概率论和数理统计

2025-2026 秋冬

黄炜
hwmath2003@zju.edu.cn

浙江大学数学科学学院

Chapter 2 Random Variables and Distributions

关键词 (Keywords):

- 随机变量 (Random Variable)
- 离散型随机变量 (Discrete Random Variable)
- 连续型随机变量 (Continuous Random Variable)
- (概率) 分布函数 ((Probability) Distribution Function)
- 随机变量的函数 (Functions of Random Variable)

§2.1 随机变量 (Random Variables)

We shall not always be interested in an experiment itself, but rather in some consequence of its random outcome. For example, many gamblers are more concerned with their losses than with the games which give rise to them. Such consequences, when real valued, may be thought of as functions which map S into a real line $\mathbb{R}$, and these functions are called random variables.

常见的两类试验结果:

- 示数的: 降雨量; 候车人数; 发生交通事故的次数; ......
- 示性的: 明天天气 (晴、多云......); 化验结果 (阴性、阳性); ......

中心问题: 根据研究的目的, 将试验结果数量化.

随机变量 (Random Variable)

Definition 1 (随机变量)

设随机试验的样本空间为 S, 若 $X = X(e)$ 为定义在样本空间 S 上的实值单值函数, 则称 $X = X(e)$ 为随机变量.

Definition 1 (Random Variable)

Let S be the sample space for an experiment. A real-valued function that is defined on S is called a random variable (abbreviated r.v.).

- 一般采用大写英文字母 X, Y, Z 等或希腊字母 ξ, η, ζ 等来表示随机变量.
- 随机变量的自变量具有随机性.

说明:

(1) 随机事件可以表示为 $A=\{e: X(e)\in I\}=\{X\in I\}$, $I\subset\mathbb{R}$.

Example 1

将一枚均匀硬币抛掷 3 次, 观测其落地情况, 则样本空间为

$$S=\{\text{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}\}.$$

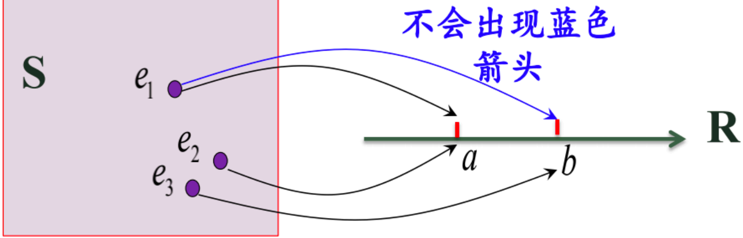
若 X 表示 3 次中出现正面 (head, H) 的次数, 则随机事件

- $A=\{\text{正面出现一次}\}=\{\text{HTT, THT, TTH}\}=\{e: X(e)=1\}=\{X=1\}$.
- $B=\{\text{3 次出现的情况相同}\}=\{\text{HHH, TTT}\}=\{X=0\text{ 或 }3\}$.
- $C=\{\text{正面至少出现一次}\}=\{X\geq 1\}$.

也可以研究随机事件发生的概率.

Example 1 (Cont.)

- $\mathrm{P}(A) = \mathrm{P}\{\text{正面出现一次}\} = \mathrm{P}\{X = 1\} = \mathrm{P}\{\mathrm{HTT, THT, TTH}\} = \dfrac{3}{8}.$
- $\mathrm{P}(B) = \mathrm{P}\{3\ \text{次出现的情况相同}\} = \mathrm{P}\{X = 0\ \text{或}\ 3\} = \mathrm{P}\{\mathrm{HHH, TTT}\} = \dfrac{2}{8} = \dfrac{1}{4}.$
- $\mathrm{P}(C) = \mathrm{P}\{\text{正面至少出现一次}\} = \mathrm{P}\{X \geq 1\} = 1 - \mathrm{P}\{X = 0\} = \dfrac{7}{8}.$



(2) 对于 $i \neq j$, 必有 $\{X = i\} \cap \{X = j\} = \varnothing$.

$$\text{常见的两类随机变量}\begin{cases}\text{离散型随机变量}\\\text{连续型随机变量}\end{cases}$$

§2.2 离散型随机变量 (Discrete Random Variable/Distribution)

Definition 2 (离散型随机变量)

取值至多可列 (即有限或者可列) 的随机变量为离散型随机变量, 其分布为离散型分布.

Definition 2 (Discrete Random Variable/Distribution)

We say that a random variable X is a discrete random variable or X has a discrete distribution that if X can take only a finite number k $(k \geq 1)$ of different values $x_1, \ldots, x_k$ or, at most, an infinite sequence of different values $x_1, x_2, \ldots$.

Definition 3 (概率分布律/列 (Probability Mass Function))

设 X 为离散型随机变量, 若其可能取值为 $x_1, x_2, \ldots, x_k, \ldots$, 即 $S = \bigcup_{k=1}^{+\infty}\{X = x_k\}$, 则称

$$P(X = x_k) = p_k, \quad k = 1, 2, \ldots$$

为 X 的*概率分布律 (Probability mass function)*. 它也可以用以下的列表方式来表示:

X	x_1	x_2	$\cdots$	x_k	$\cdots$
P	p_1	p_2	$\cdots$	p_k	$\cdots$

概率分布律需包含: (1) 离散型 r.v. 的所有可能取值; (2) 每个可能取值取到的概率

概率分布律必满足以下两条性质: (1) $p_k \geq 0, k = 1, 2, \ldots$; (2) $\sum_{k=1}^{\infty} p_k = 1$.

Definition 3 (Probability Mass Function)

Probability Mass Function/p.m.f./Support. If a random variable X has a discrete distribution, the probability mass function (abbreviated p.m.f.) of X is defined as the function $f(\cdot)$ such that for every real number x,

$$f(x) = \mathrm{P}(X = x).$$

The set $\{x : f(x) > 0\}$ is called the support of (the distribution of) X.

Some books refer to the probability mass function as the probability function, or p.f.

Example 2

某人骑自行车从学校到火车站，一路上要经过 3 个独立的交通灯，设各灯工作独立，且设各灯为红灯的概率为 p, $0 < p < 1$. 以 X 表示首次停车时所通过的交通灯数，求 X 的概率分布律.

Solution

易知 X 的可能取值为 0, 1, 2, 3 ($\{X = i\}, i = 0, 1, 2, 3$ 为样本空间 S 的一个划分). 设 $A_i = \{$第 i 个灯为红灯$\}$, 则 $P(A_i) = p$, $i = 1, 2, 3$ 且 A_1, A_2, A_3 相互独立. 于是

$$P(X = 0) = P(A_1) = p; \qquad P(X = 1) = P(\overline{A}_1 A_2) = (1 - p)p;$$

$$P(X = 2) = P(\overline{A}_1 \overline{A}_2 A_3) = (1 - p)^2 p; \qquad P(X = 3) = P(\overline{A}_1 \overline{A}_2 \overline{A}_3) = (1 - p)^3.$$

X	0	1	2	3
P	p	$(1-p)p$	$(1-p)^2 p$	$(1-p)^3$

Example 3

若随机变量 X 的概率分布律为

$$\mathsf{P}(X=k)=\frac{c\lambda^k}{k!},\quad k=0,1,2,\ldots,\lambda>0,$$

求常数 c.

Solution

由概率分布律的性质可知

$$1=\sum_{k=0}^{\infty}\mathsf{P}(X=k)=c\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}=ce^{\lambda}\quad\Longrightarrow\quad c=\mathrm{e}^{-\lambda}.$$

几个重要的离散型随机变量

(1) $0-1(p)$ 分布

Definition 4 ($0-1(p)$ 分布)

若随机变量 X 的概率分布律为

$$P(X=0)=1-p, \quad P(X=1)=p,$$

其中 $0<p<1$, 则称 X 服从参数为 p 的 $0-1$ 分布, 也称为两点分布, 记作 $X\sim 0-1(p)$ (或 $X\sim B(1,p)$).

$0-1(p)$ 分布的分布律还可以写为

$$\mathrm{P}(X=k)=p^k(1-p)^{1-k}, \quad k=0,1.$$

对于一个随机试验, 如果它的样本空间只包含两个元素, 即 $S=\{e_1,e_2\}$, 总能在 S 上定义一个服从 0-1 分布的随机变量

$$X=\begin{cases}1, & e_1\ \text{出现},\\ 0, & e_2\ \text{出现}\end{cases}$$

来描述这个随机试验的结果.

例如: 检查产品的质量是否合格、对新生婴儿的性别进行登记、检验种子是否发芽以及前面多次讨论过的“抛硬币”试验都可以用 0-1 分布的随机变量来描述.

Bernoulli 试验

对一个随机试验, 设 A 是一随机事件, 且 $\mathsf{P}(A)=p\ (0<p<1)$. 若仅考虑事件 A 发生与否, 则可定义一个服从参数为 p 的 0-1 分布的随机变量

$$X=\begin{cases}1, & 若\ A\ 发生;\\ 0, & 若\ A\ 不发生(即\ \overline{A}\ 发生)\end{cases}$$

来描述这个随机试验的结果.

只有两个可能结果的试验, 称为 Bernoulli (贝努利/伯努利) 试验, 故两点分布也称为 Bernoulli 分布.

Example 4

投掷一颗均匀的骰子，考察 6 点是否出现，用 Y 表示该试验的结果，求 Y 的概率分布律.

Solution

由题意可令 $Y = \begin{cases} 1, & \text{抛出的点数为 } 6, \\ 0, & \text{抛出的点数不为 } 6; \end{cases}$ 则 Y 的分布律为

Y	0	1
P	$5/6$	$1/6$

或写为 $\mathrm{P}(Y = k) = (\frac{1}{6})^k (\frac{5}{6})^{1-k}, k = 0, 1.$ 即 $Y \sim 0 - 1(1/6)$.

事实上，Y 也可以看作是掷一次骰子，点数为 6 的次数.

(2) 二项分布

Definition 5 (n 重 Bernoulli 试验)

设试验 E 只有两个可能的结果: A 与 $\overline{A}$, 其中 $\mathrm{P}(A)=p, 0<p<1$. 将 E 独立地重复进行 n 次, 则称这一串重复的独立试验为 n 重 Bernoulli 试验.

独立: 每次试验结果互不影响;
重复: 在相同条件下重复进行.

Example 5

- 独立重复地抛 n 次均匀硬币, 每次只有两个结果: 正面, 反面. P(正面) $=1/2$.
- 将一颗均匀的骰子抛 n 次, 仅关心是否得到 6 点, 即 $A=\{$得到 6 点$\}$, 每次只有两个结果: $A, \overline{A}$. $\mathrm{P}(A)=1/6$.
- 从 52 张牌中有放回地取 n 次, 仅关心是否取到红牌, 即 $A=\{$取到红牌$\}$, 每次只有两个结果: $A, \overline{A}$. $\mathrm{P}(A)=1/2$.(思考: 如果是不放回取呢?)

想了解: n 重 Bernoulli 试验中结果 A 发生次数的统计规律.

设 X 为 n 重 Bernoulli 试验中结果 A 发生的次数, 则 X 的可能取值为 $0, 1, \ldots, n$, 且

$$\mathsf{P}(X = k) = \mathsf{C}_n^k p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

推导 (以 $n = 3$ 为例)

设 $A_i = \{$第 i 次 A 发生$\}$, 则

$$\begin{aligned}
\mathsf{P}(X=0) &= \mathsf{P}(\overline{A}_1\overline{A}_2\overline{A}_3) = (1-p)^3, \\
\mathsf{P}(X=1) &= \mathsf{P}(A_1\overline{A}_2\overline{A}_3 \cup \overline{A}_1 A_2\overline{A}_3 \cup \overline{A}_1\overline{A}_2 A_3) = \mathsf{C}_3^1 p^1 (1-p)^{3-1}, \\
\mathsf{P}(X=2) &= \mathsf{P}(A_1 A_2\overline{A}_3 \cup A_1\overline{A}_2 A_3 \cup \overline{A}_1 A_2 A_3) = \mathsf{C}_3^2 p^2 (1-p)^{3-2}, \\
\mathsf{P}(X=3) &= \mathsf{P}(A_1 A_2 A_3) = p^3.
\end{aligned}$$

一般地, 对 n 重 Bernoulli 试验, 有 $\mathsf{P}(X = k) = \mathsf{C}_n^k p^k (1-p)^{n-k}, k = 0, 1, \ldots, n.$

Definition 6 (二项分布)

若随机变量 X 的概率分布律为

$$P(X=k)=\mathrm{C}_n^k p^k(1-p)^{n-k}, \quad k=0,1,\ldots,n,$$

其中 $n\geq 1, 0<p<1$, 则称 X 服从参数为 n,p 的二项分布 (binomial distribution), 记作 $X\sim B(n,p)$.

易见 $\mathrm{C}_n^k p^k(1-p)^{n-k}\geq 0$, 且由二项式定理可知

$$1=(p+q)^n=\sum_{k=0}^{n}\mathrm{C}_n^k p^k q^{n-k}, \text{ 其中 } q=1-p,$$

故满足概率分布律的性质.

Definition 6 (Binomial Distribution/Random Variable)

The discrete distribution represented by

$$\mathsf{P}(X=k)=\mathsf{C}_n^k p^k(1-p)^{n-k}, \quad k=0,1,\ldots,n$$

is called the binomial distribution with parameters n and p. A random variable with this distribution is said to be a binomial random variable with parameters n and p.

Example 6

有一大批产品，其验收方案如下：先作第一次检验，从中任取 10 件，无次品则接受这批产品，次品数大于 2 拒收；否则作第二次检验，从中任取 5 件，仅当 5 件中无次品才接受这批产品．设产品的次品率为 $p(0 < p < 1)$，求这批产品能被接受的概率．

Solution

设 $A = \{$接受该批产品$\}$．设 X 为第一次抽得的次品数，Y 为第二次抽得的次品数，则 $X \sim B(10, p)$，$Y \sim B(5, p)$，且 $\{X = i\}$ 与 $\{Y = j\}$ 独立．于是

$$\begin{aligned}
\mathsf{P}(A) &= \mathsf{P}(X=0) + \mathsf{P}(1 \le X \le 2 \text{ 且 } Y=0) \\
&\overset{\text{两次检验的结果是独立的}}{=\!=\!=\!=} \mathsf{P}(X=0) + \mathsf{P}(1 \le X \le 2) \cdot \mathsf{P}(Y=0) \\
&= \mathsf{P}(X=0) + [\mathsf{P}(X=1) + \mathsf{P}(X=2)] \cdot \mathsf{P}(Y=0) \\
&= (1-p)^{10} + [10p(1-p)^9 + 45p^2(1-p)^8] \cdot (1-p)^5.
\end{aligned}$$

Example 7

某人骑自行车从学校到火车站，一路上要经过 3 个独立的交通灯，设各灯工作独立，且设各灯为红灯的概率为 $p(0<p<1)$，以 Y 表示一路上遇到红灯的次数.

1. 求 Y 的概率分布律.
2. 求恰好遇到 2 次红灯的概率.

Solution

这是三重 Bernoulli 试验，即 $Y\sim B(3,p)$. 因此

1. Y 的概率分布律为 $\mathrm{P}(Y=k)=\mathrm{C}_3^k p^k(1-p)^{3-k},\ k=0,1,2,3;$
2. 恰好遇到 2 次红灯的概率为 $\mathrm{P}(Y=2)=3p^2(1-p)$.

二项分布常常要求会判别!!!

Example 8

设随机变量 $X \sim B(100, 0.05)$，求 $\mathrm{P}(X \leq 10)$ 和 $\mathrm{P}(X = 10)$.

Solution

由二项分布的概率分布律可知

$$\mathrm{P}(X \leq 10) = \sum_{k=0}^{10} \mathrm{P}(X = k) = \sum_{k=0}^{10} \mathrm{C}_{100}^{k} 0.05^{k} 0.95^{100-k};$$

$$\mathrm{P}(X = 10) = \mathrm{C}_{100}^{10} 0.05^{10} 0.95^{90}.$$

思考：若求 $\mathrm{P}(X > 10)$ 呢?

使用 Excel 表单 (以 2003 版本为例，不同版本具体做法有所不同)

1. 在 Excel 表单的任一单元格输入 “=”
2. 在主菜单中点击 “插入”，选择 “函数 (F)”，在选择类别的下拉式菜单中选择 “统计”
3. 选择 “BINOMDIST” 点击 “确定”
4. 在 “函数参数” 对话框中输入 “Number_s=10, Trials=100, Probability_s=0.05, Cumulative=TRUE” 点击 “确定”，则在单元格中出现 “0.99852759”.

另：若要计算 $P(X = 10)$，只要将上述步骤中 “Cumulative=TRUE” 改为 “Cumulative=FALSE” 即可出现 “0.016715884”

若为 Excel 2010: 在主菜单中点击 “公式”⇒“其他函数”⇒“统计”⇒“BINOM.DIST”，就出现 “函数参数” 对话框.

(3) 泊松 (Poisson) 分布

Definition 7 (泊松 (Poisson) 分布)

若随机变量 X 的概率分布律为

$$P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \quad k=0,1,2,\ldots,$$

其中 $\lambda>0$, 则称 X 服从参数为 λ 的泊松分布 (Poisson distribution), 记作 $X\sim P(\lambda)$.

泊松分布在生活中尤其是一些类似计数过程的问题中有着非常广泛的应用. 例如: 某人一天内收到的微信的数量、一段时间内某保险公司某保险业务的赔保数量、一段时间内某放射性物质发射出的粒子数量、显微镜下某区域中的白血球数量等等.

Example 9

设某汽车停靠站单位时间内候车人数 $X \sim P(4.8)$. 求:

1. 随机观察 1 个单位时间, 至少有 3 人候车的概率;
2. 随机独立观察 5 个单位时间, 恰有 4 个单位时间至少有 3 人候车的概率.

Solution

1. $\mathrm{P}(X \geq 3) = 1 - \sum_{k=0}^{2} \mathrm{P}(X = k) = 1 - \mathrm{e}^{-4.8}\left(1 + 4.8 + \frac{4.8^2}{2!}\right) = 0.8580.$
2. 设 5 个单位时间内有 Y 个单位时间是“至少有 3 人候车”, 则 $Y \sim B(5, p)$, 其中 $p = \mathrm{P}(X \geq 3) = 0.8580$, 于是

$$\mathrm{P}(Y = 4) = \mathrm{C}_5^4 p^4 (1 - p) = 0.7696.$$

二项分布与 Poisson 分布有以下近似公式:

当 $n > 10, p < 0.1$ 时,

$$\mathrm{C}_n^k p^k (1-p)^{n-k} \approx \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!},$$

其中 $\lambda = np$.

Proof

事实上, 当 n 充分大时, 对适当的 λ,

$$\frac{n(n-1)\cdots(n-k+1)}{n^k} \approx 1, \quad \left(1-\frac{\lambda}{n}\right)^k \approx 1, \quad \left[\left(1-\frac{\lambda}{n}\right)^{-n/\lambda}\right]^{-\lambda} \approx \mathrm{e}^{-\lambda}.$$

因此

$$\mathrm{C}_n^k p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \approx \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}.$$

Example 10

某地区一个月内每 200 个成年人中有 1 个会患上某种疾病，且设各人是否患病相互独立．若该地区一社区有 1000 个成年人，求某月内该社区至少有 3 人患病的概率．

Solution

设该社区 1000 人中有 X 个人患病，则 $X \sim B(1000, p)$，其中 $p = 1/200$．则

$$\mathrm{P}(X \geq 3) = 1 - \mathrm{P}(X = 0) - \mathrm{P}(X = 1) - \mathrm{P}(X = 2) = 0.8760.$$

若利用 Poisson 分布进行近似计算，$\lambda = np = 5$，则

$$\mathrm{P}(X \geq 3) \approx 1 - \mathrm{e}^{-5} - 5\mathrm{e}^{-5} - \frac{5^2}{2!}\mathrm{e}^{-5} = 0.8753.$$

使用 Excel 表单 (以 2003 版本为例，不同版本具体做法有所不同)

1. 在 Excel 表单的任一单元格输入 "="
2. 在主菜单中点击 "插入"，选择 "函数 (F)"，在选择类别的下拉式菜单中选择 "统计"
3. 选择 "POISSON" 点击 "确定"
4. 在在 "函数参数" 对话框中输入

   "X=2, Mean=5, Cumulative=TRUE"

   点击 "确定"，则在单元格中出现 "0.124652019".

进而 $\mathrm{P}(X \geq 3) = 1 - \mathrm{P}(X \leq 2) = 0.875347981.$

(4) 其它离散型随机变量

Example 11

从生产线上随机抽产品进行检测，设产品的次品率为 $p(0<p<1)$，若查到一只次品就得停机检修，设停机时已检测到 X 只产品，试写出 X 的概率分布律.

Solution

设 $A_i=\{第\ i\ 次抽到正品\}, i=1,2,\ldots$，则 $A_1, A_2, \ldots$ 相互独立. 于是 X 的概率分布律为

$$\mathrm{P}(X=k)=\mathrm{P}(A_1A_2\cdots A_{k-1}\overline{A}_k)=(1-p)^{k-1}p, \quad k=1,2,\ldots$$

此时称 X 服从参数为 p 的几何分布 (geometric distribution)，记作 $X\sim Geom(p)$.

Example 12

独立重复地进行试验, 每次试验的结果为成功或失败, 每次试验中成功的概率均为 $p(0 < p < 1)$, 试验进行到出现 r 次成功为止 $(r \geq 1)$, 以 X 表示试验次数, 则 X 的概率分布律为

$$\mathrm{P}(X = k) = \mathrm{C}_{k-1}^{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, r+2, \ldots,$$

其中 r 为正整数.

此时称 X 服从参数为 (r, p) 的巴斯卡分布 (Pascal distribution), 有时也称为负二项分布 (negative binomial distribution).

Example 13

一袋中有 a 个白球, b 个红球, $a+b=N$, 从中不放回地取 $n(\leq N)$ 个球, 设每次取到各球的概率相等, 以 X 表示取到的白球数, 则随机变量 X 的概率分布律为

$$\mathrm{P}(X=k)=\frac{\mathrm{C}_a^k\mathrm{C}_b^{n-k}}{\mathrm{C}_N^n},\quad k=l_1,l_1+1,\ldots,l_2,$$

其中 $l_1=\max(0,n-b)$, $l_2=\min(a,n)$.

此时称 X 服从超几何分布 (hypergeometric distribution).

思考题

一盒中有 2 个红球和 4 个白球, 试求如下随机变量的概率分布律.

1. 从中取一球, 取到的红球数 X 服从 0-1 分布 $B(1,1/3)$, 即

$$\mathsf{P}(X=0)=\frac{2}{3},\quad \mathsf{P}(X=1)=\frac{1}{3}.$$

2. 采用不放回抽样取 3 球, 取到的红球数 Y 服从超几何分布, 即

$$\mathsf{P}(Y=k)=\frac{\mathrm{C}_2^k\mathrm{C}_4^{3-k}}{\mathrm{C}_6^3},\quad k=0,1,2.$$

思考题 (Cont.)

3 采用放回抽样取 3 球, 取到的红球数 Z 服从二项分布 $B(3,1/3)$, 即

$$\mathrm{P}(Z=k)=\mathrm{C}_3^k\frac{2^{3-k}}{3^3},\quad k=0,1,2,3.$$

4 采用放回抽样取球, 直到取到红球为止, 取球次数 U 服从几何分布 $G(1/3)$, 即

$$\mathrm{P}(U=k)=\frac{2^{k-1}}{3^k},\quad k=1,2,\ldots.$$

5 采用放回抽样取球, 直到取到 3 个红球为止, 取球次数 V 服从巴斯卡分布, 即

$$\mathrm{P}(V=k)=\mathrm{C}_{k-1}^2\frac{2^{k-3}}{3^k},k=3,4,\ldots.$$

§2.3 Distribution Function

Definition 8 (分布函数)

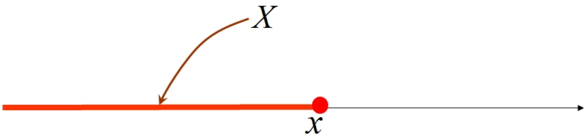
设 X 为一随机变量, 对任意实数 x, 称函数

$$F(x) = P(X \le x) = P\{X \in (-\infty, x]\}, \quad -\infty < x < \infty,$$

为 X 的概率分布函数, 简称分布函数. 有时也写为 $F_X(x) = P(X \le x)$.

任何随机变量都有相应的分布函数

分布函数的几何含义



Definition 8 ((Cumulative) Distribution Function)

The distribution function or (cumulative) distribution function (abbreviated d.f. or c.d.f.) F of a random variable X is the function

$$F(x) = \mathsf{P}(X \le x) = \mathsf{P}\{X \in (-\infty, x]\}, \quad \forall x \in \mathbb{R}.$$

It should be emphasized that the cumulative distribution function is defined as above for every random variable X, regardless of whether the distribution of X is discrete, continuous, or mixed.

分布函数的性质

由分布函数的定义，可知分布函数是事件 $\{X \le x\}$ 的概率，故显然有 $0 \le F(x) \le 1$.
此外它还具有以下三个基本性质：

(1) Nondecreasing (单调不减).
A d.f. is nondecreasing as x increases; that is, if $x_1 < x_2$, then $F(x_1) \le F(x_2)$.

因为对于任意的 $x_1 < x_2$，有 $F(x_2) - F(x_1) = \mathsf{P}(x_1 < X \le x_2) \ge 0$.
或利用：If $x_1 < x_2$, then the event $\{X \le x_1\}$ is a subset of the event $\{X \le x_2\}$.

(2) Limits at $\pm\infty$. $F(-\infty) := \lim_{x\to-\infty} F(x) = 0$ and $F(\infty) := \lim_{x\to\infty} F(x) = 1$.

(3) Continuity from the Right (右连续).
A d.f. is always continuous from the right; that is, $F(x+0) = F(x)$ at every point $x \in \mathbb{R}$.

分布函数的用途

分布函数表示出随机变量落入实数轴上任意一个范围的概率.

Let $F(x)$ be the distribution function of X. For any $a, b \in \mathbb{R}$ and $a < b$, then

$\mathrm{P}(a < X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = F(b) - F(a)$;

$\mathrm{P}(a < X < b) = \mathrm{P}(a < X \leq b) - \mathrm{P}(X = b)$

对于任意实数 x_0, 有

$$\mathrm{P}(X = x_0) = F(x_0) - F(x_0 - 0).$$

思考

What is the probability of the following events represented by F?

- $\mathrm{P}(a < X < b) = F(b-0) - F(a)$
- $\mathrm{P}(a \le X \le b) = F(b) - F(a-0)$
- $\mathrm{P}(a \le X < b) = F(b-0) - F(a-0)$

Example 14

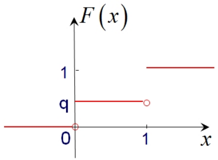
已知随机变量 X 的概率分布律为

X	0	1
P	q	p

其中 $p>0, q>0$ 且 $p+q=1$. 求 X 的概率分布函数 $F(x)$ 及 $P(X\geq 1)$ 的值.

Solution

由分布函数的定义可知 $F(x)=P(X\leq x)=\begin{cases}0, & x<0,\\ q, & 0\leq x<1,\\ 1, & x\geq 1.\end{cases}$



而 $P(X\geq 1)=p$.

比较 $P(X\geq 1)=p$ 与当 $x\geq 1$ 时, $F(x)=1$ 二者的含义区别.

一般地，离散型随机变量的分布函数为阶梯函数.

若离散型随机变量 X 的分布律为

$$\mathrm{P}(X=x_k)=p_k, \quad k=1,2,\ldots.$$

则 X 的分布函数为 $F(x)=\sum\limits_{x_k\le x} p_k$.

该分布函数 $F(x)$ 在 $x=x_k(k=1,2,\ldots)$ 处有跳跃，其跳跃度为 $p_k=\mathrm{P}(X=x_k)$.

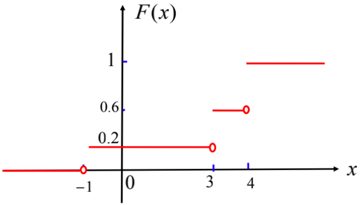
Example 15

设随机变量 X 的分布函数为 $F(x)=\begin{cases}0, & x<-1,\\ 0.2, & -1\le x<3,\\ 0.6, & 3\le x<4,\\ 1, & x\ge 4.\end{cases}$ 求 X 的概率分布律.

Solution

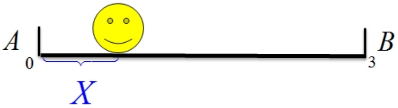
由 $F(x)$ 的定义和图像可知, $F(x)$ 只在 $-1,3,4$ 有跳跃, 跳的幅度分别是 $0.2,0.4,0.4$, 故概率分布律为

X	-1	3	4
p_k	0.2	0.4	0.4



Example 16

设一物体在 A, B 两点之间移动, A, B 之间距离 3 个单位. 该物体落在 A, B 中任一子区间的概率与区间长度成正比. 设它离 A 的距离为 X, 求 X 的分布函数.

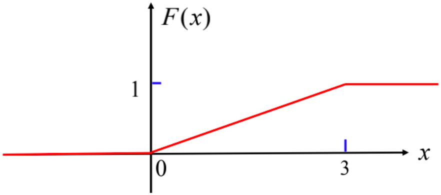


Solution

根据题意可得 $\mathrm{P}(0 \leq X \leq 3) = 1$, 故当 $x < 0$ 时, $F(x) = 0$; 当 $x \geq 3$ 时, $F(x) = 1$. 设比例系数为 k, 则 $\mathrm{P}(0 \leq X \leq 3) = 3k = 1 \Rightarrow k = 1/3$. 而当 $0 \leq x < 3$ 时,

$$F(x) = \mathrm{P}(X \leq x) = \mathrm{P}(X < 0) + \mathrm{P}(0 \leq X \leq x) = 0 + \frac{1}{3} \cdot x = \frac{x}{3}.$$

从而 X 的分布函数为 $F(x) = \begin{cases} 0, & x < 0, \\ x/3, & 0 \le x < 3, \\ 1, & x \ge 3. \end{cases}$ 其图像如下:



判断: X 是否为离散型的随机变量?

此分布函数是一个连续函数, 而非阶梯函数, 因此 X 不是离散型随机变量.

§2.4 连续型随机变量 (Continuous Random Variable/Distrubution)

Definition 9 (连续型随机变量、概率密度函数)

对于随机变量 X 的分布函数 $F(x)$, 若存在非负的函数 $f(x)$, 对于任意实数 x, 均有

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t, \quad \forall x \in \mathbb{R},$$

则称 X 为连续型随机变量 (continuous random variable), 其中 $f(x)$ 称为 X 的概率密度函数 (probability density function).

Definition 9 (Continuous Distribution/Random Variable/p.d.f./support)

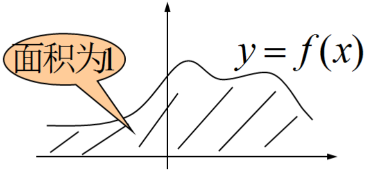
We say that a random variable X has a continuous distribution or that X is a continuous random variable if there exists a nonnegative function $f(x)$, defined on the real line, such that for every interval of real numbers (bounded or unbounded), the probability that X takes a value in the interval is the integral of $f(x)$ over the interval. $f(x)$ is called the probability density function (abbreviated p.d.f.) of X. The set $\{x : f(x) > 0\}$ is called the support of (the distribution of) X.

概率密度函数 $f(x)$ 的性质:

(1) $f(x) \geq 0$;

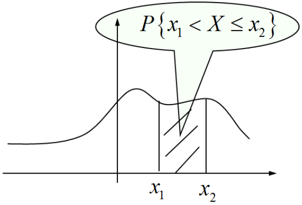
(2) $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

说明: (1) 与 (2) 为概率密度函数的两条本质性质.



(3) 对于连续型的随机变量 X, 概率密度函数为 $f(x)$, 则对于任意的实数 x_1, x_2, $x_1 < x_2$, 有

$$\mathsf{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f(x)\mathrm{d}x.$$



Proof

$$\mathsf{P}(x_1 < X \leq x_2) = F(x_2) - F(x_1) = \int_{-\infty}^{x_2} f(x)\mathrm{d}x - \int_{-\infty}^{x_1} f(x)\mathrm{d}x = \int_{x_1}^{x_2} f(x)\mathrm{d}x.$$

注: 对于连续型的随机变量 X, 概率密度函数为 $f(x)$,

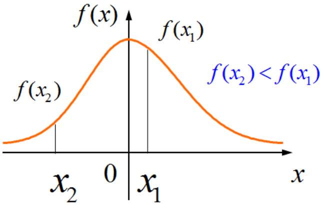
1. $\mathsf{P}(X = a) = 0, \quad \forall a \in \mathbb{R}$;
2. $\mathsf{P}(X \in I) = \int_I f(x)\mathrm{d}x, \quad I \subset \mathbb{R}$. (常用计算公式)

(4) 对于 $f(x)$ 的连续点 x, 有 $F'(x) = f(x)$.

即此时有

$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\mathsf{P}(x < X \leq x + \Delta x)}{\Delta x}.$$

说明:



1.$f(x)$ 值的含义

当 Δx 充分小时,

$$\mathrm{P}(x < X \leq x + \Delta x) \approx f(x) \cdot \Delta x;$$

2. $f(x)$ 的值有可能大于 1;
3. 连续型随机变量: 概率密度函数 $\rightleftharpoons$ 分布函数

$$f(x) \Rightarrow F(x) \quad \text{by} \quad \int_{-\infty}^{x} f(t)\,dt;$$

$$f(x) \Leftarrow F(x) \quad \text{by} \quad \frac{d}{dx}F(x).$$

Example 17

设 X 的概率密度函数为 $f(x) = \begin{cases} c, & 0 < x < 1, \\ 2/9, & 3 < x < 6, \\ 0, & \text{其他}. \end{cases}$

1. 求常数 c 的值;
2. 写出 X 的概率分布函数;
3. 要使 $\mathrm{P}(X < k) = 2/3$, 求 k 的值.

Solution

1. 由概率密度函数性质可知

$$1 = \int_{-\infty}^{+\infty} f(t)\mathrm{d}t = \int_0^1 c\mathrm{d}t + \int_3^6 \frac{2}{9}\mathrm{d}t = c + \frac{2}{3} \ \Rightarrow \ c = \frac{1}{3}.$$

由此可知

$$\mathsf{P}\{X \in (0,1) \cup (3,6)\} = 1.$$

支撑 (support): $D = \{x : f(x) > 0\}$
对于连续型随机变量 X, 有

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \int_D f(x)\mathrm{d}x + \int_{\overline{D}} f(x)\mathrm{d}x = \int_D f(x)\mathrm{d}x + 0 = \mathsf{P}(X \in D) = 1.$$

Solution (Cont.)

2 由密度函数定义

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t=\begin{cases}0, & x<0,\\ \int_0^x \frac{1}{3}\mathrm{d}t, & 0\le x<1,\\ \int_0^1 \frac{1}{3}\mathrm{d}t, & 1\le x<3,\\ \int_0^1 \frac{1}{3}\mathrm{d}t+\int_3^x \frac{2}{9}\mathrm{d}t, & 3\le x<6,\\ 1, & x\ge 6\end{cases}=\begin{cases}0, & x<0,\\ \frac{x}{3}, & 0\le x<1,\\ \frac{1}{3}, & 1\le x<3,\\ \frac{2x-3}{9}, & 3\le x<6,\\ 1, & x\ge 6.\end{cases}$$

思考：若已知 $F(x)$，如何求 $f(x)$?

3 使 $\mathsf{P}(X<k)=2/3=F(k)\Rightarrow k=4.5.$

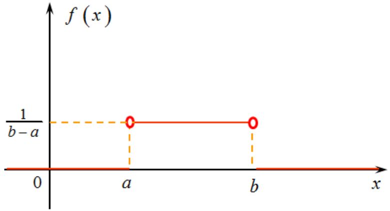
几个重要的连续型随机变量

(1) 均匀分布

Definition 10 (均匀分布)

设随机变量 X 具有概率密度函数

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in (a, b), \\ 0, & \text{其他}, \end{cases}$$



其中 $a < b$, 则称 X 服从区间 (a, b) 上均匀分布 (uniform distribution), 记作 $X \sim U(a, b)$.

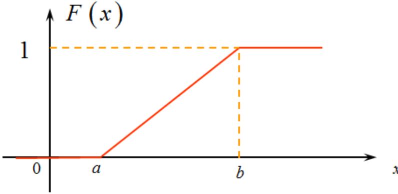
设 $a \le c < c+s \le b$, 则

$$\mathrm{P}(c < X < c+s) = \int_c^{c+s} \frac{1}{b-a} \mathrm{d}t = \frac{s}{b-a}$$

与 c 无关, 仅与区间长度 s 有关. (等可能性)

根据概率密度函数的定义, 可得均匀分布 $X \sim U(a,b)$ 的分布函数为

$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x < b, \\ 1, & x \ge b. \end{cases}$$



Example 18

在区间 $(-1,2)$ 上随机取一数 X, 试写出 X 的概率密度函数, 并求 $\mathsf{P}(X>0)$ 的值.

Solution

由题意知, X 服从区间 $(-1,2)$ 上的均匀分布, 其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{3}, & -1<x<2,\\ 0, & \text{其他}.\end{cases}$$

且

$$\mathsf{P}(X>0) \overset{\text{(法一)}}{=\!=\!=} \int_0^{+\infty} f(x)\mathrm{d}x=\int_0^2 \frac{1}{3}\mathrm{d}x=\frac{2}{3}$$

$$\overset{\text{(法二)}}{=\!=\!=} \frac{(-1,2)\cap(0,+\infty)\text{的长度}}{(-1,2)\text{的长度}}=\frac{(0,2)\text{的长度}}{(-1,2)\text{的长度}}=\frac{2}{3}.$$

Example 18 (Cont.)

若在该区间上随机取 10 个数, 求 10 个数中恰有 2 个数大于 0 的概率.

Solution (Cont.)

设 10 个数中有 Y 个数大于 0, 则 $Y \sim B\left(10, \frac{2}{3}\right)$. 故

$$\mathrm{P}(Y = 2) = \mathrm{C}_{10}^{2}\left(\frac{2}{3}\right)^{2}\left(\frac{1}{3}\right)^{8} = \frac{20}{3^{8}} \approx 0.003048.$$

Example 19

杭州某长途汽车站每天从早上 6 点 (第一班车) 开始, 每隔 30 分钟有一班车开往上海. 王先生在早上 6:20 至 7:10 随机到达车站, 设他早上 6:20 过 X 分钟到达车站, 则 X 服从 $(0,50)$ 上的均匀分布.

1. 求王先生候车时间不超过 15 分钟的概率;
2. 如果王先生一月中有两次按此方式独立地去候车, 求他一次候车不超过 15 分钟, 另一次候车大于 10 分钟的概率.

Solution

1. $\mathrm{P}(\text{候车时间不超过 } 15 \text{ 分钟}) = \frac{10+15}{50} = 0.5.$

Solution (Cont.)

2 $\mathsf{P}(\text{候车时间大于 } 10 \text{ 分钟}) = \frac{20+10}{50} = 0.6.$
设王先生两次分别在早上 6:20 过 X_1 和 X_2 分钟到达车站, 那么

$$\begin{aligned}
& \mathsf{P}(\text{一次候车时间不超过 } 15 \text{ 分钟, 另一次大于 } 10 \text{ 分钟}) \\
= \; & \mathsf{P}((X_1 < 15, X_2 > 10) \cup (X_1 > 10, X_2 < 15)) \\
= \; & \mathsf{P}(X_1 < 15, X_2 > 10) + \mathsf{P}(X_1 > 10, X_2 < 15) \\
& \quad - \mathsf{P}(10 < X_1 < 15, 10 < X_2 < 15) \\
= \; & 0.5 \times 0.6 + 0.6 \times 0.5 - 0.1 \times 0.1 \\
= \; & 0.59.
\end{aligned}$$

(2) 指数分布

Definition 11 (指数分布)

设随机变量 X 具有概率密度函数

$$f(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x > 0, \\ 0, & x \le 0, \end{cases}$$

其中 $\lambda > 0$, 则称 X 服从参数为 λ 的指数分布 (exponential distribution), 记作 $X \sim Exp(\lambda)$ 或 $X \sim E(\lambda)$.

根据概率密度函数的定义，可得指数分布 $X \sim E(\lambda)$ 的分布函数为

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$$

指数分布还具有一个重要的性质——“无记忆性”：
即对任意的 $t > 0, t_0 > 0$，有

$$\mathsf{P}(X > t_0 + t | X > t_0) = \frac{\mathsf{P}(X > t_0 + t)}{\mathsf{P}(X > t_0)} = \frac{1 - F(t_0 + t)}{1 - F(t_0)} = \mathrm{e}^{-\lambda t} = \mathsf{P}(X > t).$$

这也意味着，对于任意的 $t > 0, t_0 > 0$，在 $\{X > t_0\}$ 的条件下，$X - t_0$ 服从指数分布（即其条件分布为指数分布）.

Example 20

某大型设备在任何长度为 t 的区间内发生故障的次数 $N(t)$ 服从参数为 λt 的 Poisson 分布. 记设备无故障运行的时间为 T.

1. 求 T 的概率分布函数;
2. 已知设备已无故障运行 10 个小时, 求再无故障运行 8 个小时的概率.

Solution

1. 由题意 $N(t) \sim P(\lambda t)$, 其分布律为 $\mathrm{P}(N(t)=k)=\frac{(\lambda t)^k}{k!}\mathrm{e}^{-\lambda t}, \quad k=0,1,2,\ldots$.
无故障运行时间 T 的分布函数为 $F_T(t)=\mathrm{P}(T \leq t)$.
(1) 当 $t<0$ 时, $F_T(t)=0$;
(2) 当 $t \geq 0$ 时, $F_T(t)=1-\mathrm{P}(T>t)=1-\mathrm{P}(N(t)=0)=1-\mathrm{e}^{-\lambda t}$.

Solution (Cont.)

因此 T 的分布函数为

$$F_T(t) = \begin{cases} 1 - \mathrm{e}^{-\lambda t}, & t \geq 0, \\ 0, & t < 0. \end{cases}$$

2 注意到

$$\mathsf{P}(T > 18 | T > 10) = \frac{\mathsf{P}(T > 18)}{\mathsf{P}(T > 10)} = e^{-8\lambda} = \mathsf{P}(T > 8),$$

即 $T \sim E(\lambda)$ 具有无记忆性.

(3) 正态分布

Definition 12 (正态分布)

设随机变量 X 具有概率密度函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \quad -\infty < x < +\infty,$$

其中 $\sigma > 0, -\infty < \mu < +\infty$, 则称 X 服从参数为 μ, σ 的正态分布 (normal distribution), 或简称 X 为正态变量, 也常称为 Gauss 分布, 记作 $X \sim N(\mu, \sigma^2)$.

下面验证密度函数的归一化，即 $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$.

Proof

注意到

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x \xlongequal{\text{令}t=\frac{x-\mu}{\sigma}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{t^2}{2}}\sigma\mathrm{d}t = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t,$$

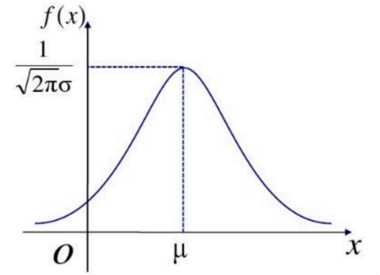
记 $I = \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$, 则 $I^2 = \iint_{\mathbb{R}^2} \mathrm{e}^{-\frac{x^2+y^2}{2}}\mathrm{d}x\mathrm{d}y = \int_0^{2\pi}\mathrm{d}\theta \int_0^{+\infty} r\mathrm{e}^{-\frac{r^2}{2}}\mathrm{d}r = 2\pi$, 于是

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi}}I = 1.$$

正态变量 $X \sim N(\mu, \sigma^2)$ 的概率密度函数 $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$

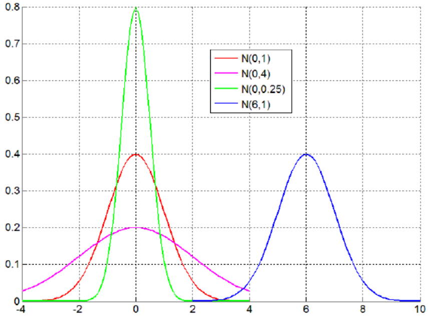
具有如下性质:

- $f(x)$ 关于 $x = \mu$ 对称;
- 当 $x > \mu$ 时, $f(x)$ 关于 x 严格单调下降;
- $f_{\max} = f(\mu) = \dfrac{1}{\sqrt{2\pi}\sigma}$;
- $\lim\limits_{|x-\mu| \to +\infty} f(x) = 0$;
- 在 $x = \mu \pm \sigma$ 处曲线有拐点.



不同的 μ 和 σ^2 对概率密度函数图像的影响 (两参数的含义)

- 当固定 σ, 改变 μ 的大小时, $f(x)$ 的图像形状不变, 只是沿着 x 轴作平移变换;
- 当固定 μ, 改变 σ 的大小时, $f(x)$ 的图像对称轴不变, 而形状在改变, σ 越小, 图形越高越瘦, σ 越大, 图形越矮越胖.



通常称 μ 为位置参数 (决定曲线对称轴位置), σ 为尺度参数 (决定曲线分散程度).

正态变量的特点

- 取值呈中间多，两头少，对称的特性;
- 当固定 μ 时，σ 越大，曲线的峰越低，落在 μ 附近的概率越小，取值就越分散，即 σ 是反映正态变量的取值分散性的一个指标;
- 在自然现象和社会现象中，大量随机变量服从或近似服从正态分布.

正态分布下的概率计算

若 $X \sim N(\mu, \sigma^2)$，则

$$\mathrm{P}\{X \leq x\} = F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}} \mathrm{d}t = ?$$

算不出来!

方法一: 利用 Excel、MATLAB、R、Python、C++、SPSS 等软件来计算;

方法二: 转化为标准正态分布, 结合标准正态分布函数表进行计算.

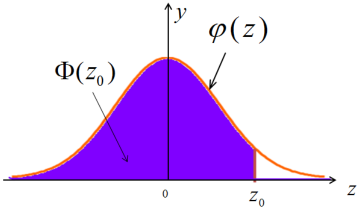
标准正态分布

Definition 13 (标准正态分布)

若 $Z \sim N(0,1)$, 则称 Z 服从标准正态分布.

概率密度函数为 $\varphi(z) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$;

分布函数为 $\Phi(z) = \displaystyle\int_{-\infty}^{z} \varphi(t)\,\mathrm{d}t = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t.$



附表 2: 标准正态分布表

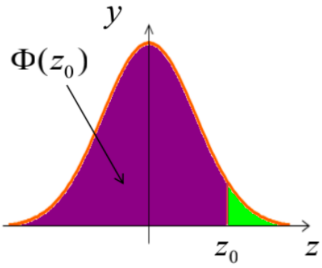
标准正态分布函数表 https://en.wikipedia.org/wiki/Standard_normal_table

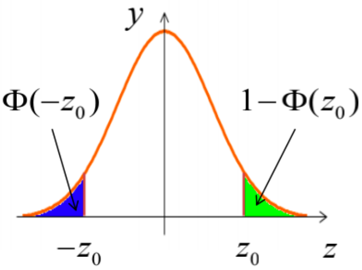
z	+ 0.00	+ 0.01	+ 0.02	+ 0.03	+ 0.04	+ 0.05	+ 0.06	+ 0.07	+ 0.08	+ 0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900

例如: $\Phi(1.23) = 0.89065$, $\Phi(1.96) = 0.97500$.

注意到 $\varphi(z)=\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{z^2}{2}}$ 是偶函数，因此标准正态分布的分布函数具有下面这个重要性质：

$$\Phi(-z_0)=1-\Phi(z_0),\quad \forall z_0\in\mathbb{R}.$$





性质

当 $X \sim N(\mu, \sigma^2)$，则有 $\dfrac{X-\mu}{\sigma} \sim N(0,1)$.

Proof

$$\begin{aligned} F_{\frac{X-\mu}{\sigma}}(z) &= \mathsf{P}\left(\frac{X-\mu}{\sigma} \leq z\right) = \mathsf{P}(X \leq \mu + z\sigma) = F_X(\mu + z\sigma) \\ &= \int_{-\infty}^{\mu+z\sigma} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}} \mathrm{d}t \xlongequal{令 s=\frac{t-\mu}{\sigma}} \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \mathrm{d}s = \Phi(z), \quad \forall z \in \mathbb{R}. \end{aligned}$$

由此可知，当 $X \sim N(\mu, \sigma^2)$ 时，对于任意实数 a，有

$$F_X(a) = \mathsf{P}(X \leq a) = \mathsf{P}\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right).$$

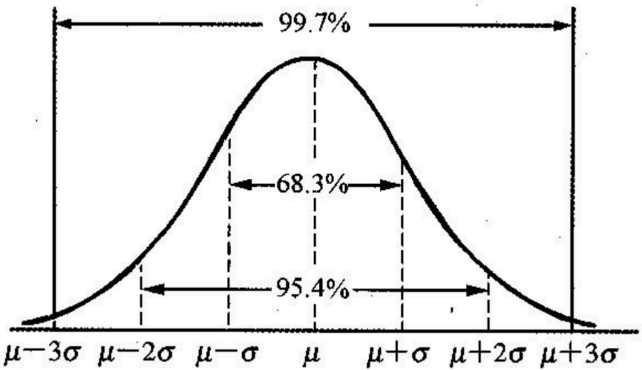
Example 21

设 $X \sim N(\mu, \sigma^2)$，则

$$\begin{aligned} \mathsf{P}(|X-\mu|<\sigma) &= \mathsf{P}(-\sigma < X-\mu < \sigma) = \mathsf{P}(\mu-\sigma < X < \mu+\sigma) \\ &= \mathsf{P}\Big(\frac{(\mu-\sigma)-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{(\mu+\sigma)-\mu}{\sigma}\Big) \\ &= \Phi(1)-\Phi(-1) = 2\Phi(1)-1 = 0.6826, \end{aligned}$$

$$\mathsf{P}(|X-\mu|<2\sigma) = 2\Phi(2)-1 = 0.9544,$$

$$\mathsf{P}(|X-\mu|<3\sigma) = 2\Phi(3)-1 = 0.9974.$$


99.7%
68.3%
95.4%
μ−3σ
μ−2σ
μ−σ
μ
μ+σ
μ+2σ
μ+3σ

Example 22

用天平称一实际重量为 a 的物体, 天平的读数为随机变量 X, 设 $X \sim N(a, 0.01^2)$.

1. 求读数 X 与 a 的误差小于 0.005 的概率;
2. 求读数 X 至少比 a 多 0.0085 的概率.

Solution

由正态分布性质知 $\dfrac{X-a}{0.01} \sim N(0,1)$. 因此

1. $\mathrm{P}(|X-a| < 0.005) = \Phi\left(\dfrac{0.005}{0.01}\right) - \Phi\left(-\dfrac{0.005}{0.01}\right) = 2\Phi(0.5) - 1$

$$\xlongequal{\text{查附表 2}} 2 \times 0.6915 - 1 = 0.3830.$$

2. $\mathrm{P}(X - a \geq 0.0085) = 1 - \Phi(0.85) = 1 - 0.8023 = 0.1977.$

使用 Excel 表单计算 P$(X - a < 0.0085)$ (以 2003 版本为例，不同版本具体做法有所不同)

1. 在 Excel 表单的任一单元格输入“=”
2. 在主菜单中点击“插入”，选择“函数 (F)”
3. 在选择类别的下拉式菜单中选择“统计”
4. 选择“NORMDIST”点击“确定”(注意区别 NORMSDIST)
5. 在函数参数表单中输入

   “X=0.0085, Mean=0, Standard_dev=0.01, Cumulative=TRUE”

   点击“确定”
6. 在单元格中出现“0.802337508”

Example 23

一批钢材 (线材) 长度 X(单位: cm) 服从正态分布 $N(\mu, \sigma^2)$.

1. 若 $\mu = 100$, $\sigma = 2$, 求这批钢材长度小于 97.8 cm 的概率.
2. 若 $\mu = 100$, 要使这批钢材的长度至少有 90% 落在区间 (97, 103) 内, 问 σ 至多取何值?

Solution

1. $$\mathrm{P}(X < 97.8) = \mathrm{P}\left(\frac{X-100}{2} < \frac{97.8-100}{2}\right) = \Phi(-1.1) = 1 - \Phi(1.1)$$
$$\xlongequal{\text{查附表 2}} 1 - 0.8643 = 0.1357.$$
2. $$\mathrm{P}(97 < X < 103) = \Phi\left(\frac{103-100}{\sigma}\right) - \Phi\left(\frac{97-100}{\sigma}\right) = 2\Phi\left(\frac{3}{\sigma}\right) - 1 \geq 90\%.$$

Solution (Cont.)

所以需 $\Phi\left(\frac{3}{\sigma}\right) \geq 0.95$, 即

$$\frac{3}{\sigma} \geq 1.645 \ \text{(利用线性插值法)} \Rightarrow \sigma \leq 1.8237.$$

在计算 Φ 的逆函数时, 也可以使用 Excel 表单.

1. 选择 "NORMINV" 点击 "确定"
2. 在函数参数表单中输入

"Probability=0.95, Mean=0, Standard_dev=1"

点击 "确定", 在单元格中出现 "1.644853627"

Example 24

设一天中经过一高速公路某一入口的重型车辆数 X 近似服从 $N(\mu,\sigma^2)$. 已知有 25% 的天数超过 400 辆, 有 33% 的天数不到 350 辆, 求 μ,σ.

Solution

由题意知

$$\mathrm{P}(X>400)\approx 1-\Phi\left(\frac{400-\mu}{\sigma}\right)=\Phi\left(\frac{\mu-400}{\sigma}\right)=0.25,$$

$$\mathrm{P}(X<350)\approx \Phi\left(\frac{350-\mu}{\sigma}\right)=0.33$$

而 $\Phi(-0.675)=0.25$, $\Phi(-0.440)=0.33$, 因此

$$\begin{cases}(\mu-400)/\sigma=-0.675,\\(350-\mu)/\sigma=-0.440\end{cases}\Rightarrow\begin{cases}\mu=369.7,\\\sigma=44.8.\end{cases}$$

Example 25

一银行服务需要等待，设等待时间 X (分钟) 的概率密度函数为

$$f(x) = \begin{cases} \dfrac{1}{10}\mathrm{e}^{-\frac{x}{10}}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

某人进了银行，且打算过会儿去办另一件事情，于是先等待，如果超过 15 分钟还没有等到服务就离开. 设他实际的等待时间为 Y.

1. 求 Y 的分布函数;
2. 问 Y 是离散型随机变量吗? 是连续型随机变量吗?

Solution

1. 由题意可知 $Y = \min\{X, 15\}$，故

$$F(y) = \mathrm{P}(Y \le y) = \begin{cases} 0, & y < 0, \\ 1 - \mathrm{e}^{-\frac{y}{10}}, & 0 \le y < 15, \\ 1, & y \ge 15. \end{cases}$$

2. Y 的取值范围是 [0, 15]，故 Y 不是离散型随机变量；
又 $\mathrm{P}(Y = 15) = \mathrm{e}^{-1.5} \ne 0$，因此 Y 也不是连续型随机变量.
(事实上，从 Y 的分布函数不是连续函数中也可知 Y 不是连续型随机变量.) 故 Y 为非连非离型的随机变量

§2.5 随机变量函数的分布 (Distributions of Functions of a Random Variable)

例如，若要测量一个圆的面积，总是测量其半径，而半径的测量值可看作随机变量 X. 假设 $X \sim N(\mu, \sigma^2)$，则面积 $S = \pi X^2$ 服从什么分布？

问题：已知随机变量 X 的概率分布，且已知 $Y = g(X)$，求 Y 的概率分布.

Now we are concerned about the distribution of functions of random variables. For the case that we have known the distribution of random variable X, $g(\cdot)$ is a measurable function, how could we get the information(distribution function) of $Y = g(X)$?

Example 26

已知 X 具有以下的概率分布

X	-1	0	1
p_i	0.2	0.5	0.3

且设 $Y = X^2$, 求 Y 的概率分布.

Solution

由题意知, Y 的所有可能取值为 0, 1, 且 $\mathrm{P}(Y=0)=\mathrm{P}(X=0)=0.5$,

$$\mathrm{P}(Y=1)=\mathrm{P}\{(X=1)\cup(X=-1)\}=\mathrm{P}(X=1)+\mathrm{P}(X=-1)=0.5.$$

即找出 $\{Y=0\}$ 的等价事件 $\{X=0\}$;
$\{Y=1\}$ 的等价事件 $\{X=1\}$ 与 $\{X=-1\}$ 的和事件.

Example 27

设随机变量 X 的概率密度函数为 $f_X(x) = \begin{cases} \dfrac{x}{8}, & 0 < x < 4, \\ 0, & \text{其他}. \end{cases}$ 求 $Y = X^2$ 的概率密度函数.

Solution

由题意知 $\mathsf{P}(0 < X < 4) = 1$, 而 $Y = X^2$, 故 $\mathsf{P}(0 < Y < 16) = 1$.
因此, 当 $y < 0$ 或 $y \geq 16$ 时, $f_Y(y) = 0$.
(当 $y < 0$ 时, 分布函数 $F_Y(y) = 0$; $y \geq 16$ 时, $F_Y(y) = 1$.)

Solution (Cont.)

当 $0 \le y < 16$ 时，Y 的分布函数为

$$F_Y(y) = \mathsf{P}(Y \le y) = \mathsf{P}(X^2 \le y) = \mathsf{P}(-\sqrt{y} \le X \le \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

故此时

$$f_Y(y) = F'_Y(y) = f_X(\sqrt{y}) \cdot \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \cdot (-\frac{1}{2\sqrt{y}}) = \frac{\sqrt{y}}{8} \cdot \frac{1}{2\sqrt{y}} - 0 \cdot (-\frac{1}{2\sqrt{y}}) = \frac{1}{16},$$

即 $f_Y(y) = \begin{cases} \frac{1}{16}, & 0 < y < 16, \\ 0, & \text{其他}. \end{cases}$ Y 服从区间 $(0, 16)$ 上的均匀分布.

一般地，若已知 X 的概率分布，求 $Y=g(X)$ 的概率分布的过程为：

1. 给出 Y 的可能取值；
2. 利用等价事件来给出概率分布.

具体来说，

- 若 Y 为离散型随机变量，则先写出 Y 的可能取值 $y_1, y_2, \ldots, y_j, \ldots$，再找出 $\{Y=y_j\}$ 的等价事件 $\{X \in I\}$，得分布律 $\mathsf{P}(Y=y_j)=\mathsf{P}(X \in I)$；
- 若 Y 为连续型随机变量，先确定 Y 的取值范围，再写出概率分布函数 $F_Y(y)=\mathsf{P}(Y \leq y)$，找出 $\{Y \leq y\}$ 的等价事件 $\{X \in I\}$，得 $F_y(y)=\mathsf{P}(X \in I)$，再求导得 Y 的概率密度函数 $f_Y(y)$.

关键是找出等价事件!

若 Y 为连续型随机变量，求其概率密度函数，通常作法为:

1. 根据 X 的取值范围以及 Y 与 X 的关系，确定 Y 的取值范围;
2. 写出 Y 的分布函数:
$$F_Y(y) = \mathsf{P}\{Y \leq y\},$$
通过 $\{Y \leq y\}$ 的等价事件$\{X \in I\}$，可将之转化为 $F_Y(y) = \mathsf{P}\{X \in I\}$;
3. 对 $F_Y(y)$ 求导得到 Y 的概率密度函数 $f_y(y)$. 常常会用到复合函数求导:

$$\frac{d(F_X(h(y)))}{dy} = f_X(h(y)) \cdot h'(y).$$

Example 28

已知 X 具有概率分布

X	-1	0	1
P	$1/3$	$1/3$	$1/3$

且设 $Y = 2X$, $Z = X^2$, 求 Y 和 Z 的概率分布.

Solution

由题意知, Y 的所有可能取值为 $-2, 0, 2$, Z 的所有可能取值为 $0, 1$, 且

$$\{Y = -2\} = \{X = -1\}, \quad \{Y = 0\} = \{X = 0\}, \quad \{Y = 2\} = \{X = 1\}$$

$$\{Z = 0\} = \{X = 0\}, \quad \{Z = 1\} = \{X = -1\} \cup \{X = 1\}$$

Solution (Cont.)

由此可以得到 Y 和 Z 的概率分布律分别为

Y	-2	0	2
P	$1/3$	$1/3$	$1/3$

Z	0	1
P	$1/3$	$2/3$

Example 29

设随机变量 X 具有概率密度 $f_X(x), -\infty < x < +\infty$, 分别求 $Y = |X|$ 和 $Z = X^2$ 的概率密度函数 $f_Y(y)$ 和 $f_Z(z)$.

Solution

记 X, Y, Z 的分布函数分别为 $F_X(x), F_Y(y), F_Z(z)$.

注意到 $\mathrm{P}(Y \geq 0) = 1$, 故当 $y < 0$ 时, $F_Y(y) = 0$; 当 $y \geq 0$ 时,

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(|X| \leq y) = \mathrm{P}(-y \leq X \leq y) = F_X(y) - F_X(-y).$$

因此

$$f_Y(y) = F'_Y(y) = \begin{cases} f_X(y) + f_X(-y), & y \geq 0, \\ 0, & y < 0. \end{cases} \tag{1}$$

Solution (Cont.)

同理, 注意到 $\mathsf{P}(Z \geq 0) = 1$, 故当 $z < 0$ 时, $F_Z(z) = 0$; 当 $z \geq 0$ 时,

$$F_Z(z) = \mathsf{P}(Z \leq z) = \mathsf{P}(X^2 \leq z) = F_X(\sqrt{z}) - F_X(-\sqrt{z}).$$

因此

$$f_Z(z) = F'_Z(z) = \begin{cases} \dfrac{1}{2\sqrt{z}}\left[f_X(\sqrt{z}) + f_X(-\sqrt{z})\right], & z \geq 0, \\ 0, & z < 0. \end{cases} \tag{2}$$

Example 30

设随机变量 $X \sim U(-1,2)$，求 $Y = |X|$ 的概率密度函数 $f_Y(y)$.

Solution

由题意可知 X 的概率密度函数为 $f_X(x) = \begin{cases} 1/3, & -1 \le x \le 2, \\ 0, & \text{其他}. \end{cases}$ 由 (1) 得 $Y = |X|$ 的概率密度函数为

$$f_Y(y) = \begin{cases} f_X(y) + f_X(-y), & y \ge 0, \\ 0, & y < 0 \end{cases}$$

$$= \begin{cases} 1/3 + 1/3, & 0 \le y < 1, \\ 1/3 + 0, & 1 \le y \le 2, \\ 0, & \text{其他} \end{cases} = \begin{cases} 2/3, & 0 \le y < 1, \\ 1/3, & 1 \le y \le 2, \\ 0, & \text{其他}. \end{cases}$$

Example 31

设随机变量 $X \sim N(0,1)$，求 $Z = X^2$ 的概率密度函数 $f_Z(z)$.

Solution

由题意知 X 的密度函数为 $f_X(x) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < +\infty$. 由 (2) 得 $Z = X^2$ 的密度函数为

$$f_Z(z) = \begin{cases} \dfrac{1}{2\sqrt{z}}\left[f_X(\sqrt{z}) + f_X(-\sqrt{z})\right], & z > 0, \\ 0, & z \le 0 \end{cases} = \begin{cases} \dfrac{1}{\sqrt{2\pi}} z^{-\frac{1}{2}} e^{-\frac{z}{2}}, & z > 0, \\ 0, & z \le 0. \end{cases}$$

在数理统计第六章中将介绍，这实际上是 $\chi^2(1)$ 分布.

Theorem 1

设随机变量 X 的概率密度函数为 $f_X(x)$，已知 $Y = g(X)$，其中 $g'(x) > 0$ (或 $g'(x) < 0$)，则 Y 的概率密度函数为

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot \left|h'(y)\right|, & \alpha < y < \beta, \\ 0, & \text{其它}, \end{cases}$$

其中 $\alpha = g(-\infty)$, $\beta = g(\infty)$ (当 $g'(x) < 0$ 时, $\alpha = g(\infty)$, $\beta = g(-\infty)$).
$h(y) = x \Leftrightarrow y = g(x)$.

Proof

不妨设 $g'(x) > 0$, 即 $g(x)$ 为单调递增函数, 且 $h'(y) > 0$.
当 $y < \alpha$ 时, $F_Y(y) = 0$;
当 $y \geq \beta$ 时, $F_Y(y) = 1$;
当 $\alpha \leq y < \beta$ 时,

$$F_Y(y) = \mathsf{P}(Y \leq y) = \mathsf{P}(g(X) \leq y) = \mathsf{P}(X \leq h(y)) = F_X(h(y)).$$

从而

$$f_Y(y) = f_X(h(y)) \cdot h'(y) = f_X(h(y)) \cdot |h'(y)|.$$

同理可证, 当 $g'(x) < 0$, 结论也成立.

推论

设随机变量 X 的概率密度函数为 $f_X(x)$, $Y = g(X)$, X 的支撑为 $\{x : f(x) > 0\} = (a, b)$. 若当 $a < x < b$ 时, 有 $g'(x) > 0$ (或 $g'(x) < 0$), 则 Y 的概率密度函数为

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot \left|h'(y)\right|, & \alpha < y < \beta, \\ 0, & \text{其它}, \end{cases}$$

其中 $\alpha = \min(g(a), g(b))$, $\beta = \max(g(a), g(b))$, $h(y) = x \Leftrightarrow y = g(x)$.

Example 32

设 X 服从正态分布 $N(\mu,\sigma^2)$，考虑随机变量 $Y=aX+b(a\neq 0)$ 的分布.

Solution

注意到 $Y=g(X)$，其中 $y=g(x)=ax+b(a\neq 0)$ 为 x 的严格单调函数 ($a>0$ 时为单调递增函数，$a<0$ 时为单调递减函数)，其反函数为 $g^{-1}(y)=\dfrac{y-b}{a}$，则根据定理，可得

$$f_Y(y)=f_X(g^{-1}(y))\cdot\left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|=\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)=\frac{1}{\sqrt{2\pi}|a|\sigma}\mathrm{e}^{-\frac{[y-(a\mu+b)]^2}{2a^2\sigma^2}}.$$

即 Y 也服从正态分布，具体为 $N(a\mu+b,a^2\sigma^2)$.

重要性质：若 $X\sim N(\mu,\sigma^2)$，$Y=aX+b(a\neq 0)$，则 $Y\sim N(a\mu+b,a^2\sigma^2)$.

Example 33

设 X 服从参数为 λ 的指数分布，$F(x)$ 为 X 的分布函数. 设 $Y = F(X)$，试证: $Y \sim U(0,1)$ (即均匀分布).

Solution

由指数分布的定义知，$F(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases}$ 现在

$$Y = F(X) = \begin{cases} 1 - \mathrm{e}^{-\lambda X}, & X > 0, \\ 0, & X \leq 0. \end{cases}$$

因此 $\mathrm{P}(0 \leq Y \leq 1) = 1$.

Solution (Cont.)

当 $y<0$ 时, $F_Y(y)=\mathsf{P}(Y\le y)=0$; 当 $y\ge 1$ 时, $F_Y(y)=\mathsf{P}(Y\le y)=1$;
当 $0\le y<1$ 时,

$$F_Y(y)=\mathsf{P}(1-\mathrm{e}^{-\lambda X}\le y)=\mathsf{P}\Big(X\le -\frac{1}{\lambda}\ln(1-y)\Big)=1-\mathrm{e}^{-\lambda[-\frac{1}{\lambda}\ln(1-y)]}=y.$$

所以 $F_Y(y)=\begin{cases}0, & y<0,\\ y, & 0\le y<1,\\ 1, & y\ge 1.\end{cases}$ 即 $Y\sim U(0,1)$.

注: 也可利用定理 (推论) 来做, 此时 $g^{-1}(y)=-\dfrac{1}{\lambda}\ln(1-y)$, $\dfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)=\dfrac{1}{\lambda}\cdot\dfrac{1}{1-y}$.

更一般地, 我们有如下结论 (见书本例 2.5.6).

Theorem 2 (Probability integral transformation)

Let X have continuous c.d.f. $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0, 1)$.

Example 34

设随机变量 X 在 $(-\pi/2, \pi/2)$ 内服从均匀分布，试求 $Y = \sin X$ 的概率密度函数.

Solution

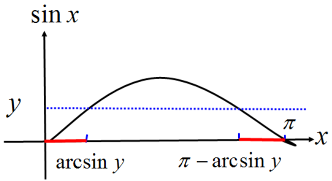
变换 $Y = \sin X$ 对应的函数为 $y = g(x) = \sin x$，此函数在 $(-\pi/2, \pi/2)$ 上恒有 $g'(x) = \cos x > 0$，且有反函数 $h(y) = \arcsin y$，其导数 $h'(y) = \dfrac{1}{\sqrt{1-y^2}}$. 因此，由定理 (其实是推论) 得 $Y = \sin X$ 的概率密度函数为

$$f_Y(y) = f_X(h(y)) \cdot \left|h'(y)\right| = \begin{cases} \dfrac{1}{\pi} \cdot \dfrac{1}{\sqrt{1-y^2}}, & -1 < y < 1, \\ 0, & \text{其他}. \end{cases}$$

Example 35

设随机变量 X 在 $(0,\pi)$ 内服从均匀分布，试求随机变量 $Y=\sin X$ 的概率密度函数.

Solution



变换 $Y=\sin X$ 对应的函数 $y=g(x)=\sin x$ 在 $(0,\pi)$ 上不单调，不能利用定理和推论.

对 $0\le y<1$,

$$F_Y(y)=\mathrm{P}(Y\le y)=\mathrm{P}(\sin X\le y)=\mathrm{P}(X\in[0,\arcsin y]\cup[\pi-\arcsin y,\pi])=\frac{2\arcsin y}{\pi}.$$

故 Y 的概率密度函数为

$$f_Y(y)=\begin{cases}\dfrac{2}{\pi\sqrt{1-y^2}}, & 0\le y<1,\\ 0 & 其他\end{cases}$$

谢谢!