MEM6810 工程系统建模与仿真

案例 软件

第二讲: 计算机仿真初步

沈海辉

中美物流研究院
上海交通大学

shenhaihui.github.io/teaching/mem6810p
shenhaihui@sjtu.edu.cn

2024年春 (MEM非全日制)

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

董浩云智能制造与服务管理研究院
CY TUNG Institute of Intelligent Manufacturing and Service Management
(中美物流研究院)
(Sino-US Global Logistics Institute)

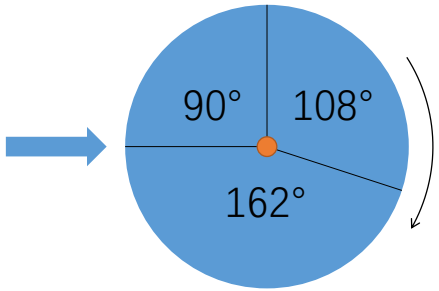
目录

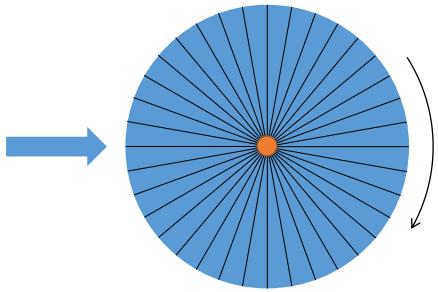
随机数的生成

- 假设某个库房有一些运输车辆到达需要卸载，经统计发现卸载所需时长的概率分布表如下：

卸载时长/分	概率
10	~~2/6~~ 0.30
20	~~3/6~~ 0.45
30	~~1/6~~ 0.25

- 如何对到达车辆的卸载时长进行模拟呢？
  1. 掷骰子
  2. 转轮盘





- 如果我们知道如何从 0 到 1 之间“随机且均匀”地抽出若干数字，那么我们便可以模拟任何分布！

随机数的生成

- 从 $(0,1)$ 区间上的连续均匀分布中抽取的独立随机样本的观测值, 被称为 $\mathrm{uniform}(0,1)$ 随机数 (random numbers), 有时也简称为随机数.
- 如果随机变量 $U \sim \mathrm{uniform}(0,1)$, 那么

$$\mathbb{E}[U] = 1/2,\ \mathrm{Var}(U) = 1/12.$$

- 使用 MATLAB 生成的 10 个 $\mathrm{uniform}(0,1)$ 随机数: 0.8147, 0.9058, 0.1270, 0.9134, 0.6324, 0.0975, 0.2785, 0.5469, 0.9575, 0.9649.
- $\mathrm{uniform}(0,1)$ 随机数的统计性质:
  - 均匀性: $(0,1)$ 区间上的每个值都有一样的可能性.
  - 独立性: 隐含前后的数相互之间无相关性.

随机数的生成

- 均匀性

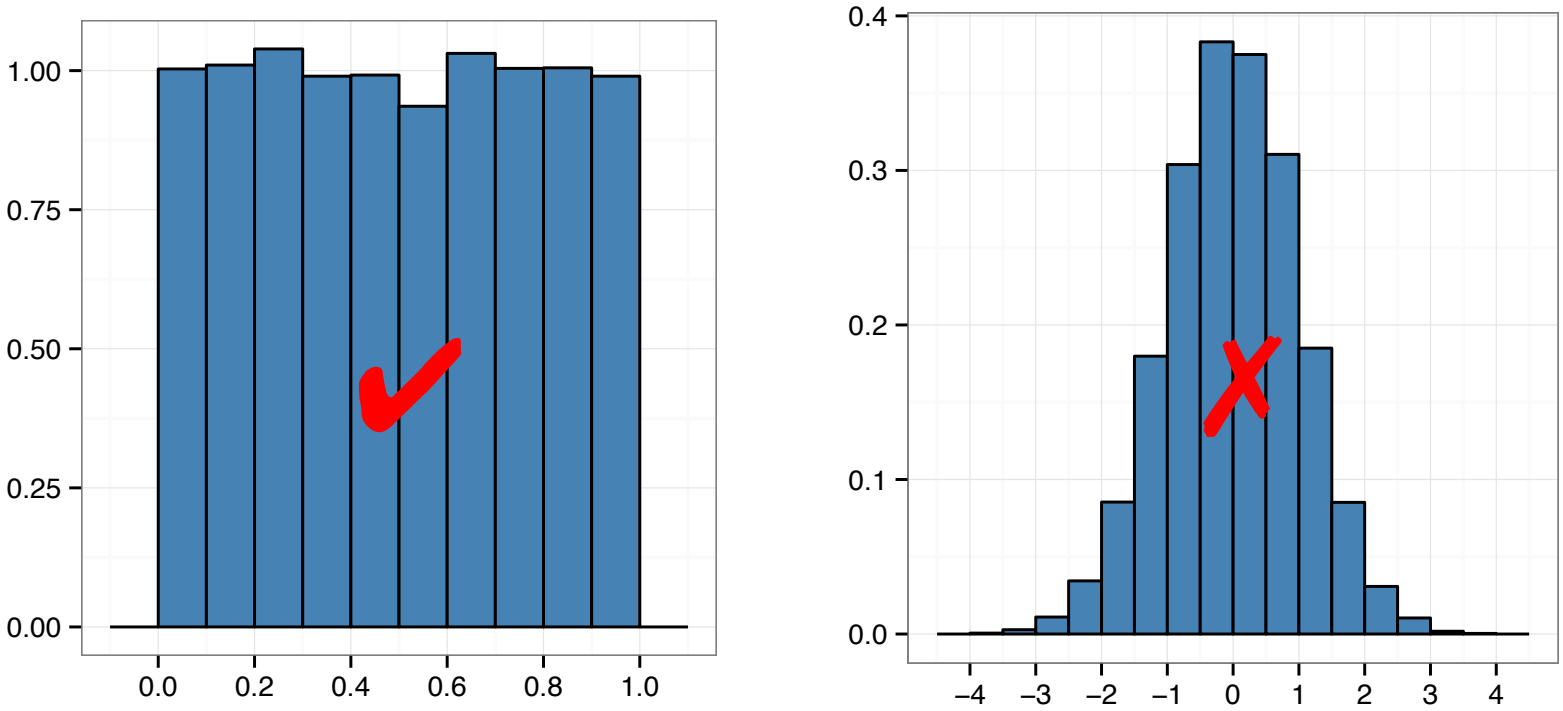


图: 经验概率密度函数 (即, 放缩后的频率直方图, 形状相同), 均匀性 vs 非均匀性 (*from* ZHANG Xiaowei)

随机数的生成

- 独立性

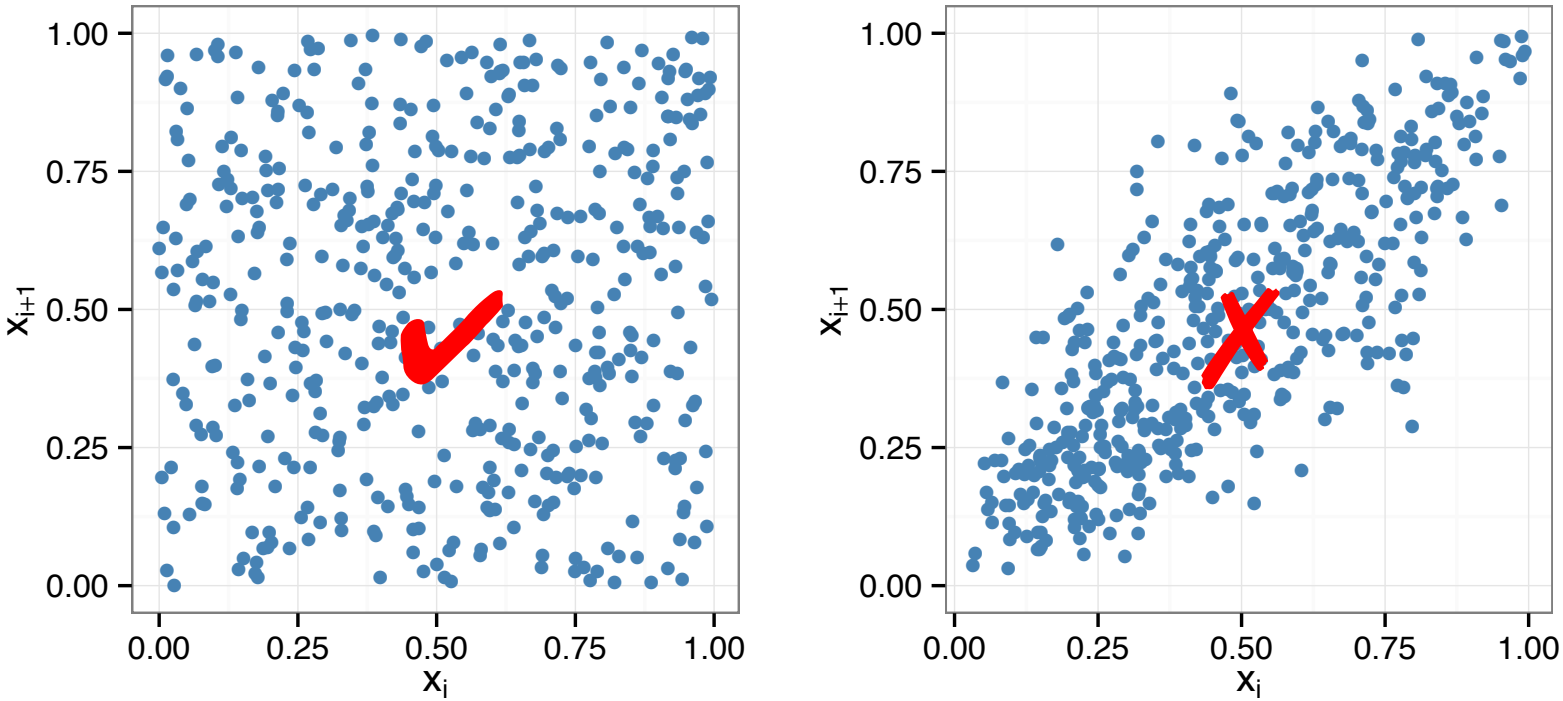


图: 散点图, 无相关性 vs 有相关性 (*from* ZHANG Xiaowei)

- 计算机无法产生真正的随机性! 它只能产生一些**伪随机**数 (pseudo-random numbers).
- "伪" 意味着不是真正的随机.
  - 随机数是通过某种算法来生成的, 这就消除了随机性.
  - 生成的随机数序列可以被复现.
- 目标: 生成 $(0, 1)$ 范围内的一系列数字, 使他们可以显示出和 $\mathrm{uniform}(0, 1)$ 随机数一样的性质.
  - 统计性质是最重要的.
  - 是否是真随机是次要的.

- 优秀的随机数发生器 (random number generator, RNG) 所需的性质:
  1. 通过统计性检验.
  2. 坚实的理论基础.
  3. 快.
  4. 足够长的周期.
  5. 可移植性好.
  6. 可复现.

- 随机数发生器的一些技术:
  - 线性同余发生器 (Linear Congruential Generator, LCG)
  - 组合线性同余发生器 (Combined LCG)
  - 多重递归发生器 (Multiple Recursive Generator, MRG)

- 线性同余发生器是一种简单的早期的随机数发生器.

1. 通过下述递归式产生一系列 0 到 $m-1$ 之间的整数 $x_0, x_1, x_2, \ldots$:

$$x_{i+1} = (ax_i + c) \bmod m, \quad i = 0, 1, 2, \ldots.$$

   - mod 表示取模操作; 初始值 x_0 称为**种子** (seed), a 称为**乘子** (multiplier), c 称为**增量** (increment), m 称为**模数** (modulus).

2. 将 x_i 变换到 0 和 1 之间的数值 u_i:

$$u_i = \frac{x_i}{m}, \quad i = 0, 1, 2, \ldots.$$

   - u_i 的可能取值: $\{0, \frac{1}{m}, \ldots, \frac{m-1}{m}\}$. (可能不完全覆盖!)

- a, c, m, 和 x_0 的选取对统计性质和周期长度有极大的影响.

- 例子: 使用 LCG, 并取 $x_0 = 27$, $a = 17$, $c = 43$, 及 $m = 100$.

$$
\begin{aligned}
x_0 &= 27 \\
x_1 &= (17 \times 27 + 43) \bmod 100 = 502 \bmod 100 = 2 \\
u_1 &= 2/100 = 0.02 \\
x_2 &= (17 \times 2 + 43) \bmod 100 = 77 \bmod 100 = 77 \\
u_2 &= 77/100 = 0.77 \\
x_3 &= (17 \times 77 + 43) \bmod 100 = 1352 \bmod 100 = 52 \\
u_3 &= 52/100 = 0.52 \\
x_4 &= (17 \times 52 + 43) \bmod 100 = 927 \bmod 100 = 27 \\
u_4 &= 27/100 = 0.27
\end{aligned}
$$

周期长度只有 4!

- 访问 `https://xiaoweiz.shinyapps.io/randNumGen` 尝试不同的参数值.

- LCG 的一个实际使用 (Lewis et al. 1969): $a = 7^5$, $c = 0$, $m = 2^{31} - 1 = 2{,}147{,}483{,}647$ (一个质数).
  - 它采用 $u_i = \frac{x_i}{m+1}$.
  - 它可通过许多标准的统计性检验.
  - 周期长度 $\approx 2^{31} - 2 \approx 2 \times 10^9$ (超过 20 亿).
- 注: 通过令模数 m 为 2 的幂 (或者接近), 取模运算可以更加高效, 因为大多数计算机是采用二进制来表示数字的.
- 随机计算机算力的增长, 简单的 LCG 如今已经无法胜任了; 实际中我们使用更加复杂的随机数发生器.

- Combined LCG: 将 J (≥ 2) 个 LCG 组合起来 (其中 $c = 0$).
- 对于 32 位计算机, L'Ecuyer (1988) 提出将 $J = 2$ 个 LCG 组合, 其中 $a_1 = 40{,}014$, $m_1 = 2{,}147{,}483{,}563$, $a_2 = 40{,}692$, 及 $m_2 = 2{,}147{,}483{,}399$.
  1. 从 $[1, m_1 - 1]$ 中为第一个发生器选择种子 $x_{1,0}$, 从 $[1, m_2 - 1]$ 中为第二个发生器选择种子 $x_{2,0}$. 令 $j = 0$.
  2. 计算
  $$x_{1,j+1} = a_1 x_{1,j} \bmod m_1,$$
  $$x_{2,j+1} = a_2 x_{2,j} \bmod m_2.$$
  3. 令 $x_{j+1} = (x_{1,j+1} - x_{2,j+1}) \bmod (m_1 - 1)$.
  (注: mod 使用 floored division, 即, $y \bmod m = y - m\lfloor \frac{y}{m} \rfloor$.)
  4. 返回
  $$u_{j+1} = \begin{cases} \frac{x_{j+1}}{m_1}, & \text{当 } x_{j+1} > 0, \\ \frac{m_1 - 1}{m_1}, & \text{当 } x_{j+1} = 0. \end{cases}$$
  5. 令 $j = j + 1$ 并跳转至第 2 步.

它的周期长度为 $(m_1 - 1)(m_2 - 1)/2 \approx 2 \times 10^{18}$.

- Multiple Recursive Generator (MRG): 通过使用更高阶的递归来拓展 LCG:

$$x_i = (a_1 x_{i-1} + a_2 x_{i-2} + \cdots + a_k x_{i-K}) \bmod m.$$

- 一个被广泛采用的特例为 MRG32k3a† (L'Ecuyer 1999), 它属于 *combined MRG*, 其中 $J = 2$ 及 $K = 3$.
  - 它的周期长度为 $\approx 3 \times 10^{57}$, 这是一个极大的数.
  - 假设你每秒可以生成 10 亿 (10^9) 个伪随机数, 那么穷尽 MRG32k3a 的周期所需的时间比当前宇宙的年龄还要长!
- 在仿真软件及编程语言中广泛使用的那些知名的随机数发生器, 其统计性质都接受过广泛的检验并被证明有效.
- 当你手中的随机数发生器并不知名或者没有任何记录, 你需要格外小心!
  - 即便是在一些大众商业软件 (如, Excel, Visual Basic) 中用了多年的发生器, 都曾被发现存在一些缺陷 (L'Ecuyer 2001).

† MRG32k3a 或其适配是 MATLAB, R, SAS, Arena 等软件所使用的随机数发生器中的一种.

- 在 Excel 中, 可以直接使用函数

```
RAND()
```

  生成 $\mathrm{uniform}(0,1)$ 随机数.

- 若要生成 $\mathrm{uniform}(a,b)$ 随机数, 其中 $a<b$, 可使用

```
a+(b-a)*RAND()
```

- 若要生成 $[a,b]$ 上的离散均匀分布的随机数 (包含端点), 其中 $a<b$, 可使用

```
RANDBETWEEN(a, b)
```

  或者

```
FLOOR(a+(b+1-a)*RAND())
FLOOR.MATH(a+(b+1-a)*RAND())
```

- Monty Hall Problem, 又称三门问题、山羊汽车问题
  - 出自美国电视游戏节目 *Let's Make a Deal*, 并以它的主持人 Monty Hall 命名.

决胜21点 (2008)

导演: 罗伯特·路克蒂克
编剧: Peter Steinfeld / 阿兰·里布
主演: 吉姆·斯特吉斯 / 凯文·史派西 / 凯特·波茨沃斯 / 艾伦·余 / 莉萨·拉皮拉 / 更多...
类型: 剧情 / 犯罪
官方网站: www.sonypictures.com/movies/21/
制片国家/地区: 美国
语言: 英语
上映日期: 2008-03-28(美国)
片长: 123分钟
又名: 玩转21点 / 斗智21点 / 攻陷拉斯维加斯 / 21 - The Movie / 21: Blackjack
IMDb链接: tt0478087

豆瓣评分
6.9 60016人评价
5星 8.8%
4星 35.3%
3星 49.3%
2星 6.0%
1星 0.6%

好于 29% 剧情片
好于 48% 犯罪片

想看 看过 评价: ☆☆☆☆☆

写短评 写影评 分享到 推荐

决胜21点的剧情简介 · · · · · ·

Ben Campbell（吉姆·斯特加斯 Jim Sturgess 饰）有着惊人的才华，身为麻省理工高材生的他学业无懈可击，他亦毫无意外地赢得了哈佛医学院的录取通知书。然而30万的高昂学费和生活费令他的大学梦摇摇欲坠。在争取奖学金的面试中，教授对他说胜出者必须要有过人的经历而不是像本这种一张白纸的学生。

Ben在一服装店打工，挣取每小时8美元的薪酬。同时和两个好友准备竞赛2.09以期获得认同和奖金。数学课上本的天才头脑被教授Mickey Rosa（凯文·史派西 Kevin Spacey 饰）发现，Mickey 希望本加入自己的21算法团队，专门去赌场依靠算牌赢得大钱。Ben并不同意，但Ben一直暗恋的女孩Jill Taylor（#凯特·波茨沃斯 Kate Bosworth 饰）也出面诱惑时，Ben开始动摇。

Ben开始了严密的训练，出师的成功让Ben尝到了金钱、虚荣、欲望的权力。同时他和旧友开始疏远，渐渐迷失在赌场的漩涡里。 ©豆瓣

- 最简单的分析:
  - 如果“不换”, 一旦选好结果便确定了,

$$\mathbb{P}(\text{选中车}) = 1/3.$$

  - 如果“换”, 一旦选好结果也确定了 (一开始选中车, 最后会选中羊; 一开始选中羊, 最后会选中车). 因此,

$$\mathbb{P}(\text{最后选中车}) = \mathbb{P}(\text{一开始选中羊}) = 2/3.$$

- 不信? 让我们来做一下仿真实验
  - 访问 http://www.rossmanchance.com/applets/MontyHall/Monty04.html 试一下!
  - 用 Excel 来实现.

- 生日问题: 假设班上有 60 名同学, 那么至少有两个同学生日为同一天 (月日) 的概率为多少? (一年按 365 天计.) 99.41%
- 分析计算
  - 先计算全班生日不同的概率:

$$\mathbb{P}(\text{全班不同}) = \frac{365 \times 364 \times \cdots \times 306}{365^{60}}.$$

  - 于是,

$$\begin{aligned}\mathbb{P}(\text{至少有两人相同}) &= 1 - \mathbb{P}(\text{全班不同}) \\ &= 1 - \frac{365 \times 364 \times \cdots \times 306}{365^{60}} \\ &= 1 - 0.0059 = 0.9941.\end{aligned}$$

- 使用 Excel 进行仿真.

- 未婚妻问题 (Fiancee Problem), 又称公主选驸马问题、秘书问题 (Secretary Problem)
  - 最早由美国数学家 Merrill M. Flood 在 1949 提出.
- 基本问题描述:
  - 要从 N 个人中挑选出一位; N 是一个已知数, 比如, $N = 10$.
  - 候选者以随机 (谁先谁后概率均等) 的顺序到来.
  - 我们看到候选者之后, 会为TA打一个分数 (不会出现同分):
    - 这个分数只与候选者的特质有关, 与出现顺序无关;
    - 可理解为候选者的客观的优秀 (匹配) 程度.
  - 看到一位候选者之后, 我们有两种选择:
    - 选择接受, 则挑选环节结束;
    - 选择拒绝, 则继续看下一位, 并且之后不能再反悔重新选TA.
  - 如果前 $N-1$ 位都没接受, 则必须接受第 N 位.
  - **问题**: 采用何种策略, 可以以最大的概率选择到真正最优秀 (最匹配) 的人?

分析计算*

- 已知最优的策略具有如下结构: 拒绝前 k 人, 从第 $k+1$ 位起, 一旦TA的分数超过一开始的 k 人, 就接受TA; 否则继续.
  - 可通过**动态规划**的方法来得出严格的证明.
- 在最优策略的结构下, 如何确定最优的 k (记为 k^*)?
  - 以 $\mathbb{P}(k)$ 表示选中最优秀 (最匹配) 者的概率.
  - 先推导出 $\mathbb{P}(k)$ 关于 k 的表达式.
  - 再求解使 $\mathbb{P}(k)$ 最大的 k, 即 k^*.
- 特殊情形
  - 若 $N=2$, 任何策略下, 选对的概率都为 $1/2$, 问题退化; 故以下只考虑 $N \geq 3$ 的情形.
  - $k=0$, 对应情况为, 一定接受第一位, 此时 $\mathbb{P}(0)=1/N$.
  - $k=N-1$, 对应情况为, 一定接受第 N 位, 此时 $\mathbb{P}(N-1)=1/N$.

分析计算 (续)*

对于 $1 \le k \le N-1$,

$$\begin{aligned}
\mathbb{P}(k) &= \sum_{i=k+1}^{N} \mathbb{P}(\text{选中第 } i \text{ 个} \cap \text{第 } i \text{ 个为最优}) \\
&= \sum_{i=k+1}^{N} \mathbb{P}(\text{选中第 } i \text{ 个} | \text{第 } i \text{ 个为最优}) \mathbb{P}(\text{第 } i \text{ 个为最优}) \\
&= \frac{1}{N} \sum_{i=k+1}^{N} \mathbb{P}(\text{选中第 } i \text{ 个} | \text{第 } i \text{ 个为最优}) \\
&= \frac{1}{N} \sum_{i=k+1}^{N} \mathbb{P}(\text{前 } i-1 \text{ 人中的最优者在前 } k \text{ 人中} | \text{第 } i \text{ 个为最优}) \\
&= \frac{1}{N} \sum_{i=k+1}^{N} \frac{k}{i-1} = \frac{k}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1}\right).
\end{aligned}$$

分析计算 (续)*

对于 $2 \le k \le N-1$, 有

$$\mathbb{P}(k) = \frac{k}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1}\right),$$

$$\begin{aligned}\mathbb{P}(k-1) &= \frac{k-1}{N}\left(\frac{1}{k-1} + \frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1}\right) \\ &= \frac{1}{N} + \frac{k-1}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1}\right).\end{aligned}$$

因此,

$$\begin{aligned}\mathbb{P}(k) - \mathbb{P}(k-1) &= \frac{1}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1}\right) - \frac{1}{N} \\ &= \frac{1}{N}\left(\frac{1}{k} + \frac{1}{k+1} + \cdots + \frac{1}{N-1} - 1\right).\end{aligned}$$

注意到, 该等式在 $k = 1$ 时, 也成立.

分析计算 (续)*

进一步注意到以下几点:

- $\mathbb{P}(k)-\mathbb{P}(k-1)$ 随着 k 增大而减小;
- $k=1$ 时, $\mathbb{P}(1)-\mathbb{P}(0)=\frac{1}{N}\left(1+\frac{1}{k+1}+\cdots+\frac{1}{N-1}-1\right)>0$;
- $k=N-1$ 时, $\mathbb{P}(N-1)-\mathbb{P}(N-2)=\frac{1}{N}\left(\frac{1}{N-1}-1\right)<0$.

因此, 必定存在一个 k, 使得 $\mathbb{P}(k)$ 取到最大值, 该值即为所求 k^*. 且 k^* 必定满足, $\mathbb{P}(k^*)-\mathbb{P}(k^*-1)\geq 0$, $\mathbb{P}(k^*+1)-\mathbb{P}(k^*)<0$. 换言之, k^* 为满足条件

$$\mathbb{P}(k)-\mathbb{P}(k-1)\geq 0,\ \text{即},$$
$$\frac{1}{k}+\frac{1}{k+1}+\cdots+\frac{1}{N-1}\geq 1,$$

的最大的 k.

- 结论: 最优的策略为, 拒绝前 k^* 人, 从第 k^*+1 位起, 一旦TA的分数超过一开始的 k^* 人, 就接受TA; 否则继续. 其中 k^* 为满足 $\frac{1}{k}+\frac{1}{k+1}+\cdots+\frac{1}{N-1}\geq 1$ 的最大的 k, 且在该策略下, 选中最优的概率为 $\mathbb{P}(k^*)=\frac{k^*}{N}\left(\frac{1}{k^*}+\frac{1}{k^*+1}+\cdots+\frac{1}{N-1}\right)$.
- 若 $N=10$, 则 $k^*=3$, $\mathbb{P}(k^*)=0.3987$;
  若 $N=50$, 则 $k^*=18$, $\mathbb{P}(k^*)=0.3743$;
  若 $N=100$, 则 $k^*=37$, $\mathbb{P}(k^*)=0.3710$.
- 通过稍微进一步的分析, 还可以证明, 当 $N\to\infty$ 时,

$$\frac{k^*}{N}\to\frac{1}{e},\quad \mathbb{P}(k^*)\to\frac{1}{e},$$

  其中 $e:=\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n\approx 2.7183$ 为自然常数, $\frac{1}{e}=0.3679$.
- 使用 Excel 进行仿真.

1 随机数的生成
- 伪随机数
- 线性同余发生器
- 更复杂的随机数发生器*
- 用Excel产生均匀分布随机数
- 简单应用实例

2 一般的随机变量及随机数生成
- 离散与连续随机变量
- 常用的分布
- 一般随机数的生成

3 输入建模

- 对于任意的随机变量 X, 我们定义累积分布函数 (cumulative distribution function, CDF) $F(x)$ 为

$$F(x) := \mathbb{P}(X \leq x), \text{ 对于任意的 } x \in \mathbb{R}.$$

- $F(x)$ 具有如下性质:
  - $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to +\infty} F(x) = 1$;
  - $F(x)$ 为 x 的非减函数;
  - $F(x)$ 为右连续, 即, 对任意 $x_0 \in \mathbb{R}$,

$$\lim_{x \downarrow x_0} F(x) = F(x_0).$$

- 离散随机变量: 可能的取值是离散的.
- 若 X 为离散随机变量, 我们定义概率质量函数 (probability mass function, pmf) $p(x)$ 为

$$p(x) := \mathbb{P}(X = x), \text{ 对于任意的 } x \in \mathbb{R}.$$

- $p(x)$ 具有如下性质:
  - $p(x) \geq 0$, 对于任意的 $x \in \mathbb{R}$;
  - $\sum_{x \in \mathbb{R}} p(x) = 1$.
- 易知, $F(x) = \sum_{y \in (-\infty, x]} p(y)$, 对于任意的 $x \in \mathbb{R}$.

- 简单例子: 假设 X 是一个离散随机变量, 它的可能取值为 0, 1 和 2, 相应的概率为 0.5, 0.3 和 0.2.
- X 的 pmf 和 CDF 如下:

$$p(x) = \begin{cases} 0.5, & x = 0, \\ 0.3, & x = 1, \\ 0.2, & x = 2, \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0, \\ 0.5, & 0 \le x < 1, \\ 0.8, & 1 \le x < 2, \\ 1, & 2 \le x. \end{cases}$$

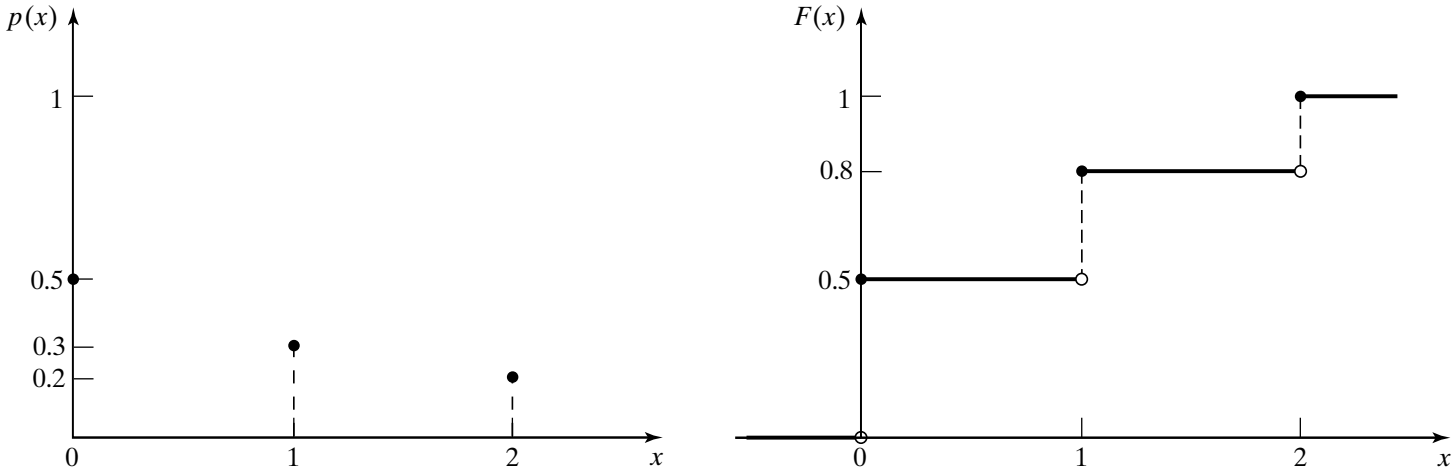


图: X 的 pmf 和 CDF 图像

- 连续随机变量: 可能的取值是连续的.
- 若 X 为连续随机变量, 我们无法定义 pmf, 因为对于任意的 $x \in \mathbb{R}$, $\mathbb{P}(X = x) = 0$.
- 我们定义概率密度函数 (probability density function, pdf) $f(x)$, 使其满足

$$\mathbb{P}(a \le X \le b) = \int_a^b f(x)\mathrm{d}x, \text{ 对于任意的 } a, b \in \mathbb{R} \text{ 且 } a < b.$$

- $f(x)$ 具有如下性质:
  - $f(x) \ge 0$, 对于任意的 $x \in \mathbb{R}$;
  - $\int_{-\infty}^{+\infty} f(t)\mathrm{d}t = 1$.
- 易知, 对于任意的 $x \in \mathbb{R}$, $F(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$, 且 $\frac{\mathrm{d}}{\mathrm{d}x}F(x) = f(x)$.

- 简单例子: 假设连续随机变量 $X \sim \text{uniform}(a, b)$, 那么它的 pdf 和 CDF 为

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{其他}, \end{cases} \qquad F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \le x \le b, \\ 1, & b < x. \end{cases}$$

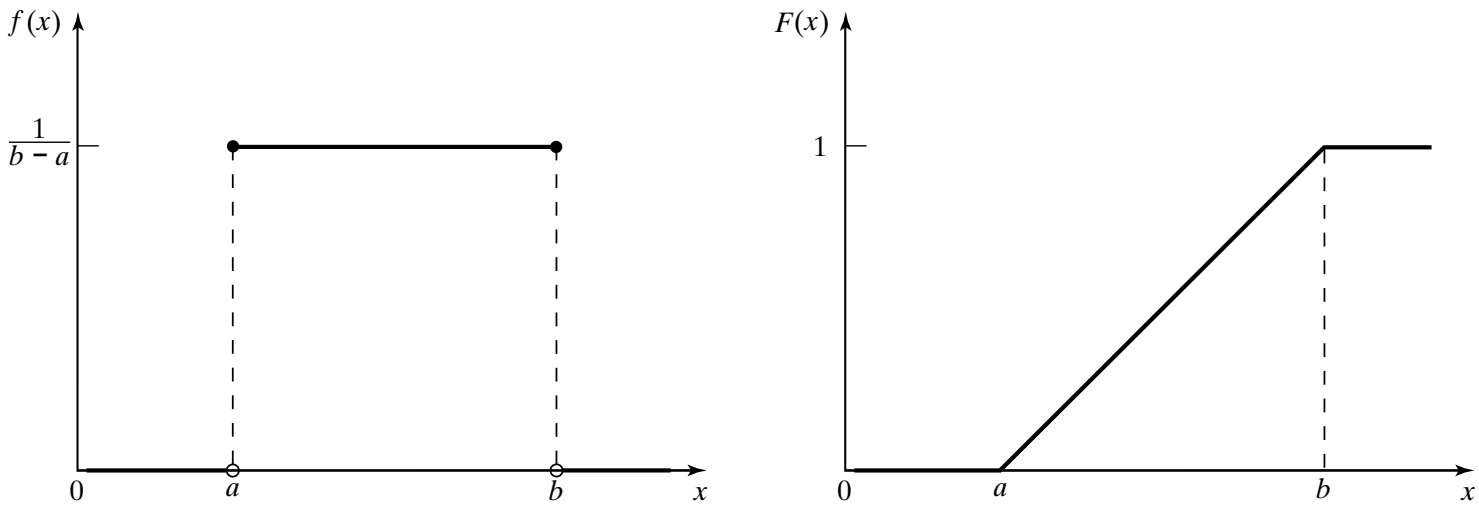


图: X 的 pdf 和 CDF 图像

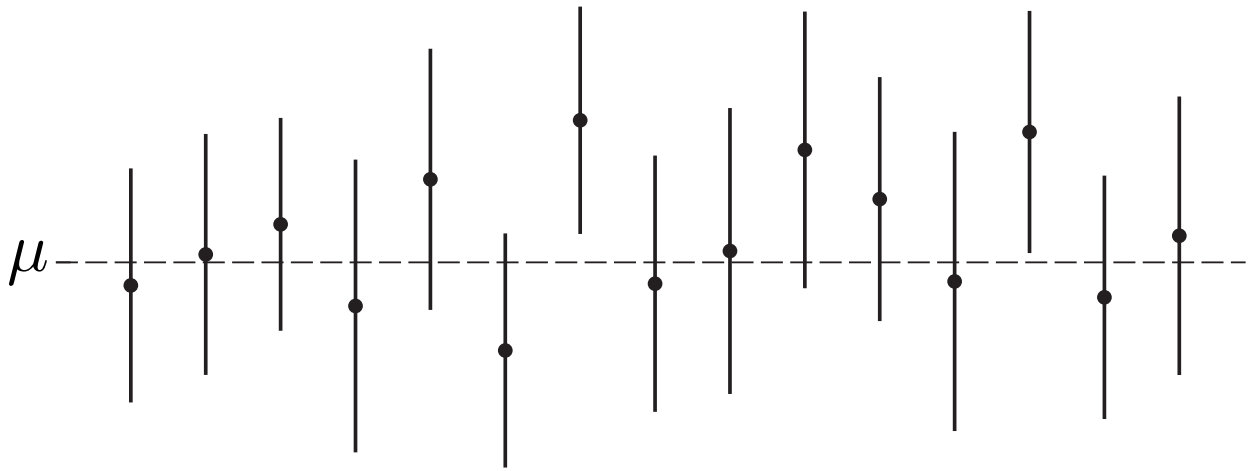
- 随机变量 X 的期望 (又称为均值), 记为 $\mathbb{E}[X]$, 有时也简记为 μ.
  - 离散: $\mathbb{E}[X] = \sum_{x\in\mathbb{R}} xp(x)$, $\mathbb{E}[h(X)] = \sum_{x\in\mathbb{R}} h(x)p(x)$;
  - 连续: $\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x$, $\mathbb{E}[h(X)] = \int_{-\infty}^{+\infty} h(x)f(x)\mathrm{d}x$.
- 随机变量 X 的方差, 记为 $\mathrm{Var}(X)$, 有时也简记为 σ^2:

$$\mathrm{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

- 随机变量 X 与 Y 之间的**线性**关联:
  - **协方差**:
    $\mathrm{Cov}(X,Y) := \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\,\mathbb{E}[Y]$.
  - **相关系数**: $\rho(X,Y) := \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$.
  - $\rho(X,Y) = 0 \iff \mathrm{Cov}(X,Y) = 0$.
- 一般地, X 与 Y 统计上独立 $\rightleftharpoons$ $\rho(X,Y) = 0$.

- 若 $x_1, x_2, \ldots, x_n$ 为随机变量 X 的一组独立观测值，即，该分布下的一组随机数，则也称它们为 X 的一组样本.
  - 该样本的样本均值为 $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$.
  - 该样本的样本方差为 $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- 注意到，当样本点 $x_1, x_2, \ldots, x_n$ 被观测到之前，它们也是随机的，都服从 X 的分布，并且相互独立.
- 记观测之前的样本点为 $X_1, X_2, \ldots, X_n$，则它们是 X 的一组 (尚未观测的) 随机样本.
  - 该随机样本的样本均值为 $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - 该随机样本的样本方差为 $S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- 可知，$\bar{X}$ 和 S^2 也是随机变量，服从一定的分布.
  - $\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \mu$, $\mathbb{E}[S^2] = \mathrm{Var}(X) = \sigma^2$.
  - $\mathrm{Var}(\bar{X}) = \sigma^2/n$.
  - 当 n 很大时，$\bar{X}$ 的随机性减弱，最终趋向常数 μ. (大数定律)

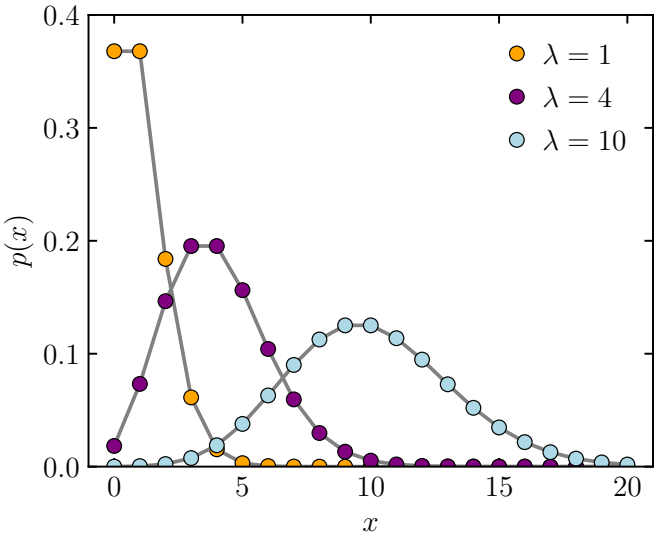
- 点估计: 用 $\bar{X}$ 来估计 μ.
- 区间估计: 用 $[\bar{X} - H, \bar{X} + H]$ 来估计 μ, 确保 $100(1-\alpha)\%$ 置信水平 (confidence level).

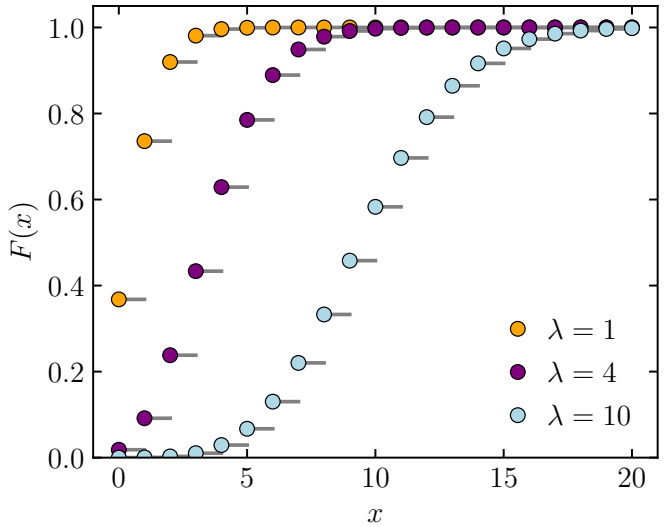


- 试验一下! http://www.rossmanchance.com/applets/ConfSim.html

- 泊松 (Poisson) 分布, 常被用来建模给定时间段内某事件发生的次数, 如
  - 电话客服系统每分钟收到的呼叫次数;
  - 每小时到达公共汽车站的乘客数.
- 记 $X \sim \text{Poisson}(\lambda)$, 其中 $\lambda > 0$, 如果

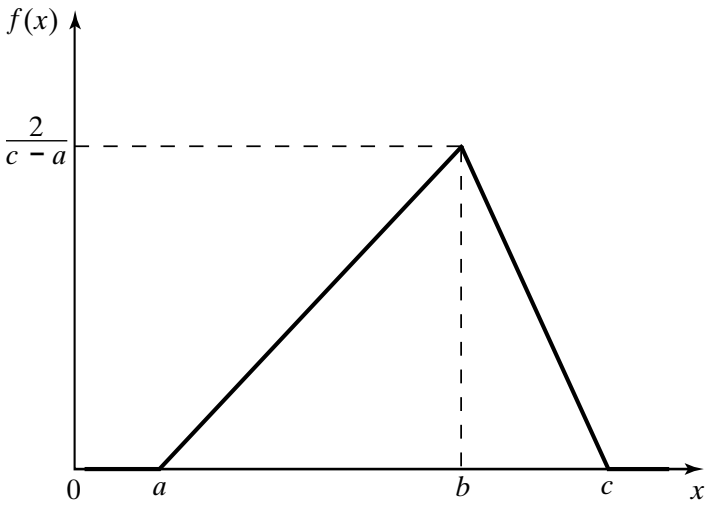
$$p(x) = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots.$$

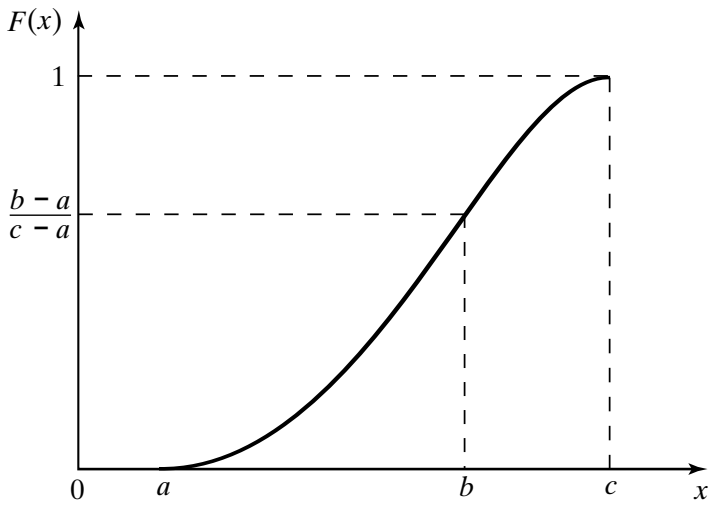




- $\mathbb{E}[X] = \lambda$, $\text{Var}(X) = \lambda$.

- 三角 (triangular) 分布, 是工程中常用的分布, 它用于当只知道某个随机变量的“最小值”、“最大值”和“ 最可能值”时.
- 若 X 服从三角分布, 且最小值为 a, 最大值为 c, 最可能值为 b, 记为 $\text{triangular}(a, b, c)$, 则它的 pdf 和 CDF 如下图所示.

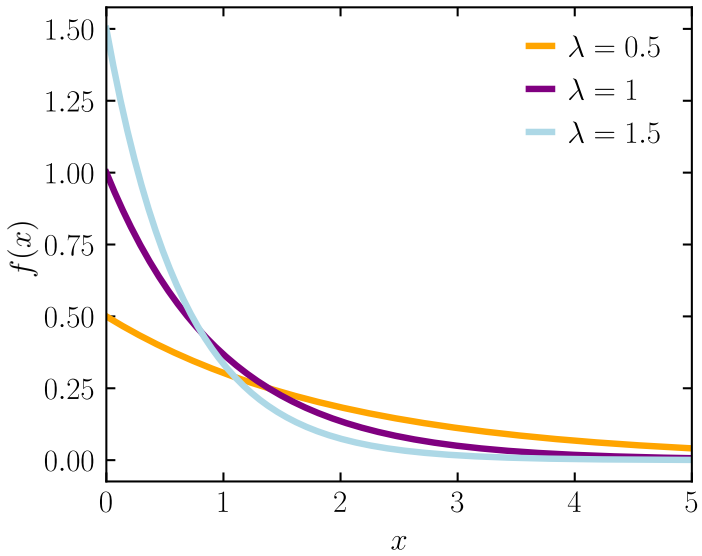


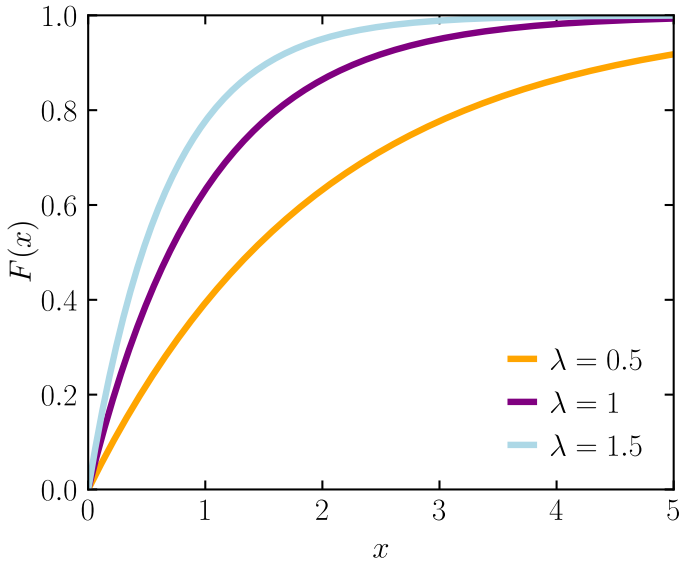


- $\mathbb{E}[X] = \frac{a+b+c}{3}$, $\text{Var}(X) = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$.

- 指数 (exponential) 分布, 常用于建模独立事件之间的时间间隔, 或者一个“**无记忆的**”过程的时间长度.
- 记 $X \sim \text{exponential}(\lambda)$, 其中 $\lambda > 0$, 如果

$$f(x) = \lambda e^{-\lambda x}, \quad F(x) = 1 - e^{-\lambda x}, \quad x \in [0, \infty).$$

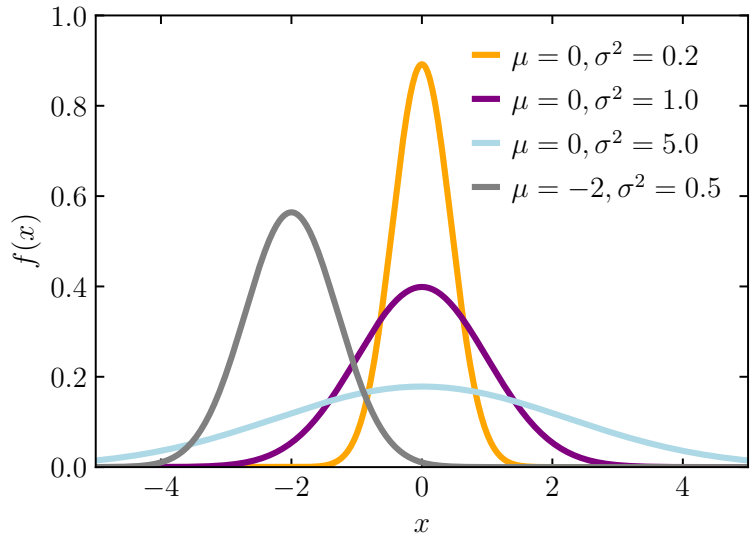


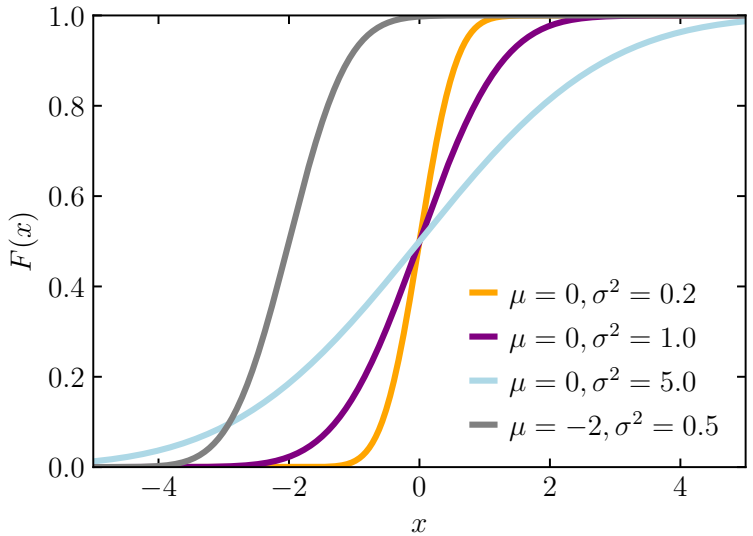


- $\mathbb{E}[X] = 1/\lambda$, $\text{Var}(X) = 1/\lambda^2$.
- 无记忆性: $\mathbb{P}(X > s | X > t) = \mathbb{P}(X > s - t)$.

- 正态 (normal) 分布, 也叫高斯分布, 是统计中**最重要**的一个分布.
- 记 $X \sim \mathcal{N}(\mu, \sigma^2)$, 其中 $\sigma > 0$, 如果

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$





- $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$.

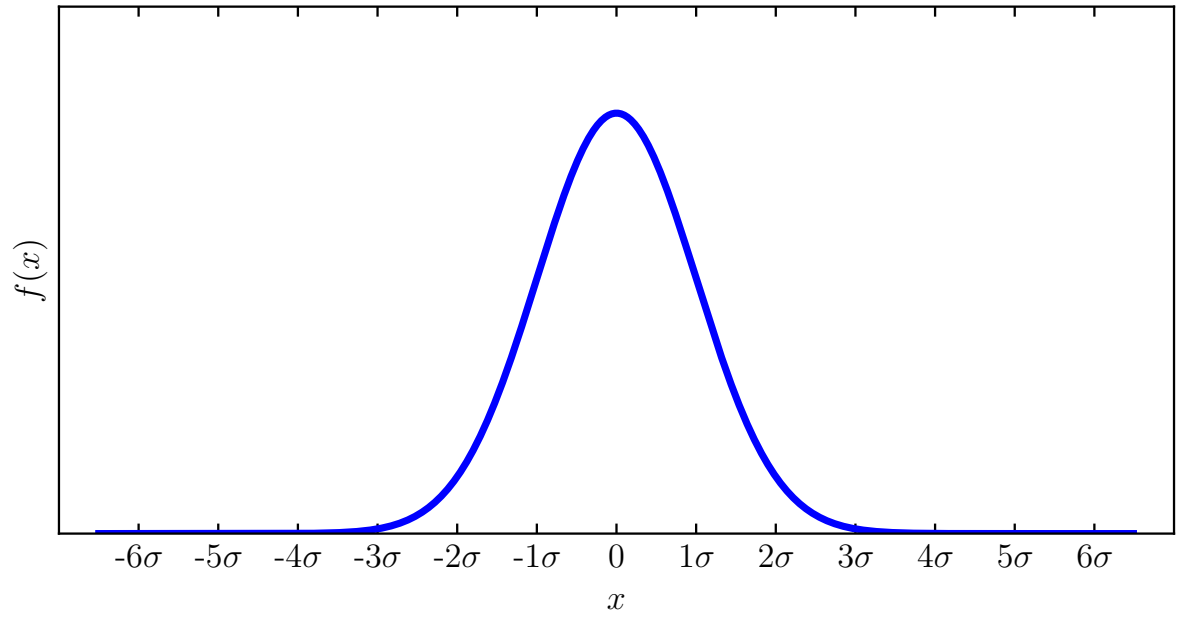
- $\mathcal{N}(0,1)$ 称为标准正态分布.
- 如果 $X \sim \mathcal{N}(\mu, \sigma^2)$, 则 $(X-\mu)/\sigma \sim \mathcal{N}(0,1)$.
- 如果 $Z \sim \mathcal{N}(0,1)$, 则 $\mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.
- 正态分布的重要性来自于**中心极限定理**!
- 假设 $X_1, \ldots, X_n \sim$ 任意一个分布, 且它们相互独立. 记

$$\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i.$$

那么中心极限定理告诉我们, 当 n 很大时,

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\text{近似}}{\sim} \mathcal{N}(0,1), \quad \bar{X} \overset{\text{近似}}{\sim} \mathcal{N}(\mu, \sigma^2/n).$$

- 正态分布下的 6σ



在 $\pm 1\sigma$ 之间	68.27%	317300 PPM (parts per million) 落在外面
在 $\pm 2\sigma$ 之间	95.45%	45500 PPM
在 $\pm 3\sigma$ 之间	99.73%	2700 PPM
在 $\pm 4\sigma$ 之间	99.9937%	63 PPM
在 $\pm 5\sigma$ 之间	99.999943%	0.57 PPM
在 $\pm 6\sigma$ 之间	99.9999998%	0.002 PPM

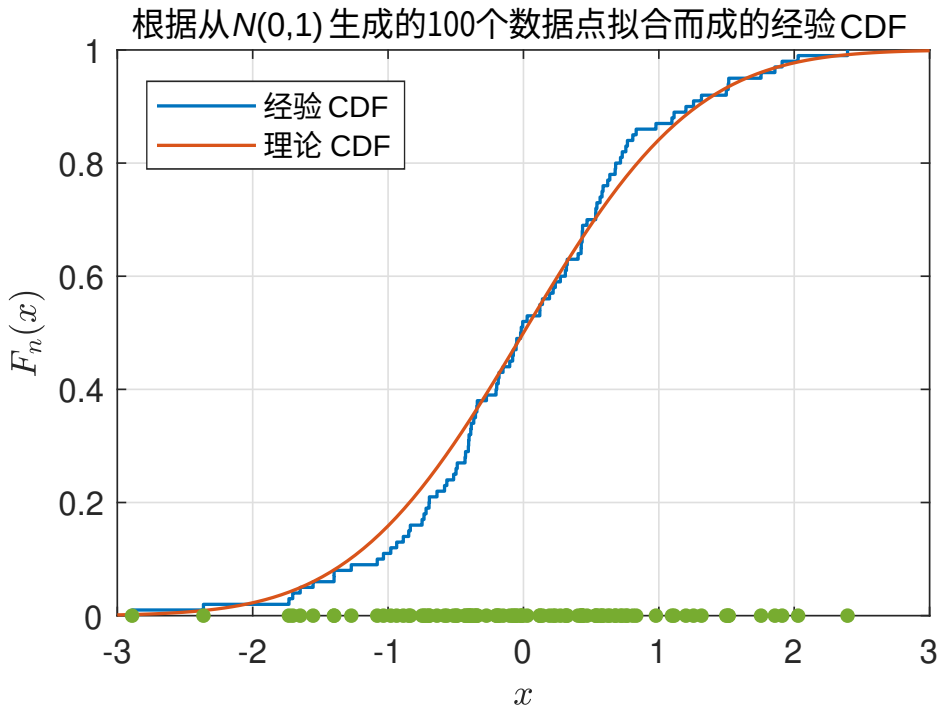
- 六西格玛 (σ) DPMO（defects per million opportunity）

西格玛水平	DPMO	废品率	合格率
1	691,462	69.77%	30.23%
2	308,538	30.88%	69.12%
3	66,807	6.68%	93.32%
4	6,210	0.62%	99.38%
5	233	0.023%	99.977%
6	3.4	0.00034%	99.99966%

- 3.4 DPMO vs 0.002 PPM? (Reason: 1.5 σ shift.)

- 经验分布 (empirical distribution), 常用于当理论分布均不适用的时候; 它的 CDF 为

$$F(x) = \frac{n \text{ 个点中小于或等于 } x \text{ 的点的数量}}{n}.$$



- 经验分布是离散的, 它的 CDF 是一个阶梯状的函数.

- 假设我们已经有了优良的 $\mathrm{uniform}(0,1)$ 随机数发生器, 即, 我们可以有一个序列的 $\mathrm{uniform}(0,1)$ 随机数.
- 如何生成一个给定的一般的分布 (如, 泊松、三角、指数、正态等) 下的随机数?
- 常用的技术
  - 逆变换法 (一种通用的方法)
  - 接受–拒绝法 (一种通用的方法)
  - 其他为某些分布专门设计的方法 (如, 用 Box–Muller 法生成 $\mathcal{N}(0,1)$ 随机数)

- 逆变换法 (Inverse-Transform Technique)

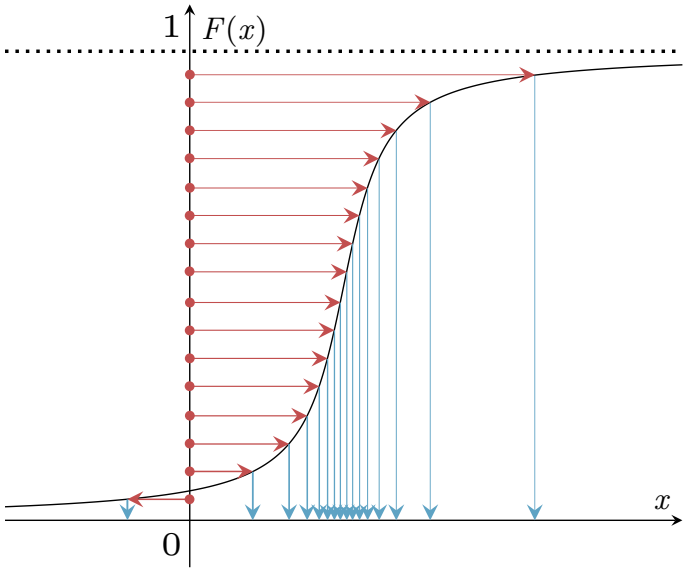


图: 连续随机变量

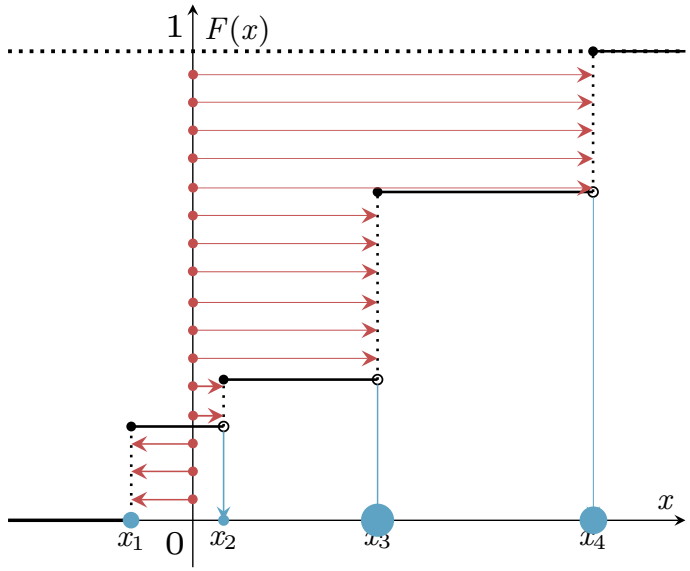


图: 离散随机变量

- 步骤
  1. 生成所需数量的 $\mathrm{uniform}(0,1)$ 随机数 (于纵坐标).
  2. 反向映射至横坐标, 所得的点即为采样自 $F(x)$ 的随机数.

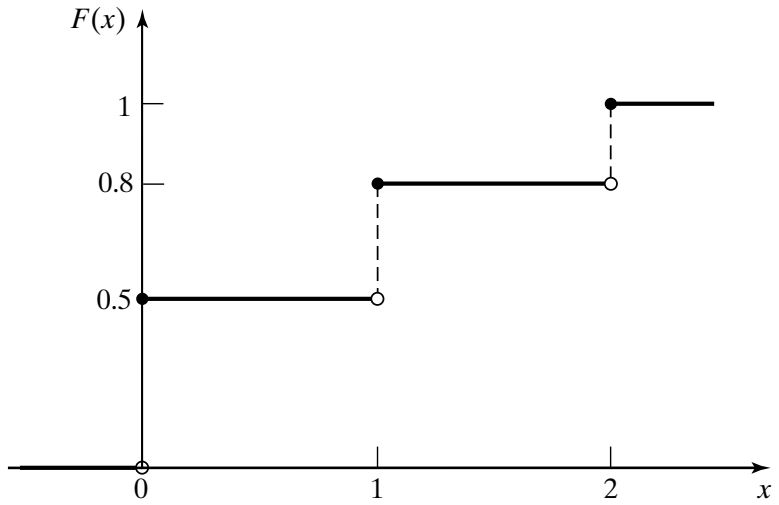
- 当一个随机变量或分布的 CDF 函数的反函数可以被**解析地求解**或者**容易地计算**时，逆变换法是一种很有用的方法.
- 它可被用于从许多连续分布中生成随机数，如
  - 均匀 (uniform) 分布
  - 指数 (exponential) 分布
  - 三角 (triangular) 分布
  - 威布尔 (Weibull) 分布
  - 柯西 (Cauchy) 分布
  - 帕累托 (Pareto) 分布
- 从原则上说，它可被用于从任意的离散分布中生成随机数，如
  - 离散均匀 (discrete uniform) 分布
  - 泊松 (Poisson) 分布
  - 任意的经验 (empirical) 分布

- 简单例子: 假设离散随机变量 X 具有如下的 pmf 和 CDF:

$$p(x) = \begin{cases} 0.5, & x = 0, \\ 0.3, & x = 1, \\ 0.2, & x = 2, \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0, \\ 0.5, & 0 \le x < 1, \\ 0.8, & 1 \le x < 2, \\ 1, & 2 \le x. \end{cases}$$

如何生成满足该分布的随机数?

- 求解 $F(x)$ 的反函数, $F^{-1}(y)$: (注: $F^{-1}(y) := \min\{x : F(x) \ge y\}$.)



$$F^{-1}(y) = \begin{cases} 0, & 0 < y \le 0.5, \\ 1, & 0.5 < y \le 0.8, \\ 2, & 0.8 < y < 1. \end{cases}$$

与最开始的直觉一致!

- 简单例子: 生成 $\mathrm{uniform}(a,b)$ 随机数.
- 直觉: 先生成 $\mathrm{uniform}(0,1)$ 随机数 u, 然后输出 $x=a+(b-a)u$, 即为所需随机数.
- 已知 $\mathrm{uniform}(a,b)$ 随机变量的 pdf 和 CDF 如下:

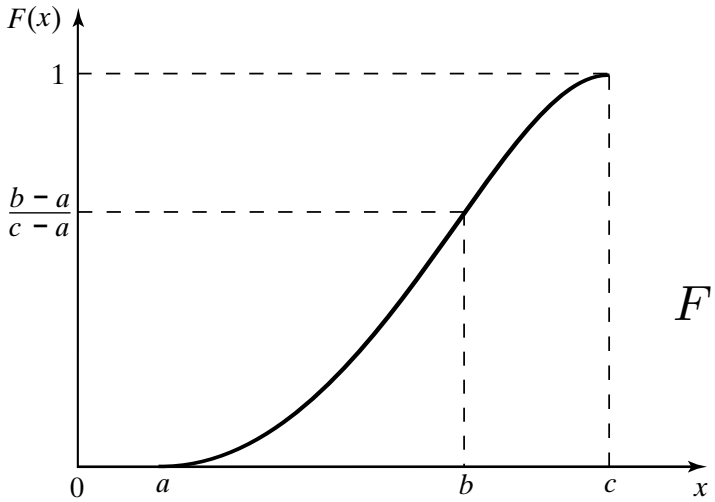
$$f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b,\\ 0, & \text{其他},\end{cases}\qquad F(x)=\begin{cases}0, & x<a,\\ \frac{x-a}{b-a}, & a\le x\le b,\\ 1, & b<x.\end{cases}$$

- 求解 $F(x)$ 的反函数:

$$F^{-1}(y)=a+(b-a)y,\quad 0<y<1.$$

- 与最开始的直觉一致!

- 例子: 生成 $\text{triangular}(a, b, c)$ 随机数.
- 已知 $\text{triangular}(a, b, c)$ 随机变量的 CDF 图像如下图所示



$$F(x) = \begin{cases} 0, & x < a, \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, & a \le x < b, \\ 1 - \dfrac{(c-x)^2}{(c-a)(c-b)}, & b \le x < c, \\ 1, & c \le x. \end{cases}$$

- 求解 $F(x)$ 的反函数:

$$F^{-1}(y) = \begin{cases} a + \sqrt{(b-a)(c-a)}\sqrt{y}, & 0 < y < \frac{b-a}{c-a}, \\ c - \sqrt{(c-b)(c-a)}\sqrt{1-y}, & \frac{b-a}{c-a} \le y < 1. \end{cases}$$

- 在 Excel 中实施.

- 例子: 生成 exponential(λ) 随机数.
- 已知 exponential(λ) 随机变量的 pdf 和 CDF 如下:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \quad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

- 求解 $F(x)$ 的反函数:

$$F^{-1}(y) = -\frac{1}{\lambda} \ln(1-y), \quad 0 < y < 1.$$

- 注: 若 $U \sim \text{uniform}(0,1) \Longrightarrow 1 - U \sim \text{uniform}(0,1)$, 因此就生成随机数而言, 只需计算 $-\frac{1}{\lambda} \ln(y)$ 即可.
- 在 Excel 中实施.

- 例子: 生成 Poisson(λ) 随机数.
- 已知 Poisson(λ) 随机变量的 pmf 和 CDF 如下:

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

$$F(x) = \mathbb{P}(X \le x) = \sum_{i=0}^{x} p(i) = \sum_{i=0}^{x} \frac{e^{-\lambda}\lambda^i}{i!}, \quad x = 0, 1, 2, \ldots.$$

- 尽管 $F(x)$ 的反函数无法写成解析形式, 但由于其离散性, 可以较简单地通过反函数的定义计算:

$$F^{-1}(y) = \min\{x : F(x) \ge y\}, \quad 0 < y < 1.$$

- 在 Excel 中, 可以用

`POISSON.DIST(x,mean,cumulative)`

分别计算 $p(x)$ (`cumulative` 为 `FALSE`) 和 $F(x)$ (`cumulative` 为 `TRUE`).

- 经验分布随机数也可这样生成, 此时 $F(x)$ 是根据数据得出.

- 当一个随机变量或分布的 CDF 函数的反函数无解析形式，且数值计算比较复杂时，逆变换法的效率便会下降.
- 例如，生成 $\mathcal{N}(\mu, \sigma^2)$ 随机数. 已知 $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 以及 $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$，且 $F(x)$ 的反函数无解析形式，只能通过复杂的数值计算得到 $F^{-1}(y)$.
  - 在 Excel 中，可以用

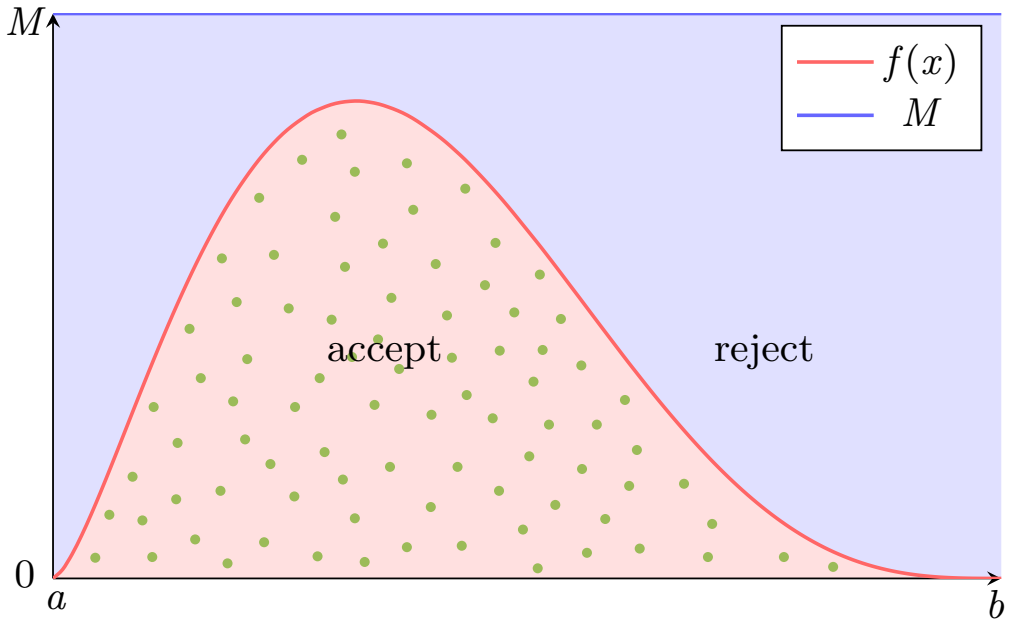
    `NORM.DIST(x,mean,standard_dev,cumulative)`

    分别计算 $f(x)$ (`cumulative` 为 `FALSE`) 和 $F(x)$ (`cumulative` 为 `TRUE`).
  - 在 Excel 中，还可以用

    `NORM.INV(probability,mean,standard_dev)`

    计算 $F^{-1}(y)$.
- 除了逆变换法之外，还有其他的随机数生成的方法可以考虑，如，接受–拒绝法 (Acceptance-Rejection Technique).

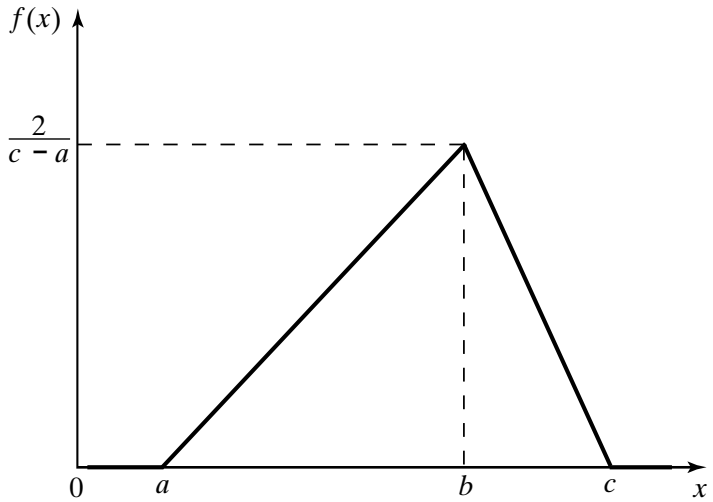
- 若我们想产生某个分布下的随机数，已知它的概率密度函数 $f(x)$ 只在 $x \in [a, b]$ 时为正值，且 $f(x) \le M$.



图：有界的定义域 (*original image from ZHANG Xiaowei*)

1. 生成 $\text{uniform}\{(y, z) : a \le y \le b,\ 0 \le z \le M\}$ 随机数对 $(y_1, z_1), (y_2, z_2), \ldots$.
   - y_i 来自 $\text{uniform}(a, b)$, z_i 来自 $\text{uniform}(0, M)$.
2. 如果 $z_i < f(y_i)$，接受这个随机数对，并且输出 y_i.

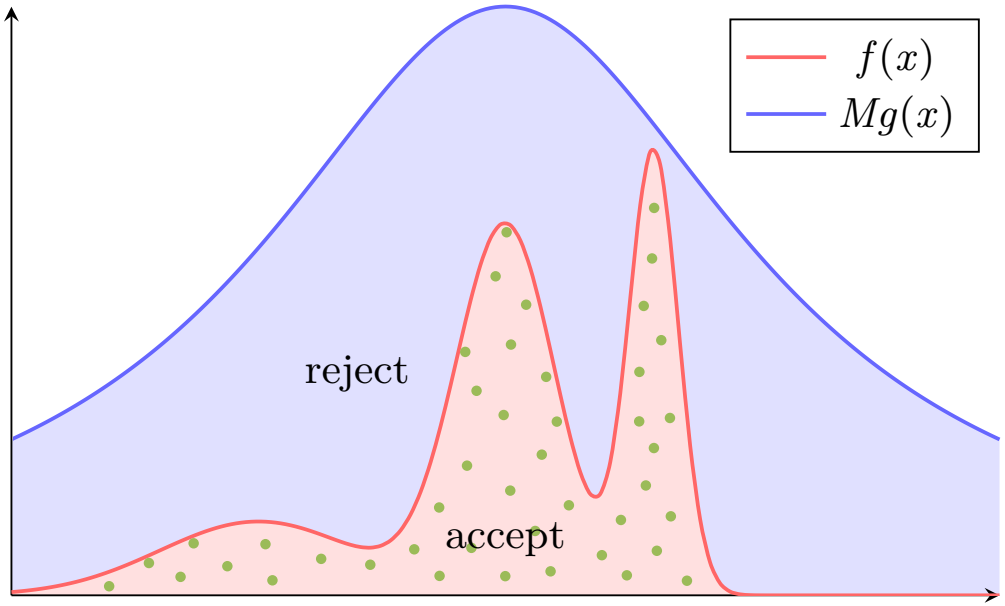
- 例子: 生成 $\text{Triangular}(a, b, c)$ 随机数.
- 已知 $\text{Triangular}(a, b, c)$ 随机变量的 pdf 图像如下图所示



1. 生成随机数对 (y, z), 其中 y 来自 $\text{uniform}(a, c)$, z 来自 $\text{uniform}(0, 2/(c-a))$.
2. 如果 $z < f(y)$, 接受这个随机数对, 并且输出 y; 否则回到第 1 步.

- 注意: 为了生成 1 个 $\text{Triangular}(a, b, c)$ 随机数, 需要生成多个其他分布的随机数!

- 若我们想产生某个分布下的随机数，已知它的概率密度函数 $f(x)$ 具有上界 $Mg(x)$，其中 $g(x)$ 是另一个概率密度函数.

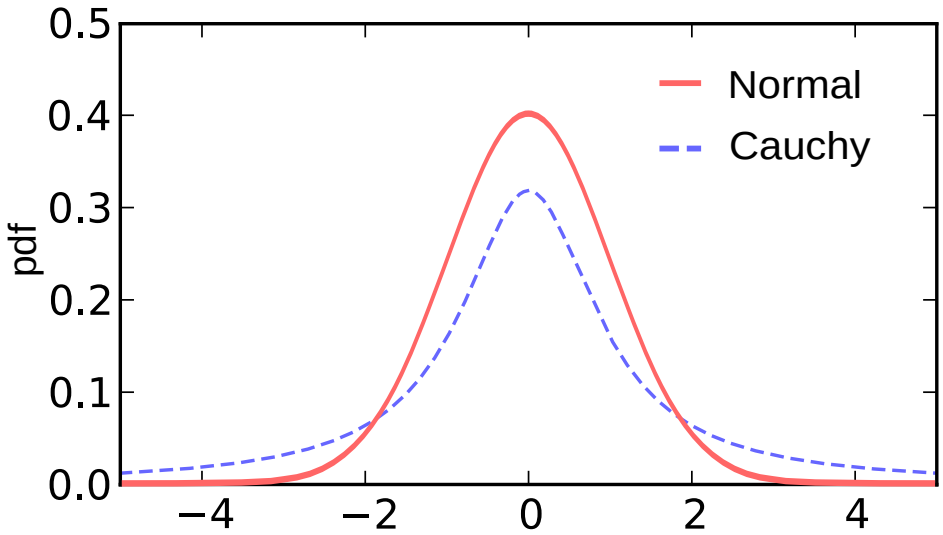


图：无界的定义域 (*original image from ZHANG Xiaowei*)

1. 生成 $\text{uniform}\{(y,z) : y \in g(\cdot)$ 定义域，$0 \le z \le Mg(y)\}$ 随机数对 $(y_1,z_1), (y_2,z_2), \ldots$.
   - y_i 来自 $Y \sim g(\cdot)$，z_i 来自 $Z \sim \text{uniform}(0, Mg(y_i))$. **(why?)**
2. 如果 $z_i < f(y_i)$，接受这个随机数对，并且输出 y_i.

- 例子: 生成 $\mathcal{N}(0,1)$ 随机数.
  - 不难发现, 若 x 为 $\mathcal{N}(0,1)$ 随机数, 则 $\mu+\sigma x$ 为 $\mathcal{N}(\mu,\sigma^2)$ 随机数.

- 采用 $\mathrm{Cauchy}(0)$ 的概率密度函数作为辅助, 该函数形式为 $g(x)=\frac{1}{\pi(1+x^2)}$, $x\in(-\infty,\infty)$.



- 需要使 $M\geq\sqrt{\frac{2\pi}{e}}$ 才可以.

- 使用 Box–Muller 法生成 $\mathcal{N}(0,1)$ 随机数.
  1. 生成独立 $\mathrm{uniform}(0,1)$ 随机数 u_1 和 u_2.
  2. 令 $z_1 = \sqrt{-2\ln u_1}\cos(2\pi u_2)$ 及 $z_2 = \sqrt{-2\ln u_1}\sin(2\pi u_2)$.
- 可以证明, z_1 和 z_2 都是 $\mathcal{N}(0,1)$ 随机数, 且它们相互独立.

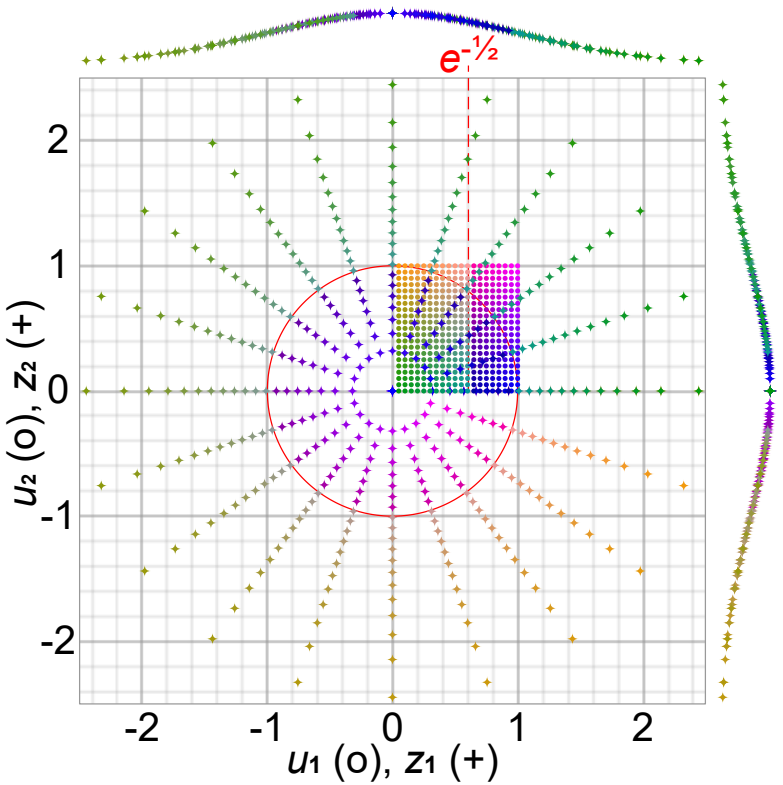


图: Box–Muller 法的可视化 (image by Cmglee / CC BY 3.0) 可交互图像

1. 随机数的生成
   - 伪随机数
   - 线性同余发生器
   - 更复杂的随机数发生器*
   - 用Excel产生均匀分布随机数
   - 简单应用实例
2. 一般的随机变量及随机数生成
   - 离散与连续随机变量
   - 常用的分布
   - 一般随机数的生成
3. 输入建模

输入建模

- 假设我们已经可以生成任意分布下的随机数，对于一个实际问题，如何决定要使用什么分布？**输入建模**
- 输入模型是仿真的驱动力.
  - 排队系统：到达时间间隔、服务时长的分布.
  - 供应链：需求、交货时长的分布.
  - 财务风险管理：资产收益的分布.
- 仿真模型的输出的质量受限于输入的质量.
  - "*Garbage in, garbage out.*"
- "All models are wrong, but some are useful." – George Box.
  - 对于随机输入，没有严格意义上的正确的输入模型.
  - 我们能做的是选择一个适当的输入模型，以获得合理的有用的结果.

输入建模

- 输入模型的一些基本要求:
  - 可以刻画系统的物理性质;
  - 可以容易地调节使其适应当前的情况;
  - 可以高效地生成相应的随机数.
- 输入建模是一种工程技术, 有时候更是一种技艺.
  - 它几乎总是要求分析师运用统计的工具, 连同他的判断.
  - 因为没有"正确的"模型, 所以一个明智的做法是将那些似乎合理的输入模型都在仿真模型中运行一下, 看看得到的结论是否对这个选择很敏感.

输入建模

- 输入建模的基本步骤.
  1. 从实际系统中采集数据.
  2. 选定一个概率分布族来刻画得到的数据.
     - 基于系统或过程的物理特性.
     - 基于对数据的图形化的审查 (如, 通过频率直方图审查分布的形状).
  3. 将选定的分布拟合至数据 (即, 确定分布中的参数值).
     - 矩估计 (method of moments, MoM).
     - 极大似然估计 (maximum likelihood estimation, MLE).
  4. 评估选定的分布及参数的拟合优良度.
     - 图像法: 频率直方图, 分位图 (quantile-quantile plot, Q-Q plot).
     - 统计学检验: 卡方检验 (chi-square test, χ^2 test), 柯尔莫哥洛夫–斯米尔诺夫检验 (Kolmogorov-Smirnov test, K-S test).
  5. 如果觉得拟合得不好, 选择另一个概率分布族, 回到第 3 步; 或者直接使用经验分布.

输入建模

- 许多软件都有一个“最优拟合”的选项 (或者按钮).
  - 它会从它的库中为你推荐“最好的”分布, 往往基于一些概括性的度量 (例如, p-值), 以及其他一些可能的因素 (例如, 离散或连续、有界无界).
- 当使用这种功能的时候, 需要谨记以下几点:
  - 软件可能对数据背后的物理基础一无所知.
  - 自动化的寻找最优拟合的程序, 往往倾向于那些函数形式更加灵活的分布族.
  - 但是, 最大限度地与数据接近未必一定能得出最合适的输入模型 (因为可能存在过拟合).
  - 那些概括性的度量 (例如, p-值), 有其局限性.
  - 将自动化的分布选择视为一个建议, 再使用图像法加以验证, 最终做出自己的选择.