正态分佈

A. K. 米特洛波尔斯基著

科学出版社

正态分佈

A. K. 米特洛波尔斯基著
施步嘉譯

科学出版社
1959

目　　录

緒　言

正态分布是統計量分布的基本类型．在建立正态分布时是以二項式分布

$$p_{n,m}=\frac{n!}{m!\,(n-m)!}p^m q^{n-m} \tag{0.1}$$

为出發点的．

二項式分布表示在 n 次独立試驗中事件 E 出現 m 次的概率，在每次試驗中，事件 E 出現的概率等于同一个量 p，而事件 E 不出現的概率是 $q=1-p$．

在研究二項式分布时，最主要的是解决两个問題：1) 确定分布的最或然值及該值的概率；2) 确定分布的这样一組值的概率，这組值与最或然值的差不超过給定的数，也就是說处于預先确定的范圍內．

第一个問題的解决，在数 n 无限增大时，是以莫阿佛尔-斯蒂尔林格 (Муавр-Стирлинг) 公式

$$n!=\sqrt{2\pi n}\,n^n e^{-n}$$

为基础的．

藉助于这个公式我們求出，当数 n 无限增加时，事件出現的最或然数 np 的概率等于

$$p_{n,np}=\frac{1}{\sqrt{2\pi npq}}. \tag{0.2}$$

同时假設概率 p 和 q 中沒有一个是很小的．

拉普拉斯定理給出第二个問題的解答．如果无限制地进行独立試驗，而在每次試驗中事件 E 出現的概率等于 p，則当試驗次数 n 无限增多时，事件 E 出現的数目 m 满足不等式

$$t_1\sqrt{npq}<m-np<t_2\sqrt{npq}$$

的概率 $P\,(n,\ t_1,\ t_2)$ 趋于極限

$$\frac{1}{\sqrt{2\pi}}\int_{t_1}^{t_2} e^{-\frac{t^2}{2}}dt, \tag{0.3}$$

其中 t_1 与 t_2 为任意的两个数($t_1<t_2$)。

特别,当

$$-t_1=+t_2=t,$$

$$P(n,\ -t,\ +t)\approx\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt. \tag{0.4}$$

§1. 正态分布

1. 正态分布的微分函数 拉普拉斯定理告訴我們說，变量

$$\frac{m-np}{\sqrt{npq}}$$

的分布趋于一个極限分布，那就是以

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

为其微分函数的連續分布。

表示这个函数的曲綫，叫做統計量的正态分布曲綫。因此，正态分布是当試驗次数无限增大，同时概率 p 和 q 中沒有一个是很小时的二項式分布的極限情况。

由于正态曲綫的特殊重要性，我們現在要細致地推导出这个曲綫的方程。

正态曲綫的方程是从二項式分布推导出来的。但是二項式分布是不連續的：从一个数值 m 一下就跳到相邻的数值 $m+1$。但是在推导正态曲綫的方程的时候，我們要考虑的不是不連續的，而是連續的分布，也就是說，我們的任务在于求出这样一个分布函数，它不仅对于整数 m 正确，而且对于任何数都正确。

設 $p_{n,np+x}$ 为在部分总体里 n 个样本中，$np+x$ 个具有某种特点，而剩下的 $n-(np+x)=nq-x$ 个不具有某种特点的概率，則根据(0.1)，将有

$$p_{n,np+x}=\frac{n!}{(np+x)!(nq-x)!}p^{np+x}q^{nq-x}=$$

$$=\frac{n!}{(np)!(nq)!}p^{np}q^{nq}\frac{nq(nq-1)\cdots(nq-x+1)}{(np+1)(np+2)\cdots(np+x)}\cdot\frac{p^x}{q^x}=$$

$$=p_{n,np}\frac{\left(1-\frac{1}{nq}\right)\left(1-\frac{2}{nq}\right)\cdots\left(1-\frac{x-1}{nq}\right)}{\left(1+\frac{1}{np}\right)\left(1+\frac{2}{np}\right)\cdots\left(1+\frac{x}{np}\right)}.$$

由此

$$\ln\frac{p_{n,np+x}}{p_{n,np}}=$$

$$=\sum_{k=1}^{x}\ln\left(1-\frac{k}{nq}\right)-\sum_{k=1}^{x}\ln\left(1+\frac{k}{np}\right)-\ln\left(1-\frac{x}{nq}\right).$$

把右方两个对数展成級数，得

$$\ln\frac{p_{n,np+x}}{p_{n,np}}=\sum_{k=1}^{x}\left(-\frac{k}{nq}-\frac{1}{2}\frac{k^2}{(nq)^2}-\frac{1}{3}\frac{k^3}{(nq)^3}-\cdots\right)-$$

$$-\sum_{k=1}^{x}\left(\frac{k}{np}-\frac{1}{2}\frac{k^2}{(np)^2}+\frac{1}{3}\frac{k^3}{(np)^3}-\cdots\right)-\ln\left(1-\frac{x}{nq}\right)=$$

$$=-\sum_{k=1}^{x}\left(\frac{k}{nq}+\frac{k}{np}\right)-\frac{1}{2}\sum_{k=1}^{x}\left(\frac{k^2}{(nq)^2}-\frac{k^2}{(np)^2}\right)-$$

$$-\frac{1}{3}\sum_{k=1}^{x}\left(\frac{k^3}{(nq)^3}+\frac{k^3}{(np)^3}\right)-\cdots-\ln\left(1-\frac{x}{nq}\right)=$$

$$=-\frac{p+q}{npq}\sum_{k=1}^{x}k-\frac{p^2-q^2}{2(npq)^2}\sum_{k=1}^{x}k^2-$$

$$-\frac{p^3+q^3}{3(npq)^3}\sum_{k=1}^{x}k^3-\cdots-\ln\left(1-\frac{x}{nq}\right).$$

把自然数各次冪的和的值代入上式，得

$$\ln\frac{p_{n,np+x}}{p_{n,np}}=-\frac{x(x+1)}{2}\frac{p+q}{npq}-\frac{1}{2}\frac{x(x+1)(2x+1)}{6}\frac{p^2-q^2}{(npq)^2}-$$

$$-\frac{1}{3}\frac{x^2(x+1)^2}{4}\frac{p^3+q^3}{(npq)^3}-\cdots+\left(\frac{x}{nq}+\frac{1}{2}\frac{x^2}{(nq)^2}+\frac{1}{3}\frac{x^3}{(nq)^3}+\cdots\right).$$

为了更进一步的推导，須注意

$$p+q=1,$$

$$p^2-q^2=(p+q)(p-q)=p-q,$$

$$p^3+q^3=(p+q)(p^2-pq+q^2)=(p+q)^2-3pq=1-3pq.$$

再令

$$x=t\sigma, \tag{A}$$

其中

$$\sigma=\sqrt{npq},$$

我們得

$$\ln\frac{p_{n,np+x}}{p_{n,np}}=-\frac{t^2\sigma^2+t\sigma}{2\sigma^2}+\frac{2t^3\sigma^3+3t^2\sigma^2+t\sigma}{12\sigma^4}(q-p)-$$
$$-\frac{t^4\sigma^4+2t^3\sigma^3+t^2\sigma^2}{12\sigma^6}(1-3pq)-\cdots+$$
$$+\left(\frac{t\sigma p}{\sigma^2}+\frac{t^2\sigma^2p^2}{2\sigma^4}+\frac{t^3\sigma^3p^3}{3\sigma^6}+\cdots\right)=$$
$$=-\frac{t^2}{2}+\frac{t}{\sigma}\left[-\frac{1}{2}+\frac{t^2}{6}(q-p)+p\right]+$$
$$+\frac{t^2}{\sigma^2}\left[\frac{q-p}{4}-\frac{t^2}{12}(1-3pq)+\frac{p^2}{2}\right]+\tau,$$

其中最后一項τ，是在分母上含有σ^3, σ^4,…的所有各項的总和。

自此以后，我們假設数n无限增大，而且仅仅考虑那些和$\sqrt{npq}$比較起来还是不可忽略的数值x。这样，我們就可以在最后一个式子中忽略所有在分母中含有σ的項(换言之，就是含有$\sqrt{n}$)。

在这种情况下，上式就变成

$$\ln\frac{p_{n,np+x}}{p_{n,np}}=-\frac{t^2}{2}.$$

由此

$$p_{n,np+x}=p_{n,np}e^{-\frac{t^2}{2}}.$$

根据(A)

$$t^2=\frac{x^2}{\sigma^2},$$

代入得

$$p_{n,np+x}=p_{n,np}e^{-\frac{x^2}{2\sigma^2}}.$$

因为根据(0.2)

$$p_{n,np}=\frac{1}{\sigma\sqrt{2\pi}},$$

因此我們看到，当n足够大时，概率$p_{n,np+x}$仅是一个变量x的函数。于是我們可以写

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}. \tag{1.1}$$

这样，我們就得到了統計量X的正态分布的微分函数。

公式(1.1)指出，如果統計量的平均值 $\bar{X}=np$ (它标志着分布曲綫的位置)为已知，則正态曲綫的形状就完全决定于唯一的参数标准差 σ. 知道了这个参数之后，用公式(1.1)計算出分布曲綫任意点的縱坐标，这样我們就能得到所研究的統計量的分布的全面形象.

正态分布的微分函数还可以用另外一些形式表示.

如果以标准差 σ 为單位来表示与平均值的差 x (亦即取 σ 作为度量單位)，則令

$$t=\frac{x}{\sigma},$$

我們得

$$f(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}. \tag{1.2}$$

如果以模数 $M=\sigma\sqrt{2}$ 为單位来表示与平均值的差，則令

$$z=\frac{x}{\sigma\sqrt{2}},$$

我們得

$$f(z)=\frac{1}{\sqrt{\pi}}e^{-z^2}. \tag{1.3}$$

在有些情况下，用准确度

$$h=\frac{1}{\sigma\sqrt{2}}$$

来代替标准差和模数.

在这种情况下，正态分布的微分函数的形式为

$$f(x)=\frac{h}{\sqrt{\pi}}e^{-h^2x^2}. \tag{1.4}$$

2. 正态分布微分函数的研究 研究分布的微分函数(1.1)，我們就能够得到关于这个函数的变化特性的完整概念，因此就可能描繪出由方程(1.1)表示的正态曲綫.

1) 首先确定函数(1.1)的存在区間. 因为正态分布的微分函数当自变量 x 为任意值时都有实数值，所以这个函数的存在区間为

$$|-\infty,\ +\infty|. \tag{1.5}$$

2) 不管我們取的 x 为何值 (正或負), 函数 (1.1) 总有正值. 卽

$$f(x)>0. \tag{1.6}$$

这个結論完全符合于我們所研究的函数的本質，因为这个函数本是概率的密度.

根据(1.6)，我們得出結論，正态曲綫分布在 OX 軸的上面.

3) 因为函数 (1.1) 是偶函数(只有变量 x 的二次冪在方程中出現)，所以以 $-x$ 代替 $+x$，函数不变：卽

$$f(+x)=f(-x). \tag{1.7}$$

所以正态曲綫对 OY 軸来說是对称的. OY 軸把曲綫下的面积分成二个相等的部分；并且中位数 $\mathring{X}$ 和坐标原点重合. 同时注意在方程(1.1)的推导过程中，取平均值 $\overline{X}$ 为計算起点. 因此在正态分布中

$$\overline{X}=\mathring{X}.$$

4) 令 $f(x)=0$，則得出的方程 $e^{-\frac{x^2}{2\sigma^2}}=0$ 沒有解；因此正态曲綫与 OX 軸沒有交点. 因为所研究的函数是負冪的，并且变量 x 是以二次冪出現的，所以 x 往正方向或負方向增加时，函数 $f(x)$ 很快的減小. 当 x 趋于 $+\infty$ 或 $-\infty$ 时，函数的冪趋于 $-\infty$，因此全式趋于零；卽

$$\lim_{x\to\pm\infty} f(x)=0. \tag{1.8}$$

所以，当从左边或右边无限制地离开坐标原点时，曲綫无限地接近于 OX 軸：所以 OX 軸是曲綫的漸近綫.

5) 当 $x=0$ 时，也就是說在对应于平均值 $\overline{X}$ 的点处，

$$f(0)=\frac{1}{\sigma\sqrt{2\pi}}. \tag{1.9}$$

因此，正态曲綫在 OY 軸上所截取的綫段等于

$$\frac{1}{\sigma\sqrt{2\pi}}=\frac{0.3989}{\sigma}\approx 0.4\frac{1}{\sigma}.$$

因为当 x 为所有的其它值时

$$f(\pm x)<f(0),$$

所以量(1.9)是正态曲綫(1.1)的最大縱坐标。

6) 为了証实这一点，我們求出函数(1.1)的一次导数：即

$$f'(x)=-\frac{x}{\sigma^3\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}. \tag{1.10}$$

因为这个导数当 $x=0$ 时为零，$x<0$ 时为正，$x>0$ 时为負，所以函数 $f(x)$ 当 $x<0$ 时上升，$x>0$ 时下降，在 $x=0$ 时有最大值。

由此看出，在正态分布中，三个特征数——平均值、中位数和众数的值彼此重合：即

$$\overline{X}=\mathring{X}=\hat{X}. \tag{1.11}$$

7) 現在我們来确定正态曲綫的向下弯曲和向上弯曲的所在区間及此曲綫的拐点。

为此，需要求出函数(1.1)的二次导数。把(1.10)微分，得

$$f''(x)=-\frac{1}{\sigma^3\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}+\frac{x^2}{\sigma^5\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}=$$

$$=\frac{1}{\sigma^5\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}(x^2-\sigma^2). \tag{1.12}$$

因为在此等式右边括号前面的因子当 x 为任意值时是正数，所以二次导数的符号决定于第二項

$$x^2-\sigma^2.$$

如果

$$|x|<\sigma,$$

即

$$-\sigma<x<\sigma,$$

則

$$f''(x)<0;$$

所以在 $-\sigma$ 到 $+\sigma$ 的范圍内，曲綫向下弯曲。

同样，如果

$$|x|>\sigma,$$

即

$$x<-\sigma \text{ 或 } x>+\sigma,$$

則

$$f''(x)>0;$$

所以在 $-\sigma$ 到 $+\sigma$ 的范圍之外，曲綫向上弯曲。

最后，令二次导数等于零，得

$$x^2-\sigma^2=0.$$

由此

$$x=\pm\sigma. \tag{1.13}$$

因为当 x 經过这些值时二次导数改变符号，所以当 x 为这些值时曲綫有拐点。

由此可知，正态曲綫的拐点分布在相当于平均值的坐标原点的左右，拐点到坐标原点的距离等于标准差。

根据上面的研究，我們可以描繪出正态曲綫的形状：这个曲綫分布在横軸的上面且对縱軸是对称的；当 $x=0$ 时有極大值；在从 $x=-\sigma$ 到 $x=+\sigma$ 的范圍內，曲綫向下弯曲，而在此范圍之外曲綫向上弯曲；$x=\pm\sigma$ 是曲綫的拐点；随着 x 絕对值的增加曲綫迅速下降，并且横坐标軸是曲綫的漸近綫。

为了描繪正态曲綫，首先要求出拐点的縱坐标及拐点切綫的斜率。

将(1.13)代入(1.1)，求出拐点的縱坐标为

$$f(\pm\sigma)=+\frac{1}{\sigma\sqrt{2\pi e}}.$$

同样将(1.13)代入(1.10)，求出拐点切綫的斜率为

$$f'(\pm\sigma)=\mp\frac{1}{\sigma^2\sqrt{2\pi e}}.$$

因此，在拐点的坐标为

$$x=+\sigma,\quad y=\frac{1}{\sigma\sqrt{2\pi e}}$$

时，切綫方程

$$Y-y=f'(x)[X-x]$$

成为

$$Y-\frac{1}{\sigma\sqrt{2\pi e}}=-\frac{1}{\sigma^2\sqrt{2\pi e}}[X-\sigma].$$

由此求出，在坐标軸上被这直綫所截取的綫段等于

$$2\sigma \text{ 和 } \frac{2}{\sigma\sqrt{2\pi e}},$$

卽比相应的拐点的坐标大一倍.

因此,为了将正态曲綫画成圖形(圖 1),我們在横軸上截取綫

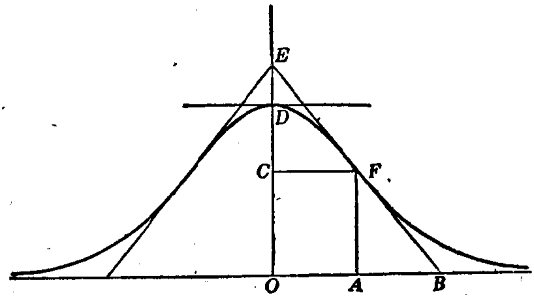

圖 1. 正态曲綫的构圖

段 OA 和 OB,各等于 σ 和 2σ,在縱軸上截取綫段 OC, OD 和 OE,使它們分别等于

$$\frac{1}{\sigma\sqrt{2\pi e}} \approx \frac{0.2420}{\sigma},$$

$$\frac{1}{\sigma\sqrt{2\pi}} \approx \frac{0.3989}{\sigma},$$

$$\frac{2}{\sigma\sqrt{2\pi e}} \approx \frac{0.4840}{\sigma}.$$

从 A 点引平行于 OY 軸的直綫,从 C 点引平行于 OX 軸的直綫.此二直綫的交点就是曲綫的拐点;通过 B 点和 F 点引一和 OY 軸相交于 E 点的直綫. 此后在三角形 FEC 中作曲綫 DF,这条曲綫在 D 点有平行于 OX 軸的切綫,而在 F 点有切綫 BE. 其次将曲綫延長和切綫 BE 相交于 F 点,这个曲綫漸近于 OX 軸.

用上法作出曲綫的右半部分,而后以 OY 軸为对称軸把曲綫反映过去;这样我們就得到了曲綫的左半部分.

为了更精确地画出正态曲綫,須用这条曲綫的縱坐标的数值表.

3. 正态分布的积分函数 具有以 1 为标准差的正态分布的

积分函数的形式是

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt, \tag{1.14}$$

它代表着統計量 X 小于 x 的概率

$$P\{X<x\}.$$

統計量 X 的值位于 $-x$ 到 $+x$ 的区間內的概率

$$P\{-x<X<+x\}$$

等于积分

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-x}^{+x}e^{-\frac{t^2}{2}}dt. \tag{1.15}$$

由于正态曲綫的对称性,我們有

$$\Phi(x)=\frac{2}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{t^2}{2}}dt. \tag{1.16}$$

函数 $\Phi(x)$ 叫作概率积分.

我們来看函数 $\Phi(x)$ 的基本特性.

1) 函数 $\Phi(x)$ 是奇函数,也就是說,当用 $-x$ 代替 $+x$ 时,函数 $\Phi(x)$ 变号而不变值.

这是因为,如果在积分

$$\Phi(-x)=\frac{2}{\sqrt{2\pi}}\int_{0}^{-x}e^{-\frac{t^2}{2}}dt$$

中作代换,使

$$t=-z,\ \ dt=-dz,$$

則当 $t=-x$ 时

$$z=x,$$

所以

$$\Phi(-x)=-\frac{2}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{z^2}{2}}dz,$$

即

$$\Phi(-x)=-\Phi(x). \tag{1.17}$$

2) 随着 x 由 0 增加到 ∞, 函数 $\Phi(x)$ 也从 0 点开始增大,起初增加得很快,然后逐渐减慢,并逐渐趋于極限值 1, 即

$$\lim_{x\to\infty}\Phi(x)=\frac{2}{\sqrt{2\pi}}\int_0^{\infty}e^{-\frac{t^2}{2}}dt=1. \qquad (1.18)$$

这是因为，如果把函数 $\Phi(x)$ 写成(1.15)的形式并且在那里讓 $x=\infty$ (也就是使积分域成为 t 的可能值的全部)，那么，作为概率相加定理的一个后果，我們得到：

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=1. \qquad (1.19)$$

从(1.19)得

$$\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=\sqrt{2\pi}. \qquad (1.20)$$

令

$$\frac{t}{\sqrt{2}}=z,$$

并且注意，当 $t\to\infty$ 时

$$z\to\infty,$$

我們将有

$$\frac{dt}{\sqrt{2}}=dz,$$

所以

$$\int_{-\infty}^{+\infty}e^{-z^2}dz=\sqrt{\pi}. \qquad (1.21)$$

最后，令

$$z=\frac{x}{\sigma\sqrt{2}},$$

則有

$$\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2\sigma^2}}dx=\sigma\sqrt{2\pi}, \qquad (1.22)$$

所以

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2\sigma^2}}dx=1. \qquad (1.23)$$

4. 貝努里定理——拉普拉斯-馬尔柯夫的証明 在解决統計学中的許多問題时，都要用到概率积分. 特別說来，概率积分是貝努里定理的分析証明的一个組成部分.

貝努里定理断言，当独立試驗的次数 n 足够大时，有

$$P\left\{-\varepsilon<\frac{m}{n}-p<+\varepsilon\right\}>1-\eta,$$

其中 ε 和 η 是任意小的正数。

为了証明这个定理，我們研究当 t 为某一值时，不等式

$$-t\sqrt{npq}<m-np<t\sqrt{npq}$$

的概率 $P(t)$。这个不等式和不等式

$$-t\sqrt{\frac{pq}{n}}<\frac{m}{n}-p<t\sqrt{\frac{pq}{n}} \tag{a}$$

是一样的。

依照拉普拉斯定理 (0.4)，这不等式的概率 $P(t)$，当 t 固定而 n 无限增大时，趋于極限

$$\frac{1}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt.$$

但是从上面的(1.20)知

$$\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=\sqrt{2\pi},$$

所以

$$\int_0^{\infty}e^{-\frac{t^2}{2}}dt=\frac{\sqrt{2\pi}}{2},$$

而这个等式指出，当 t 的值为足够大时，差

$$1-\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt$$

将变成任意地小。

因此，把正数 η 分成二个正值 η' 和 η''，亦即

$$\eta=\eta'+\eta'' \qquad (\eta'>0,\ \eta''>0),$$

我們就能够取一个 t 值使下式成立，

$$\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt=1-\eta'', \tag{б}$$

然后定出这样大的数 n_0，使当所有値 n 符合于不等式

$$n>n_0 \tag{1.24}$$

时，差

$$\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt-P(t)$$

小于 η''：

$$\frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}dt-P(t)<\eta''. \qquad (\text{в})$$

这样确定了 t 的值以后，我們再要求，除了不等式 (1.24) 以外，还要有不等式

$$n>\frac{pqt^2}{\varepsilon^2}. \qquad (1.25)$$

这时，就有

$$\varepsilon>t\sqrt{\frac{pq}{n}},$$

因此不等式

$$-\varepsilon<\frac{m}{n}-p<\varepsilon \qquad (\text{г})$$

的概率将大于不等式（a）的概率 $P(t)$，因为所有适合于不等式（a）的 m 值，同样适合于不等式（г）。

根据（б）和（в），不等式（a）的概率 $P(t)$ 大于

$$1-\eta'-\eta''=1-\eta.$$

因此，对于滿足不等式(1.24)和(1.25)的所有的 n 值，都有

$$P\left\{-\varepsilon<\frac{m}{n}-p<\varepsilon\right\}>1-\eta. \qquad (1.26)$$

这样就証明了貝努里定理。