# **置信区间如何理解和设置**

95%置信区间，意味着如果你用同样的步骤，去选样本，计算置信区间，那么100次这样的独立过程，有95%的概率你计算出来的区间会包含真实参数值，即大概会有95个置信区间会包含真值。而对于某一次计算得到的某一个置信区间，其包含真值的概率，我们无法讨论。[参源](https://blog.csdn.net/pannn0504/article/details/82455934)

## **1.点估计与区间估计**

首先我们看看点估计的含义：  
是用样本统计量来估计总体参数，因为样本统计量为数轴上某一点值，估计的结果也以一个点的数值表示，所以称为点估计。点估计虽然给出了未知参数的估计值，但是未给出估计值的可靠程度，即估计值偏离未知参数真实值的程度。  
接下来看下区间估计：  
给定置信水平，根据估计值确定真实值可能出现的区间范围，该区间通常以估计值为中心，该区间则为置信区间。

## **2.中心极限定理与大数定理**

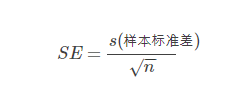
中心极限定理：  
在适当的条件下，大量相互独立随机变量的均值经适当标准化后依分布收敛于[正态分布](https://so.csdn.net/so/search?q=%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/u010058695/article/details/_blank)。例如我们要计算全中国人的平均身高。如果每次取10000个身高作为样本，对应有一个样本均值。如果再从总体中重复抽取n多次10000个样本，就对应有n个样本均值。随着n增大，把所有样本均值画出来，得到的就是一个接近正太分布的曲线。  
大数定理：  
取样数趋近无穷时，样品平均值按概率收敛于期望值。抛硬币的次数越多，越接近正反各一半。

## **3.置信区间与置信水平**

一般我们用中括号[a,b]表示样本估计总体平均值误差范围的区间。a、b的具体数值取决于你对于”该区间包含总体均值”这一结果的可信程度，因此[a,b]被称为置信区间。  
一般来说，选定某一个置信区间，我们的目的是为了让”ab之间包含总体平均值”的结果有一特定的概率，这个概率就是所谓的置信水平。  
例如我们最常用的95%置信水平，就是说做100次抽样，有95次的置信区间包含了总体均值。

## **4.标准差(standard deviation)与标准误差(standard error)**

标准差是描述观察值(个体值)之间的变异程度（例如一个人打十次靶子的成绩，这时有一个平均数8，有一个反映他成绩稳定与否的标准差）；  
标准误是描述样本均数的抽样误差（例如十次抽样，每次他成绩平均数（7,8,6,9,5,6,7,7,8,9）的标准差，也就是抽样分布的标准差）；  
样本的标准误差为:



## **5.如何理解95%的置信区间**

以上面的统计身高为例，假设全国人民的身高服从正态分布：

IMG_257

不断进行采样，假设样本的大小为n，则样本的均值为：

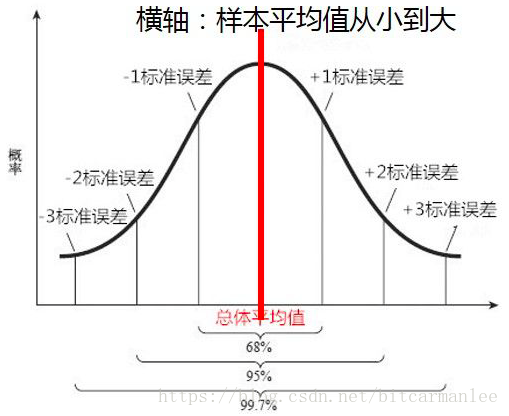
IMG_258

由大数定理与中心极限定理：

IMG_259

注意σ1的计算方法为第4部分提到的标准误差！

为什么常用95%的置信水平：



对照上图，用一句简单的话概括就是：  
有95%的样本均值会落在2个(比较精确的值是1.96)标准误差范围内。

用数学公式描述就是：

IMG_261

## **6.计算置信区间的套路**

从上面的例子来看，计算置信区间的套路如下：  
1.首先明确要求解的问题。比如我们的例子，就是想通过样本来估计全国人民身高的平均值。  
2.求抽样样本的平均值与标准误差(standard error)。注意标准误差与标准差(standard deviation)不一样。  
3.确定需要的置信水平。比如常用的95%的置信水平，这样可以保证样本的均值会落在总体平均值2个标准差得范围内。  
4.查z表，求z值。  
5.计算置信区间  
a = 样本均值 - z\*标准误差  
b = 样本均值 + z\*标准误差

用公式表示置信区间：

IMG_262  
其中，x¯表示样本的均值，z值表示有多少标准差，s为样本的方差。