概率论与数理统计基础知识

刘闯 (武汉大学计算机学院)

1659608216@qq.com

2019 年 10 月 4 日

摘要

《概率论与数理统计》齐民友, 武汉大学

《Probability and Statistics》Morris H. DeGroot , Carnegie Mellon University

《应用数理统计》邰淑彩，武汉大学

《数学物理方法》姚端正，武汉大学

目录

1 概率论基本知识

1.1 随机事件

自然界中存在两种现象：确定性现象和随机现象。对随机现象的观察，记录，实验统称为随机试验。具有几个特性：

- 可以重复进行
- 事先知道可能出现的所有的结果
- 实验前不知道那个结果会发生

一些定义：

- **样本空间**：随机事件的所有可能结果
- **基本事件**：样本空间的每一个结果
- **随机事件**：随机实验的样本空间的一个子集

1.1.1 事件的关系

- **事件的并**：事件 A，B 至少有一个发生，$A \cup B$
- **事件的交**：事件 A，B 同时发生 $A \cap B$ 或 AB
- **事件互斥**：事件 A，B 不能同时发生 $A \cap B = \varnothing$. 如图1。（图片来源于 `https://www.zhihu.com/question/25257915`）

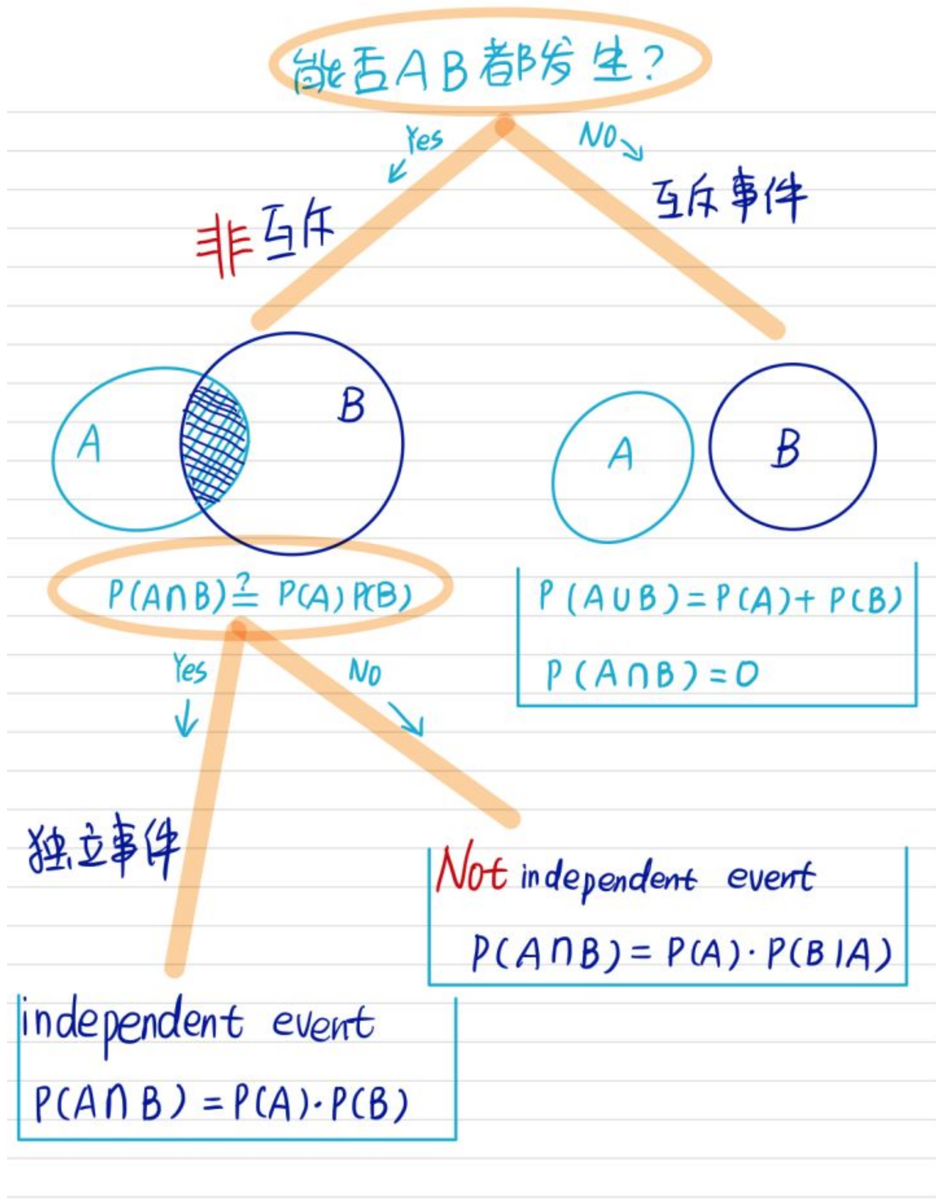


图 1: 互斥事件和独立事件

1.2 概率

1.2.1 概率的统计定义

设 E 为随机实验，A 为 E 中的事件。

相同条件下，E 重复做 n 次，A 发生的次数是 n_A ，则 $\frac{n_A}{n}$ 称为事件的频率。

当 n 很大的时候，频率稳定在一个常数 p 附近摆动，则 p 称为事件 A 的**概率**。

1.2.2 概率的公理化定义

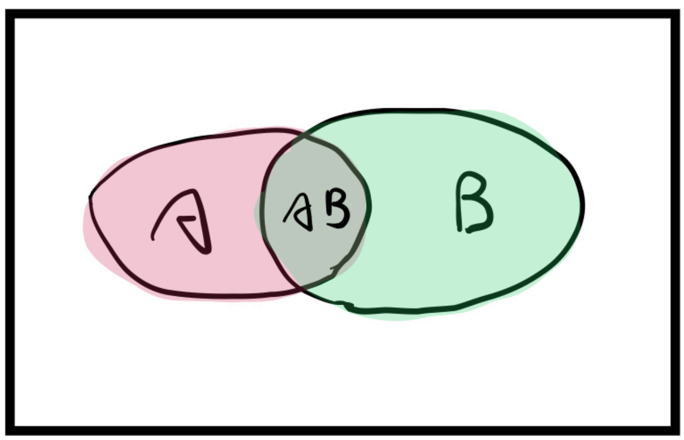
1. 对于任意事件 A , $P(A) \geq 0$, 非负性
2. $P(\Omega)=1$, 规范性
3. 对于两两互斥的事件：

$$P\left(\bigcup_{i=1}^{\infty} A_i\right)=\sum_{i=1}^{\infty} P\left(A_i\right)$$

最后一条公理：如果两事件互斥，那么这两个事件其中有一个发生的概率等于各个事件发生的（边缘）概率之和。假设我们掷出一个均匀的 6 面骰，想要知道掷出 5 点或 6 点的概率。这两个事件是互斥的，因为我们无法同时掷出 5 点和 6 点。因此掷出 5 点或 6 点的概率等于掷出 5 点的概率加上掷出 6 点的概率：$1/6+1/6=2/6=1/3$.

1.2.3 概率的性质

概率主要性质如图 2.



$$P(A \cup B) = P(A) + P(B) - P(AB)$$
$$P(A - B) = P(A) - P(AB)$$

图 2: 概率的性质

1.3 条件概率

1.3.1 定义

对于两个事件 A，B. $P(B)>0$ ，$P(A|B)$ 表示在事件 B 已经发生的条件下，A 发生的概率：

$$P(A|B)=\frac{P(AB)}{P(B)} \tag{1}$$

所以：

$$P(AB)=P(B)P(A|B) \tag{2}$$

同理：

$$P(B|A)=\frac{P(AB)}{P(A)}$$

$$P(AB)=P(A)P(B|A)$$

推广至 n 个事件：

$$\begin{aligned} P\left(A_1A_2\cdots A_n\right)=&P\left(A_1\right)P\left(A_2|A_1\right)\cdots P(A_1)\\ &P\left(A_n|A_1A_2\cdots A_{n-1}\right)\end{aligned} \tag{3}$$

1.3.2 全概率公式

如果事件 $A_1,A_2,\cdots,A_n$ 是样本空间的一个完备事件组，切 $P(A_i)>0$ 则对于任意一个事件 B：

$$P(B)=\sum_{i=1}^{n}P\left(A_i\right)P\left(B|A_i\right) \tag{4}$$

完备事件组有可以称为样本空间的一个划分。这些事件在实验中有且仅有一个发生。比如说导致一个系统（飞机，供电）发生故障的所有互不相容的原因可以构成一个划分

1.3.3 贝叶斯公式

同样的，如果事件 $A_1, A_2, \cdots, A_n$ 是样本空间的一个完备事件组，切 $P(A_i) > 0$ 则对于任意一个事件 B：

$$P\left(A_k|B\right) = \frac{P\left(A_k\right) P\left(B|A_k\right)}{\sum_{i=1}^{n} P\left(A_i\right) P\left(B|A_i\right)}, k = 1, 2, \cdots, n \tag{5}$$

贝叶斯公式可以简单的由条件概率的定义和全概率的公式得到

$$P\left(A_k|B\right) = \frac{P(BA_k)}{P(B)} = \frac{P\left(A_k\right) P\left(B|A_k\right)}{\sum_{i=1}^{n} P\left(A_i\right) P\left(B|A_i\right)}, k = 1, 2, \cdots, n \tag{6}$$

简单理解贝叶斯公式：

首先我们对条件概率公式进行简单的变形

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

我们把 P(A) 称为先验概率（Prior probability），即在 B 事件发生之前，我们对 A 事件概率的一个判断。$P(A|B)$ 称为后验概率（Posterior probability），即在 B 事件发生之后，我们对 A 事件概率的重新评估。$\frac{P(B|A)}{P(B)}$ 称为可能性函数（Likelyhood），这是一个调整因子，使得预估概率更接近真实概率。

所以条件概率可以理解为：

$$\text{后验概率} = \text{先验概率} \; x \; \text{调整因子}$$

在这里，如果可能性函数 $\frac{P(B|A)}{P(B)} > 1$，意味着先验概率被增强，事件 A 的发生的可能性变大；如果可能性函数 =1，意味着 B 事件无助于判断事件 A 的可能性；如果可能性函数 <1，意味着先验概率被削弱，事件 A 的可能性变小。

1.3.4 事件的独立性

定义： A，B 是两个事件，满足：

$$P(AB) = P(A)P(B) \tag{7}$$

则称两个事件相互独立。独立和互斥的区别参照图 1

如果 A 与 B 相互独立，那么，A 和 $\bar{B}$, $\bar{A}$ 和 B ，$\bar{A}$ 和 $\bar{B}$ 都相互独立。

2 随机变量及其概率分布

2.1 随机变量及其分布函数

1. 随机变量：因为随机实验的结果大多与数值发生自然的联系（或者人为的建立和实数的联系）

在随机实验 E 的样本空间 $\Omega = \{\omega\}$ 上定义一个实值函数 $X = X(\omega)$ 则称 X 为 **随机变量**

例如：大学生学生全体为 Ω，抽取一人为 ω ，其身高可用 $X(\omega)$ 表示。

2. 分布函数：

X 为一个随机变量，x 为任意实数，

$$F(x) = P\{X \leq x\}, -\infty < x < \infty \tag{8}$$

F(x) 即为随机变量 X 的分布函数。

3. 分布函数一个重要性质：

$$\forall a, b(a < b), P(a < X \leq b) = F(b) - F(a) \tag{9}$$

2.2 离散型随机变量及其概率分布

离散型随机变量：可能取值是有限个或者可列无穷多个

概率分布

随机变量的所有取值：$x_k(k = 1, 2...)$ 相应的各个取值的概率为

$$P\left(X = x_k\right) = p_k, k = 1, 2, \cdots$$

$$p_k \geq 0(k = 1, 2, \cdots), \sum_{k=1}^{\infty} p_k = 1$$

所以其相应的分布函数:

$$F(x)=P\{X\leq x\}=\sum_{x_k\leq x}P\left\{X=x_k\right\}=\sum_{x_k\leq x}p_k \tag{10}$$

2.2.1 常用的离散型随机变量: 二项分布

1. **二项分布**: n 重伯努利试验（试验 E 只有两种可能的结果，把 E 独立重复做 n 次）

事件 A 在任意一次试验中发生的概率为 p，事件 A 发生的次数的可能取值为 0,1,2, ...,n. X 的概率分布

$$P(X=k)=C_n^k p^k q^{n-k},(k=0,1,2,\cdots,n) \tag{11}$$

2. **泊松定理**: 上面的二项分布，当 n 很大，p 很小，$\lambda=np_n$ 大小适中的时候。

$$\lim_{n\to\infty}C_n^k p_n^k\left(1-p_n\right)^{n-k}=\frac{\lambda^k e^{-\lambda}}{k!} \tag{12}$$

也可以称为二项分布的泊松逼近 (一般要求 p < 0.1)。如果 p 较大的话，就要使用二项分布的正态逼近。

2.2.2 常用的离散型随机变量: 泊松分布

泊松分布: 随机变量 X 的分布律:

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},(k=0,1,2,\cdots) \tag{13}$$

称 X 服从参数为 λ 的泊松分布。

一般的，如果一次试验中某事件 A 发生的概率很小，则在大量的实验中事件 A 发生的次数可以近似的使用泊松分布进行描述。

理解泊松分布，详见: `https://www.zhihu.com/question/26441147`

2.3 连续型随机变量及其概率分布

很多随机变量的取值不是离散的，不可能把取值一一列出。

定义：

如果存在一个非负实值函数 f(x), 使得：

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{14}$$

则 X 为**连续型随机变量**，f(x) 为 X 的**概率分布密度函数**

基本性质

$$\begin{aligned} &(1)\ f(x) \geq 0(-\infty < x < \infty) \\ &(2)\ \int_{-\infty}^{\infty} f(t)dt = 1 \end{aligned} \tag{15}$$

某个函数满足公式 15 中的 (1)(2)，则他必定是某个概率空间的连续型随机变量的概率密度。还有其他性质见图 3.

2.3.1 常见的连续型随机变量：均匀分布

定义： 连续型随机变量的概率密度函数为：

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{o.w.} \end{cases} \tag{16}$$

X 服从 [a,b] 上的均匀分布就是指 X 在 [a,b] 中的取值是等可能性的。

概率密度和分布函数图像见图 4

2.3.2 常见的连续型随机变量：指数分布

定义： 连续型随机变量 X 的概率密度函数为：

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{17}$$

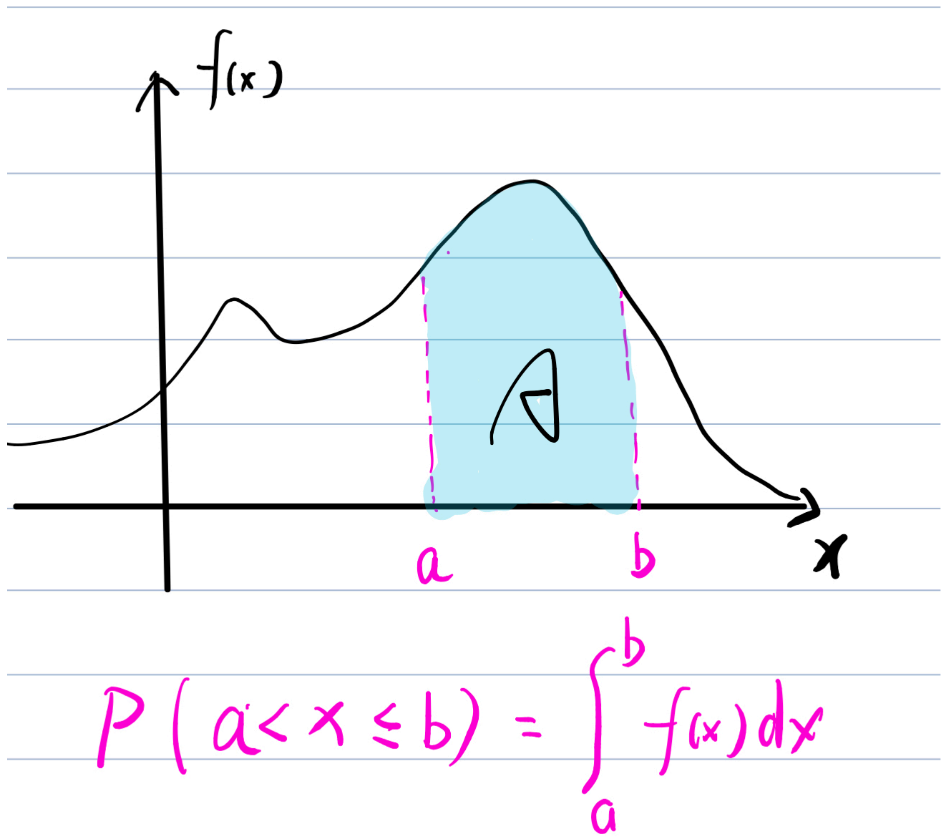


图 3: 概率密度函数性质

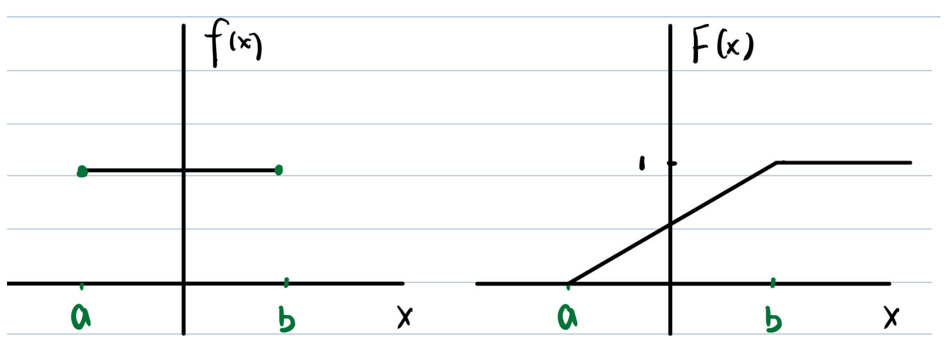
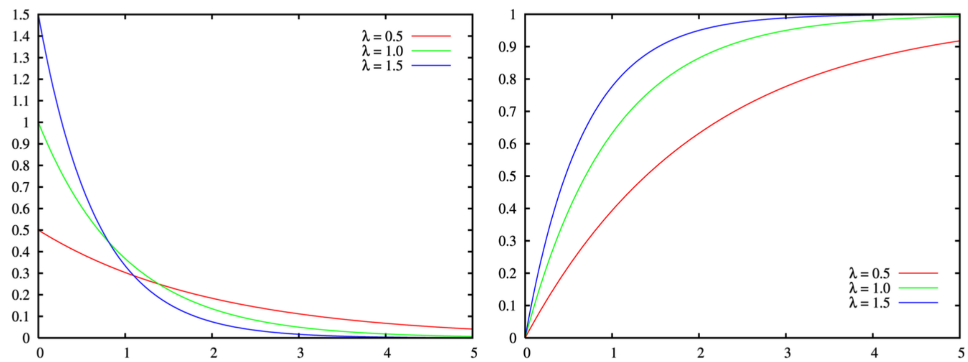


图 4: 均匀分布概率密度和分布函数

则称 X 服从参数为 λ 的指数分布。

相应的分布函数 (对 f(x) 求积分)，如图 5，(图片来源于 wiki)

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases} \tag{18}$$



(a) 指数分布概率密度　　(b) 指数分布分布函数

图 5: 指数分布

指数分布常用于各种寿命的近似估计。
指数分布的重要性表现在无记忆性

$$P(X > s + t | X > s) = P(X > t) \tag{19}$$

假设 X 为元件的工作寿命，元件已经工作了 S h 的条件下，还能工作 t 小时的概率与已经工作的时间 S 无关。

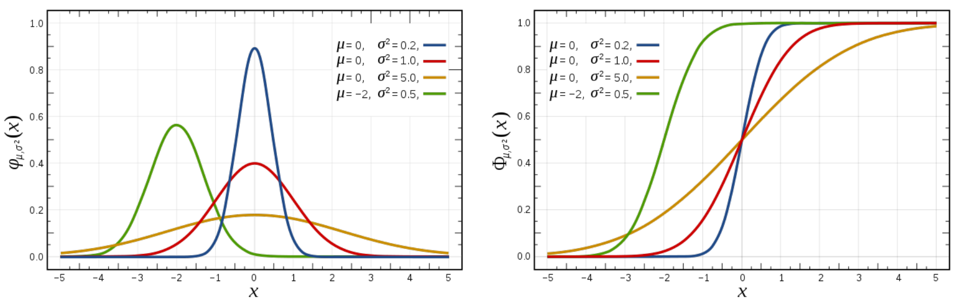
2.3.3 常见的连续型随机变量：正态分布 (高斯分布)

正态分布极其重要，许多分布都可以用正态分布逼近
定义： 连续型随机变量 X 的概率密度函数为：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (-\infty < x < \infty) \tag{20}$$

称 X 服从参数为，μ, σ 的**正态分布**，记为：$X \sim N\left(\mu, \sigma^2\right)$。
相应的**分布函数**

$$F(x)=\int_{-\infty}^{x} \frac{1}{\sqrt{2 \pi} \sigma} e^{-\frac{(t-\mu)^2}{2 \sigma^2}} d t \tag{21}$$



(a) 正态分布概率密度　　(b) 正态分布分布函数

图 6: 正态分布

如果 $\mu=0$ 并且 $\sigma=1$，这个分布被称为标准正态分布

$$f(x)=\frac{1}{\sqrt{2 \pi}} \exp \left(-\frac{x^2}{2}\right) \tag{22}$$

2.4 随机变量函数的分布

主要学习连续型：

已知连续型随机变量 X 的概率密度 $f_X(x)$, 求随机变量 $Y=g(X)$ 的概率密度 $f_Y(y)$

Y 的分布函数:

$$F_Y(y)=P(g(X)<=y)$$

,

$$f_Y(y)=F_Y^{'}(y), \quad F_Y^{'}(y)$$

我们这里可以得到一个小结论：

正态随机变量的线性函数仍然为正态随机变量

3 多维随机变量及其概率分布

3.1 二维随机变量及其联合分布函数

X, Y 为同一样本空间 Ω 上的随机变量，则由他们构成的向量 (X,Y) 称为**二维随机向量**或者**二维随机变量**。

$$F(x,y) = P\{X \le x, Y \le y\} \tag{23}$$

上式称为二维随机变量的**联合分布函数**

如图 7所示，联合分布函数 $F(x,y)$ 表示的就是随机点 (X,Y) 落 (x,y) 左下方的区域的概率

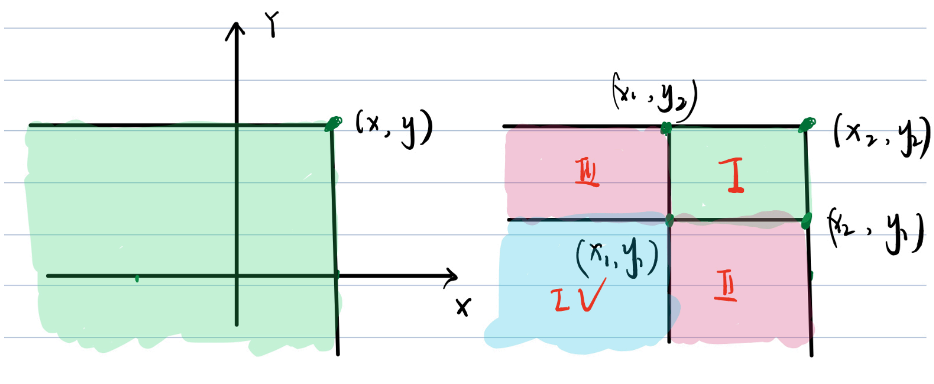


图 7: 联合分布函数

联合分布函数的性质，如图 7所示

$$\begin{aligned} & P\left\{x_1 < X \le x_2, y_1 < Y \le y_2\right\} \\ = & F\left(x_2, y_2\right) - F\left(x_1, y_2\right) - F\left(x_2, y_1\right) + F\left(x_1, y_1\right) \ge 0 \end{aligned} \tag{24}$$

边缘分布函数

$$\begin{aligned}F_X(x) &= P\{X \leq x\} = P\{X \leq x, Y < \infty\} = F(x, \infty) \\ &= \lim_{y \to \infty} F(x, y) \\ F_Y(Y) &= P\{Y \leq y\} = P\{X < \infty, Y \leq y\} = F(\infty, y) \\ &= \lim_{x \to \infty} F(x, y)\end{aligned} \tag{25}$$

边缘分布函数有随机向量 (X,Y) 的分布函数唯一确定，但是 (X,Y) 的联合分布函数由两方面的内容组成: X,Y 各自的边缘分布函数，X 和 Y 之间的关系。

3.2 二维离散型随机变量

每个分量都是离散型随机变量。

二维随机变量所有的可能取值为 (x_i, y_i)

$$P\{X = x_i, Y = y_j\} = p_{ij} \quad (i, j = 1, 2, \cdots) \tag{26}$$

称为二维随机变量的**联合概率分布**

相应的**边缘分布**

$$P(X = x_i) = p_{i\cdot} = \sum_{j=1}^{\infty} p_{ij} \quad (i = 1, 2, \cdots) \tag{27}$$

3.3 二维连续型随机变量

定义类似于一维随机变量，二维随机变量 (X,Y) 存在非负函数 f(x,y) 使得对于任意实数 x,y

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv \tag{28}$$

则，f(x,y) 是二维随机变量的**联合概率密度函数**

联合概率密度函数的性质:

1. f(x,y) 在 (x,y) 处连续，则

$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y) \tag{29}$$

2. 二维随机变量取值概率

$$P\left\{x_1 < X \le x_2, y_1 < Y \le y_2\right\} = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,y)dxdy \tag{30}$$

3. 二维随机变量取值概率转换为一个二重积分，所以该概率在数值上等于以区域 D 为底，f(x,y) 为顶面的曲顶柱体的体积

$$P\{(X,Y) \in D\} = \iint_D f(x,y)dxdy \tag{31}$$

边缘分布函数和边缘概率密度函数:

$$F_X(x) = F(x,\infty) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{x} f(u,y)du\right] dy = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(u,y)dy\right] du \tag{32}$$

概率密度函数就为:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \tag{33}$$

3.4 条件分布

3.4.1 二维离散型随机变量的条件分布律

给定二维随机变量的联合分布律，以及固定的一个随机变量 Y:

$$P\left\{X = x_i | Y = y_j\right\} = \frac{P\left\{X = x_i, Y = y_j\right\}}{P\left\{Y = y_j\right\}} = \frac{p_{ij}}{p \cdot j} \quad (i = 1, 2, \cdots) \tag{34}$$

即为 $Y=y_j$ 条件下，随机变量 X 的条件分布律。

3.4.2 二维连续型型随机变量的条件分布律

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)} \quad (-\infty<x<\infty) \tag{35}$$

即为 $Y=y$ 条件下，随机变量 X 的条件分布律。

3.5 随机变量的独立性

$F(X,Y)$, $F_X(x)$, $F_Y(y)$ 分别为联合分布函数以及边缘分布函数，对于任意的 x,y 有:

$$F(x,y)=F_X(x)F_Y(y) \tag{36}$$

则随机变量 X，Y 相互独立。

3.6 二维随机变量的函数的分布

二维随机变量 (X,Y) 概率密度 f(x,y) 。(X,Y) 的函数为 $Z=g(X,Y)$。其分布函数为:

$$F_Z(z)=P\{Z\leq z\}=\iint_{g(x,y)\leq z} f(x,y)dxdy \tag{37}$$

我们求解一个特殊情况，即和的分布 $Z=X+Y$

$$\begin{aligned}F_Z(z)=P\{X+Y\leq z\}&=\iint_{x+y\leq z} f(x,y)dxdy\\&=\int_{-\infty}^{+\infty}dx\int_{-\infty}^{z-x}f(x,y)dy\Bigg(\quad\int_{-\infty}^{+\infty}dy\int_{-\infty}^{z-y}f(x,y)dx\end{aligned} \tag{38}$$

因此得到相应的概率密度

$$f_Z(z) = \int_{-\infty}^{+\infty} f(x, z-x)dx \tag{39}$$

或者

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dy \tag{40}$$

如果 X，Y 是独立的则

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(x)f_Y(z-x)dx \tag{41}$$

或者

$$f_Z(z) = \int_{-\infty}^{+\infty} f_X(z-y)f_Y(y)dx \tag{42}$$

上述的运算称为卷积，记为：

$$f_Z = f_X * f_Y$$

4 随机变量的数字特征

4.1 随机变量的数学期望

1. 离散型随机变量 X 概率分布律 $P(X=x_k)=p_k$, $\sum_{k=1}^{\infty} x_k p_k$ 绝对收敛，其期望为

$$E(X)=\sum_{k=1}^{\infty} x_k p_k \tag{43}$$

2. 连续型随机变量 X 的概率密度为 $f(X)$, 若期望存在，则相应的期望为:

$$E(X)=\int_{-\infty}^{\infty} x f(x) d x \tag{44}$$

随机变量的数学期望的本质是加权平均数，是一个数，不再是一个随机变量

但是区别于均值，均值是指向当前试验，而期望是指向整个样本空间。**期望就是平均数随样本趋于无穷的极限**。

4.1.1 一维随机变量函数的数学期望

Y 是随机变量 X 的函数，$Y=g(X)$, g 是连续函数

1. 离散型随机变量 X 概率分布律 $P(X=x_k)=p_k$

$$E(Y)=E[g(X)]=\sum_{k=1}^{\infty} g\left(x_k\right) p_k \tag{45}$$

2. 连续型随机变量 X 的概率密度为 $f(X)$, 若期望存在，则相应的期望

为:

$$E(Y)=E[g(X)]=\int_{-\infty}^{\infty} g(x) f(x) d x \tag{46}$$

4.1.2 数学期望的性质

期望本质是积分运算，所以下面性质都可以使用积分的性质进行推导。

- E(aX+bY) = aE(X) + bE(Y)
- 两个互相独立随机变量 X 和 Y , $E(XY) = E(X)E(Y$

4.2 随机变量的方差

方差可以用来表示一个分布的离散程度，如图 8, 两个均匀分布有同样的均值，但是分布有着明显的区别。

定义 X 为一个随机变量，若方差存在，

$$D(X)=E\left[(X-E X)^{2}\right] \tag{47}$$

$$D(X)=E\left(X^{2}\right)-(E X)^{2} \tag{48}$$

D(X) 为 X 的方差，$\sqrt{D(X)}$ 为标准差

方差的性质:

1. $D(aX+b)=a^2D(X)$
2. 两个互相独立随机变量 X 和 Y , $D(X+Y)=D(X)+D(Y)$

当性质 1 中的 a 等于 1 的时候，如图 9 所示，形状不变相当于对分布进行" 搬家 “

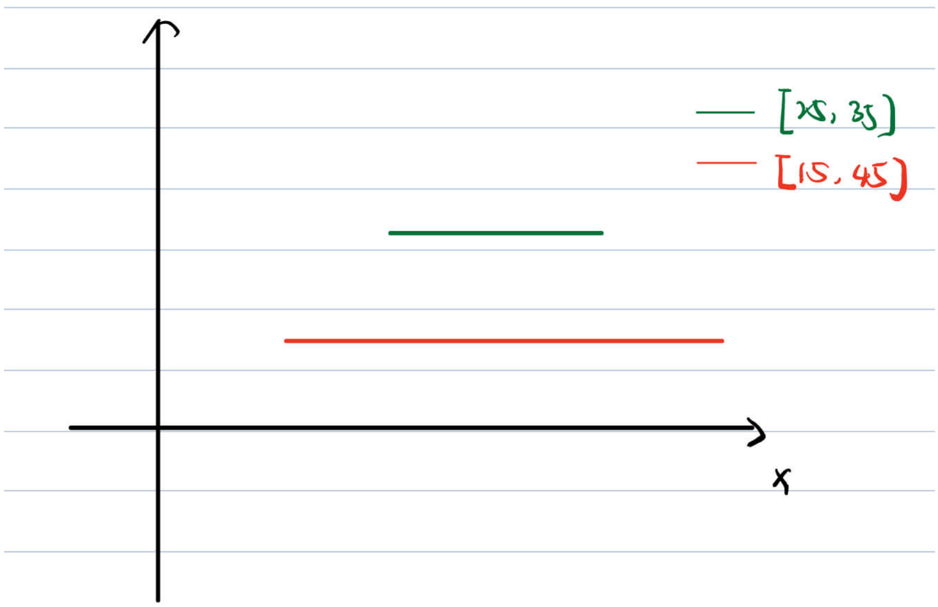


图 8: 两个均匀分布

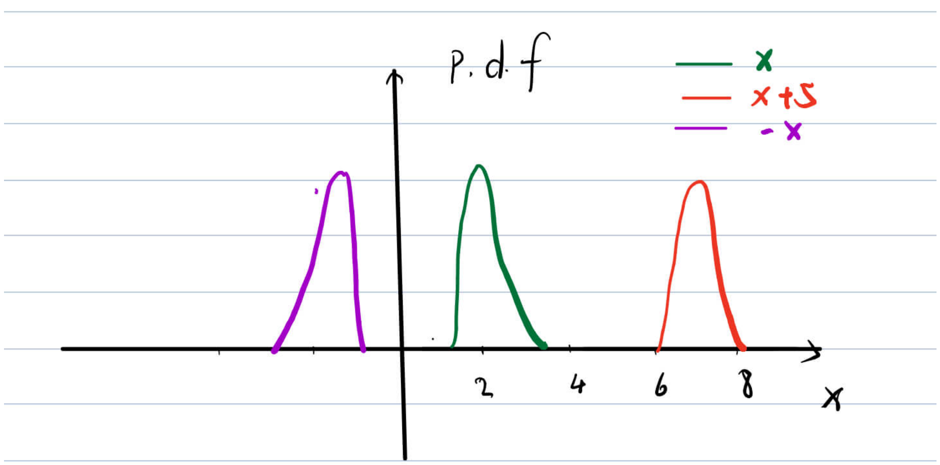


图 9: 分布“搬运”

4.3 协方差与相关系数

4.3.1 协方差

上面的期望和方差只反映变量各自的性质，不能表示 X，Y 之间的相互关系，引入协方差，刻画随机变量之间的 **线性相关性**

定义

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] \tag{49}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{50}$$

简单理解协方差：如果协方差为正，则 X，Y 同向变化，协方差越大，同向程度越高；

性质

$$\operatorname{Cov}(X+Y,Z) = \operatorname{Cov}(X,Z) + \operatorname{Cov}(Y,Z) \tag{51}$$

$$D(X+Y) = D(X) + D(Y) + 2\operatorname{Cov}(X,Y) \tag{52}$$

4.3.2 相关系数

协方差的对于变量的描述不稳定，受 X，Y 本身数值影响。X，Y 各自增大 K 倍，相互联系应该是一样的，但是协方差却增大了 K^2。还有协方差数值大小依赖于 X，Y 的度量单位。为了解决上述问题，引入相关系数。

定义

$$\rho = \rho_{XY} = \frac{\operatorname{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{53}$$

除以标准差后，相关系数是一个无量纲的绝对量，不受使用的度量单位影响。简单来说

为什么除以标准差呢？因为标准差描述了变量在整体变化过程中偏离均值的幅度

相关系数是随机变量标准化之哈偶的协方差

标准化

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}} \tag{54}$$

标准化随机变量就是把园分布中心 *E(X)* 移动至原点，不使分布中心偏左或偏右，然后扩大或者缩小坐标轴，使得分布不至于过疏或者过密

$$\rho = \rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \mathrm{Cov}(X^*, Y^*) \tag{55}$$

性质

$|\rho| <= 1$ 且 $|\rho| = 1$ 的充要条件是 X, Y 之间线性相关。

相关系数只是描述变量间线性关系强弱的一个度量。也可以叫做线性相关系数

但是注意，$|\rho| = 0$ 代表不相关，但是不代表 **独立**. 这里的不相关只是不存在线性关系，但是还有可能存在别的函数关系

若 (X,Y) 服从二维正态分布，则 X，Y 的不相关等价于 X，Y 相互独立

4.4 矩

最广泛使用的一种数字特征。

定义

k 阶原点矩:

$$E(X^k) \tag{56}$$

k 阶中心矩:

$$E((X - E(X))^k) \tag{57}$$

k+l 阶混合原点矩

$$E(X^k Y^l) \tag{58}$$

k+l 阶混合中心矩

$$E((X - E(X))^k (Y - E(Y))^l) \tag{59}$$

根据定义可知，数学期望是一阶原点矩，方差是二阶中心矩
定理
如果存在 k 阶矩，那么存在所有小于 k 阶的距。

4.4.1 矩生成函数: Moment Generating Functions

定义
X 是随机变量，t 是任意的实数，

$$\psi(t) = E\left(e^{tX}\right) \tag{60}$$

$\psi(t)$ 就是矩生产函数
矩生成函数的 n 阶导数在 t=0 处的值，刚好是随机变量的 n 阶距

性质 $X_1, X_2..X_n$ n 个独立的随机变量的和，可以用矩生成函数

$$\psi(t) = \Pi_{i=1}^{n} \psi_i(t) \tag{61}$$

$$
\begin{gathered}
\psi(t)=E\left(e^{tY}\right)=E\left(e^{tX_1+\cdots+tX_n}\right)=E\left(\Pi_{i=1}^{n}e^{tX_i}\right)\\
E\left(\Pi_{i=1}^{n}e^{tk_i}\right)=\Pi_{i=1}^{n}E\left(e^{t_i}\right)\\
\psi(t)=\Pi_{i=1}^{n}\psi_i(t)
\end{gathered}
\tag{62}
$$

4.4.2 二项分布矩生成函数: Moment Generating Functions

服从 (n,p) 的二项分布，二项分布是多个独立的伯努利分布加起来的结果

$$
\begin{aligned}
\psi_i(t)&=E\left(e^{tX_i}\right)\\
&=p\times e^t+(1-p)\times e^0\\
&=pe^t+1-p
\end{aligned}
\tag{63}
$$

根据:

$$
\psi(t)=\Pi_{i=1}^{n}\psi_i(t)
\tag{64}
$$

得到:

$$
\psi(t)=\left(pe^t+1-p\right)^n
\tag{65}
$$

4.4.3 二项分布的可加性

X_1，X_2 是独立的随机变量。假如 X_i 是服从参数 n_i p 的二项分布，则，X_1+X_2 服从参数 n_1+n_2 ，p 的二项分布。

$$
\psi_i(t)=\left(pe^t+1-p\right)^{n_i}
\tag{66}
$$

$$
\psi(t)=\left(pe^t+1-p\right)^{n_1+n_2}
\tag{67}
$$

4.5 中位数与分位数

常用于连续型随机变量。

定义：

连续型随机变量 X 的分布函数为 $F(x)$，满足条件

$$F(x_{0.5}) = P(X <= x_{0.5}) = 0.5$$

数 $x_{0.5}$ 为 X 的中位点。

$$F(x_\alpha) = P(X <= x_\alpha) = \alpha$$

x_α 为 X 的下 α 分位点。

4.6 条件数学期望

联合概率密度为 $f(x, y)$, $f_{Y|X}(y|x)$ 表示 X= x 条件下 Y 的条件概率密度

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \tag{68}$$

E(Y|X = x) 是 x 的函数，记为 $E(Y|X) = g(X)$ ，称为 Y 对 X= x 的回归函数。

$E(Y|X) = g(X)$ 也是随机变量，可以对其求数学期望，

$$E(E(Y|X=x)) = \int_{-\infty}^{\infty} E(Y|X=x) f_X(x) dx = E(Y) \tag{69}$$

称为 **全期望公式**，可以理解为分两步计算数学期望的一种方法。

先计算条件期望 $E(Y|X=x)$, 再借助 X 的分布，通过公式 69对 X 求期望得到 E(Y)

条件期望和 MSE 之间的关系

Theorem The prediction $d(X)$ that minimizes $E[Y-d(X)]^2$ is $d(X) = E(Y|X)$

详见 https://face2ai.com/math-probability-4-7-conditional-expectation/

5 大数定律和中心极限定理

5.1 马尔可夫不等式,Markov Inequality

定义:

随机变量 X 的期望 E(X) 存在,

$$P(X \geq \varepsilon) \leq \frac{E(X)}{\varepsilon} \tag{70}$$

马尔科夫不等式得到概率和期望之间关系

如果一个随机变量的均值是 1,那么其取到大于等于 100 的概率是

$$P(X \geq 100) \leq \frac{1}{100} = 0.01$$

5.2 切比雪夫不等式

定义:

随机变量 X 的期望 E(X) 和方差 D(X) 均存在,任意的实数 $\varepsilon > 0$

$$P\{|X - E(X)| \geq \varepsilon\} \leq \frac{D(X)}{\varepsilon^2} \tag{71}$$

方差是反应随机变量分布的离散情况的描述。不等式表明,随机变量值与其均值之间的距离的概率受到其方差的制约。

例如:令 $D(X) = \sigma^2, \varepsilon = 3\sigma$

$$P(|X - E(X)| \geq 3\sigma) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}$$

超过 3σ 的部分概率小于 $\frac{1}{9}$

5.3 大数定律

5.3.1 随机变量序列的依概率收敛

定义：

X_n 为一个随机变量序列，X 为一个随机变量

$$\lim_{n \to \infty} P\{|X_n - X| < \varepsilon\} = 1 \qquad \lim_{n \to \infty} P\{|X_n - X| \geq \varepsilon\} = 0 \tag{72}$$

称 X_n 依概率收敛于 X

直观理解：对于任意的 $\varepsilon > 0$ 当 n 充分大的时候，X_n 与 X 的偏差大于 ε 这个时间发生的概率很小，收敛到 0.

性质：

$$X_n \xrightarrow{P} a, Y_n \xrightarrow{\quad P \quad} b$$

函数 $g(x,y)$ 在点 (a,b) 处连续，则

$$g(X_n, Y_n) \xrightarrow{P} g(a,b) \tag{73}$$

5.3.2 切比雪夫大数定律

定义

$X_1, X_2 ... X_n$ 是**独立**的随机变量序列。每个随机变量的数学期望存在。存在常数 C ，使得 $D(X_n <= C)$, 则：

$$\lim_{n \to \infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right\} = 1 \tag{74}$$

可由切比雪夫不等式证明

特殊情况

$X_1, X_2 ... X_n$ 是**独立同分布**的随机变量序列，即有相同的期望和方差。$E(X_i) = \mu$

$$\lim_{n \to \infty} P\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right\} = 1 \tag{75}$$

理解：

试验次数 n 趋向于无穷的时候，平均值 $\frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛与数学期望。

在测量中使用多次重复测量结果的算术平均值来作为测量值的近似

5.3.3 伯努利大数定律

定义

n_A 为 n 重伯努利实验中事件 A 发生的次数，A 每次发生的概率为 p，则：

$$\lim_{n \to \infty} P\left\{\left|\frac{n_A}{n} - p\right| < \varepsilon\right\} = 1 \tag{76}$$

证明

$X_1, X_2 ... X_n$ 是**独立同分布**的随机变量序列。$E(X_n) = p$

$$n_A = \sum i^n X_i$$

带入切比雪夫大数定律，证完

简单理解

事件趋于无穷的时候，事件 A 发生的频率依概率收敛到 A 发生的概率

5.3.4 辛钦大数定律

5.4 中心极限定理

若被研究随机变量是大量的独立的随机变量的和，其中每一随机变量对于总和只有微小的作用，则**可以人为这个随机啊变量近似的服从正态分布**

现实中很多随机变量具有上述性质，例如人的身高，都是由大量的独立随机因素综合影响的结果

中心极限定理：随机变量序列的极限分布是正态分布的结果

5.4.1 随机变量序列的依分布收敛

最弱的收敛方式

定义：

X_n 为一个随机变量序列，对应的分布函数列为 $F_n(x)$ X 为一个随机变量，分布函数为 $F(x)$，如果对于 F(X) 的任意连续点 x ，有

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{77}$$

则称 X_n 依分布收敛到 X

5.4.2 列维-林德伯格中心极限定理

$X_1, X_2 ... X_n$ 是**独立同分布**的随机变量序列。$E\left(X_i\right) = \mu, D\left(X_i\right) = \sigma^2 < \infty (i = 1, 2, \cdots)$. 则

$$\lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{78}$$

从上面公式中看出，不管 X_n 服从什么分布，只要 n 足够大，$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$ 这个随机变量就服从**标准正态分布**，线性变换后，$\sum_{i=1}^{n} X_i$ 服从正态分布 $N(n\mu, n\sigma^2)$

5.4.3 棣莫夫-拉普拉斯中心极限定理

n_A 为 n 重伯努利实验中事件 A 发生的次数，A 每次发生的概率为 p，$X_1, X_2 ... X_n$ 是**独立同分布**的随机变量序列。$E(X_n) = p$

$$n_A = \sum_{i}^{n} X_i$$

带入林德伯格中心极限定理

$$\lim_{n\to\infty} P\left\{\frac{n_A - np}{\sqrt{npq}} \le x\right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{79}$$

$$n_A = \sum_{i}^{n} X_i$$

服从二项分布 $B(n,p)$。

所以 **对于二项分布，当 n 很大的时候，根据依分布收敛，二项分布可以用正态分布进行近似**

6 结语

笔记中有些定义因为时间关系写的不严谨，不规范；证明因为公式太多 (懒惰) 也没有给出；相应使用例子欠缺 (导致读起来不会那么友好)；这些会慢慢补充，未完不待续...

本篇笔记主要是概率论，数理统计相关的见下一篇