粒子物理与核物理实验中的数据分析

陈少敏
清华大学

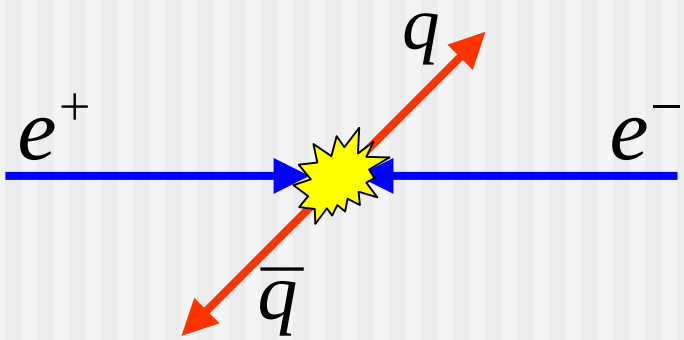
第一讲：基本概念

http://hep.tsinghua.edu.cn/~chensm/lectures/lecture_1.ppt

本次讲座的要点

- 概率
- 随机变量与函数
- 期待值
- 误差传递

实验的目的是什么？



观察某一过程
的 **n** 个事例

实验测量出每个事例的特征量(能动量，末态粒子数...)。

理论预言出上述各特征量的分布，而且可能还会包含某些如相互作用耦合常数等自由参数。

收集数据
统计分析 → 估计参数值与相应的误差围，检验在何种程度上理与实验数据相符。

问题：如何评价这种检验？ → 使用概率来量化结论！

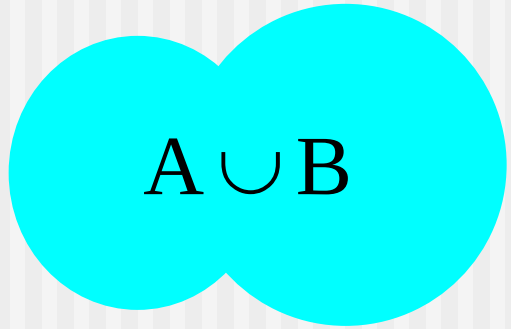
随机事例

在一定的实验条件下，现象**A**可能发生，也可能不发生，并且只有发生或不发生这样两种可能性，这是偶然现象中一种比较简单的形态，我们把发生了现象**A**的事例称为随机事例**A**，简称事例**A**。

随机事例之间的相互关系

A与B之并事例 $A \cup B$

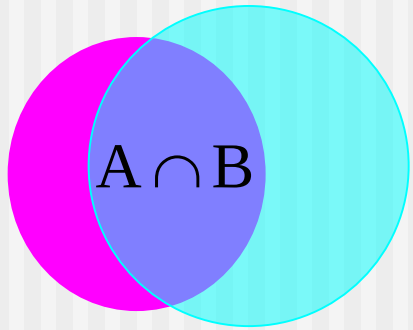
指事例A与B中至少有一个出现的事例



如果A与B互斥，则 $A \cup B = A + B$

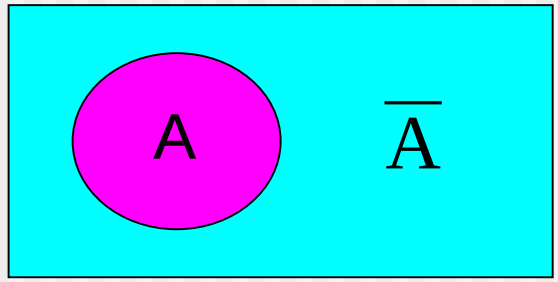
A与B之积(交)事例 $A \cap B$

指事例A与B中同时出现的事例



A之逆事例 $\overline{A}$

指事例A不出现的事例



概率的定义

柯尔莫哥洛夫公理：考虑一全集**S**具有子集**A**，**B**，...

$$A \subset S, P(A) \geq 0$$

$$P(S) = 1$$

$$A \cap B = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$$

P(A)称为事例**A**的概率

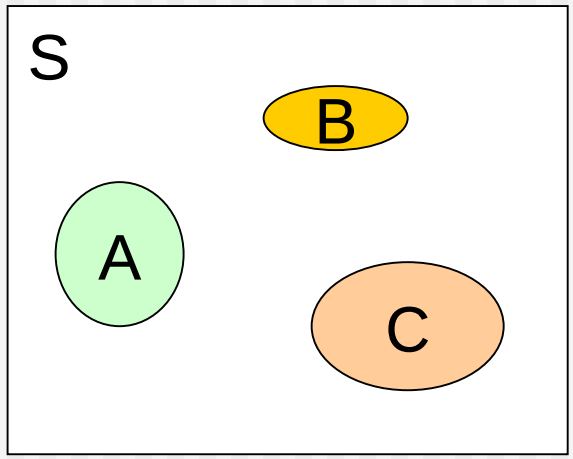
从该公理可以导出下列概率公式

$$P(\overline{A}) = 1 - P(A)$$

$$P(A \cup \overline{A}) = 1$$

$$A \subset B \Rightarrow P(A) \leq P(B)$$

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



条件概率

假设**B**出现的概率不为零，在给定**B**的情况下出现**A**的条件概率定义为

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

如果 $P(A \cap B) = P(A)P(B)$ 则表明**A**与**B**相互独立。

如果**A**与**B**相互独立，则有

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

结果与**B**无关

注意：与不相交的子集定义不同 $A \cap B$

概率的含义

- 相对频率

假设**A**，**B**，**...**是一可重复实验的结果，则概率就是

$$P(A) = \lim_{n\to\infty} \frac{\text{结果为 A}}{n\text{次实验}}$$

- 主观概率

如果**A**，**B**，**...**是假设**(**是真或是假的各种陈述**)**，那么概率

$$P(A) = \text{对A为真的信心程度}$$

✓两种解释皆与柯尔莫哥洛夫公理相符。

✓概率的频率解释在数据分析中用起来比较自然，但是...

频率概率中的问题

- 实际问题中，统计量总是有限的。**P(A)**完全取决于**A**的划分与总统计量的大小。

概率大小会出现波动。

需要解决好
- **A**的定义
- 适当的误差

- 该定义不适用于某些特殊情况

例如：我们可以说“明天有雨”。但是，如果我们根据概率频率定义说“明天可能有雨”，却是一个毫无科学意义的预报。

主观概率中的问题

- 主观性：在对同一随机现象的描述中，我的**P(理论)**与你的**P(理论)**可能不同

理论家甲
之理论A

理论家乙
之理论B

- 使用主观概率的原因
  - 出于绝望 ✓
  - 出于无知 ×
  - 出于懒惰 ?

主观概率的一些特点

主观概率有一些吸引人的地方，例如对于不可重复现象的处理中，显得比较自然

- 系统误差(重复实验时仍保持不变)；
- 在该事例出现的粒子是正电子；
- 自然界是超对称的；
- 明天将下雨(将来事件的不确定性)；
- 公元1500年元月一日北京下雨(过去事件的不确定性)。

结论中包含了主观上对事件为真的信念!

频率论者与主观概率

P(938.27195<质子质量<938.27211MeV)是什么？

当以质量来判断一实际为质子的粒子类别时

- 频率论者：质子或非质子（不知道是哪个）
- 主观主义者(贝叶斯论者):68%是质子(对知识的陈述)

对主观概率而言，意味着

质子质量的不确定性与从100只球中有68只白球的球筐里能拿出白球的不确定性一样。

频率论者与主观概率(续)

如果大多数贝叶斯论者说

- 巴西赢得2006年世界杯冠军的概率为68%
- 质子质量在938.27195-938.27211MeV内的概率为68%
- 希拉里.克林顿2009年入主白宫的概率为68%

那么上述论断的68%就应该理解为结果为真的概率。

能否在频率定义中将质子质量在938.27195-938.27211MeV内理解成：在整个宇宙中，自然界给出了各种不同的质子质量，而它们中有68%在938.27195与938.27211MeV之间?

没问题...只不过这是对信心程度的一种表达。

贝叶斯定理

根据条件概率的定义

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

而 $P(A\cap B)=P(B\cap A)$，故

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

贝叶斯定理由 **Reverend Thomas Bayes (1702-1761)** 首先提出。

贝叶斯理论与主观概率

贝叶斯理论通常用于主观概率问题

$$P(\text{理论}|\text{实验}) = \frac{\text{P(实验}|\text{理论)}}{\text{P(实验)}}\text{P(理论)}$$

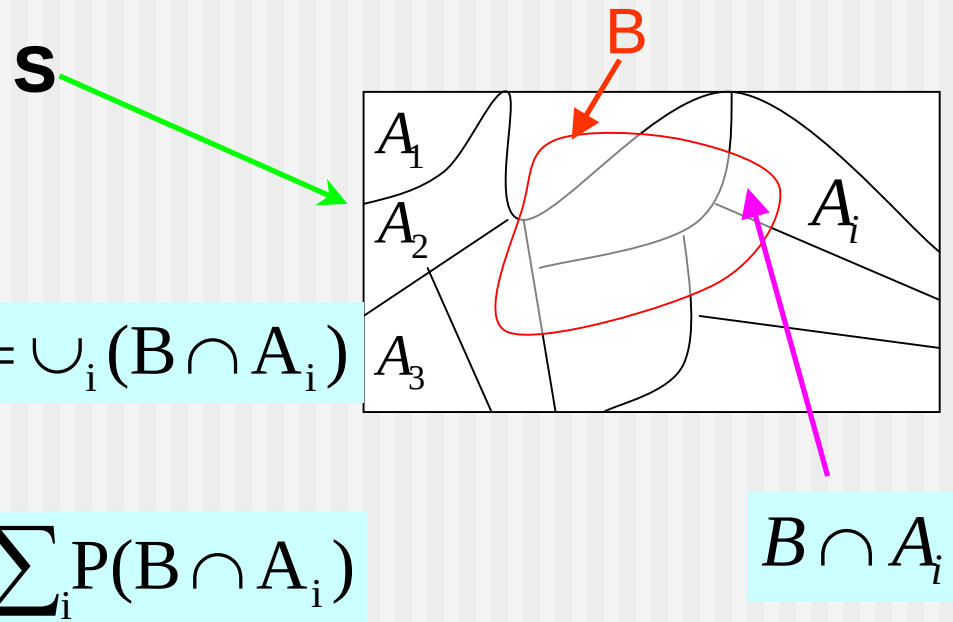
通过实验结果改进基于某一理论的信念(后验性的)

- 如果实验证明P(实验|理论)=0，则表明理论不能接受。
- 大的P(实验|理论)会增加对理论的信任度。
- 通过实验结果可以改进P(理论)。
- 改进的P(理论)可应用于对重复实验结果的预测。
- P(实验|理论)对先验理论的依赖将最终消失。

全概率事例

考虑在样本空间**S**中有一子集**B**。将样本空间分为互斥的子集$\mathbf{A_i}$，使得

$$\cup_i A_i = \sum_i A_i = S$$



因此，

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$$

表示成概率的形式为

$$P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

得到全概率事例公式

$$P(B) = \sum_i P(B \mid A_i) P(A_i)$$

贝叶斯定理

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{\sum_i P(B \mid A_i) P(A_i)}$$

例子：如何利用贝叶斯定理

假设对任意一个人而言，感染上**AIDS**的概率为

$P(AIDS) = 0.001$　　验前概率,即任何检验之前

$P(no\ AIDS) = 0.999$

考虑任何一次**AIDS**检查的结果只有阴性**(-)**或阳性**(+)**两种

$P(+|AIDS) = 0.98$　　AIDS感染患者阳性的概率

$P(-|AIDS) = 0.02$　　AIDS感染患者阴性的概率

$P(+|no\ AIDS) = 0.03$　　AIDS未感染者阳性的概率

$P(-|no\ AIDS) = 0.97$　　AIDS未感染者阴性的概率

如果你的检查结果为阳性**(+)**，而你却觉得自己无明显感染渠道。那么你是否应担心自己真的感染上了**AIDS**?

例子：如何利用贝叶斯定理(续)

利用贝叶斯定理，阳性结果条件下是**AIDS**患者的概率为

$$P(AIDS|+) = \frac{P(+|AIDS)P(AIDS)}{P(+|AIDS)P(AIDS)+P(+|no\ AIDS)P(no\ AIDS)}$$

$$= \frac{0.98\times 0.001}{0.98\times 0.001+0.03\times 0.999}$$

$$= 0.032 \quad (\text{验后概率})$$

AIDS患者阳性
所有为阳性结果的人

也就是说，你可能没什么问题！

从你的观点上看：对自己染上**AIDS**结果的可信度为**3.2%**。

从医生角度上看：象你这样的人有**3.2%**感染上了**AIDS**。

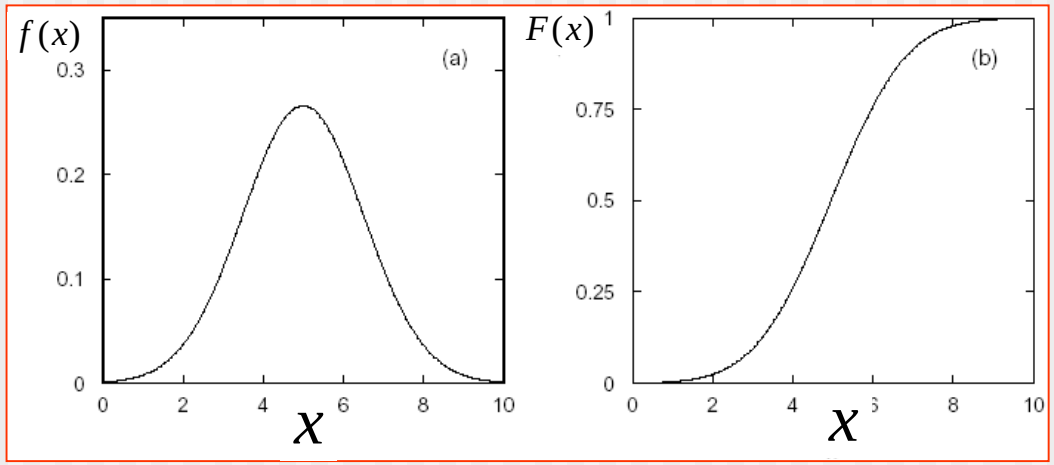
随机变量与概率密度函数

假设实验结果为**x** **(**记作样本空间中元素**)**

$$P(\text{观测到}\,x\,\text{在}[x, x+dx]\text{范围内}) = f(x)dx$$

那么概率密度函数 **p.d.f.** 定义为 $f(x)$，它满足

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



定义累积分布函数为

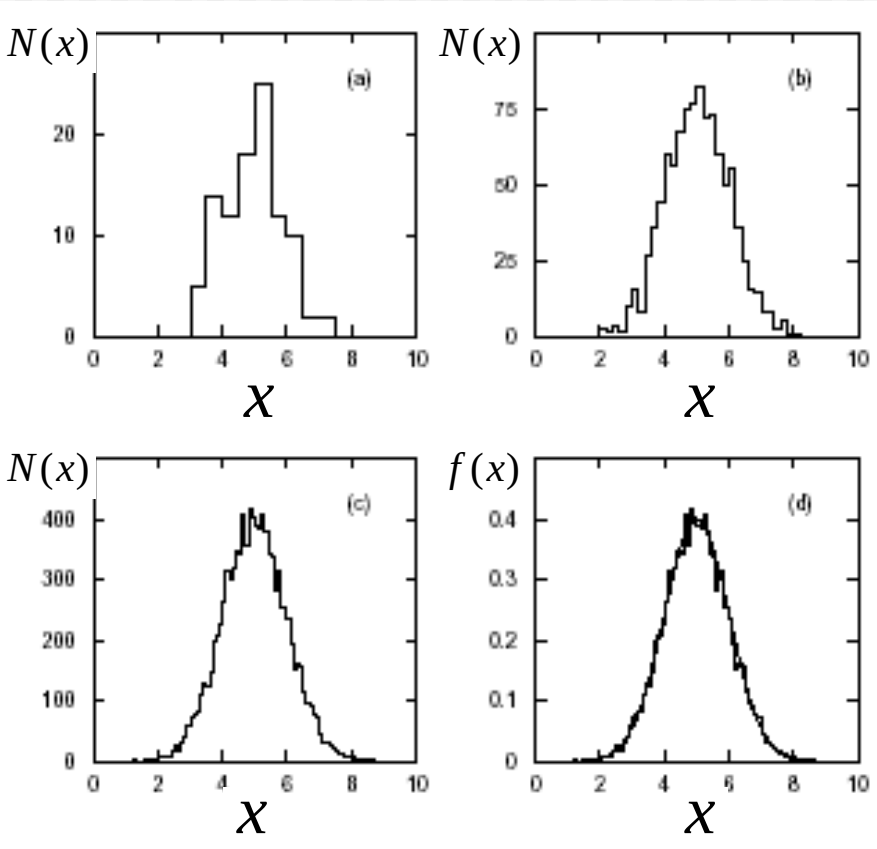
$$F(x) = \int_{-\infty}^{x} f(x')dx'$$

对于离散型随机变量

$$f_i = P(x_i), \qquad \sum_{i=1}^{n} f_i = 1, \qquad F(x) = \sum_{x_i \le x} P(x_i)$$

直方图与概率密度函数

概率密度函数 **p.d.f.** 就是拥有无穷大样本，区间宽度为零，而且归一化到单位面积的直方图。



$$f(x)=\frac{N(x)}{n\Delta x}$$

$N(x)=$每个区间的事例数(频数)

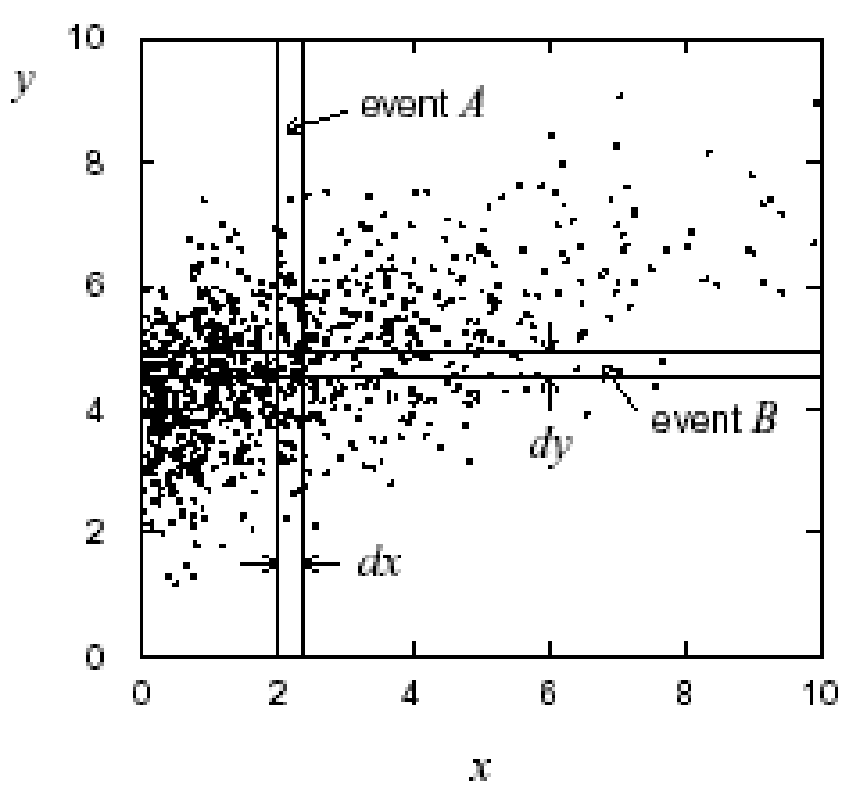
$n=$添入直方图的总事例数

$\Delta x=$区间的宽度

直方图在统计分析中非常重要，应准确理解它的含义。

多变量情形

如果观测量大于一个，例如 x 与 y



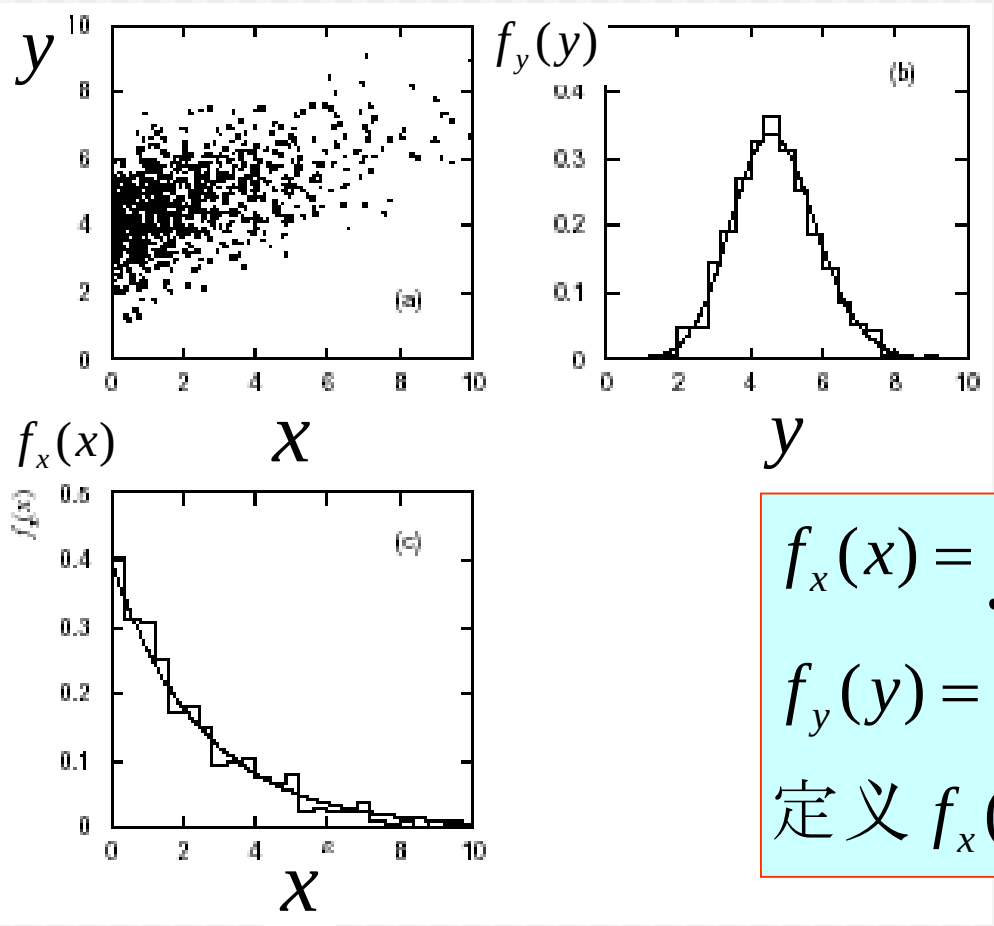
$$P(A \cap B) = \iint f(x,y)dxdy$$

$$f(x,y) = \text{联合的p.d.f.}$$

$$\iint f(x,y)dxdy = 1$$

投影分布

将联合概率密度函数 **p.d.f.** 投影到 x, y 轴**(**如图所示**)**



$$f_x(x) = \int f(x, y)dy$$

$$f_y(y) = \int f(x, y)dx$$

定义 $f_x(x), f_y(y)$ = 投影的p.d.f.

条件概率密度函数

利用条件概率的定义，可得到

$$P(B|A)=\frac{P(A\cap B)}{P(A)}=\frac{f(x,y)dxdy}{f_x(x)dx}$$

定义条件概率的密度函数 **p.d.f.** 为

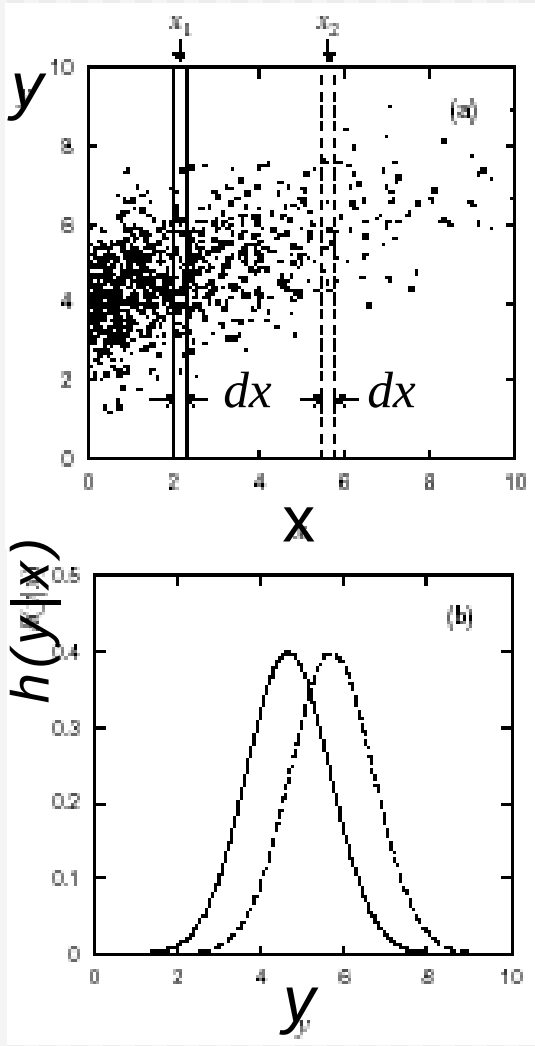
$$h(y|x)=\frac{f(x,y)}{f_x(x)},\quad g(x|y)=\frac{f(x,y)}{f_y(y)}$$

则贝叶斯定理可写为

$$g(x|y)=\frac{h(y|x)f_x(x)}{f_y(y)}$$

若 x,y相互独立，则

$$f(x,y)=f_x(x)f_y(y)$$



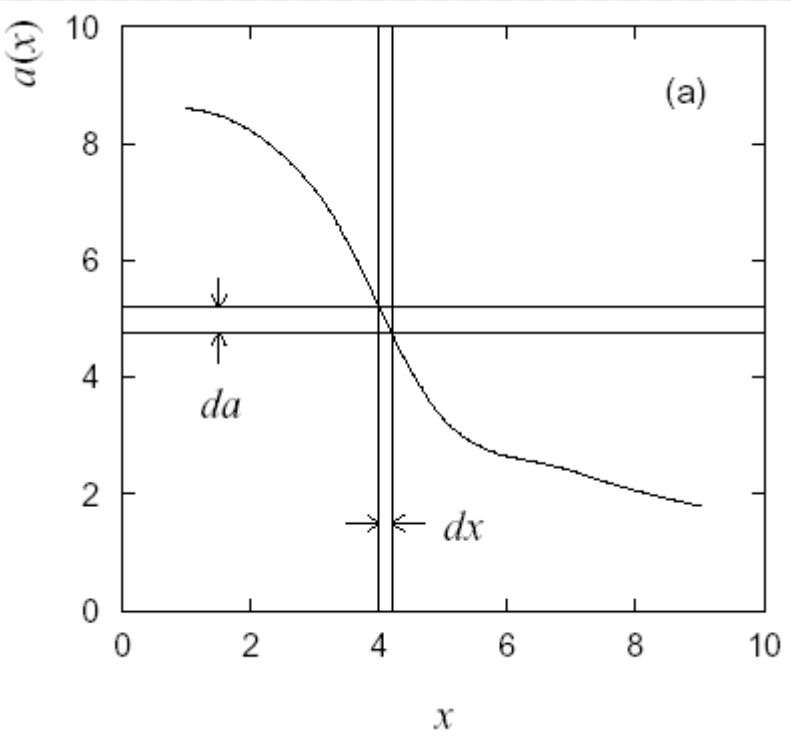
随机变量的函数*

随机变量的函数自身也是一个随机变量。

例如：

θ与$\cos\theta$

假设x服从 **p.d.f.** $f(x)$，对于函数$a(x)$，其**p.d.f.** $g(a)$为何？



$$g(a)da = \int_{dS} f(x)dx$$

$$dS = a\text{在}[a, a+da]\text{内的}x\text{空间范围}$$

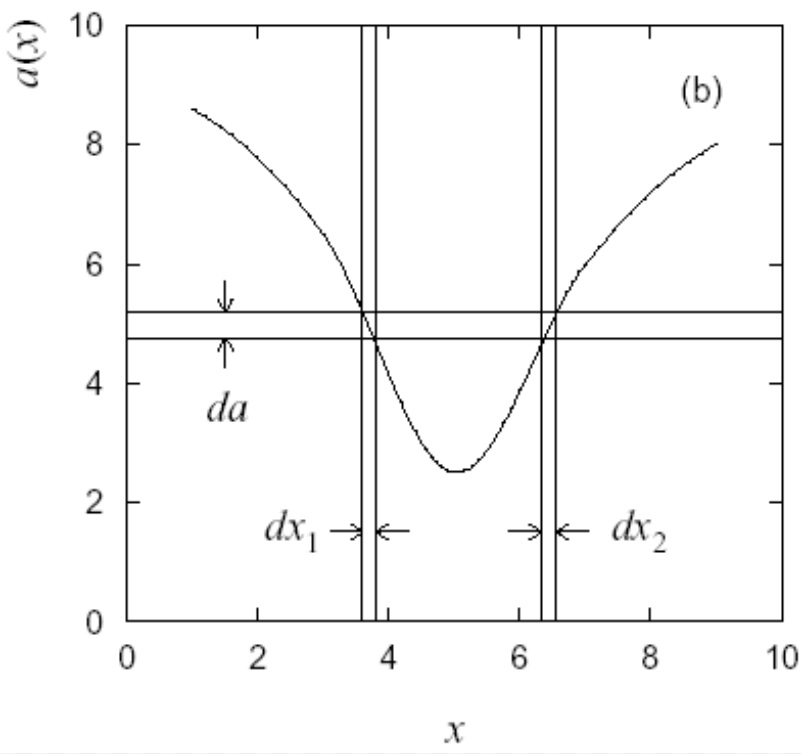
$$g(a)da = \left|\int_{x(a)}^{x(a+da)} f(x')dx'\right|$$

$$= \int_{x(a)}^{x(a)+\left|\frac{dx}{da}\right|da} f(x')dx'$$

$$\Rightarrow g(a) = f(x(a))\left|\frac{dx}{da}\right|$$

函数的逆不唯一情况*

假如$a(x)$的逆不唯一，则函数的 **p.d.f.** 应将dS中对应于da的所有dx的区间包括进来



例如:$a = x^2,\ x = \pm\sqrt{a},\ dx = \pm\dfrac{da}{2\sqrt{a}}$

$$g(a)da = \int_{dS} f(x)dx$$

$$dS = \left[\sqrt{a}, \sqrt{a} + \frac{da}{2\sqrt{a}}\right] \cup \left[-\sqrt{a} - \frac{da}{2\sqrt{a}}, -\sqrt{a}\right]$$

$$g(a) = \frac{f(\sqrt{a})}{2\sqrt{a}} + \frac{f(-\sqrt{a})}{2\sqrt{a}}$$

多个随机变量的函数*

考虑随机矢量 $\vec{x}=(x_1,...,x_n)$ 与函数 $a(\vec{x})$ 对应的 **p.d.f.**

$$g(a')da' = \int\cdots\int_{dS} f(x_1,...,x_n)dx_1...dx_n$$

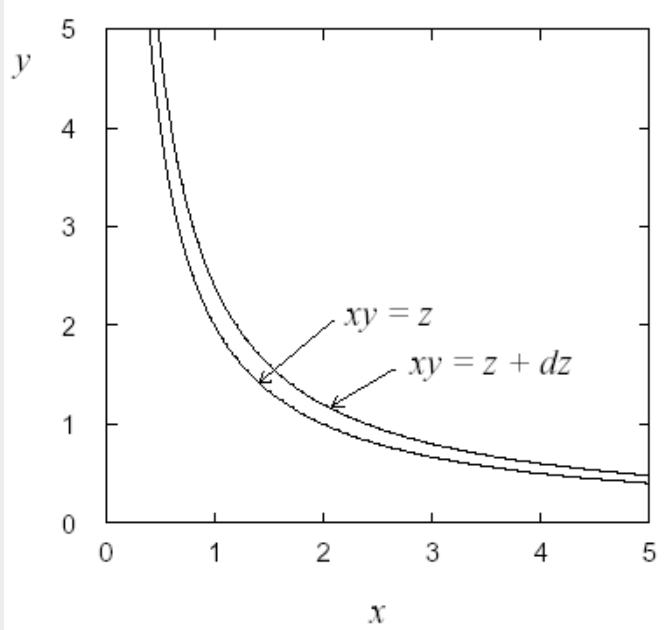
$$dS = \text{在}\, a(\vec{x})=a' \text{与}\, a(\vec{x})=a'+da' \text{定义的曲面}\, \vec{x}\, \text{空间范围}$$

例如随机变量 $x, y > 0$ 服从联合的 **p.d.f.** $f(x,y)$，考虑函数 $z=xy$，其 $g(z)$ 应是何种形式

$$g(z)dz = \int...\int_{dS} f(x,y)dxdy$$

$$= \int_0^{\infty} dx \int_{z/x}^{(z+dz)/x} f(x,y)dy$$

$$g(z) = \int_0^{\infty} f(x,\frac{z}{x})\frac{dx}{x} = \int_0^{\infty} f(\frac{z}{y},y)\frac{dy}{y}$$



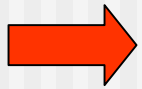
多个随机变量的函数(续)*

考虑具有联合的 **p.d.f.** 的随机矢量 $\vec{x}=(x_1,\ldots,x_n)$，构造 n 个线性独立的函数：$\vec{y}(\vec{x})=(y_1(\vec{x}),\ldots,y_n(\vec{x}))$，而且其逆函数 $x_1(\vec{y}),\ldots,x_n(\vec{y})$ 存在。那么 $\vec{y}$ 的联合 **p.d.f.** 为

$$g(\vec{y})=|J|\,f(\vec{x})$$

这里 J 是雅可比行列式

$$J=\begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & & & \vdots \\ & & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$



任意一个函数 $g_i(y_i)$ 均可通过对函数 $g(\vec{y})$ 积分掉其它不用的变量而得到。

期待值

考虑具有 **p.d.f.** $f(x)$ 的随机变量 x，定义期待(平均)值为

$$E[x]=\int x\, f(x)dx$$

通常记为： $E[x]=\mu$

注意: 它不是 x 的函数，而是 $f(x)$ 的一个参数。

对离散型变量，有 $E[x]=\sum_{i=1}^{n} x_i P(x_i)$

对具有 **p.d.f.** $g(y)$ 的函数 $y(x)$，有

$$E[y]=\int yg(y)dy=\int y(x)\, f(x)dx$$

方差定义为

$$V[x]=E[(x-E[x])^2]=E[x^2]-\mu^2$$

通常记为：$V[x]=\sigma^2$

标准偏差： $\sigma\equiv\sqrt{\sigma^2}$

协方差与相关系数

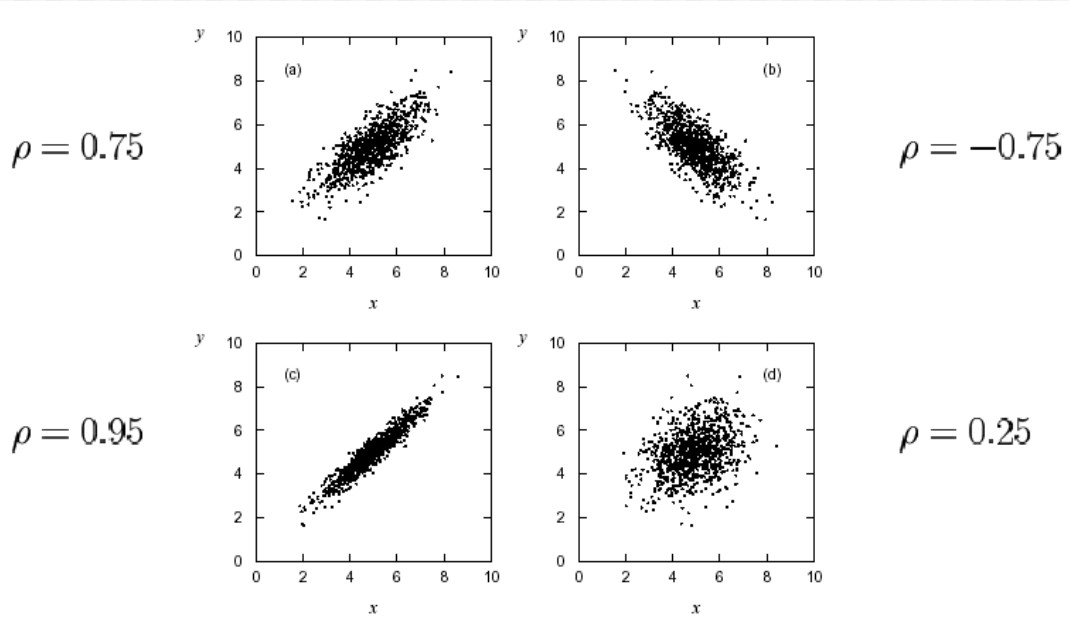
定义协方差 $\mathrm{cov}[x, y]$（也可用矩阵表示 V_{xy}）为

$$\mathrm{cov}[x, y] = E[(x-\mu_x)(y-\mu_y)] = E[xy] - \mu_x \mu_y$$

相关系数定义为

$$\rho_{xy} = \frac{\mathrm{cov}[x, y]}{\sigma_x \sigma_y},$$

$$-1 \le \rho_{xy} \le 1$$



如果 x, y 独立，即

$$f(x, y) = f_x(x) f_y(y)$$

则

$$\mathrm{cov}[x, y] = 0$$

误差传递

假设 $\vec{x}=(x_1,...,x_n)$服从某一联合 **p.d.f.** $f(\vec{x})$，我们也许并不全部知道该函数形式，但假设我们有协方差

$$V_{ij}=\mathrm{cov}[x_i,x_j]$$

和平均值 $\vec{\mu}=E[\vec{x}]$

现考虑一函数 $y(\vec{x})$，方差 $V[y]=E[y^2]-(E[y])^2$是什么？

将$y(\vec{x})$在 $\vec{\mu}$ 附近按泰勒展开到第一级

$$y(\vec{x})\approx y(\vec{\mu})+\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}}(x_i-\mu_i)$$

然后，计算 $E[y]$与 $E[y^2]$ **...**

误差传递(续一)

由于 $E[x_i-\mu_i]=0$ 所以

$$E[y(\vec{x})]\approx y(\vec{\mu})$$

$$E[y^2(\vec{x})]\approx y^2(\vec{\mu})+2y(\vec{\mu})\cdot\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}}E[x_i-\mu_i]$$

$$+E\left[\left(\sum_{i=1}^{n}\left[\frac{\partial y}{\partial x_i}\right]_{\vec{x}=\vec{\mu}}(x_i-\mu_i)\right)\left(\sum_{j=1}^{n}\left[\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}}(x_j-\mu_j)\right)\right]$$

$$=y^2(\vec{\mu})+\sum_{i,j=1}^{n}\left[\frac{\partial y}{\partial x_i}\frac{\partial y}{\partial x_j}\right]_{\vec{x}=\vec{\mu}}V_{ij}$$

误差传递(续二)

两项合起来给出 $y(\vec{x})$ 的方差

$$\sigma_y^2 \approx \sum_{i,j=1}^{n} \left[\frac{\partial y}{\partial x_i} \frac{\partial y}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

如果 x_i 之间是无关的，则 $V_{ij} = \sigma_i^2 \delta_{ij}$ ，那么上式变为

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left[\frac{\partial y}{\partial x_i} \right]_{\vec{x}=\vec{\mu}}^2 \sigma_i^2$$

类似地，对于 m 组函数

$$\vec{y}(\vec{x}) = (y_1(\vec{x}), \ldots, y_m(\vec{x}))$$

误差传递(续三)

$$U_{kl} = \mathrm{cov}[y_k, y_l] \approx \sum_{i,j=1}^{n} \left[\frac{\partial y_k}{\partial x_i} \frac{\partial y_l}{\partial x_j} \right]_{\vec{x}=\vec{\mu}} V_{ij}$$

或者记为矩阵形式

$$U = AVA^T, \qquad A_{ij} = \left[\frac{\partial y_i}{\partial x_j} \right]_{\vec{x}=\vec{\mu}}$$

注意：上式只对$\vec{y}(\vec{x})$为线性时是精确的，近似程度在函数非线性区变化比σ_i要大时遭到很大的破坏。另外，上式并不需要知道x_i的 **p.d.f.** 具体形式，例如，它可以不是高斯的。

误差传递的一些特殊情况

$$y = x_1 + x_2 \qquad \sigma_y^2 = \sigma_1^2 + \sigma_2^2 + 2\,\mathrm{cov}[x_1, x_2]$$

$$y = x_1 x_2 \qquad \frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\mathrm{cov}[x_1, x_2]}{x_1 x_2}$$

注意在相关的情况下，最终的误差会有很大的改变，例如当

$$y = x_1 - x_2,\ \mu_1 = \mu_2 = 10,\ \sigma_1 = \sigma_2 = 1$$

$$\begin{cases} \rho = 0: & E[y] = \mu_1 - \mu_2 = 0,\ V[y] = 1^2 + 1^2 = 2,\ \sigma_y = 1.4 \\ \rho = 1: & E[y] = \mu_1 - \mu_2 = 0,\ V[y] = 1^2 + 1^2 - 2 = 0,\ \sigma_y = 0 \end{cases}$$

这种特征有时候是有益的：将公共的或难以估计的误差，通过适当的数学处理将它们消掉，达到减小误差的目的。

小结

1. 概率

a) 定义：柯尔莫哥洛夫公理 + 条件概率
b) 解释：频率或信心程度
c) 贝叶斯定理

2. 随机变量

a) 概率密度函数 **p.d.f.**
b) 累积分布函数
c) 联合，投影与条件的 **p.d.f.**

3. 随机变量函数

a) 函数自身也是随机变量
b) 几种方法找出 **p.d.f.**

4. 误差传递

函数方差的计算方法是基于一阶泰勒展开，只对线性方程精确。

习题

习题1.1:一束光子束流含有10^{-4}的电子.当它们通过一双层的探测器会给出无击中,单层击中或双层击中的信号.对于可能是电子和光子的概率为

$$P(0|e)=0.001 \qquad P(0|\gamma)=0.99899$$
$$P(1|e)=0.01 \qquad P(1|\gamma)=0.001$$
$$P(2|e)=0.989 \qquad P(2|\gamma)=10^{-5}$$

(a)求在只观测到一层击中的情况下,是光子的概率;

(b)求在观测到两层都击中的情况下,是电子的概率.

习题(续)

习题1.2:证明对于一个随机变量 x 和常数 α 与 β 有下式

$$E[\alpha x+\beta]=\alpha E[x]+\beta$$

$$V[\alpha x+\beta]=\alpha^2 V[x]$$

习题1.3:对于两个变量x与y的情况

(a)证明变量 αx+y 存在

$$V[\alpha x+y]=\alpha^2 V[x]+V[y]+2\alpha\,\mathrm{cov}[x,y]$$

$$=\alpha^2 V[x]+V[y]+2\alpha\rho\sigma_x\sigma_y$$

这里α是任意常数,

(b)利用结果(a),证明相关系数总是在[-1,1]之间.(利用变量V[αx+y]总是大于或等于零并考虑 $\alpha=\pm\sigma_y/\sigma_x$)