引理 4.8 若随机变量 $X_1, X_2, \cdots, X_n$ 是相互独立的、且分别服从参数为 $\lambda_1, \lambda_2, \cdots, \lambda_n$ 的指数分布, 则有

$$X = \min\{X_1, X_2, \cdots, X_n\} \sim e(\lambda_1 + \lambda_2 + \cdots + \lambda_n) \ .$$

解 这里随机变量的相互独立性可以理解为随机变量取不同值的随机事件相互独立. 计算随机变量 X 的分布函数

$$\begin{aligned} F_X(x) &= P(X \leqslant x) = 1 - P(\min(X_1, X_2, \cdots, X_n) > x) \\ &= 1 - \prod_{i=1}^{n} P(X_i > x) = 1 - \prod_{i=1}^{n} \exp(-\lambda_i x) = 1 - \exp\left(-x \sum_{i=1}^{n} \lambda_i\right) \ , \end{aligned}$$

由此完成证明.

4.4.3 正态分布

正太分布是概率统计中最重要的一种分布, 最早由法国数学家棣莫弗 (De Moivre, 1667-1754) 在 1730s 提出, 用于近似抛硬币试验中随机事件的概率, 即中心极限定理的雏形. 德国数学家高斯 (Gauss, 1777-1855) 在 1800s 首次将正太分布应用于预测天文学中星体的位置, 由此才展示出正太分布的应用价值, 后来发现很多随机现象可以通过正太分布来描述, 正太分布因此被称为高斯分布.

正太分布在概率统计中的重要性主要体现在以下几方面:

- 现实生活中很多随机现象需要用正太分布进行描述, 如人的身高或体重, 某地区的降雨量等;
- 很多分布可以通过正太分布来进行近似计算, 如后面所学的中心极限定理;
- 数理统计中常用的统计分布是由正太分布导出的, 如后面所学的 χ^2 分布、t 分布和 F 分布.

定义 4.7 给定任何实数 u 和 $\sigma > 0$, 若随机变量 X 的概率密度为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad x \in (-\infty, +\infty) \ ,$$

称随机变量 X 服从 **参数为** (μ, σ^2) **的正态分布** (normal distribution), 又称为 **高斯分布** (Gaussian distribution), 记为 $X \sim \mathcal{N}(\mu, \sigma^2)$.

特别地, 当 $\mu = 0$ 和 $\sigma = 1$ 时的正太分布 $\mathcal{N}(0, 1)$ 被称为 **标准正态分布**, 此时密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad x \in (-\infty, +\infty) \ .$$

对任意 $x \in (-\infty, +\infty)$ 有 $f(x) \geqslant 0$, 利用极坐标变换 ($x = r\cos\theta, y = r\sin\theta$) 有

$$\left(\int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx\right)^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-\frac{x^2+y^2}{2}} dxdy$$

$$= \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} r dr = \int_0^{2\pi} d\theta \int_0^{+\infty} e^{-\frac{r^2}{2}} d\frac{r^2}{2} = 2\pi \ ,$$

由此验证了 $\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = 1$, 利用简单的变量替换可验证一般正太分布的密度函数.

关于标准正太分布和一般的正太分布, 有如下关系:

定理 4.5 若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则有 $Y = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$; 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $Y = \sigma X + \mu \sim \mathcal{N}(\mu, \sigma^2)$.

证明 若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则 $Y = (X-\mu)/\sigma$ 的分布函数

$$F_Y(y) = P[Y \leqslant y] = P[X - \mu \leqslant y\sigma] = P[X \leqslant y\sigma + \mu] = \int_{-\infty}^{\mu + y\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \ .$$

令 $x = (t-\mu)/\sigma$, 代入上面的分布函数有

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \ ,$$

由此可证 $Y \sim \mathcal{N}(0,1)$.

另一方面, 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $Y = \sigma X + \mu$ 的分布函数

$$F_Y(y) = P(Y \leqslant y) = P(\sigma X + \mu \leqslant y) = P(X \leqslant (y-\mu)/\sigma) = \int_{-\infty}^{(y-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \ .$$

令 $t = (x-\mu)/\sigma$, 代入上面的分布函数有

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \ ,$$

由此可证 $Y \sim \mathcal{N}(\mu, \sigma^2)$.

关于正太分布的数字特征有

定理 4.6 若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则有 $E(X) = \mu$ 和 $\mathrm{Var}(X) = \sigma^2$; 特别地, 若随机变量 $X \sim \mathcal{N}(0,1)$, 则有 $E(X) = 0$ 和 $\mathrm{Var}(X) = 1$.

正太分布的两个参数分别表示正太分布的期望和方差.

证明 这里仅仅证明标准正太分布的期望为 0 和方差为 1, 结合定理 4.5 直接可得 $X \sim \mathcal{N}(\mu, \sigma^2)$ 的期望和方差. 若随机变量 $X \sim \mathcal{N}(0,1)$, 根据奇函数在对称的区间上积分为 0 有

$$E(X) = \int_{-\infty}^{+\infty} \frac{t}{\sqrt{2\pi}} e^{-t^2/2} dt = 0 \ .$$

根据方差的定义和分部积分有

$$\mathrm{Var}(X)=\int_{-\infty}^{+\infty}\frac{t^2}{\sqrt{2\pi}}e^{-t^2/2}dt=-\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}tde^{-t^2/2}=\left[\frac{te^{-t^2/2}}{-\sqrt{2\pi}}\right]_{t=-\infty}^{+\infty}+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-t^2/2}dt=1\,.$$

由此完成证明.

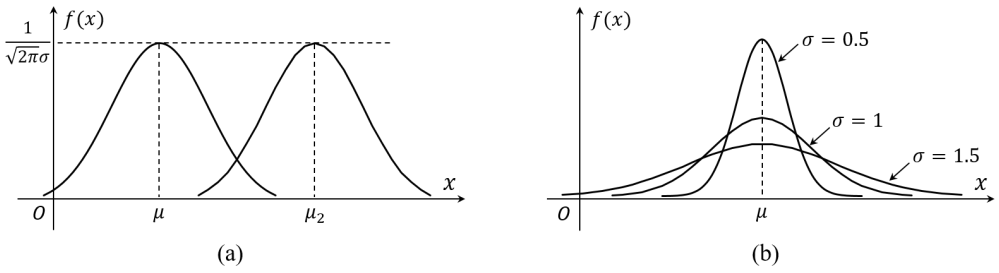

图 4.6 正太分布的密度函数

正太分布的密度函数如图 4.7(a) 所示, 具有以下一些特点:

1) 曲线 $f(x)$ 关于 $x=\mu$ 对称, 先单调递增, 之后单调递减, 在 $x=\mu$ 处取最大值 $1/\sqrt{2\pi}\sigma$. 说明随机变量 X 的取值主要集中在 $x=\mu=E(X)$ 附近, 离 $x=\mu$ 越远的区间概率越小.

2) 根据 $\lim_{x\to\pm\infty}f(x)=0$ 可知曲线 $f(x)$ 以 x 轴为渐近线; 根据 $f''(x)=0$ 可知曲线 $f(x)$ 的拐点为 $x=\mu\pm\sigma$.

3) 固定标准差 σ 而改变期望 μ 的值, 曲线 $f(x)$ 形状不变, 仅沿 x 轴左右平行移动, 如图 4.7(a).

4) 固定期望 μ 而改变标准差 σ 的值, 曲线 $f(x)$ 的对称点不变, 但最大值 $1/\sqrt{2\pi}\sigma$ 和拐点 $x=\mu\pm\sigma$ 发生了改变. 如图 4.7(b) 所示: 当 σ 越小, 曲线顶峰越高, 曲线越陡峭, 分布越集中, 方差越小; 反之 σ 越大, 曲线顶峰越地, 曲线越平坦, 分布越分散, 方差越大.

关于正太分布的概率估计, 有下面的不等式:

定理 4.7 若 $X\sim\mathcal{N}(0,1)$, 对任意 $\epsilon>0$ 有

$$\begin{aligned}P(X\geqslant\epsilon)&\leqslant\frac{1}{2}e^{-\epsilon^2/2}\\P(|X|\geqslant\epsilon)&\leqslant\min\left\{1,\sqrt{\frac{2}{\pi}}\frac{1}{\epsilon}e^{-\epsilon^2/2}\right\}.\end{aligned}$$

在上面的定理中, 第一个不等式具有广泛的应用, 在 $\epsilon\in(0,1)$ 时对真实的概率有更好的估计; 第二个不等式被称为 Mill 不等式, 在 $\epsilon\in(1,+\infty)$ 时对真实的概率有更好的估计.

证明 针对第一个不等式, 我们有

$$\begin{aligned} P(X \geqslant \epsilon) &= \int_{\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(x+\epsilon)^2/2} dx \\ &\leqslant e^{-\epsilon^2/2} \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \frac{1}{2} e^{-\epsilon^2/2} . \end{aligned}$$

对于 Mill 不等式, 根据 $\mathcal{N}(0,1)$ 的概率密度 $f(x) = e^{-x^2/2}/\sqrt{2\pi}$ 有 $f'(x) = -xf(x)$, 进一步可得

$$\begin{aligned} P(|X| \geqslant \epsilon) &= 2\int_{\epsilon}^{+\infty} f(t) dt = 2\int_{\epsilon}^{+\infty} \frac{tf(t)}{t} dt \\ &\leqslant 2\int_{\epsilon}^{+\infty} \frac{tf(t)}{\epsilon} dt = -2\int_{\epsilon}^{+\infty} \frac{f'(t)}{\epsilon} dt = -\frac{2}{\epsilon} \left[f(t) \right]_{\epsilon}^{+\infty} = \frac{2}{\sqrt{2\pi}\epsilon} e^{-\epsilon^2/2} . \end{aligned}$$

由此完成证明.

若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则有分布函数

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

该分布函数没有显示的表达式, 只能求数值解, 函数如图 4.7(a) 所示. 为便于研究正太分布的分布函数, 利用定理 4.5 可将其它正太分布都转化为标准正太分布 $\mathcal{N}(0,1)$, 设其分布函数为

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

表 4.1 给出了标准正太分布 $\Phi(x)$ 的函数表, 在计算具体的概率时可供查询. 下面给出关于分布函数 $\Phi(x)$ 的一些性质:

1) 根据对称性有 $\Phi(x) + \Phi(-x) = 1$.

2) 若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则对任意实数 $a < b$ 有

$$\begin{aligned} P(X < a) &= P\left(\frac{X-\mu}{\sigma} \leqslant \frac{a-\mu}{\sigma} \right) = \Phi\left(\frac{a-\mu}{\sigma} \right) , \\ P(X > b) &= 1 - P\left(\frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma} \right) = 1 - \Phi\left(\frac{b-\mu}{\sigma} \right) , \\ P(a \leqslant X \leqslant b) &= P\left(\frac{a-\mu}{\sigma} \leqslant \frac{X-\mu}{\sigma} \leqslant \frac{b-\mu}{\sigma} \right) = \Phi\left(\frac{b-\mu}{\sigma} \right) - \Phi\left(\frac{a-\mu}{\sigma} \right) . \end{aligned}$$

3) 若随机变量 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则对任意实数 $k > 0$ 有

$$P(|x-\mu| < k\sigma) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1 .$$

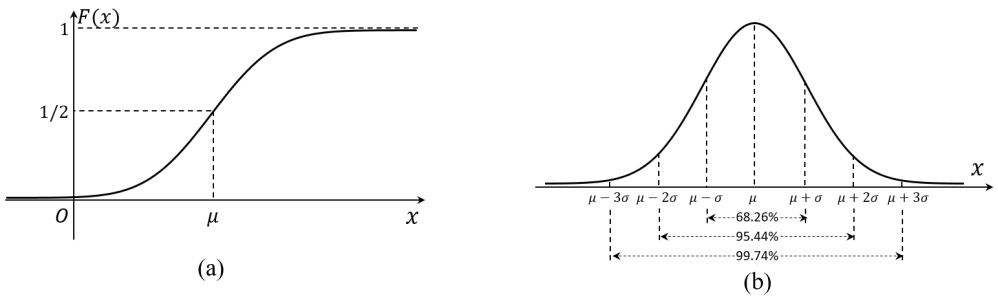

图 4.7 正太分布函数和 3σ 原则

特别的, 当 $k=1,2,3$ 时通过表 4.1 有

$$P(|x-\mu|<\sigma)=0.6826,\quad P(|x-\mu|<2\sigma)=0.9544,\quad P(|x-\mu|<3\sigma)=0.9974\ .$$

如图 4.7(b) 所示, 尽管随机变量 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的取值范围为整个实数域 $\mathbb{R}$, 但其取值落在 $[\mu-3\sigma,\mu+3\sigma]$ 之外的概率不超过千分之三, 也就是 $X\sim\mathcal{N}(\mu,\sigma^2)$ 的取值几乎总在 $[\mu-3\sigma,\mu+3\sigma]$ 之内, 这就是人们所说的"3σ 原则", 在实际的统计推断, 特别是产品质量检测中具有重要的应用.

4) 若随机变量 $X\sim\mathcal{N}(\mu,\sigma^2)$, 且已知 $P(X<c)=p$, 则有

$$p=P(X<c)=P\left(\frac{X-\mu}{\sigma}<\frac{c-\mu}{\sigma}\right)=\Phi\left(\frac{c-\mu}{\sigma}\right)\ ,$$

由此可反解出 $c=\mu+\sigma\Phi^{-1}(p)$. 这里 $\Phi^{-1}(x)$ 表示标准正太分布函数 $\Phi(x)$ 的反函数, 可根据表 4.1 由里向外查得, 例如 $\Phi^{-1}(0.5871)=0.22$.

例 4.11 已知某公司员工每个月的工资服从正太分布 $\mathcal{N}(6000,\sigma^2)$, 问题:

i) 若已知标准差 $\sigma=500$, 求工资在 5000 与 7000 之间的员工在公司中占比多少?

ii) 当标准差 σ 为何值时, 工资在 5000 与 7000 之间的员工在公司中占比为 0.803?

解 用随机变量 X 表示公司员工每个月的工资, 则 $X\sim\mathcal{N}(6000,\sigma^2)$. 针对问题 i), 当 $\sigma=500$ 时通过查询表 4.1 有

$$P(5000\leqslant X\leqslant 7000)=P\left(-2\leqslant\frac{X-6000}{500}\leqslant 2\right)=\Phi(2)-\Phi(-2)=2\Phi(2)-1=0.9544.$$

针对问题 ii) 有

$$P(5000\leqslant X\leqslant 7000)=P\left(-\frac{1000}{\sigma}\leqslant\frac{x-6000}{\sigma}\leqslant\frac{1000}{\sigma}\right)=2\Phi(\frac{1000}{\sigma})-1=0.803\ ,$$

于是得到 $\Phi(1000/\sigma)=0.9015$, 通过由内自外查表 4.1 有 $\sigma\approx 775.2$.

表 4.1 标准正态分布表 $\Phi(x)=\int_{-\infty}^{x}e^{-t^2/2}/\sqrt{2\pi}dt$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

4.5 连续随机变量函数的分布

当知道一个随机变量的概率分布后, 经常会考虑它的一些函数的分布, 例如当知道一个圆的直径 X 服从均匀分布 $U(a,b)$, 需要考虑圆的面积 $Y=\pi(X/2)^2$ 的分布. 一般地, 若已知随机变量 X 的分布函数, 以及 $g(x)$ 是定义在随机变量 X 所有可能取值的集合上的函数, 则称 $Y=g(X)$ 为随机变量 X 的函数, 很显然 Y 也是随机变量. 研究的问题可以归纳为: 若已知随机变量 X 的概率分布和函数 $g(x)$, 如何求解随机变量 $Y=g(X)$ 的概率分布.