第 5 章 无失真信源编码

CHAPTER 5

由于信源符号承载了需要传输的信息量，因此通信的根本问题是在信宿端尽量准确地恢复信道所传输的信源符号。为此，首先需要解决两个问题：

（1）**信源到底输出多少信息量**。此为信息度量问题，已经在第 2 章中解决。

（2）**如何让消息来承载信息量**。此为信源编码问题，将在本章中介绍。

若信宿要求精确地复现信源的输出，则需保证信源产生的全部信息量无损地传送给信宿，此时的信源编码就是**无失真信源编码**。为分析方便，将信道编码看成信道的一部分，不考虑信道编码。

5.1 信源编码的相关概念

本节简单介绍与信源编码有关的基本概念和数学模型。

5.1.1 无失真信源编码的数学模型

由于无失真信源编码可以不考虑抗干扰问题，所以它的数学模型比较简单。根据符号个数可以分为单符号无失真信源编码和多符号无失真信源编码。

1. 单符号无失真信源编码

单符号的无失真信源编码，其数学模型为

$$S=\{s_1,s_2,\cdots,s_Q\}\longrightarrow\boxed{\text{编 码 器}}\longrightarrow C=\{\omega_1,\omega_2,\cdots,\omega_Q\}$$
$$\uparrow$$
$$X=\{x_1,x_2,\cdots,x_R\}$$

根据这个数学模型，简要介绍信源编码中的几个基本概念：

（1）**信源符号**：无失真信源编码的对象是原始的信源符号 S，其样本空间为信源符号集 $\{s_1,s_2,\cdots,s_Q\}$，共有 Q 个信源符号。

（2）**码字**：无失真信源编码的结果。码字与信源符号一一对应，因此共有 Q 个码字 $\omega_i(i=1,2,\cdots,Q)$。

（3）**代码组**：码字的集合称，简称**码**，表示为 $\boldsymbol{C}=\{\omega_1,\omega_2,\cdots,\omega_Q\}$。

（4）**码元**：码字 ω_i 的基本组成单位，即码字 ω_i 是由若干码元组成的码元序列。所有码元组成的集合称为码元集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_R\}$。码元又称为**码符号**。

（5）**码长**：组成码字 ω_i 的码元符号个数，用符号 l_i 表示。

（6）**编码器**：将信源符号 s_i 变换成码字 ω_i 的单元，表示为

$$s_i \leftrightarrow \omega_i = x_{i_1}, x_{i_2}, \cdots, x_{i_{l_i}} \qquad (i=1,2,\cdots,Q) \tag{5.1}$$

式中，$x_{i_k} \in \boldsymbol{X}(k=1,2,\cdots,l_i)$。编码器位于信源处，与此相对应，信宿处有一个**译码器**完成相反的功能（译码）。

由此可知，信源编码是信源符号和输出码字之间的一种映射。若要实现无失真信源编码，则这种映射必须一一对应并且可逆。

例5.1 二元码。

设有二元信道，其信道输入的符号分布（信源概率空间）为

$$\begin{bmatrix}\boldsymbol{S}\\ P\end{bmatrix}=\begin{bmatrix}s_1 & s_2 & \cdots & s_Q\\ p(s_1) & p(s_2) & \cdots & p(s_q)\end{bmatrix}$$

二元信道是指信道中传输的消息，由基本符号集 $\{0,1\}$ 中的元素组成。因此，若信源发出的消息要通过二元信道传输，则需将信源符号 $s_i(i=1,2,\cdots,Q)$ 变换为由基本符号 0 和 1 组成的符号序列，此过程即为信源编码。

二元信道中信源编码的码元为 0 和 1，码元集 $\boldsymbol{X}=\{0,1\}$。实际上，存在多种信源编码方案使码元序列（码字）ω_i 与信源符号 s_i 一一对应，表 5.1所列的**代码组** 1 和**代码组** 2 都是**二元码**（码元集有 2 个元素）。

表 5.1　二元码

信源符号 s_i	信源符号概率 $p(s_i)$	代码组 1（及其码长）	代码组 2（及其码长）
s_1	$p(s_1)$	00（2）	0　（1）
s_2	$p(s_2)$	01（2）	01　（2）
s_3	$p(s_3)$	10（2）	001（3）
s_4	$p(s_4)$	11（2）	111（3）

2. 多符号无失真信源编码

多符号无失真信源编码的数学模型为（以 N 次扩展信源为例）

$$S^N=\{\alpha_1,\alpha_2,\cdots,\alpha_{Q^N}\} \longrightarrow \boxed{\text{编码器}} \longrightarrow \boldsymbol{C}=\{\omega_1,\omega_2,\cdots,\omega_{Q^N}\}$$
$$\uparrow$$
$$X=\{x_1,x_2,\cdots,x_R\}$$

编码器将 N 长信源符号序列 α_i 转换成码长为 l_i 的码字 ω_i。按照 N 次扩展信源的工作原理，输入的信源符号序列 α_i 共有 Q^N 个，输出的码字也有 Q^N 个。码字 ω_i 与信源符号序列 α_i 一一对应。因此，多符号无失真信源编码的编码器表示为

$$\alpha_i \leftrightarrow \omega_i = x_{i_1}, x_{i_2}, \cdots, x_{i_{l_i}} \quad (i = 1, 2, \cdots, Q^N)$$
$$\alpha_i = s_{i_1}, s_{i_2}, \cdots, s_{i_N}$$
$$s_{i_n} \in S = \{s_1, s_2, \cdots, s_Q\} \quad (n = 1, 2, \cdots, N)$$
$$a_{i_k} \in \boldsymbol{X} = \{x_1, x_2, \cdots, x_R\} \quad (k = 1, 2, \cdots, l_i)$$

例5.2 二次扩展码。

表 5.1中代码组 2 的二次扩展码如表 5.2所示。

表 5.2　二次扩展码

二次扩展信源符号 α_i	二次扩展码字 ω_i
$\alpha_1 = s_1 s_1$	00
$\alpha_2 = s_1 s_2$	001
$\alpha_3 = s_1 s_3$	0001
$\cdots$	$\cdots$
$\alpha_{16} = s_4 s_4$	111111

5.1.2　信源编码的分类

由例 5.1可知，同一个信源可以有不同的编码方案（代码组）。这个多个编码方案可以根据编码特性将信源编码分为以下类别：

（1）**分组码和非分组码**。

信源编码过程可以抽象为一种映射，即将信源符号集 $S = \{s_i, i = 1, 2, \cdots, Q\}$ 中的每个符号 s_i 映射为码字 $\omega_i (i = 1, 2, \cdots, Q)$。

定义5.1 分组码

信源符号集中的每个信源符号 s_i 固定地映射为相对应的码字 ω_i，这样的码称为**分组码**，分组码又称为**块码**。

与分组码对应的是**非分组码**，又称为**树码**。树码编码器输出的码符号通常与所有信源符号都有关。例 5.1中的编码就是分组码，信源输出的符号序列分成组，从信源符号序列到码字序列的变换是在分组的基础上进行的，即特定的信源符号组唯一地确定了特定的码字组。

（2）**奇异码**和**非奇异码**。

定义5.2 奇异码和非奇异码

若一种分组码中的所有码字都不相同，则称此分组码为**非奇异码**；否则，称为**奇异码**。

表 5.3中，码一是奇异码，码二是非奇异码。非奇异码是分组码能够正确译码的必要条件，而不是充分条件。例如，传输分组码二，当接收端接收到序列 00 时，并不能确定发送端发送的消息是 s_1s_1 还是 s_3。

表 5.3　奇异码和非奇异码

信源符号 s_i	码一	码二
s_1	0	0
s_2	11	10
s_3	00	00
s_4	11	01

（3）**唯一可译码和非唯一可译码**。

定义5.3 唯一可译码

任意有限长的码元序列，若只能唯一地分割成一个个码字，则称为**唯一可译码**。

唯一可译码不但要求不同的码字表示不同的信源符号，而且要求对信源符号序列进行编码时，在接收端仍能正确译码而不发生混淆。唯一可译码首先是非奇异码，且任意有限长的码字序列不能相同。

若对于任意有限的整数 N，分组码的 N 次扩展码均为非奇异的，则为唯一可译码；否则，为**非唯一可译码**。

（4）**即时码和非即时码**。

定义5.4 即时码

不考虑后续码元就可以从码元序列中译出码字，这样的唯一可译码称为**即时码**。

表 5.4列出了两组唯一可译码。当传输码一时，信宿接收到一个码字后不能立即译码，还需等到下一个码字接收到时才能判断是否可以译码；传输码二时，无此限制，接收到一个完整码字后立即可以译码，这是一种即时码，是唯一可译码的一种。

表 5.4　唯一可译码与即时码

信源符号 s_i	码一	码二
s_1	1	1
s_2	10	01
s_3	100	001
s_4	1000	0001

信源编码分类，如图 5.1所示。

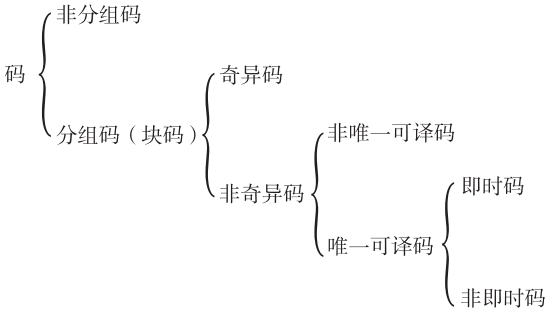

图 5.1　信源编码分类

除了上述的分类，还可以根据码长和码元的性质分类如下：

（1）**定长码和变长码**。

若码字集合 $\boldsymbol{C}$ 中所有码字的码长相同，则称为**定长码**或**等长码**；反之，码长不同，则称为**变长码**。

（2）**二元码和多元码**。

若码元集合有两个元素 $\boldsymbol{X}=\{0,1\}$，则对应的码字为**二元码**。二元码是数字通信系统中常用的一种码。若码元集合中的码元有多个，则称为**多元码**。

5.1.3　即时码存在定理和构造方法

1. 即时码存在定理

定义5.5 码字前缀

设 $\omega_i=x_{i_1}x_{i_2}\cdots x_{i_l}$ 为一码字，对于任意的 $1\leqslant j\leqslant l$，码元符号序列的前 j 个元素 $x_{i_1}x_{i_2}\cdots x_{i_j}$ 称为码字 ω_i 的前缀。

根据码字前缀的定义，有下述定理：

定理5.1 即时码存在

唯一可译码成为即时码的充要条件是代码组中的任何码字都不是其他码字的前缀。

证明：（1）**充分性**。若所有码字都不是其他码字的前缀，则接收到一个码元符号序列（长度为一个完整码字的码长）后，便可以立即译码，而无须考虑后面的码元。

（2）**必要性**。设码字 ω_i 是码字 ω_j 的前缀，若接收到一个码元符号序列，其长度为码字 ω_i 的码长，则不能立即判定这个码元序列是一个完整的码字，还必须参考后续的码元。这与即时码的定义相矛盾。因此，即时码存在的必要条件为任何一个码字都不是其他码字的前缀。

2. 即时码构造方法

即时码可以利用码树来构造，非常方便。码树中有**树枝**和**节点**。码树最顶部的节点称为**根节点**，不再产生分支的节点称为**终端节点**，其他节点称为**中间节点**。每个节点能够生成的树枝数目等于码元数 r。一个 r 进制码树的生成过程如下：

（1）从根节点出发，延伸出 r 条树枝，产生 r 个**一阶节点**；

（2）在树枝旁边分别标以码元 $0, 1, \cdots, r-1$；

（3）每个一阶节点再延伸出 r 条树枝，产生 r 个**二阶节点**，因此共有 r^2 个二阶节点；

（4）重复上述过程，从 $N-1$ 阶节点出发再延伸出 r 条树枝，产生 r 个 N **阶节点**，因此共有 r^N 个节点。

若指定某个 N 阶节点为终端节点，表示一个信源符号，则该节点不再延伸。从根节点到指定终端节点所经过的树枝形成一条路径，该路径所对应的码元序列就构成一个码字。

根据码树构造的码字满足即时码的条件，因为从根节点到每个终端节点所经过的路径均不相同，并且中间不安排码字，故一定满足**即时码**对前缀的限制。若有 Q 个信源符号，则在码树上就要选择 Q 个终端节点，相应地有 Q 个码字。

例5.3 二元即时码的码树。

已知一个码 C 包含 4 个码字，码字集合为 $\{1, 01, 000, 001\}$，试用码树表示。

解：码树采用二进制码，如图 5.2所示。

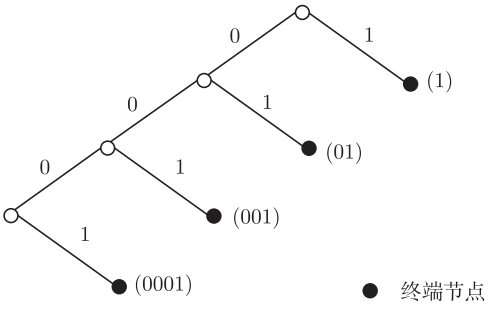

图 5.2　码的二进制码树表示

5.2　定长码及定长编码定理

一般而言，实现无失真信源编码，码必须是唯一可译码；否则，就会因译码带来错误与失真。本节主要讨论定长码实现无失真编码的条件和性质。

5.2.1　唯一可译定长码存在的一般条件

对定长码来说，若定长码是非奇异码，则它的任意有限长 N 次扩展码一定也是非奇异码，因此定长非奇异码一定是唯一可译码。

1. 一般条件

若信源 S 有 Q 个信源符号，则信源 S 存在唯一可译定长码的条件为

$$Q \leqslant r^l \tag{5.2}$$

式中，r 为码元集合中的码元数；l 为定长码的码长。

例如，表 5.1中，信源 S 共有 $Q=4$ 个信源符号，进行二元等长编码时，$r=2$，则信源存在唯一可译定长码的条件是 $l \geqslant 2$。

2. N 次扩展信源存在唯一可译定长码的条件

对信源 S 的 N 次扩展信源 S^N 进行定长编码，若要编得的定长码是唯一可译码，则必须满足

$$Q^N \leqslant r^l \tag{5.3}$$

式中，Q 为信源 S 的符号个数，Q^N 为 N 次扩展信源 S^N 的符号个数；r 为码元集合 $\boldsymbol{X}$ 的码元数。

当把 N 次扩展信源 S^N 的信源符号 $\alpha_i(i=1,2,\cdots,Q^N)$ 变换成长度为 l 的码字 $\omega_i = x_{i_1}x_{i_2}\cdots x_{i_l}$ 时，若要求编得的定长码是唯一可译码，则每个信源符号 α_i 都有一个码字对应，所以必须满足式 (5.3)。换句话说，只有当 l 长的码元序列个数 r^l 不小于 N 次扩展信源的信源符号序列个数 Q^N 时，才能存在定长非奇异码。

对式 (5.3) 两边取对数，可得

$$N\log Q \leqslant l\log r$$

即

$$\frac{l}{N} \geqslant \frac{\log Q}{\log r} = \log_r Q \tag{5.4}$$

说明：（1）$\dfrac{l}{N}$ 表示扩展信源 S^N 中，单个原始信源符号所需的平均码元个数；

（2）对定长唯一可译码而言，平均每个原始信源符号至少需要 $\log_r Q$ 个码元表示；

（3）当 $r=2$ 时，$\dfrac{l}{N} \geqslant \log_2 Q$ 表示对二元定长唯一可译码，平均每个原始信源符号至少需要用 $\log_2 Q$ 个二元符号表示；

（4）当 $N=1$ 时，$l \geqslant \log_r Q$。

例5.4 英文电报编码。

已知英文电报有 32 个符号（26 个英文字母和 6 个标点符号）。若利用二元码对英文电报符号进行编码，至少需要多少个码元？

解：据式（5.4），$Q=32$。如果采用二元码，则 $r=2$。对信源 S 的每个符号 $s_i(i=1,2,\cdots,32)$ 进行二元编码，则 $N=1$，即有

$$l = \frac{l}{N} \geqslant \log_2 32 = 5$$

这说明，每个英文电报符号至少要用 5 位的二元码进行编码才能得到唯一可译码。

由例 3.17 可知，对实际英文电报信源，在考虑了符号出现概率以及符号之间的相关性后，平均每个英文电报符号所能提供的信息量约为 1.4bit，远小于 5bit。因此，定长编码后，每个码字只载荷约 1.4bit 信息量（5 个二元码符号只携带 1.4bit 信息量），而 5 个二元码符号却可以携带的最大信息量为 5bit。因此，定长编码的信息传输效率较低。那么如何才能提高信息传输效率？

（1）**考虑符号之间的依赖关系**。对信源 S 的扩展信源进行编码，考虑符号间的依赖关系后，某些信源符号序列不会出现，这样可能出现的信源符号序列个数会小于 Q^N 个。

（2）**小概率符号序列不予编码**。对概率为 0 或者非常小的符号序列可以不予编码，这样可能会造成一定的误差。但是，当 N 足够长以后，这种误差概率可以任意小，即可以做到**几乎无失真编码**。

5.2.2 定长编码定理

下面将讨论的定长编码定理给出了定长编码所需码长的理论极限值。根据渐近等同分割性和 ε 典型序列（见**附录** E）的性质，可以得到以下**定长编码定理**。

定理5.2 定长编码

设离散无记忆信源的熵为 $H(S)$。若对信源长为 N 的序列进行定长编码，码元集合 X 中有 r 个符号，码长为 l。对于任意 $\varepsilon>0$，只要满足 $\dfrac{l}{N}\geqslant\dfrac{H(S)+\varepsilon}{\log r}$，当 N 足够大时，可实现**几乎无失真编码**，即译码错误概率 δ 任意小；若 $\dfrac{1}{N}\leqslant\dfrac{H(S)-2\varepsilon}{\log r}$，则不可能实现**几乎无失真编码**，即当 N 足够大时，译码错误概率为 1。

证明：（1）ε **典型序列**与**非** ε **典型序列**。

离散无记忆信源的 N 次扩展信源的输出序列可以分为高概率的ε **典型序列**和低概率的**非** ε **典型序列**两类。当 $N\to\infty$ 时，ε 典型序列出现的概率趋于 1，非 ε 典型序列出现的概率趋于 0。

（2）ε **典型序列的等长编码**。由于 ε 典型序列在全部序列中的比例很小，因此只对少数的 ε 典型序列进行一一对应的等长编码，这就要求码字总数不小于 M_{G}，即

$$r^l\geqslant M_{\mathrm{G}} \tag{5.5}$$

$$r^l\geqslant 2^{N\left[H(S)+\varepsilon\right]}\geqslant M_{\mathrm{G}} \tag{5.6}$$

$$l\log r\geqslant N\left[H(S)+\varepsilon\right] \tag{5.7}$$

$$\frac{l}{N}\geqslant\frac{H(S)+\varepsilon}{\log r} \tag{5.8}$$

这样就能使 S^N 中所有ε **典型序列**都有不同的码字与之对应。尽管**非** ε **典型序列**的总概率很小，但这些**非** ε **典型序列**仍可能出现，因而会造成译码错误，其错误概率 P_{E} 就是 $\overline{G_\varepsilon}$ 出现的概率。故

$$P_{\mathrm{E}} = P\left[\overline{G_\varepsilon}\right] \leqslant \delta(N,\varepsilon) = \frac{D\left[I(s_i)\right]}{N\varepsilon^2} \tag{5.9}$$

当 $N \to \infty$ 时，$P_{\mathrm{E}} \to 0$。

- **几乎无失真编码的条件**

若 $\dfrac{l}{N} \leqslant \dfrac{H(S)-2\varepsilon}{\log r}$，则有

$$r^l \leqslant 2^{N[H(S)-2\varepsilon]} \leqslant 2^{-N_\varepsilon} 2^{N[H(S)-\varepsilon]} \leqslant \left[1-\delta(N,\varepsilon)\right] 2^{N[\boldsymbol{H}(S)-\varepsilon]}$$

此时选取的码字总数小于ε **典型序列**的序列数，因而ε **典型序列**集将有部分序列没有码字与之对应。把有码字对应的序列的发生概率之和记作 $P\left[G_\varepsilon\text{中}r^l\text{个}s_j\right]$，则其满足

$$P\left[G_\varepsilon\text{中}r^l\text{个}s_j\right] \leqslant r^l 2^{-N[H(S)-\varepsilon]} \leqslant 2^{N[H(S)-2\varepsilon]} 2^{-N[H(S)-\varepsilon]} = 2^{-N_\varepsilon}$$

G_ε 中的 r^l 个信源符号序列由于有不同的码字与之对应，所以在译码时能得到正确恢复。其他没有码字对应的信源符号序列中译码时均会产生错误，因而正确译码概率 $\overline{P}_{\mathrm{E}}$ 满足

$$\overline{P}_{\mathrm{E}} = 1 - P_{\mathrm{E}} = P\left[G_\varepsilon\text{中}r^l\text{个}s_j\right]$$

由于 $1-P_{\mathrm{E}} \leqslant \mathrm{e}^{-N_\varepsilon}$，则 $P_{\mathrm{E}} \geqslant 1-2^{-N_\varepsilon}$。当 $N \to \infty$ 时，$P_{\mathrm{E}} \to 1$。所以定长编码时，如果码长 l 满足 $\dfrac{l}{N} \leqslant \dfrac{H(\boldsymbol{S})-2\varepsilon}{\log r}$，当 N 很大时，许多经常出现的 ε 典型序列被舍弃而没有编码，这样会造成很大的译码错误，不可能实现**几乎无失真编码**。

说明：（1）定理 5.2是在**离散平稳无记忆信源**的条件下得到的，但它同样适用于**平稳有记忆信源**，只需将式中的 $H(S)$ 修改为极限熵 H_∞ 即可，条件是**有记忆信源的极限熵** H_∞ **和极限方差** σ_∞^2 **存在**。

（2）定长编码定理给出了定长编码时每个信源符号所需的最少码元的理论极限，该极限值由信源熵 $H(S)$ 决定。

（3）对二元编码，$r=2$，有

$$\frac{l}{N} \geqslant H(S)+\varepsilon \tag{5.10}$$

这是定长编码时平均每个信源符号所需的二元码元数的理论极限。

（4）编码效率的提高

$$\text{式 (5.4)：}\quad \frac{l}{N} \geqslant \underbrace{\log_2 Q}_{\text{当符号等概率分布时 } H(S)=\log_2 Q} \qquad H(S)+\varepsilon \leqslant \frac{l}{N} \quad \text{：定理 5.2}$$

当信源符号等概率分布时，根据式（5.4）得到的码元个数，与根据定理 5.2所得到的结果是一致的。但一般情况下信源符号并非等概率分布，这就意味着根据式（5.4）得到的结果偏大；同时，信源符号之间存在相关性，某些信源符号间的相关性会很强，信源熵 H_∞ 会远小于信源熵最大值 $\log_2 Q$，因此式（5.4）得到的结果又会进一步变大。故而根据定理 5.2，单个信源符号平均所需的二元码元可大大减少，从而使编码效率提高。

（5）定理 5.2中的条件 $\underline{\dfrac{l}{N} \geqslant \dfrac{H(S)+\varepsilon}{\log r}}$ 还可以写为下面的形式：

$$\underbrace{l \log r}_{l \text{ 个码元所能携带的最大信息量}} > \underbrace{N H(\boldsymbol{S})}_{N \text{ 长信源符号序列所携带的信息量}} \quad (5.11)$$

式（5.11）表明，只要 l 长码元序列所能携带的最大信息量大于 N 长信源符号序列的信息熵，就可以实现无失真传输，条件是N 足够大。

例5.5 提高英文电报编码效率。

已知英文电报有 32 个符号（26 个英文字母和 6 个标点符号）。

若考虑信源符号间的相关性，利用二元码对英文电报符号进行编码时，至少需要多少个码元？

解：考虑英文信源符号之间的相关性，信源熵 $H_\infty \approx 1.4$ 比特/符号。根据定理 5.2，码长 l 需满足 $\dfrac{l}{N} = l > 1.4$ 个二元码元符号/信源符号。这说明，平均每个英文电报符号只需大约 1.4 个二元符号来编码，与例 5.4计算的 5 个二元符号相比减少了许多，信息传输效率得到提高。

定义5.6 编码速率

熵为 $H(S)$ 的离散无记忆信源，若对 N 长的信源符号序列进行定长编码，码元集的码元个数为 r。

码长为l，定义**编码速率**$R' = \dfrac{l}{N} \log r$ (5.12)

编码速率表示平均每个信源符号编码后所能载荷的最大信息量。

根据**编码速率**的概念，可以将定理 5.2中的条件 $\underline{\dfrac{l}{N} \geqslant \dfrac{H(S)+\varepsilon}{\log r}}$ 改写为 $R' \geqslant H(S)+\varepsilon$。这说明，编码速率大于信源熵才能实现**几乎无失真编码**。

定义5.7 编码效率

编码效率定义为