概率分布

维基百科，自由的百科全书

概率分布（德語：Wahrscheinlichkeitsverteilung；英語：probability distribution）或簡稱**分布**，是概率論的一個概念。使用時可以有以下兩種含義：

- 廣義地，它指稱隨機變量的概率性質－－當我們說概率空間$(\Omega, \mathcal{F}, \mathbb{P})$中的兩個隨機變量$X$和$Y$具有同樣的分布時，我們是無法用概率$\mathbb{P}$來區別他們的。換言之：

稱X和Y為同分布的隨機變量，當且僅當對任意事件$A \in \mathcal{F}$，有$\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$成立。

但是，不能認為同分布的隨機變量是相同的隨機變量。事實上即使X與Y同分布，也可以没有任何點ω使得$X(\omega)=Y(\omega)$。在這個意義下，可以把隨機變量分類，每一類稱作一個分布，其中的所有隨機變量都同分布。用更簡要的語言來說，同分布是一種等價關係，每一個等價類就是一個分布。需注意的是，通常談到的離散分布、均匀分布、伯努利分布、正態分布、泊松分布等，都是指各種類型的分布，而不能視作一個分布。

- 狹義地，它是指**隨機變量的概率分布函數**。設X是樣本空間$(\Omega, \mathcal{F})$上的隨機變量，$\mathbb{P}$為概率測度，則稱如下定義的函數是X的分布函數（德語：Verteilungsfunktion，英語：distribution function），或稱累積分布函数（德語：kumulative Verteilungsfunktion，英語：cumulative distribution function，簡稱CDF）：

$F_X(a) = \mathbb{P}(X \leq a)$，對任意實數$a$定義。

具有相同分布函數的隨機變量一定是同分布的，因此可以用分布函數來描述一個分布，但更常用的描述手段是概率密度函數（德語：Wahrscheinlichkeitsdichtefunktion，英語：probability density function, pdf）。

- 在常用的文獻中，「分布」一詞可指其廣義和狹義，而「累計分布函數」或「分布函數」一詞只能指稱後者。為了不致混淆，下文中談及上述的廣義時使用「分布」一詞；狹義時使用「分布函數」一詞。

目录

分布函數的性質刻劃

對於特定的隨機變量 X，其分布函數F_X是單調不減及右連續，而且$F_X(-\infty)=0$，$F_X(\infty)=1$。這些性質反過來也描述了所有可能成為分布函數的函數：

- 設 $F:[-\infty,\infty]\to[0,1], F(-\infty)=0, F(\infty)=1$且單調不減、右連續，則存在概率空間 $(\Omega,\mathcal{F},\mathbb{P})$及其上的隨機變量 X ，使得 F 是 X 的分布函數，即 $F_X=F$

隨機變量的分布

設 P 為概率測度，X 為隨機變量，則函數 $F(x)=P(X\leq x),(x\in\mathbb{R})$ 稱為 X 的概率分布函數。如果将 X 看成是数轴上的随机点的坐标，那么，分布函数 $F(x)$ 在 x 处的函数值就表示 X 落在区间 $(-\infty,x]$ 上的概率。

例如，設隨機變量 X 為擲兩次骰子所得的點數差，而整个樣本空間由 36 个元素组成。

数量	$(i,j)\in S$	x	$P(X=x)$	$F(x)$
6	(1,1)，(2,2)，(3,3) (4,4)，(5,5)，(6,6)	0	6/36	6/36
10	(1,2)，(2,3) (3,4)，(4,5)，(5,6) (2,1)，(3,2)，(4,3) (5,4)，(6,5)	1	10/36	16/36
8	(1,3)，(2,4)，(3,5) (4,6)，(3,1)，(4,2) (5,3)，(6,4)	2	8/36	24/36
6	(1,4)，(2,5)，(3,6) (4,1)，(5,2)，(6,3)	3	6/36	30/36
4	(1,5)，(2,6) (5,1)，(6,2)	4	4/36	34/36
2	(1,6)，(6,1)	5	2/36	36/36

其分布函数是：

$$F(x)=\begin{cases}0,x<0\\6/36,0\le x<1\\16/36,1\le x<2\\24/36,2\le x<3\\30/36,3\le x<4\\34/36,4\le x<5\\1,5\le x\end{cases}$$

离散分布

上面所列举的例子属于离散分布，即分布函数的值域是離散的，比如只取整数值的隨機變量就是屬於離散分布的。$F(x)$表示随机变量$X\le x$的概率值。如果X的取值只有$x_1<x_2<\ldots<x_n$，则：

- $F_X(x_i)=\sum_{j=1}^{i}P(x_j)$
- $\sum_{k=1}^{n}P(x_k)=1$

均匀分布

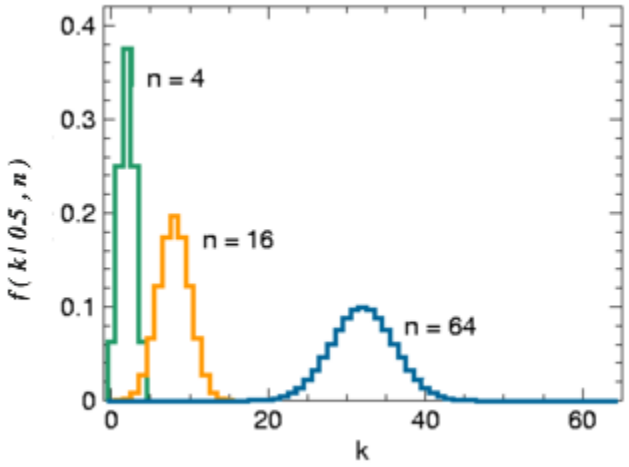
二项分布

二项分布是最重要的离散概率分布之一，由瑞士数学家雅各布·伯努利（Jakob Bernoulli）所发展，一般用二项分布来计算概率的前提是，每次抽出样品后再放回去，并且只能有两种试验结果，比如黑球或红球，正品或次品等。二项分布指出，随机一次试验出现的概率如果为p，那么在n次试验中出现k次的概率为：

$$f(n,k,p)=\binom{n}{k}p^k(1-p)^{n-k}$$

例如，在掷3次骰子中，不出现6点的概率是：$f(3,0,\frac{1}{6})=\binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3=0.579$

在连续两次的轮盘游戏中，至少出现一次红色的概率为：$f(2,1,\frac{18}{37})+f(2,2,\frac{18}{37})=0.736$



二项分布在$p=0.5$时的对称性（自变量为k）

二项分布在$p=0.5$时表现出图像的对称性，而在p取其它值时是非对称的。另外二项分布的期望值$\mathrm{E}(X)=np$，以及方差$\mathrm{var}(X)=n\,p\,(1-p)$

几何分布

超几何分布

作为离散概率分布的超几何分布尤其指在抽样试验时抽出的样品不再放回去的分布情况。在一个容器中一共有N个球，其中M个黑球，$(N-M)$个红球，通过下面的超几何分布公式可以计算出，从容器中抽出的n个球中（抽出的球不放回去）有k个黑球的概率是多少：

$$f(k,n;M;N):=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

例如，容器中一共10个球，其中6个黑色，4个白色，一共抽5次（抽出的球不放回去），在这5个球中有3个黑球的概率是：$f(k=3)=\frac{\binom{6}{3}\binom{10-6}{5-3}}{\binom{10}{5}}=0.476$

超几何分布与二项分布的关系

和二项分布不同的是，在超几何分布中，特别强调的是抽出的样品在下一次抽取前不再放回去，但是如果抽取的次数n和总共样品数N相比很小（大约$n/N < 0,05$），这时在计算上二项分布和超几何分布相互间则没有主要的区别，此时人们更愿意采用二项分布的方法，因为在数学计算上二项分布要简单一些。

泊松近似

泊松近似是二项分布的一種極限形式。其强调如下的试验前提：一次抽样的概率值p相对很小，而抽取次数n值又相对很大。因此泊松分布又被称之为罕有事件分布。泊松分布指出，如果随机一次试验出现的概率为p，那么在n次试验中出现k次的概率按照泊松分布应该为：

$$f(n,k,p) = \frac{(n \cdot p)^k}{e^{n \cdot p} \cdot k!}$$

其中，数学常数$e = 2.71828...$(自然對數的底数)
例如，某工厂在生产零件时，每200个成品中会有1个次品，那么在100个零件中最多出现2个次品的概率按照泊松分布应该是：$f(100,0,\frac{1}{200}) + f(100,1,\frac{1}{200}) + f(100,2,\frac{1}{200}) = 0.986$

在实践中如果遇到n值很大導致二项分布難於計算时，可以考慮使用泊松分布，但前提是$n \cdot p$必須趨於一個有限极限。采用泊松分布的一个不太严格的规则（通过展开二项分布，并在形式上化简为类似泊松分布后，利用极限化简即可得）是：

1. $n \geq 100$
2. $p \leq 0.1$

连续分布

設X是具有分布函数F的连续随机变量，且F的一階導數處處存在，則其導函數

$$f(x) = \frac{\mathrm{d}\,F(x)}{\mathrm{d}\,x}$$

称为X的機率密度函数。
每个機率密度函数都有如下性质：

- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$
- $\int_{a}^{b} f(x)\,\mathrm{d}x = \mathrm{P}(a \leq X \leq b) = F(b) - F(a)$

第一个性质表明，機率密度函数与x轴形成的区域的面积等于1，第二个性质表明，连续随机变量在区间$[a,b]$的概率值等于密度函数在区间$[a,b]$上的积分，也即是与X轴在$[a,b]$内形成的区域的面积。因为$0 \leq F(x) \leq 1$，且$f(x)$是$F(x)$的导数，因此按照积分原理不难推出上面两个公式。

正态分布、指数分布、t-分布，F-分布以及Ξ^2-分布都是連續分布。

均匀分布

指数分布

伽马分布

正态分布

连续随机变量的機率密度函数如果是如下形式，

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

那么这个连续分布被称之为正态分布，或者高斯分布。其密度函数的曲线呈对称钟形，因此又被称之为钟形曲线，其中μ是平均值，σ是标准差。正态分布是一种理想分布，许多典型的分布，比如成年人的身高，汽车轮胎的运转状态，人类的智商值（IQ），都属于或者说至少接近正态分布。同样按照连续分布的定义，常態機率密度函数具有和普通機率密度函数类似的性质：

- $\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1$
- $F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)}\,\mathrm{d}t$

如果给出一个正态分布的平均值μ以及标准差σ，可以根据上面的第二个公式计算出任一区间的概率分布情况。但是如上的计算量是相当庞大的，没有计算机的辅助基本是不可能的，解决这一问题的方法是借助z-变换以及标准正态分布表格（z-表格）。

中间值$\mu = 0$以及标准差$\sigma = 1$的正态分布被称之为标准正态分布，其累積分布函数是

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{-\frac{1}{2}t^2}\,\mathrm{d}t$$

将普通形式的正态分布变换到标准正态分布的方法是

$$z = \frac{x - \mu}{\sigma}$$

例如，已知一正态分布的$\mu = 5$，$\sigma = 3$，求区间概率值$P(4 < X \leq 7)$?计算过程如下，

$$\frac{4-5}{3} < Z \leq \frac{7-5}{3}$$
$$-1/3 < Z \leq 2/3$$
$$P(4 < X \leq 7) = P(-1/3 < Z \leq 2/3) = \Phi(2/3) - \Phi(-1/3) = 0.7475 - 0.3694 = 0.3781$$

其中$\Phi(z)$值通过查z-表格获得。

正态分布与二项分布的关系

在离散分布中如果试验次数n值非常大，而且单次试验的概率p值又不是很小的情况下，正态分布可以用来近似的代替二项分布。一个粗略的使用正态分布的近似规则是：$n \cdot p \cdot (1-p) \geq 9$。
从二项分布中获得μ和σ值的方法是

- 期望值$\mu = n \cdot p$
- 標準差$\sigma = \sqrt{n \cdot p \cdot (1-p)}$

如果$\sigma > 3$，则必须采用下面的近似修正方法：

$$P(x_1 \leq X \leq x_2) = \underbrace{\sum_{k=x_1}^{x_2} \binom{n}{k} \cdot p^k \cdot (q)^{n-k}}_{\mathrm{EF}} \approx \underbrace{\Phi\left(\frac{x_2 + 0.5 - \mu}{\sigma}\right) - \Phi\left(\frac{x_1 - 0.5 - \mu}{\sigma}\right)}_{\mathrm{ZF}}$$

（注：$q = 1 - p$；EF：二项分布；ZF：正态分布）

上（下）临界值分别增加（减少）修正值0.5的目的是在σ值很大时获得更精确的近似值，只有σ很小时，修正值0.5可以不被考虑。

例如，随机试验为连续64次掷硬币，获得的国徽数位于32和42之间的概率是多少？用正态分布计算如下，

$$\begin{aligned}
\mu &= n \cdot p = 64 \cdot 0.5 = 32 \\
\sigma &= \sqrt{n \cdot p \cdot (1-p)} = \sqrt{64 \cdot 0.5 \cdot 0.5} = 4
\end{aligned}$$

$n \cdot p \cdot q = 16 \geq 9$，符合近似规则，应用$z$-变换：

$$\begin{aligned}
P(32 \leq X \leq 42) &\approx \Phi\left(\frac{42 + 0.5 - 32}{4}\right) - \Phi\left(\frac{32 - 0.5 - 32}{4}\right) \\
&= \Phi\left(2.63\right) - \Phi\left(-0.13\right) = 0.0517 + 0.4957 = 0.5474
\end{aligned}$$

z-表格，区域：从 0 到 z

z	0	1	2	3	4	5	6
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979

标准正态分布$N(0,1)$下的z-表格

在运用z-表格时注意到利用密度函数的对称性来求出z为负值时的区域面积。

参考文献

- 彼得·缺菲尔（Peter Zoefel）：《统计和经济学家》（德文）. PEASON Studium出版社，2003年. ISBN 3-8273-7062-0.
- 约瑟夫·西拉（Josef Schira）：《统计理论与企业管理》（德文）. PEASON Studium出版社，2003年. ISBN 3-8273-7041-8.
- 汉斯-底特·黑伯曼（Hans-Dieter Hippmann）：《统计学》（德文）. SCHAEFFER POESCHEL出版社，2003年. ISBN 3-7910-2119-2.

外部链接

概率分布Java演示

- 二项分布Java演示 (https://web.archive.org/web/20070613031452/http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-bin.html)
- 超几何分布Java演示 (https://web.archive.org/web/20070613194829/http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-hyp.html)
- 泊松分布Java演示 (https://web.archive.org/web/20070613195109/http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-poi.html)
- 正态分布Java演示 (https://web.archive.org/web/20070613194335/http://www.uni-konstanz.de/FuF/wiwi/heiler/os/vt-norm.html)

参見

- 概率论
- 隨機變數
- 累积分布函数
- 概率密度函数
- 概率質量函数
