的边缘分布为正态分布，但是反过来，两个正态分布在一起不一定就是联合正态分布。比如，如果 $X_1\sim N(0,1)$，而给定一个常数 c，定义

$$X_2=\begin{cases}X_1 & if\ |X_1|>c\\ -X_1 & else\end{cases}$$

可以计算，X_2 也为正态分布，但是 (X_1,X_2) 显然不是联合正态分布。

以上二元情况还可以推广，如果 $X=(X_1,X_2)'\sim N(\mu,\Sigma)$，其中 X_1 为 $k\times 1$ 向量，X_2 为 $(n-k)\times 1$ 向量，$\mu=(\mu_1,\mu_2)'$，

$$\Sigma=\begin{bmatrix}\Sigma_{k\times k} & \Sigma_{k\times(n-k)}\\ \Sigma_{(n-k)\times k} & \Sigma_{(n-k)\times(n-k)}\end{bmatrix}$$

那么边缘分布 $X_1\sim N(\mu_1,\Sigma_{k\times k})$，条件分布 $X_1|X_2\sim N\left(\widetilde{\mu},\widetilde{\Sigma}\right)$，其中：

$$\begin{aligned}\widetilde{\mu}&=\mu_1+\Sigma_{k\times(n-k)}\Sigma^{-1}_{(n-k)\times(n-k)}(X_2-\mu_2)\\ \widetilde{\Sigma}&=\Sigma_{k\times k}-\Sigma_{k\times(n-k)}\Sigma^{-1}_{(n-k)\times(n-k)}\Sigma_{(n-k)\times k}\end{aligned}$$

现在如果令 $X\sim N(\mu,\Sigma)$，令 $Y=\Sigma^{-\frac{1}{2}}(X-\mu)$，可以得到 $Y\sim N(0,I)$，进而可以得到

$$\begin{aligned}(X-\mu)'\Sigma^{-1}(X-\mu)&=Y'Y\\ &=\sum_{i=1}^{n}Y_i^2\end{aligned}$$

由于 $Y_i\sim N(0,1)$ 且 Y_i 之间相互独立，从而 $(X-\mu)'\Sigma^{-1}(X-\mu)=\sum_{i=1}^{n}Y_i^2\sim\chi^2_n$。

前面我们介绍了投影矩阵的概念，现在考虑一个投影矩阵 P，其必然可以分解为 $P=\Gamma'\Lambda\Gamma$，其中 Γ 为正交矩阵，而 Λ 为对角矩阵，且对角元只能为 1 或者 0。现在考虑一个联合正态分布 $X\sim N(0,I)$，那么：

$$\begin{aligned}X'PX&=X'\Gamma'\Lambda\Gamma X\\ &=(\Gamma X)'\Lambda(\Gamma X)\end{aligned}$$

根据之前的推理，$Y=\Gamma X\sim N(0,I)$，因而：

$$\begin{aligned}X'PX&=Y'\Lambda Y\\ &=\sum_{i=1}^{k}Y_k^2\end{aligned}$$

其中 $k=\operatorname{tr}(P)=\operatorname{tr}(\Lambda)$，因而 $X'PX\sim\chi^2_k$。

例 18. 在例 (1) 中，我们定义了 $P_0=\frac{1}{n}\iota\iota'$ 以及 $M_0=I-P_0=I-\frac{1}{n}\iota\iota'$，并有 $\operatorname{tr}(M_0)=n-1$。对于联合正态分布 $X\sim N(0,I)$，有：

$$X'M_0X=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\sim\chi^2_{n-1}$$

对于两个投影矩阵 M 和 P，我们有如下定理：

定理 4. 如果 n 维随机向量 $X \sim N(0, I)$，矩阵 M 和 P 为投影矩阵，那么二次型 $X'MX$ 和 $X'PX$ 独立的充要条件是 $MP = 0$。

例 19. 接上例，我们有 $P_0 M_0 = 0$，因而二次型 $X'P_0X$ 和 $X'M_0X$ 是相互独立的。

在以上定理的基础之上，回顾 F 分布的定义，我们有如下定理：

定理 5. 如果 n 维随机向量 $X \sim N(0, I)$，矩阵 M 和 P 为投影矩阵且 $MP = 0$，$tr(M) = k_1, tr(P) = k_2$，那么

$$\frac{X'PX/k_2}{X'MX/k_1} \sim F_{k_2,k_1}$$

类似的，对于一个向量 $L_{n\times 1}$，我们也有如下定理：

定理 6. 如果随机向量 $X \sim N(0, I)$，矩阵 P 为投影矩阵，那么二次型 $X'PX$ 和随机变量 $L'X$ 独立的充要条件是 $PL = 0$。

回顾 t 分布的定义，相应的我们有如下定理：

定理 7. 如果随机向量 $X \sim N(0, I)$，矩阵 P 为投影矩阵，向量 L 满足 $PL = 0$，$tr(P) = k$，且 $L'L = 1$，那么

$$\frac{L'X}{\sqrt{X'PX/k}} \sim t_k$$

例 20. 如果 n 维随机向量 $X \sim N(0, I)$，取 $L = \frac{1}{\sqrt{n}}\iota$ 以及 $M_0 = I - P_0 = I - \frac{1}{n}\iota\iota'$，可以得到：

$$\begin{aligned} M_0 L &= \left(I - \frac{1}{n}\iota\iota'\right)\frac{1}{\sqrt{n}}\iota \\ &= \frac{1}{\sqrt{n}}\left(\iota - \frac{1}{n}\iota\iota'\iota\right) \\ &= \frac{1}{\sqrt{n}}\left(\iota - \frac{1}{n}\iota n\right) \\ &= \frac{1}{\sqrt{n}}(\iota - \iota) = 0 \end{aligned}$$

且 $\mathbb{E}(L'X) = 0$，$\mathrm{Var}(L'X) = L'IL = \frac{1}{n}\iota'\iota = 1$，因而 $L'X \sim N(0, 1)$。根据例 (18)，$X'M_0X \sim \chi^2(n-1)$，因而：

$$\frac{L'X}{\sqrt{X'M_0X/(n-1)}} = \frac{\sqrt{n}\bar{X}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}} \sim t_{n-1}$$

更加一般的，如果 $X \sim N\left(\mu, \sigma^2 I\right)$ 为独立同分布的随机向量，那么 $\frac{1}{\sigma}(X-\mu) \sim N(0, I)$，因而：

$$\begin{aligned}\frac{L'\left(\frac{1}{\sigma}(X-\mu)\right)}{\sqrt{\left[\frac{1}{\sigma}(X-\mu)\right]' M_0\left[\frac{1}{\sigma}(X-\mu)\right]/(n-1)}} &= \frac{\frac{1}{\sigma}L'((X-\mu))}{\sqrt{\frac{1}{\sigma^2}X'M_0X/(n-1)}} \\ &= \frac{\frac{1}{\sigma}\sqrt{n}\left(\bar{X}-\mu\right)}{\frac{1}{\sigma}\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}} \\ &= \frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}} \sim t_{n-1}\end{aligned}$$

其中 $M_0(X-\mu) = M_0(X-\mu\iota) = M_0X - \mu M_0\iota = M_0X$。即如果随机向量 X 为独立且同分布的正态分布，那么：

$$\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}} \sim t_{n-1}$$

6.3 指数分布族

在上一节中我们讨论了单参数指数分布族，这一节中我们把指数分布族进一步推广。更加一般化的指数分布族的定义如下：

定义 7.（**指数分布族**）对于一个参数族 $\{P_\theta, \theta \in \Theta\}$，如果其概率密度（质量）函数可以写成如下形式：

$$f(x|\theta) = h(x) \cdot \exp\left\{\sum_{i=1}^{k}\left[\eta_i(\theta) \cdot T_i(x)\right] - B(\theta)\right\} \tag{7}$$

那么我们称 $\{P_\theta, \theta \in \Theta\}$ 为**指数分布族**（**Exponential family**）。

如果使用向量的形式，令

$$\eta(\theta) = \begin{bmatrix}\eta_1(\theta) \\ \eta_2(\theta) \\ \vdots \\ \eta_k(\theta)\end{bmatrix}, T(x) = \begin{bmatrix}T_1(x) \\ T_2(x) \\ \vdots \\ T_k(x)\end{bmatrix}$$

为列向量[1]，那么方程 (2) 也可以写为：

$$f(x|\theta) = h(x) \cdot \exp\left\{\eta(\theta)' T(x) - B(\theta)\right\}$$

[1]根据惯例，向量一般写为列向量的形式。

例 21. 正态分布的密度函数:

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

如果令 $\theta=\begin{pmatrix}\mu\\ \sigma\end{pmatrix}\in\Theta=\mathbb{R}\times\mathbb{R}^{+}$，那么其密度函数可以写为:

$$\begin{aligned}f\left(x|\mu,\sigma\right) &= \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\\ &= \frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{\left(x-\mu\right)^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\\ &= \frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{x^2-2\mu x+\mu^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\\ &= \frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\ln\left(\sigma\right)\right\}\end{aligned}$$

令 $h\left(x\right)=\frac{1}{\sqrt{2\pi}}$，$\eta\left(\theta\right)=\begin{pmatrix}-\frac{1}{2\sigma^2}\\ \frac{\mu}{\sigma^2}\end{pmatrix}$，$T\left(x\right)=\begin{pmatrix}x^2\\ x\end{pmatrix}$，$B\left(\theta\right)=\frac{\mu^2}{2\sigma^2}+\ln\left(\sigma\right)$，可以得到正态分布也属于指数分布族。

需要注意的是，在指数分布族中，其密度函数:

$$\begin{aligned}f\left(x|\theta\right)&=h\left(x\right)\cdot\exp\left\{\sum_{i=1}^{k}\left[\eta_i\left(\theta\right)\cdot T_i\left(x\right)\right]-B\left(\theta\right)\right\}\\ &=h\left(x\right)\cdot\exp\left\{-B\left(\theta\right)\right\}\cdot\exp\left\{\sum_{i=1}^{k}\left[\eta_i\left(\theta\right)\cdot T_i\left(x\right)\right]\right\}\\ &\triangleq\frac{1}{\mathcal{B}\left(\theta\right)}\cdot h\left(x\right)\exp\left\{\sum_{i=1}^{k}\left[\eta_i\left(\theta\right)\cdot T_i\left(x\right)\right]\right\}\end{aligned}$$

而由于 $\int f\left(x|\theta\right)dx=1$，因而

$$\mathcal{B}\left(\theta\right)=\int h\left(x\right)\exp\left\{\sum_{i=1}^{k}\left[\eta_i\left(\theta\right)\cdot T_i\left(x\right)\right]\right\}dx$$

这意味着指数分布族密度函数中的四个函数: $h\left(x\right),T\left(x\right),\eta\left(\theta\right),B\left(\theta\right)$ 并不是独立任意选取的。

与单参数的指数分布族类似，我们通常会把密度函数重新参数化，即对于指数分布族

$$f\left(x|\theta\right)=h\left(x\right)\cdot\exp\left\{\eta\left(\theta\right)'T\left(x\right)-B\left(\theta\right)\right\}$$

我们令 k 维向量 $\lambda=\eta(\theta)$，那么指数分布族可以写为：

$$f(x|\theta)=h(x)\cdot\exp\{\lambda' T(x)-C(\lambda)\} \tag{8}$$

我们将指数分布族重新参数化为式 (8) 的形式，并将这种形式成为**规范形式**（**Canonical form**），新的参数称之为**自然参数**（**Natural parameter**），而新的参数的参数空间 Λ 为**自然参数空间**（**Natural parameter space**）。

例 22. 在正态分布例 (21) 中，可以令 $\lambda=(\lambda_1,\lambda_2)'=\left(-\frac{1}{2\sigma^2},\frac{\mu}{\sigma^2}\right)'$，而

$$C(\lambda)=-\frac{\lambda_2^2}{4\lambda_1}-\frac{\ln(-2\lambda_1)}{2}$$

其中 $\mu=-\frac{\lambda_2}{2\lambda_1},\sigma^2=-\frac{1}{2\lambda_1}$。由此我们写出了正态分布的规范形式。

在有了指数分布族的规范形式和向量导数的概念之后，我们可以介绍一下定理：

定理 8. 对于一个规范形式的指数分布族的随机变量 $X\sim P_\lambda\in\{P_\lambda(x),\lambda\in\Lambda\}$，有：

1. Λ 为一个凸集
2. $C(\lambda)$ 为凸函数（$\frac{\partial^2 C(\lambda)}{\partial\lambda\partial\lambda'}$ 为正定矩阵）
3. $\mathbb{E}[T(X)]=\frac{\partial C(\lambda)}{\partial\lambda}$，$Var[T(X)]=\mathbb{E}\left[T(X)T(X)'\right]=\frac{\partial^2 C(\lambda)}{\partial\lambda\partial\lambda'}$

例 23. 例 (22) 中我们得到了正态分布的规范形式，其中 $T(X)=\left(X^2,X\right)'$，因而使用上述定理：

$$\mathbb{E}[T(X)]=\mathbb{E}\left(\begin{bmatrix}X^2\\X\end{bmatrix}\right)=\frac{\partial C(\lambda)}{\partial\lambda}=\begin{bmatrix}\frac{\lambda_2^2}{4\lambda_1^2}-\frac{1}{2\lambda_1}\\-\frac{\lambda_2}{2\lambda_1}\end{bmatrix}=\begin{bmatrix}\mu^2+\sigma^2\\\mu\end{bmatrix}$$

在贝叶斯统计中，经常需要计算两个密度函数的乘积，诸如以下形式：

$$f(x|\theta)\cdot\pi(\theta)$$

其中 $\pi(\theta)$ 为为参数 θ 的先验分布。如果概率密度函数 $f(x|\theta)$ 可以写为指数分布族的形式，即：

$$f(x|\theta)=h(x)\cdot\exp\left\{\sum_{i=1}^{k}\left[\eta_i(\theta)\cdot T_i(x)\right]-B(\theta)\right\}$$

现在我们将以上密度中的参数（θ）视为变量，而将 T_i 视为参数，那么令：

$$\pi_t(\theta) = \exp\left\{\sum_{i=1}^{k}[t_i \cdot \eta_i(\theta)] - t_{k+1}B(\theta) - \ln[k(t)]\right\}$$

其中 $\ln[k(t)]$ 使得 $\int \pi(\theta)\, d\theta = 1$。可以得到

$$\begin{aligned} f(x|\theta)\cdot\pi_t(\theta) &= h(x)\cdot\exp\left\{\sum_{i=1}^{k}[(T_i(x)+t_i)\cdot\eta_i(\theta)] - (t_{k+1}+1)B(\theta) - \ln(k(t))\right\} \\ &= h(x)\cdot\exp\left\{\sum_{i=1}^{k}[s_i(x)\cdot\eta_i(\theta)] - s_{k+1}B(\theta) - \ln(k(t))\right\} \end{aligned}$$

其中 $s_i(x) = (T_i(x)+t_i), i = 1, ..., k$，$s_{k+1} = t_{k+1}+1$。可以发现两者相乘之后得到的密度函数与 $\pi_t(\theta)$ 有着相同的密度函数。我们称 $\pi_t(\theta)$ 为 $f(x|\theta)$ 的**共轭先验**（**Conjugate prior**）。

例 24. 对于一个方差已知的正态分布 $X \sim N(\mu, \sigma_0^2)$，其密度函数：

$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\cdot\exp\left\{-\frac{1}{2\sigma_0^2}x^2 + \frac{\mu}{\sigma_0^2}x - \frac{\mu^2}{2\sigma_0^2} - \ln(\sigma_0)\right\}$$

如果视 μ 为变量，那么：

$$\begin{aligned} \pi(\mu) &\propto \exp\left\{\frac{\mu}{\sigma_0^2}t_1 - \frac{\mu^2}{2\sigma_0^2}t_2\right\} \\ &= \exp\left\{-\frac{t_2\left(\mu^2 - 2\mu\frac{t_1}{t_2} + \frac{t_1^2}{t_2^2}\right) - \frac{t_1^2}{t_2}}{2\sigma_0^2}\right\} \\ &\propto \exp\left\{-\frac{\left(\mu - \frac{t_1}{t_2}\right)^2}{2\left(\frac{\sigma_0}{t_2}\right)^2}\right\} \end{aligned}$$

因而 $\pi(\mu)$ 为 $N\left(\frac{t_1}{t_2}, \left(\frac{\sigma_0}{t_2}\right)^2\right)$ 的正态分布。

习题

练习 1. 对于一个向量 $x \in \mathbb{R}^n$ 以及一个权重向量 $w \in \mathbb{R}^n$，$\sum_{i=1}^{n} w_i = \iota' w = 1$，我们希望计算其加权平均：

$$\bar{x}_w = \sum_{i=1}^{n} w_i \cdot x_i$$

请写出一个幂等矩阵 P_w 使得 $P_w x = \bar{x}_w \iota$。

练习 2. 若随机向量 (U,V) 的分布函数为:

$$F_{U,V}(u,v)=\frac{uv}{1-\theta(1-u)(1-v)},\theta\in[-1,1)$$

其中 $P(U\in[0,1])=1,P(V\in[0,1])=1$，求其边缘分布函数和边缘密度函数。

练习 3. 如果一个随机变量 $X\sim N(0,1)$，现如下定义随机变量 Y:

$$Y=\begin{cases}X-2 & with\ prob\ 0.5\\ X+2 & with\ prob\ 0.5\end{cases}$$

求 $\mathrm{Var}(Y)$。

练习 4. 证明 $g(X)\cdot\mathbb{E}(Y|X)=\arg\min_{h\in\mathbb{H}}\left\{\mathbb{E}\left[(g(X)\cdot Y-h(X))^2\right]\right\}$。

练习 5. 证明 $\mathrm{Var}(Y)=\mathrm{Var}[\mathbb{E}(Y|X)]+\mathbb{E}[\mathrm{Var}(Y|X)]$

练习 6. 使用练习 (5) 中的结论，计算例 (13) 中的 $\mathrm{Var}(M)$。

练习 7. 如果随机变量 X 和 Y 相互独立，求 $\mathbb{E}(Y|X)$。

练习 8. 使用上述结论，产生一组二维正态随机变量 $X=(X_1,X_2)$，使得第一个分量方差为 1，第二个分量方差为 2，且其相关系数为 0.5。

练习 9. Γ 分布是否属于指数分布族?

练习 10. 使用正态分布的规范形式求 $\mathbb{E}X^3$ 及 $\mathbb{E}X^4$，并验证 $\frac{\partial^2 C(\lambda)}{\partial\lambda\partial\lambda'}$ 的正定性。

练习 11. 对于一个二项分布 $X\sim Bi(N,p)$，N 已知，那么若将 p 视为变量，那么其共轭先验是什么分布?

参考文献

[1] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.

[2] Bickel, P.J., Doksum, K.A., 2001. Mathematical Statistics: Basic Ideas and Selected Topics. Prentice-Hall, Inc, New Jersey.

[3] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[4] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[5] Greene, W.H., 2013. Econometric analysis, Seventh Ed. ed. Pearson Education.

[6] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

[7] Wooldridge, J.M., 2010. Econometric Analysis of Cross Sectional and Panel Data, 2nd ed. The MIT Press, Cambridge.

第五节 · 统计与统计量

司继春

上海对外经贸大学统计与信息学院

在这一节中我们将讨论统计学的一些基本概念，这些概念是我们后面学习统计学理论的基础。我们首先介绍统计学中总体、样本和模型的概念，进而介绍统计量的概念及性质。

1 统计的基本概念

1.1 统计学中的数据

统计学是一门关于数据的学科，所有的统计方法都是围绕着数据展开的，因而我们从数据的分类入手，介绍统计学的一些基本概念。

现实生活中碰到的数据是多种多样的，针对同一个个体，我们可以通过很多特征对其进行刻画。比如，对于一个人来说，其性别、年龄、身高等都是其个人的特征；而对于一家企业来说，其所有权性质、企业年龄、注册资本等也是其特征。我们通常把这些描述个体的特征称为变量（Variable）。然而注意到，这些变量的性质并不一样。比如我们可以比较两个人的身高、年龄的大小，然而我们却不能比较性别的大小。因而尽管我们经常把数据全都编码为数值型（比如男 =1, 女 =0 等），然而这些数值的大小并不是都有意义。根据数据度量的层次，一般可以将数据分为以下三类：

1. **分类变量**（**Categorical variable**）：指数据仅仅用于区分类别，而数据没有数值上的意义，比如性别、企业注册类型等。

2. **顺序变量**（**Ordinal variable**）：指数据的值不仅仅用于区分类别，还可以用于排序。比如奖学金等级（一二三等），空气污染等级（重度污染，轻度污染，良好）等。

3. **数值变量**（**Numerical variable**）：指不仅仅数据的排序有意义，而且数据值的差是有意义的。通常又可以将数值型数据分为离散变量和连续变量，前者如次数、人数、年龄等，后者如温度、长度、金额等等。

当然，数据的分类方法并不唯一。比如有的分类方法将数据分为定类数据、定序数据、定距数据和定比数据。而还有一些数据是复合类型，比如对于**截尾数**

据（**Censored data**），就结合了顺序变量和数值变量的特点。针对不同类型的数据，使用的统计方法经常有很大的差别。

而根据时间和个体进行划分，我们经常使用的数据一般有两种最基本的数据类型：**横截面数据**（**Cross-sectional data**）与**时间序列数据**（**Time series data**）。

其中，横截面数据，或者简称截面数据，指同一时间点或者时间段，对不同主体的某些变量进行观测。比如在实验中，对于某一次实验，不同的实验对象的不同观察指标组成的数据即横截面数据。再比如，在调查数据中，很多家庭的多个变量组成的数据也是横截面数据。横截面数据只有个体上的差异而没有时间上的差别。一般我们用 N 记为数据中个体的个数。

而时间序列数据是对于一个或者多个变量在不同时间上的观测。比如 2000 年到现在我国每个季度的 GDP 即时间序列数据。再比如 2000 年到现在我国每个季度的货币供给 M0、M1、M2 也是时间序列数据。一般我们用 T 记为数据中时间的长度。时间序列数据只有时间上的差异而不存在个体上的差异。

除这两种外，还有这两种类型数据的合并数据，如常用的**面板数据**（**Panel data**）或者**纵向数据**（**Longitudinal data**）、**重复面板数据**（**Repeated cross-sectional data**）等等。其中面板数据指的是同时观测多个个体，同时对于每个个体，在不同时间段对某些变量进行观测。比如，单独看上海市从 2000 年到现在每年的 GDP 是时间序列数据，然而如果我们可以观察到全国每个省从 2000 年到现在每年的 GDP，那么就是面板数据。面板数据既有时间上的信息又有不同个体的信息，我们一般把 $N \gg T$ 的面板数据称为长面板数据。

1.2 统计模型

在获得数据之后，我们需要对这些数据建立概率模型。一般而言，一个典型的统计学问题可以如下描述：进行一次或者一系列的随机试验，并且从这些随机试验中收集了一些数据，而统计的任务就是使用这些数据提取我们希望得到的信息，得到一些结论。在统计学中，我们希望得到的结论是多种多样的，比如在机器学习中，最关心的结论是预测是否正确，而在计量经济学中，很多时候识别因果关系则是最希望得到的结论。一般来说，统计方法可以分为描述性统计方法和推断统计。

针对已经有的数据，首先可以进行一些**描述性统计**（**descriptive statistics**）的工作。描述性统计通过表和图的形式对数据加以展示，常用的描述性统计量一般包括均值、标准差、最大值、最小值、中位数、四分位数等。描述性统计经常作为初步的研究，研究者可以通过描述性统计对数据的分布情况做初步的了解，检查数据中可能有的错误，帮助研究者发现数据中特有的现象等等。

描述性统计虽然重要，但是只能作为初步的观察，而更进一步的结论则需要借助**统计推断**（**Statistical inference**）来实现。统计推断通过概率建模的方法，用已知的样本对未知的总体进行推断，一般包含着**参数估计**（**Estimation**）、**假设检验**（**Hypothesis testing**）等内容。因而在这里，理解总体和样本的概

念是非常重要的。

在统计推断理论中，数据集 $\{x_i, i=1,...,N\}$ 被视为是概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中一系列随机变量（向量）$\{X_i, i=1,...,N\}$ 的实现，概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中的概率函数 $\mathscr{P}$ 则被称为**总体**（**Population**）。

在统计推断中，我们经常将数据集 $\{x_i, i=1,...,N\}$ 建模为样本空间 $\Omega = \mathbb{R}^n$ 中随机试验的一个实现。如果随机变量 $\{X_i, i=1,...,N\}$ 中 X_i 是相互独立的，且具有相同的分布函数，那么我们称 $\{X_i, i=1,...,N\}$ 为**独立同分布的**（**independent and identically dstributed**，**i.i.d**）。如果 $\{X_i, i=1,...,N\}$ 是来自于 $\prod_{i=1}^{N}(\mathbb{R}, \mathscr{B}, P)$ 的一组随机变量，且独立同分布，我们称 $\{X_i, i=1,...,N\}$ 为**随机样本**（**Random sample**），其中总体为 P，而随机变量的个数 N 称为**样本量**（**Sample size**）。

由于随机样本其分布都相同且相互独立，因而总体 P 由 X_i 的边缘分布 $F_X(\cdot)$ 确定，其联合分布可以写为：

$$F(x) = \prod_{i=1}^{N} F_{X_i}(x_i) = \prod_{i=1}^{N} F_X(x_i)$$

其中第一个等号使用了独立的假设，而第二个等号使用了同分布的假设。

例 1. 如果我们希望使用统计手段调查一条生产线上的次品率，我们经常会对这个产品线上的产品进行抽样。如果我们抽取了 N 个样本 $\{X_i, i=1,...,N\}$，假设 $X_i = 1$ 为次品，否则为 0，X_i 独立同分布，那么这里 $\Omega = \{0,1\}^N$，而总体 P 为一个伯努利分布 $P(X_i = 1) = p$。我们的目的即希望通过样本 $\{X_i, i=1,...,N\}$ 对总体 P 做出推断。由于每个 X_i 其概率质量函数为：

$$P(X_i = x) = p^x (1-p)^{1-x}, x = 0, 1$$

因而样本 $\{X_i, i=1,...,N\}$ 的联合密度函数为：

$$P(x) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i} = p^{N_1}(1-p)^{N-N_1}$$

其中 $N_1 = \sum_{i=1}^{N} x_i$。

例 2. 有的时候我们会对某些测量感兴趣（如体温、距离、长度等），如果我们对其进行 N 次观测，假设每次观测误差都是独立同分布的，记每次观测为 X_i，那么这里 $\Omega = \mathbb{R}^N$，而其联合分布函数：

$$F(x) = \prod_{i=1}^{N} F_{X_i}(x_i) = \prod_{i=1}^{N} F_X(x_i) = \prod_{i=1}^{N} P(X_i \leq x_i)$$

其中 $F_X(\cdot)$ 为每次测量的边缘分布。在这里，我们的总体 P 与边缘分布 $F_X(\cdot)$ 是等价的。

而在一些问题中，我们所关注的个体是有限的。比如如果我们只对一批产品的合格率感兴趣，或者当我们关注全国人民的收入分布时，全国人民是一个有限的集合。然而在这些情况下，调查每一个个体经常是不现实的，所以我们需要在所关注的个体中找到一个子集进行研究。当然，对所有关注的个体进行调查也是有可能的，比如**普查**（**Census**），包括人口普查、经济普查等。

一般的，如果令 $\mathcal{P} = \{y_1, y_2, ..., y_M\}$ 为我们所关心的全体，而全体不能一一进行调查，我们通常对其一个子集 $S \subset \mathcal{P}$ 进行调查。因而这里就涉及到我们如何从 $\mathcal{P}$ 中挑选出子集 S，即**抽样**（**Sampling**）问题。

抽样方法分为两种：**概率抽样**（**Probability sampling**）和**非概率抽样**（**Nonprobability sampling**）。其中概率抽样指每个个体都有正的概率被抽中，且该概率已知（或者可以被计算）。概率抽样的统计性质良好，因而我们下面主要集中在概率抽样的条件下进行讨论。

一个最简单的概率抽样方法即从 M 个元素中等可能的不放回地抽取（**sampling without replacement**）N 个样本 $\{X_i, i = 1, ..., N\}$，即**简单随机抽样**（**simple random sampling**）。由于在这种情况下，样本的分布由 $\mathcal{P}$ 和每个个体被抽中的概率决定，因而我们通常将 $\mathcal{P}$ 称之为总体。

除了简单随机抽样之外，综合考虑调查成本和其他因素，还有其他的抽样方法，如：

1. 系统抽样（Systematic sampling）：指先根据一定规则对个体排序，再根据一定的规则选取样本。比如对一个班的学生进行抽样，可以抽取学号尾数为 1 的所有个体。

2. 分层抽样（Stratified sampling）：先对所有个体分组，再在每个组内进行抽样。比如如果需要抽取全国的样本，可以在每个城市单独抽样然后汇总。

3. 整群抽样（Cluster sampling）：先对所有个体分组，再抽取组别，进而调查被抽中的组的所有个体。比如想要在全校学生中抽样，可以抽取班级，再调查被抽中的班级的所有学生。

4. 多阶段抽样（Multi-stage sampling）：先对所有个体分组，抽取分组，再在每个组内部抽样。与分层抽样的差别是，分层抽样要求每个组内都进行抽样，而多阶段抽样只对抽中的组进行抽样。比如可以对全国所有城市中抽取 20 个城市，再在这 20 个城市中进行抽样。

5. 面板抽样（Panel sampling）：指先随机抽取样本，之后隔一段时间对同一批个体进行反复的调查，可以获得面板数据。

注意以上抽样方法并不能一定保证每个个体被抽中的概率相等。

在这里需要特别注意的是，在抽样的情况下，我们虽然可以认为样本 X_i 是同分布的，但是样本未必是独立的，因而样本的联合分布不一定可以写成样本边缘分布的乘积的形式。比如，如果我们关心全国的家庭总消费，不同地区可能

有不同的消费和物价水平，因而每个家庭的消费可能满足如下形式：

$$C_{ri} = \alpha_r + \epsilon_{ri}$$

其中 r 代表地区，而 i 代表每个家庭。如果假设 ϵ_{ri} 为独立同分布的，且 $\mathrm{Cov}(\epsilon_i, \alpha_r) = 0$，那么对于同一地区的不同个体，$\mathrm{Cov}(C_{ir}, C_{jr}) = \mathrm{Cov}(\alpha_r + \epsilon_i, \alpha_r + \epsilon_j) = \mathrm{Var}(\alpha_r)$，因而消费水平可能不是独立的，除非 $\alpha_r = \alpha$ 为所有地区都相等的常数。

以上定义了总体，即一个概率函数 P，然而现实中总体 P 是不能被观测的，而统计推断的任务就是使用可以观测的样本 $\{X_i\}$ 对未知的总体进行推断。而所谓的**统计模型**（**Statistical model**）即通过对总体 P 做一系列的假设，简化问题并对总体进行推断。

一般来说，统计模型分为**参数模型**（**parametric model**）和**非参数模型**（**nonparametric model**），以及介于两者之间的**半参数模型**（**semi-parametric model**）。

参数模型即假设总体 P 属于某一个参数族 $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$，其中 $\Theta \subset \mathbb{R}^d$，且一旦 θ 确定了，那么 P_θ 为一个概率函数。其中 Θ 被称为参数空间（parameter space），而 d 称为参数空间的维数。

对于参数族 $\{P_\theta, \theta \in \Theta\}$，如果当 $\theta_1 \neq \theta_2$ 时，必然有 $P_{\theta_1} \neq P_{\theta_2}$，那么我们称其为可识别的（identifiable）。如果参数族不可识别，意味着存在多于一个 θ 代表了同一个概率函数，或者模型的解不唯一。因而一般我们要求参数模型必须为可识别的。

例 3. 在例 (2) 中，我们可以假设每一次测量 $X_i \sim N(\mu, \sigma^2)$，其中 μ 为测量的真实值（如真实体温、长度），而 σ^2 代表每次测量可能的误差大小，且假设误差服从正态分布。在这里，我们假设总体 $P \in \{N(\mu, \sigma^2)\}$，参数空间为 $\Theta = \mathbb{R} \times \mathbb{R}^+$。只要 μ 和 σ^2 确定了，那么总体 P 也就确定了。

相反，如果对总体 P 不做任何参数上的假定，我们称其为非参数模型。此类模型并不假设 P 属于某一个参数族，而是对 P 做了其他的假定，如对分布函数的连续型、光滑性、对称性等做出假定。

2 统计量及其抽样分布

在统计理论中，所有的统计方法都是通过**统计量**（**Statistic**）实现的。一般的，对于概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 中的一组样本 $\{X_i, i = 1, 2, ..., N\}$，$X = [X_1, ..., X_N]'$，统计量即样本的一个不依赖于其具体实现的函数 $T(X)$。由于 X 为随机向量，因而统计量 $t = T(X)$ 作为随机向量的函数仍然是概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的随机变量，所以统计量 s 同样具有期望、方差、分布等随机变量所具有的特征。其中统计量的分布称为抽样分布（sampling distribution）。

统计量是所有统计方法的基本工具，在不同的统计方法中有不同的称谓，如在描述性统计中，统计量一般称之为**描述性统计量**（**descriptive statistic**）；在

参数估计中，我们称其为**估计量**（**estimator**）；而在假设检验中，我们称其为**检验统计量**（**test statistic**）。

这里需要注意参数 θ 和统计量 t 的差别。参数是总体的特征，因而是不可观测的，而估计量是样本的函数，由于样本是可以观测的，因而估计量也必须是可以计算可以观测的。统计推断的目标即使用有限的样本，通过计算样本统计量，对总体的参数做出推断。

为了符号统一，在不引起歧义的情况下，接下来我们将不区分样本 $\{X_i\}$ 及其实现 $\{x_i, i = 1, 2, ..., N\}$，即当我们使用 $\{x_i\}$ 时，仍然代表来自于总体的一组样本，即概率空间中的一系列随机变量。

接下来我们以样本均值和样本方差为例，介绍一些简单的统计量的抽样分布。

2.1 样本均值

样本均值（Mean）是对数据平均水平的最常用的度量，其定义为：

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = L'x$$

其中 $x = (x_1, x_2, ..., x_N)'$，$L = \frac{1}{N}\iota$。注意样本均值使得样本均方误差最小化，即：

$$\bar{x} = \arg\min_{c} \frac{1}{N}\sum_{i=1}^{N} (x_i - c)^2$$

如果假设 $\{x_i\}$ 为独立同分布的样本，且 $\mathbb{E}(x_i) = \mu$，$\mathrm{Var}(x_i) = \sigma^2$（或记为 $x_i \sim (\mu, \sigma^2)\ i.i.d$），那么我们有：

$$\mathbb{E}(\bar{x}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(x_i) = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu$$

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(x_i) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2 = \frac{\sigma^2}{N}$$

更进一步，如果 $\{x_i\}$ 来自于正态总体且独立同分布，即：$x_i \sim N(\mu, \sigma^2)\ i.i.d$，或者等价的记为 $x \sim N(\mu\iota, \sigma^2 I)$，那么可以得到 $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$，或者：

$$\sqrt{N}\frac{\bar{x} - \mu}{\sigma} = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} \sim N(0, 1) \tag{1}$$

即正态总体的样本均值服从正态分布。

2.2 样本方差

样本方差是数据离散程度最常用的度量，其定义为:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}x'M_0x$$

其中 $M_0 = I - P_0 = I - \frac{1}{N}\iota\iota'$。相应的，样本标准差定义为: $s = \sqrt{s^2}$。由于:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 + \bar{x}^2 - 2\bar{x}x_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 + \bar{x}^2 - \frac{1}{N}\sum_{i=1}^{N}2\bar{x}x_i \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 + \bar{x}^2 - 2\bar{x}\cdot\frac{1}{N}\sum_{i=1}^{N}x_i \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 + \bar{x}^2 - 2\bar{x}^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2
\end{aligned}$$

从而样本方差可以写为:

$$s^2 = \frac{N}{N-1}\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2\right) = \frac{N}{N-1}\left(\overline{x^2} - \bar{x}^2\right)$$

即样本均值可以写为样本平方的均值减去样本均值的平方，乘以 $\frac{N}{N-1}$。注意到在计算样本均值时，我们使用 N 做为分母，然而在计算样本方差时，我们使用 $N-1$ 作为分母，这是由于使用 $N-1$ 做分母可以保证样本方差的期望 $\mathbb{E}s^2 = \sigma^2$，

即如果假设 $x_i \sim \left(\mu, \sigma^2\right)$ $i.i.d$，那么：

$$\begin{aligned}
\mathbb{E}s^2 &= \frac{N}{N-1}\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2\right) \\
&= \frac{N}{N-1}\left[\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left(x_i^2\right) - \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}\left(\sum_{i=1}^{N}x_i^2 + 2\sum_{1\le i<j\le N}x_ix_j\right)\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N^2}\mathbb{E}\left(\sum_{i=1}^{N}x_i^2\right) - \frac{2}{N^2}\mathbb{E}\left(\sum_{1\le i<j\le N}x_ix_j\right)\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N}\left(\mu^2+\sigma^2\right) - \frac{2}{N^2}\frac{N^2-N}{2}\mu^2\right] \\
&= \frac{N}{N-1}\left[\mu^2+\sigma^2 - \frac{1}{N}\left(\mu^2+\sigma^2\right) - \frac{N-1}{N}\mu^2\right] \\
&= \sigma^2
\end{aligned}$$

而如果更进一步，假设 $x \sim N\left(\mu\iota, \sigma^2 I\right)$，由于 $M_0\mu\iota = \mu\left(I - \frac{1}{N}\iota\iota'\right)\iota = \mu\left(\iota - \frac{1}{N}\iota\iota'\iota\right) = 0$，那么

$$\frac{(N-1)\,s^2}{\sigma^2} = \frac{1}{\sigma^2}x'M_0x = \left(\frac{x-\mu}{\sigma}\right)'M_0\left(\frac{x-\mu}{\sigma}\right)$$

其中 $\frac{x-\mu}{\sigma} \sim N(0, I)$，且 M_0 为投影矩阵，$\mathrm{tr}\left(M_0\right) = \mathrm{tr}\left(I - \frac{1}{N}\iota\iota'\right) = \mathrm{tr}\left(I\right) - \frac{1}{N}\mathrm{tr}\left(\iota\iota'\right) = N - \frac{1}{N}\mathrm{tr}\left(\iota'\iota\right) = N-1$，因而可以得到：

$$\frac{(N-1)\,s^2}{\sigma^2} \sim \chi^2_{N-1}$$

即正态总体的样本方差标准化之后服从卡方分布，且自由度为 $N-1$。注意这一结论是在正态总体且独立同分布的假定下得到的。

如果我们将式 (1) 中的总体方差 σ^2 替换为样本方差 s^2，即：

$$\begin{aligned}\sqrt{N}\frac{\bar{x}-\mu}{s} &= \sqrt{N}\frac{\bar{x}-\mu}{\sqrt{s^2}} \\ &= \sqrt{N}\frac{L'x-\mu}{\sqrt{\frac{1}{N-1}x'M_0x}} \\ &= \sqrt{N}\frac{L'\left(x-\mu\right)}{\sqrt{\frac{1}{N-1}x'M_0x}} \\ &= \sqrt{N}\frac{L'\left(\frac{x-\mu}{\sigma}\right)}{\sqrt{\frac{1}{N-1}\frac{1}{\sigma^2}x'M_0x}}\end{aligned}$$

由于 $\frac{x-\mu}{\sigma} \sim N\left(0,I\right)$，因而 $\sqrt{N}L'\frac{x-\mu}{\sigma} \sim N\left(0,NL'L\right) = N\left(0,1\right)$，而分母上 $\frac{1}{\sigma^2}x'M_0x \sim \chi^2_{N-1}$，且 $LM = \frac{1}{N}\iota\left(I-\frac{1}{N}\iota\iota'\right) = 0$（分子与分母独立，即 $\bar{X}$ 与 s^2 是独立的[1]），因而：

$$\sqrt{N}\frac{\bar{x}-\mu}{s} = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}} \sim t_{N-1} \tag{2}$$

由于 t_{N-1} 分布在 $N\to\infty$ 时趋向于标准正态分布，因而当样本足够大时，上述分布趋向于正态分布。对于样本 $\{x_i\}$，如果我们对每个样本都减去样本均值，再除以样本标准差，即：

$$x_i^s = \frac{x_i-\bar{x}}{s}$$

我们称这个过程为**标准化**（**standardrize**），标准化之后的数据 $\{x_i^s\}$ 其样本均值为 0，而样本方差为 1。注意由于：

$$\frac{\bar{x}-\mu}{s} = \frac{1}{N}\sum_{i=1}^{N}\frac{x_i-\mu}{s} = \frac{1}{N}\sum_{i=1}^{N}x_i^s$$

因而式 (2) 表明，标准化之后的样本均值乘以 $\sqrt{N}$ 服从 t 分布，且其自由度为 $N-1$。仍然，这一结论只有在独立同分布的正态总体下才成立。

2.3 分位数与次序统计量

上一节中我们介绍了使用样本均值度量平均水平，使用样本方差度量数据的离散程度。而除了平均值，**中位数**（**median**）是平均水平的另外一种度量方法。

对于一个总体 P，如果其分布函数为 $F\left(x\right)$，那么中位数定义为 $F^{-1}\left(\frac{1}{2}\right)$。例如，由于对称性，正态分布 $N\left(\mu,\sigma^2\right)$ 的中位数为 μ。注意中位数是以下最小化问题的解：

$$\min_c \mathbb{E}\left|X-c\right| \tag{3}$$

[1] 注意这一结论的前提是总体为正态分布。

为证明以上结论，注意当以上目标函数取最小值时，一阶条件意味着：

$$\begin{aligned}
0 &= \frac{\partial \mathbb{E}\left|X-c\right|}{\partial c} \\
&= \frac{\partial \int_{\mathbb{R}} \left|x-c\right| dF\left(x\right)}{\partial c} \\
&= \int_{\mathbb{R}} \frac{\partial \left|x-c\right|}{\partial c} dF\left(x\right) \\
&= \int_{c}^{\infty} \left(-1\right) dF\left(x\right) + \int_{-\infty}^{c} 1 dF\left(x\right) \\
&= -\left[1-F\left(c\right)\right] + \left[F\left(c\right)-0\right] \\
&= -1+2F\left(c\right)
\end{aligned}$$

因而当 $c=F^{-1}\left(\frac{1}{2}\right)$ 时，上式成立。

对于一组样本 $\{x_1, x_2, ..., x_N\}$，记最小的样本为 $x_{(1)}$，第二小的样本为 $x_{(2)}$，以此类推，最大的样本记为 $x_{(N)}$。我们称 $x_{(n)}$ 为**次序统计量**（**order statistics**）。**样本中位数**（**sample median**，记为 M）定义为：

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n\text{为奇数} \\ \frac{\left(x_{\left(\frac{n}{2}\right)}+x_{\left(\frac{n}{2}+1\right)}\right)}{2} & n\text{为偶数} \end{cases}$$

注意实际上中位数是以下最小化问题的解：

$$\min_{c} \frac{1}{N} \sum_{i=1}^{N} \left|x_i - c\right|$$

以上最小化问题是式 (3) 的样本等价形式。

更进一步，我们还可以定义其他的**分位数**（**quantiles**）。q 分位数（q-Quantiles）即将实轴分为概率相等的 q 部分。$q-1$ 个分位数值将实轴分为 q 个概率相等的部分。比如，四分位数（quartiles，记为 Q），即 $Q_1 = F^{-1}\left(0.25\right), Q_2 = F^{-1}\left(0.5\right), Q_3 = F^{-1}\left(0.75\right)$。此外，百分位数（percentiles，记为 P），即 $P_p = F^{-1}\left(\frac{p}{100}\right), p=1,2,...,99$。

如果令：

$$\psi_q\left(x\right) = \begin{cases} qx & x>0 \\ \left(q-1\right)x & x\leq 0 \end{cases}$$

其中 $0<q<1$，那么可以证明，$F^{-1}\left(q\right)$ 是以下最小化问题的解：

$$\min_{c} \mathbb{E}\psi_q\left(X-c\right)$$

类似的，对于 $0<q<1$，令 $\{Nq\}$ 代表 Nq 的四舍五入，那么样本的分位

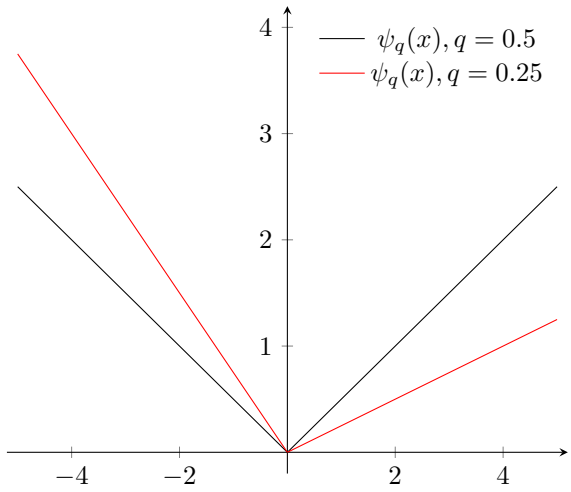


图 1: $\psi_q(x)$ 函数示意图

数即 $x_{(\{Nq\})}$。同样，样本分位数也是以下最小化问题的解：

$$\min_c \frac{1}{N}\sum_{i=1}^{N}\psi_q(x_i - c)$$

以上我们使用次序统计量 $x_{(n)}$ 定义了中位数、分位数，接下来我们讨论次序统计量的分布问题。对于次序统计量，我们有如下定理：

定理 1. 如果 $x_{(1)}, x_{(2)}, ..., x_{(n)}$ 为独立同分布随机样本 $\{X_i, 1 \le i \le N\}$ 的次序统计量，且总体分布函数为 $F(x)$，密度函数为 $f(x)$，那么次序统计量 $x_{(n)}$ 的密度函数为：

$$f_{x_{(n)}}(x) = \frac{N!}{(n-1)!\,(N-n)!} f(x) [F(x)]^{n-1} [1 - F(x)]^{N-n}$$

Proof. 可以首先计算 $x_{(n)}$ 的分布函数 $F_{x_{(n)}}(x) = P\left(x_{(n)} \le x\right)$，密度函数即其分布函数的导数。现在令 Y 为小于等于 x 的样本数，即 $Y = \#\{x_i \le x\}$，那么 $Y \sim Bi(N, F(x))$。而 $x_{(n)} \le x$ 等价于 $Y \ge n$，因而：

$$F_{x_{(n)}} = P(Y \ge n) = \sum_{k=n}^{N} \left(\begin{array}{c} N \\ k \end{array} \right) [F(x)]^k [1 - F(x)]^{N-k}$$

对以上分布函数求导可得结论。 □

此外，我们还可以计算不同次序统计量之间的联合分布。所有次序统计量

的联合密度函数为:

$$f_{x_{(1)}\ldots x_{(n)}}(x_1,...,x_n)=\begin{cases}N!\prod_{i=1}^{N}f(x_i) & -\infty<x_1<...<x_n<\infty\\ 0 & otherwise\end{cases}$$

其中 $N!$ 是由于对于一个随机样本，有 $N!$ 中情况可以得到 $-\infty<x_1<...<x_n<\infty$ 的实现。比如样本 $\{x_1=1,x_2=2\}$ 和样本 $\{x_1=2,x_2=1\}$ 都可以产生相同的次序统计量，因而有 2! 中情况。根据以上的密度函数，可以得到任意两个次序统计量 $\left(x_{(i)},x_{(j)}\right),1\leq i<j\leq N$ 的联合密度函数为:

$$f_{x_{(i)},x_{(j)}}(u,v)=\begin{cases}\frac{N![F(u)]^{i-1}[F(v)-F(u)]^{j-i-1}[1-F(v)]^{N-j}f(u)f(v)}{(i-1)!(j-i-1)!(N-j)!} & u<v\\ 0 & otherwise\end{cases}$$

3 描述性统计

正如前文所述，统计方法包括描述性统计和统计推断两部分。虽然统计推断是数理统计的核心，然而在数据分析之前，做好描述性统计是非常重要的。

描述性统计即使用图和表的形式对数据的分布特征进行度量。描述性统计可以帮助研究者初步掌握数据的分布情况，并发现数据中潜在的问题，比如异常值的存在、数据的不一致性等。此外，描述性统计给研究者提供了一些需要解释的现象，比如研究人员发现城市的大小、姓氏的分布等都服从幂律（Power law），该如何解释这些分布规律成为了一些学科的研究热点。最后，描述性统计的结果有利于对统计推断结果的理解，如回归分析中，回归结果经常需要与描述性统计相配合才能得到回归系数的影响大小。

下面我们分描述性统计量和统计图表两部分介绍描述性统计的初步知识。

3.1 描述性统计量

针对已有的数据，可以使用一些样本统计量对数据进行描述。而针对不同的数据类型，使用的描述性统计量也不尽相同，比如，对于分类数据，就不能计算其均值和中位数等。下面我们将分别对这些描述性统计量做简要介绍。

3.1.1 平均水平度量

常用的度量平均水平的描述性统计量有众数、中位数、平均数等。

1. **众数**（**mode**）：适用于分类数据和顺序数据，指样本数量最多的类别。如一个班中男生 30 人，女生 20 人，那么众数即男生。

2. **中位数**（**median**）：适用于顺序数据和数值型数据，但是不适用于分类数据。对于顺序数据，如一等奖 3 人、二等奖 5 人、三等奖 7 人，那么中位数即为二等奖。中位数具有不易受异常值影响的优点。此外，中位数对于

单调变换有不变性，比如一组数据 $\{x_i\}$ 的中位数为 M，那么经过单调变换之后的数据，比如 $\{\ln(x_i)\}$ 的中位数为 $\ln(M)$。

3. **平均数**（**mean**）：仅适用于数值型数据。与中位数相比，平均数比较容易受到异常值的影响，然而其具有非常良好的性质，因而是最常使用的平均水平的度量。

3.1.2 离散程度度量

常见的度量平均水平的描述性统计量有：异众比率、四分位差、极差、标准差、离散系数等。

1. **异众比率**（**variation ratio**）：适用于分类数据和顺序数据，指样本中非众数组的比例。

2. **四分位差**（**quartile deviation**）、**极差**（**range**）：适用于顺序数据和数值型数据，其中四分位差定义为 $Q_3 - Q_1$，而极差定义为 $x_{(N)} - x_{(1)}$。同样，这两个变量不太容易受到极端值的影响。

3. **标准差**（**standard deviation**）：适用于数值型数据，是离散程度的最常用的度量。

4. **离散系数**（**coefficient of variation**）：适用于数值型数据，定义为样本标准差与样本均值的比值：
$$v = \frac{s}{\bar{x}}$$
其优点是消除了单位以及平均水平大小的影响。

3.1.3 偏度系数与峰度系数

偏度系数度量了数据分布的对称性，而峰度系数则度量了分布的厚尾性。常用的偏度系数有：

1. **样本偏度系数**（**sample skewness**）：是偏度系数的最常用度量，定义为：
$$b_1 = \frac{N^2}{(N-1)(N-2)} \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{s^3}$$
当样本偏度系数大于 0 时为右偏，小于 0 时为左偏。

2. **非参数偏度**（**nonparametric skew**）：定义为：
$$b_2 = \frac{\bar{x} - M}{s}$$
即当样本均值大于中位数时为右偏，小于时为左偏。

当分布对称时，两个偏度系数都等于 0。注意两种定义下的偏度系数可能会出现符号相反的情况，即样本偏度系数大于 0 并不一定代表均值大于中位数。

而峰度系数度量了数据分布的厚尾特性（tailedness），其定义为：

$$k = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{s^4}$$

由于正态分布的峰度系数为 3，因而应用中经常将峰度系数减 3 处理，即定义：

$$k = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^4}{s^4} - 3$$

需要注意的是，虽然峰度系数中文名称似乎与分布的峰有关，然而其度量的是分布的尾巴的厚度。虽然一般情况下，厚尾伴随着尖峰，然而情况并不总是如此。

3.1.4 相关系数

我们经常会关心两组数据 $\{(x_i, y_i), i = 1, ..., N\}$ 的相关性，此时需要使用相关系数。常用的相关系数包括：

1. **Pearson 相关系数（Pearson correlation coefficient）**：是最常用的相关系数，简称样本相关系数（sample correlation coefficient），其计算公式为：
$$\rho = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{(N-1)\, s_x s_y}$$
注意样本相关系数只能度量变量之间的线性相关性。

2. **Spearman 秩相关系数（Spearman's rank correlation coefficient）**：即数据排序的相关系数，度量了两个变量单调变换的相关性。如果记 $r(x_i)$ 为 x_i 在样本中的排序，$r(y_i)$ 为 y_i 在样本中的排序，那么秩相关系数被定义为 $r(x_i)$ 与 $r(y_i)$ 的样本相关系数。

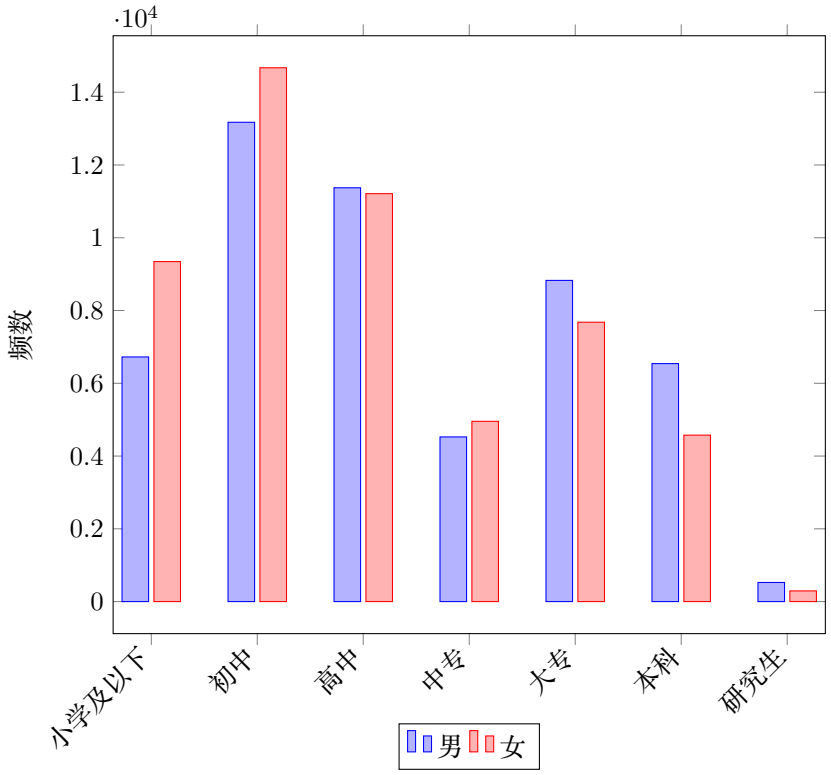
例 4. 一组数据：$\{(1,1), (2,4), (3,9)\}$，可以得到 $\bar{x} = 2, \bar{y} = \frac{14}{3}, s_x = 1, s_y = \frac{7}{\sqrt{3}}$，其样本相关系数为：

$$\rho = \frac{(1 + 8 + 27) - 3 \times 2 \times \frac{14}{3}}{2 \times 1 \times \frac{7}{\sqrt{3}}} \approx 0.9897$$

而两列数据的排序分别为：$\{(1,1), (2,2), (3,3)\}$，因而其秩相关系数为：

$$r = \frac{(1 + 4 + 9) - 3 \times 2 \times 2}{2 \times 1 \times 1} = 1$$

即两组数据存在着完全负向的单调关系，然而并不存在完全单调的负向线性关系。



数据来源：2009 年中国城镇住户调查

图 2: 条形图示例：我国人口教育程度分布

3.2 数据的图表展示

虽然以上描述性统计量从各个方面对数据进行了描述，然而这些统计量仍然不够直观。而图、表可以以更加直观的方式呈现数据。下面我们分别介绍一些简单的数据图表。

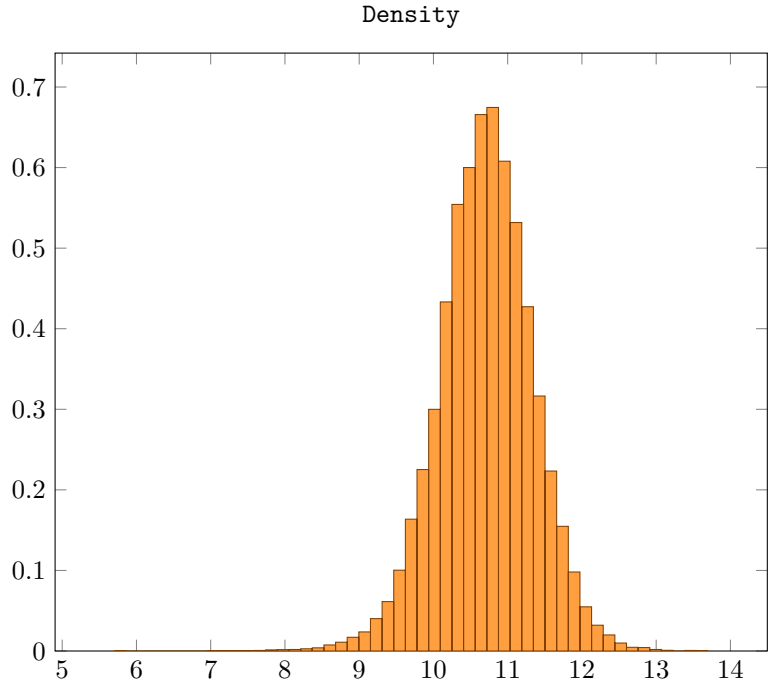
3.2.1 统计图

图形可以非常直观的展示数据的分布等情况。以下介绍几种最为常见的统计表：

1. 条形图（bar chart）：为一个二维图，其中 x 轴为分类数据或者顺序数据，纵轴可以为频数、频率或者其他数值型数据等。当纵轴为为频数时，条形图的高度表示频数，而条形图的宽度没有意义。

2. 饼图（pie chart）：使用圆形及扇形角度表示比例，一般用于展示分类数据的比例。

3. 直方图（histogram）：用来表示数值型数据密度分布的一种图，使用矩形**面积**代表落在某一区间内概率的大小，是数值型数据最常用的分布图形。注



数据来源：2009 年中国城镇住户调查

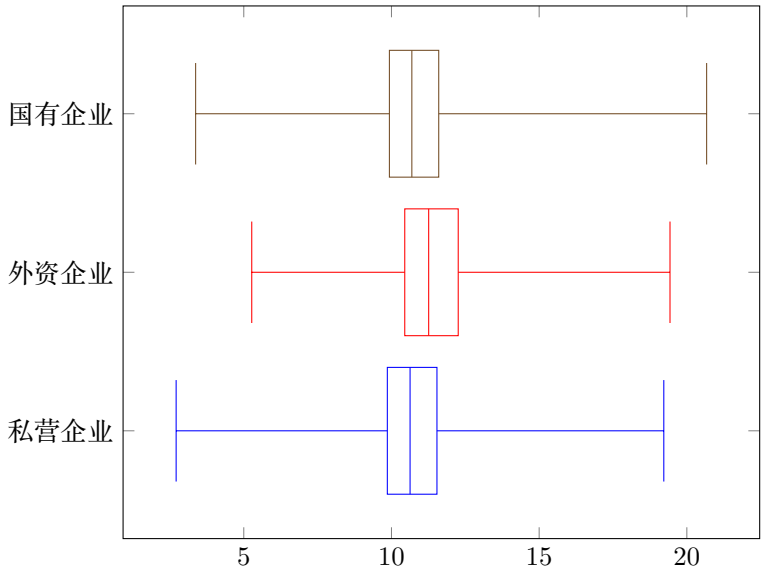
图 3: 直方图示例：我国家庭收入对数的分布

意直方图与条形图的差别，条形图针对分类数据，其宽度没有意义，只有长度有意义，而直方图每个柱形的宽度代表区间大小，其面积代表落入该区间的概率；条形图针对分类数据，因而不同类别之间矩形是分开的，而直方图代表的是连续的区间，因而矩形之间是紧密相连的。

4. 箱线图(box chart)：即将数据的最小值、最大值、中位数、四分位数(Q_1, Q_3)画在图中，两个四分位数作为「箱子」，并在中间表示中位数，再将最大值、最小值与箱子连接所做的图（如图 (4) 所示）。箱线图可以比较直观的观察数据的平均水平、离散程度以及偏态。

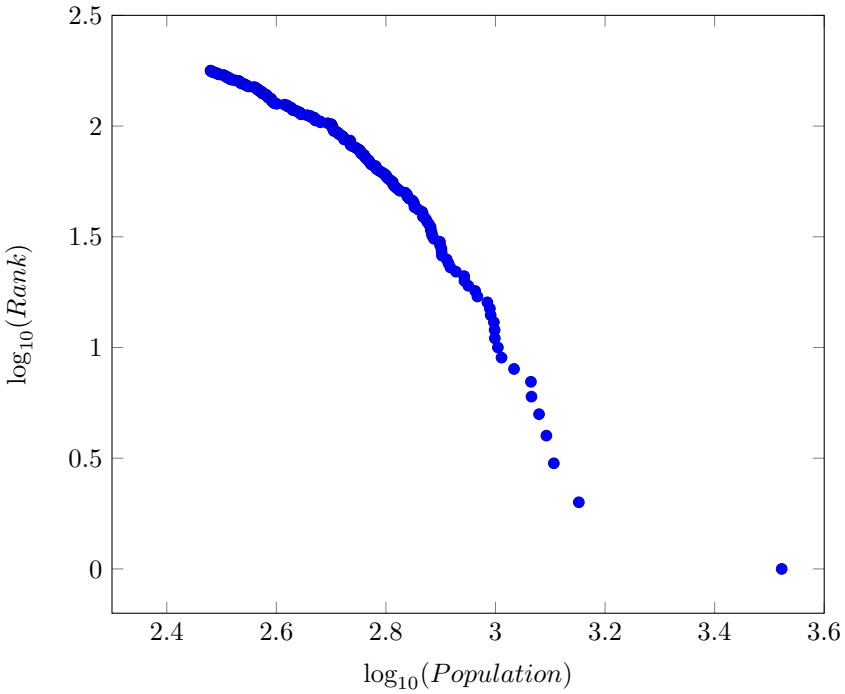
5. 散点图（scatter diagram）：一种二维图，x 轴与 y 轴分别表示一个变量，将数据的 $x-y$ 组合以散点的形式画在图上，表示两个变量之间的相关关系。例如，图 (5) 展示了 2011 年我国人口大于 300 万的城市的人口数与其人口数排名之间的关系，可以看到除了某些异常点之外，两个变量之间有着近似的线性关系。此外，散点的颜色、大小等也可以表示第三维数据，如气泡图等。

6. 线图（line plot）：横轴为某一序列（如时间），纵轴为数值型数据，一般用来描述时间序列数据。



数据来源：2013 年中国工业企业数据库

图 4: 箱线图示例：分所有制企业规模（ln (总资产)）



数据来源：2011 年《中国城市统计年鉴》

图 5: 散点图示例：我国城市人口排名与人口数的关系

表 1: 描述性统计表示例

变量	(1) 样本量	(2) 均值	(3) 标准差	(4) 最小值	(5) 最大值
总人口	286	439.5	312.0	19.50	3,330
人均生产总值	286	38,812	24,247	6,457	163,014
第一产业比重	286	13.06	8.130	0.0600	48.64
第二产业比重	286	51.96	10.49	17.02	89.34
第三产业比重	286	34.98	9.056	10.15	76.07

注：数据来源：2011 年《中国城市统计年鉴》

在画图时需要特别注意图的坐标轴。例如，为了便于比较，一般条形图、直方图、线图的横轴应该以 0 开始，不以 0 为起始的图经常给人以误导。

除了以上介绍的这些图之外，还有很多其他类型的图，在这里我们不一一赘述。

3.2.2 统计表

除了图之外，统计表格也是经常使用的描述性统计工具。一般来说，一张统计表应该包含表头、行标题、列标题、数值、附注等部分。此外，在格式上，统计表格一般要遵循一定的规范，如表格除了上下两条横线用粗线之外，其他线一般用细线；统计表格一般两边不封口。表 (1) 展示了一张典型的描述性统计表格。

4 充分统计量

以上我们介绍了统计量的概念。我们知道，在参数模型中，我们假设总体 P 属于某一个参数族 $\mathcal{P}=\{P_\theta:\theta\in\Theta\}$，其中 $\theta\in\Theta$ 为参数。在获得了一个样本 $x=(x_1,...,x_N)$ 之后，我们通常会使用一些统计量 $T(x)$ 来描述总体 P 的分布，一般而言，这些统计量的个数小于样本数量。一个很自然的问题是，这些统计量 $T(x)$ 在何种程度上代表了样本 x 所包含的信息呢？是不是存在有限个统计量使得这些统计量能够完全代表样本的信息呢？这里我们需要所谓的**充分统计量**（**Sufficient statistics**）的概念。

定义 1. （充分统计量）若 $x=(x_1,...,x_N)$ 为来自于未知总体 $P\in\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ 的一组样本，如果给定一组统计量 $T(x)$，样本的条件分布 $f(x|T(x)=t)$ 不依赖于 θ，那么我们称 $T(x)$ 为充分统计量。

以上定义意味着，如果我们计算得到了充分统计量 $T(x)$，那么样本 x 不包含除了 $T(x)$ 之外的关于 θ 的任何信息，或者说，充分统计量 $T(x)$ 包含了使用样本 x 对总体 P（或者等价的，θ）进行推断所需要的所有信息。这也就意味着，如果两个样本 x^1 和 x^2，有 $T\left(x^1\right)=T\left(x^2\right)$，那么我们对于 θ 的所有推断应该是等价的，而不管 $x^1=x^2$ 是否成立。

例 5. 假设 $x_i \sim Ber(p)$ $i.i.d$，$x = (x_1, ..., x_N)$，那么样本的联合分布为：

$$f(x = \tilde{x}) = \prod_{i=1}^{N} p^{\tilde{x}_i} (1-p)^{1-\tilde{x}_i} = p^{\sum_{i=1}^{N} \tilde{x}_i} (1-p)^{N-\sum_{i=1}^{N} \tilde{x}_i} \tag{4}$$

如果记统计量 $N_1(x) = \sum_{i=1}^{N} x_i$，那么 $N_1(x) \sim Bi(N, p)$。由于 $f(x|N_1(x)) = \frac{f(x, N_1(x))}{f_{N_1}(N_1(x))}$，因而我们需要计算 $f(x, N_1(x))$ 的联合分布。可得：

$$f(x = \tilde{x}, N_1(x) = n) = \begin{cases} p^n (1-p)^{N-n} & if \ \sum_{i=1}^{N} \tilde{x}_i = n \\ 0 & if \ \sum_{i=1}^{N} \tilde{x}_i \neq n \end{cases}$$

因而：

$$f(x|N_1(x) = n) = \begin{cases} \frac{p^n (1-p)^{N-n}}{\begin{pmatrix} N \\ n \end{pmatrix} p^n (1-p)^{N-n}} & if \ \sum_{i=1}^{N} x_i = n \\ 0 & if \ \sum_{i=1}^{N} x_i \neq n \end{cases}$$

$$= \begin{cases} \frac{1}{\begin{pmatrix} N \\ n \end{pmatrix}} & if \ \sum_{i=1}^{N} x_i = n \\ 0 & if \ \sum_{i=1}^{N} x_i \neq n \end{cases}$$

注意以上条件分布并不依赖于未知参数 p，因而 $N_1(x)$ 是 p 的一个充分统计量。

以上我们通过观察猜测的方式找到了伯努利总体的充分统计量，然而该过程比较繁琐。实际上，在寻找充分统计量时，我们有如下简单的定理可以使用：

定理 2. （因子分解定理）若 $x = (x_1, ..., x_N)$ 为来自于未知总体 $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 的一组样本，令 $f(x|\theta)$ 样本的联合概率密度函数。统计量 $T(x)$ 为充分统计量的充要条件是存在函数 $g(t|\theta)$ 和 $h(x)$，使得：

$$f(x|\theta) = g(T(x)|\theta) \cdot h(x)$$

例如，通过观察式 (4) 可以发现，$f(x|p) = p^{N_1(x)} (1-p)^{N-N_1(x)}$，因而 $N_1(x)$ 是其充分统计量。

例 6. 假设 $x_i \sim U(0,\theta)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{N}\left[\frac{1}{\theta}1_{(0,\theta)}(x_i)\right] \\ &= \frac{1}{\theta^N}\prod_{i=1}^{N}1_{(0,\theta)}(x_i) \\ &= \frac{1}{\theta^N}1\left\{\max_i x_i \le \theta\right\} \end{aligned}$$

因而 $T(x)=\max_i\{x_i\}$ 即其充分统计量。

例 7. 若 $x_i \sim P(\lambda)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\lambda) &= \prod_{i=1}^{N}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \\ &= e^{-N\lambda}\lambda^{\sum_{i=1}^{N}x_i}\prod_{i=1}^{N}\frac{1}{x_i!} \end{aligned}$$

其中可以令 $h(x)=\prod_{i=1}^{N}\frac{1}{x_i!}$，$T(x)=\frac{1}{N}\sum_{i=1}^{N}x_i$，$g(T(x)|\lambda)=e^{-N\lambda}\lambda^{NT(x)}$，因而 $T(x)=\frac{1}{N}\sum_{i=1}^{N}x_i$ 是泊松分布的充分统计量。

例 8. 假设 $x_i \sim N\left(\mu,\sigma^2\right)$ $i.i.d$，那么样本 x 的联合密度函数为：

$$\begin{aligned} f(x|\theta) &= \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2+\mu^2-2\mu x_i\right)\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{N}x_i^2+N\mu^2-2\mu\sum_{i=1}^{N}x_i\right)\right\} \\ &= (2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{x^2}+\mu^2-2\mu\bar{x}\right)\right\} \end{aligned}$$

因而 $T(x)=\left(\bar{x},\overline{x^2}\right)'$ 是正态分布的充分统计量。由于 $s^2=\frac{N}{N-1}\left(\overline{x^2}-\bar{x}^2\right)$，因而 $T'(x)=\left(\bar{x},s^2\right)$ 也是正态分布的充分统计量。

回忆指数分布族的定义，我们发现以上两例中分布都属于指数分布族。实际上，根据指数分布族的定义：

$$f(x_i|\theta)=h(x_i)\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot T_k(x_i)\right]-B(\theta)\right\}$$

那么在独立同分布的假定下，样本 $x=(x_1,...,x_N)$ 的联合分布为：

$$f(x|\theta)=\left[\prod_{i=1}^{N}h(x_i)\right]\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot\sum_{i=1}^{N}T_k(x_i)\right]-NB(\theta)\right\}$$

因而 $T(x)=\left[\sum_{i=1}^{N}T_1(x_i),...,\sum_{i=1}^{N}T_K(x_i)\right]$ 为其充分统计量。

例 9. 对于伯努利分布，其分布函数：

$$\begin{aligned}f(x_i|\lambda)&=\exp\{x_i\ln p+(1-x_i)\ln(1-p)\}\\&=\exp\left\{x_i\ln\frac{p}{1-p}+\ln(1-p)\right\}\end{aligned}$$

因而 $T(x)=\sum_{i=1}^{N}x_i$ 为其充分统计量。

进一步，在得到充分统计量之后，我们会关心充分统计量的分布。对于指数分布族，我们有如下定理：

定理 3. 若 $x=(x_1,...,x_N)$ 为来自于未知总体 $P\in\mathcal{P}=\{P_\theta:\theta\in\Theta\}$ 的一组独立同分布的样本，若 $\{P_\theta:\theta\in\Theta\}$ 为指数分布族，即其密度函数为：

$$f(x_i|\theta)=h(x_i)\cdot\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot T_k(x_i)\right]-B(\theta)\right\}$$

那么 $T(x)=\left[\sum_{i=1}^{N}T_1(x_i),...,\sum_{i=1}^{N}T_K(x_i)\right]$ 为其充分统计量。若集合 $\{(\eta_1(\theta),...,\eta_K(\theta)),\theta\in\Theta\}$ 包含了 $\mathbb{R}^K$ 中的一个开集，那么 $T(x)$ 的联合密度函数同样属于指数分布族，且具有如下形式：

$$f_T(t_1,...,t_K|\theta)=H(t_1,...,t_K)\exp\left\{\sum_{k=1}^{K}\left[\eta_k(\theta)\cdot t_k\right]-NB(\theta)\right\}$$

注意集合 $\{(\eta_1(\theta),...,\eta_K(\theta)),\theta\in\Theta\}$ 包含了 $\mathbb{R}^K$ 中的一个开集的假设排除了诸如 $N\left(\mu,\mu^2\right)$ 这样的指数分布族。

例 10. 例 (9) 中，我们得到伯努利分布的一个充分统计量为 $T(x)=\sum_{i=1}^{N}x_i$。根据上述定理，充分统计量 $T(x)$ 的密度函数为：

$$\begin{aligned}f_T(t)&=H(t)\exp\left\{t\ln\frac{p}{1-p}+N\ln(1-p)\right\}\\&=H(t)\exp\{t\ln p-t\ln(1-p)+N\ln(1-p)\}\\&=H(t)\exp\{t\ln p+(N-t)\ln(1-p)\}\\&=H(t)\,p^t(1-p)^{N-t}\end{aligned}$$

实际上，我们知道 $T(x)$ 服从二项分布，即 $T(x) \sim Bi(N,p)$，可以验证，以上的形式当 $H(t) = \begin{pmatrix} N \\ t \end{pmatrix}$ 时，上述密度函数即得到了二项分布，验证了以上定理。

实际上，对于某一个总体，可能有不止一组充分统计量。比如，次序统计量 $T(x) = \left(x_{(1)}, ..., x_{(N)}\right)$ 包含了样本所有的信息，因而一定是充分统计量，然而这样的充分统计量并没有达到数据压缩的目的。因而自然的想法是，在所有的充分统计量中，我们是不是可以找到一组最少的充分统计量。为此，我们定义**最小充分统计量**（**minimal sufficient statistics**）的概念：

定义 2. 一个充分统计量 $T(x)$，如果对于任何其他的充分统计量 $T'(x)$，$T(x)$ 都是 $T'(x)$ 的函数，那么我们称 $T(x)$ 为一个最小充分统计量。

以上定义意味着，$T(x)$ 本身就是充分统计量，而且 $T(x)$ 可以由其他的任何充分统计量 $T'(x)$ 计算得到，因而从这个意义上说，$T(x)$ 比 $T'(x)$ 要「小」。例如，对于泊松分布，我们已经找到其一个充分统计量为 $T(x) = \sum_{i=1}^{N} x_i$，而次序统计量也是其充分统计量，我们可以使用次序统计量计算出 $T(x)$，但是不能通过 $T(x)$ 计算出次序统计量，因而 $T(x)$ 是比次序统计量要「小」的充分统计量。

以下定理可以帮助我们寻找最小充分统计量：

定理 4. 对于任意的两组来自于未知总体 $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$ 的样本 $x = (x_1, ..., x_N)$，$y = (y_1, ..., y_N)$，如果存在统计量 $T(x)$，使得当两个样本联合密度函数可以写为 $f(x|\theta) = f(y|\theta)\,\phi(x,y)$ 的形式时，必然有 $T(x) = T(y)$，那么 $T(x)$ 为 θ 的最小充分统计量。

例 11. 假设 $x_i \sim N\left(\mu, \sigma^2\right)$ $i.i.d$，由于密度函数在任何一个点处都大于 0，那么任意两个样本 x 和 y 联合密度函数的比例 $\frac{f(x|\theta)}{f(y|\theta)}$ 为：

$$\begin{aligned}\frac{f(x|\theta)}{f(y|\theta)} &= \frac{(2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{x^2} + \mu^2 - 2\mu\bar{x}\right)\right\}}{(2\pi)^{-\frac{N}{2}}\sigma^{-N}\exp\left\{-\frac{N}{2\sigma^2}\left(\overline{y^2} + \mu^2 - 2\mu\bar{y}\right)\right\}} \\ &= \exp\left\{-\frac{N}{2\sigma^2}\left[\left(\overline{x^2} - \overline{y^2}\right) - 2\mu\left(\bar{x} - \bar{y}\right)\right]\right\}\end{aligned}$$

如果上述比例与 θ 无关则必然有 $\overline{x^2} = \overline{y^2}, \bar{x} = \bar{y}$，所以 $T(x) = \left(\bar{x}, \overline{x^2}\right)$ 为正态分布的最小充分统计量。

实际上，最小充分统计量并非只有一组，比如由于 $s^2 = \frac{N}{N-1}\left(\overline{x^2} - \bar{x}^2\right)$，因而 $\left(\bar{x}, s^2\right)$ 是 $\left(\bar{x}, \overline{x^2}\right)$ 的函数，同时反过来 $\left(\bar{x}, \overline{x^2}\right)$ 也是 $\left(\bar{x}, s^2\right)$ 的函数，所以 $\left(\bar{x}, s^2\right)$ 也是正态分布的最小充分统计量。实际上，如果找到了一组最小充分统计量，那么这组最小充分统计量的所有一一映射都是最小充分统计量。

习题

练习 1. 等式 $\mathbb{E}s = \sigma$ 是否成立？如果成立，请证明，如果不成立，请指出其大小关系。

练习 2. 证明 $F^{-1}(q)$ 是以下最小化问题的解：

$$\min_c \mathbb{E}\psi_q(X - c)$$

练习 3. 求以下分布的充分统计量：

1. 泊松分布
2. 指数分布
3. 正态分布
4. Beta 分布

参考文献

[1] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[2] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

第六节 · 大样本理论

司继春

上海对外经贸大学统计与信息学院

大样本理论（Large sample theory）在现代统计理论中占据着核心位置。一般而言，大样本理论告诉我们当样本容量趋向于无穷时统计量的分布特征，而当数据足够多时，统计量的分布特征与这个极限特征仅有很小的误差。由于有限样本的统计性质通常难以计算，因而多数时候我们需要使用大样本理论对有限样本进行近似计算。下面我们从最简单的数列极限开始入手，进而介绍常用的大样本理论。

1 收敛的概念

首先我们先回顾一下数列极限的概念：

定义 1. 若 $\{a_n, n = 1, 2, ...\}$ 为实数序列，如果对于任意的 $\epsilon > 0$，存在 $n_0 = n_0(\epsilon)$ 使得：

$$|a_n - a| < \epsilon, \forall n > n_0$$

那么我们称数列 $\{a_n\}$ 的极限为 a，或者 $\{a_n\}$ **收敛到**（**converges to**）a，记为

$$\lim_{n\to\infty} a_n = a$$

或者 $a_n \to a \; as \; n \to \infty$。

数列极限有一些简单的性质，比如，如果 $a_n \to a, b_n \to b$，那么 $(c \cdot a_n + d \cdot b_n) \to ca + db$，$a_n \cdot b_n \to ab$，如果 $b \neq 0$，那么 $\frac{a_n}{b_n} \to \frac{a}{b}$。

如一个经常使用的数列极限：

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n = e^c$$

而：

$$\lim_{n\to\infty} \frac{\left(1 + \frac{c}{n}\right)^n}{\left(1 + \frac{d}{n}\right)^n} = e^{c-d}$$

此外，数列可能随着 $n \to \infty$ 时，并不趋向于一个常数，而是趋向于 ∞。更严谨的，我们可以定义 $a_n \to \infty$ 如下：

表 1: $\frac{1}{n}$ 和 $\frac{1}{n^2}$ 的收敛速度

n	1	2	5	10	100	1000	10000
$\frac{1}{n}$	1	0.5	0.2	0.1	1×10^{-2}	1×10^{-3}	1×10^{-4}
$\frac{1}{n^2}$	1	0.25	0.04	0.01	1×10^{-4}	1×10^{-6}	1×10^{-8}

定义 2. 若 $\{a_n, n=1,2,...\}$ 为实数序列，如果对于任意的 M，存在一个 $n_0 = n_0(M)$，使得：

$$a_n > M, \forall n > n_0$$

那么我们称 $\{a_n\}$ 趋向于 ∞，记为

$$\lim_{n\to\infty} a_n = \infty$$

或者 $a_n \to \infty \; as \; n \to \infty$。

如经常遇到的数列：$\ln n, n^k, e^n$ 等都趋向于正无穷。相反，如果数列不趋向于无穷，那么我们称其为有界的。其定义如下：

定义 3. 若 $\{a_n, n=1,2,...\}$ 为实数序列，如果存在常数 $b<\infty$，使得 $|a_n|<b$，那么我们称数列 $\{a_n\}$ 为**有界**的（**bounded**），否则称之为**无界**的（**unbounded**）。

显然，如果一个数列收敛到一个常数，那么其必然是有界的，而一个趋向于 ∞ 的数列必定是无界的。而反过来则不成立，一个有界的数列并不一定是收敛的，例如 $a_n = (-1)^n$，虽然是有界的，但其极限并不存在。同时，无界的序列也不一定趋向于 ∞，例如 $a_n = n\cdot[1+(-1)^n]$，当 n 为奇数时 $a_n = 0$，因而这个数列并不趋向于 ∞。

对于两个序列 $\{a_n\}, \{b_n\}$，我们经常会关心两个序列收敛的速度问题。比如，如果令 $a_n = \frac{1}{n^2}, b_n = \frac{1}{n}$，我们有 $a_n \to 0, b_n \to 0$，然而两个序列收敛到 0 的速度是不一样的。表 (1) 列出了随着 n 的增大，两个序列趋向于 0 的速度，可以看到 $\frac{1}{n^2}$ 比 $\frac{1}{n}$ 以更快的速度趋向于 0。

一般的，为了比较两个序列收敛速度的问题，我们做如下定义：

定义 4. 对于两个序列 $\{a_n\}, \{b_n\}$，如果随着 $n\to\infty$，有：

$$\frac{a_n}{b_n} \to 0$$

那么我们记为 $a_n = o(b_n)$。特别的，如果令 $b_n = 1$，那么 $a_n = o(1)$ 等价为 $a_n \to 0$。

如在上例中，

$$\frac{a_n}{b_n} = \frac{\frac{1}{n^2}}{\frac{1}{n}} = \frac{1}{n} \to 0$$

因而 $\frac{1}{n^2} = o\left(\frac{1}{n}\right)$，即 $\frac{1}{n^2}$ 以更快的速度收敛到 0。如果两个序列 $a_n \to 0, b_n \to 0$，且 $a_n = o(b_n)$，那么我们称 a_n 为比 b_n 高阶的无穷小量。

假设有两个数列，

$$a_n = \frac{1}{n} + \frac{6}{n^2} - \frac{8}{n^3}$$

而另外一个序列：

$$b_n = \frac{1}{n}$$

如果定义 $R_n = \frac{6}{n^2} - \frac{8}{n^3}$，显然 $R_n = o\left(\frac{1}{n}\right)$，因而 $a_n = b_n + o\left(\frac{1}{n}\right)$，即：

$$\frac{a_n}{b_n} = \frac{b_n + o\left(\frac{1}{n}\right)}{b_n} \to 1$$

因而尽管两个序列 a_n 和 b_n 并不相等，但是当 $n \to \infty$ 时，两者误差趋向于 0，因而我们可以舍去无穷小量 R_n，使用更简单的序列 b_n 去逼近 a_n。

一般的，如果两个序列 a_n, b_n 随着 $n \to \infty$ 满足：

$$\frac{a_n}{b_n} \to 1$$

那么我们称这两个序列是**渐进等价**（**asymptotically equivalent**）的，记为 $a_n \sim b_n$。渐进意味着随着 $n \to \infty$，而等价意味着两个序列的误差很小。实际上，如果我们把相对误差写为：

$$\left|\frac{a_n - b_n}{b_n}\right| = \left|\frac{a_n}{b_n} - 1\right|$$

那么渐进等价意味着随着 n 的增大，相对误差趋向于 0。实际上，如果 $a_n = o(b_n)$，那么 $b_n + a_n = b_n + o(b_n) \sim b_n$，即一个序列加上这个序列的无穷小量，渐进等价于这个序列本身。

当然，由于 $a_n = \frac{1}{n} + \frac{6}{n^2} + o\left(\frac{1}{n^2}\right)$，我们也可以使用 $\frac{1}{n} + \frac{6}{n^2}$ 作为 a_n 的更加精确的逼近。

这种逼近比较常用的即**泰勒级数**（**Taylor series**）。当 $x \to a$ 时，$(x-a) = o(1)$，同时我们有 $(x-a)^{k+1} = o\left((x-a)^k\right)$，即当 $x \to a$ 时，$(x-a)$ 的高阶幂是低阶幂的无穷小量。对于一个单变量实值函数 $f(x)$ 且 k 阶可微，那么有：

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \frac{f^{(k)}(a)}{k!}(x-a)^k + o\left(|x-a|^k\right)$$

因而对于一个难以计算的函数 f，我们经常使用其前 k 阶泰勒多项式对其进行逼近。

例 1. 函数 $f(x) = \ln(1+x)$ 在 $x = 0$ 处的泰勒展开为：

$$f(x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots$$

因而当 x 充分靠近 0 时，我们可以使用前 k 阶泰勒展开对其进行逼近。特别的，

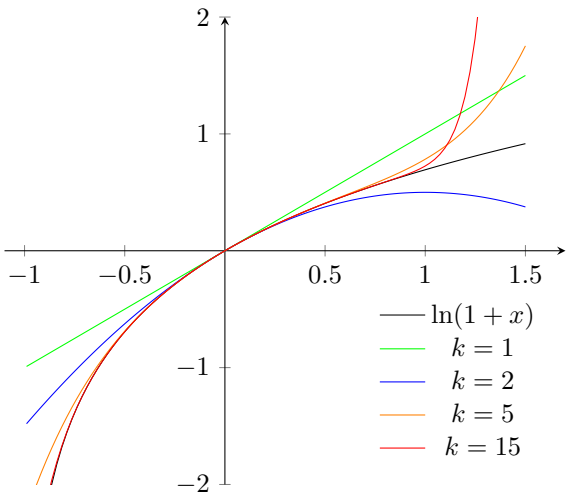


图 1: $\ln(1+x)$ 的泰勒展开

如果令 $k=1$，$\ln(1+x)=x+o(x)\approx x$。图 (1) 展示了使用不同阶数的多项式逼近 $\ln(1+x)$ 的结果。

值得注意的是，在图 (1) 中，只有当 x 在 $(-1,1)$ 区间之内，随着 k 的增加多项式逐渐逼近 $\ln(1+x)$。注意我们使用泰勒级数进行逼近的前提条件是 x 充分的接近于 a，因而如果 x 不在泰勒级数的收敛半径之内，泰勒级数不能用于逼近原始函数。

更一般的，如果 $f:\mathbb{R}^n\to\mathbb{R}$ 为多元实值函数，那么其泰勒级数为:

$$f(x)=f(a)+\frac{\partial f}{\partial x'}(a)(x-a)+\frac{1}{2!}(x-a)'\frac{\partial^2 f}{\partial x\partial x'}(a)(x-a)+o\left(\|x-a\|^2\right)$$

其中 x 和 a 为 $n\times 1$ 向量。

例 2. 令 $f(x)=e^{x_1}\ln(1+x_2)$，其中 $x=(x_1,x_2)'$。那么:

$$\frac{\partial f}{\partial x}=\begin{bmatrix} e^{x_1}\ln(1+x_2) \\ \frac{e^{x_1}}{1+x_2}\end{bmatrix},\frac{\partial^2 f}{\partial x\partial x'}=\begin{bmatrix} e^{x_1}\ln(1+x_2) & \frac{e^{x_1}}{1+x_2} \\ \frac{e^{x_1}}{1+x_2} & -\frac{e^{x_1}}{(1+x_2)^2}\end{bmatrix}$$

那么其在 $a=(0,0)'$ 处的二阶泰勒展开:

$$\begin{aligned} p_2(x)&=[0,1]\begin{bmatrix}x_1\\x_2\end{bmatrix}+\frac{1}{2}[x_1,x_2]\begin{bmatrix}0&1\\1&-1\end{bmatrix}\begin{bmatrix}x_1\\x_2\end{bmatrix} \\ &=x_2+\frac{1}{2}\left(2x_1x_2-x_2^2\right)\end{aligned}$$

如果令 $a_n = n^\alpha$，$b_n = n^\beta$，其中 α, β 为常数，那么：

$$\frac{a_n}{b_n} = \frac{n^\alpha}{n^\beta} = \frac{1}{n^{\beta-\alpha}}$$

当 $\beta > \alpha$ 时，以上极限趋向于 0，即 $a_n = o(b_n)$，a_n 相比于 b_n 以更慢的速度趋向于 ∞。

例 3. （指数增长）对于任意的 $k > 0$ 以及 $a > 1$，有 $n^k = o(a^n)$。为了证明，取 $c = a - 1 > 0$。考虑 $k = 2$ 的情形，欲证明

$$\frac{n^k}{a^n} \to 0$$

等价于证明

$$\frac{a^n}{n^k} \to \infty$$

由于：

$$a^n = (1+c)^n = 1 + nc + \begin{pmatrix} n \\ 2 \end{pmatrix} c^2 + \begin{pmatrix} n \\ 3 \end{pmatrix} c^3 + \cdots > \begin{pmatrix} n \\ 3 \end{pmatrix} c^3$$

因而

$$\frac{a^n}{n^2} > \frac{\begin{pmatrix} n \\ 3 \end{pmatrix} c^3}{n^2} = \frac{c^3}{6}\left(n - 3 + \frac{2}{n}\right) \to \infty$$

对于其他整数 k 可以类似证明。如果 k 不为整数，则取 k' 为比 k 大的最小整数进行证明。

类似的，还可以证明 $n^k = o(\log n)$。这三个序列，$a^n, n^k, \log n$ 是经常使用的三种增长速度，三者趋向于无穷的增长速度逐渐递减，分别代表着快速、适当和慢速的增长。相应的，其倒数，$a^{-n}, n^{-k}, 1/\log n$ 趋向于 0 的速度逐渐递减。

与无穷小的逼近类似，如果 $a_n = a^n + n^k$，$b_n = a^n$，那么 $a_n = b_n + o(b_n)$，因而我们同样可以使用 b_n 对 a_n 进行近似逼近。

对于小 o 符号，有如下性质：

定理 1. （小 o 的性质）

1. 若 $a_n = o(b_n), b_n = o(c_n)$，那么 $a_n = o(c_n)$
2. 对于任意的常数 $c \neq 0$，及 $a_n = o(b_n)$，有 $ca_n = o(b_n)$
3. 对于任意的数列 $c_n \neq 0$，及 $a_n = o(b_n)$，有 $c_n a_n = o(c_n b_n)$
4. 如果 $d_n = o(b_n), e_n = o(c_n)$，那么 $d_n e_n = o(b_n c_n)$
5. 如果 $a_n, b_n > 0, c_n, d_n > 0$，$a_n = o(b_n), c_n = o(d_n)$，那么 $a_n + c_n = o(b_n + d_n)$。

与小 o 符号相对应，我们还可以定义大 O 符号：

定义 5. 对于两个序列 $\{a_n\},\{b_n\}$，如果随着 $n\to\infty$，$\left|\frac{a_n}{b_n}\right|$ 是有界的，即存在一个 M 使得：

$$\left|\frac{a_n}{b_n}\right|<M$$

那么我们记为 $a_n=O(b_n)$。特别的，如果令 $b_n=1$，那么 $a_n=O(1)$ 等价为 a_n 是有界的。

根据以上定义，如果 $a_n=o(b_n)$，那么必然有 $a_n=O(b_n)$。进而，我们可以使用大 O 符号定义序列同阶。

定义 6. 对于两个序列 $\{a_n\},\{b_n\}$，如果 $a_n=O(b_n)$，且同时 $b_n=O(a_n)$，那么我们称两个序列是同阶的，简记为 $a_n\asymp b_n$。

下面的例子展示了 $\sim,\asymp,o,O$ 的区别：

例 4. 对于序列 $a_n=\frac{1}{n}+\frac{b}{n\sqrt{n}}+\frac{c}{n^2}+\frac{d}{n^2\sqrt{n}}$，同时定义 $R_n=\frac{b}{n\sqrt{n}}+\frac{c}{n^2}+\frac{d}{n^2\sqrt{n}}$ 那么：

1. $a_n\sim\frac{1}{n}$
2. 若 $b=0$，$R_n=O\left(\frac{1}{n^2}\right)$
3. 若 $b=0$，$R_n\asymp\frac{1}{n^2}$
4. 若 $b\neq0$，$R_n\sim\frac{b}{n\sqrt{n}}$
5. 若 $b=c=0$，$R_n=o\left(\frac{1}{n^2}\right)$

大 O 符号有如下性质：

定理 2. （大 O 的性质）

1. 若 $a_n=O(b_n),b_n=O(c_n)$，那么 $a_n=O(c_n)$

2. 对于任意的常数 $c\neq0$，及 $a_n=O(b_n)$，有 $ca_n=O(b_n)$

3. 对于任意的数列 $c_n\neq0$，及 $a_n=O(b_n)$，有 $c_na_n=O(c_nb_n)$

4. 如果 $d_n=O(b_n),e_n=O(c_n)$，那么 $d_ne_n=O(b_nc_n)$

5. 如果 $a_n=o(b_n)$，$c_n=O(b_n)$，那么 $a_nc_n=o(b_n)$

6. 如果 $a_n=o(b_n)$，$c_n=O(b_n)$，那么 $a_n+c_n=O(b_n)$

例如，如果 $a_n=\frac{1}{n}+\frac{b}{n\sqrt{n}}+\frac{c}{n^2}+\frac{d}{n^2\sqrt{n}}=O\left(\frac{1}{n}\right)$，那么根据性质 (3)，$n\cdot a_n=\frac{b}{\sqrt{n}}+\frac{c}{n}+\frac{d}{n\sqrt{n}}=O\left(n\cdot\frac{1}{n}\right)=O(1)$。

例 5. 如果令 $a_n = c \cdot (nh)^{-1}, b_n = g \cdot h^4$，其中 c, g 为非零常数，$h = n^q, q < 0$，那么 $a_n = O\left((nh)^{-1}\right) = O\left(n^{-q-1}\right)$，$b_n = O\left(h^4\right) = O\left(n^{4q}\right)$，所以 $a_n + b_n = O\left(n^{-q-1} + n^{4q}\right)$。例如，当 $q = -\frac{1}{2}$ 时，$a_n + b_n = O\left(n^{-\frac{1}{2}} + n^{-2}\right)$，由于 $n^{-2} = o\left(n^{-\frac{1}{2}}\right)$，因而 $a_n + b_n = O\left(n^{-\frac{1}{2}}\right)$。类似的，当 $\frac{n^{-q-1}}{n^{4q}} = n^{-1-5q} \to 0$，即 $-1-5q < 0$ 时，$n^{-q-1} = o\left(n^{4q}\right)$，$a_n + b_n = O\left(n^{4q}\right)$；当 $-1-5q > 0$ 时，$n^{4q} = o\left(n^{-q-1}\right)$，$a_n + b_n = O\left(n^{-q-1}\right)$；当 $q = -\frac{1}{5}$ 时，$a_n + b_n = O\left(n^{-\frac{4}{5}}\right)$。因而可以证明，当 $q = -\frac{1}{5}$ 时，使得 $a_n + b_n$ 以最快的速度趋向于 0。

2 概率收敛的概念

上面讨论了数列收敛的概念。在概率统计中，最经常使用的工具是随机变量，因而在这里我们将讨论随机变量的收敛。在这里我们将主要介绍四种收敛的概念：**几乎必然收敛**（**almost sure convergence**）、**依概率收敛**（**convergence in probability**）、**均方收敛**（**convergence in mean square** or **convergence in quadratic mean**）以及**依分布收敛**（**convergence in distribution** or **convergence in law**）。

2.1 几乎必然收敛

假设 $\{X_n\}$ 为在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列的随机变量。由于随机变量 X_i 是定义在样本空间 Ω 上的函数，因而定义随机变量收敛的一个最简单的想法是对于每一个 $\omega \in \Omega$，都有 $X_n(\omega) \to X(\omega)$，那么我们可以称随机变量序列 $\{X_n\}$ 收敛到随机变量 X。

然而我们不必要求如此严格，我们可以要求在某一些点 $\omega \in \Omega$ 处，$X_n(\omega) \nrightarrow X(\omega)$，只要这样的点「不多」就可以了。更进一步，我们可以使用概率来描述「不多」这一概念，这样就催生了第一个收敛的定义：

定义 7.（几乎必然收敛）如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量 $\{X_n\}$ 满足：

$$\mathscr{P}\left(\left\{\lim_{n\to\infty} X_n(\omega) = X\right\}\right) = 1$$

那么我们称 X_n 几乎必然收敛于 X，记为 $X_n \xrightarrow{\text{a.s.}} X$，或者 $X_n \to X$ $a.s.$。

回忆几乎处处（$a.e$）和几乎必然（$a.s.$）的概念，几乎必然收敛意味着的确存在某些情况使得 $\lim_{n\to\infty} X_n(\omega) = X$ 不成立，然而不成立的概率为 0。

特别的，如果令 $X = c$ 为常数，是一个退化的随机变量，那么当 $n \to \infty$ 时，随机变量趋向于一个非随机的常数。在统计中，如果我们的估计量作为一个随机变量趋向于一个常数（通常是真值），那么这个估计量被称为**一致的**（**consistent**）。接下来我们将会经常碰到随机变量收敛到常数的情况。

例 6. 令 $\Omega=\mathbb{R}$，分布函数为：

$$F(\omega)=\begin{cases}0 & \omega<-1\\ \frac{1}{2}\omega+\frac{1}{2} & -1\leq\omega\leq 1\\ 1 & \omega>1\end{cases}$$

即在 $[-1,1]$ 上的均匀分布，进而使用此分布函数构建 $\mathbb{R}$ 上的概率测度 P。随机变量

$$X_n(\omega)=\begin{cases}0 & if\ |\omega|>\frac{1}{n}\\ n & if\ |\omega|\leq\frac{1}{n}\end{cases}$$

因而 $\{\lim_{n\to\infty}X_n(\omega)\neq 0\}=\{0\}$，即只有在 $\omega=0$ 处 X_n 不收敛到 0，进而 $\mathscr{P}\{\lim_{n\to\infty}X_n(\omega)\neq 0\}=\mathscr{P}\{0\}=0$，因而 $X_n\overset{\text{a.s.}}{\to}0$。

对于几乎必然收敛，我们有如下命题。

定理 3. $X_n\overset{a.s.}{\to}X$ 等价于对于任意的 $\epsilon>0$，当 $n\to\infty$ 时，对于任意的 $k>n$，有：

$$\mathscr{P}(\{|X_k-X|<\epsilon\})\to 1$$

Proof. 令

$$A_{n,\epsilon}=\{\omega\in\Omega:|X_k-X|<\epsilon\ \forall k\geq n\}$$

那么根据几乎必然收敛的定义，$X_n\to X$ 即对于任意的 ϵ，存在一个 n 使得对于任意的 $k>n$ 有 $|X_k-X|<\epsilon$，因而收敛的点可以表述为：

$$\bigcap_{\epsilon>0}\bigcup_{n=1}^{\infty}A_{n,\epsilon}$$

因而证明 $X_n\overset{\text{a.s.}}{\to}X$ 等价于证明 $\mathscr{P}(\cup_{\epsilon>0}\cap_{n=1}^{\infty}A_{n,\epsilon})\to 1$。对于 $0<\epsilon_1<\epsilon_2$，由于 $\bigcup_{n=1}^{\infty}A_{n,\epsilon_1}\subset\bigcup_{n=1}^{\infty}A_{n,\epsilon_2}$，因而随着 $\epsilon\to 0$，$\bigcap_{\epsilon>0}\bigcup_{n=1}^{\infty}A_{n,\epsilon}\downarrow\bigcup_{n=1}^{\infty}A_{n,\epsilon}$。而由于 $A_{n,\epsilon}\subset A_{n+1,\epsilon}$ 对于 n 是单调递增的，因而随着 $n\to\infty$，$A_{n,\epsilon}\uparrow\bigcup_{n=1}^{\infty}A_{n,\epsilon}$。因而 $\mathscr{P}(\cup_{\epsilon>0}\cap_{n=1}^{\infty}A_{n,\epsilon})\to 1$ 等价于 $\mathscr{P}(A_{n,\epsilon})\to 1$，即命题得证。 □

2.2 依概率收敛

几乎必然收敛关注的是点态收敛，只要不收敛的点不要太多即可。而换一种思路，我们也可以关注随机变量 X_n-X 之间的误差，如果两者之间的误差趋向于 0，我们也可以定义其为收敛，这就诞生了依概率收敛的概念。

定义 8. （依概率收敛）如果对于任意的 $\epsilon>0$，概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的一系列随机变量序列 $\{X_n\}$ 满足：

$$\mathscr{P}(|X_n-X|>\epsilon)\to 0$$

那么我们称 X_n 依概率收敛于 X，记为 $X_n \xrightarrow{p} X$，或 $\text{plim}X_n = X$。

即随着 $n \to \infty$，X_n 与 X 之间误差比较大的点的概率趋向于 0。

例 7. 在例 (7) 中，对于任意的 $\epsilon > 0$，可以得到 $|X_n - 0| < \epsilon$ 的点集为：$\left\{\omega : |\omega| \leq \frac{1}{n}\right\}$，因而对于任意的 $\epsilon > 0$，都有 $\mathscr{P}(|X_n - 0| < \epsilon) = \frac{2}{n} \to 0$，因而 $X_n \xrightarrow{p} 0$。

注意依概率收敛和几乎必然收敛是两个不同的概念，依概率收敛并不一定能得到几乎必然收敛。

例 8. 令概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 如例 (7) 中定义，定义随机变量 $X_{j,n}, 0 \leq j < n-1$ 为：

$$X_{j,n} = \begin{cases} 1 & \omega \in [\frac{j}{n}, \frac{j+1}{n}) \\ 0 & else \end{cases}$$

对于任意的 i，令 $n = \sup_n \left\{i > \frac{n(n+1)}{2}\right\}$，$j = (i \mod n) + 1$，随机变量 $X_i = X_{j,n}$。对于任意的 $0 < \epsilon < 1$：

$$\mathscr{P}(|X_i| \geq \epsilon) = \frac{1}{n}$$

因而随着 $i \to \infty$，$n \to \infty$，$\mathscr{P}(|X_i| \geq \epsilon) \to 0$，因而 $X_i \xrightarrow{p} 0$。然而随着 $i \to \infty$，除了 $\omega = 1$ 之外，没有任何一个点收敛于 0，因而 X_i 并不几乎必然收敛于 0。

而相反，观察依概率收敛的定义，根据定理 (3)，对于任意的 $\epsilon > 0$，

$$\mathscr{P}(|X_n - X| < \epsilon) \geq \mathscr{P}(\{|X_k - X| < \epsilon, \forall k > n\}) \to 1$$

因而几乎必然收敛可以得到依概率收敛。因而几乎必然收敛是比依概率收敛更强的一个结论。

与数列极限相似，我们也可以在概率收敛的语境下定义小 o 符号。

定义 9. $\{X_n\}$ 与 $\{Y_n\}$ 为定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两个随机变量序列，如果

$$\frac{X_n}{Y_n} \xrightarrow{p} 0$$

那么我们记为 $X_n = o_p(Y_n)$。特别的，当 $Y_n = 1$ 时，即 $X_n = o_p(1)$，等价于 $X_n \xrightarrow{p} 0$。

小 o_p 符号是对小 o 符号的推广，他们的性质非常相似。比如，对于定义在同一概率空间上的三个随机变量序列 $\{X_n\}, \{Y_n\}, \{Z_n\}$，如果 $X_n = o_p\{Y_n\}$，那么 $X_n Z_n = o_p(Y_n Z_n)$。特别的，如果 $Z_n = a_n$ 为退化的随机变量序列，即实数序列，那么 $a_n X_n = o_p(a_n Y_n)$。例如，如果 $X_n = o_p(n)$，那么 $\frac{1}{n} X_n = o_p(1)$。

类似的，小 o_p 符号允许我们做近似的逼近。比如如果随机变量 $Z_n = X_n + o_p(1)$，那么我们可以使用 X_n 对 Z_n 进行近似，因为两者当 $n \to \infty$ 时是等价的。

类似的，我们还可以定义大 O_p 符号。

定义 10. $\{X_n\}$ 与 $\{Y_n\}$ 为定义在概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两个随机变量序列，如果对于任意的 $\epsilon > 0$，存在一个 C_ϵ 使得：

$$\sup_n \mathscr{P}(|X_n| \geq C_\epsilon |Y_n|) < \epsilon$$

那么我们记 $X_n = O_p(Y_n)$。特别的，当 $Y_n = 1$ 时，我们称 X_n **依概率有界**（**bounded in probability**）。

同样，大 O_p 符号是对大 O 符号的推广。如果 $X_n = O_p(1)$，那么意味着 X_n 不能太大，尽管允许 $X_n(\omega)$ 的某些值趋向于 ∞，但是这样的点随着 $n \to \infty$，其概率也慢慢变为 0。例如，在例 (7) 中，尽管 $X_n(0) \to \infty$，但是对于任意的 ϵ，令 $C_\epsilon = \frac{2}{\epsilon} + 1$，那么上式成立，因而 $X_n = O_p(1)$。

注意根据切比雪夫不等式，如果 $\mathrm{Var}(X_n) < M$，即随机变量序列 $\{X_n\}$ 的方差有界，那么对于任意的 $\epsilon > 0$，取 $C_\epsilon = \sqrt{\mathbb{E}(X_n^2)/\epsilon + 1}$，那么：

$$\mathscr{P}(|X_n| \geq C_\epsilon) \leq \frac{\mathbb{E}(X_n^2)}{C_\epsilon^2} = \frac{\mathbb{E}(X_n^2)}{\mathbb{E}(X_n^2)/\epsilon + 1} < \epsilon$$

因而 $X_n = O_p(1)$。

与大 O 符号类似，大 O_p 符号有如下性质：

定理 4.（大 O_p 性质）如果 $X_n = o_p(1), Y_n = o_p(1), Z_n = O_p(1), W_n = O_p(1)$，那么：

1. $X_n + Y_n = o_p(1)$
2. $X_n + Z_n = O_p(1)$
3. $Z_n + W_n = O_p(1)$
4. $X_n Y_n = o_p(1)$
5. $X_n Z_n = o_p(1)$
6. $Z_n W_n = O_p(1)$

以上特征可以分别简记为：$o_p(1) + o_p(1) = o_p(1), O_p(1) + O_p(1) = O_p(1), O_p(1) + o_p(1) = O_p(1), o_p(1) \cdot o_p(1) = o_p(1), O_p(1) \cdot o_p(1) = o_p(1), O_p(1) \cdot O_p(1) = O_p(1)$。

2.3 均方收敛

依概率收敛是讨论当 $n \to \infty$ 时 X_n 与 X 的误差，而类似于条件期望的定义，我们也可以使用平方的期望作为误差的一个度量，这就催生了均方收敛的概念。

定义 11.（均方收敛）如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一系列随机变量序列 $\{X_n\}$ 随着 $n \to \infty$ 满足：

$$\mathbb{E}\left(|X_n - X|^2\right) \to 0$$

那么我们称 X_n 均方收敛于 X，记为 $X_n \xrightarrow{L^2} X$。

均方收敛意味着，随着 $n \to \infty$，X_n 与 X 之间误差的平方的期望是趋向于 0 的。同样，均方收敛也是一个比依概率收敛更强的收敛。

定理 5. 如果随机变量序列 $X_n \xrightarrow{L^2} X$，那么 $X_n \xrightarrow{p} X$。

Proof. 根据切比雪夫不等式，对于任意的 $\epsilon > 0$，有：

$$\mathscr{P}\left(|X_n - X| > \epsilon\right) \leq \frac{\mathbb{E}\left(|X_n - X|^2\right)}{\epsilon^2} \to 0$$

□

实际上均方收敛的概念可以扩展到任意的 $r > 0$，如果 $\mathbb{E}\left(|X_n - X|^r\right) \to 0$，那么我们称 X_n 依 r 阶均值收敛于 X（**convergence in the rth mean**），记为 $X_n \xrightarrow{L^r} X$。可以证明，如果 $X_n \xrightarrow{L^r} X$，那么 $X_n \xrightarrow{p} X$。

然而相反则不成立，如在在例 (7) 中，$\mathbb{E}\left(|X_n|^2\right) = 2n$，并不趋向于 0，尽管 $X_n \xrightarrow{\text{a.s.}} 0$ 进而 $X_n \xrightarrow{p} 0$，然而 X_n 并不均方收敛于 0。因而均方收敛也是比依概率收敛更强的一个结论。而尽管均方收敛和几乎必然收敛都比依概率收敛要强，但是这两者之间并没有强弱的关系，也不等价，只有在一定条件下，几乎必然收敛才与均方收敛才同时成立。

2.4 依分布收敛

之前讨论的收敛分别从点态和误差的角度讨论了随机变量收敛的问题，接下来我们从随机变量的分布函数的角度讨论随机变量的收敛性。

如果 $\{X_n\}$ 为一系列随机变量，其对应的分布函数为 $F_n(x)$，那么其极限：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

如果存在，那么一个自然的问题是，$F(x)$ 是否还能构成一个分布函数。极限的单调性显然成立，那么接下来需要验证：

$$\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1 \tag{1}$$

以及 $F(x)$ 的右连续性。

首先，针对式 (1)，有以下定理：

定理 6. 对于随机变量序列 $\{X_n\}$ 及其对应的分布函数 $\{F_n(x)\}$，如果

$$\lim_{n\to\infty} F_n(x) = F(x)$$

那么式 *(1)* 成立的充分必要条件是 $X_n = O_p(1)$。

例 9. 如果令随机变量

$$X_n = \begin{cases} 0 & with\ prob\ 1-p \\ n & with\ prob\ p \end{cases}$$

可知 X_n 并不是有界的，而

$$F_n(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \le x < n \\ 1 & x = n \end{cases}$$

那么显然 $\lim_{n\to\infty} F_n(x) = 0 = F(x)$，因而 $\lim_{x\to\infty} F(x) = 0 \neq 1$。

以上定理解决了极限函数 $F(x)$ 的极限问题，然而对于右连续的要求，却并不能保证。比如令 $X_n = X + \frac{1}{n}$，X 的分布函数为 $G(x)$，那么 $F_n(x) = G\left(x - \frac{1}{n}\right)$，因而 $F(x) = \lim_{n\to\infty} F_n(x) = G(x-)$，如果 $G(x)$ 在 x 处不连续，那么 $F(x)$ 在 x 处为左连续函数。

因而为了避免右连续的问题，依分布收敛的定义通常只考虑分布函数的连续处的点。正式的，依分布收敛的定义如下：

定义 12. （依分布收敛）令 F_n, F 为分布函数，如果对于每一个 $F(x)$ 连续的点 x，有：

$$\lim_{n\to\infty} F_n(x) = F(x)$$

那么我们称 $F_n(x)$ 弱收敛于 $F(x)$，记为 $F_n \xrightarrow{w} F$。如果一系列随机变量 $\{X_n\}$ 的分布函数 $F_{X_n}(x) \xrightarrow{w} F_X$，我们称 X_n 依分布收敛于 X，记为 $X_n \xrightarrow{D} X$。

注意以上定义只要求随机变量 X_n 的分布函数收敛，而并没有对 X_n 和 X 之间的关系做出限定。实际上，根据依分布收敛的定义，$\{X_n\}$ 甚至不需要来自于同一个概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$。同时需要注意，弱收敛的定义中也没有要求对于每个 x 都有 $\lim_{n\to\infty} F_n(x) = F(x)$，而仅仅只要求了 $F(x)$ 在 x 处连续。

上面我们介绍了只有当 $X_n = O_p(1)$ 时，X_n 的分布函数的极限才可能是分布函数。而反过来依然成立，即如果 X_n 依分布收敛，那么 $X_n = O_p(1)$。由于依分布收敛于 O_p 符号的这种关系，依分布收敛的很多性质可以与 O_p 类比，比

如 $O_p(1)+o_p(1)=O_p(1)$，$O_p(1)\cdot o_p(1)=o_p(1)$ 等。令 $Y_n \xrightarrow{p} a, Z_n \xrightarrow{p} b$，即 $Y_n-a=o_p(1), Z_n-b=o_p(1)$，其中 a,b 为常数，那么：

$$Z_nX_n+Y_n \xrightarrow{D} bX+a$$

实际上依分布收敛是一个比依概率收敛还要弱的收敛。我们有如下结论：

定理 7. 如果 $X_n \xrightarrow{p} X$，那么 $X_n \xrightarrow{D} X$。

而相反，依分布收敛则不一定可以得到依概率收敛。但当随机变量收敛到一个常数时，我们有如下结论：

定理 8. 如果 $X_n \xrightarrow{D} c$，那么 $X_n \xrightarrow{p} c$。

此外，依分布收敛的概念与期望的收敛密不可分。我们有如下定理：

定理 9. $X_n \xrightarrow{D} X$ 的充分必要条件为，对于任意的有界连续函数 g，有：

$$\mathbb{E}[g(X_n)] \to \mathbb{E}[g(X)]$$

注意以上定理成立的前提是函数 $g(x)$ 必须为连续且有界的函数。比如如下两个例子中，如果违背了两个假定，结论并不一定成立。

例 10. 令 $g(x)=x$，随机变量

$$X_n=\begin{cases} n & with\ prob\ \frac{1}{n} \\ 0 & with\ prob\ 1-\frac{1}{n} \end{cases}$$

那么 $X_n \xrightarrow{D} 0$，然而 $\mathbb{E}[g(X_n)]=n\cdot\frac{1}{n}=1\neq 0=\mathbb{E}(0)$。

例 11. 令 $X_n=\frac{1}{n}$ 为退化的随机变量，那么 $X_n \xrightarrow{D} 0$。令

$$g(x)=\begin{cases} 1 & if\ x>0 \\ 0 & if\ x=0 \end{cases}$$

那么 $\mathbb{E}[g(X_n)]=1 \nrightarrow 0=\mathbb{E}[g(0)]$。

2.5 几种收敛之间的关系

以上我们介绍了四种收敛的概念，下面我们将集中收敛之间的关系整理如下：

定理 10. （四种收敛之间的关系）$\{X_n\}$ 为概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上的一系列随机变量，那么

1. $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X$

2. $X_n \xrightarrow{L^r} X \Rightarrow X_n \xrightarrow{p} X, r>0$，特别的，$X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{p} X$
3. $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{D} X$
4. 如果 $X_n \xrightarrow{p} X$，那么存在 X_n 的一个子列 $\{X_{n_j}\}$，当 $j \to \infty$ 时，$X_{n_j} \xrightarrow{a.s.} X$
5. 如果 $X_n \xrightarrow{D} X$ 且 $P(X=c)=1$，那么 $X_n \xrightarrow{p} c$
6. 若 $X_n \xrightarrow{a.s.} X$，且对于 $r>0$ 以及一个正的随机变量 Z，满足 $\mathbb{E}(Z)<\infty$，如果 $|X_n|^r \le Z$，那么 $X_n \xrightarrow{L^r} X$
7. 若 $X_n \xrightarrow{a.s.} X$，$X_n \ge 0$，且 $\mathbb{E}(X_n) \to \mathbb{E}(X) < \infty$，那么 $X_n \xrightarrow{L^1} X$
8. 对于任意的 $\epsilon>0$，如果 $\sum_{n=1}^{\infty} \mathscr{P}(|X_n - X| \ge \epsilon) < \infty$，那么 $X_n \xrightarrow{a.s.} X$。
9. $X_n \xrightarrow{p} 0 \Longleftrightarrow \mathbb{E}\left(\frac{|X_n|}{1+|X_n|}\right) \to 0$

此外，我们上面讨论的都是一元随机变量的收敛，所有概念都可以扩展到随机向量中，仅需要把所有的绝对值替换为欧几里得范数，即 $\|x\| = \sqrt{x'x}$。

3 大数定律

在实际应用中，我们经常关注一系列随机变量的和的极限情况，即：

$$S_n = \sum_{i=1}^{n} X_i$$

的极限情况。实际上我们下面要讨论的**大数定律**（**Law of Large Numbers, LLN**）即在讨论随机变量和的极限行为。特别的，在大数定律中，我们最为关注的是样本均值的极限情况，即：

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{p} 0$$

是否成立。实际上根据上一节的讨论，如果均方收敛那么则必然有依概率收敛，所以我们不妨从均方收敛入手。

实际上，$S_n - \mathbb{E}(S_n) = \sum_{i=1}^{n} [X_i - \mathbb{E}(X_i)]$，因而：

$$\begin{aligned}
\mathbb{E}[S_n - \mathbb{E}(S_n)]^2 &= \mathbb{E}\left(\sum_{i=1}^{n} [X_i - \mathbb{E}(X_i)]\right)^2 \\
&= \mathbb{E}\left(\sum_{i=1}^{n} [X_i - \mathbb{E}(X_i)]^2 + 2 \sum_{1 \le j < i \le n}^{n} [X_i - \mathbb{E}(X_i)][X_j - \mathbb{E}(X_j)]\right) \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le j < i \le n}^{n} \mathrm{Cov}(X_i, X_j)
\end{aligned}$$

如果假设每个 X_i 都有有限的二阶矩，即存在一个 M 使得对于所有的 $i = 1, 2, \ldots$，都有 $\operatorname{Var}(X_i) < M$，那么 $\sum_{i=1}^{n} \operatorname{Var}(X_i) < nM$，所以 $\sum_{i=1}^{n} \operatorname{Var}(X_i) = O(n)$。而根据 Cauthy-Schwartz 不等式，如果二阶矩有限，任意两个随机变量的协方差也必然有界，所以 $\sum_{1 \le j < i \le n}^{n} \operatorname{Cov}(X_i, X_j) = O\left(n^2\right)$。进而：

$$\mathbb{E}\left[\frac{S_n - \mathbb{E}(S_n)}{n}\right]^2 = \frac{1}{n^2} O(n) + \frac{1}{n^2} O\left(n^2\right) = o(1) + O(1)$$

因而当 $\operatorname{Cov}(X_i, X_j) = 0$，即 $\{X_i\}$ 之间两两不相关时，上式趋向于 0。因而我们有如下定理：

定理 11. 如果概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}$ 两两不相关，且存在一个 M 使得对于所有的 $i = 1, 2, \ldots$，都有 $Var(X_i) < M$，那么：

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{L^2} 0$$

从而

$$\frac{S_n - \mathbb{E}(S_n)}{n} \xrightarrow{p} 0$$

注意在以上定理中，我们并没有限定 X_i 具有相同的均值或者相同的方差，实际上，如果令

$$\bar{\mu}_n = \frac{\mathbb{E}(S_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i) \tag{2}$$

那么可以得到均值 $S_n/n \xrightarrow{p} \bar{\mu}_n$，即样本均值收敛到期望的平均。

推论 1. 如果 $\{X_i\}$ 为两两独立且同分布的随机变量序列，且其方差存在，记 $\mu = \mathbb{E}(X_i)$，那么：

$$\frac{S_n - \mu}{n} \xrightarrow{L^2} 0$$

或者：

$$\frac{S_n}{n} \xrightarrow{L^2} \mu$$

即样本均值收敛到其期望。

实际上，定理 (11) 不仅在均方收敛和依概率收敛的意义下成立，在几乎必然收敛的意义下也成立，即满足定理 (11) 的假设下，有：

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mu$$

证明见 Ch. 5.1, Chung (2001)。

以上的大数定律建立在不相关和二阶矩有界的条件下，同时得到了依概率收敛和几乎必然收敛的结果。实际上，根据结论中收敛方式的不同，大数定律分为「**强大数定律（Strong Law of Large Numbers，SLLN）**」和「**弱大数**

定律（Weak Law of Large Numbers，WLLN）」，两者的区别在于，SLLN 需要得到几乎必然收敛这一更强的结果，而弱大数定律只要求得到依概率收敛这一比较弱的结果。相应的，强大数定律需要的假设条件更强，而弱大数定律需要的假设更弱。

实际上，大数定律通常在几种不同的假设之间做权衡，比如随机变量之间的相关性（独立、两两独立、不相关）、是否同分布以及是否存在高阶矩。通常情况下，为了同样得到某种收敛，如果放松了某个假设，则必须在另外的假设上加强。而在这其中，**独立同分布（independent and identically distributed,** $i.i.d$）是相关性和是否同分布两个假设条件中最宽松的假设条件。

注意 $\{X_i\}$ 相互独立是比两两独立更强的假设，$\{X_i\}$ 相互独立必须要求任意数量的随机变量组合挑出来都是独立的，而两两独立只要求任意两个随机变量 X_i 和 X_j 是独立的。

下面我们就分别探讨弱大数定律和强大数定律。

3.1 弱大数定律

虽然依概率收敛结论比较弱，但是在通常情况下，依概率收敛是最简单，而且在很多应用中也已经足够的收敛形式，而弱大数定律就是解决随机变量之和的依概率收敛的问题。正式的，我们定义如下：

定义 13. 概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}_{i\geq 1}$，如果对于数列 $\{a_n\}_{n\geq 1}$ 和 $\{b_n\}_{n\geq 1}$，随着 $n \to \infty$，满足：

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0$$

那么我们称 $\{X_i\}$ 服从弱大数定律。

下面的例子是在独立同分布和二阶矩有限的情况下得到的最简单的弱大数定律。

例 12. 令 $\{X_i\}$ 为一系列 $i.i.d$ 的的随机变量，且 $\mathbb{E}\left(X_i^2\right) < \infty$，令 $\mu = \mathbb{E}(X_i), \sigma^2 = \mathrm{Var}(X_i)$，那么根据切比雪夫不等式：

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} = \frac{1}{\epsilon^2}\frac{n\sigma^2}{n^2} = \frac{\sigma^2}{\epsilon^2}\frac{1}{n} = o(1)$$

从而 $S_n/n \xrightarrow{p} \mu$。实际上，直接使用定理 (11) 可以得到同样的结果。

例 13. 如果令 $\{X_i\}$ 为一系列 $i.i.d$ 的随机变量，且 $X_i \sim Ber(p)$，那么 $\mathbb{E}(X_i) = p, \mathrm{Var}(X_i) = p(1-p) < \infty$，定义：

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

即成功的比例，那么根据上例，可以得到 $\hat{p} \xrightarrow{p} p$。

以上的弱大数定律是在二阶矩（方差）有限、独立同分布的条件下得到的，而这些假设仍然可以放宽。例如，根据定理 (11)，在二阶矩有限的条件下，同分布的假设可以不用，而独立的条件可以放宽为两两不相关。而以下的定理放松了二阶矩有限的假定以及独立的假定，保留了独立同分布的假定：

定理 12. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两两独立且同分布的随机变量序列，若 $\mathbb{E}|X_i| < \infty$，那么 $S_n/n \xrightarrow{p} \mu$，其中 $\mu = \mathbb{E}(X_i)$。

而以下的定理则同时放宽了同分布的假定以及二阶矩的假定。

定理 13. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的独立的随机变量序列，如果存在一个常数 $p \in [1, 2]$，随着 $n \to \infty$，使得：

$$\frac{1}{n^p} \sum_{i=1}^{n} \mathbb{E}|X_i|^p \to 0$$

那么 $S_n/n \xrightarrow{p} \mu_n$，其中 μ_n 根据式 *(2)* 定义。

3.2 强大数定律

以上讨论了弱大数定律，然而很多时候我们仍然需要更强的结论，如几乎必然收敛。因而引入强大数定律就非常必要了。

定义 14. 概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的一个随机变量序列 $\{X_i\}_{i\geq 1}$，如果对于数列 $\{a_n\}_{n\geq 1}$ 和 $\{b_n\}_{n\geq 1}$，随着 $n \to \infty$，满足：

$$\frac{S_n - a_n}{b_n} \xrightarrow{\text{a.s.}} 0$$

那么我们称 $\{X_i\}$ 服从强大数定律。

下面的例子是在独立同分布和四阶矩有限的情况下得到的最简单的强大数定律。

例 14. （Borel's SLLN）令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的 $i.i.d$ 的随机变量

序列，且 $\mathbb{E}X_i^4 < \infty$。根据切比雪夫不等式：

$$\begin{aligned} P\left(\left|\frac{S_n}{n}-\mu\right|>\epsilon\right) &\leq \frac{\mathbb{E}\left[\left(\frac{S_n}{n}-\mu\right)^4\right]}{\epsilon^4} \\ &= \frac{\mathbb{E}\left[\left(S_n-n\mu\right)^4\right]}{n^4\epsilon^4} \\ &= \frac{\mathbb{E}\left[\left(\sum_{i=1}^n\left(X_i-\mu\right)\right)^4\right]}{n^4\epsilon^4} \\ &= \frac{n\mathbb{E}\left[\left(X_i-\mu\right)^4\right]+3n\left(n-1\right)\left[\mathbb{E}\left(X_i-\mu\right)^2\right]^2}{n^4\epsilon^4} \\ &= O\left(\frac{1}{n^2}\right) \end{aligned}$$

根据定理 (10.8)，可以得到 $S_n/n \xrightarrow{\text{a.s.}} \mu$，其中 $\mu = \mathbb{E}\left(X_i\right)$。

可以看到，为了得到更强的结论（几乎必然收敛），需要使用更强的假设（四阶矩有限而非二阶矩有限）。回忆一下定理 (11) 也可以得到几乎必然收敛的结论，然而其中的条件仍然可以继续放宽。比如下面的 SLLN 就放宽了矩的假设，然而将强了独立性的假设以及同分布的假设。

定理 14. （*Etemadi's SLLN*）令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的两两独立且同分布的随机变量序列，且 $\mathbb{E}\left|X_i\right| < \infty$，那么 $S_n/n \xrightarrow{a.s.} \mu$。

实际上，上述定理的条件与定理 (12) 中的条件是一样的，而几乎必然收敛可以推出依概率收敛，因而上述定理实际上可以导出定理 (12)。

而以下定理相对于定理 (11) 则放宽了同分布及二阶矩有限的假定，加强了独立性的假定。

定理 15. 令 $\{X_i\}$ 为概率空间 $(\Omega, \mathscr{F}, \mathscr{P})$ 上的独立的随机变量序列，且 $\mathbb{E}\left|X_i\right| < \infty$，如果存在一个常数 $p \in [1,2]$，随着 $n \to \infty$，使得：

$$\sum_{i=1}^{\infty}\frac{\mathbb{E}\left|X_i\right|^p}{i^p} < \infty$$

那么 $S_n/n \xrightarrow{a.s.} \mu_n$。

例 15. 在例 (13) 中，使用定理 (14)，我们同样可以得到 $\hat{p} \xrightarrow{\text{a.s.}} p$。现在假设我们有一系列 $i.i.d$ 的随机变量 $\{X_i, i = 1, ..., n\}$，其分布函数为 $F\left(x\right)$，那么定义**经验分布函数**（**empirical distribution funciton**）为：

$$\hat{F}_n\left(x\right) = \frac{1}{n}\sum_{i=1}^n 1_{[X_i,\infty)}\left(x\right) = \frac{1}{n}\sum_{i=1}^n 1\left\{X_i \leq x\right\}$$

可以得到随机变量 $1_{[X_i,\infty)}(x) \sim Ber(F(x))$，因而 $\hat{F}_n(x) \overset{\text{a.s.}}{\to} F(x)$。实际上，我们可以得到更强的结论，即

$$P\left\{\sup_x \left|\hat{F}_n(x) - F(x)\right| \to 0\right\} = 1$$

首先令 $\epsilon > 0$ 为任意小的常数，令整数 $k > \frac{1}{\epsilon}$，并令 $-\infty = x_0 < x_1 \leq x_2 \leq \cdots \leq x_{k-1} < x_k = \infty$，使得对于 $j = 1, ..., k-1$，有 $F(x_j-) \leq \frac{j}{k} \leq F(x_j)$。注意如果 $x_{j-1} < x_j$，那么有 $F(x_j-) - F(x_{j-1}) \leq \epsilon$。由于 $\hat{F}_n(x) \overset{\text{a.s.}}{\to} F(x), \hat{F}_n(x-) \overset{\text{a.s.}}{\to} F(x-)$，因而：

$$\Delta_n = \max\left\{\left|F_n(x_j) - F(x_j)\right|, \left|F_n(x_j-) - F(x_j-)\right|, j = 1, ..., k-1\right\} \overset{\text{a.s.}}{\to} 0$$

令任意的 $x \in (x_{j-1}, x_j]$，那么

$$\begin{aligned}\hat{F}_n(x) - F(x) &\leq \hat{F}_n(x_j-) - F(x_{j-1}) \\ &\leq \hat{F}_n(x_j-) - F(x_j-) + \epsilon\end{aligned}$$

以及

$$\begin{aligned}\hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_{j-1}) - F(x_j-) \\ &\geq \hat{F}_n(x_{j-1}) - F(x_{j-1}) + \epsilon\end{aligned}$$

因而

$$\sup_x \left|\hat{F}_n(x) - F(x)\right| \leq \Delta_n + \epsilon \overset{\text{a.s.}}{\to} \epsilon$$

由于对于任意 $\epsilon > 0$，上式都成立，因而

$$P\left\{\sup_x \left|\hat{F}_n(x) - F(x)\right| \to 0\right\} = 1$$

3.3 一致大数定律

以上大数定律讨论的是一些随机变量的和的收敛，而**一致大数定律**（**Uniform law of large numbers, ULLN**）讨论的则是函数的收敛。

现在假设有一个函数 $g(x, \theta), \theta \in \Theta$，如果我们有一系列 $i.i.d$ 的随机变量 $\{X_i\}$，那么根据 SLLN，在可积性的条件下可以得到，对于任意的 $\theta \in \Theta$，有：

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i, \theta) \overset{\text{a.s.}}{\to} \mathbb{E}g(X_i, \theta) \triangleq g(\theta)$$

然而这个结论是基于点态收敛，仍然不足以支撑更多更有用的结论，很多时候

我们需要这个收敛对 θ 是一致（uniformly）收敛，即：

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}g\left(X_i,\theta\right)-g\left(\theta\right)\right|\xrightarrow{\text{a.s.}}0$$

对于以上结论，我们有以下定理可以使用：

定理 16. 对于 $i.i.d$ 的随机变量 $\{X_i\}$ 以及函数 $g\left(x,\theta\right)$，如果：

1. Θ 为紧集
2. 对于所有的 x，$g\left(x,\theta\right)$ 对 θ 都是连续的
3. 存在一个不依赖于 θ 的函数 $K\left(x\right)$ 满足 $\mathbb{E}\left(K\left(x\right)\right)<\infty$，使得对于所有的 x 和 θ，有：$\left|g\left(x,\theta\right)\right|\leq K\left(x\right)$

那么：

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}g\left(X_i,\theta\right)-g\left(\theta\right)\right|\xrightarrow{a.s.}0$$

4 中心极限定理

以上我们讨论了不同的大数定律。除了 S_n 的收敛特性之外，我们还关心 S_n 极限时的分布情况，此时我们需要使用**中心极限定理**（**Central limit theorem, CLT**）。

根据之前对依分布收敛的讨论，我们需要找到一个 $O_p\left(1\right)$，进而可以得到依分布收敛。如果假设 $\{X_i\}$ 支架两两不相关，那么：

$$\begin{aligned}\text{Var}\left(S_n\right)&=\mathbb{E}\left[S_n-\mathbb{E}\left(S_n\right)\right]^2\\&=\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i-\mathbb{E}\left(X_i\right)\right)\right]^2\\&=\sum_{i=1}^{n}\text{Var}\left(X_i\right)\\&=O\left(n\right)\end{aligned}$$

因而 $S_n/\sqrt{n}=O_p\left(1\right)$。如果记 $\bar{X}_n=S_n/n$，那么 $\sqrt{n}\bar{X}_n=O_p\left(1\right)$。

对于 $i.i.d$ 的随机变量 $\{X_i\}$，我们有如下定理：

定理 17. 令 $\{X_i\}$ 为概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上 $i.i.d$ 的随机变量序列，且 $\mathbb{E}\left(X_i\right)=\mu, Var(X_i)=\sigma^2$，那么：

$$\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{D}N\left(0,\sigma^2\right)$$

或：

$$\sqrt{n}\left(\frac{\bar{X}_n-\mu}{\sigma}\right)\xrightarrow{D}N(0,1)$$

以上定理意味着，只要 X_i 有有限的二阶矩，那么不管 X_i 服从何种分布，其均值在极限的条件下都服从正态分布。

例 16. 如果 $\{X_i\}$ 为 $i.i.d$ 的随机变量，且 $X_i\sim Ber(p)$，令 $\hat{p}_n$ 如前定义，那么：

$$\sqrt{n}(\hat{p}_n-p)\xrightarrow{D}N(0,p(1-p))$$

如果 $\{X_i\}$ 为 $i.i.d$ 的随机变量，且 $X_i\sim N(0,1)$，那么可知 $\mathbb{E}\left(X_i^2\right)=1,\mathbb{E}\left(X_i^4\right)=3$，因而：

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2-1\right)\xrightarrow{D}N(0,2)$$

以上定理还可以对随机向量进行扩展。

定理 18. 令 $\{X_i\}$ 为概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上 $i.i.d$ 的随机向量序列，且 $\mathbb{E}(X_i)=\mu, Var(X_i)=\Sigma$，那么：

$$\sqrt{n}\left(\bar{X}_n-\mu\right)\xrightarrow{D}N(0,\Sigma)$$

或：

$$\sqrt{n}\Sigma^{-\frac{1}{2}}\left(\bar{X}_n-\mu\right)\xrightarrow{D}N(0,I)$$

以上中心极限定理仅适用于独立同分布的情况，而如果扩展到独立但不同分布的情况，情况就稍微复杂一点。为此，我们先引入一个随机变量的三角向量，即形如：

$$\begin{array}{llllll}
X_{11} & & & & \sum\Rightarrow & Z_1\\
X_{21} & X_{22} & & & \sum\Rightarrow & Z_2\\
X_{31} & X_{32} & X_{33} & & \sum\Rightarrow & Z_3\\
X_{41} & X_{42} & X_{43} & X_{44} & \sum\Rightarrow & Z_4\\
\cdots & & & & & \cdots
\end{array}$$

其中每一行的随机变量 X_{nj} 假设为相互独立的。我们有如下定理：

定理 19. （*Lindeberg-Feller*）对于 $n=1,2,...$，令 $X_{nj},j=1,2,...,n$ 为独立的随机变量，$\mathbb{E}(X_{nj})=0$，$Var(X_{nj})=\sigma_{nj}^2$。令

$$Z_n=\sum_{j=1}^{n}X_{nj}$$

并令

$$\sigma_n^2 = \sum_{j=1}^{n} \sigma_{nj}^2$$

如果 ***Lindeberg*** 条件成立，即对于任意的 $\epsilon > 0$，随着 $n \to \infty$，有：

$$\frac{1}{\sigma_n^2} \sum_{j=1}^{n} \mathbb{E}\left[X_{nj}^2 \cdot 1\left\{|X_{nj}| \geq \epsilon \sigma_n^2\right\}\right] \to 0$$

那么有

$$\frac{Z_n}{B_n} \xrightarrow{D} N(0,1)$$

以上定理被称为 Lindeberg-Feller CLT。其中 Lindeberg 条件的一个推论是，随着 $n \to \infty$，

$$\max_{j \leq n} \frac{\sigma_{nj}^2}{\sigma_n^2} \to 0$$

也就是当 n 趋向于无穷时，任何随机变量的方差都是小到可以忽略不计的，即在 Z_n 中，每个随机变量 X_{nj} 对 Z_n 的影响可以不一样，但是没有一个 X_{nj} 对 Z_n 有决定性的影响。

实际上由于 Lindeberg 条件比较难以验证，很多时候我们会直接使用其充分条件，即如果存在 $\delta > 0$，有：

$$\sum_{j=1}^{n} \mathbb{E}\left|X_{nj} - \mathbb{E}X_{nj}\right|^{2+\delta} = o\left(\sigma_n^{2+\delta}\right)$$

那么 Lindeberg 条件即满足。

5 变换的收敛

以上讨论了随机变量的收敛，很多时候我们会关心随机变量的变换的收敛情况。比如，我们知道在大样本条件下，样本均值渐进服从正态分布，那么样本均值的平方服从何种分布呢？为此我们引入以下定理：

定理 20. 令 $\{X_i\}$ 为 k 维随机向量，$g(x) : \mathbb{R}^k \to \mathbb{R}^l$ 为连续函数，那么：

1. $X_n \xrightarrow{a.s.} X \Rightarrow g(X_n) \xrightarrow{a.s.} g(X)$
2. $X_n \xrightarrow{p} X \Rightarrow g(X_n) \xrightarrow{p} g(X)$
3. $X_n \xrightarrow{D} X \Rightarrow g(X_n) \xrightarrow{D} g(X)$

例 17. 根据中心极限定理，$i.i.d$ 的 k 维随机变量 $\{X_i\}$ 满足：

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{D} N(0, \Sigma)$$

那么 $\sqrt{n}\Sigma^{-\frac{1}{2}}\left(\bar{X}_n - \mu\right) \xrightarrow{D} N(0, I)$，从而 $n\left(\bar{X}_n - \mu\right)' \Sigma^{-1}\left(\bar{X}_n - \mu\right) \xrightarrow{D} \chi_k$。

例 18. 对于二维随机向量 (X,Y)，其相关系数定义为

$$\operatorname{Corr}(X,Y)=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}=\frac{\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}$$

令 (X_i,Y_i) 为 $i.i.d$ 的样本，那么在可积性条件下，

$$\begin{cases}\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{p}\mathbb{E}(X)\\ \frac{1}{n}\sum_{i=1}^{n}Y_i\xrightarrow{p}\mathbb{E}(Y)\\ \frac{1}{n}\sum_{i=1}^{n}X_iY_i\xrightarrow{p}\mathbb{E}(XY)\\ \frac{1}{n}\sum_{i=1}^{n}X_i^2\xrightarrow{p}\mathbb{E}\left(X^2\right)\\ \frac{1}{n}\sum_{i=1}^{n}Y_i^2\xrightarrow{p}\mathbb{E}\left(Y^2\right)\end{cases}$$

从而

$$\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\frac{1}{n}\sum_{i=1}^{n}X_i\frac{1}{n}\sum_{i=1}^{n}Y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}Y_i^2-\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)^2}}\xrightarrow{p}\operatorname{Corr}(X,Y)$$

定理 21. （*Slutsky*）如果随机变量 $X_n\xrightarrow{D}X$，$R_n=o_p(1)$，那么 $X_n+R_n\xrightarrow{D}X$。同时如果 $Y_n\xrightarrow{p}a\neq 0$，那么 $\frac{X_n}{Y_n}\xrightarrow{D}X/a$。如果 $Y_n\xrightarrow{p}a$，那么 $X_nY_n\xrightarrow{D}aX$。

使用以上定理可以非常方便的推导各类极限分布，比如：

例 19. 之前曾讨论过，如果 $(X_1,...,X_n)'\sim N\left(\mu\iota,\sigma^2I\right)$，那么：

$$\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}}\sim t_{n-1}$$

现在我们不假设 X_i 服从正态分布，而是假设其独立同分布且具有有限的二阶矩，那么我们有

$$\bar{X}\xrightarrow{p}\mathbb{E}(X),\frac{1}{n}\sum_{i=1}^{n}X_i^2\xrightarrow{p}\mathbb{E}\left(X^2\right)$$

进而：

$$\begin{aligned}\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1} &= \frac{\sum_{i=1}^{n}\left(X_i^2+\bar{X}^2-2\bar{X}X_i\right)}{n-1}\\ &= \frac{\sum_{i=1}^{n}X_i^2+n\bar{X}^2-2\bar{X}\sum_{i=1}^{n}X_i}{n-1}\\ &= \frac{\sum_{i=1}^{n}X_i^2+n\bar{X}^2-2n\bar{X}^2}{n-1}\\ &= \frac{\sum_{i=1}^{n}X_i^2+n\bar{X}^2-2n\bar{X}^2}{n-1}\\ &= \frac{\sum_{i=1}^{n}X_i^2-n\bar{X}^2}{n-1}\\ &= \frac{\sum_{i=1}^{n}X_i^2}{n-1}-\frac{n}{n-1}\bar{X}^2\\ &\xrightarrow{p}\mathbb{E}\left(X^2\right)-\left[\mathbb{E}\left(X\right)\right]^2=\mathrm{Var}\left(X\right)\end{aligned}$$

进而：

$$\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}}\xrightarrow{p}\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\mathrm{Var}\left(X\right)}}=\sqrt{n}\left(\frac{\bar{X}-\mu}{\sqrt{\mathrm{Var}\left(X\right)}}\right)\xrightarrow{D}N\left(0,1\right)$$

因而当样本足够大时，即使 X_i 不服从正态分布，以上的 $\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n-1}}}$ 仍然服从正态分布。

现在假设 $a_n\left(X_n-c\right)\xrightarrow{D}Y$，$\lim_{a_n\to\infty}a_n=\infty$，那么 $a_nX_n=O_p\left(1\right),X_n-c=o_p\left(1\right)$。对于任意的连续二阶可微的函数 $g\left(x\right)$，都可以对其泰勒展开：

$$\begin{aligned}a_n\left[g\left(X_n\right)-g\left(c\right)\right] &= \frac{\partial g}{\partial x'}\left(c\right)a_n\left(X_n-c\right)+\frac{1}{2}a_n\left(X_n-c\right)'\frac{\partial^2 g}{\partial x\partial x'}\left(c\right)\left(X_n-c\right)+\cdots\\ &= \frac{\partial g}{\partial x'}\left(c\right)a_n\left(X_n-c\right)+\frac{1}{2}O_p\left(1\right)O\left(1\right)o_p\left(1\right)\\ &= \frac{\partial g}{\partial x'}\left(c\right)a_n\left(X_n-c\right)+o_p\left(1\right)\\ &\xrightarrow{D}\frac{\partial g}{\partial x'}\left(c\right)Y\end{aligned}$$

因而 $a_n\left[g\left(X_n\right)-g\left(c\right)\right]\xrightarrow{D}\frac{\partial g}{\partial x'}\left(c\right)Y$。特别的，如果 $Y\sim N\left(0,\Sigma\right)$，那么：

$$a_n\left[g\left(X_n\right)-g\left(c\right)\right]\xrightarrow{D}N\left(0,\frac{\partial g}{\partial x'}\left(c\right)\Sigma\frac{\partial g}{\partial x}\left(c\right)\right)$$

以上过程我们称之为 **delta 方法**（**delta method**）。

例 20. 令 $\{X_i\}$ 为概率空间 $(\Omega,\mathscr{F},\mathscr{P})$ 上 $i.i.d$ 的随机变量序列，且 $\mathbb{E}\left(X_i\right)=$

$\mu, \mathrm{Var}(X_i) = \sigma^2$，那么根据中心极限定理：

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right)$$

如果我们关心 $Y = \exp\left(\bar{X}_n\right)$ 的分布，那么可以对其进行泰勒展开：

$$\begin{aligned}
\sqrt{N}Y - \sqrt{N}\exp(\mu) &= \sqrt{N}\exp\left(\bar{X}_n\right) - \sqrt{N}\exp(\mu) \\
&= \sqrt{N}\exp(\mu)\left(\bar{X}_n - \mu\right) + \frac{1}{2}\sqrt{N}\exp(\mu)\left(\bar{X}_n - \mu\right)^2 + \cdots \\
&= \sqrt{N}\exp(\mu)\left(\bar{X}_n - \mu\right) + \frac{1}{2}\exp(\mu) O_p(1) o_p(1) \\
&= \sqrt{N}\exp(\mu)\left(\bar{X}_n - \mu\right) + o_p(1) \\
&\xrightarrow{D} \sqrt{N}\exp(\mu)\left(\bar{X}_n - \mu\right) \\
&= N\left(0, \exp(2\mu)\sigma^2\right)
\end{aligned}$$

因而 $Y \xrightarrow{D} N\left(0, \frac{\exp(2\mu)\sigma^2}{N}\right)$。

习题

练习 1. 写程序近似逼近正态分布的分布函数并画图。

练习 2. 请找出一项表达式使其与如下序列渐进等价：

1. $\ln n + \frac{1}{2}n$
2. $\ln n + \ln(\ln n)$
3. $n^2 + e^n$

练习 3. 请确定 $a_n = \sqrt{\log n}$ 与 $b_n = \log(\sqrt{n})$ 的阶。

练习 4. 请问如下命题是否成立？若成立，请给出证明，若不成立，请给出反例：

1. $a_n = o(b_n), c_n = o(b_n)$，那么 $a_n + c_n = o(b_n)$。
2. $a_n = o(b_n), c_n = o(d_n)$，那么 $a_n + c_n = o(b_n + d_n)$。

练习 5. 如果 $h = n^q, -1 < q < 0$，$a_n = \frac{1}{n^2h^2} + \frac{10}{n^3h}$，$b_n = 3h^3 + 10h^4$，求 q 使得 $a_n + b_n$ 以最快的速度趋向于 0。

练习 6. 程序题：给定一个 p 和一个 n，重复的生成 n 个服从伯努利分布的随机变量，并计算其均值 $\hat{p}_n$。对于 $n = 10, 20, 30, ..., 1000$ 重复以上过程，并将结果以 n 为 x 轴，将 $\hat{p}_n$ 画在一张图上观察其收敛性。将伯努利分布换成 Cauthy 分布，再次观察其收敛性。

练习 7. 程序题：给定一个 p 和一个 n，重复的生成 n 个服从伯努利分布的随机变量，并计算 $\hat{p}_n$。对于每一个 n，重复计算出 500 个 $\hat{p}$。对于 $n = 3, 10, 30, 100$ 重复以上过程，并画出每个 n 的情况下 500 个 $\hat{p}$ 的直方图。

练习 8. 程序题：将上述练习中的伯努利分布换成正态分布的平方，重复上述练习的过程。换成 Cauthy 分布，继续重复上述练习的过程。

参考文献

[1] Athreya, K.B., Lahiri, S.N., 2006. Measure Theory and Probability Theory. Springer, New York.

[2] Chung, K.L., 2001. A Course in Probability Theory, 3rd editio. ed. Elsevier Ltd., Singapore.

[3] Ferguson, T.S., 1996. A Course in Large Sample Theory, Chapman &. ed. New York.

[4] Lehmann, E.L., 1999. Elements of Large-Sample Theory. Springer Science & Business Media, New York.

[5] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

[6] Wooldridge, J.M., 2010. Econometric Analysis of Cross Sectional and Panel Data, 2nd ed. The MIT Press, Cambridge.

第七节 · 参数估计

司继春

上海对外经贸大学统计与信息学院

在这一节中我们将学习统计推断中参数估计的相关内容，包括点估计和区间估计两部分。其中，我们将介绍两种点估计的方法：矩估计和极大似然估计。

1 参数估计

1.1 参数估计的基本概念

在参数模型中，我们假设总体 P 属于某一个参数族 $\{P_\theta, \theta \in \Theta\}$，从而推断总体等价于找到一个参数 θ_0，使得 $P_{\theta_0} = P$。我们一般把 θ_0 成为真值（true value）。而由于总体是不可观测的，我们只能通过样本对总体进行推断，因而我们不可能得到 θ_0 的精确值，只能对其进行估计，即**参数估计**（**Estimation**）。

参数估计包含两部分，即**点估计**（**Point estimation**）和**区间估计**（**interval estimation**）。其中点估计即找到一个统计量 $\hat{\theta}(x)$，对总体参数 θ_0 进行推断，而统计量 $\hat{\theta}$ 我们一般称为**估计量**（**estimator**）。而区间估计即找到一组统计量 $L(x), U(x)$，使得由其组成的区间包含总体参数 θ_0 的概率为已知的，即 $P(L(x) \leq \theta_0 \leq U(x)) = p$。其中统计量 $L(x), U(x)$ 被成为**区间估计量**（**interval estimator**）。

此外，类似于统计量及其实现的差别，我们还需要区分估计量和估计。如上所述，估计量即样本的一个函数，即用于估计参数的统计量，而估计（estimate）是对于某一个样本，估计量的实现。

1.2 评价估计量的标准

对于同一个参数，经常我们有不同的估计量，比如对于正态总体 $x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d$，自然地，$\sigma^2$ 的一个估计量为：

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

而类似的，我们也可以使用：

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{1}$$

作为 σ^2 的一个估计。以上两个估计量的差别在于分母的不同，很显然，以上两个估计量具有不同的抽样分布。那么，在有很多统计量可供选择时，该如何评价这些统计量呢?

一个常用的标准是**均方误差**（**Mean squared error**, MSE），即对于一个参数 θ 和它的估计量 $\hat{\theta}$，其误差平方的期望 $\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2$ 为估计量 $\hat{\theta}$ 的均方误差。注意由于：

$$\begin{aligned}
\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 &= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}+\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)^2 + \mathbb{E}\left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 + 2\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right)\left(\mathbb{E}\hat{\theta}-\theta_0\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 + 2\left(\mathbb{E}\hat{\theta}-\theta_0\right)\mathbb{E}\left(\hat{\theta}-\mathbb{E}\hat{\theta}\right) \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left(\mathbb{E}\hat{\theta}-\theta_0\right)^2 \\
&= \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2
\end{aligned}$$

其中定义**偏差**（**bias**）$\mathrm{Bias}\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta}-\theta_0\right)$，从而 $\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2$，即均方误差等于估计量的方差与偏差平方的和。

因而，降低均方误差有两种途径：降低估计的方差以及降低偏差。此外，根据均方收敛的定义，只要 $\mathbb{E}\left(\hat{\theta}-\theta_0\right)^2 = \mathrm{Var}\left(\hat{\theta}\right) + \left[\mathrm{Bias}\left(\hat{\theta}\right)\right]^2 \to 0$，那么 $\hat{\theta} \xrightarrow{L^2} \theta_0$，从而 $\hat{\theta} \xrightarrow{p} \theta_0$。尽管小样本情况下估计量的偏差不为 0，但是我们希望当样本量趋向于无穷时，估计量收敛到真值，也是可以接受的。这就引出了评价估计量的三条标准：**无偏性**（**unbiasedness**）、**有效性**（**efficiency**）、**一致性**（**consistency**）。

1.2.1 无偏性

无偏性要求估计量的偏差为 0。当估计量的偏差为 0，即 $\mathbb{E}\left(\hat{\theta}\right) = \theta_0$ 时，我们称估计量 $\hat{\theta}$ 为无偏的（unbiased）。无偏性意味着，尽管对于每个样本，对 θ_0 的估计不可能完全准确，但是平均而言，估计量 $\hat{\theta}$ 总是围绕在真值 θ_0 的周围，不会有系统性的偏差。

例 1. 根据之前的计算，我们知道对于正态总体，$x_i \sim N\left(\mu, \sigma^2\right)$ $i.i.d$：

$$\mathbb{E}\left(s^2\right) = \sigma^2$$

因而式 (1) 中定义的 $\hat{\sigma}^2$ 的期望为:

$$\mathbb{E}\left(\hat{\sigma}^2\right) = \mathbb{E}\left(\frac{N-1}{N}s^2\right) = \frac{N-1}{N}\sigma^2$$

因而 $\text{Bias}\left(s^2\right) = 0, \text{Bias}\left(\hat{\sigma}^2\right) = \frac{1}{N}\sigma^2$，只有 s^2 是 σ^2 的无偏估计量。

1.2.2 有效性

为了降低 MSE，除了降低偏差以外，降低估计量的方差 $\text{Var}\left(\hat{\theta}\right)$ 也是非常重要的手段。一般而言，如果两个估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$，如果 $\text{Var}\left(\hat{\theta}_1\right) < \text{Var}\left(\hat{\theta}_2\right)$，那么我们称 $\hat{\theta}_1$ 相对于 $\hat{\theta}_2$ 是有效的。

例 2. 在例 (1) 中，因为 $\hat{\sigma}^2 = \frac{N-1}{N}s^2$，从而

$$\text{Var}\left(\hat{\sigma}^2\right) = \left(\frac{N-1}{N}\right)^2 \text{Var}\left(s^2\right) < \text{Var}\left(s^2\right)$$

因而 $\hat{\sigma}^2$ 是相对于 s^2 更有效的估计量。

我们注意到，尽管 s^2 比 $\hat{\sigma}^2$ 的偏差更小，但是 s^2 比 $\hat{\sigma}^2$ 的方差更大。在很多应用问题，比如非参数回归或者监督学习（supervised learning）中，我们都会碰到类似的偏差-方差权衡（bias-variance tradeoff），即很多时候，同时降低偏差和方差是不可能的。

1.2.3 一致性

很多时候，尽管在有限样本下，一个估计量的偏差不为零，但是如果样本量足够大时，估计量与真值之间的误差充分的小，我们也可以接受。如果一个估计量 $\hat{\theta}$ 依概率收敛到真值 θ_0，即 $\hat{\theta} \xrightarrow{p} \theta_0$，那么我们称估计量 $\hat{\theta}$ 为一致估计量。如果一个估计量是不一致的，也就是说即便我们拥有无限多的样本，我们也不能获得真值 θ_0 的估计，因而一致性是对一个估计量的最低要求。

例 3. 在例 (1) 中，由于:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}x_i^2 - \bar{x}^2$$

其中 $\bar{x} \xrightarrow{p} \mu$，而:

$$\frac{1}{N}\sum_{i=1}^{N}x_i^2 \xrightarrow{p} \mathbb{E}\left(x^2\right) = \mu^2 + \sigma^2$$

从而 $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$，即 $\hat{\sigma}^2$ 是 σ^2 的一致估计量。而:

$$s^2 = \frac{N}{N-1}\hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

因而 s^2 也是 σ^2 的一致估计量。

需要注意的是，无偏性关注的是估计量的期望，而一致性则是当样本足够大时估计量的性质，两者并没有任何必然联系，无偏性和一致性并不是彼此的充分或者必要条件。

2 矩估计

矩估计（**method of moments**）是使用历史最长的参数估计方法，其思路是使用样本矩代替总体矩对参数进行估计。如果样本 $x=(x_1,...,x_N)'$ 是来自于总体 P_{θ_0} 的独立同分布的样本，那么其一阶样本矩和一阶总体矩可以分别定义为：

$$\begin{cases} m_1(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \\ \mu_1(\theta)=\mathbb{E}_\theta x_i \end{cases}$$

其中 $\mathbb{E}_\theta$ 表示给定一个参数 θ，使用总体 P_θ 计算得到的理论的总体期望。由于真值为 θ_0，因而真实的期望 $\mathbb{E}x_i=\mathbb{E}_{\theta_0}x_i$。

我们知道，在一定比较宽松的条件下，根据大数定律有：

$$m_1(x)=\frac{1}{N}\sum_{i=1}^{N}x_i \xrightarrow{p} \mathbb{E}_{\theta_0}x_i=\mu_1(\theta_0)$$

如果 $\mu_1(\cdot)$ 是一个连续且可逆的函数，那么真实参数 θ_0 可以写为：

$$\theta_0=\mu_1^{-1}(\mu_1(\theta_0))$$

那么我们可以使用样本矩 $m_1(x)$ 代替上式中的总体矩 $\mu_1(\theta_0)$，由于 $m_1(x)\xrightarrow{p}\mu_1(\theta_0)$，而 $\mu_1^{-1}(\cdot)$ 为连续函数，从而估计量：

$$\hat{\theta}\triangleq\mu_1^{-1}(m_1(x))\xrightarrow{p}\mu_1^{-1}(\mu_1(\theta_0))=\theta_0$$

从而 $\hat{\theta}$ 是 θ_0 的一致估计。

更加形象的理解是，给定任何一个 θ，总体 P_θ 是一个确定的概率函数，因而可以计算在 θ 情况下的样本矩 $\mathbb{E}_\theta x_i$。理论上，样本矩 $m_1(x)$ 和总体矩 $\mu_1(\theta)$ 在样本量足够大的情况下应该是充分接近的，那么我们可以找到一个 $\hat{\theta}$ 使得 $\mu_1\left(\hat{\theta}\right)$ 与 $m_1(x)$ 的差距最小，从而得到对真值 θ_0 的估计。以上就是矩估计的思想。

例 4. 如果样本 $x_i\sim P(\lambda_0)$ $i.i.d$，我们知道样本矩 $m_1(x)=\bar{x}$，比如，如果我们的样本观测值为 $x=(3,5,7,2,3)$，那么样本矩为

$$m_1(x)=\bar{x}=\frac{3+5+7+2+3}{5}=4$$

加入任意给定一个 λ，比如令 $\lambda=2$，总体的期望为 $\mathbb{E}_\lambda x_i=\lambda=2\neq 4$，因而如果认为 $\lambda_0=2$，那么总体 $P(2)$ 所产生的总体矩与样本矩仍然有差异。只有当 $\lambda=4$ 时，总体矩 $\mathbb{E}_\lambda x_i=4=m_1(x)$，总体矩与我们观察到的样本矩相等，因而我们可以推断 $\hat{\lambda}=4$。一般的，对于泊松分布总体，我们可以直接令总体矩等于样本矩得到估计，即：

$$\hat{\lambda}=m_1(x)=\bar{x}$$

下面我们分别讨论该估计量的无偏性和一致性。首先，对于无偏性，由于：

$$\mathbb{E}\hat{\lambda}=\mathbb{E}\bar{x}=\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}x_i=\lambda_0$$

因而 $\hat{\lambda}$ 是 λ_0 的无偏估计。而对于一致性，根据大数定理：

$$\hat{\lambda}=\bar{x}\xrightarrow{p}\mathbb{E}x_i=\lambda_0$$

因而 $\hat{\lambda}$ 是 λ_0 的一致估计。当然，一致性还可以通过分析 $\hat{\lambda}$ 的偏差与方差来证明。根据以上讨论，该估计量的偏差为 $\text{Bias}\left(\hat{\lambda}\right)=\mathbb{E}\left(\hat{\lambda}\right)-\lambda_0=0$，而其方差为：

$$\text{Var}\left(\hat{\lambda}\right)=\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)=\frac{1}{N^2}\sum_{i=1}^{N}\text{Var}(x_i)=\frac{\lambda_0}{N}$$

从而 $\mathbb{E}\left(\hat{\lambda}-\lambda_0\right)^2=\text{Var}\left(\hat{\lambda}\right)+\left[\text{Bias}\left(\hat{\lambda}\right)\right]^2\to 0$，从而 $\hat{\lambda}\xrightarrow{L^2}\lambda_0$，从而 $\hat{\lambda}\xrightarrow{p}\lambda_0$。进一步，根据中心极限定理，有：

$$\sqrt{N}\left(\hat{\lambda}-\lambda_0\right)=\sqrt{N}\left(\bar{x}-\lambda_0\right)\xrightarrow{D}N(0,\lambda_0)$$

因而 $\hat{\lambda}\xrightarrow{D}N\left(\lambda_0,\frac{\lambda_0}{N}\right)$。

例 5. 如果样本 $x_i\sim LN(\mu_0,2)\ \ i.i.d$，即总体为对数正态分布，且一个参数 $\sigma^2=2$ 已知。类似的，样本矩 $m_1(x)=\bar{x}$，而总体矩 $\mathbb{E}_\mu x_i=e^{\mu_0+1}$。根据矩估计的思想，令总体矩等于样本矩，即：

$$e^{\hat{\mu}+1}=m_1(x)=\bar{x}$$

可以得到 μ_0 的矩估计值：

$$\hat{\mu}=\ln\bar{x}-1$$

现在讨论该估计量的无偏性和一致性。首先根据 Jensen 不等式：

$$\mathbb{E}(\hat{\mu})=\mathbb{E}(\ln\bar{x})-1\leq\ln(\mathbb{E}\bar{x})-1=\ln e^{\mu_0+1}-1=\mu_0$$

因而 $\hat{\mu}$ 并不是 μ_0 的无偏估计。而根据大数定律，$\bar{x}\xrightarrow{p}\mathbb{E}x_i=e^{\mu_0+1}$，由于 ln 为

连续函数，从而：

$$\hat{\mu} = \ln \bar{x} - 1 \xrightarrow{p} \ln e^{\mu_0+1} - 1 = \mu_0$$

因而 $\hat{\mu}$ 是 μ_0 的一致估计。此外，我们还可以使用 delta 方法计算 $\hat{\mu}$ 的极限分布。根据中心极限定理，$\sqrt{N}\left(\bar{x} - e^{\mu_0+1}\right) \xrightarrow{D} N\left(0, \mathrm{Var}\left(x_i\right)\right)$，其中 $\mathrm{Var}\left(x_i\right) = e^{2(\mu_0+2)} - e^{2\mu_0+2}$。因而，对估计量进行泰勒展开：

$$\begin{aligned}
\sqrt{N}\left(\hat{\mu} - \mu_0\right) &= \sqrt{N}\left(\ln \bar{x} - 1 - \ln e^{\mu_0+1} + 1\right) \\
&= \sqrt{N}\left(\ln \bar{x} - \ln e^{\mu_0+1}\right) \\
&= \sqrt{N}\left(\frac{1}{e^{\mu_0+1}}\left(\bar{x} - e^{\mu_0+1}\right) + o_p\left(1\right)\right) \\
&\xrightarrow{D} \frac{1}{e^{\mu_0+1}}\sqrt{N}\left(\bar{x} - e^{\mu_0+1}\right) \\
&= N\left(0, e^{-2(\mu_0+1)}\mathrm{Var}\left(x_i\right)\right)
\end{aligned}$$

因而 $\hat{\mu} \xrightarrow{D} N\left(\mu_0, \frac{e^2-1}{N}\right)$。

更一般的，如果我们有 k 个未知参数，即 θ 为 k 维向量，那么我们可以联立前 k 个样本矩和总体矩对 θ 进行估计，其中前 k 个样本矩和总体矩定义为：

$$\begin{cases}
m_1\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} x_i^1 & \mu_1\left(\theta\right) = \mathbb{E}_\theta x_i^1 \\
m_2\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 & \mu_2\left(\theta\right) = \mathbb{E}_\theta x_i^2 \\
\vdots & \vdots \\
m_k\left(x\right) = \frac{1}{N}\sum_{i=1}^{N} x_i^k & \mu_k\left(\theta\right) = \mathbb{E}_\theta x_i^k
\end{cases}$$

一般而言，如果我们有 k 个参数，$\theta = \left(\theta_1, ..., \theta_k\right)'$，那么我们使用前 k 个矩，解方程：

$$\begin{cases}
m_1\left(x\right) = \mu_1\left(\hat{\theta}\right) \\
m_2\left(x\right) = \mu_2\left(\hat{\theta}\right) \\
\vdots \\
m_k\left(x\right) = \mu_k\left(\hat{\theta}\right)
\end{cases}$$

如果该联立方程有解，即可得到参数 θ_0 的估计。

例 6. 对于正态总体 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中未知总体参数 $\theta = \left(\mu, \sigma^2\right)$，其一阶样本矩为 $m_1\left(x\right) = \bar{x}$，二阶样本矩为 $m_2\left(x\right) = \overline{x^2}$。我们知道对于正态分布，$\mu_1\left(\theta\right) = \mu, \mu_2\left(\theta\right) = \mu^2 + \sigma^2$，从而矩估计为：

$$\begin{cases}
m_1\left(x\right) = \bar{x} = \hat{\mu} \\
m_2\left(x\right) = \overline{x^2} = \hat{\mu}^2 + \hat{\sigma}^2
\end{cases}$$

解得：

$$\begin{cases}\hat{\mu}=\bar{x}\\ \hat{\sigma}^2=\overline{x^2}-\bar{x}^2\end{cases}$$

下面分析其无偏性和一致性。根据之前的结论，$\mathbb{E}\hat{\mu}=\mu_0, \mathbb{E}\left(\hat{\sigma}^2\right)=\frac{N-1}{N}\sigma_0^2$，因而 $\hat{\mu}$ 是无偏估计量而 $\hat{\sigma}^2$ 并非无偏估计量。而由于 $\bar{x}\xrightarrow{p}\mu_0$，$\overline{x^2}\xrightarrow{p}\mu_0^2+\sigma_0^2$，从而 $\hat{\sigma}^2\xrightarrow{p}\mu_0^2+\sigma_0^2-\mu_0^2=\sigma_0^2$，因而 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 都是一致估计量。

3 极大似然估计

3.1 极大似然估计量

极大似然估计量（**maximum likelihood estimator**）是目前为止最常见的得到估计量的方法，其思想是，如果我们要对未知参数总体 P_θ 做推断，估计 θ_0，那么我们就寻找一个 $\hat{\theta}$，使得这组数据出现的概率最高，则 $\hat{\theta}$ 理应是 θ_0 的一个合理估计。

如果假设一组独立同分布的样本 $x=(x_1,...,x_N)$ 来自于参数总体 P_θ，且密度函数为 $f\left(x_i|\theta\right)$，那么样本的联合分布函数为：

$$f\left(x|\theta\right)=\prod_{i=1}^{N}f\left(x_i|\theta\right)$$

现在，将未知参数 θ 视为变量，x 为给定的样本，由于对数函数为单调函数，因而可以将联合分布函数取对数，得到对数似然函数（log-likelihood function）：

$$L\left(\theta|x\right)=\ln f\left(x|\theta\right)=\sum_{i=1}^{N}\ln f\left(x_i|\theta\right)$$

极大似然估计即找到一个 $\hat{\theta}$ 使得对数似然函数最大化：

$$\hat{\theta}=\arg\max_{\theta}L\left(\theta|x\right)$$

从而我们得到了极大似然估计量 $\hat{\theta}$。

例 7. 如果 $x_i\sim Ber\left(p_0\right)\ i.i.d$，那么其联合密度函数为：

$$f\left(x|p\right)=\prod_{i=1}^{N}p^{x_i}\left(1-p\right)^{1-x_i}$$

对数似然函数为:

$$\begin{aligned} L\left(p|x\right) &= \sum_{i=1}^{N}\left[x_i \ln p + \left(1-x_i\right)\ln\left(1-p\right)\right] \\ &= \left(\sum_{i=1}^{N} x_i\right)\ln p + \left(N - \sum_{i=1}^{N} x_i\right)\ln\left(1-p\right) \end{aligned}$$

现在欲得到 p 的极大似然估计值，只要对上述对数似然函数求最大值，即:

$$\frac{\partial L\left(p|x\right)}{\partial p} = \frac{\sum_{i=1}^{N} x_i}{p} - \frac{\left(N - \sum_{i=1}^{N} x_i\right)}{1-p} = 0$$

从而得到:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

现在讨论该估计量的无偏性和一致性。对于无偏性，我们有:

$$\mathbb{E}\left(\hat{p}\right) = \mathbb{E}\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}x_i = p_0$$

而对于一致性，根据大数定律，我们有:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i \xrightarrow{p} \mathbb{E}x_i = p_0$$

因而极大似然估计量 $\hat{p}$ 是真值 p_0 的无偏、一致估计量。

例 8. 如果 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d$，其中 $\theta = \left(\mu, \sigma^2\right)$，那么其联合密度函数为:

$$f\left(x|\theta\right) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right\}$$

对数似然函数为:

$$\begin{aligned} L\left(\theta|x\right) &= \sum_{i=1}^{N}\left[-\frac{1}{2}\ln\left(2\pi\right) - \ln\sigma - \frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right] \\ &= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i-\mu\right)^2 \\ &= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_i^2 + \mu^2 - 2\mu x_i\right) \\ &= -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{N\mu^2}{2\sigma^2} - \frac{N}{2\sigma^2}\overline{x^2} + \frac{N\mu}{\sigma^2}\bar{x} \end{aligned}$$

对其求极大值，得到：

$$\frac{\partial L\left(\theta|x\right)}{\partial\theta}=\begin{pmatrix}-\frac{\mu}{\sigma^2}+\frac{\bar{x}}{\sigma^2}\\-\frac{1}{\sigma}+\frac{\mu^2}{\sigma^3}+\frac{\overline{x^2}}{\sigma^3}-\frac{2\mu\bar{x}}{\sigma^3}\end{pmatrix}=0$$

解得：

$$\begin{cases}\hat{\mu}=\bar{x}\\\hat{\sigma}^2=\overline{x^2}-\bar{x}^2\end{cases}$$

这与矩估计的估计量是一样的。我们之前已经证明了，$\hat{\mu}$ 是 μ_0 的无偏、一致估计量，而 $\hat{\sigma}^2$ 是 σ_0^2 的一致估计量，$\frac{N}{N-1}\hat{\sigma}^2$ 是 σ_0^2 的无偏估计量。

例 9.（截尾数据）现在正在进行一项调查，其中一项调查为收入（y_i）调查，其中关于收入的问题为：

- 请问您的收入是多少?
  - 小于 1000
  - 大于 10000
  - 其他 _____（请填写具体数值）

如果假设收入的对数（$x_i^*=\log_{10}y_i$）服从正态分布，即 $x_i^*\sim N\left(\mu,\sigma^2\right)\ i.i.d$，那么我们观察到的数据为：

$$x_i=\begin{cases}3 & y_i\leq 1000\\4 & y_i\geq 10000\\x_i^* & otherwise\end{cases}$$

我们称数据存在截尾（censoring）现象。为了估计以上问题，我们可以计算：

$$P\left(x_i=3\right)=P\left(x_i^*\leq 3\right)=P\left(\frac{x_i^*-\mu}{\sigma}\leq\frac{3-\mu}{\sigma}\right)=\Phi\left(\frac{3-\mu}{\sigma}\right)$$

同理 $P\left(x_i=4\right)=1-\Phi\left(\frac{4-\mu}{\sigma}\right)$。因而 x_i 的密度函数为：

$$f\left(x_i|\theta\right)=\left[\Phi\left(\frac{3-\mu}{\sigma}\right)\right]^{1\{x_i=3\}}\left[1-\Phi\left(\frac{4-\mu}{\sigma}\right)\right]^{1\{x_i=4\}}\left[\phi\left(\frac{x_i-\mu}{\sigma}\right)\right]^{1\{3<x_i<4\}}$$

因而其对数似然函数为：

$$\begin{aligned}L\left(\theta|x\right)=&N_3\ln\Phi\left(\frac{3-\mu}{\sigma}\right)+N_4\ln\left[1-\Phi\left(\frac{3-\mu}{\sigma}\right)\right]\\&+\sum_{i=1}^{N}1\left\{3<x_i<4\right\}\ln\phi\left(\frac{x_i-\mu}{\sigma}\right)\end{aligned}$$

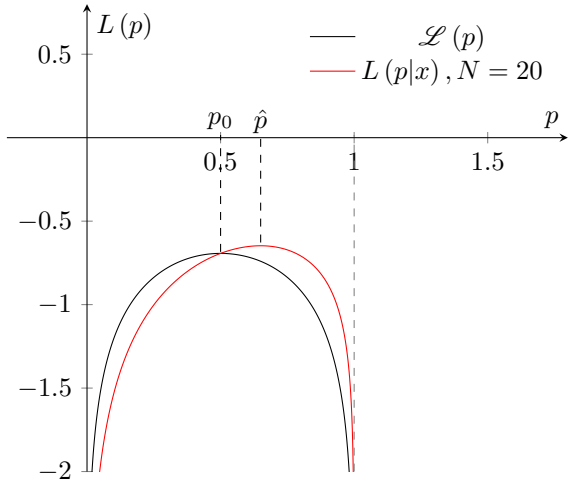


图 1: 伯努利分布的总体和样本似然函数

最大化以上对数似然函数，我们就得到了正态分布总体的极大似然估计。

3.2 一致性与 Kullback-Leiber 信息

在以上两个例子中，我们发现，尽管极大似然估计不能保证无偏性，但是所有的估计量都是一致的。那么是不是极大似然估计一定能保证一致性呢?

为了回答这个问题，我们分两步来看。首先，观察对数似然函数: $\frac{1}{N}L(\theta|x) = \frac{1}{N}\sum_{i=1}^{N}\ln f(x_i|\theta)$，对于任意一个给定的 θ（不一定是真值 θ_0），在一定的条件下，根据大数定律，有:

$$\frac{1}{N}L(\theta|x) \xrightarrow{p} \mathbb{E}\ln f(x_i|\theta) \triangleq \mathscr{L}(\theta)$$

即样本似然函数收敛到总体的似然函数[1]。其次，极大似然估计的方法是最大化 $L(\theta|x)$ 获得估计，即 $\hat{\theta} = \arg\max_\theta L(\theta|x)$，那么如果真值 $\theta_0 = \arg\max_\theta \mathscr{L}(\theta)$，由于 $\frac{1}{N}L(\theta|x) \xrightarrow{p} \mathscr{L}(\theta)$，那么样本似然函数最大值 $\hat{\theta}$ 应该也会无限趋向于总体似然函数最大值 θ_0，从而是 θ_0 的一致估计。那么接下来的问题就是，我们的真值 θ_0 是不是的确能够最大化总体似然函数 $\mathscr{L}(\theta) = \mathbb{E}\ln f(x_i|\theta)$ 呢? 我们先看一个例子。

例 10. 若 $x_i \sim Ber(p_0)$ $i.i.d$，那么其联合密度函数为:

$$f(x|p) = \prod_{i=1}^{N} p^{x_i}(1-p)^{1-x_i}$$

[1]注意严谨来说，在这里依概率收敛是不够的，我们需要一致收敛。

对数似然函数为:

$$L(p|x)=\sum_{i=1}^{N}\left[x_i\ln p+(1-x_i)\ln(1-p)\right]$$

根据大数定律，对于任意的 p，上述似然函数

$$\begin{aligned}\frac{1}{N}L(p|x)\xrightarrow{p}\mathscr{L}(p)&\triangleq\mathbb{E}\left[x_i\ln p+(1-x_i)\ln(1-p)\right]\\&=\mathbb{E}(x_i)\ln p+\mathbb{E}(1-x_i)\ln(1-p)\\&=p_0\ln p+(1-p_0)\ln(1-p)\end{aligned}$$

其中 p_0 为真值。那么接下来的问题是，是不是只有当 $p=p_0$ 时，$\mathscr{L}(p)$ 达到了最大值呢？为求最大值，我们对 $\mathscr{L}(p)$ 求导数并令其等于 0 得到:

$$\frac{\partial\mathscr{L}(p)}{\partial p}=\frac{p_0}{p}-\frac{1-p_0}{1-p}=0$$

从而只有当 $p=p_0$ 时，以上导数等于 0。因而真值 p_0 最大化了总体似然函数 $\mathscr{L}(p)$。

以上伯努利分布的例子并不是个例，实际上，我们可以证明，真值 θ_0 总是可以最大化总体似然函数。为了证明这一点，我们可以从总体似然函数出发:

$$\begin{aligned}\mathscr{L}(\theta)&=\mathbb{E}\ln f(x_i|\theta)\\&=\mathbb{E}_{\theta_0}\ln f(x_i|\theta)\\&=\int_{\mathbb{R}}\ln f(x|\theta)\cdot f(x|\theta_0)\,dx\end{aligned}$$

求其最大值，得到:

$$\begin{aligned}\frac{\partial\mathscr{L}(\theta)}{\partial\theta}&=\frac{\partial}{\partial\theta}\int_{\mathbb{R}}\ln f(x|\theta)\cdot f(x|\theta_0)\,dx\\&=\int_{\mathbb{R}}\frac{\partial}{\partial\theta}\ln f(x|\theta)\cdot f(x|\theta_0)\,dx\\&=\int_{\mathbb{R}}\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\cdot f(x|\theta_0)\,dx\end{aligned}$$

当 $\theta = \theta_0$ 时，有：

$$\begin{aligned}\frac{\partial \mathscr{L}(\theta)}{\partial \theta} &= \int_{\mathbb{R}} \frac{1}{f(x|\theta_0)} \frac{\partial f(x|\theta_0)}{\partial \theta} \cdot f(x|\theta_0)\, dx \\ &= \int_{\mathbb{R}} \frac{\partial f(x|\theta_0)}{\partial \theta} dx \\ &= \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x|\theta_0)\, dx \\ &= 0\end{aligned}$$

其中最后一步由于 $\int_{\mathbb{R}} f(x|\theta_0)\, dx = 1$。我们通常称对数似然函数的一阶导数 $\frac{\partial}{\partial \theta} \ln f(x_i|\theta)$ 为得分函数（score function），记为 $s_i(\theta) = \frac{\partial}{\partial \theta} \ln f(x_i|\theta)$。以上结论意味着得分函数的期望等于 0，即 $\mathbb{E}s_i(\theta_0) = 0$。我们知道，一阶导数等于 0 是最大化的必要条件，因而我们猜测，真值 θ_0 很有可能最大化了总体似然函数 $\mathscr{L}(\theta)$。

为了更进一步得到真值 θ_0 是否最大化了总体似然函数，我们引入 Kullback-Leiber 信息的概念。

定义 1. 令 P 和 Q 为同一概率空间中的两个概率函数，p 和 q 分别为其概率密度函数。Kullback-Leiber 信息被定义为：

$$\mathscr{K}(P, Q) = \int_{\mathbb{R}} \ln \frac{p(x)}{q(x)} p(x)\, dx$$

当 P 和 Q 属于参数族 P_θ 和 Q_η 时，Kullback-Leiber 信息即：

$$\mathscr{K}(\theta, \eta) = \int_{\mathbb{R}} \ln \frac{f(x|\theta)}{g(x|\eta)} f(x|\theta)\, dx = \mathbb{E}_\theta \left[\ln \frac{f(x|\theta)}{g(x|\eta)} \right]$$

实际上，Kullback-Leiber 信息度量的是两个概率函数的「距离」，可以证明，Kullback-Leiber 信息 $\mathscr{K}(P, Q) \geq 0$，当且仅当 $P = Q$ 时等号成立。

对于极大似然函数，给定任意一个 θ，其代表的概率函数与真值代表的概率函数之间的距离，即 Kullback-Leiber 信息为：

$$\mathscr{K}(\theta_0, \theta) = \mathbb{E}_{\theta_0} \left[\ln \frac{f(x|\theta_0)}{f(x|\theta)} \right] \geq 0$$

当且仅当 $\theta = \theta_0$ 时等式成立，因而 θ_0 最小化了：

$$\mathbb{E}_{\theta_0} \left[\ln \frac{f(x|\theta_0)}{f(x|\theta)} \right] = \mathbb{E}_{\theta_0} \left[\ln f(x|\theta_0) - \ln f(x|\theta) \right]$$

或者等价的，最大化了 $\mathbb{E}_{\theta_0} \left[\ln f(x|\theta) \right] = \mathscr{L}(\theta)$。

3.3 极限分布与 Fisher 信息

上一节中，我们知道在一定条件下，极大似然估计量 $\hat{\theta}$ 是真值 θ_0 的一致估计量，进一步的，我们希望知道估计量 $\hat{\theta}$ 的抽样分布。我们下面将从极大似然函数的一阶条件开始，使用 delta 方法得到估计量 $\hat{\theta}$ 的极限分布。

由于我们计算极大似然估计量时最大化了极大似然函数，其一阶条件为：

$$\frac{\partial L\left(\hat{\theta}|x\right)}{\partial\theta}=\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\ln f\left(x_i|\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=0$$

即样本得分函数的均值等于 0。我们对上式在 $\theta=\theta_0$ 处进行泰勒展开，得到：

$$0=\sum_{i=1}^{N}s_i\left(\hat{\theta}\right)=\sum_{i=1}^{N}s_i\left(\theta_0\right)+\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\left(\hat{\theta}-\theta_0\right)+O\left(\left(\hat{\theta}-\theta_0\right)^2\right)$$

我们记 $H_i\left(\theta\right)=\frac{\partial}{\partial\theta'}s_i\left(\theta\right)=\frac{\partial}{\partial\theta\partial\theta'}\ln f\left(x_i|\theta\right)$ 为对数似然函数的海塞矩阵，我们有：

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta'}s_i\left(\theta_0\right)\xrightarrow{p}\mathbb{E}H_i\left(\theta_0\right)\triangleq-H_0$$

而由于 $\mathbb{E}s_i\left(\theta_0\right)=0$，因而 $\mathrm{Var}\left(s_i\left(\theta_0\right)\right)=\mathbb{E}\left[s_i\left(\theta_0\right)s_i'\left(\theta_0\right)\right]\triangleq\mathcal{I}_0$，因而根据中心极限定理：

$$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)\xrightarrow{D}N\left(0,\mathcal{I}_0\right)$$

因而：

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=H_0^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}s_i\left(\theta_0\right)+o_p\left(1\right)\xrightarrow{D}N\left(0,H_0^{-1}\mathcal{I}_0H_0^{-1}\right)\qquad(2)$$

注意其中：

$$\begin{aligned}
-H_0 &= \mathbb{E}H_i\left(\theta_0\right) \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta\partial\theta'} \ln f\left(x|\theta_0\right) \cdot f\left(x|\theta_0\right) dx \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta'} \left[\frac{1}{f\left(x|\theta_0\right)} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta}\right] \cdot f\left(x|\theta_0\right) dx \\
&= \int_{\mathbb{R}} \left[\frac{1}{f\left(x|\theta_0\right)} \frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'} - \frac{1}{f^2\left(x|\theta_0\right)} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta'}\right] \cdot f\left(x|\theta_0\right) dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'} - \frac{1}{f\left(x|\theta_0\right)} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta'} dx \\
&= \int_{\mathbb{R}} \frac{\partial^2 f\left(x|\theta_0\right)}{\partial\theta\partial\theta'} dx - \int_{\mathbb{R}} \left(\frac{1}{f\left(x|\theta_0\right)} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta} \frac{\partial f\left(x|\theta_0\right)}{\partial\theta'}\right)^2 f\left(x|\theta_0\right) dx \\
&= \frac{\partial^2}{\partial\theta\partial\theta'} \int_{\mathbb{R}} f\left(x|\theta_0\right) dx - \mathbb{E}_{\theta_0} \frac{\partial \ln f\left(x|\theta_0\right)}{\partial\theta} \frac{\partial \ln f\left(x|\theta_0\right)}{\partial\theta'} \\
&= -\mathbb{E}_{\theta_0}\left[s_i\left(\theta_0\right) s_i\left(\theta_0\right)'\right] \\
&= -\mathrm{Var}\left[s_i\left(\theta_0\right)\right] \\
&= -\mathcal{I}_0
\end{aligned}$$

因而我们有：$H_0 = \mathcal{I}_0$，即对数似然函数海塞矩阵的期望等于得分函数的方差。将以上等式带入式 (2)，可以得到：

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N\left(0, \mathcal{I}_0^{-1}\right)$$

即大样本条件下，极大似然估计量的极限分布为正态分布，且其渐进方差为对数似然函数海塞矩阵倒数的逆矩阵。特别的，当 θ 为一维标量时，

$$\sqrt{N}\left(\hat{\theta} - \theta_0\right) \xrightarrow{D} N\left(0, -\frac{1}{\mathbb{E}\left(\frac{d^2}{d\theta^2} \ln f\left(x|\theta\right)\right)}\right)$$

例 11. 在例 (7) 中，我们已经得到伯努利分布的极大似然估计为：

$$\hat{p} = \bar{x}$$

根据中心极限定理，$\hat{p} = \bar{x} \xrightarrow{D} N\left(p_0, \frac{p_0(1-p_0)}{N}\right)$。另一方面，直接使用以上结论也可以得到 $\hat{p}$ 的极限分布。注意对数似然函数为：

$$L\left(p|x\right) = \sum_{i=1}^{N} \left[x_i \ln p + \left(1 - x_i\right) \ln\left(1 - p\right)\right]$$

因而得分函数为：

$$s_i\left(p\right) = \frac{x_i}{p} - \frac{1 - x_i}{1 - p}$$

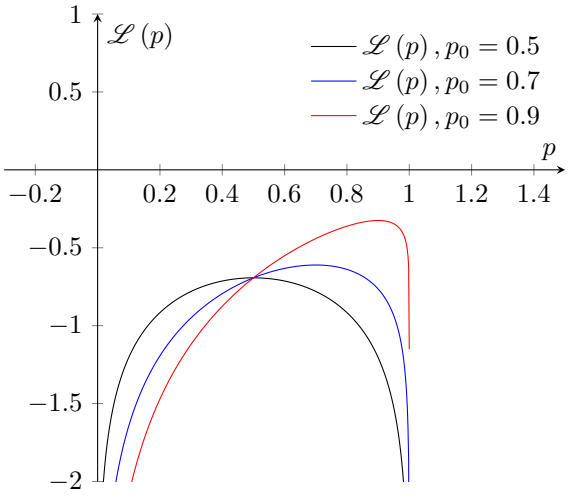


图 2: 不同 p_0 下伯努利分布的总体似然函数

进而海塞矩阵（二阶导）：

$$-H_i(p) = -\frac{x_i}{p^2} - \frac{1-x_i}{(1-p)^2}$$

因而带入真值之后其期望为：

$$\mathcal{I}_0 = H_0 = \mathbb{E}H_i(p_0) = \frac{1}{p_0} + \frac{1}{1-p_0} = \frac{1}{p_0(1-p_0)}$$

从而 $H_0^{-1} = p_0(1-p_0)$，从而 $\sqrt{N}(\hat{p}-p) \sim N(0, p_0(1-p_0))$。

实际上，可以证明，以上的 $\mathcal{I}_0^{-1}$ 是渐进无偏估计量所能达到的最小方差，我们称方差为 $\mathcal{I}_0^{-1}$ 的估计量为渐进有效（asymptotic efficient）估计量。而 $\mathcal{I}_0$ 实际上度量了数据中所包含的「信息量」的大小，因而我们称 $\mathcal{I}_0$ 为**费雪信息矩阵**（**Fisher information matrix**）。$\mathcal{I}_0$ 越大，意味着所包含的信息越多，极大似然估计所得到的方差也越小。如图 (2) 所示，当 $p_0 = 0.5$ 时，信息矩阵 $\mathcal{I}_0$ 达到了最小值，而 $\hat{p}$ 的方差达到了最大值，表现在图中即对数似然函数在真值 p_0 处非常平缓。当真值 p_0 逐渐接近 0 或者 1 时，信息矩阵 $\mathcal{I}_0$ 逐渐变大，$\hat{p}$ 的方差也逐渐变小，图中对数似然函数在真值 p_0 处也更加尖锐。因而同样是伯努利分布，真值越接近于 0 或者 1 的伯努利分布实际上携带了更多的信息。

3.4 条件极大似然估计

以上介绍了极大似然估计法，需要设定数据 X 的完整的分布情况才能得到估计。然而很多时候，我们观察到一系列数据 $w_i \in \mathbb{R}^k, i = 1, ..., N$，其中 $w_i = (y_i', x_i')', y_i \in \mathbb{R}^{k_1}, x_i \in \mathbb{R}^{k_2}$，很多时候我们仅仅希望研究 x 和 y 之间的关

系，而不关心随机向量 x 之间的关系，如果使用极大似然估计，我们就必须设定 x 的联合分布。然而，设定 x 的联合分布很多时候是多于的，实际上，如果我们能够找到 y 给定 x 的条件分布，即 $f(y|x,\theta)$，那么基于条件分布的极大似然估计仍然能够得到参数 θ 的一致估计。

例 12. （线性回归）如果 $(y_i, x_i')', i = 1, ..., N, x_i \in \mathbb{R}^K$ 为一系列独立同分布的随机向量。为了使用 x_i 预测或者拟合 y_i，我们可以假设 $y_i|x_i \sim N\left(x_i'\beta, \sigma^2\right)$，即给定 x，y 服从正态分布[2]。或者等价的，以上关系可以写成：

$$y_i = x_i'\beta + u_i$$

其中 $u_i \sim N\left(0, \sigma^2\right)$。在上述条件下，条件密度函数为：

$$f\left(y_i|x_i,\beta\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{\sigma^2}\right\}$$

因而似然函数为：

$$L\left(\beta|y,x\right) = -\frac{N}{2}\ln\left(2\pi\right) - N\ln\sigma - \frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\beta\right)^2$$

如果对 β 求导，可以得到：

$$-\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(y_i - x_i'\hat{\beta}\right)x_i = 0$$

解得：

$$\hat{\beta} = \left[\sum_{i=1}^{N}\left(x_i x_i'\right)\right]^{-1}\sum_{i=1}^{N}\left(x_i y_i\right)$$

如果令：

$$X = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}$$

以及：

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

[2]注意尽管 $y|x$ 服从正态分布，y 有可能不服从正态分布。

那么 $\hat{\beta}=(X'X)^{-1}X'Y$。以上就是所谓的最小二乘估计（Ordinary least squares, OLS）。特别的，如果 $K=2$，$x_{i1}=1$，即解释变量中存在截距项（常数项），以及一个额外的解释变量，我们称这种情况为一元线性回归，可以得到：

$$\hat{\beta}_2=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$
$$\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x}$$

例 13. （Logistic 回归）如果 $(y_i,x_i')',i=1,...,N,x_i\in\mathbb{R}^K$ 为一系列独立同分布的随机向量，而其中 y_i 为二元变量，即 $y_i\in\{0,1\}$，那么此时使用线性回归模型就不再适合。如果假设

$$p(y_i=1|x_i,\beta)=\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$$

那么条件密度函数为：

$$f(y_i|x_i,\beta)=\left[\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]^{1\{y_i=1\}}\left[\frac{1}{1+e^{x_i'\beta}}\right]^{1\{y_i=0\}}$$

因而极大似然函数为：

$$L(\beta|y,x)=\sum_{i=1}^{N}\left[1\{y_i=1\}\ln\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)+1\{y_i=0\}\ln\left(\frac{1}{1+e^{x_i'\beta}}\right)\right]$$

最大化以上似然函数，就可以得到 β 的一致估计，进而得到 $p(x_i)=P(y_i=1|x_i)$，即给定 x_i，$y_i=1$ 的概率的估计。

4 区间估计

上一节中我们讨论了参数的点估计。尽管我们可以使用点估计方法对参数值进行推断，然而我们知道，参数的点估计值 $\hat{\theta}$ 与真值 θ_0 相等的概率一般为 0，即 $P\left(\hat{\theta}=\theta_0\right)=0$。因而更进一步的，我们很多时候希望得到一个区间，使得这个区间能够以正的概率包含真值 θ_0。这就诞生了区间估计（interval estimation）的概念。

区间估计，即对于样本 $x=(x_1,...,x_N)$，通过一对统计量 $L(x)$ 和 $U(x)$，满足 $L(x)\leq U(x)$，我们可以使用区间 $[L(x),U(x)]$ 对未知参数 θ_0 进行推断。

例 14. 如果样本 $x_i\sim N(\mu_0,1)$ $i.i.d,i=1,...,N$，那么区间 $[\bar{x}-0.5,\bar{x}+0.5]$

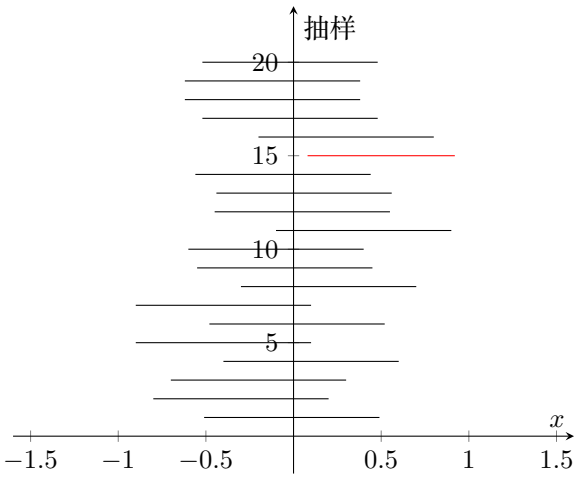


图 3: 不同抽样下长度为 1 的置信区间估计

包含真值 μ_0 的概率为:

$$\begin{aligned} P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) &= P\left(\mu_0 - 0.5 \leq \bar{x} \leq \mu_0 + 0.5\right) \\ &= P\left(-\frac{0.5}{\sqrt{\frac{1}{N}}} \leq \frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}} \leq \frac{0.5}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

由于 $\bar{x} \sim N\left(\mu_0, \frac{1}{N}\right)$，因而

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}} \sim N(0,1)$$

因而

$$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - \Phi\left(-\frac{0.5}{\sqrt{\frac{1}{N}}}\right) = 2\Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right) - 1$$

例如，当 $N = 16$ 时，查表可得，$P\left(\mu_0 \in [\bar{x}-0.5, \bar{x}+0.5]\right) = 2\Phi(2) - 1 \approx 2 \times 0.9772 - 1 = 0.9544$，即区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 包含真值 μ_0 的概率为 95.44%。如图 (3) 画出了当 $\mu_0 = 0$ 时，20 次不同抽样的区间。平均而言，每抽 100 次大概有 5 次区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 不能包含真实的参数 μ_0（图中红色区间）。

注意由于未知参数 θ_0 是一个未知的常数，而统计量 $L(x)$ 和 $U(x)$ 是随着抽样的变化而变化的，因此我们不能说「θ_0 落入区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 的概率是多少」，而只能说「区间 $[\bar{x}-0.5, \bar{x}+0.5]$ 包含 θ_0 的概率是多少」。我们把概率 $P\left(\theta_0 \in [L(x), U(x)]\right)$ 称为覆盖概率（coverage probability）。注意由于总体参数 θ_0 未知，因而概率 $P\left(\mu_0 \in [L(x), U(x)]\right)$ 可能依赖于未知的参数 θ_0，

因而我们通常将而覆盖概率的下界，即 $\inf_\theta P_\theta\left(\theta_0 \in [L(x), U(x)]\right)$ 称为**置信度**（**confidence coefficient**）或者**置信水平**，通常用 $1-\alpha$ 表示。在某一置信度下，区间 $[L(x), U(x)]$ 又被称为**置信区间**（**confidence interval**）。因而在例 (14) 中，我们可以说在 95.44% 的置信水平下，置信区间为 $[\bar{x}-0.5, \bar{x}+0.5]$。

此外还需要注意的是，在例 (14) 中，为了求得置信区间和覆盖概率，我们首先将统计量 $\bar{x}$ 做了标准化处理，即使用 $\frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}}$ 推算概率，而不是直接使用 $\bar{x}$。使用 $\frac{\bar{x}-\mu_0}{\sqrt{\frac{1}{N}}}$ 的好处是，此统计量不依赖于任何未知参数，因而其分布不会随着未知参数的变化而变化，即服从一个标准的分布，这样一来，我们得到的覆盖概率不依赖于任何未知参数，因而就等于置信度。一般的，我们把分布不依赖于未知参数的统计量成为**基准统计量**（**pivotal statistic**）。

例 15. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i=1,...,N$，那么：

1. 统计量 $\bar{x} \sim N\left(\mu_0, \frac{\sigma_0^2}{N}\right)$，其分布依赖于两个未知参数；
2. 统计量 $\bar{x}-\mu_0 \sim N\left(0, \frac{\sigma_0^2}{N}\right)$，其分布仍然依赖于未知参数 μ_0；
3. 统计量
$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{\sigma_0^2}{N}}} \sim N(0,1)$$
分布不依赖于任何未知参数，因而是基准统计量；
4. 统计量
$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \sim t_{N-1}$$
分布不依赖于任何未知参数，因而是基准统计量；

例 16. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i=1,...,N$，那么

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi^2_{N-1}$$

，其分布不依赖于任何未知参数，因而是基准统计量。

在例 (14) 中，我们首先给出了区间，进而计算了该区间的置信度。然而现实中，我们经常希望得到在一定置信水平下的置信区间，即一般的区间估计过程。

例 17. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i=1,...,N$，为了得到 μ_0 的 95% 的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有 μ_0 是未知的，其他都是已知的（包括已知常数以及已知统计量）。在例 (15) 中，只有统计量

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \sim t_{N-1}$$

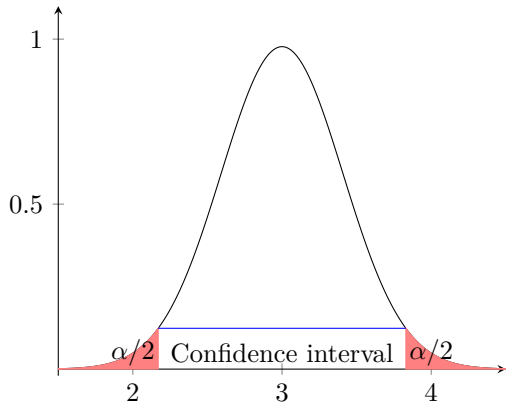


图 4: 正态分布或 t 分布置信区间

满足以上条件。如果令 $t_{\alpha/2} = F_t^{-1}\left(\frac{\alpha}{2}\right)$，我们有：

$$\begin{aligned} P\left(-t_{\alpha/2} \leq \frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \leq t_{\alpha/2}\right) &= F_t\left(t_{\alpha/2}\right) - F_t\left(-t_{\alpha/2}\right) \\ &= 1 - 2F_t\left(t_{\alpha/2}\right) \\ &= 1 - 2F_t\left(F_t^{-1}\left(\frac{\alpha}{2}\right)\right) \\ &= 1-\alpha \end{aligned}$$

因而我们可以得到：

$$P\left(\bar{x} - t_{\alpha/2}\sqrt{\frac{s^2}{N}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right) = 1-\alpha$$

从而 $\left[\bar{x} - t_{\alpha/2}\sqrt{\frac{s^2}{N}}, \bar{x} + t_{\alpha/2}\sqrt{\frac{s^2}{N}}\right]$ 就是我们想要的置信区间。例如，对于一个 $N = 30$ 的正态样本，$\bar{x} = 3$，$s^2 = 5$，如果我们想要得到 95% 置信水平下的置信区间，查表$(d.f. = 29)$得到 $t_{\alpha/2} = 2.0452$，因而置信下界为 $3 - 2.0452 \times \sqrt{5/30} \approx 2.17$，置信上界为 $3 + 2.0452 \times \sqrt{5/30} \approx 3.83$。如图 (4) 所示，其中红色区域为左右两个概率为 $\alpha/2$ 的区域，中间的一块即为所要求的置信区间。

例 18. 如果样本 $x_i \sim N\left(\mu_0, \sigma_0^2\right)$ $i.i.d, i = 1, ..., N$，为了得到 σ_0^2 的 95% 的置信区间，我们首先找到基准统计量，要求在基准统计量中，只有 σ_0^2 是未知的，其他都是已知的。在例 (16) 中，统计量

$$(N-1)\frac{s^2}{\sigma_0^2} \sim \chi_{N-1}^2$$

满足以上条件。如果令 $\chi^2_{\alpha/2} = F^{-1}_{\chi^2}\left(\frac{\alpha}{2}\right)$，$\chi^2_{1-\alpha/2} = F^{-1}_{\chi^2}\left(1-\frac{\alpha}{2}\right)$，我们有：

$$P\left(\chi^2_{\alpha/2} \leq (N-1)\frac{s^2}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}\right) = 1-\alpha$$

因而：

$$P\left(\frac{(N-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma_0^2 \leq \frac{(N-1)s^2}{\chi^2_{\alpha/2}}\right) = 1-\alpha$$

σ_0^2 的 95% 的置信区间为：$\left[\frac{(N-1)s^2}{\chi^2_{1-\alpha/2}}, \frac{(N-1)s^2}{\chi^2_{\alpha/2}}\right]$。

总结上述两个置信区间的计算，一般而言我们得到置信区间的步骤如下：

1. 给定置信度 $1-\alpha$；
2. 找到一个基准统计量，其中只有所要求的参数是未知的，其他都是已知的；
3. 找到这个基准统计量的分布函数 $F(\cdot)$；
4. 查表或使用计算机计算 $F^{-1}\left(\frac{\alpha}{2}\right)$ 以及 $F^{-1}\left(1-\frac{\alpha}{2}\right)$；
5. 通过不等式变换得到置信区间。

因而，计算置信区间最关键的步骤即找到基准统计量，并得到此统计量的分布。

尽管上两例给出了正态总体的均值和方差的置信区间的计算方法，然而很多时候我们的总体并不是一定来自于正态总体，很多时候我们很难计算在非正态总体下样本均值的精确分布。然而根据中心极限定理，在一定条件下，有：

$$\sqrt{N}(\bar{x}-\mu_0) \overset{a}{\sim} N(0, \mathrm{Var}(x))$$

因而大样本条件下，我们可以使用中心极限定理近似样本均值的分布，从而得到区间估计。

例 19. 根据 2009 年中国城镇住户调查，在 37480 户家庭中，已知家庭年收入均值为 54157.63 元，标准差为 38533.96 元，那么全国家庭家庭平均收入的 95% 置信区间是多少？首先一般而言，收入一般不服从正态分布，但是在大样本条件下，我们知道样本均值近似服从正态分布，因而：

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{N}}} \overset{a}{\sim} N(0,1)$$

为基准统计量。如果记 $z_{\alpha/2} = \Phi_t^{-1}\left(\frac{\alpha}{2}\right)$，查表知 $z_{2.5\%} = 1.96$，因而置信下界为：$54157.63 - 1.96 \times \sqrt{\frac{38533.96^2}{2963}} \approx 53766.88$，同理置信上界约为 54547.12，因而全国家庭家庭平均收入的 95% 置信区间为 $[53766.88, 54547.12]$。

例 20. 根据 2013 年中国家庭金融调查，样本 7711 户家庭中，有 6% 的家庭有信用卡，请问全国持有信用卡的家庭比例的 95% 置信区间是多少？同样的，比例一般不服从正态分布，但是如果把每个家庭是否持有信用卡假设为独立同分布的伯努利分布，即 $x_i \sim Ber(p_0)$，那么 $x_i^2 = x_i$，因而 $\overline{x^2} = \bar{x}$，从而：

$$\frac{s^2}{N} = \frac{N-1}{N}\frac{\overline{x^2} - \bar{x}^2}{N} \approx \frac{\overline{x^2} - \bar{x}^2}{N} = \frac{\bar{x} - \bar{x}^2}{N} = \frac{\bar{x}(1-\bar{x})}{N}$$

其中比例 $\hat{p} = \bar{x}$，从而

$$\frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}} \overset{a}{\sim} N(0,1)$$

从而置信下界为 $\hat{p} - z_{2.5\%}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} = 0.06 - 1.96 \times \sqrt{\frac{0.06 \times (1-0.06)}{7711}} \approx 5.47\%$，同理置信上界约为 6.53%，因而全国家庭持有信用卡比例的 95% 置信区间为 $[5.47\%, 6.53\%]$。

此外，很多时候我们还对两个样本的差值感兴趣。如果假设两个独立的样本 x_1 和 x_2，其均值分别为 $\bar{x}_1$ 和 $\bar{x}_2$，且 $x_{1i} \sim N(\mu_1, \sigma_1^2)$，$x_{2i} \sim N(\mu_2, \sigma_2^2)$，那么 $\bar{x}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{N_1}\right)$，$\bar{x}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{N_2}\right)$，其差值：

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)$$

因而可以使用以上分布对 $\mu_1 - \mu_2$ 进行区间估计。即使两个样本不来自于正态总体，仍然可以使用上面的中心极限定理，通过渐进正态性得到相似的结论。

例 21. 在 2009 年中国城镇住户调查中，共有 23440 位 20-50 岁的男性，以及 21184 位 20-50 岁的女性。已知男性年平均收入为 28367.96 元，标准差为 21811.88 元；女性年平均收入为 20145.77 元，标准差为 16541.08 元。如果假设男女收入独立，请问男女收入差异的 95% 置信区间是多少？同上，尽管收入不服从正态分布，但是大样本情况下可以使用正态分布近似。从而：

$$\bar{x}_1 - \bar{x}_2 \overset{a}{\sim} N\left(28367.96 - 20145.77, \frac{21811.88^2}{23440} + \frac{16541.08^2}{21184}\right)$$

因而其差值的置信区间为 $[7864.99, 8579.38]$。

最后，根据区间估计，我们还能确定为了达到某一精度所需要样本量的大小。根据中心极限定理，在一定条件下，有：

$$\sqrt{N}(\bar{x} - \mu_0) \overset{a}{\sim} N(0, \mathrm{Var}(x))$$

因而大样本条件下，均值在 $1-\alpha$ 置信水平下的置信区间为：$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right]$，区间长度应为：$\frac{2z_{\alpha/2}\sigma}{\sqrt{N}}$。可以看到，区间大小随着样本量的增加而减小。如果我们

要求在 $1-\alpha$ 置信水平下的置信区间的长度为 l，那么样本量应为 $N=\left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2$，即样本量与精度成二次方关系。

例 22. 在例 (20) 中，为了使 95% 置信区间长度不超过 1%，需要的样本量为：

$$N=\left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2=\left[\frac{2\times 1.96\times\sqrt{0.06\times(1-0.06)}}{0.01}\right]^2\approx 8667$$

即需要 8667 户样本。然而实际情况中，我们不太可能知道 σ，所以有时预调研是非常重要的。不过在这个例子中，我们可以得到一个样本量的上界：

$$N=\left[\frac{2z_{\alpha/2}\sigma}{l}\right]^2=\left[\frac{2z_{\alpha/2}\sqrt{p(1-p)}}{l}\right]^2\leq\left[\frac{2z_{\alpha/2}\sqrt{0.5(1-0.5)}}{l}\right]^2=\left[\frac{z_{\alpha/2}}{l}\right]^2$$

即在这个例子中，如果我们不知道持有信用卡的家庭约为 6%，那么需要的样本数量上界为 38416 户家庭。

习题

1. 计算例 (1) 中两个估计量的 MSE。
2. 求以下分布总体的矩估计，并验证其无偏性和一致性。
   (a) $x_i\sim Ber(p)$
   (b) $x_i\sim N\left(\mu,\sigma^2\right)$
   (c) $x_i\sim P(\lambda)$
3. 若 $x_i\sim U(a,b)$ $i.i.d$，求其矩估计，并验证其一致性。
4. 求以下分布总体的极大似然估计，证明其一致性并计算估计量的极限分布。
   (a) $x_i\sim P(\lambda)$
   (b) $x_i\sim N(\mu,1)$
   (c) $x_i\sim N\left(0,\sigma^2\right)$
5. 若 $x_i^*\sim N\left(\mu,\sigma^2\right)$，但是当 $x_i^*\leq 100$ 时，我们观察不到 x_i^*，即我们观察到 x_i 满足：
   $$x_i=\begin{cases}100 & x_i^*\leq 100\\ x_i^* & otherwise\end{cases}$$
   请写出以上问题的对数似然函数。

6. 编程题：若 $x_i \sim Beta(\alpha, \beta)$ $i.i.d$，请写出其矩估计和极大似然估计的实现。(Beta 分布随机数可以使用 random.betavariate(alpha, beta) 来生成)。

参考文献

[1] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[2] Schervish, M.J., 1995. Theory of Statistics. Springer-Verlag, New York.

[3] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.

第八节 · 假设检验

司继春

上海对外经贸大学统计与信息学院

在上一节中，我们讨论了对于未知总体参数 θ 的参数估计问题，包括点估计和区间估计。很多时候，我们不仅仅需要回答总体参数 θ 是多少的问题，或者 θ 在什么区间范围以内的问题，还要回答「是不是」的问题，比如 $\theta = \theta_0$ 是不是成立？例如，在上一节中，我们发现在中国城镇住户调查的数据中，男性收入平均比女性年收入多 8222 元，那么我们是不是可以推断总体男性收入的确比女性高呢？还是仅仅因为抽样的巧合导致了我们计算的男性收入比女性收入高？我们可以使用**假设检验**（**Hypothesis testing**）的方法回答此类问题。

1 假设检验

为了讨论假设检验的问题，我们首先介绍「假设（hypothesis）」的概念。在假设检验中，假设指的是关于总体参数 θ 的一个命题。比如，对于不同总体，我们可能有如下假设：

1. 山东成年男性的平均身高为 175cm（$\theta = \theta_0$）
2. 某生产线次品率控制在 0.1% 范围内（$\theta \leq \theta_0$）
3. 北方人平均身高高于南方人（$\theta_1 \geq \theta_2$）

以上的例子都是关于未知总体参数的一些猜想。由于总体参数是未知的，我们只能观察到样本，因而我们不能确切的知道以上命题究竟是否成立，而只能使用样本对以上命题进行推断。

需要注意的是，这里的假设（hypothesis）与数学命题中的假设（assumption）是不同的。假设检验中的假设是我们要验证或者推翻的某个命题，而数学命题中的假设则是结论的前提条件。

假设检验中有两个互补的假设：**原假设**（**null hypothesis**）和**备择假设**（**alternative hypothesis**），分别用 H_0 和 H_1 来表示。如果参数的范围为 Θ，即总体参数 $\theta \in \Theta$，而原假设为 $\theta \in \Theta_0$，那么备择假设即为原假设的补集，即 $\theta \in \Theta_0^c$。比如，如果 $\Theta = \mathbb{R}$，若原假设是 $\theta = \theta_0$，那么备择假设就是 $\theta \neq \theta_0$；若原假设是 $\theta \geq \theta_0$，那么备择假设就是 $\theta < \theta_0$。注意原假设一般包含等号。

假设检验的过程就是使用数据作为「证据」试图推翻原假设的过程，这个过程与法官判案的过程类似。在法律中，有所谓的「无罪推定原则（presumption

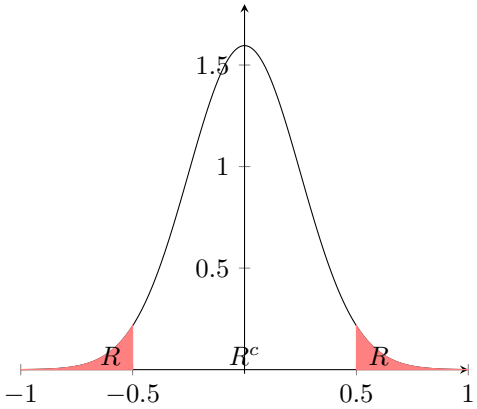


图 1: 拒绝域与第 I 类错误的概率

of innocence)」，即对于犯罪嫌疑人，必须先假设其无罪，原告方有义务提出证据证明其犯罪，而不得强迫嫌疑人自证其罪。使用以上术语，即原假设（H_0）为被告无罪，备择假设（H_1）为被告有罪，假设检验的目的就是使用证据（样本数据）试图推翻原假设（无罪）。如果现有证据可以推翻原假设，那么我们称为拒绝原假设（rejecting H_0），即可以认为原假设为假；而如果现有证据不能推翻原假设，即没有充足的证据证明原假设为假，那么我们称不能拒绝原假设（not rejecting H_0）。注意「接受原假设（accepting H_0）」的说法与「不能拒绝原假设」的说法有细微差别，如果不能拒绝原假设，可能是由于我们的证据不够充分，因而「不能拒绝原假设」的说法更加准确。基于上述原因，我们一般会把想要推翻的结论放在原假设上。

由于统计方法总会存在误差，因而基于以上两类假设的推断也会存在犯错的可能性。在假设检验中，有两种错误可能会发生：

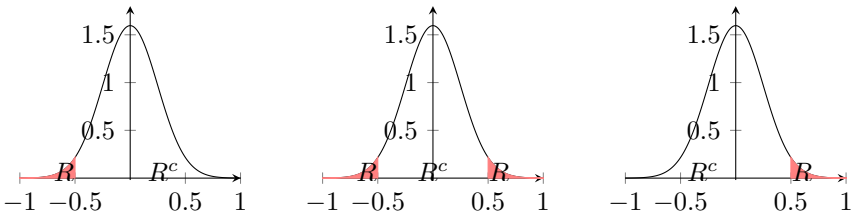
1. 第 I 类错误：原假设为真，但是拒绝原假设，即「弃真错误」；

2. 第 II 类错误：备择假设为真，但是接受原假设，即「取伪错误」。

比如，如果一个被告本来无罪，但是错误地判其有罪，那么就犯了第 I 类错误；而如果一个被告的确犯罪，但是却判其无罪，那么就犯了第 II 类错误。

假设检验一般通过设定一个检验统计量 $T(x)$，以及一个拒绝域 R，当 $T(x) \in R$ 时拒绝原假设。如果记 $H_0 : \theta \in \Theta_0$，而 $H_1 : \theta \in \Theta_0^c$，那么第 I 类错误，即原假设为真但是拒绝原假设的概率为：$P_{\theta \in \Theta_0}(T(x) \in R)$。

例 1. 如果样本 $x_i \sim N(\mu_0, 1)$ $i.i.d, i = 1, ..., N$，μ 为未知总体参数。如果原假设为 $H_0 : \mu_0 = 0$，备择假设为 $H_1 : \mu_0 \neq 0$，那么 $\Theta_0 = \{0\}$。令检验统计量 $T(x) = \bar{x}$，如果在原假设条件下，即假设 $\mu_0 = 0$，那么 $\bar{x} \sim N\left(0, \frac{1}{N}\right)$，即如果原假设成立，那么样本均值应该分布在 0 附近。而如果我们看到了样本均值距离 0 比较远，那么说明原假设有可能是不成立的。如图 (1) 所示，如果取拒绝



$H_0 : \mu \geq \mu_0$ 左侧检验　　$H_0 : \mu = \mu_0$ 双侧检验　　$H_0 : \mu \leq \mu_0$ 右侧检验

图 2: 单侧检验与双侧检验

域为 $R = (-\infty, -0.5) \cup (0.5, \infty)$，那么：

$$\begin{aligned} P_{\theta_0=0}(\bar{x} \in R) &= P_{\theta_0=0}(|\bar{x}| > 0.5) \\ &= P_{\theta_0=0}\left(\left|\frac{\bar{x}}{\sqrt{\frac{1}{N}}}\right| > \frac{0.5}{\sqrt{\frac{1}{N}}}\right) \\ &= 2\left(1 - \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right)\right) \end{aligned}$$

如果令 $N = 16$，那么 $P_{\theta_0=0}(\bar{x} \in R) = 2(1 - \Phi(2)) \approx 4.56\%$。注意以上概率是我们在原假设的假设上进行计算的，这意味着，如果原假设成立，那么得到均值落在拒绝域，即 $\bar{x} \in R = (-\infty, -0.5) \cup (0.5, \infty)$ 的概率为 4.56%。这样，如果我们采取「如果均值落入拒绝域就拒绝原假设」这一策略，那么在原假设的条件下，我们犯错的概率就是 4.56%，即犯第 I 类错误的概率为 4.56%。

此外，如图 (2) 所示，根据原假设的不同，检验还可以分为单侧检验和双侧检验。在上例中我们讨论的是双侧检验，并计算了给定拒绝域的情况下犯第 I 类错误的概率。然而如果原假设是不等于号，我们通常不能精确计算犯第 I 类错误的概率。此时，值得关注的是犯第 I 类错误的概率的上界，即 $\sup_{\theta \in \Theta_0} P_\theta(T(x) \in R)$。

例 2. 如果样本 $x_i \sim N(\mu_0, 1)$ $i.i.d, i = 1, ..., N$，μ 为未知总体参数。如果原假设为 $H_0 : \mu_0 \leq 0$，那么当 $\bar{x}$ 足够大时，可以拒绝原假设。在原假设的条件下，

当 $\mu_0 = \mu < 0$ 时，如果取拒绝域为 $R = (0.5, \infty)$，那么：

$$\begin{aligned} P_\mu(\bar{x} \in R) &= P_\mu(\bar{x} > 0.5) \\ &= P_\mu\left(\frac{\bar{x} - \mu}{\sqrt{\frac{1}{N}}} > \frac{0.5 - \mu}{\sqrt{\frac{1}{N}}}\right) \\ &= 1 - \Phi\left(\frac{0.5 - \mu}{\sqrt{\frac{1}{N}}}\right) \end{aligned}$$

注意以上概率随着 μ 的增加而增加，由于 $\mu \in \Theta_0 = (-\infty, 0]$，因而其上界：

$$\sup_{\theta \in \Theta_0} P_\theta(T(x) \in R) = 1 - \Phi\left(\frac{0.5 - 0}{\sqrt{\frac{1}{N}}}\right) = 1 - \Phi\left(\frac{0.5}{\sqrt{\frac{1}{N}}}\right)$$

当 $N = 16$ 时，上式为 $1 - \Phi(2) \approx 2.28\%$，即在原假设 $H_0 : \mu_0 \leq 0$ 的假设下，错误拒绝原假设的概率上界为 2.28%，或者等价的，犯第 I 类错误的概率上界为 2.28%。

在假设检验中，我们希望控制犯第 I 类错误的概率进行推断，即给定一个 α，找到一个拒绝域 R_α，使得 $\sup_{\theta \in \Theta_0} P_\theta(T(x) \in R_\alpha) \leq \alpha$。如此，我们便保证了使用拒绝域 R_α 进行假设检验，犯第 I 类错误的概率不超过 α。我们称 α 为**显著性水平**（**level of significance**），一般取 $\alpha = 0.01, 0.05, 0.1$，而 R_α 为一个区间，区间的断点成为临界值（critical value）。更小的 α 代表我们对犯第 I 类错误更加不能容忍，因而我们会更大的概率接受原假设，即拒绝域也会越小。

此外，我们还可以定义 p 值的概念。p 值指的是，给定检验统计量 $T(x) = t$，在原假设的条件下，能够取到 t 或者比 t 更加极端的值的概率。或者等价的，p 值可以被定义为能够拒绝原假设的最小的显著性水平，即给定 $T(x) = t$：

$$p = \inf\{\alpha \in (0, 1) : t \in R_\alpha\}$$

由于检验统计量 $T(x)$ 为随机变量，因而 $p(x) = \inf\{\alpha \in (0, 1) : T(x) \in R_\alpha\}$ 也是一个随机变量。当 $p \leq 0.01, 0.05, 0.1$ 时，可以拒绝原假设。如图 (3) 所示，绿色区域面积即 p 值。

综上，一般的假设检验的步骤可以总结如下：

1. 确定原假设 H_0 和备择假设 H_1；
2. 找到一个检验统计量 $T(x)$（通常为基准统计量）；
3. 确定检验统计量 $T(x)$ 在原假设 H_0 下的分布；
4. 设定显著性水平 α，并根据 α 确定拒绝域 R_α，若 $T(x) \in R_\alpha$ 则在 α 的显著性水平下拒绝原假设；或者根据 $T(x)$ 计算 p 值，若 $p < \alpha$ 则拒绝原

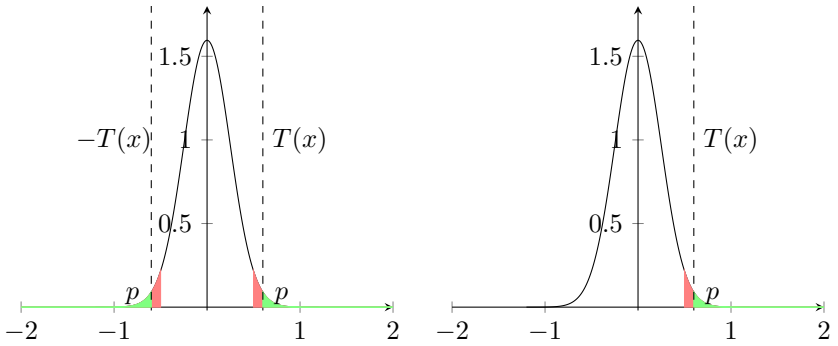


图 3: p 值的定义

假设。

例 3. 如果样本 $x_i \sim N\left(\mu, \sigma^2\right)\ i.i.d, i = 1, ..., N$，为了检验 $H_0 : \mu = \mu_0$，在 H_0 的假定下，$\bar{x} \sim N\left(\mu_0, \frac{\sigma^2}{N}\right)$，因而可以构建统计量：

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} \sim t_{N-1}$$

由于是双侧检验，因而临界值为 $t_{\alpha/2}$，拒绝域为 $R_\alpha = \left(-\infty, -t_{\alpha/2}\right) \cup \left(t_{\alpha/2}, \infty\right)$，当 $|t| > t_{\alpha/2}$ 时拒绝原假设，即认为 $\mu \neq \mu_0$，否则不能拒绝原假设。

例 4. 如果样本 $x_{1i} \sim N\left(\mu_1, \sigma_1^2\right)\ i.i.d, i = 1, ..., N_1$，$x_{2i} \sim N\left(\mu_2, \sigma_2^2\right)\ i.i.d, i = 1, ..., N_2$，且两个样本独立。为了检验 $H_0 : \sigma_1^2 = \sigma_2^2$，在 H_0 的假定下，

$$\frac{\frac{(N_1-1)s_1^2}{\sigma_1^2} / (N_1 - 1)}{\frac{(N_2-1)s_2^2}{\sigma_2^2} / (N_2 - 1)} = \frac{s_1^2}{s_2^2} \sim F\left(N_1 - 1, N_2 - 1\right)$$

因而其拒绝域为 $R_\alpha = \left(0, F_{\alpha/2}\right) \cup \left(F_{1-\alpha/2}, \infty\right)$

例 5. 根据 2009 年中国城镇住户调查，在 37480 户家庭中，已知家庭年收入均值为 54157.63 元，标准差为 38533.96 元，请问在 1%、5% 的显著性水平下，是否可以认为家庭年收入均值为 53000 元？在这里，原假设为 $H_0 : \mu_0 = 53000$，在原假设条件下，我们有：

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{N}}} = \frac{\sqrt{37480}\left(\bar{x} - 53000\right)}{38533.96} \overset{a}{\sim} N\left(0, 1\right)$$

计算可得 $z = 5.82$，查表得到在 1% 的显著性水平下，拒绝域为 $(-\infty, -2.58) \cup (2, 58, \infty)$，显然 z 在拒绝域范围内，因而可以拒绝原假设。在 1% 的显著性水平下拒绝原假设，因而在 5% 的显著性水平下必然也拒绝原假设。

例 6. 根据 2013 年中国家庭金融调查，样本 7711 户家庭中，有 6% 的家庭有信用卡，请问在 5% 的显著性水平下，我们是否可以认为我国家庭持有信用卡的比例超过了 5%？为了解决这一问题，原假设为：$H_0 : p \leq 5\%$，在原假设的条件下，我们有：

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}} = \frac{\hat{p} - 5\%}{\sqrt{\frac{5\%(1-5\%)}{7711}}} \overset{a}{\sim} N(0,1)$$

带入计算可得检验统计量 $z = 4.03$。由于是右侧检验，因而拒绝域为 $\left(\Phi^{-1}(0.95), \infty\right)$，即 $(1.65, \infty)$，$4.93 > 1.65$，因而拒绝原假设，可以认为我国家庭持有信用卡的比例超过了 5%。

例 7. 在 2009 年中国城镇住户调查中，共有 23440 位 20-50 岁的男性，以及 21184 位 20-50 岁的女性。已知男性年平均收入为 28367.96 元，标准差为 21811.88 元；女性年平均收入为 20145.77 元，标准差为 16541.08 元。如果假设男女收入独立，请问在 5% 的显著性水平下，是否可以认为男女收入差异没有超过 10000 元？这里，原假设为 $H_0 : \mu_1 - \mu_2 \geq 10000$，在原假设的条件下，我们有：

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \overset{a}{\sim} N(0,1)$$

由于 $\bar{x}_1 - \bar{x}_2 = 28367.96 - 20145.77 = 8222.19$，$\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} = \frac{21811.88^2}{23440} + \frac{16541.08^2}{21184} = 33212.60$，$\mu_1 - \mu_2 = 10000$，因而 $z = -9.76$。由于是左侧检验，因而拒绝域为 $(-\infty, -1.65)$，可以拒绝原假设，即可以认为男女收入差异没有超过 10000 元。

2 评价检验的方法

根据以上的假设检验步骤，我们知道假设检验可以控制犯第 I 类错误的概率，然而对于假设检验而言，还有犯第 II 类错误的可能性，即当我们无法拒绝原假设的时候，备择假设可能是成立的。为了研究第 II 类错误的概率，我们引入假设检验的势（power）的概念。

定义 1. 对于一个假设检验及其拒绝域 R，检验的**势函数**（**power function**）即给定 θ 拒绝原假设的概率，即 $\beta(\theta) = P_\theta(T(x) \in R)$。

注意在以上定义中我们并没有限定 $\theta \in \Theta_0$ 或者 $\theta \in \Theta_0^c$，因而理想的情况下，当 $\theta \in \Theta_0$ 时，$\beta(\theta)$ 应该接近于 0，而当 $\theta \in \Theta_0^c$ 时，$\beta(\theta)$ 应该接近于 1。当 $\theta \in \Theta_0^c$，即备择假设为真时，没有拒绝原假设的概率，即 $1 - \beta(\theta)$，即犯第 II 类错误的概率。

例 8. 若样本 $x_i \sim N\left(\mu, \sigma^2\right)$ $i.i.d, i = 1, ..., N$，且 σ^2 已知，设原假设为 $H_0 : \mu \leq \mu_0$，备择假设为 $H_1 : \mu > \mu_0$，那么可以构建检验统计量为：

$$T(x) = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{N}}} \sim N(0,1)$$

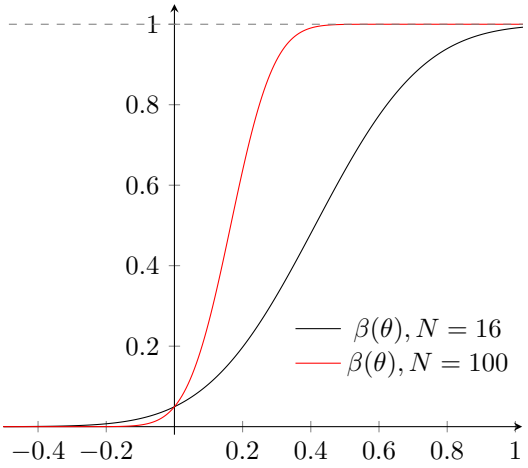


图 4: 势函数

在原假设的条件下，若给定 $\alpha = 0.05$，那么 $R_{0.05} = (z_{0.95}, \infty)$。因而，给定 μ，势函数为：

$$\begin{aligned}
\beta(\theta) &= P_\mu\left(T(x) \in R_{0.05}\right) \\
&= P_\mu\left(\frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{N}}} > z_{0.95}\right) \\
&= P_\mu\left(\frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{N}}} > z_{0.95} + \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{N}}}\right) \\
&= 1 - \Phi\left(z_{0.95} + \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{N}}}\right)
\end{aligned}$$

图 (4) 给出了当 $\mu_0 = 0$，$N = 16$，$\sigma^2 = 1$ 时的势函数。注意当 $\mu \to -\infty$ 时，$\beta(\mu) \to 0$；当 $\mu = \mu_0$ 时，$\beta(\mu) = 0.05$；而当 $\mu \to \infty$ 时，$\beta(\mu) \to 1$，且 $\beta(\theta)$ 是 μ 的单调递增函数。这意味着，在当原假设为真时，即 $\mu \leq \mu_0$ 时，拒绝原假设的概率总是小于等于 $\alpha = 0.05$ 的，这与我们假设检验控制第 I 类错误的概率是一致的。而当 $\mu > \mu_0$ 时，随着 μ 的增大，犯第 II 类错误的概率也随之降低。此外，注意到，犯第 II 类错误的概率是随着样本量 N 的增大而减小的。

观察图 (4) 我们会发现，不管样本量 N 再大，当真值 $\theta > \theta_0$，但是差异很小时，第 II 类错误概率 $1 - \beta(\theta)$ 仍然会无线趋向于 $1 - \alpha$。因而当备择假设为真，但是真值与原假设非常接近时，仍然需要样本量非常大才能够正确的拒绝原假设。

介于此种情况，我们可以引入无差异区域（indifference region），即虽然备择假设为真，但是 θ 与 θ_0 的差异足够的小，我们认为在这个区域里面错误接

受原假设也是可以接受的。例如，如果我们设计实验研究 wifi 会不会致癌，即看照射组合非照射组的实验对象患癌症的概率是否显著大于 0，即 $H_0 : \mu \leq 0$。假设 wifi 的确会致癌，但是致癌的概率充分的接近于 0，根据图 (4) 我们会发现，如果我们想要正确的拒绝原假设，需要非常大的样本量才能保证以一个比较大的概率 β 拒绝原假设。然而如果我们认为，致癌概率在一定的范围内，比如 $[0, \Delta)$ 内，是可以接受的，那么我们可以据此设计样本量，保证以一个确定的概率拒绝原假设。

如果我们希望，当 $\mu = \Delta$ 时，至少以 β 的概率拒绝原假设，那么

$$\beta = 1 - \Phi\left(z_\alpha + \frac{0 - \Delta}{\sqrt{\frac{\sigma^2}{N}}}\right)$$

从而

$$N = \left[\frac{\sigma}{\Delta}\left(z_\alpha - z_{1-\beta}\right)\right]^2$$

例 9. 在上例中，记实验组（wifi 照射）得癌症的概率为 p_1，对照组（无 wifi 照射）得癌症的概率为 p_2，我们关注的关键变量为 $\mu = p_1 - p_2$，原假设为 $H_0 : \mu \leq 0$。如果拒绝原假设，即可以认为 wifi 照射会致癌。如果假设两组的样本量相同，那么：

$$\mathrm{Var}\left(\hat{\mu}\right) = \mathrm{Var}\left(\hat{p}_1 - \hat{p}_2\right) = \mathrm{Var}\left(\hat{p}_1\right) + \mathrm{Var}\left(\hat{p}_2\right) = \frac{p_1\left(1 - p_1\right) + p_2\left(1 - p_2\right)}{N}$$

在原假设下，可以认为 $p_1 = p_2$，那么 $\mathrm{Var}\left(\hat{\mu}\right) = \frac{2p_1(1-p_1)}{N}$。如果取 $\Delta = 0.001, \beta = 0.8$，记我们希望当 wifi 致癌的概率为千分之一时，我们希望以 80% 的概率拒绝原假设，取 $\alpha = 0.05$，那么所需样本量为：

$$N = \left[\frac{2\sqrt{p_1\left(1 - p_1\right)}}{\Delta}\left(1.65 - 0.15\right)\right]^2 = 4\left(\frac{1.65 - 0.15}{0.0001}\right)^2\left[p_1\left(1 - p_1\right)\right]$$

如果自然条件下，癌症发病率为 0.2%，即当 $p_1 = 0.002$ 时，上述样本量大概需要 $N = 17964$ 个。而反过来，如果样本量足够大，那么 μ 与 μ_0 的一些非常细微的差别也足以导致统计上的显著性，尽管很多时候这种细微的差别几乎没有现实意义。因而，即使数据中可以得到统计上显著的结果，特别是在样本量非常大的情况下，我们仍然要注意这些结果是不是有「经济显著性（economic significancy）」，即这些差别在现实中是不是足以引起重视。

3 构造假设检验的方法

在以上两节中我们介绍了假设检验的一般概念和思路。我们知道，如果在原假设 H_0 的条件下得到检验统计量及其抽样分布，我们就可以使用上节中给出的步骤进行假设检验。尽管上一节中我们讨论了单个样本、多个样本均值的

假设检验，然而很多时候我们可能希望对不止一个参数进行假设检验，或者对参数的函数进行假设检验。一般的，记我们的原假设为：

$$H_0 : C(\theta) = 0$$

其中 $\theta \in \mathbb{R}^k$，$C(\theta) \in \mathbb{R}^r$，且 $C(\theta)$ 为 θ 的连续可微函数。那么一般的，我们可以通过如下两个方法构造假设检验。

3.1 Wald 检验

如果对于 θ 的估计，我们已经有 $\hat{\theta} \sim N(\theta, \Sigma)$，那么使用 Delta 方法：

$$C\left(\hat{\theta}\right) = C(\theta) + \ddot{C}\left(\hat{\theta} - \theta\right) + o\left(\left|\hat{\theta} - \theta\right|\right)$$

其中 $\ddot{C} = \partial C(\theta)/\partial\theta$ 为 $r \times k$ 的矩阵，且假设 $\operatorname{rank}\left(\ddot{C}\right) = r$，那么在原假设的条件下，：

$$C\left(\hat{\theta}\right) = \ddot{C}\left(\hat{\theta} - \theta\right) + o_p(1) \sim N\left(0, \ddot{C}\Sigma\ddot{C}'\right)$$

进而我们可以构建检验统计量：

$$C'\left(\hat{\theta}\right)\left[\ddot{C}\Sigma\ddot{C}'\right]^{-1} C\left(\hat{\theta}\right) \sim \chi^2_r$$

因而可以使用以上检验统计量对原假设进行假设检验。

3.2 似然比检验

在原假设的条件下，如果可以使用极大似然估计对 θ 进行估计，那么我们可以在 H_0 的约束下对 θ 进行估计，即：

$$\tilde{\theta} = \arg\max_{\theta} L(\theta|x)$$
$$s.t. C(\theta) = 0$$

与此同时，我们还可以估计无约束的极大似然估计：

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x)$$

由于 $\tilde{\theta}$ 是在约束条件下得到的，而 $\hat{\theta}$ 是在无约束的条件下得到的，因而 $L\left(\hat{\theta}|x\right) \geq L\left(\tilde{\theta}|x\right)$。如果原假设成立，即 $C(\theta) = 0$，那么 $\tilde{\theta}$ 是和 $\hat{\theta}$ 应该充分接近，因而 $L\left(\hat{\theta}|x\right)$ 和 $L\left(\tilde{\theta}|x\right)$ 也应该充分接近；但是如果 $L\left(\hat{\theta}|x\right) > L\left(\tilde{\theta}|x\right)$ 成立，那么

我们可以认为 $C(\theta)=0$ 是不成立的。实际上，我们可以得到：

$$LR=2\left[L\left(\hat{\theta}|x\right)-L\left(\tilde{\theta}|x\right)\right]\sim\chi_r^2$$

因而我们可以根据以上结论对 $C(\theta)=0$ 进行检验。

例 10. 如果 $x_i\sim P(\lambda)$，$i=1,...,N$，如果原假设为 $H_0:\lambda=1$，那么无约束时，$\hat{\lambda}=\bar{x}$，而有约束时，$\tilde{\lambda}=1$。其似然函数分别为：

$$L\left(\hat{\lambda}|x\right)=\sum_{i=1}^{N}\left[x_i\ln\left(\hat{\lambda}\right)-\ln(x_i!)-\hat{\lambda}\right]=\sum_{i=1}^{N}\left[x_i\ln(\bar{x})-\ln(x_i!)-\bar{x}\right]$$
$$L\left(\tilde{\lambda}|x\right)=\sum_{i=1}^{N}\left[x_i\ln\left(\tilde{\lambda}\right)-\ln(x_i!)-\tilde{\lambda}\right]=\sum_{i=1}^{N}\left[x_i\ln(1)-\ln(x_i!)-1\right]$$

因而检验统计量为：

$$\begin{aligned}LR&=2\left[\sum_{i=1}^{N}\left[x_i\ln(\bar{x})-\ln(x_i!)-\bar{x}\right]-\sum_{i=1}^{N}\left[x_i\ln(1)-\ln(x_i!)-1\right]\right]\\&=2\left[\sum_{i=1}^{N}\left[x_i\ln(\bar{x})-\bar{x}+1\right]\right]\\&=2N\left[\ln(\bar{x})\,\bar{x}-\bar{x}+1\right]\sim\chi_1^2\end{aligned}$$

习题

1. （程序题）生成一系列正态分布 $x_i\sim N\left(\mu,\sigma^2\right)$，给定不同的 $\mu\in[-2,2]$，在 $H_0:\mu=0$ 的条件下，对于固定的 μ，重复 1000 次假设检验的过程，并计算对于每一个 μ，在 $H_0:\mu=0$ 的条件下拒绝原假设的比率（power function）。

2. （程序题）对于比例的检验，在 $H_0:p=p_0$ 的原假设下，我们有如下两个渐进等价的方法：

   $$\frac{\hat{p}-p}{\sqrt{\frac{p_0(1-p_0)}{N}}}\overset{a}{\sim}N(0,1)$$

   以及：

   $$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}}\overset{a}{\sim}N(0,1)$$

   分别模拟两种检验方法在小样本（$N=15$）和大样本（$N=50$）条件下的势函数，并进行比较。

3. （程序题）请使用对数似然比检验，在 $H_0:p=p_0$ 的原假设下，写出对样本比例的检验统计量，并使用数值模拟计算出其势函数，并于上题中的势

函数进行比较。

参考文献

[1] Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

[2] Shao, J., 2007. Mathematical Statistics, 2nd ed. Springer, New York.