2016 威海暑期学校结业报告

统计及数据分析的学习

单心钰
中国科学技术大学
导师：彭海平　　**专业：**粒子物理实验　　**Email:** shanxy@mail.ustc.edu.cn

本次暑期学校的报告涵盖了粒子物理的很多方面，从理论到探测器加速器的设计到实验数据的处理，经过各个专题的学习，加深了我对粒子物理研究领域的了解。本次暑期学校使我收获最大的莫过于 Dorigo 教授的报告，报告中关于统计及数据分析的内容使我对统计及数据处理有了初步的了解。因此，我将在报告简单阐述下我学到的关于 Neyman 置信区间估计的知识。

一.　置信区间的 Neyman 方法

对于一个实验测量的随机变量 x，其概率密度函数为$f(x,\theta)$，θ为待估计的参数，区间估计的就是通过研究测量的样本来推断出一个区间，使其包含θ真值的概率为一给定常数。

Neyman 方法是一种经典的置信区间的估计方法。假定实验测量得到参量为 x，对于任一特定μ值，可以得到 x 的接受区域[x1,x2]，满足关系式

$$P(x \in [x1, x2] | \mu) = 1 - \alpha, \qquad (1)$$

α为第一类误差，所有可能的μ值相应的接受区间[x1,x2]的集合构成置信度为$1-\alpha$的置信带，如图 1 所示。假设μ的真值为$\mu 0$，当 x 位于

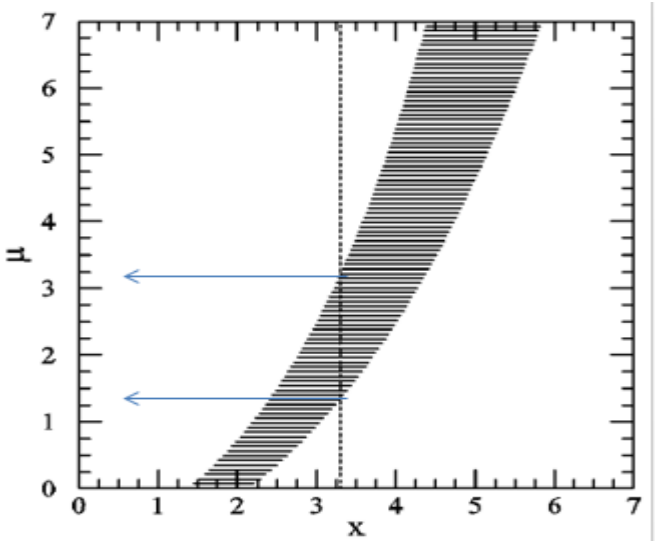

图 1

x1(μ0)和 x2(μ0)之间时，μ0才位于μ1(x)和μ2(x)之间。因此有

$$1-\alpha = P\left(x \in [x1(\mu), x2(\mu)]\right) = P\left(\mu \in [\mu1(x), \mu2(x)]\right) \qquad (2)$$

这样可以看出μ的两个端点μ1(x)和μ2(x)为随机变量，而μ为一个未知参量。如果重复多次实验，随机区域$[\mu1, \mu2]$发生变化，而此区域覆盖真实值μ的概率为$1-\alpha$。显然可以发现满足式 1 的接受区域有无穷多个，取决于你想提取的信息，通常而言,对于一观测量 x，对于上限置信区域$[\mu2, \infty)$定义为$P(\mu < \mu2\ |x) = 1-\alpha$，对于下限置信区域$(-\infty, \mu1]$定义为$P(\mu > \mu1\ |x) = 1-\alpha$，对于中心置信区域$[\mu1, \mu2]$定义为$P(\mu < \mu1\ |x) = P(\mu > \mu2\ |x) = \alpha/2$。

二． 接近边界的参数的置信区间和突变方法(flip-flopping)

当参数只能取一个有限范围的数值时，Neyman 置信区域的构建会遇到一些困难，例如我们做一个无偏的高斯分布测量，由于物理需要限定μ大于 0，而我们实验测得值为 x，分布为

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \quad (3)$$

x 可能取为负值，此时我们倾向于报道上限，以高斯分布为例 ，对于x<D(3σ或5σ)，报道上限，对于$x \geq D$报道中心区间，如图 2 所示，这种方式称为突变方式(flip-flopping)，然而这种方式存在两个问题，一方面，对于待估计参量μ的某些值涵盖概率不足，如对于$\mu = 4$，区间涵盖概率小于 90%，另一方面，当观测值小于某些值，例如$x = -1.8$时，置信度 90%的置信区间为空集。

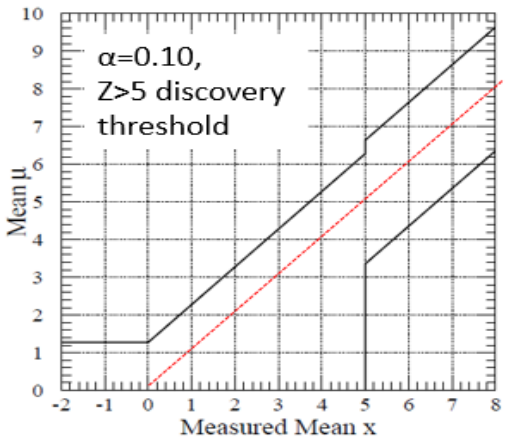

图 2

由此可见，在讨论带估计参数接近物理边界的估计区间时，经典的 Neyman 方法遇到了一些困难，需要一些新的方法例如 F-C 方法(似然比方法)来处理这类问题。

通过这期的暑期学校，使我了解了关于粒子物理的方方面面，相信从中学习到了知识会对我今后的学习研究有所帮助。