粒子物理与核物理实验中的数据分析

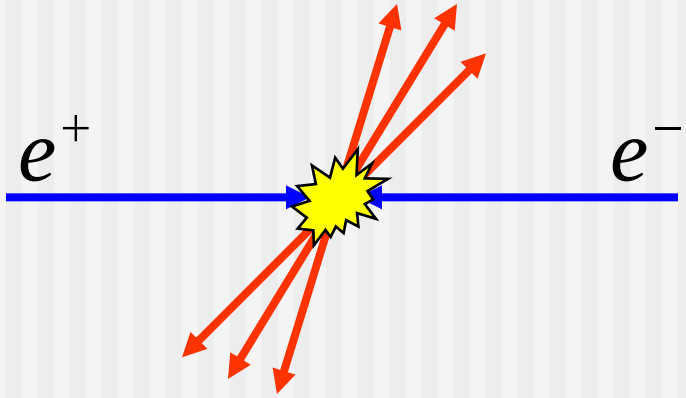
陈少敏
清华大学

第一讲：基本概念

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

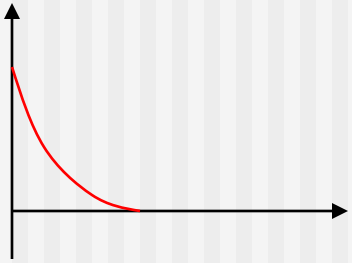
实验测量

给出每个事例的特征量(能动量，末态粒子数...)。

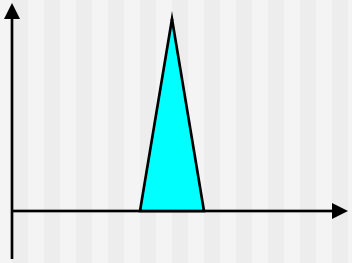
理论预言

给出上述各特征量的分布，而且可能还会包含自由参数。

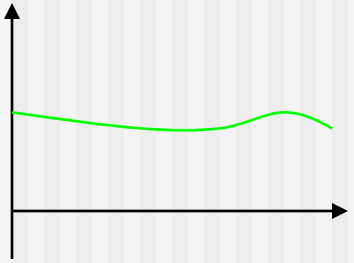
数据背后的物理图像是什么？



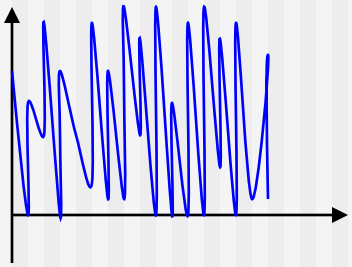
原初物理



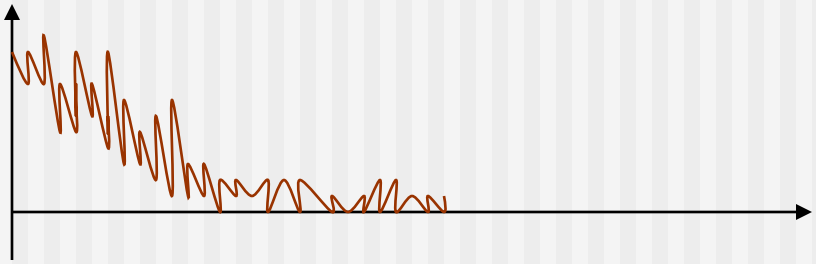
分辨率



探测效率



本底噪音



实验数据

数据分析专业术语：

事例选择，粒子鉴别，CUT条件，信噪比优化，无偏选择，效率修正，卷积分辨率，解谱（像）还原...

如何科学地给出物理结论？

收集数据

↓

数据分析

↓

估计参数值与相应的误差范围，检验在何种程度上理论与实验数据相符。

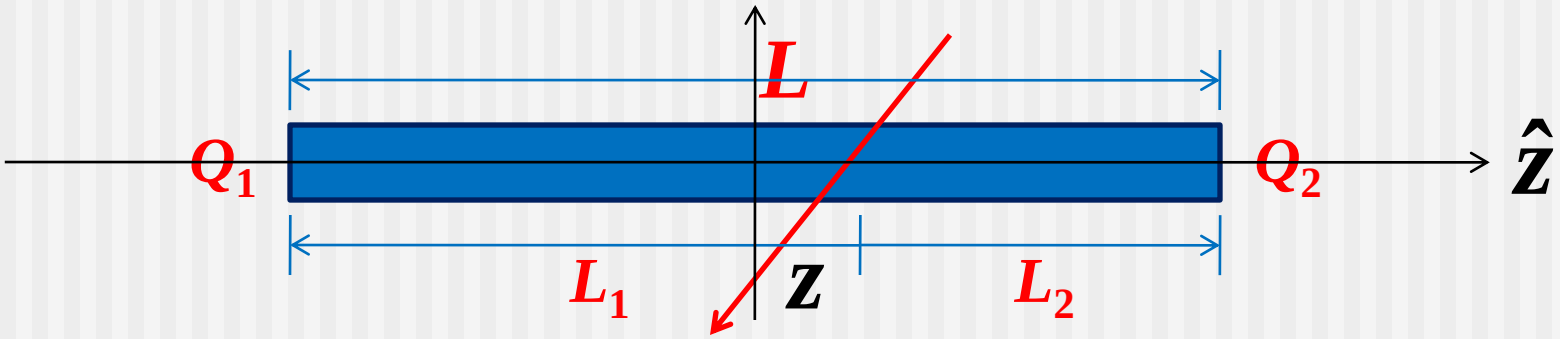
问题：如何评价这种检验？

举例：测量闪烁体衰减长度

光在闪烁体中传播时，具有下列衰减关系

$$Q = Q_0 \exp(-L / L_0)$$

其中，L_0 是闪烁体的衰减长度，它是表征闪烁体质量的一项重要指标。实验上测量衰减长度的方法如下图所示



$$Q_0 \propto E, \quad Q_1 = 0.5Q_0 \exp(-L_1 / L_0), \quad Q_2 = 0.5Q_0 \exp(-L_2 / L_0)$$

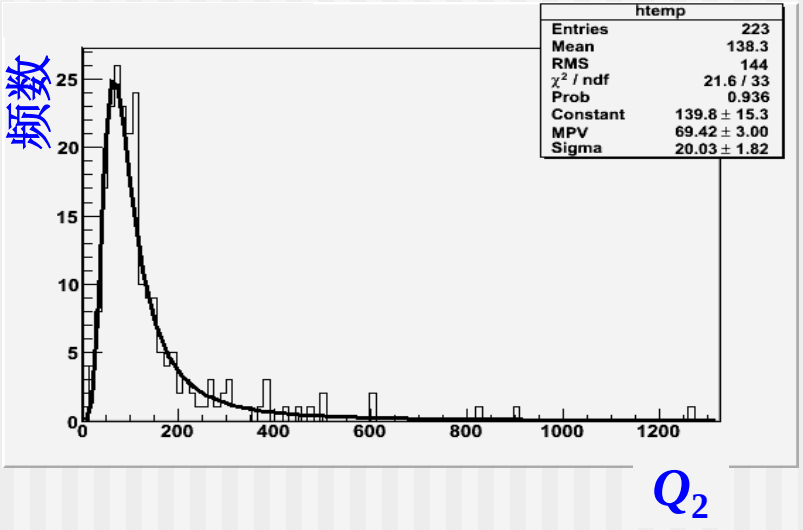
$$L_1 = 0.5L + z, \quad L_2 = 0.5L - z,$$

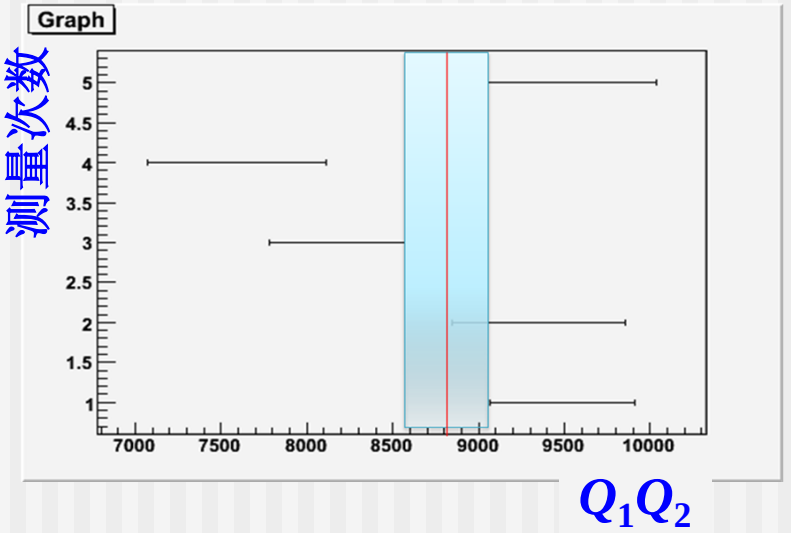
$$Q_1 Q_2 = 0.25Q_0^2 \exp(-L / L_0), \quad L_0 = 2z \ln(Q_1 / Q_2)$$

举例：测量闪烁体衰减长度（续）

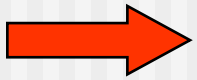
$$Q_1 Q_2 = 0.25 Q_0^2 \exp(-L / L_0), \quad L_0 = 2z \ln(Q_1 / Q_2)$$

实验采用恒定光源，因此 Q_0 为常数，对待测闪烁体 L_0 也为常数。理论上只要在给定一个位置 z，测量闪烁体两端的电荷输出量即可。但在实际中，往往需要做多点测量。





理论上是不变的 Q_1Q_2值，为什么每次测量都不相同？能否认为 L_0不是常数？



使用概率来量化结论！

随机事例

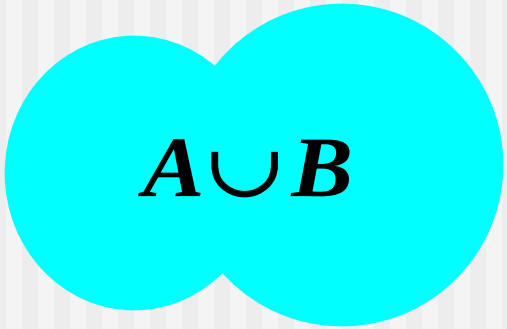
在一定的实验条件下，现象 A 可能发生，也可能不发生，并且只有发生或不发生这样两种可能性，这是偶然现象中一种比较简单的形态，我们把发生了现象 A 的事例称为随机事例 A，简称事例 A。

随机事例之间的相互关系

A 与 B 之并事例 $A \cup B$

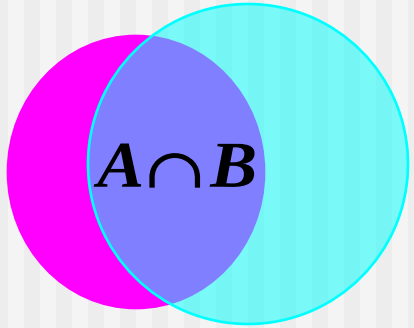
指事例 A 与 B 中至少有一个出现的事例

如果 A与 B 互斥，则 $A \cup B = A + B$



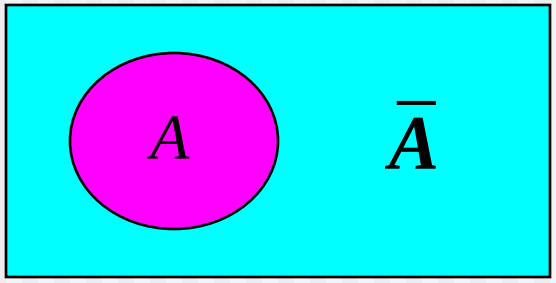
A 与 B 之积(交)事例 $A \cap B$

指事例 A 与 B 中同时出现的事例

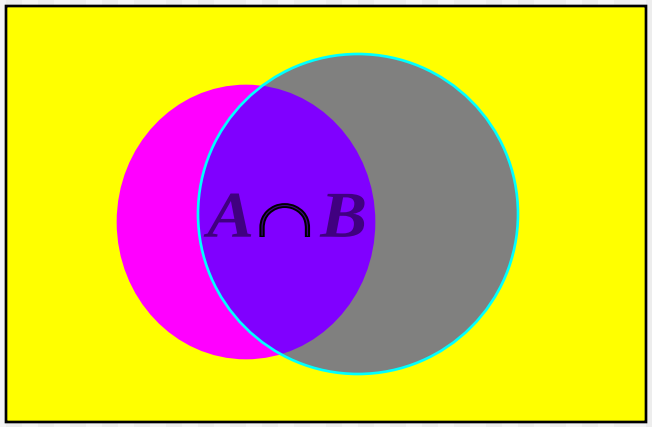


A 之逆事例 $\overline{A}$

指事例 A 不出现的事例 $A \cap \overline{A} = 0$



文恩图（Venn diagram）检验



$$\overline{A \cap B} = \overline{A} \cup \overline{B}$$

$$A \cup (A \cap B) = A$$

$$(A \cap B) \cup (A \cap \overline{B}) = A$$

$$A \cup B = (A \cap B) \cup (A \cap \overline{B}) \cup (\overline{A} \cap B)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

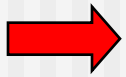
概率的定义

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

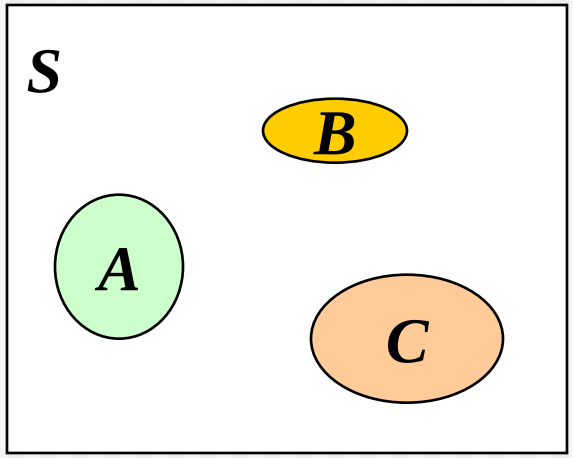
从该公理与文恩图给出的结论可以导出下列概率公式

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

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

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

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



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

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

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

$$P(A|B)=\frac{P(A)P(B)}{P(B)}=P(A)$$ ⇒ 结果与 ***B*** 无关

注意：与不相交的子集定义不同 $A\cap B$

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

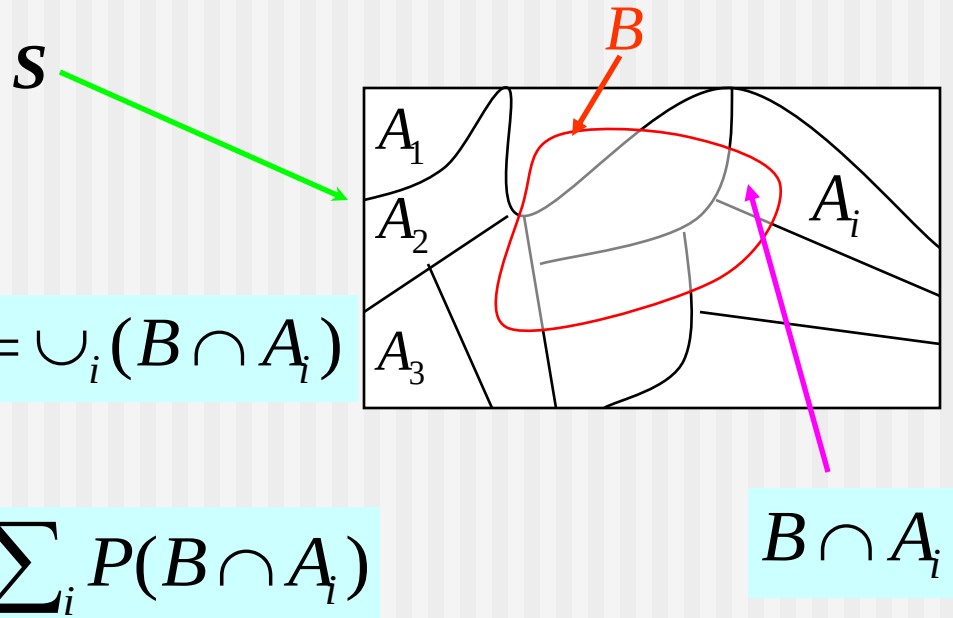
$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

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

全概率事例与贝叶斯定理

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 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

$P(+ \mid AIDS) = 0.98$　　AIDS感染患者阳性的概率

$P(- \mid AIDS) = 0.02$　　AIDS感染患者阴性的概率

$P(+ \mid no\ AIDS) = 0.03$　　AIDS未感染者阳性的概率

$P(- \mid no\ AIDS) = 0.97$　　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**。

涉及到如何诠释结果（概率）的问题！

概率含义的诠释

- **相对频率（频率论者）**

  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（贝叶斯论者）**

  如果A，B，…是假设(是真或是假的各种陈述)，那么概率

$$P(A)=\text{对 }A\text{ 为真的信心程度}$$

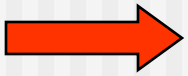
✓两种解释皆与柯尔莫哥洛夫公理相符。

✓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。

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

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

先验概率: $P(\text{理论})$； 验后概率: $P(\text{理论}\mid\text{实验})$

似然性: $P(\text{实验}\mid\text{理论})$

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$P(\text{实验}\mid\text{理论})=0$，则表明理论不能接受。
- 大的$P(\text{实验}\mid\text{理论})$会增加对理论的信任度。
- 通过实验结果可以修改 $P(\text{理论})$。
- 改进的$P(\text{理论})$可应用于对重复实验结果的预测。
- $P(\text{实验}\mid\text{理论})$对先验理论的依赖将最终消失。

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 < 938.27211 MeV)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2010年足球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%
- 大陆中国人2020年获诺贝尔奖的概率为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。

艾滋病检验结果再认识

$$P(AIDS) = 0.001 \quad \text{(验前概率)}$$

$$P(AIDS \mid +) = 0.032 \quad \text{(验后概率)}$$

对于个人而言，0.032 是主观概率。如果没有其它额外的信息时，应把 0.001 当作相对频率解释。但是往往在病毒检验前，该相对频率被当作一种信念来处理个人是否患病。

如果还有其它额外的信息，应该给出不同的先验概率。这种贝叶斯统计的特点必定是主观的。例如，受检者有过吸毒历史。一旦验前概率改变，贝叶斯定理就会告诉患病的可能性。对阳性结果的诠释就会改变。

问题：能否构造含自变量的概率？

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）

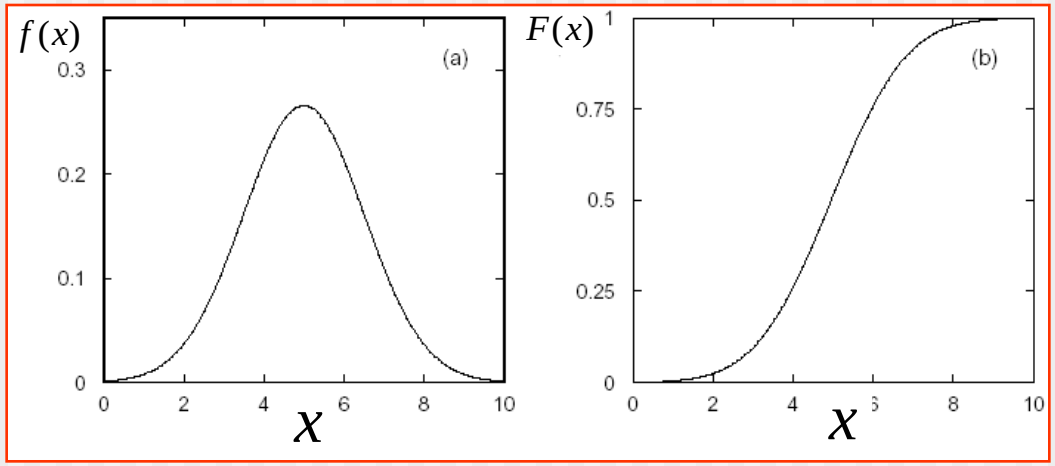
$$P(\text{观测到}x\text{在}[x, x+dx]\text{范围内}) = f(x)dx$$

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

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

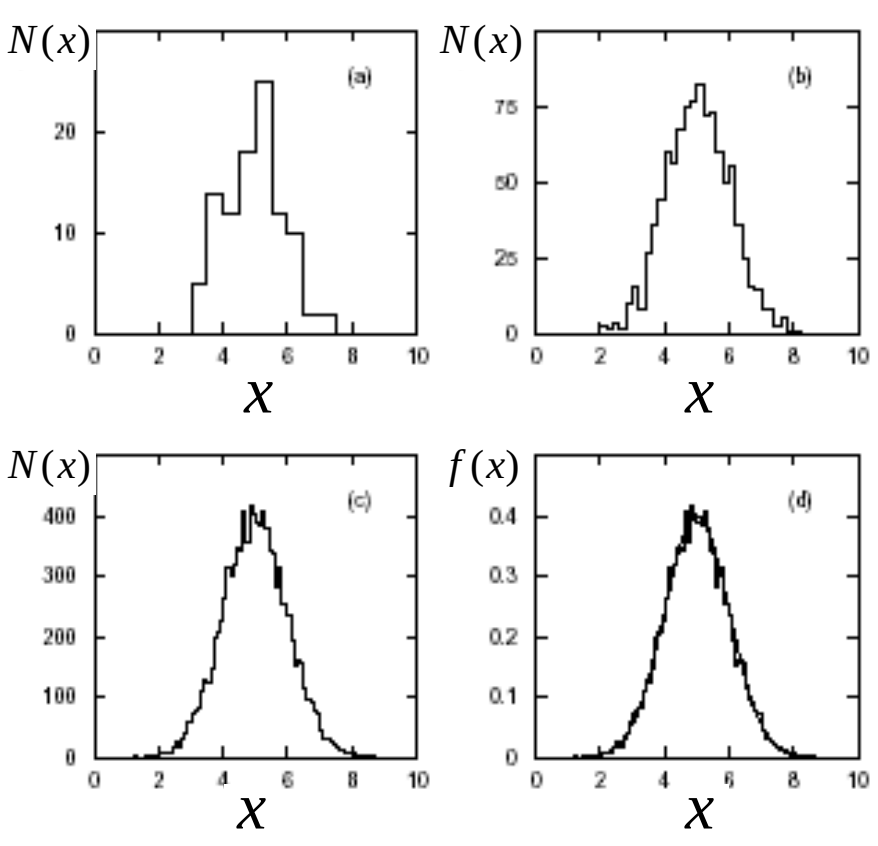


对于离散型随机变量

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

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)

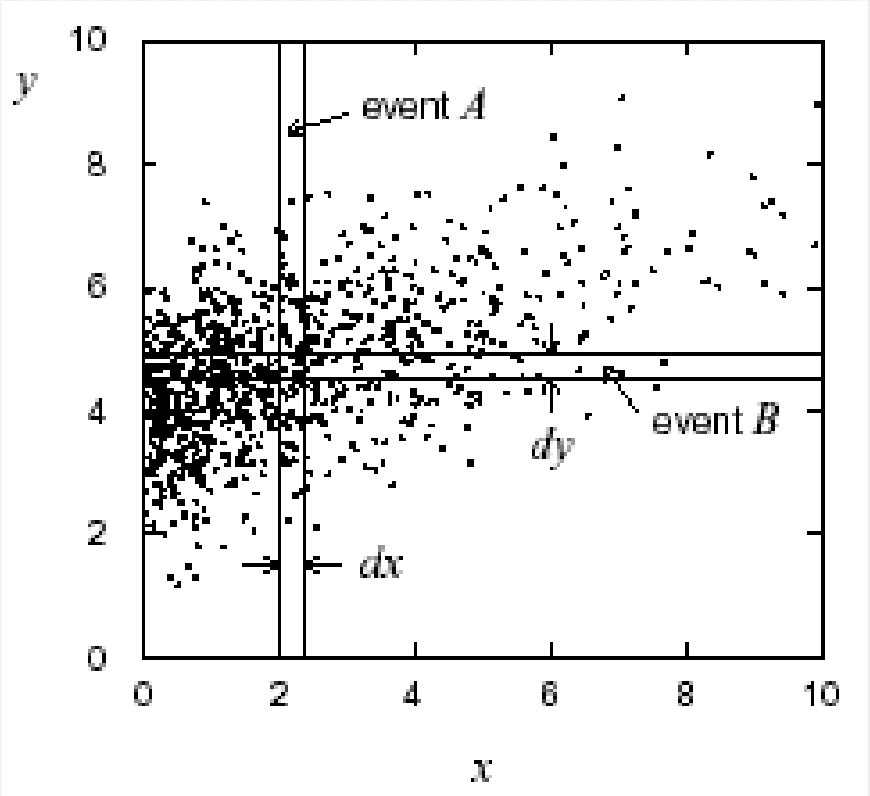
$n=$添入直方图的总事例数

$\Delta x=$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



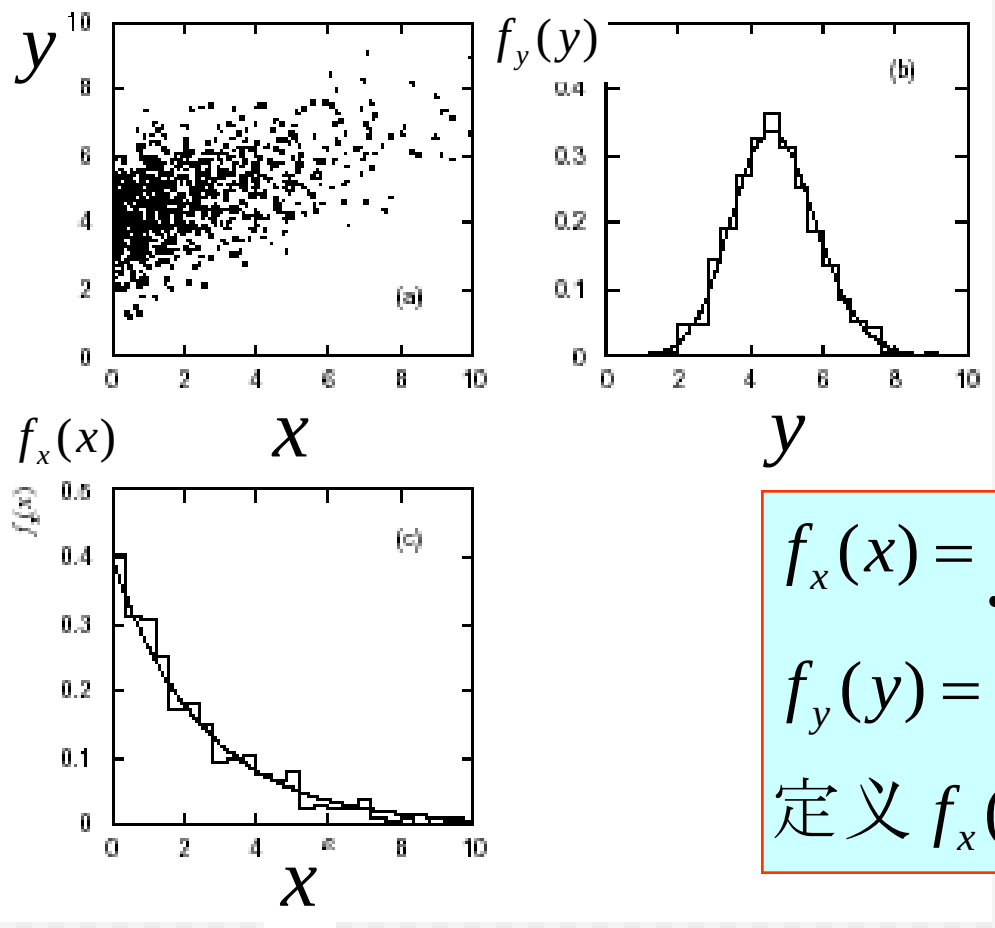
$$P(A \cap B) = f(x,y)dxdy$$

$$f(x,y) = \text{联合的 p.d.f.}$$

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

边缘分布

将联合概率密度函数 **p.d.f.** 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 $f_x(x), f_y(y)$ = 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

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

定义条件概率的密度函数 **p.d.f.** 为

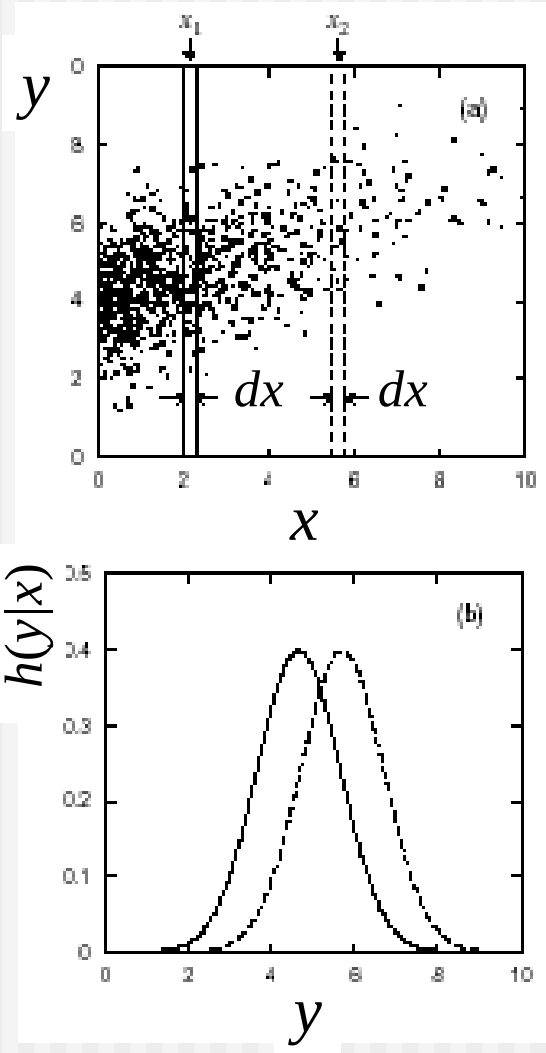
$$h(y\mid x)=\frac{f(x,y)}{f_x(x)},\quad g(x\mid y)=\frac{f(x,y)}{f_y(y)}$$

则贝叶斯定理可写为

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

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

问题

条件概率

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

如果 **A** 与 **B** 相互独立，则从文恩图上得到

$$A\cap B=0$$

因此

$$P(A\cap B)=0 \quad \Rightarrow \quad P(A|B)=\frac{P(A\cap B)}{P(B)}\neq P(A) \quad \Rightarrow \quad =0 \quad ???$$

解答：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

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

解答：互斥与相互独立

互斥的定义为

$$A \cup B = A + B$$

也就是两个事例的定义没有交集。所给出的推论为

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

相互独立的定义为

如果 $P(A \cap B) = P(A)P(B)$ 则 A 与 B 相互独立。

因此，根据定义两个相互独立的事例不意味着是互斥的。前面的问题属于把两者定义混淆了。

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

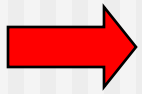
$$P(\bar{A}) = 1 - P(A)$$

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

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



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

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

2) $h(x)=\dfrac{x^2}{25}$ 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。

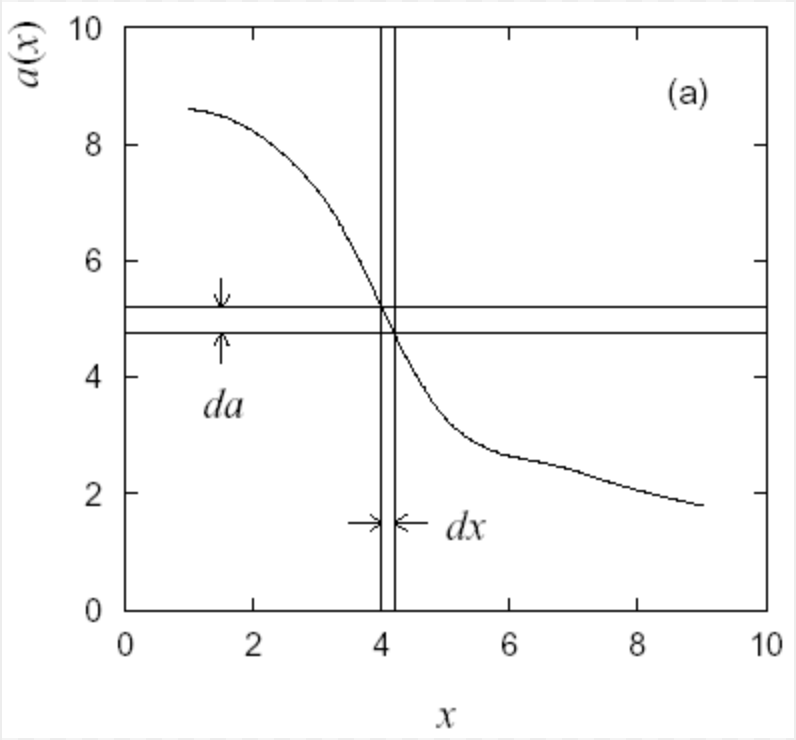
一维随机变量的函数*

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？



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

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

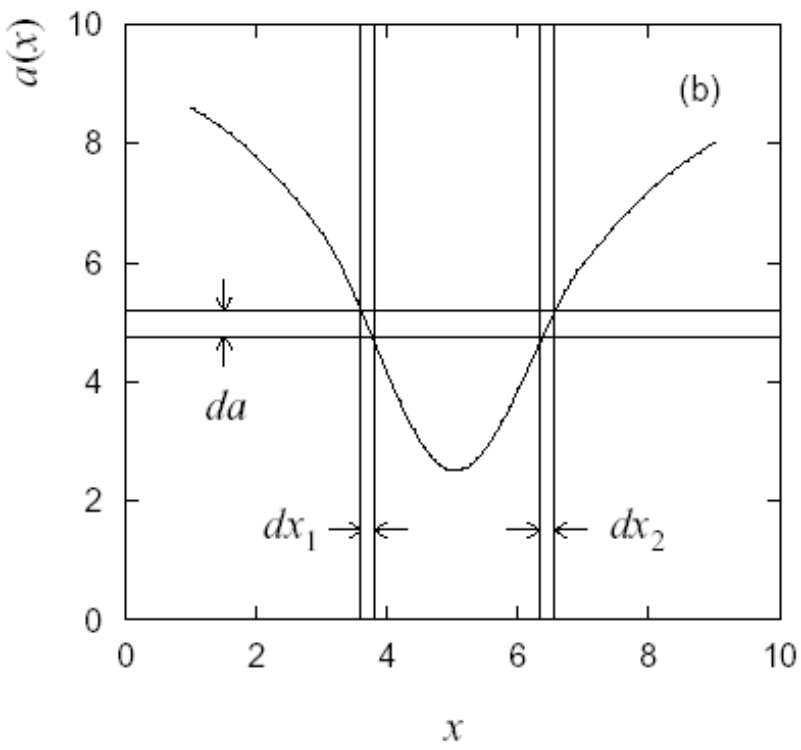
$$g(a)da = \left| \int_{x(a)}^{x(a+da)} f(x')dx' \right|$$

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

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

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

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

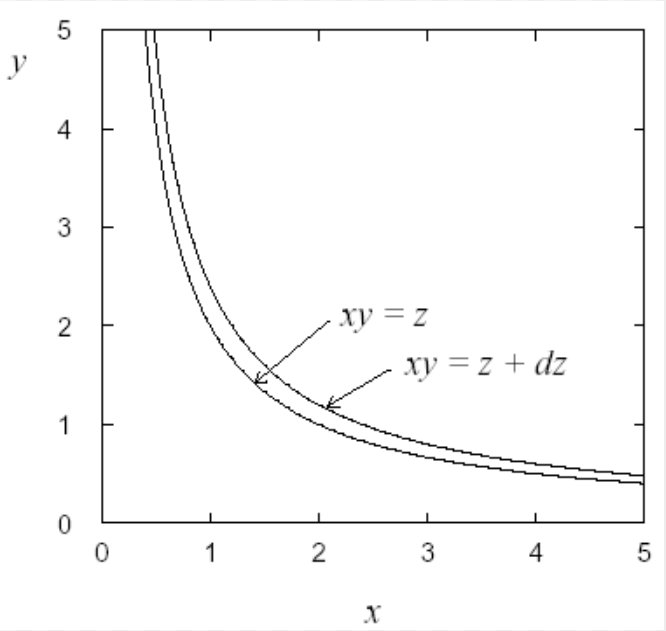
dS = 在 $a(\vec{x}) = a'$ 与 $a(\vec{x}) = a' + da'$ 定义的曲面 $\vec{x}$ 空间范围

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

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

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

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



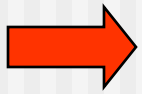
多维随机变量的函数(续)*

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。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，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

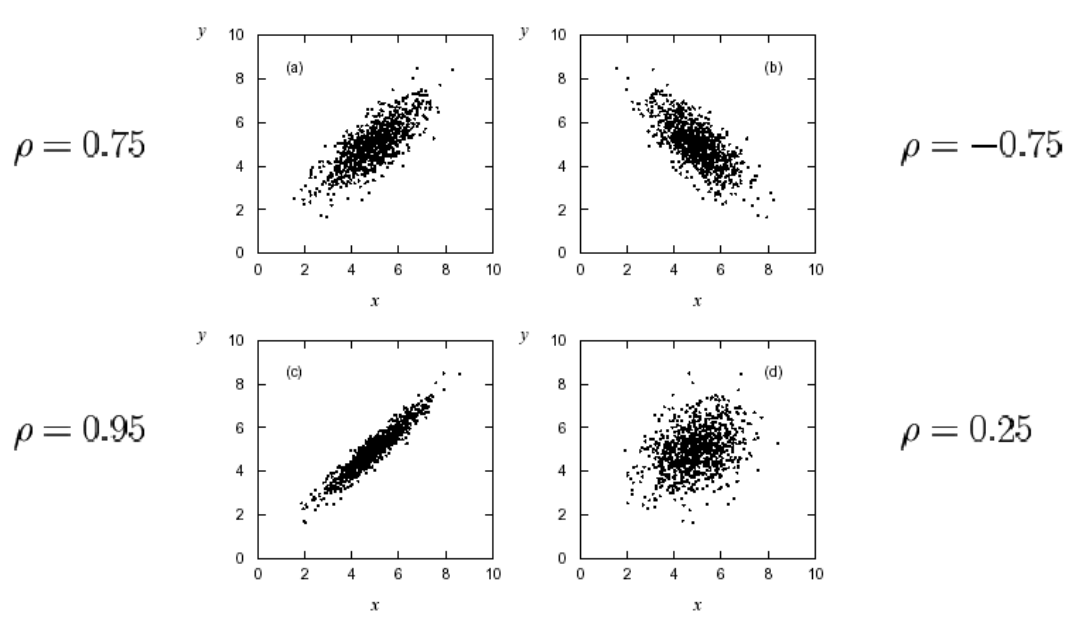
定义协方差 $\mathrm{cov}[x, y]$（也可用矩阵表示 V_{xy}）为

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

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

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

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 $$E[t]=\sum_{i=1}^{n} t_i P(t_i)$$

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？

举例：两衰变分支比测量相关性

$$Br = \frac{\text{观测到某一衰变的事例数}}{A\text{ 的所有衰变事例数} \times \text{探测效率}} = \frac{N_{B或C}}{N_A \times \varepsilon_{B或C}}$$

假设在探测的两种不同衰变事例 B 与 C中，有部分重叠ΔN，试估计相关系数的大小。假设对 B 与 C，探测效率不变。

根据分支比的定义，得到（相对频率）

$$x = Br（A \to B）$$

和

$$y = Br（A \to C）。$$

根据概率的相对频率定义以及该定义存在的问题，我们需要估计对应的方差。

假设（以后再讲）已经估计出对应的方差 $V[x]$ 与 $V[y]$，或者以标准偏差表示：σ_x 与 σ_y，如何研究相关性？

举例：相关性

由于事例 B 与 C 中，有部分重叠ΔN，因此两分支比测量值之间存在相关。

该相关性的存在会造成因为 ΔN 的变化，使得 x 与 y 的变化存在可以定量预见到某种程度上的正（反）比关系。例如，

情况1：分子比计算中，扣除重叠部分ΔN

情况2：分子比计算中，包含重叠部分ΔN

情况3：分子比计算中，重叠部分ΔN 只算在 B 或 C 中。

如果对应的标准偏差 σ_x 与 σ_y 中重叠部分贡献为 Δ_{xy}，能定量估计相关性吗？

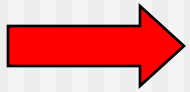
举例：相关系数估计

方法：重复实验测量分支比 x 与 y，或者在不失去统计意义的情况下在已有的样本中分成 m 个子样本，使得可以 m 次独立计算分支比 x 与 y，按照定义计算协方差

$$x_m = \sum_{i=1}^{n_x} \frac{1}{N_m}, \quad y_m = \sum_{i=1}^{n_y} \frac{1}{N_m}$$

按照定义计算协方差 $\mathrm{cov}[x,y] = E[xy] - \mu_x \mu_y$

$$E[xy] = \frac{1}{m} \sum_{i}^{m} x_i y_i$$



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

问题：分成子样本后 μ_x，μ_y，σ_x，σ_y 是什么？
注意：如已知 x 与 y 的p.d.f 和 Δ_{xy}，还可以有别的方法。

误差传递

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

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

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

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

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

误差传递(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

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

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。

坐标变换下的误差矩阵

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。

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

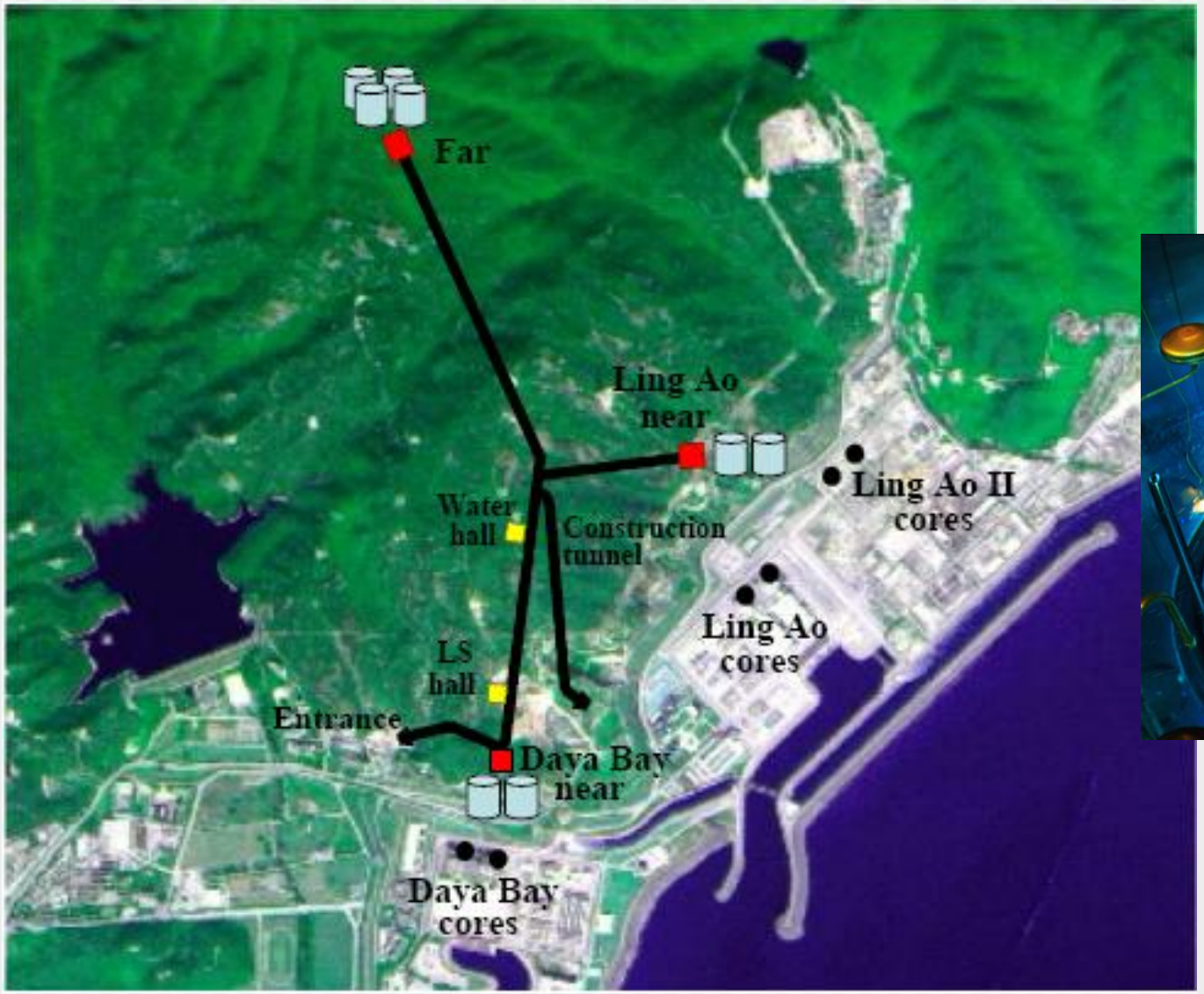
由于

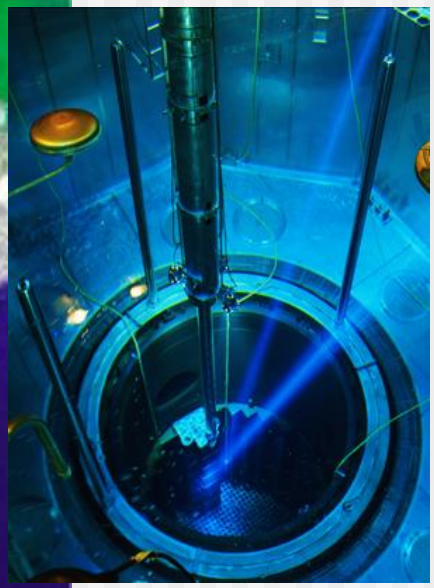
$$U(r,\theta) = AV(x,y)A^T$$

因此，坐标变换后的误差矩阵为

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

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

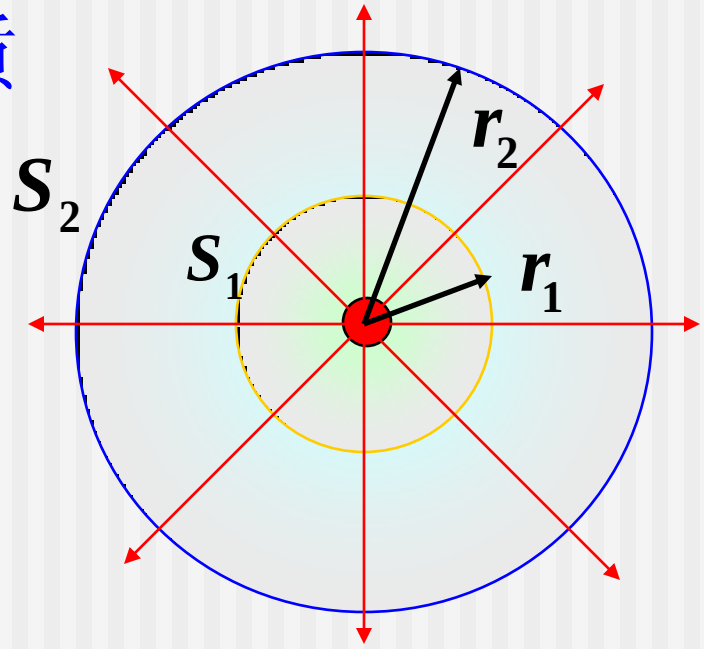
3 GW 热功率反应堆 $n \rightarrow p + e^- + \bar{\nu}_e$

$\Rightarrow$ 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

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

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

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？

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

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

$\Leftrightarrow$

非线性情况

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

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

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

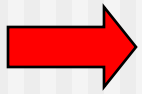
变换矩阵 A 由本征矢量 $\vec{r}$ 给出，即

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

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

$$
\begin{aligned}
U_{ij} &= \sum_{k,l=1}^{n} A_{ik} V_{kl} A_{lj}^{T} \\
&= \sum_{k,l=1}^{n} r_k^i V_{kl} r_l^j \\
&= \sum_{k=1}^{n} r_k^i \lambda_j r_k^j \\
&= \lambda_j \vec{r}^{\,i} \cdot \vec{r}^{\,j} \\
&= \lambda_j \delta_{ij}
\end{aligned}
$$



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。

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. 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。