粒子物理与核物理实验中的数据分析

杨振伟
清华大学

第二讲：基本概念（续）

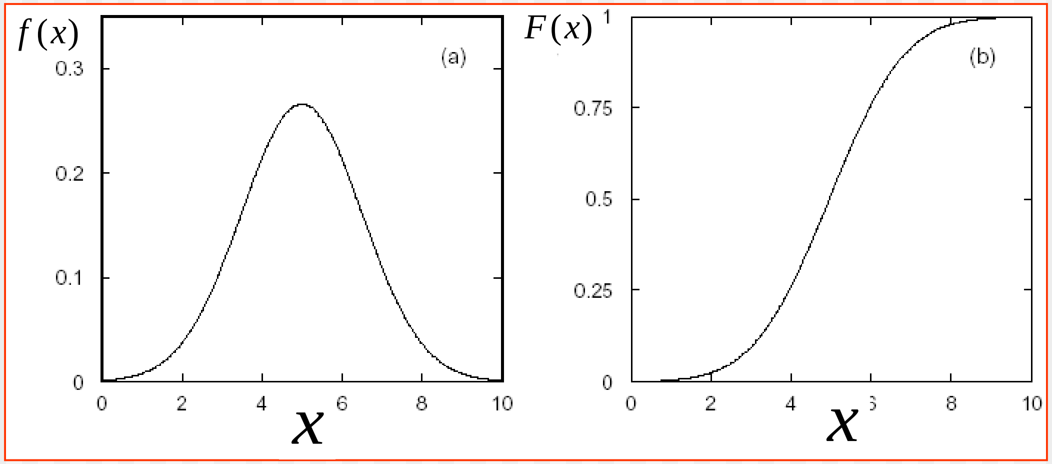
随机变量与概率密度函数

假设实验结果为 x（记作样本空间中元素)的概率为

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

那么概率密度函数 **p.d.f.** 定义为 $f(x)$，它对全部样本空间 S 满足

$$\int_S f(x)dx = 1$$



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

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

α 分位数、中值与模

分位点 x_α 定义为随机变量 x 的值，它使得

$$F(x_\alpha) = \alpha$$

这里 $0 \le \alpha \le 1$。因此可以容易求出分位点

$$x_\alpha = F^{-1}(\alpha)$$

随机变量 x 的中值定义为

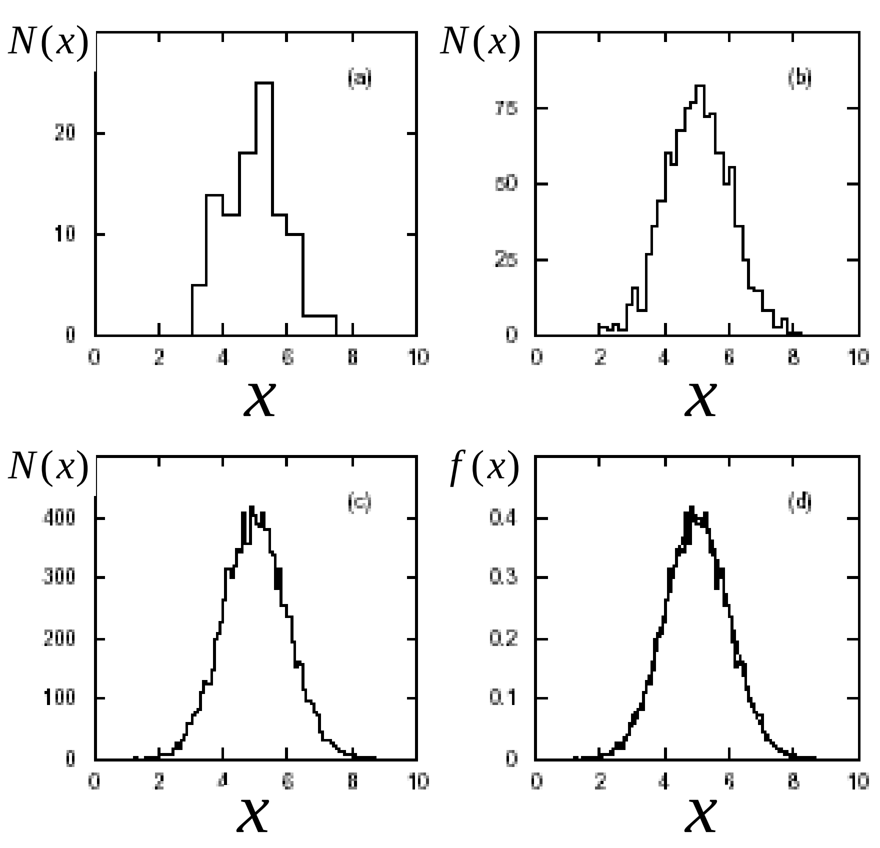
$$x_{1/2} = F^{-1}(1/2)$$

随机变量 x 被观测到大于或小于中值的概率是相等的。

模定义为使概率密度函数值达到极大的随机变量值。

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)

$n =$ 填入直方图的总事例数

$\Delta x =$ 区间的宽度

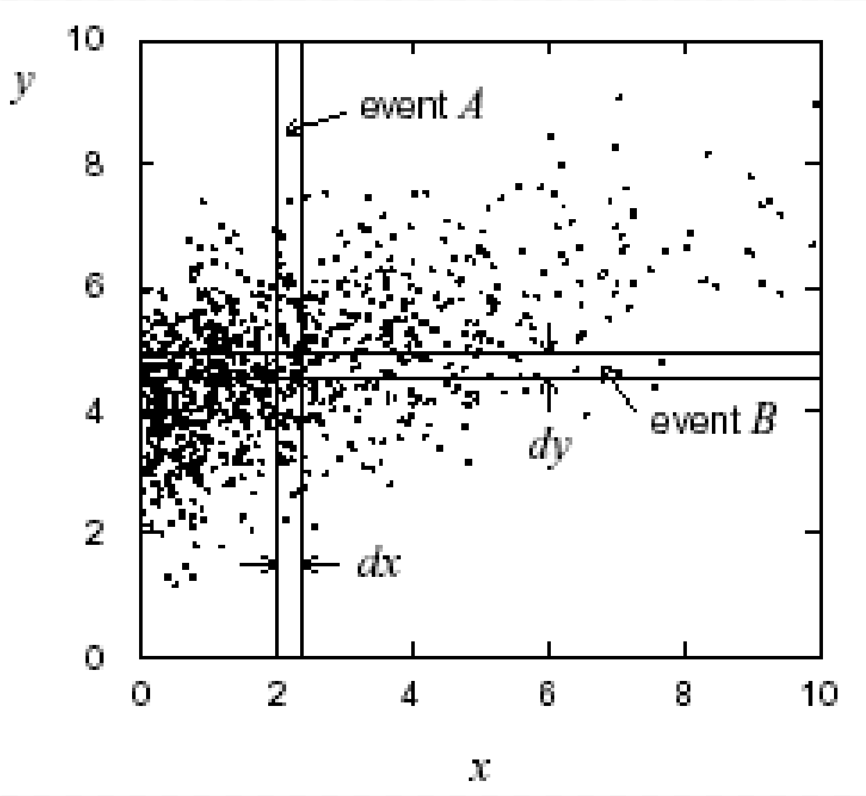
$n \to \infty, \Delta x \to 0$

$n \cdot \Delta x$有限

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

观测量大于一个，例如 x 与 y



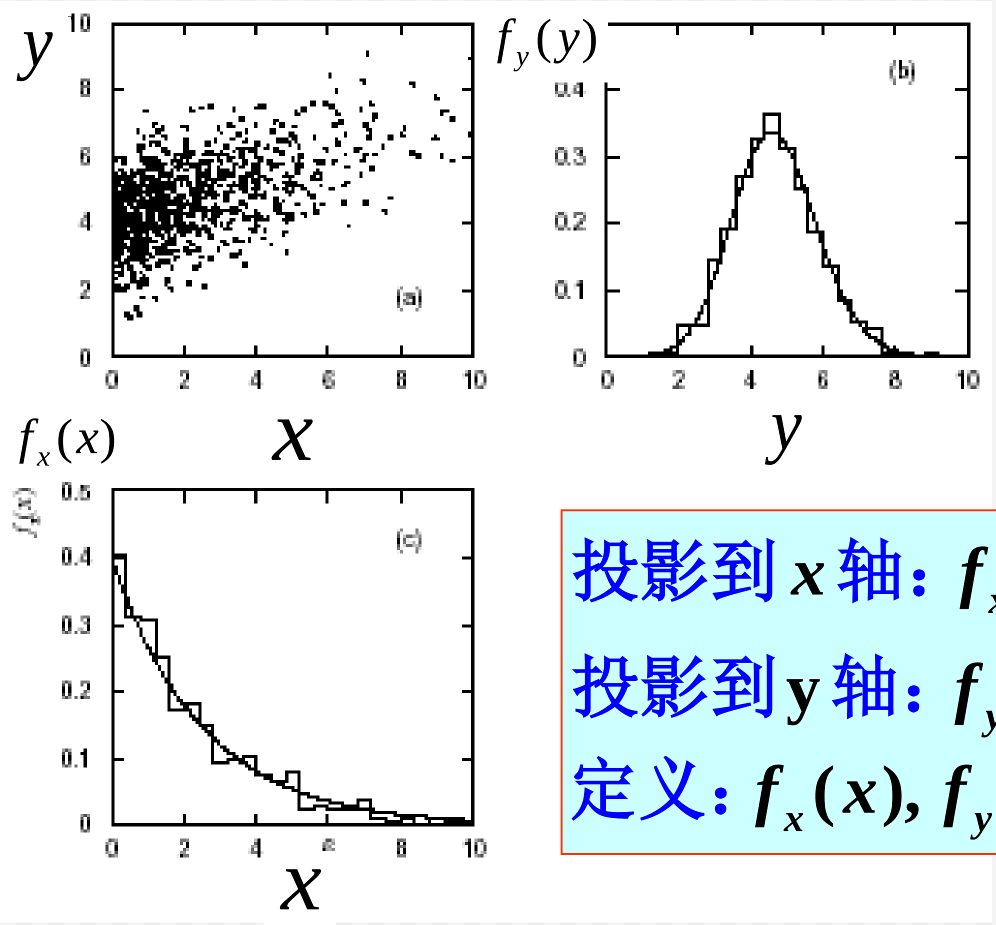
$$P(A\cap B)=f(x,y)dxdy$$

$$f(x,y)=\text{联合的p.d.f.}$$

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

边缘分布

将联合概率密度函数 **p.d.f.** 分别投影到 x 与 y 轴



投影到 x 轴：$f_x(x) = \int f(x,y)dy$

投影到 y 轴：$f_y(y) = \int f(x,y)dx$

定义：$f_x(x), f_y(y) =$ 边缘的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

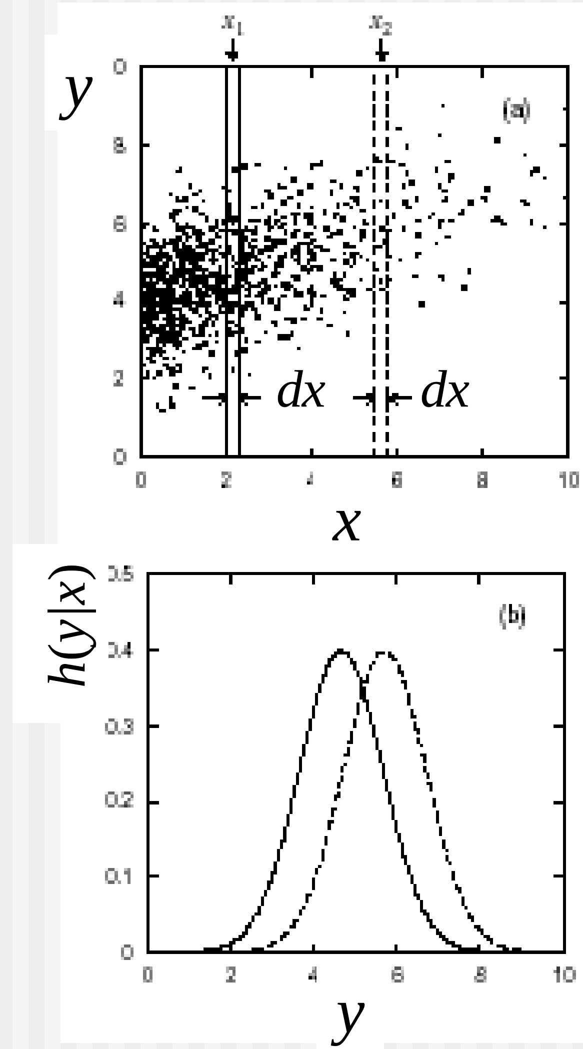
$$h(y|x)=\frac{f(x,y)}{f_x(x)},\quad g(x|y)=\frac{f(x,y)}{f_y(y)}$$

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，则可构造**2-维p.d.f**

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



名词总汇

随机事例

概率　　相对频率与主观概率

条件概率　　贝叶斯定理

随机变量

概率密度函数

条件密度函数

直方图

提醒：概率都是条件概率

由柯尔莫哥洛夫公理，我们定义了概率 ***P*(*A*)**。

但在实际应用中，我们总是对 ***A*** 相对于许多样本空间的概率感兴趣，而不仅仅只是一个空间。因此，通常以记号

$$P(A|S)$$

来表示所进行的研究是在特定的样本空间 ***S*** 中，也就是 ***A*** 相对于 ***S*** 的条件概率。

因此，所有概率在实际应用中都是条件概率。

只有当 ***S*** 的选择是明白无误时，才能简单记为

$$P(A|S) \Rightarrow P(A)$$

证明举例：事例与逆事例

如果 A 是在 S 中的任意一个事例，则

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

证明：由于 A 与 $\bar{A}$ 根据定义是互斥的，并且从文恩图得到

$$A \cup \bar{A} = S$$

因此可以写出

$$\begin{aligned} P(A) + P(\bar{A}) &= P(A \cup \bar{A}) \\ &= P(S) \\ &= 1 \end{aligned}$$

$\Rightarrow$

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

举例：检查给定概率的合理性

如果一个实验有三种可能并且互斥的结果 A，B 和 C ，检查下列各种情况给出的概率值是否是合理的：

1) $P(A)=1/3, P(B)=1/3, P(C)=1/3$

2) $P(A)=0.64, P(B)=0.38, P(C)=-0.02$

3) $P(A)=0.35, P(B)=0.52, P(C)=0.26$

4) $P(A)=0.57, P(B)=0.24, P(C)=0.19$

结论：只有1）与4）是合理的。

评论：作为一个合格的实验研究人员，一定要具备判断结果是否合理的能力！

举例：检查经验概率密度函数

实验上经常经验性地从直方图中给出概率密度函数（例如通过拟合直方图分布等等），但是需要确定得到的函数是否满足概率密度函数的定义，例如

$$1)\ f(x)=\frac{x-2}{2} \quad \text{对于}\ x=1,2,3,4$$

$$2)\ h(x)=\frac{x^2}{25} \quad \text{对于}\ x=0,1,2,3,4$$

试判断哪一个可以用作概率密度函数？

答案：1）有负概率值；2）累积函数值大于1。因此，两者在给定的随机变量范围内都不能用作概率密度函数。

数据分析中的问题

粒子与核物理实验中对动量的测量通常是分别测量

$$p_{xy} \qquad p_z \qquad f(p_{xy}, p_z)$$

在已知两分量测量值的概率密度函数情况下，总动量为

$$p = \sqrt{p_{xy}^{\ 2} + p_z^{\ 2}}$$

如何导出总动量的测量值的概率密度函数?

$$g(p)$$

是研究随机变量函数的**p.d.f**问题。

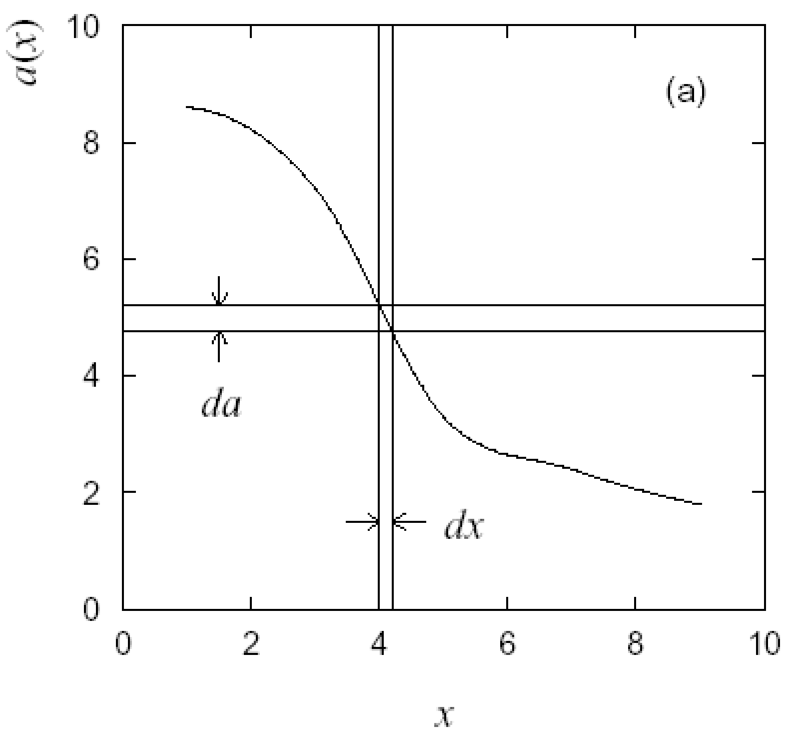
一维随机变量的函数

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？



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

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

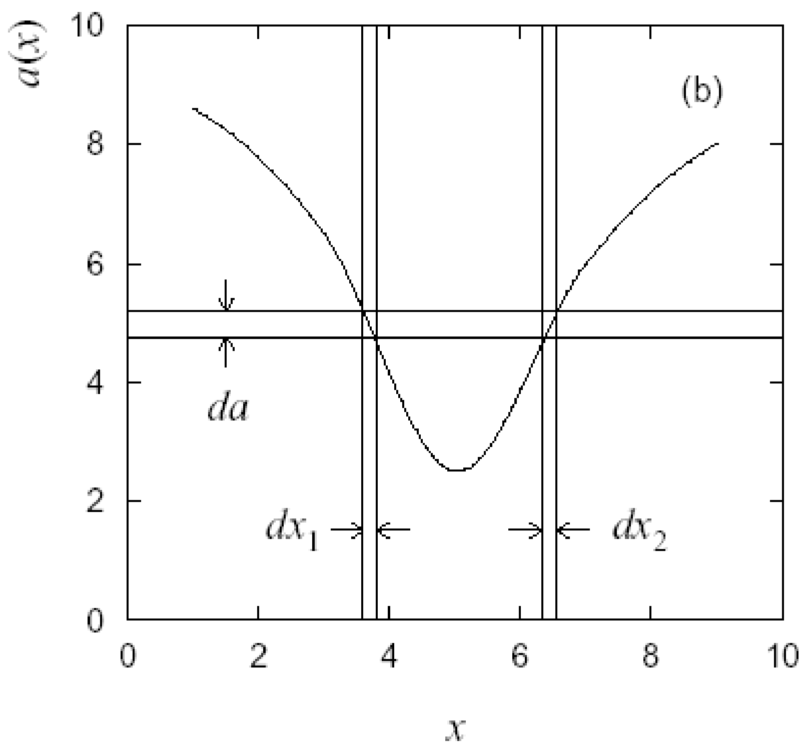
$$g(a)da = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right|$$

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

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

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

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

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

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

多维随机变量的函数

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

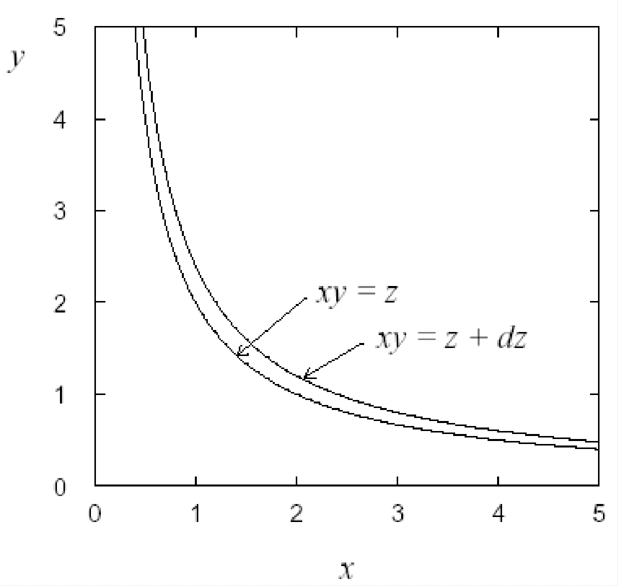
$$g(a')da'=\int\cdots\int_{dS} f(x_1,\ldots,x_n)dx_1\ldots dx_n$$

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

如果两个独立变量 x 与 y，分别按 $g(x)$ 与 $h(y)$ 分布，那么函数 $z=xy$ 应具有何种形式？

$$f(x,y)=g(x)h(y)$$

$$f(z)dz=\iint_{dS} f(x,y)dxdy$$
$$=\iint_{dS} g(x)h(y)dxdy$$
$$=\int_{-\infty}^{\infty} g(x)dx\int_{z/|x|}^{(z+dz)/|x|} h(y)dy$$



多维随机变量的函数(续一)

$$f(z)=\int_{-\infty}^{\infty} g(x)h(\frac{z}{x})\frac{dx}{|x|}=\int_{-\infty}^{\infty} g(\frac{z}{y})h(y)\frac{dy}{|y|}$$

记作 g 与 h 的Mellin卷积 ➡ $f=g\otimes h$

如果函数为 $z=x+y$，则应具有何种形式?

$$f(z)=\int_{-\infty}^{\infty} g(x)h(z-x)dx=\int_{-\infty}^{\infty} g(z-y)h(y)dy$$

记作 g 与 h 的傅立叶卷积 ➡ $f=g\otimes h$

注意：通常将两者皆称为 g 与 h 的卷积，已相同记号表示。

多维随机变量的函数(续二)

考虑具有联合 **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:$\vec{a}(\vec{x})=(a_1(\vec{x}),\ldots,a_n(\vec{x}))$，而且其逆函数 $x_1(\vec{a}),\ldots,x_n(\vec{a})$存在。那么$\vec{a}$ 的联合 **p.d.f.** 为

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

这里 J 是雅可比行列式

$$J=\begin{vmatrix} \frac{\partial x_1}{\partial a_1} & \frac{\partial x_1}{\partial a_2} & \cdots & \frac{\partial x_1}{\partial a_n} \\ \frac{\partial x_2}{\partial a_1} & \frac{\partial x_2}{\partial a_2} & \cdots & \frac{\partial x_2}{\partial a_n} \\ \vdots & & & \vdots \\ & & \cdots & \frac{\partial x_n}{\partial a_n} \end{vmatrix}$$

➡ 任意一个函数 $g_i(a_i)$ 均可通过对函数 $g(\vec{a})$ 积分掉其它不用的变量而得到。是数据处理中误差传递的基础。

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。

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

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

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

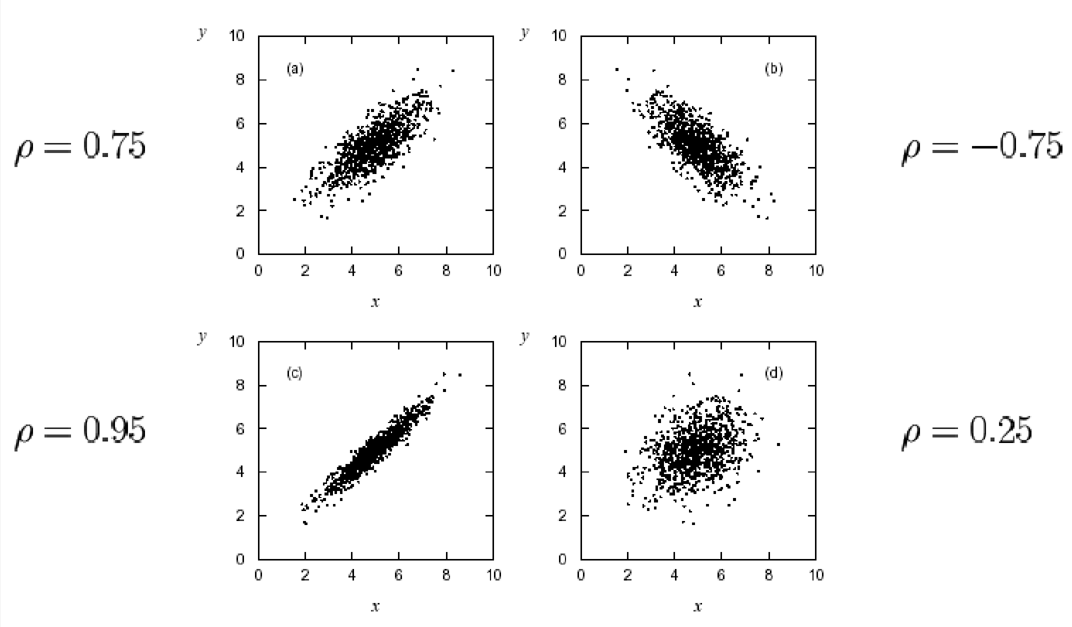
定义协方差 $\mathrm{cov}[x,y]$ (也可用矩阵表示 V_{xy})为

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

关联系数定义为

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

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



如果 **x**, **y** 独立，即

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

则

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

举例：样本平均值

假设实验上研究一核素衰变寿命，在探测效率为100%的情况下，每次探测到的寿命为 t_i，一共测量了 n 次，求平均寿命（也就是寿命的期待值）。

根据离散型期待值的定义 $$E[t]=\sum_{i=1}^{n}t_iP(t_i)$$

问题的关键是 t_i 的概率密度函数是什么？

根据概率的相对频率定义，在 n 次测量中出现 t_i 频率为一次

$$P(t_i)=\frac{1}{n}$$

因此，期待值（或平均寿命）为 $$E[t]=\sum_{i=1}^{n}t_i\frac{1}{n}=\frac{1}{n}\sum_{i=1}^{n}t_i$$

思考：如果频率为 m_i 次，结果会不同吗？

不确定度传递

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

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

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 $E[y^2]$…

不确定度传递(续一)

由于 $E[x_i-\mu_i]=0$ 所以利用泰勒展开式可求

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

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

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

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

不确定度传递(续二)

两项合起来给出 $y(\vec{x})$ 的方差

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

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

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

类似地，对于 m 组函数

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

不确定度传递(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。

不确定度传递的一些特殊情况

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

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

注意在相关的情况下，最终的误差会有很大的改变，例如当

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

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

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

坐标变换下的不确定度矩阵

实验上经常通过测量粒子在探测器中各点的击中坐标（x, y）来拟合在极坐标下的径迹（r, θ）。通常情况下，（x, y）的测量是不关联的。

$$r^2 = x^2 + y^2$$
$$\tan\theta = y / x$$

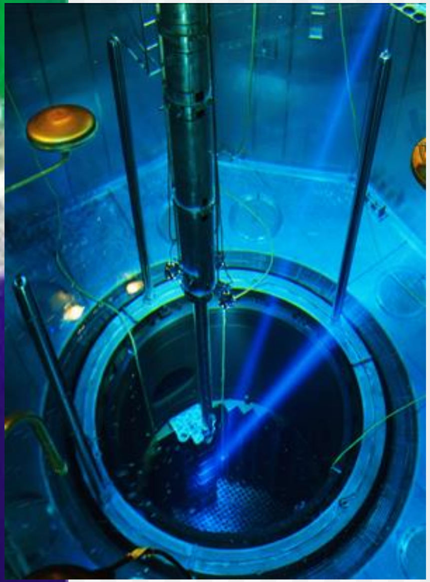
由于

$$U(r,\theta) = AV(x,y)A^T$$

因此，坐标变换后的不确定度矩阵为

$$\begin{pmatrix} \sigma_r^2 & \mathrm{cov}(r,\theta) \\ \mathrm{cov}(r,\theta) & \sigma_\theta^2 \end{pmatrix} = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ -\frac{y}{r^2} & \frac{x}{r^2} \end{pmatrix} \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} \frac{x}{r} & -\frac{y}{r^2} \\ \frac{y}{r} & \frac{x}{r^2} \end{pmatrix} = \frac{1}{r^2} \begin{pmatrix} x^2\sigma_x^2 + y^2\sigma_y^2 & \frac{xy}{r}(\sigma_y^2 - \sigma_x^2) \\ \frac{xy}{r}(\sigma_y^2 - \sigma_x^2) & \frac{1}{r^2}(y^2\sigma_x^2 + x^2\sigma_y^2) \end{pmatrix}$$

大亚湾反应堆中微子实验



反应堆中微子

- 反应堆能产生大量反电子型中微子

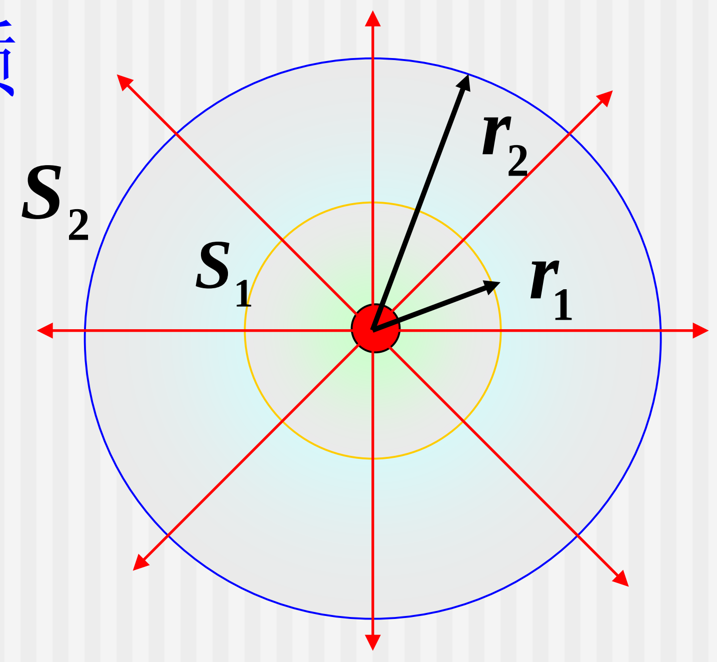
3 GW 热功率反应堆 $n \rightarrow p + e^- + \bar{\nu}_e$

$\Rightarrow$ 6×10^{20} 个反电子中微子/秒

- 中微子几乎无损穿透物质

假设产生的中微子以球面波传播，那么在任一地方任一给定面元的中微子流强为

$$\Delta I_r = \frac{\Delta S}{4\pi r^2}\cdot I$$



大亚湾中微子振荡

- 中微子振荡

  中微子在运动过程中自己不断改变形态

- 测量中微子形态随运动距离的改变

$$\Delta I_{r_1} = \frac{\Delta S}{4\pi r_1^2} \cdot I \quad \Rightarrow \quad \Delta I_{r_2} = \frac{\Delta S}{4\pi r_2^2} \cdot I$$

- 中微子形态随运动距离的改变理论预言

$$\Delta I_r \sim \frac{\Delta S}{4\pi r^2} \cdot I \cdot P(\bar{\nu}_e \to \bar{\nu}_e)$$

$$= \frac{\Delta S}{4\pi r^2} \cdot I \cdot f(\Delta m, \sin\theta_{13}) \cdot \sigma_{截面} \cdot \varepsilon_{效率}$$

如何保证1%精度？

- 测量中微子振荡的影响

方案1： ΔI_r

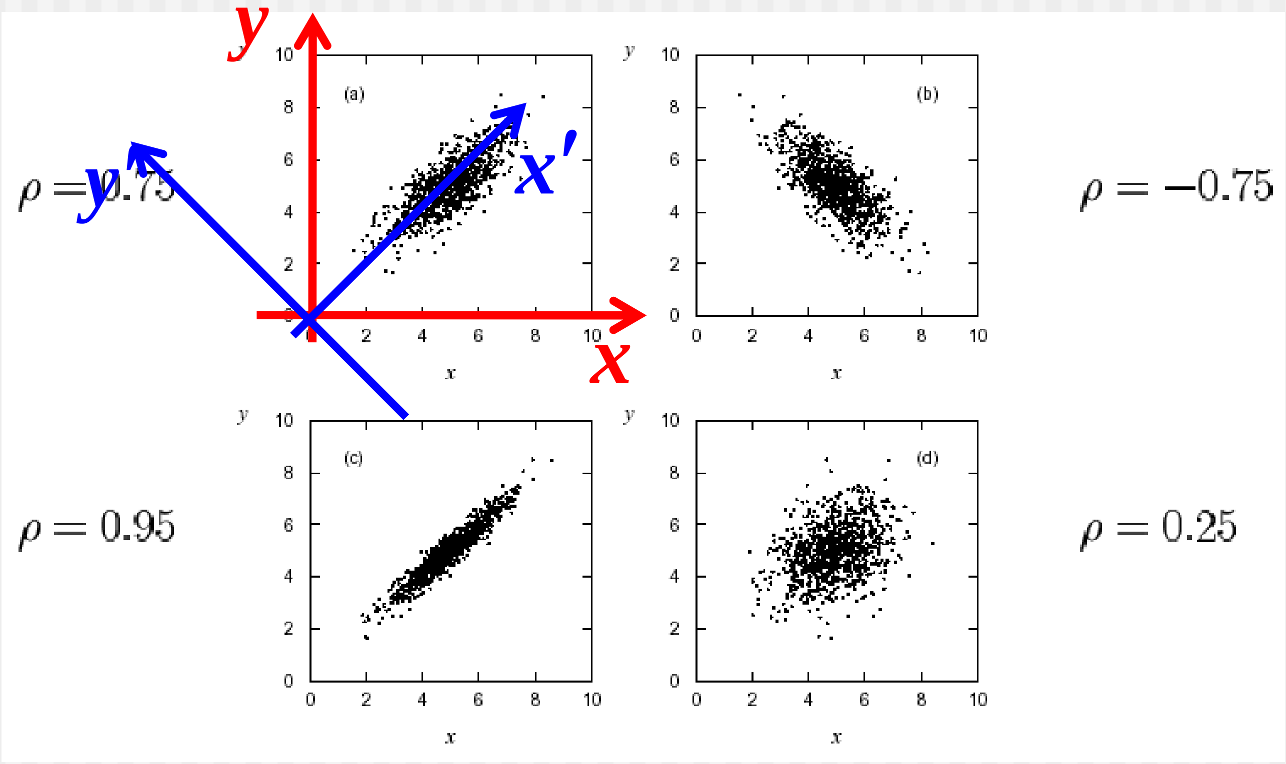
方案2: $\dfrac{\Delta I_2}{\Delta I_1}$

$$\Delta I_r \sim \frac{\Delta S}{4\pi r^2} \cdot I \cdot f(\Delta m, \sin\theta_{13}) \cdot \sigma_{\text{截面}} \cdot \varepsilon_{\text{效率}}$$

那一种方案更易实现**1%**精度的测量？为什么？

不同坐标系下相关性的变化

通过转动坐标，随机变量的相关性会发生改变。



显然，通过将坐标系转动 45^0，上面的相关性在新坐标系下消失。

随机变量作正则变换去除相关性

假设有 n 个随机变量 $x_1,\ldots,x_n$ 以及协方差矩阵$V_{ij}=\mathrm{cov}[x_i, x_j]$，可以证明有可能通过线性变换重新定义 n 个新的变量 $y_1,\ldots,y_n$ 使得对应的协方差矩阵$U_{ij}=\mathrm{cov}[y_i, y_j]$非对角元为零。令

$$y_i = \sum_{j=1}^{n} A_{ij} x_j$$

对应的协方差矩阵为

$$\begin{aligned} U_{ij} &= \mathrm{cov}[y_i, y_j] \\ &= \mathrm{cov}\left[\sum_{k=1}^{n} A_{ik} x_k, \sum_{k=1}^{n} A_{jl} x_l\right] \\ &= \sum_{k,l=1}^{n} A_{ik} A_{jl}\, \mathrm{cov}[x_k, x_l] \\ &= \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^{T} \end{aligned}$$

$\Longleftrightarrow$

非线性情况

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

变换后的变量协方差矩阵对角化

为了使协方差矩阵 U 对角化

$$U = AVA^T$$

可先确定协方差矩阵 V 的本征列矢量 $\vec{r}^{\,i}$，$i=1,\ldots,n$。解方程

$$V\vec{r}^{\,i} = \lambda_i \vec{r}^{\,i} \quad \text{或} \quad V_{kl} r_l^i = \lambda_i r_k^i$$

由于协方差矩阵总是对称的，因此可知本征矢量是正交的

$$\vec{r}^{\,i} \cdot \vec{r}^{\,j} = \sum_{k=1}^{n} r_k^i r_k^j = \delta_{ij}$$

变换矩阵 A 由本征矢量 $\vec{r}$ 给出，即

$$A_{ij} = r_j^i, \quad A_{ij}^T = r_i^j, \quad \sum_{j=1}^{n} A_{ij} A_{jk}^T = \sum_{j=1}^{n} r_j^i r_j^k = \vec{r}^{\,i} \cdot \vec{r}^{\,k} = \delta_{ik}$$

正则变换后变量的协方差矩阵

因此，正则变换的协方差矩阵为

$$U_{ij} = \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^{T}$$

$$= \sum_{k,l=1}^{n} r_k^i V_{kl} r_l^j$$

$$= \sum_{k=1}^{n} r_k^i \lambda_j r_k^j$$

$$= \lambda_j \vec{r}^{\,i} \cdot \vec{r}^{\,j}$$

$$= \lambda_j \delta_{ij}$$

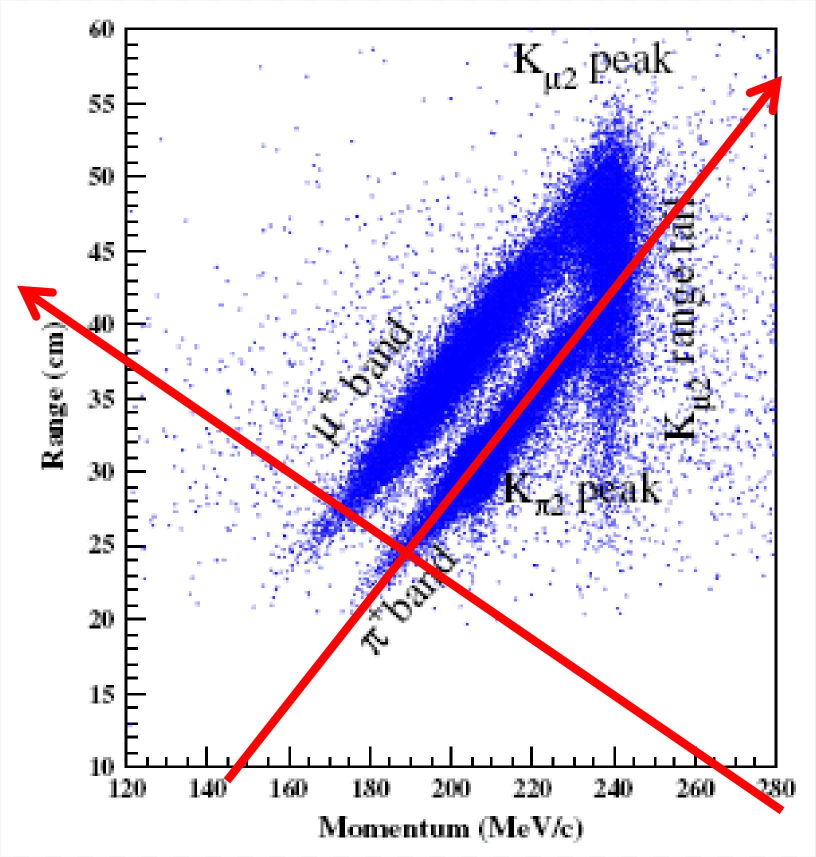
变量作正则变换后，其方差由原协方差矩阵 V 的本征值给出。

对应于矢量的转动不改变模的大小。

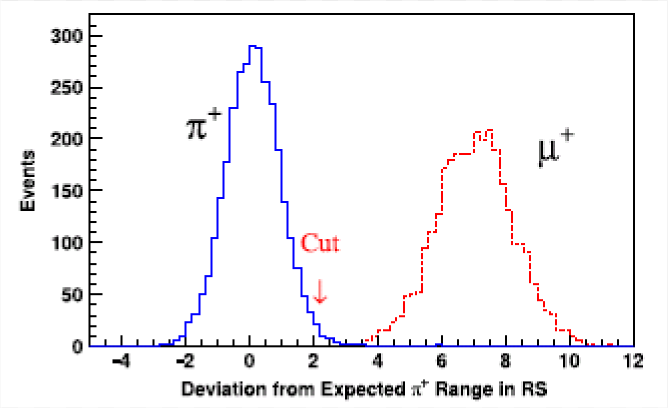
$|y|^2 = y^T y = x^T A^T A x = |x|^2$

尽管非关联变量经常容易处理，但是对经过变换的变量的理解不一定容易。

带电粒子在闪烁体的射程



PHYSICAL REVIEW D 77, 052003 (2008)



在原来的定义下，可以得到粒子射程随动量大小的变化关系。通过转动变换，粒子的射程与动量发生了改变，无物理含义，但是提供了一个很好的粒子类型甄别变量。

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，边缘与条件的 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. 不确定度传递

随机变量的函数的方差，计算方法是基于一阶泰勒展开，只对线性方程精确。