



清华大学
Tsinghua University



粒子物理与核物理实验中的 数据分析

第四章：统计检验

杨振伟
清华大学



回顾

概率的基本概念

随机变量与概率密度函数

随机变量的均值与方差

常见的概率分布

蒙特卡罗方法

本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

概率与统计

统计的含义可以通过比较概率理论来理解

概率	统计(参数估计与假设检验)
从理论到数据	从数据到理论
根据理论计算某些可观测量(如, 均值, 分布等), 给出预期的实验分布。	进行所谓的假设检验, 比较理论预期的参量值或分布。从观察的实验数据中给出所研究参数的观测值和误差, 并且在某一置信水平上检验理论的正确与否。
例如: 若宇称守恒, 对某个特定衰变的分布有什么影响?	例如: 观测到一特定衰变分布, 是否可以断定宇称守恒?

统计分析的目标

假设检验



检验数据与某一特定理论是否相符(注意, 理论可包含一些自由参数)



相符程度由显著性水平表示

参数拟合



利用数据确定自由参数的大小



参数的准确程度由对应的不确定度表示

假设检验无处不在

- 如何判断某人是否感染了新冠病毒？
- 如何判断某病患是新冠肺炎还是普通肺炎？

“感染了新冠病毒” 和 “未感染新冠病毒” → 假设

“是新冠肺炎” 和 “不是普通肺炎” → 假设

根据观测结果确定 “假设” 真否 → 假设检验

关键：找到不同假设之间的典型差异

	新型冠状病毒感染的肺炎	流感	普通感冒
病原体	新型冠状病毒(2019-nCoV)	流感病毒	病毒、细菌、支原体、衣原体等多种病原体
主要症状	发热、乏力、干咳为主，部分患者可无发热，或出现胸闷、呼吸困难	高热、咳嗽、咽痛、头痛、肌肉疼痛等。流感也可引起肺炎，但是并不常见	鼻塞、流鼻涕等，多数患者症状较轻，一般不引起肺炎症状
是否有疫苗可预防	否	是，建议每年接种一次	否

新型冠状病毒的肺部CT表现也是和其他肺炎的肺部CT表现是不一样的，最初可能会表现斑片状阴影，随着病情的进展会出现磨玻璃状阴影，直到双肺出现实质性的病变而出现白色的变化。

普通的肺炎不会出现上述CT上的变化，新型冠状病毒肺炎也会很少出现胸腔积液的问题。

本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

假设 (hypotheses)

- 假设 H 可以预测数据的概率，即观测的结果（用 x 表示）
 - 例如， $x \sim f(x|H)$
 - x 可以是单变量，也可以是多变量；可以连续，也可以离散
 - x 可以代表观测的一个粒子、一个事例甚至整个“实验”
- x 的可能值构成样本空间或数据空间 S

简单假设： $f(x|H)$ 完全确定

复合假设： H 包含未确定的参数

给定 H 时 x 的概率又称作假设 H 的似然值： $L(x|H)$

什么是检验

检验的目标是，根据观测数据 x 对可能的假设的正确性给出某种论断

考虑简单假设 H_0 和备择假设 H_1 。对 H_0 的检验定义为：

对数据样本指定一个临界域 W ，使得在 H_0 正确的情况下，观测到这个数据的概率不超过某个（小）概率 α ，即

$$P(x \in W | H_0) \leq \alpha$$

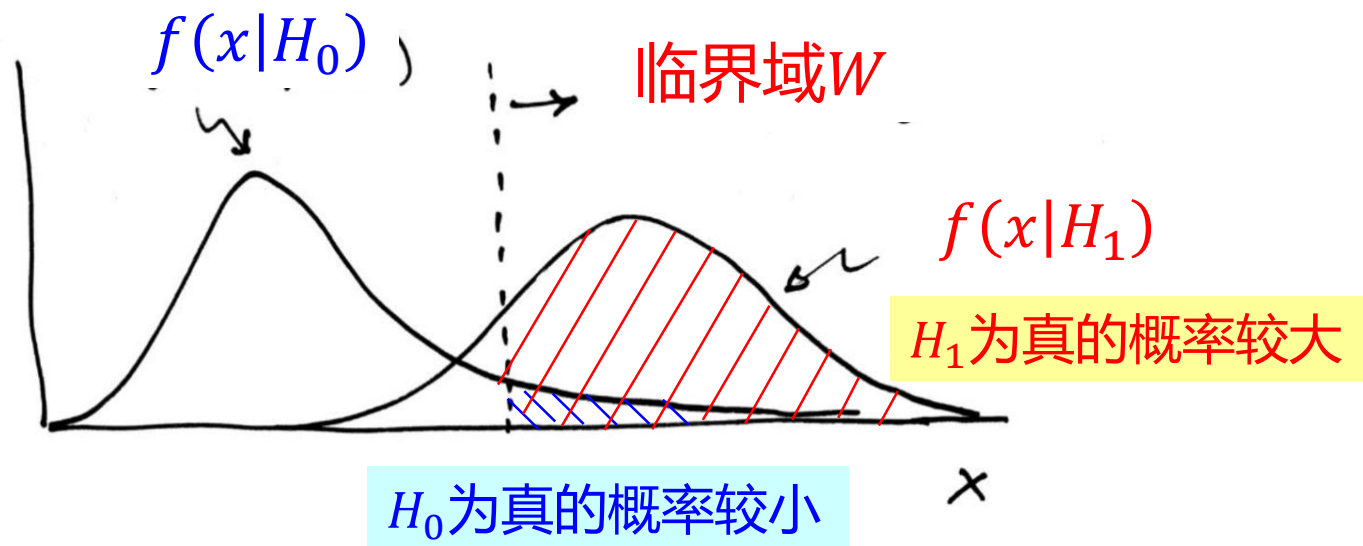
- 如果在临界域观测到 x ，则拒绝 H_0
- α ：检验的显著性水平或检验的大小
- 临界域又称为“拒绝域”，其补集称为接受域

什么是检验（续）

通常存在无穷多个可能的临界域可给出相同的显著性水平 α 。

所以，对 H_0 检验的临界域的选择需要考虑备择假设 H_1 。

大致说来，临界域的选择应当满足：临界域内 H_0 为真的概率较小， H_1 为真的概率较大。



拒绝假设 H_0

需要注意的是，拒绝假设 H_0 并不必然等价于我们相信 H_0 为假而 H_1 为真。对于相对频率论来说，统计只能把可重复观测的结果（数据）与概率联系起来。

在贝叶斯统计中，假设的概率（即信心度）可由贝叶斯定理给出：

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)dH} \quad \text{依赖于先验概率}\pi(H)$$

相对频率检验能做的是，在认为某个假设为真或者某个备择假设为真的条件下，计算接受这个假设或拒绝这个假设的概率。

第一类错误和第二类错误

选择有风险！

如果假设 H_0 为真而被拒绝，称为**第一类错误**，或**弃真错误**。

第一类错误的最大概率等于检验的显著性水平：

$$P(x \in W | H_0) \leq \alpha$$

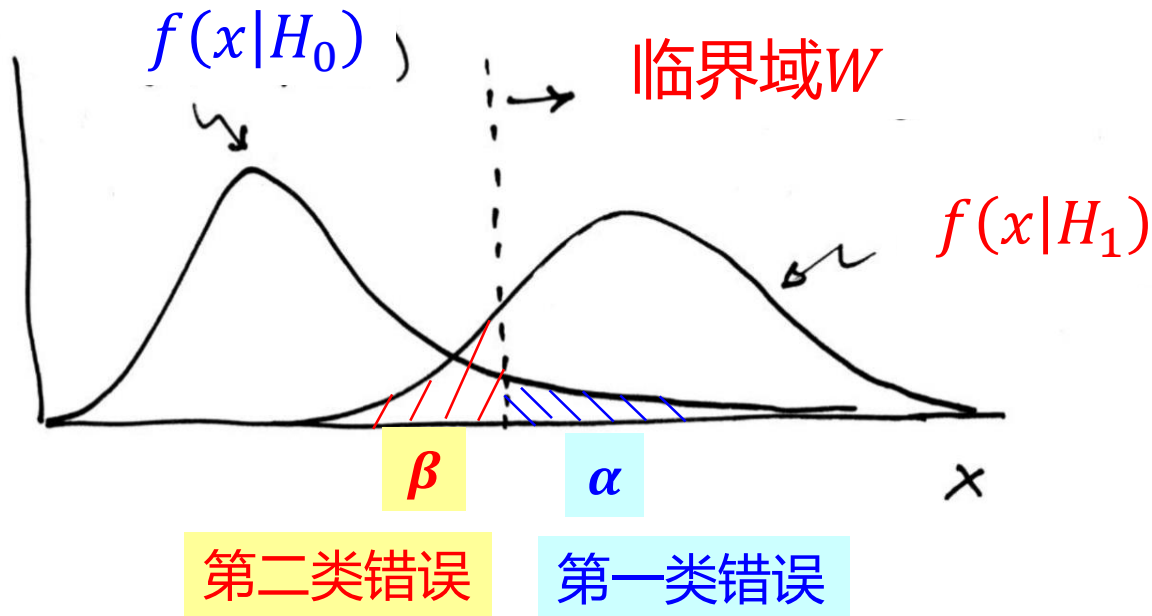
也可能 H_0 为假而 H_1 为真，但我们接受了 H_0 ，这称为**第二类错误**，或**取伪错误**，概率为

$$P(x \in S - W | H_1) = \beta$$

$1 - \beta$ 称为相对于备择假设 H_1 的检验的**功效** (Power)

$$\text{功效} = 1 - \beta$$

第一类错误和第二类错误 (续)



临界域的选择

要构造假设 H_0 的一个检验，可以问的一个问题是：想得到高功效的相关的备择假设是什么？

最大化相对于 H_1 的效力 = 最大化 H_1 为真时拒绝 H_0 的概率

有时，需要的备择假设不是简单假设，而是复合假设，例如，根据对 $x \sim N(\mu, \sigma^2)$ 的观测，可以检验

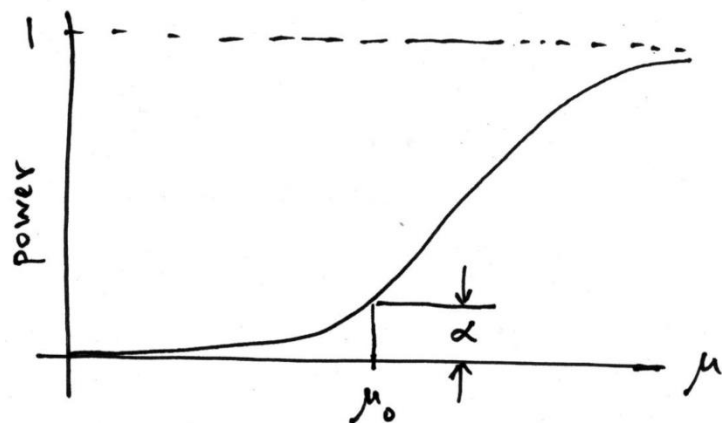
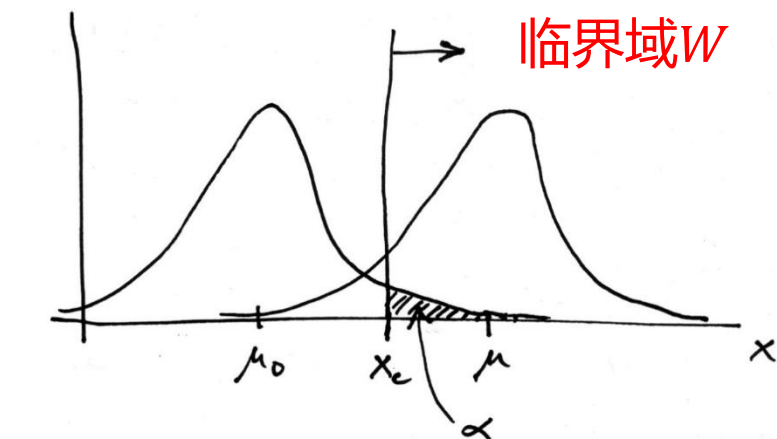
$$H_0: \mu = \mu_0 \quad \text{v.s.} \quad H_1: \mu > \mu_0$$

取临界域 $x \geq x_c$ 得到相对于任意 $\mu > \mu_0$ 的最高功效， x_c 由显著性水平 α 确定

$$\alpha = P(x \geq x_c | \mu_0)$$

例：高斯样本 $x \sim N(\mu, \sigma^2)$ 的均值检验

➤ 检验 $H_0: \mu = \mu_0$ v.s. $H_1: \mu > \mu_0$



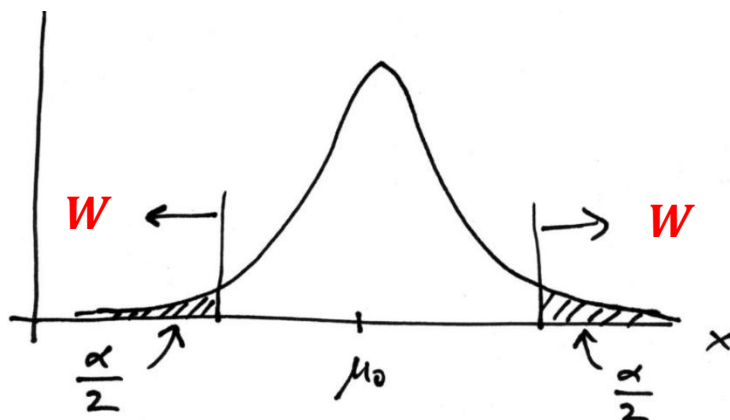
标准高斯的累积分布

$$\alpha = 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma}\right)$$
$$x_c = \mu_0 + \sigma \Phi^{-1}(1 - \alpha)$$

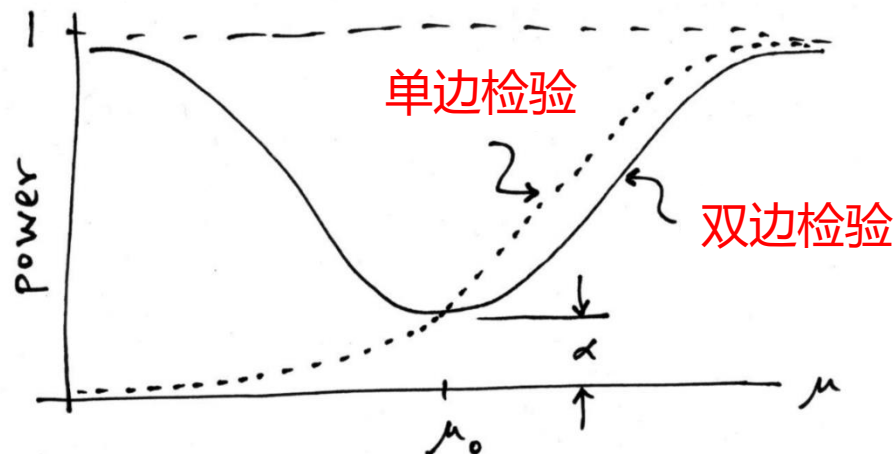
标准高斯的分位数

$$\begin{aligned} \text{功效} &= 1 - \beta = P(x > x_c | \mu) \\ &= 1 - \Phi\left(\frac{x_c - \mu_0}{\sigma} + \Phi^{-1}(1 - \alpha)\right) \end{aligned}$$

基于功效的临界域选择



有时我们会将 $\mu > \mu_0$ 和 $\mu < \mu_0$ 都看作可能的备择假设，临界域将同时包含 x 较大和较小的区域（双边检验）



在这个双边检验中，临界域对于 $\mu < \mu_0$ 给出合理的功效，但是对于 $\mu > \mu_0$ 给出的功效低于单边检验的结果

一般来说，无法找到一个临界域对所有可能的备择假设都给出最大功效

本章要点

- 假设, 检验, 显著水平, 功效, 临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度, p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

粒子物理中的统计检验

假设单个事例的测量结果用矢量 $\vec{x} = (x_1, \dots, x_n)$ 表示

$$x_1 = \mu \text{轻子的数目}$$

$$x_2 = \text{喷注的平均 } p_T$$

$$x_3 = \dots\dots$$

假设 \vec{x} 服从某个 n 维的联合概率密度, 这个分布依赖于产生的事例类型, 例如

$$pp \rightarrow t\bar{t}, \quad pp \rightarrow \tilde{g}\tilde{g}, \dots$$

考虑的每个反应都有一个关于 \vec{x} 的概率密度的假设, 如 $f(\vec{x}|H_0)$, $f(\vec{x}|H_1)$, $\dots\dots$

例如, 称 H_0 为本底假设, H_1 为信号假设

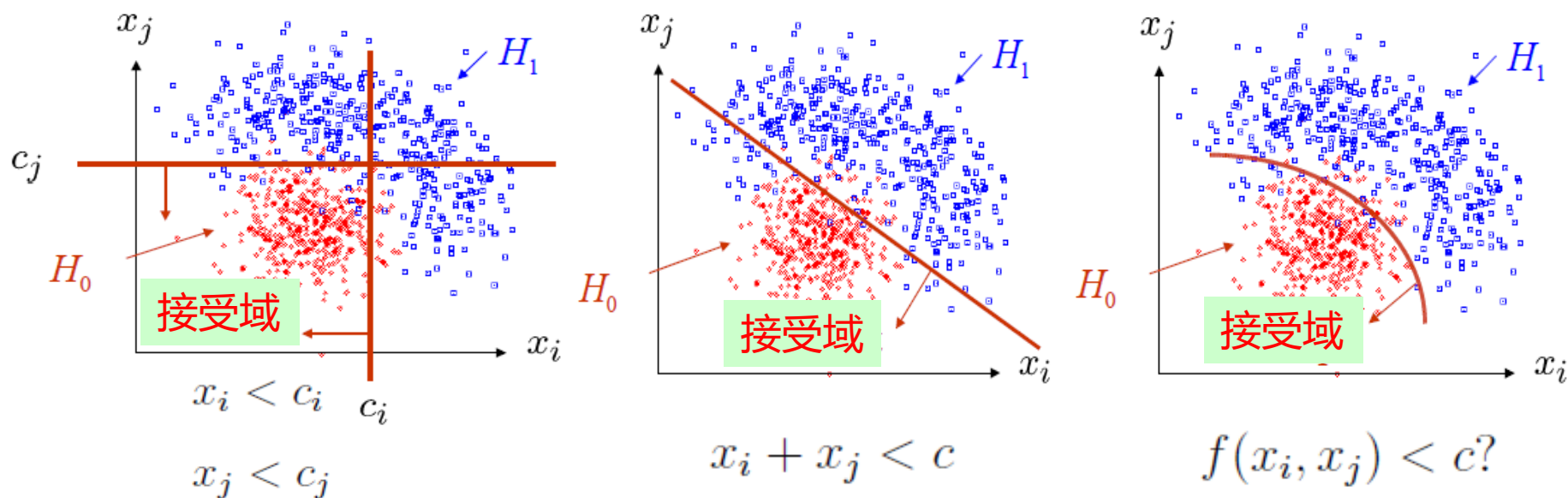
希望拒绝的事例类型

希望保留的事例类型

事例筛选

假设实验数据包含两种不同事例，对应于假设 H_0 和 H_1 ，我们希望选择出 H_1 类的事例（位于临界域）。

每个事例对应于 \vec{x} 空间的一个点。应当如何选择判别边界？



- 1) 如何给出“最优化”的选择？
- 2) 多维空间有何困难？

检验统计量

n 维数据空间 $\vec{x} = (x_1, \dots, x_n)$ 的临界域的边界

$$t(x_1, \dots, x_n) = t_{\text{cut}}$$

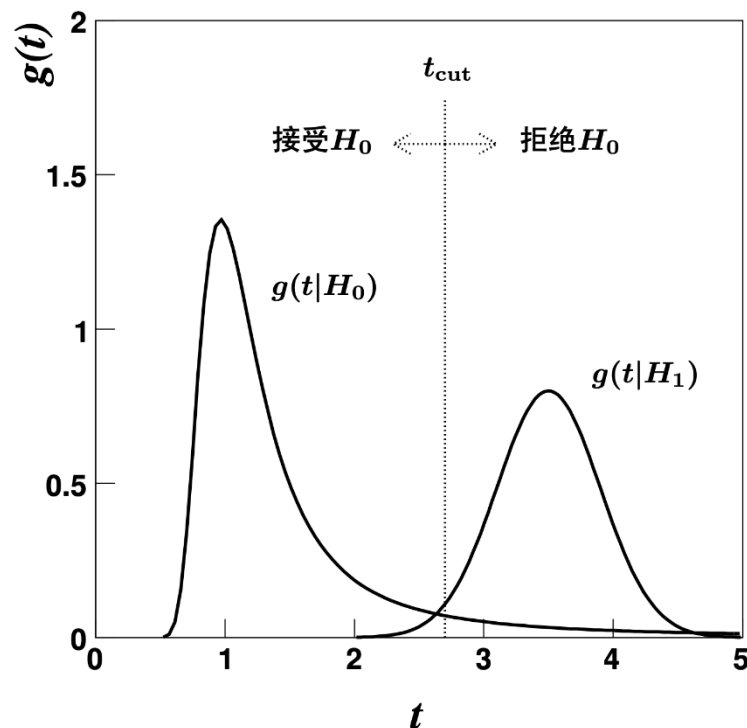
其中 $t(x_1, \dots, x_n)$ 是一个标量检验统计量

可以算出概率密度

$$g(t|H_0), g(t|H_1), \dots$$

边界的确定现在变成对单个变量 t 的处理, t_{cut} 确定了临界域

→ n 维问题约化为一维问题



从统计检验的视角看分类

H_0 为真却拒绝 H_0 的概率（弃真错误）： $\alpha = \int_W f(\vec{x}|H_0) d\vec{x}$

α = 检验的显著性水平，检验的大小，发现的错误率

H_1 为真却接受 H_0 的概率（取伪错误）： $\beta = \int_{\bar{W}} f(\vec{x}|H_1) d\vec{x}$

$1 - \beta$ = 检验对于 H_1 的效力

等效地，如果 H_0 = 本底事例， H_1 = 信号事例， ε_b 和 ε_s 分别为本底和信号的选择效率

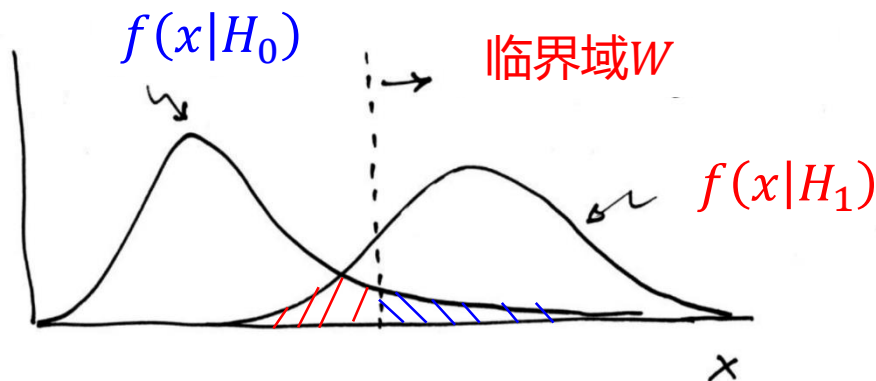
$$\varepsilon_b = \int_W f(\vec{x}|H_0) d\vec{x} = \alpha$$

$$\varepsilon_s = \int_W f(\vec{x}|H_1) d\vec{x} = 1 - \beta = \text{效力}$$

纯度/误鉴别率

考虑信号事例 (s) 被正确分类的概率，即事例选择的纯度
利用贝叶斯定理：

$$\begin{aligned} \underbrace{P(s|\vec{x} \in W)}_{\text{验后概率}} &= \frac{\overbrace{P(\vec{x} \in W|s)}^{\epsilon_s} \underbrace{P(s)}_{\text{先验概率}}}{P(\vec{x} \in W|s)P(s) + \underbrace{P(\vec{x} \in W|b)}_{\epsilon_b} P(b)} \\ &= \text{信号纯度} = 1 - \text{信号误鉴别率} \end{aligned}$$



注意：纯度依赖于给定事例为信号或本底先验概率，也依赖于信号和本底的效率

例：粒子鉴别

一束只包含 K/π 两种介子的束流穿过2厘米厚的闪烁体，电离能损的大小可以用于粒子鉴别。构造能量沉积观测量 t 。

$H_0 = \pi$ (信号)

$H_1 = K$ (本底)

$t < t_{\text{cut}}$ 选择 π 介子，效率为

$$\varepsilon_{\pi} = \int_{-\infty}^{t_{\text{cut}}} g(t|\pi) dt = 1 - \alpha$$

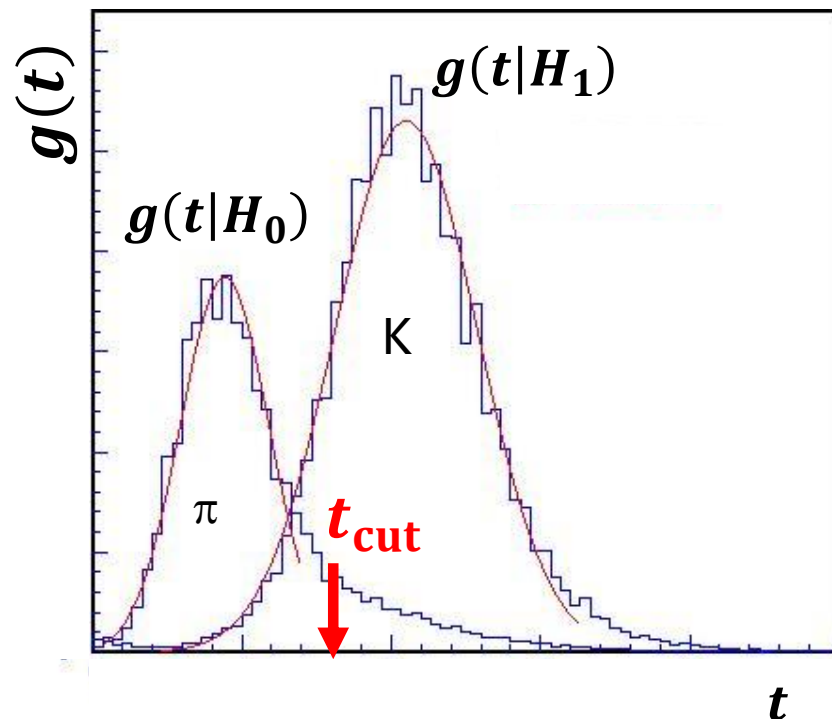
$$\varepsilon_K = \int_{-\infty}^{t_{\text{cut}}} g(t|K) dt = \beta$$

宽松选择：效率高，但 K 本底高；

严格选择：纯度高，但效率低。

π 的份额 a_{π} 可从 t 的分布估计

$$f(t; a_{\pi}) = a_{\pi} g(t|\pi) + (1 - a_{\pi}) g(t|K)$$



粒子鉴别的概率问题

对于测量值为 t 的粒子，如何估计其为 K 或 π 的概率？

贝叶斯定理



$$h(K|t) = \frac{a_K g(t|K)}{a_K g(t|K) + a_\pi g(t|\pi)}$$
$$h(\pi|t) = \frac{a_\pi g(t|\pi)}{a_K g(t|K) + a_\pi g(t|\pi)}$$

贝叶斯论：上式是粒子为 K 或 π 的可信程度

频率论：给定 t 条件下，粒子是 K 或 π 的比率



两种解释
均有道理

通常情况下，需要给出选择样本的纯度

$$p_\pi = \frac{N_\pi(t < t_{\text{cut}})}{N_{\text{all}}(t < t_{\text{cut}})} = \frac{\int_{-\infty}^{t_{\text{cut}}} a_\pi g(t|\pi) dt}{\int_{-\infty}^{t_{\text{cut}}} [a_\pi g(t|\pi) + (1 - a_\pi) g(t|\pi)] dt} = \frac{\int_{-\infty}^{t_{\text{cut}}} h(\pi|t) f(t) dt}{\int_{-\infty}^{t_{\text{cut}}} f(t) dt}$$

= π 介子在区间 $(-\infty, t_{\text{cut}}]$ 的概率

注意: $h(\pi|t)$ 有时会被解释为检验统计量。

粒子鉴别

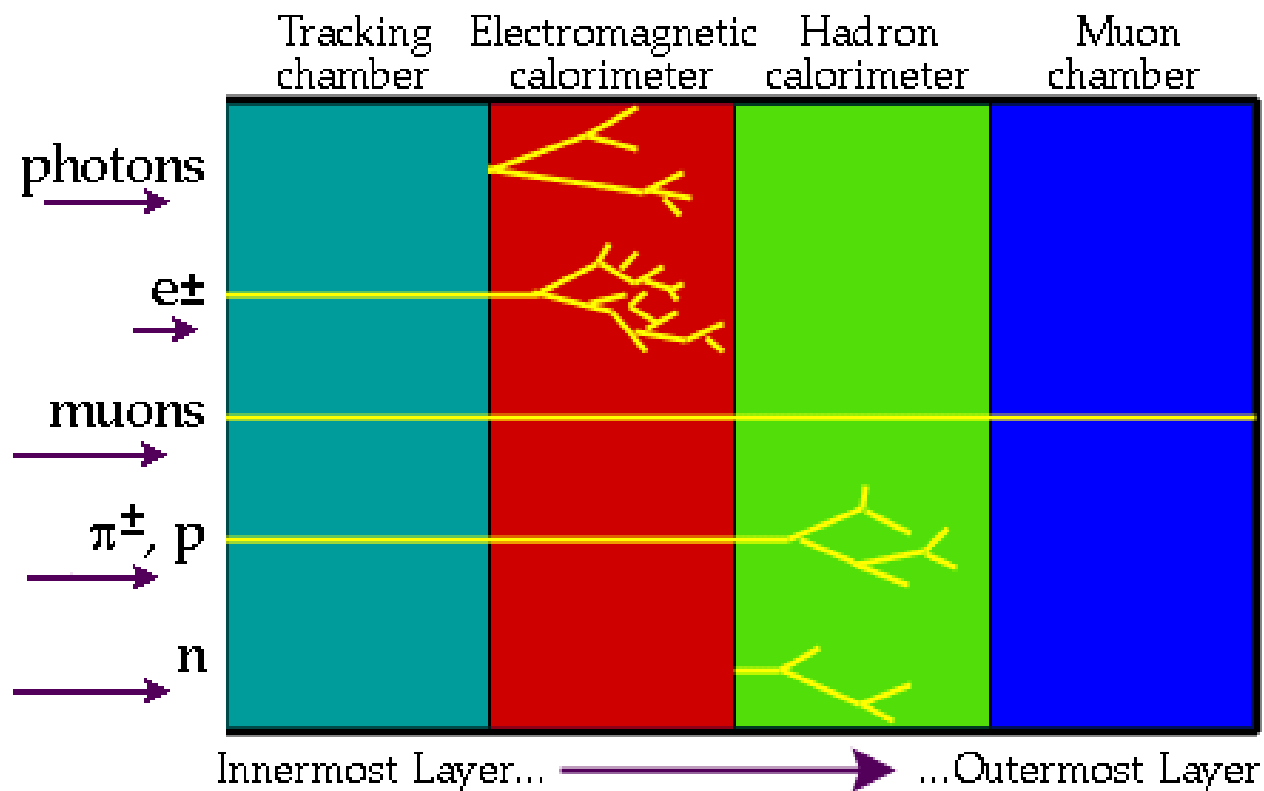
探测器中的稳定粒子($\tau \geq 10^{-9}\text{s}$)

$\gamma, e^{\pm}, \mu^{\pm}, \pi^{\pm}, K^{\pm}, K_L$

p, n, Λ

$$l = \frac{pc^2}{E} \cdot \tau \cdot \frac{E}{mc^2}$$

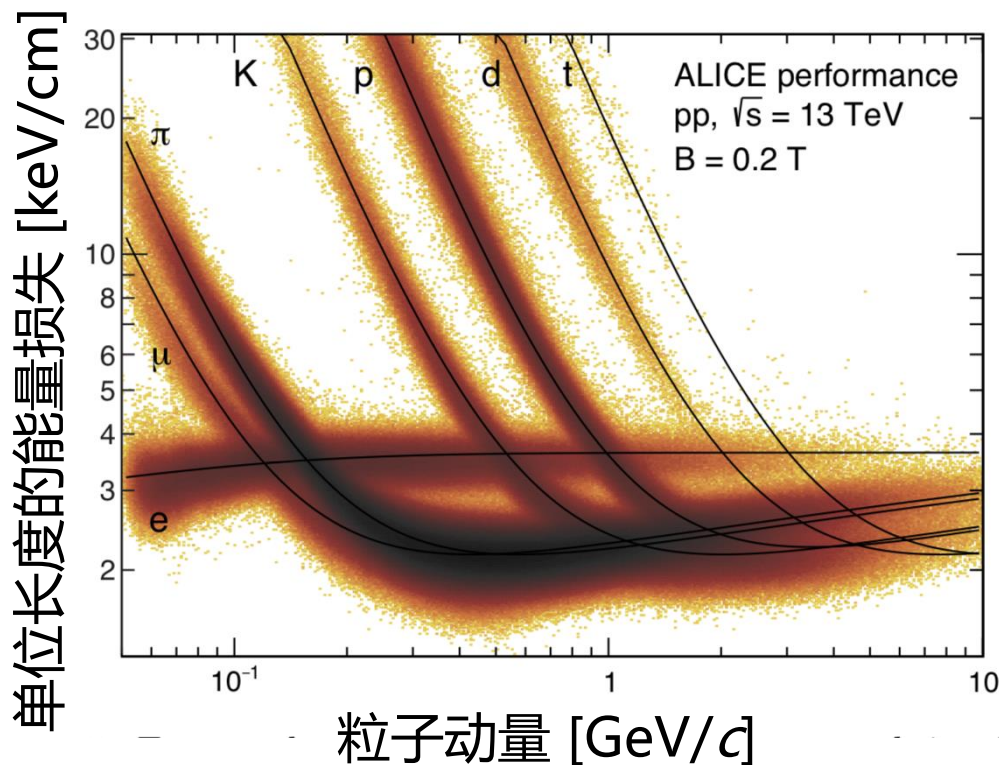
$$= \tau \frac{p}{m}$$



粒子鉴别的例子

粒子在物质中的任何相互作用原则上都可用来作粒子鉴别，区别在于效果好坏。

带电粒子在物质中电离造成能量损失，可用来鉴别带电粒子：

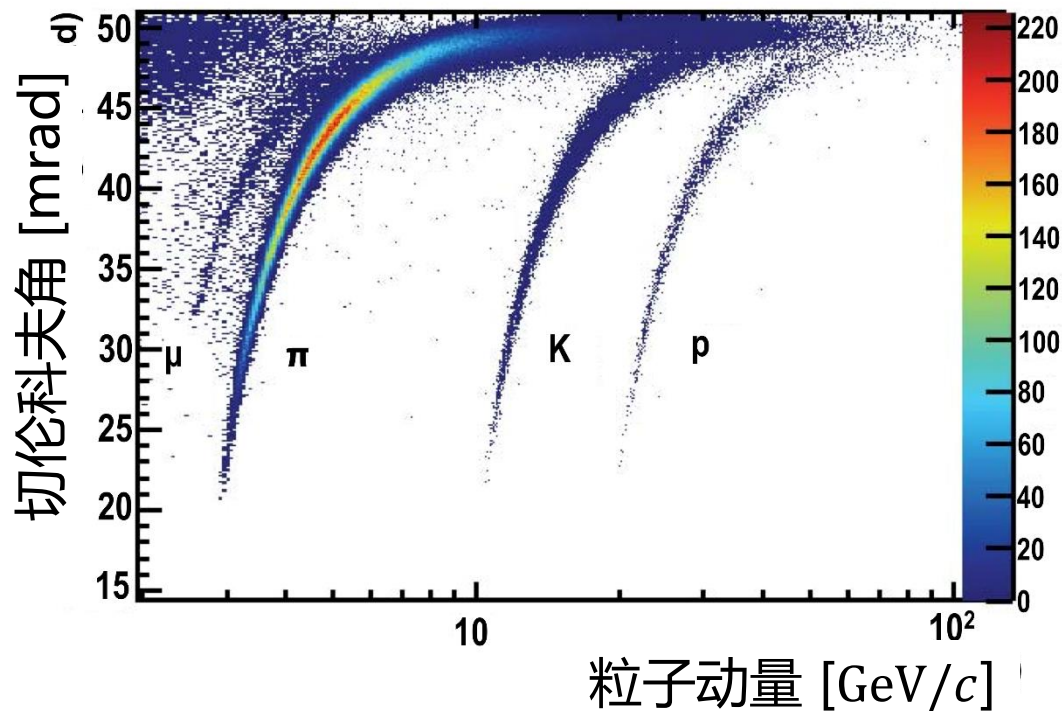


局限？

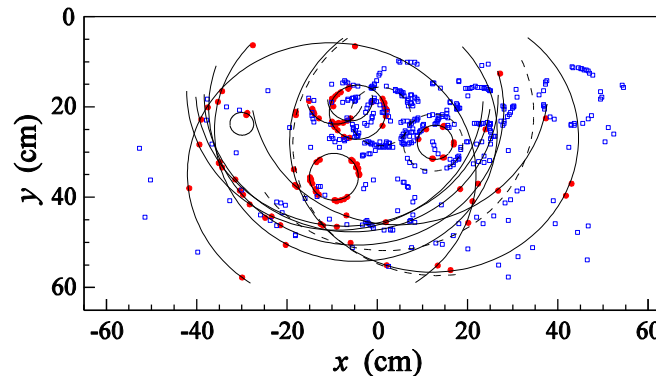
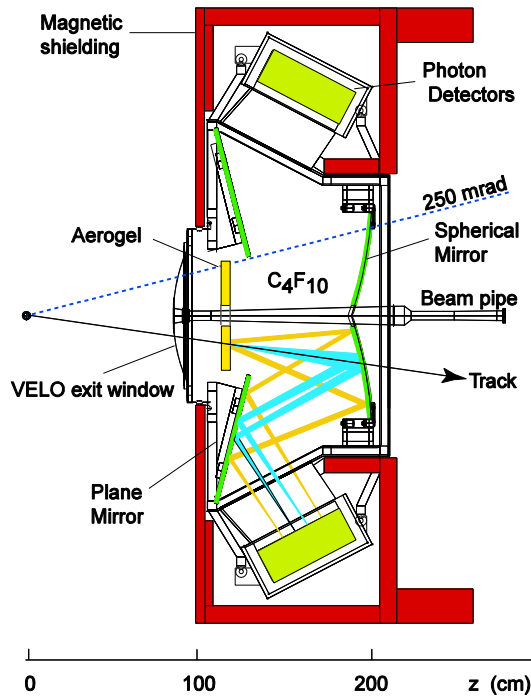
粒子鉴别的例子（续）

粒子在物质中的任何相互作用原则上都可用来作粒子鉴别，区别在于效果好坏。

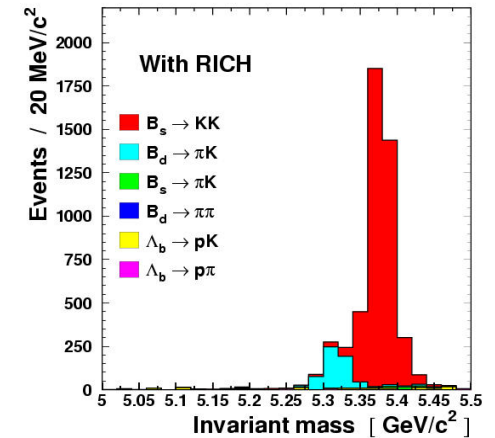
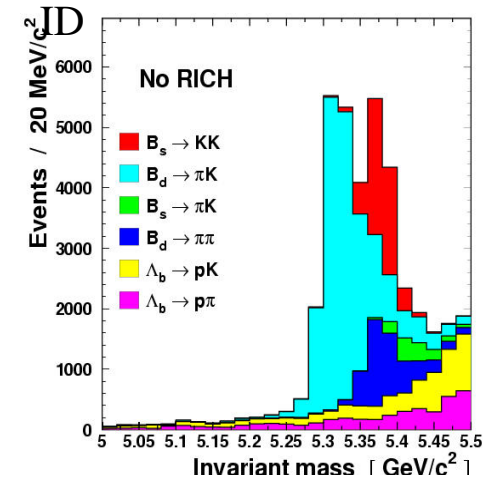
带电粒子在介质中发射切伦科夫光：



LHCb切伦科夫探测器

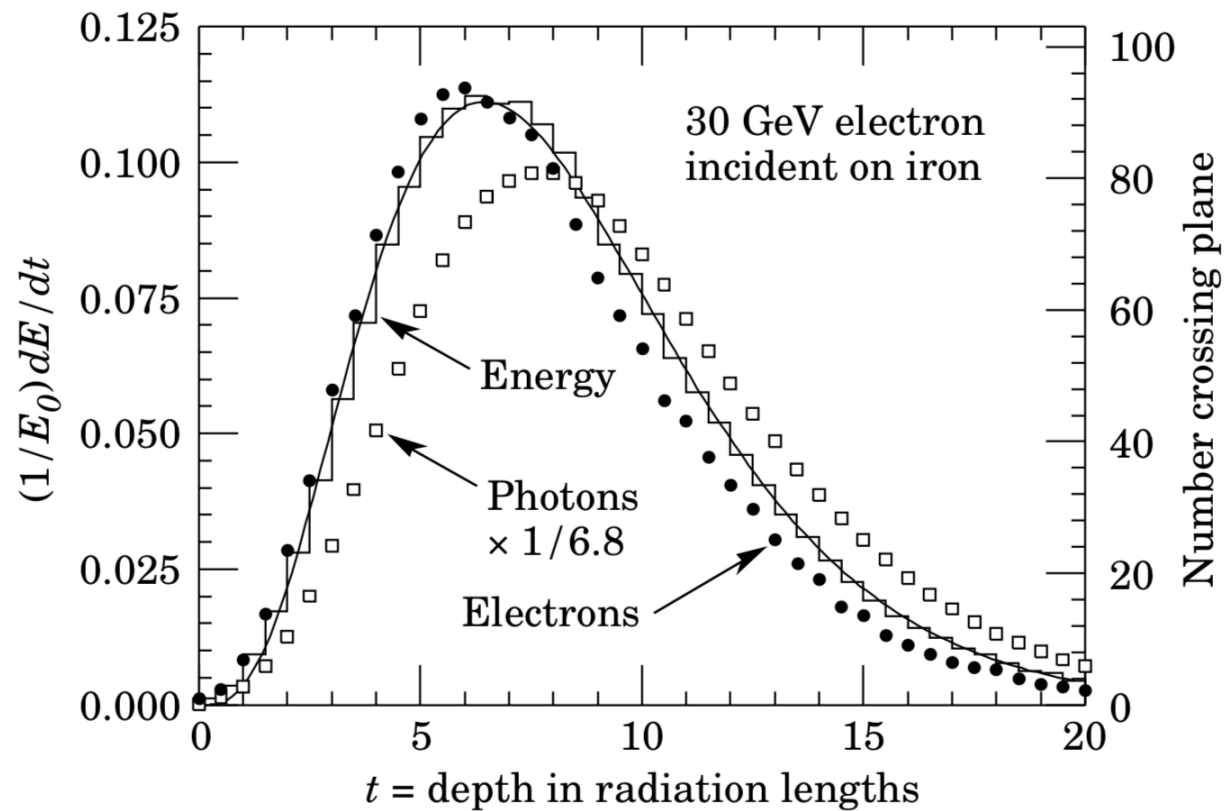


Performance of particle



光子鉴别

高能光子 ($> 1 \text{ GeV}/c$) 在物质中发生电磁簇射



本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

奈曼-皮尔逊引理与临界域

多维检验统计量 $\vec{t} = (t_1, \dots, t_m)$, 原假设 H_0 , 备择假设 H_1 。

问题：如何选择一个最佳的临界域或者 cut?

奈曼-皮尔逊引理：在给定效率条件下，要得到最高纯度的信号样本，或者在给定显著性水平下得到最高功效，可以选择下列接受域来实现

$$\frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} > c, \quad c \text{ 为常数, 与效率有关。}$$

对于不含未定参量的最优化一维检验统计量,

$$r \equiv \frac{g(\vec{t}|H_0)}{g(\vec{t}|H_1)} \longrightarrow \text{简单假设 } H_0 \text{ 与 } H_1 \text{ 的似然之比}$$

实际应用中, r 最好是单调函数。

基于似然比的检验统计量

在只考虑两种假设的情况下，对于每个事例，测量

$$\vec{x} = (x_1, \dots, x_n)$$

根据纽曼-皮尔森引理，为了选择事例，可构造检验统计量

$$t(\vec{x}) \equiv \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)}$$

思考：这个检验统计量的任何单调函数给出的检验相同吗？

问题：如何知道这两个不同假设下的概率密度函数？

如何得到不同假设下概率密度函数

实际应用中，可用蒙特卡罗方法模拟物理过程与探测器响应，产生大量事例，近似得到概率密度函数。

分别产生信号与本底事例，并经过探测器模拟。

对每个事例，得到 \vec{x} ，并填入 n 维直方图。如果 M 为每个分量的区间数，则总的区间单元数为 M^n 。

$$\frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)}$$

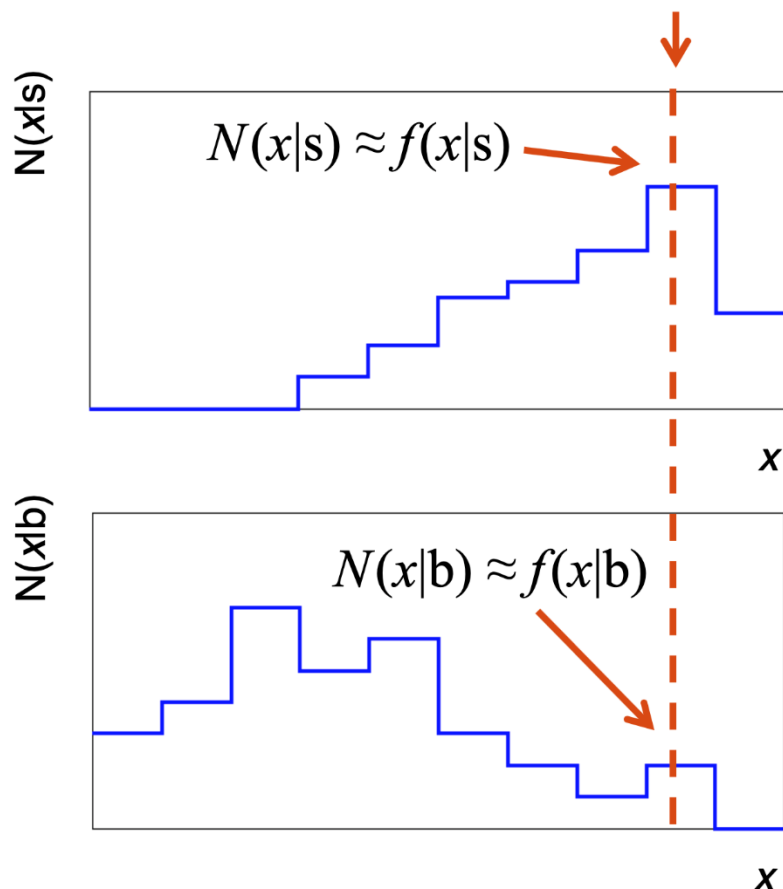
这种模拟往往不便宜（LHC上一个完全模拟事例大约需要1个CPU 运算1分钟）

如果 n 太大时，实际运用会很困难。

奈曼-皮尔逊引理实际上往往不可行

利用直方图近似得到似然比

希望得到 $t(x) = \frac{f(x|s)}{f(x|b)}$



可以产生MC数据并得到信号和本底的直方图。

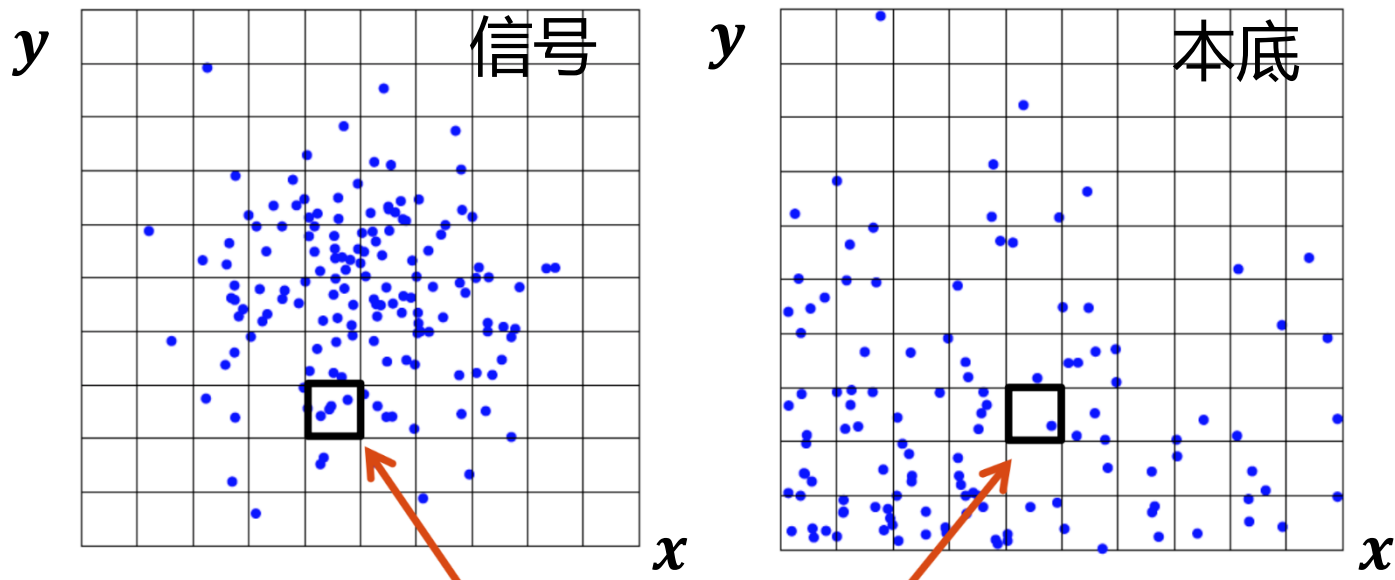
利用归一化的直方图的值近似得到似然比

$$t(x) = \frac{N(x|s)}{N(x|b)}$$

对于单变量问题，这种方法很有效。

利用二维直方图近似得到似然比

假设问题中有两个变量(x, y), 可以使用二维直方图



利用对应网格中的 $N(x, y|s)$ 和 $N(x, y|b)$ 近似二维概率密度。

如果希望每个变量分 M 个区间, 对于 n 维问题, 需要 M^n 个网格; 很难产生足够的训练数据。

➔ 对于 $n > 1$ 的问题, 直方图方法往往不适用

多变量分析的策略

- 奈曼-皮尔逊引理给出了优化的结果，但是往往无法直接使用，因为我们通常没有 $f(\vec{x}|\mathbf{s})$ 和 $f(\vec{x}|\mathbf{b})$
- 对于直方图方法， n 个变量 M 个区间需要我们估计 M^n 个参数（即每个网格中概率密度的值），往往不适用
- 一种折中的办法是，假设检验统计量 $t(\vec{x})$ 的概率密度可以用参数很少的某种函数表示，并确定其形式，给出信号和本底的最佳区分
- 或者，用比直方图更好的办法估计概率密度 $f(\vec{x}|\mathbf{s})$ 和 $f(\vec{x}|\mathbf{b})$ ，然后用估计的概率密度构造近似的似然比

多变量方法

➤ 多变量方法很多

- 费舍尔判别量 (Fisher discriminant)
- 神经网络 (Neural networks)
- 核密度方法 (Kernel density methods)
- 支持向量机 (Support Vector Machines)
- 决策树 (Decision trees)
 - ✓ Boosting
 - ✓ Bagging

多变量方法的参考资料

- C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009
- R. Duda, P. Hart, D. Stork, *Pattern Classification*, 2nd ed., Wiley, 2001
- A. Webb, *Statistical Pattern Recognition*, 2nd ed., Wiley, 2002
- Ilya Narsky and Frank C. Porter, *Statistical Analysis Techniques in Particle Physics*, Wiley, 2014
- 朱永生（编著），实验数据多元统计分析，科学出版社，北京，2009

多变量方法的软件

- TMVA, Höcker, Stelzer, Tegenfeldt, Voss, Voss, Physics/0703039
 - tmva.sourceforge.net, 包含于ROOT发行版
 - 手册好用, 广泛应用于粒子物理实验
- Scikit-learn
 - 基于Python的机器学习工具
 - scikit-learn.org
 - 广泛的用户群

发展非常快, 包括软件更新和新软件发布

线性检验统计量

维数 $n > 2$ 时，用蒙特卡罗法求多维概率密度有困难。假设每一维均需分 M 个区间，共需 M^n 个单元格近似概率密度函数。为了简化问题，可以采用线性变换方法给出包含少量参数的检验统计量，并确定参数，最大限度地区分 H_0 与 H_1 。

线性变换

$$t(\vec{x}) = \sum_{i=1}^n a_i x_i = \vec{a}^T \vec{x}$$

给定变换系数 \vec{a} ，可以得到相应的概率密度 $g(t|H_0), g(t|H_1)$ 。

通过选择 \vec{a} ，达到最大程度区分 $g(t|H_0)$ 和 $g(t|H_1)$ 的目的。

不同甄别量的定义会导致确定系数的规则不同，因此



必须明确定义所谓的甄别量

$t(\vec{x})$ 在不同假设下的均值与方差

观测量 \vec{x} 的均值与协方差

$$(\mu_k)_i = \int x_i f(\vec{x}|H_k) d\vec{x}$$

$$(V_k)_{ij} = \int (x - \mu_k)_i (x - \mu_k)_j f(\vec{x}|H_k) d\vec{x}$$

$k = 0, 1$ (假设)

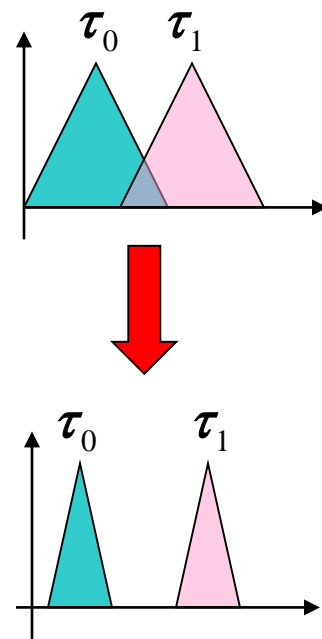
$i, j = 1, \dots, n$ (\vec{x} 的分量)

类似地, 可以得到 $t(\vec{x})$ 的均值与方差

$$\tau_k = \int t(\vec{x}) f(\vec{x}|H_k) d\vec{x} = \vec{a}^T \vec{\mu}_k$$

$$\Sigma_k^2 = \int (t(\vec{x}) - \tau_k)^2 f(\vec{x}|H_k) d\vec{x} = \vec{a}^T V_k \vec{a}$$

要求 $|\tau_0 - \tau_1|$ 大, 而 Σ_0^2 和 Σ_1^2 小, 使得概率密度分布 $g(t|H_0)$ 和 $g(t|H_1)$ 都集中在各自的均值附近且均值相差较大。



费舍尔判别法

费舍尔定义判别函数

$$J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}$$

$$\sum_{i,j=1}^n a_i a_j (\mu_0 - \mu_1)_i (\mu_0 - \mu_1)_j = \sum_{i,j=1}^n a_i a_j B_{ij} = \vec{a}^T B \vec{a}$$

$$B = (\vec{\mu}_0 - \vec{\mu}_1)(\vec{\mu}_0 - \vec{\mu}_1)^T$$

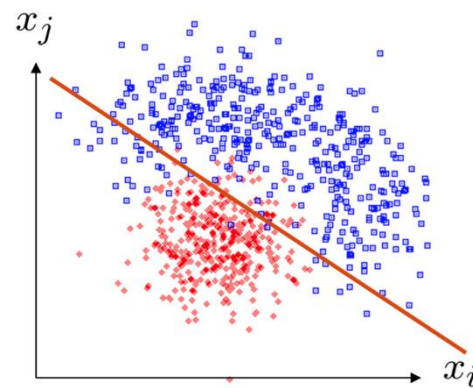
$$\sum_{i,j=1}^n a_i a_j (V_0 + V_1)_{ij} = \vec{a}^T W \vec{a}$$

$$W = V_0 + V_1$$

$$J(\vec{a}) = \frac{\vec{a}^T B \vec{a}}{\vec{a}^T W \vec{a}}$$

$$\text{令 } \frac{\partial J}{\partial a_i} = 0$$

$$\vec{a} \propto W^{-1}(\vec{\mu}_0 - \vec{\mu}_1)$$



因此定义了可求极值的费舍尔线性判别函数 J 。

求费舍尔甄别函数的最大值

将检验统计量 $t(\vec{x})$ 写为

$$t(\vec{x}) = a_0 + \sum_{i=1}^n a_i x_i$$

这样可以用任意标度和偏倚量 a_0 去固定 τ_0 和 τ_1

于是, 求 $J(\vec{a}) = \frac{(\tau_0 - \tau_1)^2}{\Sigma_0^2 + \Sigma_1^2}$ 的最大值, 意味着最小化其分母

$$\Sigma_0^2 + \Sigma_1^2 = E_0[(t - \tau_0)^2] + E_1[(t - \tau_1)^2]$$

与假设 H_0 和 H_1 对应的期待值

求费舍尔函数 $J(\vec{a})$ 最大值是后面将介绍的最小二乘法的例子

高斯分布下费舍尔甄别量的特点

假设 $f(\vec{x}|H_k)$ 是多变量高斯分布，均值为

$\vec{\mu}_0$ ：假设 H_0 的均值 $\vec{\mu}_1$ ：假设 H_1 的均值

并假设二者的协方差矩阵为 $V_0 = V_1 \equiv V$

含偏置量的费舍尔甄别量为 $t(\vec{x}) = a_0 + (\vec{\mu}_0 - \vec{\mu}_1)^T V^{-1} \vec{x}$

利用前面所述的似然比在给定效率条件下的最大纯度

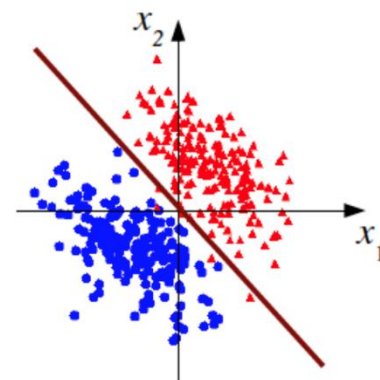
$$r = \frac{f(\vec{x}|H_0)}{f(\vec{x}|H_1)} = \exp \left[-\frac{1}{2} (\vec{x} - \vec{\mu}_0)^T V^{-1} (\vec{x} - \vec{\mu}_0) + \frac{1}{2} (\vec{x} - \vec{\mu}_1)^T V^{-1} (\vec{x} - \vec{\mu}_1) \right]$$

$\propto e^t$

→ $t \propto \log r + \text{常数}$ (单调变化)

费舍尔甄别量
与似然比等价

提问：这个费舍尔甄别量是否最优检验统计量？



验后概率与逻辑函数

如果多维变量 \vec{x} 在不同假设下协方差相同，则验后概率有简单的表达式，例如

$$P(H_0|\vec{x}) = \frac{f(\vec{x}|H_0)P(H_0)}{f(\vec{x}|H_0)P(H_0) + f(\vec{x}|H_1)P(H_1)} = \frac{1}{1 + \frac{P(H_1)}{P(H_0)} r}$$

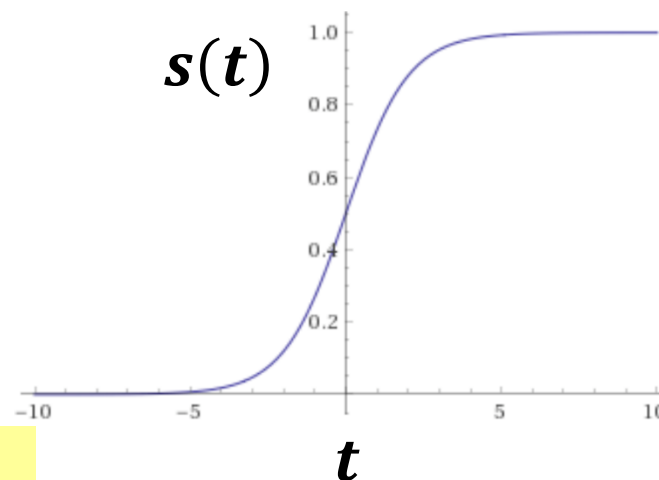
↑
贝叶斯定理

验前概率

选择恰当的偏置量 a_0 ，利用高斯分布下费舍尔甄别量的特点 ($r \propto e^t$)：

$$P(H_0|\vec{x}) = \frac{1}{1 + e^{-t}} \equiv s(t)$$

这就是所谓的“逻辑S型”函数



思考：要得到上式， a_0 应选为什么形式？

输入变量的变换

如果数据不是方差相同的高斯分布，线性决策边界不是最优的。但是可以考虑对数据作变换

$$\vec{\varphi} = (\varphi_1(\vec{x}), \dots, \varphi_m(\vec{x}))$$

然后以 φ_i 为新输入变量。 $\vec{\varphi}$ 经常被称为“特征空间”， φ_i 是“基函数”。基函数可以固定，也可以包含可调参数，由训练数据优化可调参数得知（参见神经网络）。

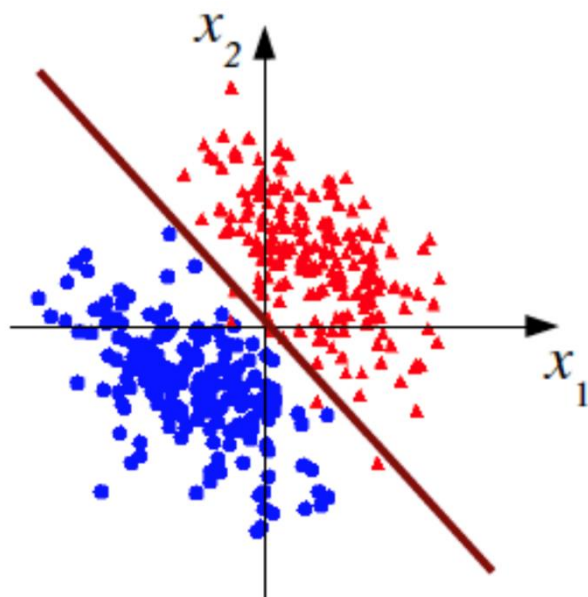
在某些情况下，基函数以点积的形式出现，即

$$\vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) = K(\vec{x}_i, \vec{x}_j)$$

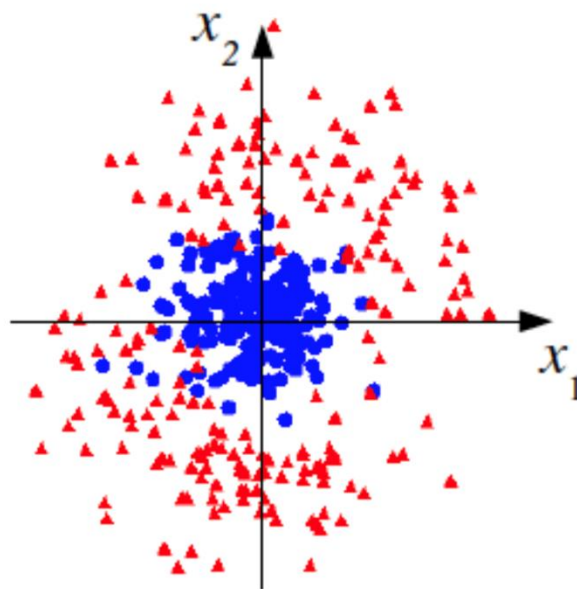
此时只需要“核函数”（kernel function）： $K(\vec{x}_i, \vec{x}_j)$ 。

线性决策边界

仅当信号和本底都服从方差相同均值不同的多变量高斯分布时，线性决策边界才是最优的



在某些情况下，线性决策边界几乎没有任何用处

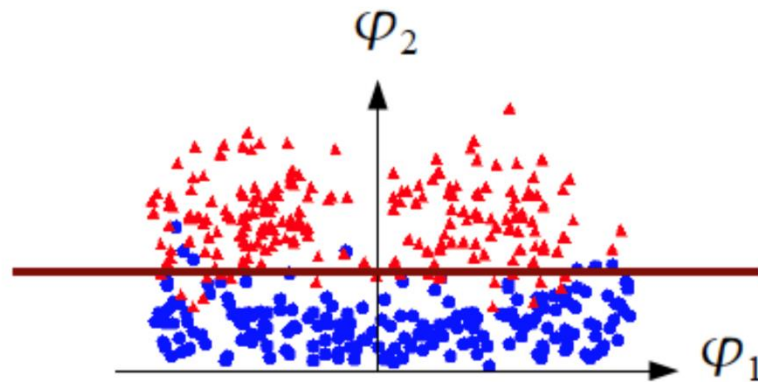
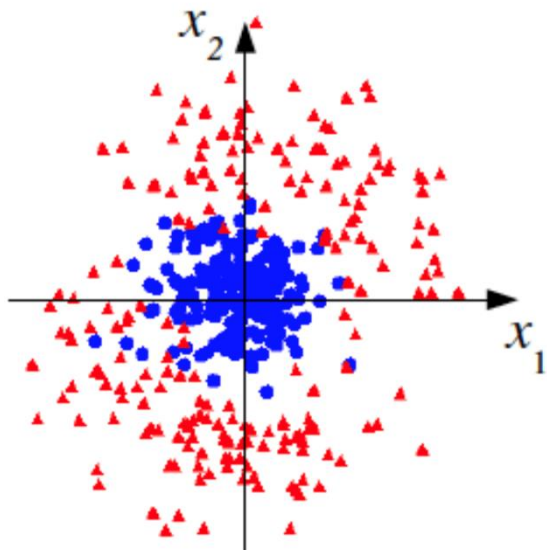


输入量的非线性变换

我们可以找到一种变换 $(x_1, \dots, x_n) \rightarrow (\varphi_1(\vec{x}), \dots, \varphi_m(\vec{x}))$, 使得变换后的“特征空间”变量可以用线性边界更好地区分:

$$\varphi_1 = \tan^{-1}(x_2/x_1)$$

$$\varphi_2 = \sqrt{x_1^2 + x_2^2}$$



神经网络

神经网络源于对神经过程的模拟 (McCulloch and Pitts, 1943; Rosenblatt, 1962)

应用领域广泛，很多年来一度是在粒子物理实验中流行的唯一“先进的”多变量方法

可以把神经网络方法看作用来定义特征空间变换的参数化基函数的一个特殊方法

用训练数据调整这些参数，使得到的甄别量函数具有最佳性能

单层感知器

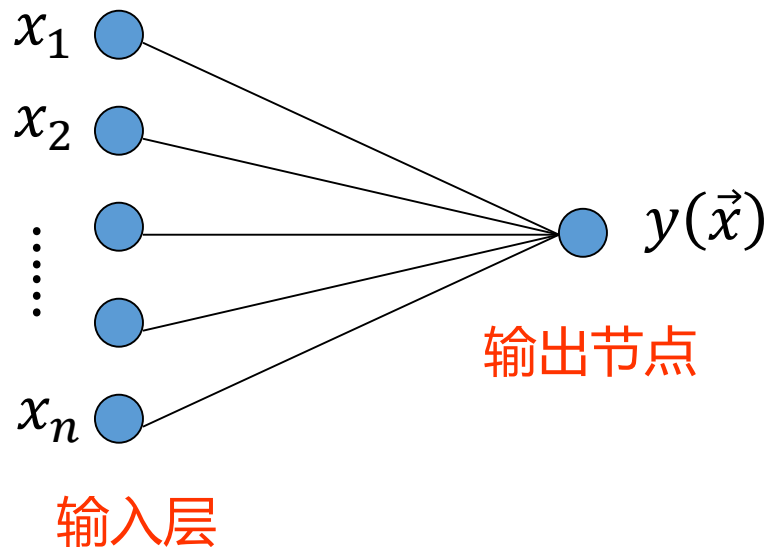
利用非线性函数 $y(\vec{x})$ 定义判别量：
$$y(\vec{x}) = h\left(\omega_0 + \sum_{i=1}^n \omega_i x_i\right)$$

其中 h 是非线性的单调激活函数，例如，逻辑S型函数

$$h(x) = (1 + e^{-x})^{-1}$$

如果激活函数是单调的，得到的 $y(\vec{x})$ 等价于原始的线性判别量。

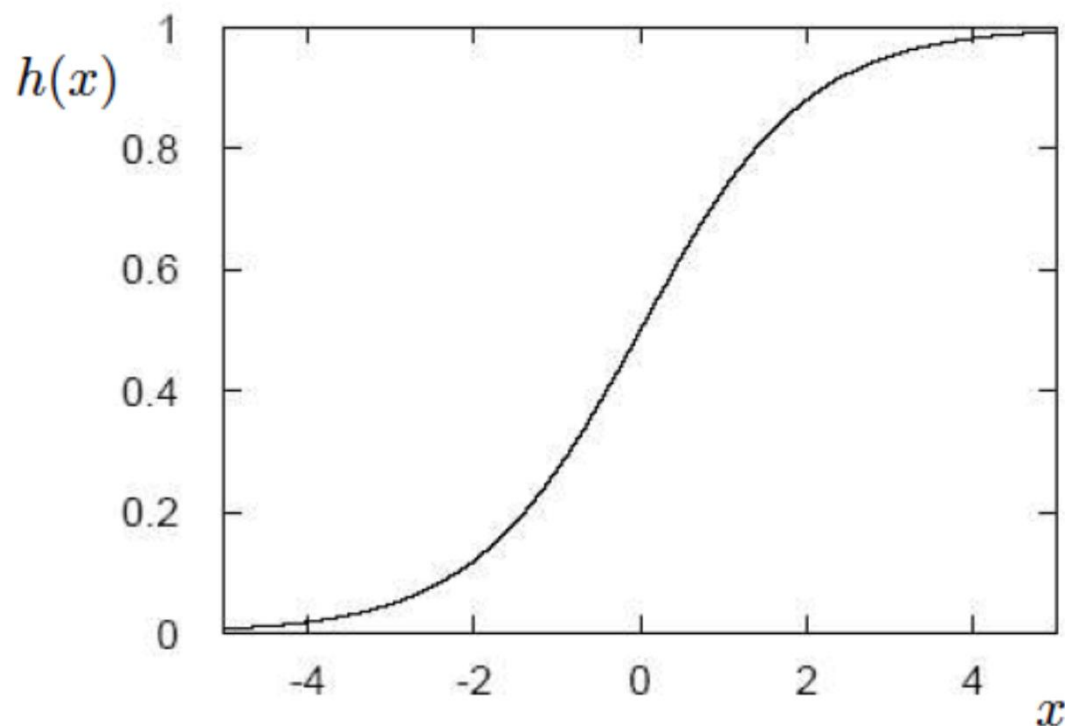
这是“推广的线性模型”的一个例子，称为单层感知器。



激活函数

激活函数 $h(\cdot)$ 经常采用逻辑S型函数

$$h(x) = \frac{1}{1 + e^{-x}}$$



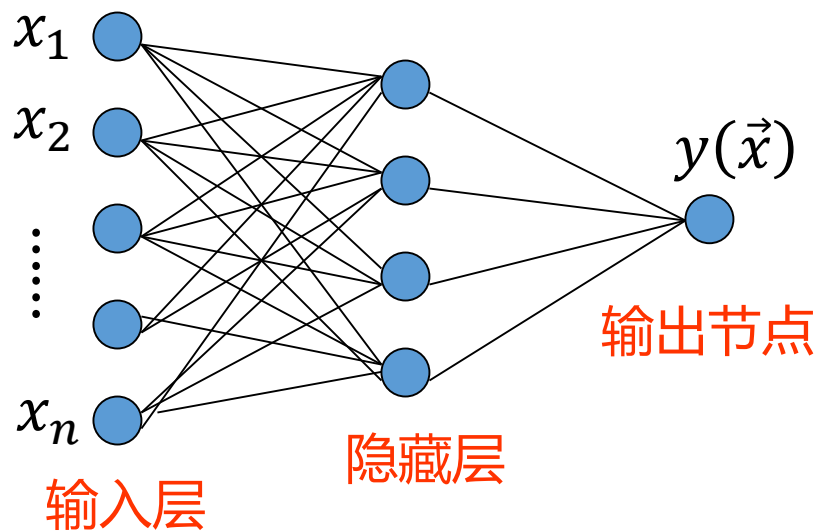
多层感知器 (MLP)

利用同样的想法定义，既可以定义输出 $y(\vec{x})$ ，也可以定义构成“隐藏层”的一组变换输入量 $\varphi_1(\vec{x}), \dots, \varphi_m(\vec{x})$ ：

权重中的上标
表示第几层

$$\varphi_i(\vec{x}) = h \left(\omega_{i0}^{(1)} + \sum_{j=1}^n \omega_{ij}^{(1)} x_j \right)$$

$$y(\vec{x}) = h \left(\omega_{10}^{(2)} + \sum_{j=1}^m \omega_{1j}^{(2)} \varphi_j(\vec{x}) \right)$$



这种构造称为多层感知器，是基本的神经网络模型；很容易推广到更多隐藏层。

神经网络的整体结构

定理：如果单隐藏层的MLP的隐藏层有足够多的节点，那么它可以任意好地近似最优的决策边界。

这个定理对于任意连续的非多项式激活函数都成立
(Leshno, Lin, Pinkus and Schocken (1993) Neural Networks 6, 861-867)

然而，需要的节点数目可能非常大，从而无法利用有限的训练数据实现好的训练。

深度神经网络的最新进展表明，多个隐藏层具有重要的优越性。

对于深度学习在粒子物理中的应用，参见：
Baldi, Sadowski and Whiteson, Nature Communications 5 (2014); arXiv:1402.4735

网络训练

每个训练事例的类型都已知，例如，对于事例 a ,

$\vec{x}_a = (x_1, \dots, x_n)$ 输入变量

$t_a = 0, 1$ 每个事例类型的数值标记（目标值）

令 $\vec{\omega}$ 表示网络所有权重的集合，可以通过最小化“误差函数”的平方和确定权重的最佳值

$$E(\vec{\omega}) = \frac{1}{2} \sum_{a=1}^N |y(\vec{x}_a, \vec{\omega}) - t_a|^2 = \sum_{a=1}^N E_a(\vec{\omega})$$

每个事例对误差函数的贡献

$E(\vec{\omega})$ 的数值最小化

考虑梯度下降法：从权重空间的某个猜测的初值， $\vec{\omega}^{(1)}$ 在最大下降的方向取一个小步长。对于从 τ 到 $\tau + 1$,

$$\vec{\omega}^{(\tau+1)} = \vec{\omega}^{(\tau)} - \eta \nabla E(\vec{\omega}^{(\tau)})$$

如果我们对整个误差函数 $E(\vec{\omega})$ 这么做，梯度下降的效果差得惊人，最好使用“共轭梯度”。

但是，梯度下降对于在线方法是有用的，即当我们对于每个训练事例 a 更新 $\vec{\omega}$ 时（对所有训练事例作循环）：

$$\vec{\omega}^{(\tau+1)} = \vec{\omega}^{(\tau)} - \eta \nabla E_a(\vec{\omega}^{(\tau)})$$

误差的向后传递

误差的向后传递是梯度下降法所需要的求导数的一种算法

神经网络输出可以写为 $y(\vec{x}) = h(u(\vec{x}))$, 其中

$$u(\vec{x}) = \sum_{j=0} \omega_{1j}^{(2)} \varphi_j(\vec{x}), \quad \varphi_j(\vec{x}) = h\left(\sum_{k=0} \omega_{jk}^{(1)} x_k\right)$$

其中, 定义 $\varphi_0 = x_0 = 1$, 并且对前一层节点的求和从零开始, 以包含偏移量。

以事例 a 为例,

$$\frac{\partial E_a}{\partial \omega_{1j}^{(2)}} = (y_a - t_a) h'(u(\vec{x})) \varphi_j(\vec{x})$$

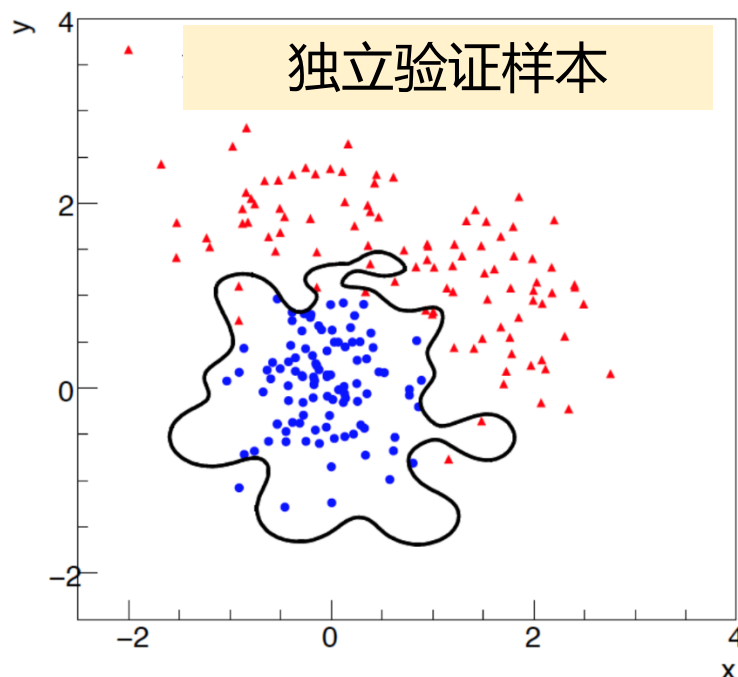
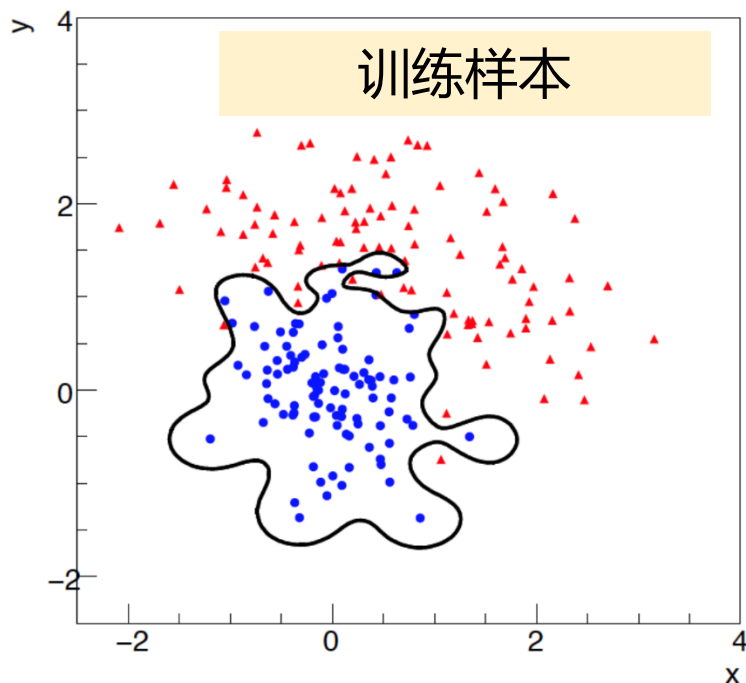
激活函数的导数

求导的链式法则给出所有需要用到的导数

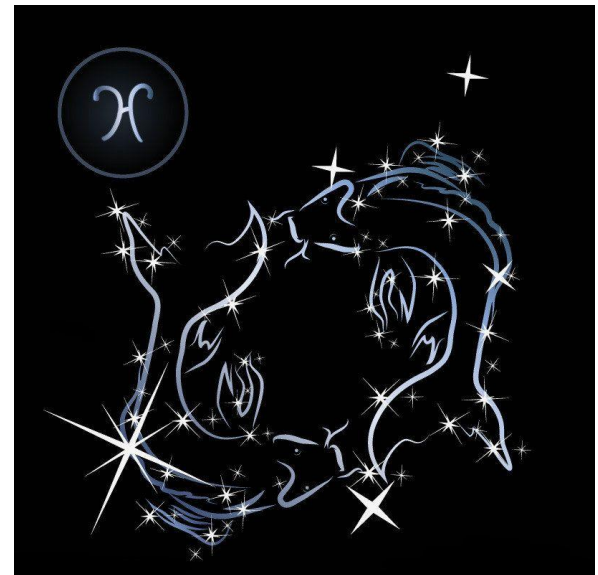
过度训练 (overtraining)

在分类器中包含更多参数可以使得决策边界非常灵活，例如，在神经网络中采用更多节点和层数。

“灵活”的分类器可能会过于迎合训练样本，相同的决策边界在独立的验证数据样本中性能不会太好 → “过度训练”

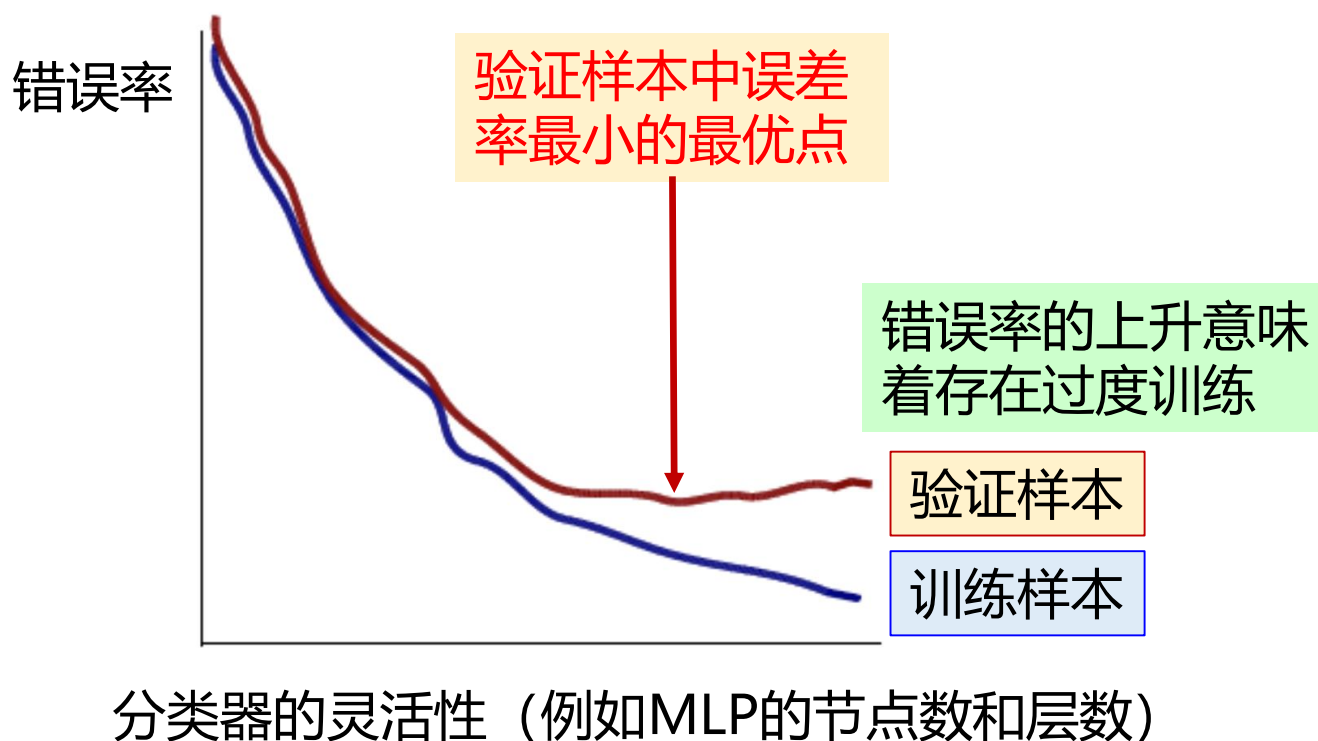


过度训练/过拟合




监控过度训练

如果监控验证样本和训练样本中误鉴别事例的比率（或者，误差函数 $E(\vec{\omega})$ ），当决策边界变得更灵活时，两个样本的误鉴别比率通常都会下降：



关于神经网络的输入变量问题

问题：是否输入量越多越好？

较少的输入量  较少的可调参数

利用有限的训练样本，可以很好地确定参数

如果输入量之间中有很强相关情形，往往可以只保留一个。

如果输入量对甄别没有太大影响，一般应弃之。

神经网络利用了较高阶矩的联合概率密度函数 $f(\vec{x}|H)$ ，也许训练样本无法得到较好的模型描述它们：

 最好简化 $t(\vec{x})$ ，只要它还能恰当地描述样本。

避免输入量和要研究的信号特征量相关。

本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

极端情况下的拟合优度检验

前面讲了统计检验的甄别问题，但在实际情况中还要处理极端情况下无效假设的拟合优度检验问题。

任意投掷一枚硬币，结果为正面与反面的概率都是0.5。

如果有人声称对此进行了检验。投了20次，得到了17次正面的结果。那么能否断定得到正面的概率应该是

$$p_h = 0.85 \pm 0.08$$

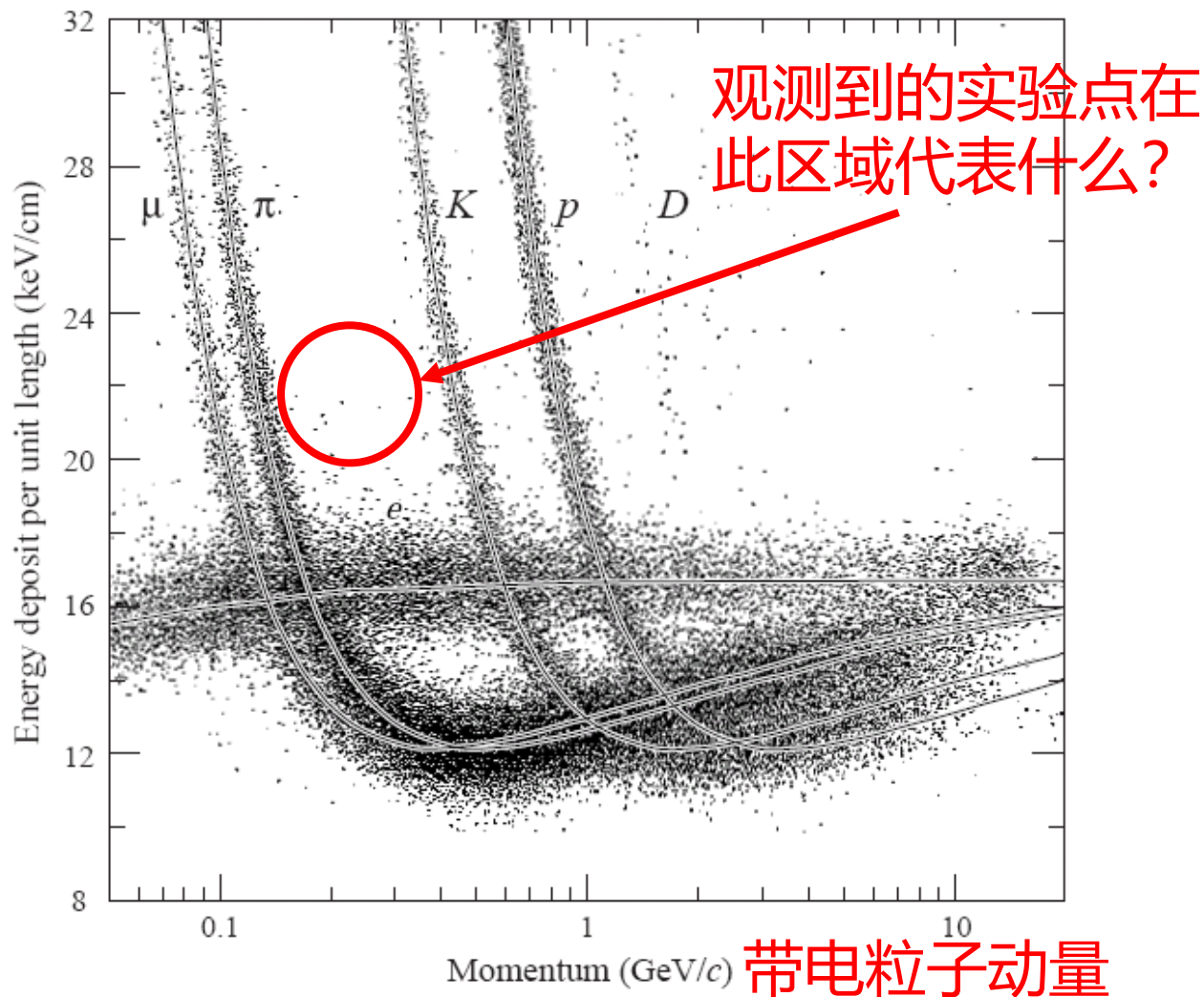
也就是说与预期值 0.5 有 4 倍标准差呢？

问题：理论上允许这样的极端情况出现吗？

或者说，与这种极端情况相等或更高的概率有多大？

例：粒子鉴别的常见问题

粒子在介质中单位长度的能量沉积



检验拟合优度

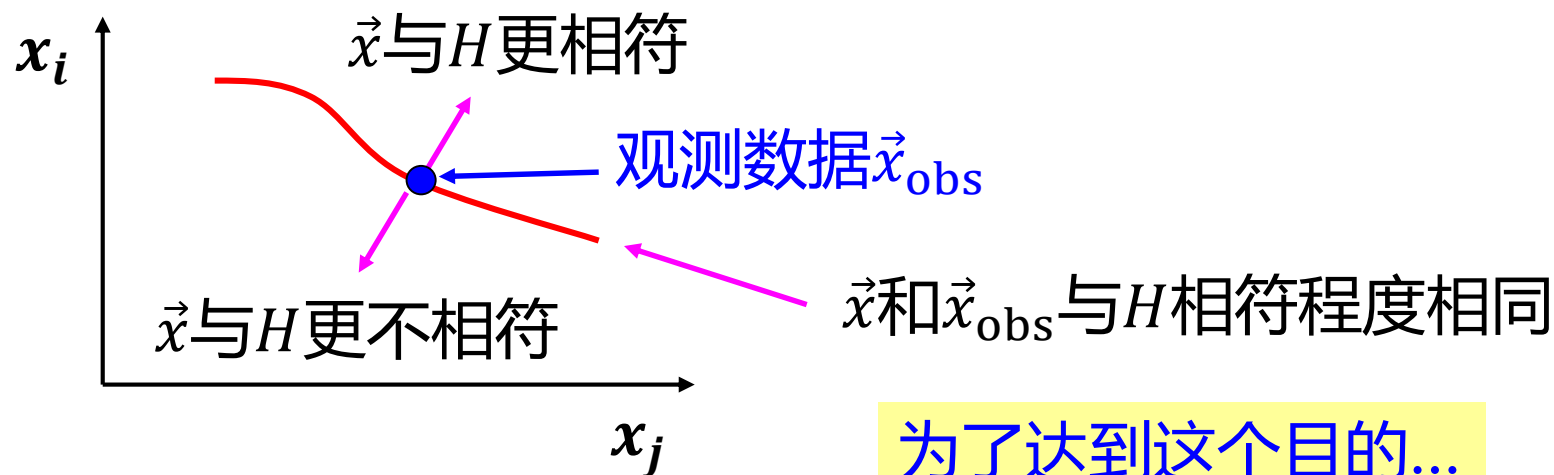
如果假设 H 对一组观测量 $\vec{x} = (x_1, x_2, \dots, x_n)$ 的预言为 $f(\vec{x}|H)$,

我们在 \vec{x} -空间观测到数据点: \vec{x}_{obs} 。

从数据来看, 对假设 H 的正确与否能得出什么结论呢?

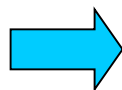
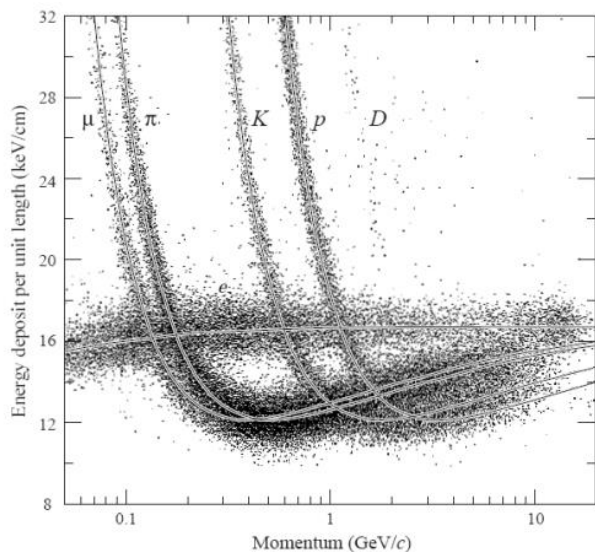


需要在 \vec{x} -空间确定一个曲面, 曲面一侧的点相比于观测点 \vec{x}_{obs} 与 H 符合程度更好, 另一侧符合程度更差。曲面上的点跟观测点 \vec{x}_{obs} 与 H 的符合程度相同。



检验统计量与拟合优度

通常需要构造检验统计量 $t(\vec{x})$ ，其大小可反映出 \vec{x} 与 H 的符合程度，例如



$$t = \left(\frac{(\mathrm{d}E/\mathrm{d}x)_K^{\mathrm{Th}} - (\mathrm{d}E/\mathrm{d}x)_K^{\mathrm{obs}}}{\sigma} \right)^2$$

小 t



数据与 H 更符合

大 t



数据与 H 更不符合

由于概率密度函数 $f(\vec{x}|H)$ 已知，因此在 H 假设条件下检验统计量 t 的概率密度函数 $g(t|H)$ 完全可以确定。

p 值定义

用 p 值表示假设检验的拟合优度：

p = 观测到数据 \vec{x} 与假设 H 的符合程度不好于实际数据 \vec{x}_{obs} 与 H 的符合程度的概率



注意: 这不是 H 为真的概率。

经典统计不讨论 $P(H)$ ，除非 H 表示可重复观测。

贝叶斯统计把 H 当成随机变量，并利用贝叶斯定理得到

$$p(H|t) = \frac{P(t|H)\pi(H)}{\int P(t|H)\pi(H)dH}$$

$\pi(H)$: H 的先验概率

对所有可能作归一化积分

p 值的分布

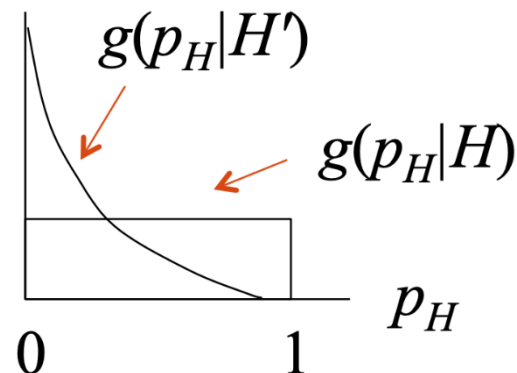
p 值是数据的函数，其本身也是有一定分布的随机变量。
如果从检验统计量 $t(\vec{x})$ 得到假设 H 的 p 值：

$$p_H = \int_t^{\infty} f(t'|H) dt'$$

在 H 的假设下， p 值的概率密度函数为

$$g(p_H|H) = \frac{f(t|H)}{|\partial p_H / \partial t|} = \frac{f(t|H)}{f(t|H)} = 1 \quad (0 \leq p_H \leq 1)$$

对于连续数据，在 H 的假设下，
 $p_H \sim U(0,1)$ ；
在很多备择假设下，聚集于零附近。



利用 p 值定义 H_0 的检验

H_0 假设的 p 值 (p_0) 小于 α 的概率为

$$P(p_0 \leq \alpha | H_0) = \alpha$$

先在原始的数据空间 (\vec{x}) 定义临界域, 然后表示为标量检验统计量 $t(\vec{x})$ 。

我们可以进一步定义显著性水平为 α 的 H_0 检验的临界域为 $p_0 \leq \alpha$ 的数据空间的集合。

形式上, p 值仅与 H_0 有关, 但是得到的检验还与相对于给定的备择假设 H_1 的功效有关。

例：拟合优度检验

投硬币 N 次，观测到正面朝上的次数 n_h 服从二项分布：

$$f(n_h; p_h, N) = \frac{N!}{n_h! (N - n_h)!} p_h^{n_h} (1 - p_h)^{N - n_h}$$

假设 H ：硬币是公平的 (正面朝上的 $p_h = p_t = 0.5$)

取拟合优度检验统计量： $t = |n_h - N/2|$

投 $N = 20$ 次，观测到 $n_h = 17$ ，则 $t_{\text{obs}} = \left| 17 - \frac{20}{2} \right| = 7$

在 t 空间，相比于 t_{obs} ，与 H 符合程度相同或更差的区域为

$$t = (n_h - N/2) \geq 7$$

$$p\text{值} = P(n_h = 0, 1, 2, 3, 17, 18, 19, 20) = \sum_{i \leq 8} f_i \approx 0.0026$$

拟合优度检验中的问题

问题： p 值 = 0.0026时，是否意味着 H 假设是错的？

p 值并不回答此问题。它只是给出与观察到的结果一样，与 H 假设不符或者高于 H 假设($p_h = p_t = 0.5$)的概率。

p 值 = “偶然” 得到如此奇怪结果的概率

一种实用的检验方法是在同样的假设下，多次重复试验，每次产生同样数目的事例。检查如此奇怪的结果发生的概率是否与 p 值相当。

本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

泊松计数实验

假设做一个计数实验，观测到 n 个事例（可能是信号，也可能是本底）。

泊松模型：

$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

s = 信号事例的均值(期待值)

b = 本底事例的均值(期待值)

目标：对 s 给出一个统计推断（假设检验），例如

$H_0: s = 0$ （拒绝 H_0 意味着“发现了信号过程”）

$H_0: s \neq 0$ （未被拒绝的值就是置信区间）

两种情况都需要考虑相关的备择假设是什么。

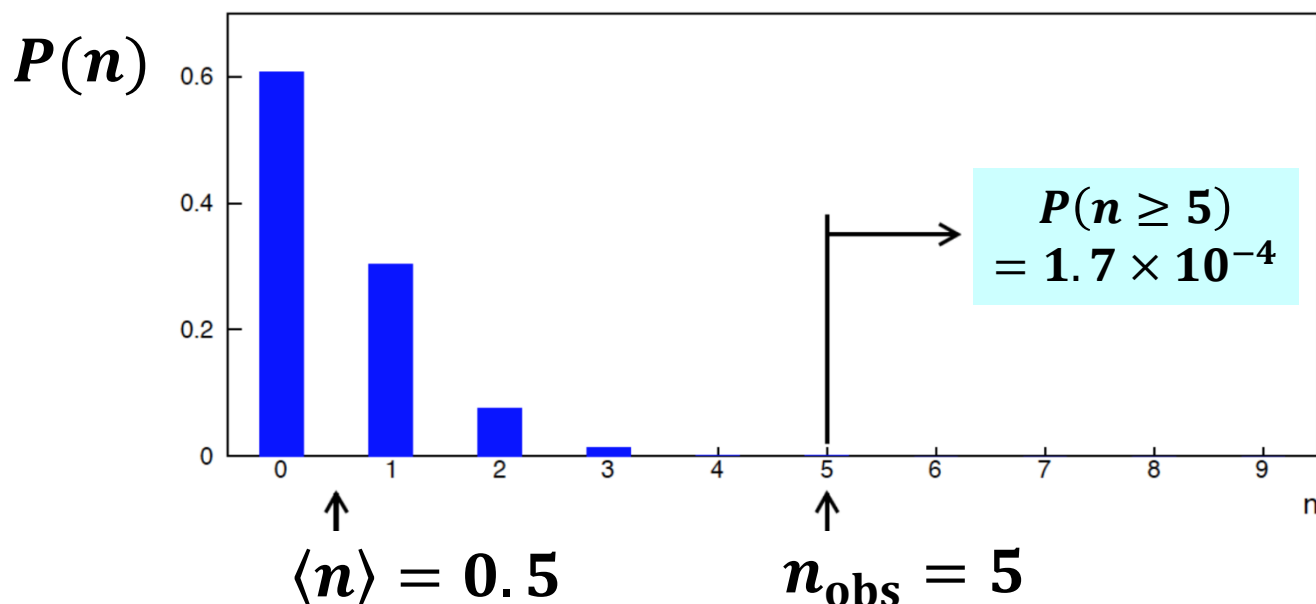
泊松计数实验：新发现的 p 值

假设 $b = 0.5$ （已知），观测到事例数 $n_{\text{obs}} = 5$ 。

问题：我们是否应当宣称存在新发现的迹象（evidence）？

计算 H_0 假设 ($s = 0$) 的 p 值：

$$p\text{值} = P(n \geq 5; b = 0.5, s = 0) \approx 1.7 \times 10^{-4} \neq P(s = 0)!$$

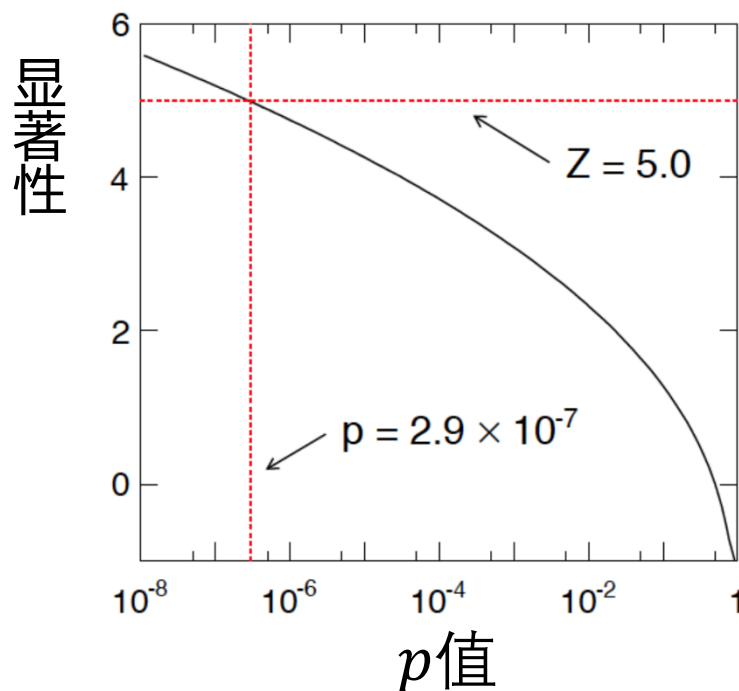


泊松计数实验：新发现的显著性

与 $p = 1.7 \times 10^{-4}$ 等价的显著性： $Z = \Phi^{-1}(1 - p) = 3.6$

通常在 $Z > 5$ ($p < 2.9 \times 10^{-7}$) 时宣称新发现： 5σ 效应

计算 H_0 假设 ($s = 0$) 的 p 值：

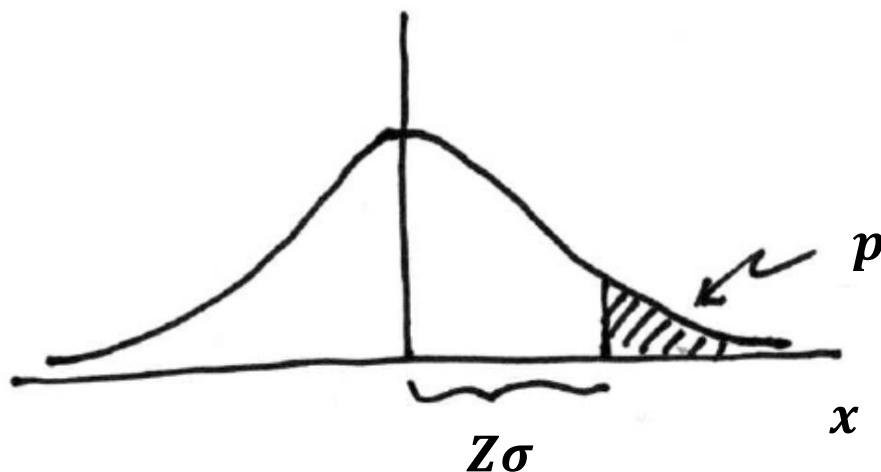


p 值想定量评估纯本底假设下数据涨落出信号形状的概率。

p 值并不涵盖以下问题：隐藏的系统误差，信号模型的合理性，数据与信号模型的符合程度，“查看别处效应” (look-elsewhere effect)，等等。

从 p 值得到显著性

我们经常将显著性 Z 定义为，高斯变量在一个方向涨落得到相同 p 值所对应的标准差的倍数。



$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 1 - \Phi(Z) \quad 1 - \text{TMath}::\text{Freq}(\text{double } Z)$$

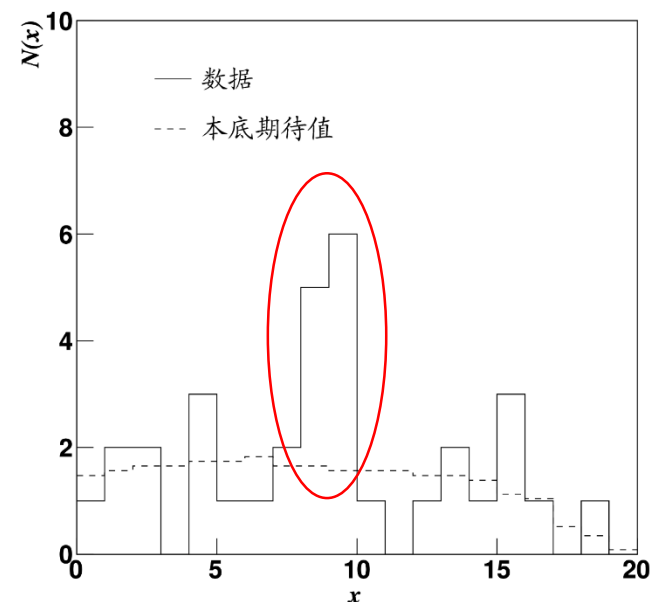
$$Z = \Phi^{-1}(1 - p)$$

$$\text{TMath}::\text{NormQuantile}(\text{double } p)$$

直方图中峰结构的显著性

假设我们对每个事例测量 x 的值：

观测的每个区间频数都是泊松随机变量，其均值由图中虚线给出。



对于峰所在的两个区间，共有11个事例，本底为 $b = 3.2$ 。

$s = 0$ 假设的 p 值为：

$$p(n \geq 11; b = 3.2, s = 0) = 5.0 \times 10^{-4}$$

直方图中峰结构的显著性（续）

但是，我们知道到哪里去寻找“信号峰”吗？

➡ 查看别处效应：求在直方图任意位置看到同样显著的峰的概率

我们查看过多少区间与分布？

➡ 查看上千个区间与分布，我们将发现大约 10^{-3} 的效应

观测到的宽度与期待的 x 的分辨率是否一致？

➡ 选取 x 的信号窗口应当等于几倍于期待分辨率

我们是否调整过事例筛选条件以“增强”这个“信号峰”？

➡ 冻结筛选条件，用新数据重复整个分析

我们应当发表这个结果吗？如何发表？

为什么选取 5σ 为标准?

粒子物理与核物理实验在 $p \geq 2.9 \times 10^{-7}$ 时宣称有新发现，
对应于显著性 $Z = \Phi^{-1}(1 - p) = 5$: 5σ 效应

为什么选取这么高的门槛?

宣称错误发现的“代价”很高

对于模型中的系统不确定度不完全有把握

对于查看别处效应不完全有把握

数据暗示的信号从先验上看极不可能，例如洛伦兹不变性破坏

我们还需要考虑数据与新物理相符的程度，而不仅仅是与零假设不一致的程度： p 值仅仅是新发现的第一步!

为什么选取 5σ 为标准（续）？

p 值的主要作用是量化纯本底模型因统计涨落得到跟观测到的信号一样甚至更显著的概率。

p 值本身不是为了防止隐藏的系统问题或者为宣称重要发现设置很高的标准。

在确立新发现的过程中， p 值告诉我们观测到的现象不是简单的统计涨落而是一种“效应”，然后问题变为：这种效应是新物理导致的还是系统问题导致的？

如果做了查看别处效应（LEE）的研究，宣称新发现的阈值可能接近 3σ 而不是 5σ 。

本章要点

- 假设，检验，显著水平，功效，临界域
- 粒子物理中的统计检验
- 奈曼-皮尔逊引理和检验统计量的构造
 - 费舍尔甄别函数与神经网络
- 检验拟合优度， p 值定义与应用
- 信号观测的显著程度
- 皮尔逊的 χ^2 检验

皮尔逊卡方统计量

观测数据（实验）： $\vec{n} = (n_1, \dots, n_N)$, n_i 相互独立

预期均值（模型）： $\vec{v} = (v_1, \dots, v_N)$

如何比较 \vec{n} 与 \vec{v} 是否相符？

皮尔逊卡方统计量：

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - v_i)^2}{\sigma_i^2}, \quad \sigma_i^2 = V[n_i]$$

如果 $n_i \sim \pi(v_i)$, 则 $V[n_i] = v_i$,
皮尔逊 χ^2 统计量变为

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - v_i)^2}{v_i}$$

皮尔逊卡方检验

如果 $n_i \sim N(\nu_i, \sigma_i^2)$, 则皮尔逊 χ^2 统计量 (记为 z) 服从自由度为 N 的卡方分布 ($z \sim \chi^2(N)$) :

$$f_{\chi^2}(z; N) = \frac{1}{2^{N/2} \Gamma(N/2)} z^{N/2-1} e^{-z/2}$$

如果 $n_i \sim \pi(\nu_i)$, 并且 $\nu_i \gg 1$ (实际应用中 $\nu_i > 5$ 即可), 则泊松分布可以近似为高斯分布, 因此, 皮尔逊 χ^2 统计量也服从卡方分布。

从数据得到的 χ^2 值可以给出对应的 p 值:

$$p = \int_{\chi^2}^{\infty} f_{\chi^2}(z; N) dz$$

`TMath::Prob(double chi2, int ndf)`

皮尔逊卡方检验 (续)

$$z \sim \chi^2(N) \rightarrow E[z] = N, V[z] = 2N$$

→ 经常有人以 χ^2/N 展示数据与模型的符合程度

最好同时给出 χ^2 和 N ，而不仅仅是比值 χ^2/N ，例如，同样是 $\chi^2/N = 1.5$ ， p 值可能相差很大：

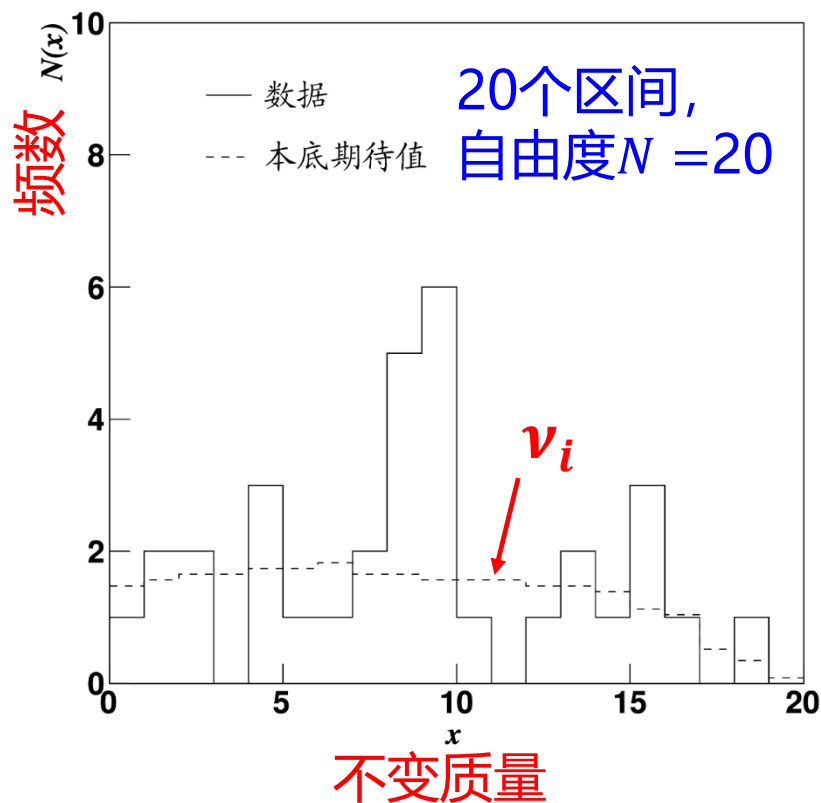
$$\begin{aligned}\chi^2/N = 15/10 &\Rightarrow p\text{值} = 0.13 \\ \chi^2/N = 150/100 &\Rightarrow p\text{值} = 9.0 \times 10^{-4}\end{aligned}$$

若 N 很大，即使 χ^2/N 稍大于 1， p 值也很小，即拟合优度很差。

如果 $n_{\text{tot}} = \sum_{i=1}^N n_i$ 固定， n_i 服从二项分布， $p_i = v_i/n_{\text{tot}}$ ，则

$$\chi^2 = \sum_{i=1}^N \frac{(n_i - p_i n_{\text{tot}})^2}{p_i n_{\text{tot}}} \sim \chi^2(N-1) \quad \text{要求 } p_i n_{\text{tot}} \gg 1$$

例：皮尔逊卡方检验



$$\chi_{\text{obs}}^2 = \sum_{i=1}^N \frac{(n_i - v_i)^2}{v_i} = 29.8$$

$$p\text{值} = \int_{\chi_{\text{obs}}^2}^{\infty} f_{\chi^2}(z; 20) dz = 0.073$$

问题：这个结果对吗？

例：皮尔逊卡方检验（续）

答： p 值计算不对。许多区间只有很少或没有计数，此时皮尔逊卡方统计量不服从自由度为 N 的 χ^2 分布。

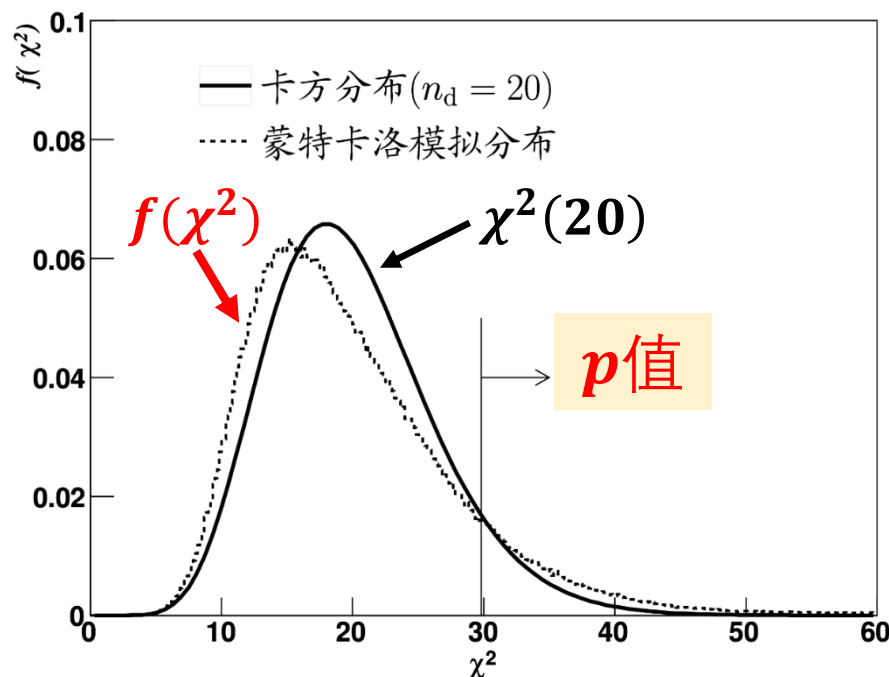
利用蒙特卡罗方法得到皮尔逊 χ^2 统计量的概率密度 $f(\chi^2)$ 。

模拟产生样本 $n_i \sim \pi(v_i)$,
($i = 1, \dots, N$)
每次计算 χ^2 值并填入直方图。

➡ $f(\chi^2)$

MC pdf: $p\text{值} = 0.11$

$\chi^2(20)$ pdf: $p\text{值} = 0.073$



对于统计检验的评论

在实际问题中，我们常常需要在低统计量的情况下，判断观测到的现象是否为真正的物理信号。利用 p 值的大小可以表示结果是否为已知过程的极端情形，但每个人的信心不同，会造成同一个 p 值，结论却完全不一样的现象。

在统计误差范围内无新迹象

结果虽然在统计误差范围，但有可能是新物理的信号

发现了新物理的信号，误差为...

历史上类似故事的发生很多：

J/ψ 介子的发现， W 玻色子的发现，顶夸克的发现...

本章总结

□ 统计检验:

检验在何种程度上, 数据与假设相符。

□ 检验统计量:

将矢量 \vec{x} 简化为一个或几个分量的矢量 $t(\vec{x})$

□ 检验的要点:

临界域, 显著性水平, 功效, 纯度, 效率。

□ 奈曼-皮尔逊引理:

在给定效率条件下, 给出纯度最大区。

□ 构造检验统计量:

最好是似然比, 但通常需太多待定参数。

□ 统计分析中两种方法:

费舍尔甄别函数(线性的); 神经网络(非线性的)。

本章总结（续）

❑ 检验拟合优度， p 值定义与应用

p 值为得到数据像已观测的结果一样与假设不符或更不符合的概率。

❑ 信号观测的显著程度

很复杂，许多具有 10^{-4} 效应的结果最终证明是统计涨落的受害者。

❑ 皮尔逊 χ^2 检验

广泛用于检验统计量。对于小样本数据，它将不服从 χ^2 的概率密度函数分布。但仍可用蒙特卡罗得到概率密度函数分布。

p 值与假设检验

根据 p 值的定义，对 H 假设拟合优度的检验可以通过计算 p 值的大小完成。但是应注意以下两点：

- 在 p 值定义中不涉及别的假设
 - p 值是一个随机变量，前面的显著性水平在检验时已经被指定为常数
-
- ➡ 若 H 为真，则对于连续的 \vec{x} ， p 在 $[0,1]$ 范围内均匀分布
 - ➡ 若 H 非真，则 p 的概率密度函数通常很接近零

概率密度估计 (PDE)

对于包含两类事例的数据样本 \vec{x} ，可以构造概率密度函数 $p(\vec{x}|H_0)$ 和 $p(\vec{x}|H_1)$ 的非参数化估计量，并以此构造似然比作为甄别量函数：

$$y(\vec{x}) = \frac{\hat{p}(\vec{x}|H_1)}{\hat{p}(\vec{x}|H_0)}$$

n 维直方图是构造这种非参数化估计量最粗暴的方法，实际上可以有更好的方法。

相关与独立

一般来说，多变量分布 $p(\vec{x})$ 不能因子化成每个变量的边缘分布的乘积：

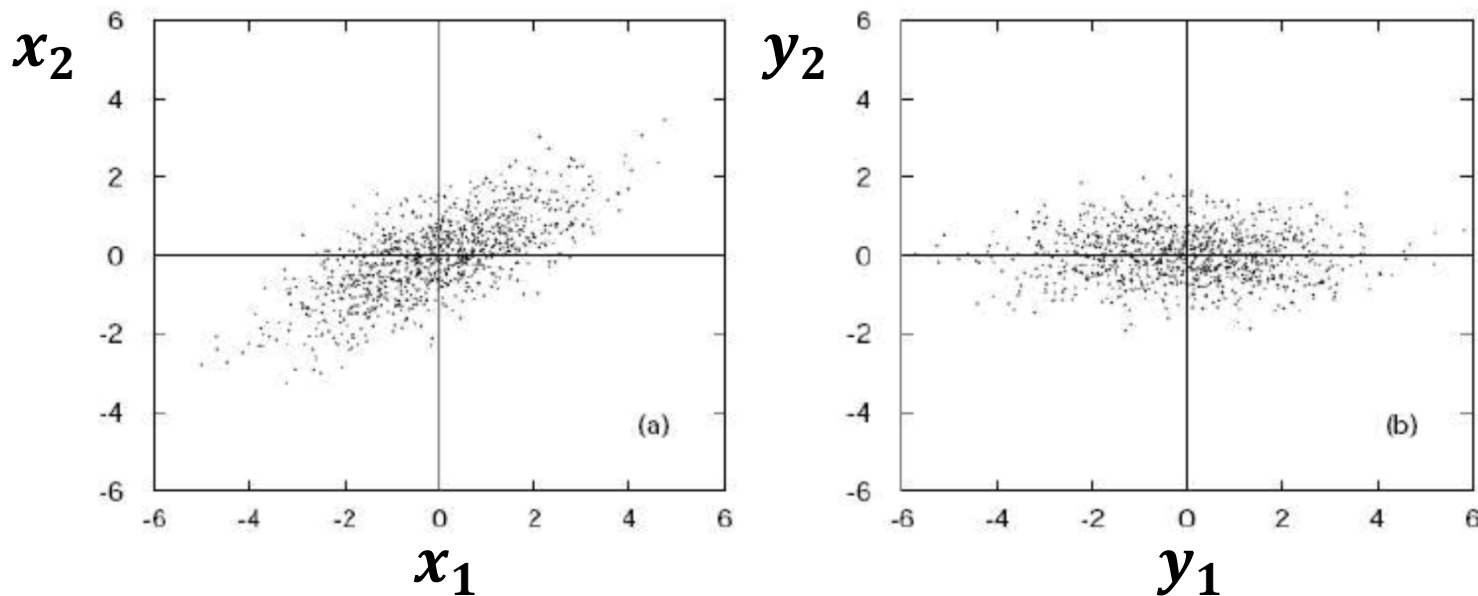
$$p(\vec{x}) = \prod_{i=1}^n p_i(x_i) \quad \text{仅当}\vec{x}\text{的分量相互独立时才成立}$$

最重要的一点是， \vec{x} 的分量之间通常方差不为零，即它们之间是相关的：

$$V_{ij} = \text{cov}[x_i, x_j] = E[x_i x_j] - E[x_i]E[x_j] \neq 0$$

输入变量的退相关

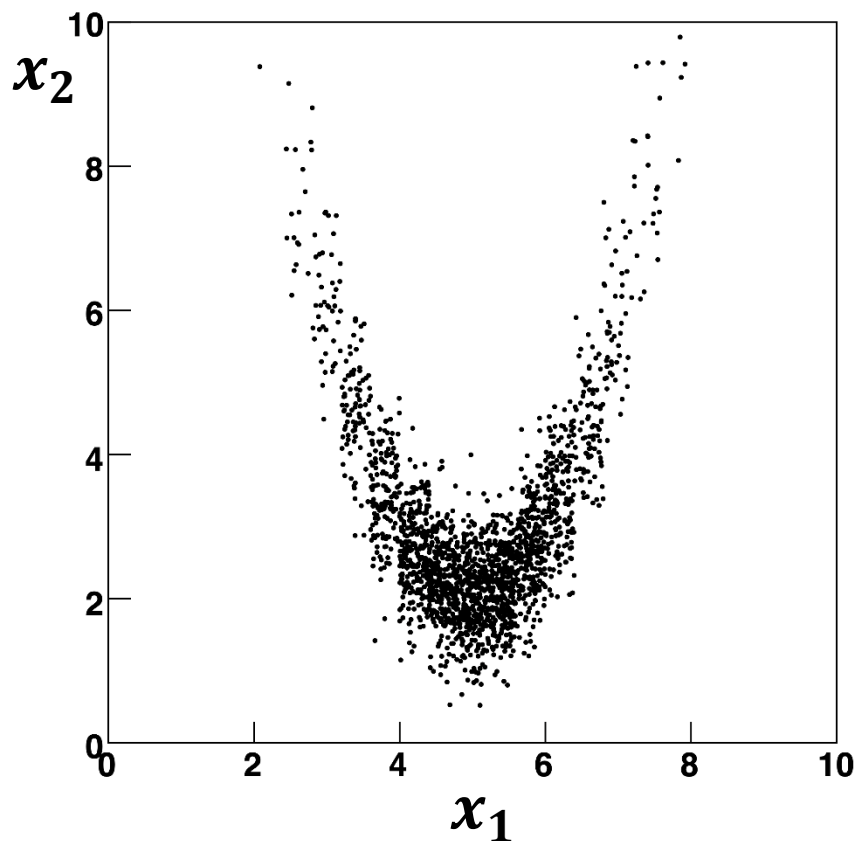
我们可以通过线性变换定义一组不相关的输入变量，即寻找矩阵 A ，使得 $\vec{y} = A\vec{x}$ 的协方差为零： $\text{cov}[y_i, y_j] = 0$



我们可以通过线性变换定义一组不相关的输入变量，即寻找矩阵 A ，使得 $\vec{y} = A\vec{x}$ 的协方差为零： $\text{cov}[y_i, y_j] = 0$

退相关并不足够

通常来说，多变量概率密度函数 $p(\vec{x})$ 是非线性的，不相关的变量不意味着相互独立。



协方差为零，但不独立

$$p(x_2|x_1) \equiv \frac{p(x_1, x_2)}{p_1(x_1)} \neq p_2(x_2)$$

$$p(x_1, x_2) \neq p_1(x_1)p_2(x_2)$$

线性检验统计量

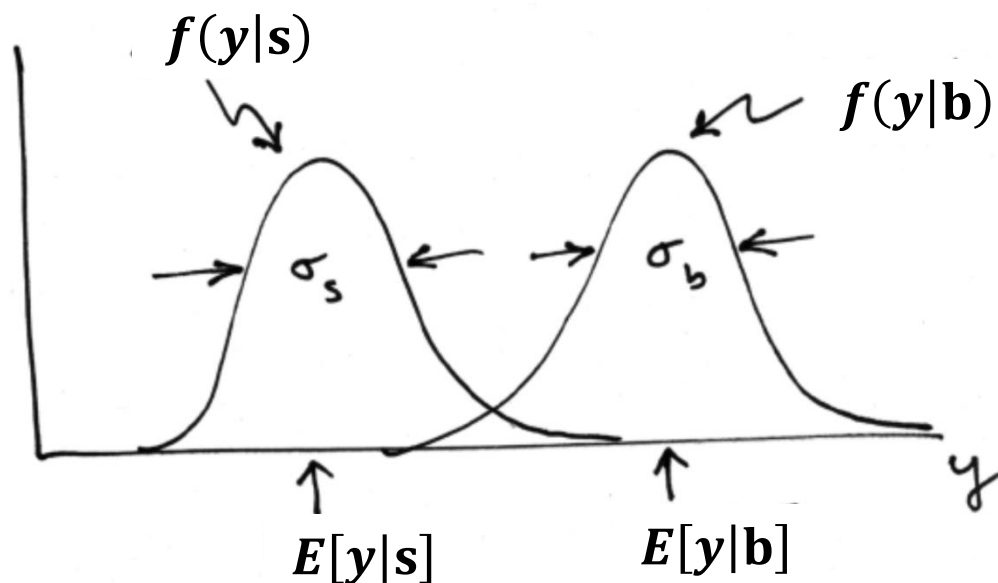
假设存在 n 个输入变量: $\vec{x} = (x_1, \dots, x_n)$

考虑线性函数: $y(\vec{x}) = \sum_{i=1}^n \omega_i x_i$

对于给定的系数 $\vec{\omega} = (\omega_1, \dots, \omega_n)$, 可得到概率密度 $f(y|s)$ 和 $f(y|b)$:

$\vec{\omega}$ 的选择应当满足

- $|E[y|s] - E[y|b]|$ 尽可能大
- σ_s 和 σ_b 尽可能小



线性检验统计量：费舍尔判别量

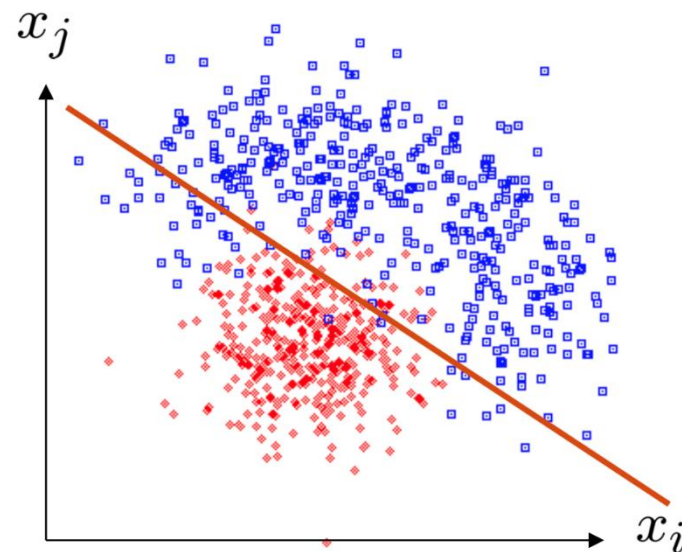
选择 $\vec{\omega}$ 使得 $J(\vec{\omega})$ 最大化：
$$J(\vec{\omega}) = \frac{(E[y|s] - E[y|b])^2}{V[y|s] + V[y|b]}$$

$$\frac{\partial J}{\partial \omega_i} = 0 \Rightarrow \begin{cases} \vec{\omega} \propto W^{-1}(\vec{\mu}_b - \vec{\mu}_s) \\ W_{ij} = \text{cov}[x_i, x_j|s] + \text{cov}[x_i, x_j|b] \\ \mu_{i,s} = E[x_i|s], \quad \mu_{i,b} = E[x_i|b] \end{cases}$$

费舍尔判别量：

$$y(\vec{x}) = \omega_0 + \sum_{i=1}^n \omega_i x_i$$

这个检验的临界域的边界是常数 $y(\vec{x})$ 确定的曲面。



例：高斯数据的费舍尔判别量

假设输入变量的概率密度 $f(\vec{x}|\mathbf{s})$ 和 $f(\vec{x}|\mathbf{b})$ 都是多维高斯分布，其方差相同但是均值不同：

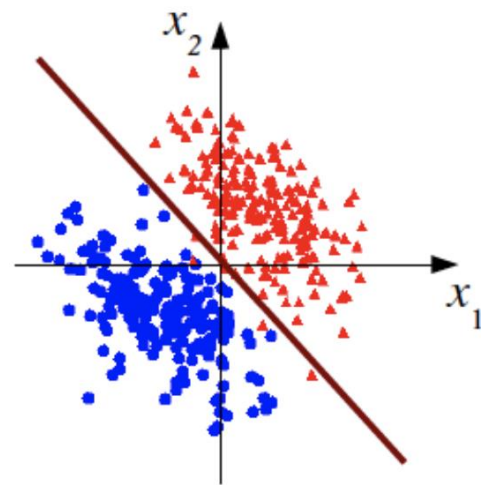
$$f(\vec{x}|\mathbf{s}) = N(\vec{\mu}_{\mathbf{s}}, V)$$

$$f(\vec{x}|\mathbf{b}) = N(\vec{\mu}_{\mathbf{b}}, V)$$

方差相同
 $V_{ij} = \text{cov}[x_i, x_j]$

此时可以证明费舍尔判别量为

$$y(\vec{x}) \sim \ln \frac{f(\vec{x}|\mathbf{s})}{f(\vec{x}|\mathbf{b})}$$



即，这是似然比的单调函数，因此给出相同额临界域。
所以，这种情况下费舍尔判别量是最优的统计检验。

非线性统计检验：神经网络

如果不同假设下观测量的概率密度函数 $f(\vec{x}|H_0)$ 与 $f(\vec{x}|H_1)$ 不是高斯或协方差矩阵不同，费舍尔甄别方法不再适用。此时可以采用非线性统计检验方法，如神经网络

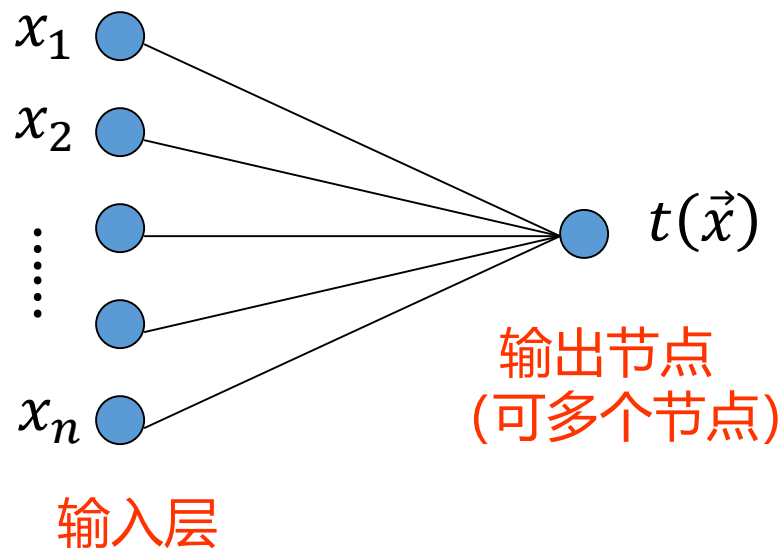
假设统计检验量
$$t(\vec{x}) = s \left(a_0 + \sum_{i=1}^n a_i x_i \right)$$

激活函数

$$s(u) = \frac{1}{1 + e^{-u}}$$

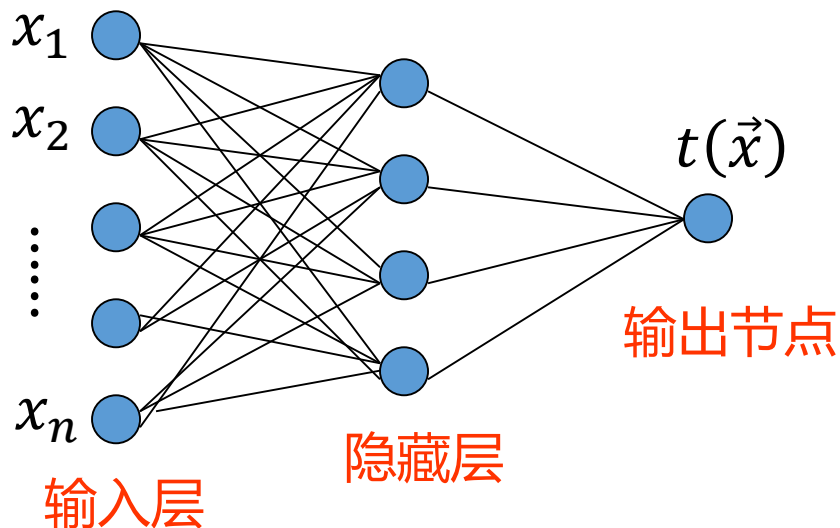
单层感知器

$s(u)$ 是单调函数，所以等效于线性的 $t(\vec{x})$



多层感知器的神经网络

推广到多层感知器



输出定义为

$$t(\vec{x}) = s \left(a_0 + \sum_{i=1}^n a_i h_i(\vec{x}) \right)$$

a_i, w_{ij} 为权重或者联结强度。

上一层节点函数可写为

$$h_i(\vec{x}) = s \left(\omega_{i0} + \sum_{j=1}^n \omega_{ij} x_j \right)$$

a_i, ω_{ij} : 权重或联结强度

节点越多



神经网络越接近优化的 $t(\vec{x})$

但参数更多，需要更大的训练样本！

神经网络中的误差函数最小化

参数取值通常由最小化误差函数确定

$$\varepsilon = E_0 \left[(t - t^{(0)})^2 \right] + E_1 \left[(t - t^{(1)})^2 \right]$$

$t^{(0)}$ 和 $t^{(1)}$ 为目标值，例如选 0 和 1 的逻辑S型函数的值

实际应用中，通常以训练样本的样本均值取代均值。

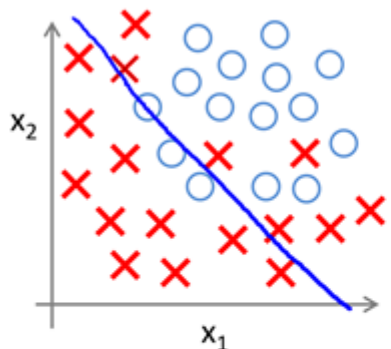
(调整参数值 = 神经网络的学习过程)

在粒子物理与核物理研究中，定义信号与本底两个样本，从样本中给出每个事例的相关观测量(动量、飞行时间...), 然后直接调用TMVA软件包(基于ROOT), 得到训练后的参数与输出量，并将它们用于待分析的事例，推断其是本底还是信号。具体应用参见下列网站

ROOT用户: <https://root.cern/tmva>

过度训练/过拟合

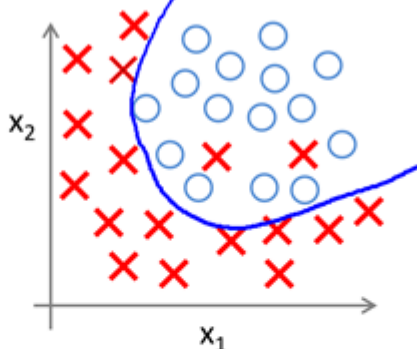
Example: Logistic regression



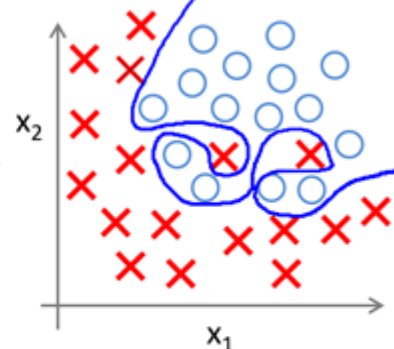
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)

"Underfit"

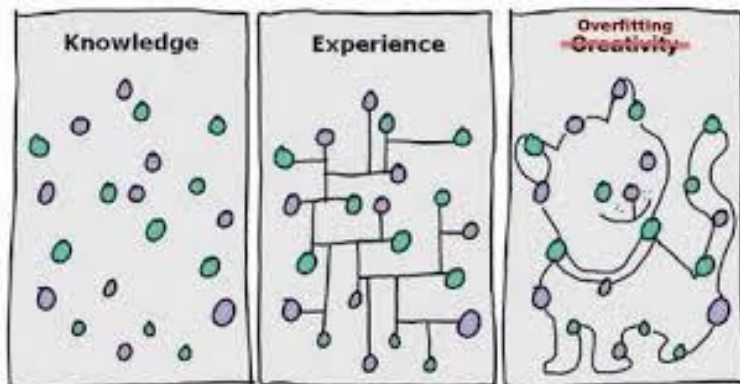


$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

"Overfit"



观测到一个信号的显著程度

假设观测 n 个事例，包含了

n_b = 已知过程(本底)的事例数 n_s = 新过程(或信号)的事例数

如果 n_b 和 n_s 服从泊松分布，均值为 b 和 s ，其和 $n = n_b + n_s$ 也服从泊松分布，均值为 $\nu = b + s$ ：

$$P(n; s, b) = \frac{(s + b)^n}{n!} e^{-(s+b)}$$

如果 $b = 0.5$ ，而且观测到 $n_{\text{obs}} = 5$

→ 可否据此声称该迹象为新的发现？

假设 $H: s = 0$ ，即只有本底过程的假设

→ 所谓的“无效假设”

观测到一个信号的显著程度(续)

对应的 p 值

$$p\text{值} = P(n \geq n_{\text{obs}}) = \sum_{n=n_{\text{obs}}}^{\infty} \frac{b^n}{n!} e^{-b} = 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n}{n!} e^{-b} \\ \approx 1.7 \times 10^{-4}$$

$p\text{值} \neq P(s = 0) !$



给出了得到这种极端结果的概率：
虽然很小但不为零！

潜在的问题之一

一个误导读者但又经常被使用的结果表达...

估计 ν_s 得到: $n_{\text{obs}} = 5$
估计 n 的标准差: $\sqrt{n} = 2.2$ } 信号
 $b = 0.5$



→ s 的估计值: $n_{\text{obs}} - b = 4.5 \pm 2.2$

与零仅有两倍的标准差?

实际想要的是: 均值 $b = 0.5$ 的泊松变量给出观测量大于 5 的概率是多少?

→ 概率为 1.7×10^{-4}

但上面的结果表达隐含了均值为4.5, $\sigma = 2.2$ 的高斯变量给出零或更少的概率:

→
$$\int_{-\infty}^0 \frac{1}{\sqrt{2\pi} \times 2.2} \exp\left(-\frac{(x - 4.5)^2}{2 \times 2.2^2}\right) dx = 0.021$$

如果 $s \gg 1$, 没有问题, 此时 n 服从高斯分布。

潜在的问题之二

实际问题会涉及系统误差，例如 $\nu_b = 0.8$ ，则概率变为

$$\begin{aligned} p\text{值} &= P(n \geq 5; b = 0.8, s = 0) = \sum_{n=n_{\text{obs}}}^{\infty} P(n; b = 0.8, s = 0) \\ &= 1 - \sum_{n=0}^{n_{\text{obs}}-1} \frac{b^n}{n!} e^{-b} = 1.4 \times 10^{-3} \end{aligned}$$

虽然本底只增大了0.3， p 值却比 $\nu_b = 0.5$ 时小了一个量级。



建议给出与 ν_b 合理变化相对应的 p 值范围