- 在概率论中，该不等式给出随机变量的和与其期望值偏差的概率上限。

伯努利随机变量的特例

- 霍夫丁不等式经常被用于一些分布的伯努利随机变量的重要特例中。
- 硬币 A面 p，B面 $1-p$，n次后A面朝上期望 np.

A面次数不超过 k 次的概率为：

$$P(H(n)\le k) = \sum \binom{n}{i} p^i (1-p)^{n-i}.$$

表示A面朝上的次数.

对某一 $\varepsilon>0$，当 $k=(p-\varepsilon)n$ 时，上面不等式确定的霍夫丁上界将会按照指数级变化：

$$P(H(n)\le(p-\varepsilon)n) \le \exp(-2\varepsilon^2 n).$$

相似的，对某一 $\varepsilon>0$，当 $k=(p+\varepsilon)n$ 时，霍夫丁不等式的概率也界同样可以确定为：

$$P(H(n)\ge(p+\varepsilon)n) \le \exp(-2\varepsilon^2 n).$$

根据上两式可以得到：

$$P((p-\varepsilon)n \le H(n) \le (p+\varepsilon)n) \ge 1-2\exp\{-2\varepsilon^2 n\}.$$

现在令 $\varepsilon=\sqrt{\ln n/n}$

那么可以得到：

$$P(|H(n)-pn| \le \sqrt{n\ln n}) \ge 1-2\exp\{-2\ln n\} = 1-\frac{2}{n^2},$$

普遍情况

- 现在令 $X_1, x_2, \cdots, x_n$ 为[0,1]的独立随机变量，即 $0 \le x_i \le 1$，我们定义变量的经验均值为：

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$$

- 霍夫丁定理一中的不等式为：

$$P(\bar{x} - E[\bar{x}] \ge t) \le \exp(-2nt^2)$$

当知道 x_i 严格边界范围 a_i, b_i 时，霍夫丁定理二中更加广泛。

有界区间，

$$P(\bar{x} - E[\bar{x}] \ge t) \le \exp\left\{-\frac{2n^2t^2}{\sum(b_i - a_i)^2}\right\}$$

$$P(|\bar{x} - E[\bar{x}]| \ge t) \le 2\exp\left\{-\frac{2n^2t^2}{\sum(b_i - a_i)^2}\right\}$$

证明

霍夫丁引理

- 假设 X 为一均值为0的实数随机变量并且满足，

$$P(X \in [a,b]) = 1$$

那么有如下不等式

$$E[e^{sX}] \le \exp\{\frac{1}{8}s^2(b-a)^2\}$$

- 假如 $x_1, \cdots x_n$ 为 n 个独立分布随机变量并且

$$P(x_i \in [a_i, b_i]) = 1 \qquad 1 \le i \le n$$

令

$$S_n = x_1 + \cdots + x_n$$

对于，$s, t \ge 0$，马尔可夫不等式以及 x_i 独立性说明

$$\begin{aligned} P(S_n - E[S_n] \ge t) &= P(e^{s(S_n - E[S_n])} \ge e^{st}) \\ &\le e^{-st} \cdot E[e^{s(S_n - E[S_n])}] \\ &= e^{-st} \prod E[e^{s(X_i - E[X_i])}] \\ &\le e^{-st} \prod e^{\frac{s^2(b_i - a_i)^2}{8}} \\ &= \exp\{-st + \frac{1}{8}s^2 \Sigma (b_i - a_i)^2\} \end{aligned}$$

为了得到最好的概率上限，将不等式右边等于为一个关于 s 的函数，

$$\begin{cases} g: R_+ \longrightarrow R \\ g(s) = -st + \frac{s^2}{8} \Sigma (b_i - a_i)^2 \end{cases}$$

注意到 g 是二次函数，要取最小值需满足

$$s = \frac{4t}{\Sigma (b_i - a_i)^2}$$

这样我们有

$$P(S_n - E[S_n] \ge t) \le \exp\{-\frac{2t^2}{\Sigma (b_i - a_i)^2}\}$$

举例

置信区间

- 霍夫丁不等式被用来分析样本的置信区间，

$$P(\bar{X} - E[\bar{X}] \geq t) \leq e^{-2nt^2}$$

这个不等式说明估计值大t的概率被指数边界控制，

$$P(-\bar{X} + E[\bar{X}] \geq t) \leq e^{-2nt^2}$$

将两式合并

$$P(|\bar{X} - E[\bar{X}]| \geq t) \leq 2e^{-2nt^2}.$$

上述不等式可以理解为：

$$\alpha = P(\bar{X} \notin [E[\bar{X}] - t, E[\bar{X}] + t]) \leq 2e^{-2nt^2}.$$

即真值估计范围。其中

$$n \geq -\frac{\log(\alpha/2)}{2t^2}.$$

所以我们要求至少上述不等式右边式子的样本数量从而使估值区间更加靠近真值。