软件工程统计方法

随机变量及其分布(二)

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

内 容 提 纲

- ❏ 统计学导论
- ❏ 描述统计
- ❏ 概率计算基础
- ❏ 随机变量及其分布
- ❏ 统计量及其抽样分布
- ❏ 参数估计
- ❏ 参数假设检验
- ❏ 非参数假设检验
- ❏ 方差分析
- ❏ 回归分析

1 本节内容

- 随机变量
- 分布函数
- 数学期望
- 方差
- 常用分布
- 二维随机向量

2 连续随机变量

Example 1 一个靶子是半径为2米的圆盘, 设击中靶上任意同心圆盘上的点的概率与该圆盘的面积成正比, 并设射击都能击中靶, 以X表示弹着点于圆心的距离. 试求随机变量X的分布函数.

解: 若$X < 0$, 则$\{X \leq x\}$是不可能事件, 于是$F(x) = P\{X \leq x\} = 0$. 若$0 \leq x \leq 2$, 由题意, $P\{0 \leq X \leq x\} = kx^2$, k是某一常数, 为确定k的值, 取$x = 2$, 有$P\{0 \leq X \leq 2\} = 2^2 k$, 但已知$P\{0 \leq X \leq 2\} = 1$, 故得$k = 1/4$, 即$P\{0 \leq X \leq x\} = x^2/4$, 于是

$$F(x) = P\{X \leq x\} = P\{X < 0\} + P\{0 \leq X \leq x\} = x^2/4$$

若$X > 2$, 由题意, 有$F(x) = P\{X \leq x\} = 1$. 综合上述, 即得X的分布函数为

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{4}, & 0 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

X的分布函数也可以表示为

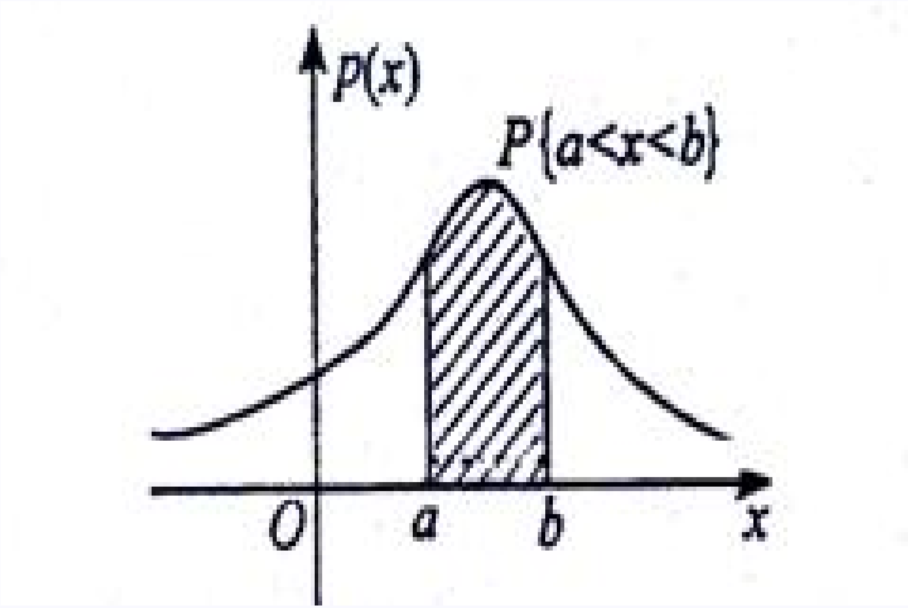
$$F(x) = \int_{-\infty}^{x} f(t)dt$$

其中$f(t) = t/2, 0 < t < 2$; $f(t) = 0$, 其他.

连续随机变量

连续型随机变量及其密度函数的定义:

Definition 1 (概率密度) 对于随机变量X, 其分布函数为$F(x)$, 如存在非负可积函数$f(x)$, 使得对于任意实数x, 有$F(x) = \int_{-\infty}^{x} f(t)dt$, 则称$X$为连续型随机变量, $f(x)$称为X的概率密度函数.



连续随机变量

概率密度函数的性质:

1. $f(x) \geq 0$;
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$;
3. 对于任意实数a,b($a < b$)，都有$P(a < X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$;
4. F 若$f(x)$在点x处连续, 则有$F'(x) = f(x)$, 即分布函数$F(x)$是概率密度的一个原函数.
5. 对于连续性随机变量X, X取任一指定实数值a的概率均为0, 即$P\{X = a\} = 0$.

例 子

Example 2 设随机变量X具有概率密度

$$f(x)=\begin{cases} kx, & 0\le x<3 \\ 2-\frac{x}{2}, & 3\le x\le 4 \\ 0, & \text{其他} \end{cases}$$

*(1)*确定常数k; *(2)*求X的分布函数$F(x)$; *(3)*求$P\{1<X\le \frac{7}{2}\}$.

解: (1)由$\int_{-\infty}^{+\infty} f(x)dx=1$得

$$\int_0^3 kxdx+\int_3^4 (2-\frac{x}{2})dx=1$$

解得$k=\frac{1}{6}$.
(2) 所以分布函数为

$$F(x)=\begin{cases} 0, & x<0 \\ \int_0^x \frac{t}{6}dt, & 0\le x<3 \\ \int_0^3 \frac{t}{6}dt+\int_3^x 2-\frac{t}{2}dt, & 3\le x<4 \\ 1, & x\ge 4 \end{cases}$$

(3)$P\{1<X\le 7/2\}=F(7/2)-F(1)=\frac{41}{48}$.

3 数字特征

连续随机变量数学期望

设连续型随机变量X的概率密度为$f(x)$, 如果广义积分$\int_{-\infty}^{+\infty} xf(x)dx$绝对收敛, 则称此积分值为随机变量$X$的数学期望, 记作$E(X)$或$EX$, 即$E(X) = \int_{-\infty}^{+\infty} xf(x)dx$.

如果上述级数或积分不绝对收敛, 则称此随机变量的数学期望不存在.
Theorem 1 设$Y = g(X)$是随机变量X的函数, g是连续函数. 若X为具有密度函数$f(x)$的连续型随机变量, $\int_{-\infty}^{+\infty} g(x)f(x)dx$ 绝对收敛, 则

$$E(Y) = E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

数 字 特 征

数学期望是一种最基本的数字特征。在物理学中, 矩是用于物体形状识别的重要参数指标。数学中矩的概念来自于物理学。在统计学中, 矩用来描述随机变量的数字特征。常用的矩有原点矩, 中心矩, 混合矩, 混合中心矩。

Definition 2 (矩, Moment) 设X,Y为两个随机变量, 不同矩的定义如下。

1. 若$E(X^k)$存在($k = 1, 2, \cdots$), 称$E(X^k)$为X的k阶原点矩或k阶矩, 记为μ_k。
2. 若$E((X - EX)^k)$存在($k = 1, 2, \cdots$), 称$E((X - EX)^k)$ 为X的k阶中点矩, 记为υ_k。
3. 若$E(X^kY^l)$存在(k,$l = 1$,2,$\cdots$), 则$E(X^kY^l)$ 称为X, Y的$k + l$混合矩。
4. 若$E((X - E(X))^k(Y - E(Y))^l)$存在($k$,$l = 1$,$2$,$\cdots$), 则$E((X - E(X))^k(Y - E(Y))^l)$ 称为X,Y的$k + l$混合中心矩。

数 字 特 征

Theorem 2 (中心矩的原点矩表示定理)

$$\upsilon_k = \sum_{i=0}^{k} C_k^i \mu_i (-\mu_1)^{k-i} \tag{1}$$

根据中心矩的原点矩表示定理, 前面四个k阶中心矩可以由原点矩表示如下：

$$\upsilon_1 = 0$$

$$\upsilon_2 = \mu_2 - \mu_1^2$$

$$\upsilon_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$$

$$\upsilon_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$

数 字 特 征

我们介绍两外两个常用的矩：偏度和峰度。

Definition 3 (偏度, Skewness) 若随机变量X的三阶距存在, 则

$$\gamma_1 = \frac{\upsilon_3}{\upsilon_2^{(\frac{3}{2})}} \tag{2}$$

称为X的偏度, 或偏度系数, 其中υ_2, υ_3分别为二阶和三阶中心矩。

偏度γ_1是一个描述分布对称程度的特征量。当偏度$\gamma_1 = 0$时, 概率密度呈对称分布(注意不一定是以$x = 0$对称)。当偏度$\gamma_1 < 0$时, 概率分布的左尾偏长, 称为负偏或左偏。当偏度$\gamma_1 > 0$时, 概率分布的右尾偏长, 称为正偏或右偏。由于标准正态分布的三阶中心矩为0, 所以它的偏度为0。

数 字 特 征

Definition 4 (峰度, Kurtosis) 若随机变量X的三阶距存在, 则

$$\gamma_2 = \frac{\upsilon_4}{\upsilon_2^2} \tag{3}$$

称为X的峰度, 或峰度系数, 其中υ_2, υ_4分别为二阶和四阶中心矩。

峰度γ_2是一个描述分布尖锐程度(或平坦程度)的特征量。峰度γ_2越大, 表示概率分布越尖锐, 反之越平坦均匀。由于标准正态分布的四阶中心矩为3, 二阶中心矩为1, 所以它的峰度为3。当$\gamma_2 = 3$, 说明随机变量的尖锐程度跟标准正态分布相当。当$\gamma_2 < 3$, 我们称之为低峰度, 说明随机变量的分布比标准正态分布平坦。当$\gamma_2 > 3$, 我们称之为高峰度, 说明随机变量的分布比标准正态分布尖锐。有些文献为了方便跟标准正态分布比较, 将峰度的定义改为3中的比值减去3。

数 字 特 征

连续随机变量的概率密度函数的$[-\infty,\infty]$积分为1, 即累积面积为1。在统计推断中, 我们经常用到一个数字x_p, 使得在x_p左边或右边的面积为$p(0<p<1)$。这种数字特征我们成为分位数。

Definition 5 (分位数) 设连续随机变量X的概率密度函数为$f(X)$, 对于任意$0<p<1$, 则满足条件

$$\int_{\alpha_p}^{\infty} f(x) = p$$

的α_p称为上侧p分位数。

满足条件

$$\int_{-\infty}^{\beta_p} f(x) = p$$

的β_p称为下侧p分位数。

在统计应用中, 一些常用的分布通常预先计算好分位数表。需要使用时, 直接通过概率查询对应分位数或者通过分位数查询概率, 而不需要计算繁琐的积分。上侧分位数和下侧分位数具有如下对应关系。

$$\alpha_p = \beta_{1-p}, \beta_p = \alpha_{1-p}$$

本书采用上侧分位数作为查表计算方法。当$p=0.5$时, 上侧分位数和下侧分位数相等, 此时称为中位数。

4 均匀分布

Definition 6 (均匀分布) 如果随机变量X的概率密度为:

$$f(x)=\begin{cases}\frac{1}{b-a} & a<x<b\\ 0 & x\le a, or, x\ge b\end{cases}\tag{4}$$

则称X在区间(a,b)内服从均匀分布,记为$X\sim U(a,b)$, 其分布函数为

$$F(x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x<b\\ 1 & x\ge b\end{cases}\tag{5}$$

数字特征:

$$E(X)=\int_{-\infty}^{+\infty}xf(x)dx=\int_{-\infty}^{+\infty}\frac{x}{b-a}dx=\frac{a+b}{2}$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)dx==\int_{-\infty}^{+\infty}\frac{x^2}{b-a}dx$$

$$=\frac{b^3-a^3}{3(b-a)}=\frac{a^2+b^2+ab}{3}$$

$$Var(X)=\frac{a^2+b^2+ab}{3}-\frac{a^2+b^2+2ab}{4}=\frac{(b-a)^2}{12}$$

均 匀 分 布

Example 3 设电阻值R是一个随机变量, 均匀分布在$900-1100$欧. 求R的概率密度及R落在$950-1050$的概率.

$$f(r)=\frac{1}{1100-900}, 900\le r\le 1100$$

则有

$$F(1050)-F(950)=\int_{950}^{1050}f(r)dr=0.5$$

思考：离散均匀分布？

5 指数分布

Definition 7 如果随机变量X的概率密度为:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{6}$$

或者

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{7}$$

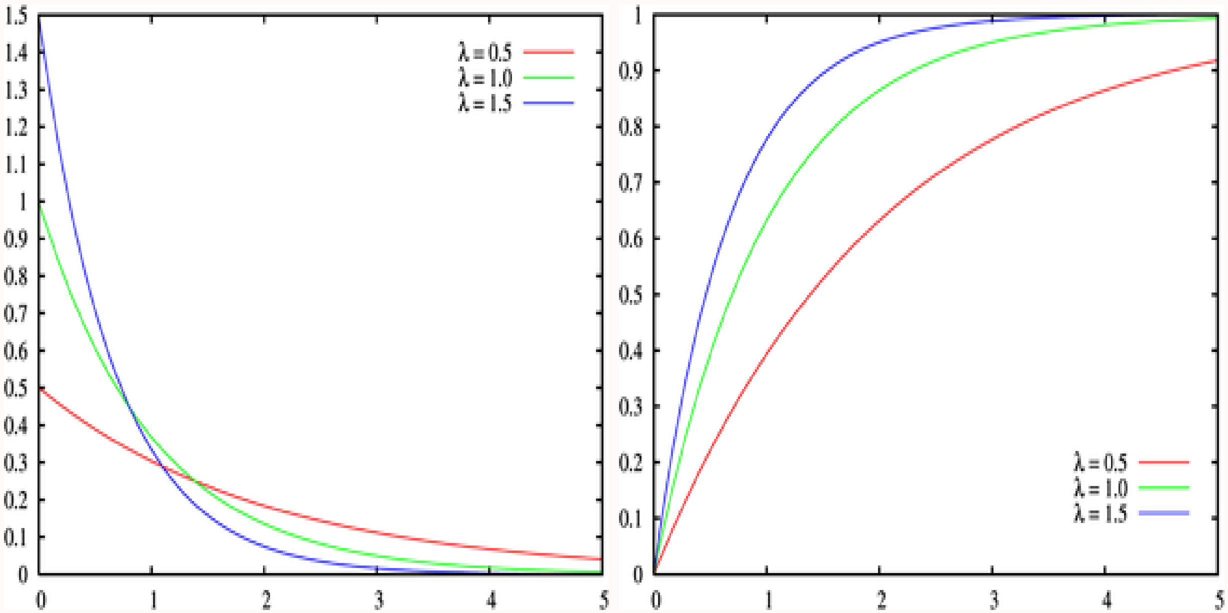
则称X服从参数为$\lambda(\lambda > 0)$的指数分布, 记为$X \sim E(\lambda)$, 其分布函数为:

$$F(x) = \begin{cases} 1 - e^{-\frac{x}{\theta}} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{8}$$

或者

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{9}$$

指 数 分 布


1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5
λ = 0.5
λ = 1.0
λ = 1.5
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5
λ = 0.5
λ = 1.0
λ = 1.5

指 数 分 布

$$P\{X > s + t | X > s\} = P\{X > t\}$$

这一性质称为指数分布的无记忆性. 事实上可以证明指数分布是唯一具有上述性质的连续型分布.

证明:

$$P\{X > s + t | X > s\} = \frac{P\{(X > s + t) \cap (X > s)\}}{P\{X > s\}} = \frac{P\{X > s + t\}}{P\{X > s\}}$$

$$= \frac{P\{X > s + t\}}{P\{X > s\}} = \frac{1 - E(s + t))\}}{1 - E(s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda(s)}} = e^{-\lambda t} = P\{X > t\}$$

指数分布数字特征

Theorem 3 设随机变量X指数分布, 即概率密度为

$$f(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, x > 0$$

则$E(X) = \theta = \frac{1}{\lambda}$, $Var(X) = \theta^2 = \frac{1}{\lambda^2}$.

证:

$$E(X) \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} x\frac{1}{\theta}e^{-\frac{x}{\theta}}dx$$

$$= -xe^{-\frac{x}{\theta}}|_0^{\infty} + \int_0^{\infty} e^{-\frac{x}{\theta}} = \theta$$

$$E(X^2) \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^{\infty} x^2\frac{1}{\theta}e^{-\frac{x}{\theta}}dx$$

$$= -x^2e^{-\frac{x}{\theta}}|_0^{\infty} + \int_0^{\infty} 2xe^{-\frac{x}{\theta}} = 2\theta^2$$

$$Var(X) = E(X^2) - E(X)^2 = \theta^2$$

指数分布

泊松过程是一种重要的随机过程。泊松过程中，第k次随机事件与第$k+1$次随机事件出现的时间间隔服从指数分布。这是因为，第k次随机事件之后长度为t的时间段内，第$k+1$次随机事件出现的概率等于1减去这个时间段内没有随机事件出现的概率。而根据泊松过程的定义，长度为t的时间段内没有随机事件出现的概率等于λt的泊松分布概率

$$\frac{(\lambda t)^0}{0!}e^{-(\lambda t)} = e^{-(\lambda t)}$$

Example 4 已知设备无故障运行*10*个小时, 求再无故障运行*8*小时的概率.

解:

$$P\{T \geq 18|T > 10\} = \frac{P\{T > 18\}}{P\{T > 10\}} = P\{T > 8\} = e^{-8t}$$

指 数 分 布

Example 5 电子元件的寿命X(年）服从参数为$\lambda = 3$的指数分布:

(1) 求该电子元件寿命超过2年的概率。

(2) 已知该电子元件已使用了1.5年，求它还能使用两年的概率为多少？

6 正态分布

Definition 8 如果随机变量X 的概率密度为:

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

则称X 服从参数为μ, σ^2 的正态分布,记为$X \sim N(\mu, \sigma^2)$,其中参数$\mu \in R, \sigma > 0$.

正态分布最早是棣莫弗在1718年著作的书籍的及1734年发表的一篇关于二项分布文章中提出的，当二项随机变量的位置参数n很大及形状参数为1/2时，则所推导出二项分布的近似分布函数就是正态分布。拉普拉斯在1812年发表的《分析概率论》中对棣莫弗的结论作了扩展到二项分布的位置参数为n及形状参数为p时。现在这一结论通常被称为棣莫佛—拉普拉斯定理。拉普拉斯在误差分析试验中使用了正态分布。勒让德于1805年引入最小二乘法这一重要方法；而高斯则宣称他早在1794年就使用了该方法，并通过假设误差服从正态分布给出了严格的证明。

正 态 分 布

当$\mu = 0, \sigma = 1$,称X服从标准正态分布, 记为$X \sim N(0,1)$, 分布函数记为$\Phi(x)$. 有

$$\Phi(-x) = 1 - \Phi(x)$$

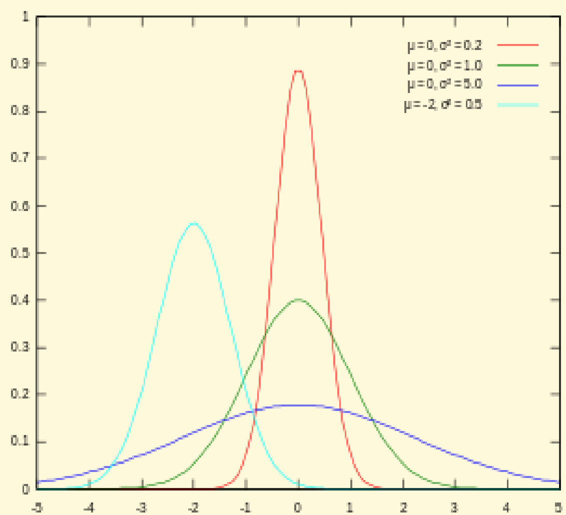
练习：标准正态分布查表。

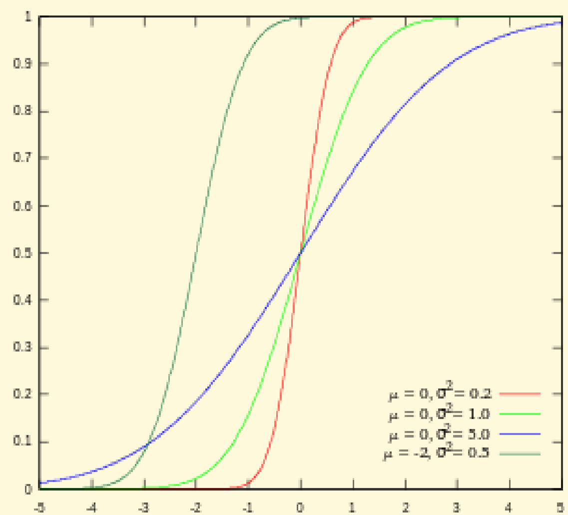
在自然现象和社会现象中, 大量随机变量都服从或近似服从正态分布. 如人的身体特征指标(身高、体重), 学习成绩, 产品的数量指标等等都服从正态分布. 许多较复杂的指标, 只要在受到的大量因素作用下每个因素的影响都不显著, 且因素相互独立, 也可认为近似服从正态分布. 又如二项分布、泊松分布在n很大时, 也以正态分布为极限分布. 因此, 可以说正态分布是最重要的分布.

正 态 分 布

正态分布性质:

- 曲线关于$x=\mu$对称.
- 当$x=\mu$时取到最大值$\frac{1}{\sqrt{2\pi}\sigma}$.
- 固定σ, 改变μ, 曲线沿O_x轴平移;
- 固定μ, 改变σ, 由于最大值为$\frac{1}{\sqrt{2\pi}\sigma}$. 曲线变得越尖, 因而$X$落在$\mu$附近的概率越大.





正 态 分 布

正态分布标准化:

Theorem 4 若$X \sim N(\mu, \sigma^2)$, 则$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

证明:

$$P\{Z \le x\} = P\{\frac{X-\mu}{\sigma} \le x\} = P\{X \le \mu + \sigma x\}$$

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

令$y = \frac{t-u}{\sigma}$, 则

$$P\{Z \le x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy = \Phi(x)$$

所以$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

正态分布数字特征

Theorem 5 $Z \sim N(0,1)$, 证明$E(X)=0$, $Var(X)=1$.

证明:

$$E(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} te^{-\frac{t^2}{2}}dt=0$$

$$Var(Z)=E(Z^2)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t^2e^{-\frac{t^2}{2}}dt$$

$$=-\frac{1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}}|_{-\infty}^{+\infty}+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}}dt=1$$

对于任意$X \sim N(\mu,\sigma^2)$, $Z=\frac{X-\mu}{\sigma} \sim N(0,1)$. 所以

$$E(X)=E(\mu+\sigma Z)=\mu$$

$$Var(X)=Var(\mu+\sigma Z)=\sigma^2 D(Z)=\sigma^2$$

正 态 分 布

Example 6 一种电子元件的使用寿命$X \sim \mathbb{N}(100, 15^2)$,某仪器上装有*3*个这种元件，三个元件损坏与否是相互独立的.求：使用的最初*90*小时内无一元件损坏的概率.

正 态 分 布

Example 7 一批钢材(线材)长度$X \sim N(\mu, \sigma^2)$
*(1)*若$\mu = 100$, $\sigma = 2$, 求这批钢材长度小于*97.8cm*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?

解: (1)

$$P\{X < 97.8\} = \Phi(\frac{97.8-100}{2}) = 1-\Phi(1.1) = 1-0.8643 = 0.1357$$

(2)

$$0.90 \le P\{97 < X < 103\} = \Phi(\frac{103-100}{\sigma}) - \Phi(\frac{97-100}{\sigma})$$

$$= 2\Phi(\frac{3}{\sigma}) - 1 \Rightarrow \Phi(\frac{3}{\sigma}) \ge 0.95 \Rightarrow \frac{3}{\sigma} \ge 1.645$$

所以$\sigma \le 1.8237$.

正 态 分 布

Example 8 将一温度调节器放置在存储着某种液体的容器内, 调节器定在d, 液体的温度X是一个随机变量, 且$X \sim N(d, 0.5^2)$. *(1)*若$d = 90$, 求$X < 89$的概率; *(2)*若要求保持液体的温度至少为*80*的概率不低于*0.99*, 问d至少为多少*?*

正态分布

解: (1)

$$P\{X < 89\} = P\{\frac{X-90}{0.5} < \frac{89-90}{0.5}\} = \Phi(\frac{89-90}{0.5})$$
$$= \Phi(-2) = 1 - \Phi(2) = 1 - 0.9772 = 0.228$$

(2)

$$0.99 \le P\{X \ge 80\} = P\{\frac{X-d}{0.5} \ge \frac{80-d}{0.5}\}$$
$$= 1 - P\{\frac{X-d}{0.5} < \frac{80-d}{0.5}\} = 1 - \Phi(\frac{80-d}{0.5})$$
$$\Phi(\frac{80-d}{0.5}) \le 1 - 0.99 = 1 - \Phi(2.327) = \Phi(-2.327)$$

即$\frac{80-d}{0.5} \le -2.327$, $d > 81.1635$.