**四、统计学**

**核心知识点：**

**1、中心极限定理**

**(样本量n足够大时，样本均值的分布近似服从正态分布)**

独立同分布中心极限(Lindburg-Levy)定理：设随机变量独立同分布，且，，记，则有,即n趋于无穷时，服从标准正态分布。

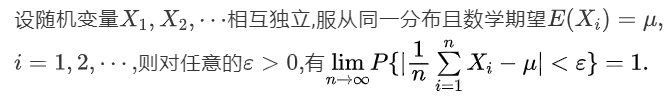
应用至总体与样本中：

设从均值为，方差为的任意一个总体中抽取样本量为n的样本，当n充分大的时候，样本均值的抽样分布近似服从均值为，方差为的正态分布，也就是说不管总体的分布是什么，样本量足够大时，样本均值的分布总是接近正态分布

**2、大数定律**

**(样本量n足够大时，样本均值逐渐接近总体均值)**

辛钦大数定律



随着样本容量n不断增加，样本平均数将越来越接近于总体平均数(期望)

**3、假设检验、显著性水平、置信区间**

**假设检验**(hypothesis testing)是数理统计学中根据一定假设条件由样本推断总体的一种方法。其先对总体要得到的结论提出假设，然后利用样本信息去检验这个假设是否成立。

**显著性检验**是假设检验中最常用的一种方法，也是一种最基本的统计推断形式，其基本原理是先对总体的特征做出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是接受做出推断。常用的假设检验方法有Z检验、t检验、卡方检验、F检验等。  
**假设检验的基本思想:**利用“小概率事件”原理，即小概率事件在一次试验中基本上不会发生。为了检验一个假设H0是否正确，首先假定该假设H0正确，然后根据样本对假设H0做出接受或拒绝的决策。如果样本观察值发生了“小概率事件”，就应拒绝假设H0，否则应接受假设H0。

**显著性水平：**犯第一类错误(拒真错误)的概率，也即原假设为真时被拒绝的概率，常用表示

**置信水平:** 相信变量在设定的置信区间的概率，与显著性水平加和为1，用1-表示。  
**置信区间**：在区间估计中，由样本统计量所构造的总体参数的估计区间称为置信区间。  
**假设检验的基本步骤：**   
 （1）提出原假设H0，备择假设H1。   
 H0：样本与总体或样本与样本间的差异是由抽样误差引起的；   
 H1：样本与总体或样本与样本间存在本质差异；   
 预先设定的检验水平α（当检验假设为真但被错误地拒绝的概率），通常取α=0.05。   
 （2）选定统计方法，由样本观察值按相应的公式计算出统计量的大小，如X2值、t值等。根据数据的类型和特点，可分别选用Z检验，T检验，秩和检验和卡方检验等。   
 （3）根据统计量的大小及其分布确定检验假设成立的可能性P的大小并判断结果。若P>α，结论为按α所取水平不显著，不拒绝H0，即认为差别很可能是由于抽样误差造成的，在统计上不成立；如果P≤α，结论为按所取α水平显著，拒绝H0，接受H1，则认为此差别不大可能仅由抽样误差所致，很可能是实验因素不同造成的，故在统计上成立。

**4、T检验 F检验 卡方检验**

**T检验：T检验（T-test）是为了比较数据样本之间是否具有显著性的差异，一般用于定量数据的检测（定类数据采用卡方检验），主要用于样本含量较小（n < 30），总体标准差σ未知的正态分布的样本。**T检验的前提条件是假设样本服从或者近似服从正态分布。

只能比较两组数据，衍生出几种t检验的应用场景。

**1）单样本均值检验**

用于检验总体方差未知、正态数据或近似正态的单样本的均值是否与已知的总体均值相等。

例1：从某厂生产的零件中随机抽取若干件，检验其某种规格的均值是否与要求的规格相等（双侧检验）--- 总体均值和样本均值不相等就用双侧检验，因为不相等包括总体均值大于样本均值和总体均值小于样本均值两种情况

例2：在某偏远地区随机抽取若干健康男子，检验其脉搏均数是否高于全体健康男子平均水平（单侧检验） --- 研究的事总体均值是不是大于(小于)样本均值就用单侧检验

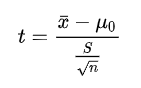
例2进行t检验计算步骤

a. 提出假设：H0-该山区成年男子脉搏与一般成年男子脉搏数没差别

H1-该山区成年男子脉搏比一般成年男子脉搏数高

b. 做t检验，计算统计量

计算单样本t统计量



c.查表得到p值

查看t值表，若对应的单侧p值小于0.05，在显著性水平为0.05的前提下，拒绝原假设，承认备择假设；否则承认原假设

**2）两独立样本均值检验**

用于检验两对独立的正态数据或近似正态的样本的均值是否相等，两组样本的数量可以不等，这里可根据总体方差是否相等分类讨论。

例1：检验两工厂生产同种零件的规格是否相等（双侧检验）

例2：为研究某种治疗儿童贫血新药的疗效，以常规药作为对照，治疗一段时间后，检验施以新药的儿童血红蛋白的增加量是否比常规药的大（单侧检验）

两个独立样本的T检验，通常需要先进行F检验（方差齐次检验），检验两个独立样本的方差是否相同，若两总体方差相等，则直接用t检验，若不等，可采用t’检验或变量变换或秩和检验等方法。也就是说进行两独立样本的T检验时，需首先验证两样本的方差是否相同，进行t检验步骤与1）类似，t统计量换为



**3）配对样本均值检验**

用于检验一对配对样本的均值的差是否等于某一个值，要求两组样本数据必须相等，总体方差相等，并近似正态分布。这种情况常常出现在生物医学研究中，常见的情形有：

例1：配对的受试对象分别接受不同的处理（如将小白鼠配对为两组，分别接受不同的处理，检验处理结果的差异）

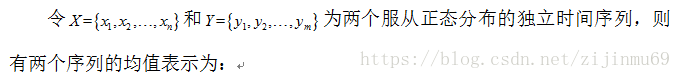
例2：同一受试对象的两个部分接受不同的处理（如对于一批血清样本，将其分为两个部分，利用不同的方法接受某种化合物的检验，检验结果的差异）

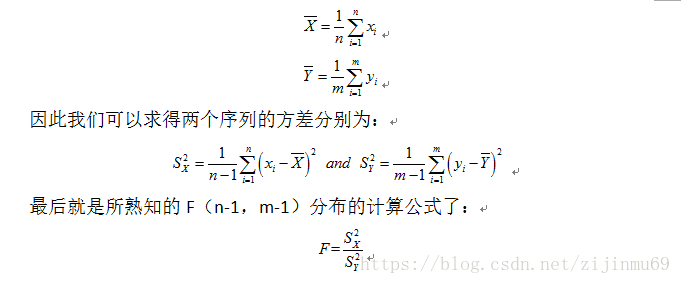
例3：同一受试对象的自身前后对照（如检验癌症患者术前、术后的某种指标的差异）

什么叫配对？就是消除了每一对自身的差异，比如对照组是一比一挑出来的，比如年龄，性别，学历，都和另一个组的分布是一模一样的，这种情况就是用配对检验，如果是随机挑出来的，那么就用独立样本检验。

**F检验：**F检验（F-test），最常用的别名叫做联合假设检验（英语：joint hypotheses test），此外也称方差比率检验、方差齐性检验。它是一种在零假设（null hypothesis, H0）之下，统计值服从F-分布的检验。**要判断两总体方差是否相等，就可以用F检验。**

**计算过程：**

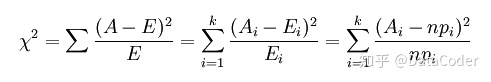




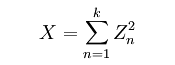
计算得到的F值，再与对应F分布表查询，若大于表中的值，则接受原假设（两样本的方差相同），否则拒绝原假设。

**卡方检验：**卡方检验（chi-square test），也就是χ2检验，**用来验证两个总体间某个比率之间是否存在显著性差异**。卡方检验属于**非参数假设检验**，适用于布尔型或二项分布数据（即**只适用于分类变量**），基于两个概率间的比较，早期用于生产企业的产品合格率等。

卡方检验是以χ2分布为基础的一种常用假设检验方法，它的原假设H0是：观察频数与期望频数没有差别。该检验的基本思想是：首先假设H0成立，基于此前提计算出χ2值，它表示观察值与理论值之间的偏离程度。根据χ2分布及自由度可以确定在H0假设成立的情况下获得当前统计量及更极端情况的概率P。如果当前统计量大于P值，说明观察值与理论值偏离程度太大，应当拒绝无效假设，表示比较资料之间有显著差异；否则就不能拒绝无效假设，尚不能认为样本所代表的实际情况和理论假设有差别。（摘自智库百科）。  
　卡方检验的基本公式为：



A 为观察值，E为理论值，k为观察值的个数，最后一个式子实际上就是具体计算的方法了 n 为总的频数，p为理论频率，那么n\*p自然就是理论频数（理论值）。  
卡方检验是以χ2分布为基础的一种常用假设检验方法，χ2分布，就叫做卡方分布。若k 个随机变量Z1、……、Zk 相互独立，且数学期望为0、方差为 1(即服从标准正态分布)，则随机变量X：



记作：



在做卡方检验时，通常针对的数据主要为定类数据，例如样本中是男生的人数，研究的问题通常是两个类别之间是否有显著关联。

**5、参数估计**

**定义：**参数估计（parameter estimation），统计推断的一种。根据从总体中抽取的随机样本来估计总体分布中未知参数的过程。从估计形式看，区分为点估计与区间估计；

常见点估计方法：**矩估计、最小二乘估计、极大似然估计、贝叶斯估计**

区间估计：**利用已知的抽样分布、利用区间估计与假设检验的联系、利用大样本理论**

**1) 点估计**

a、**矩估计**

[矩估计](https://so.csdn.net/so/search?q=%E7%9F%A9%E4%BC%B0%E8%AE%A1&spm=1001.2101.3001.7020)法的理论依据是大数定律。矩估计是基于一种简单的“替换”思想，即用**样本矩估计总体矩**

优点：简单易行, 并不需要事先知道总体是什么分布。（根据均值方差来计算未知参数）

缺点：当总体类型已知时，没有充分利用分布提供的信息（有一定随意性）

b、**最小二乘估计**

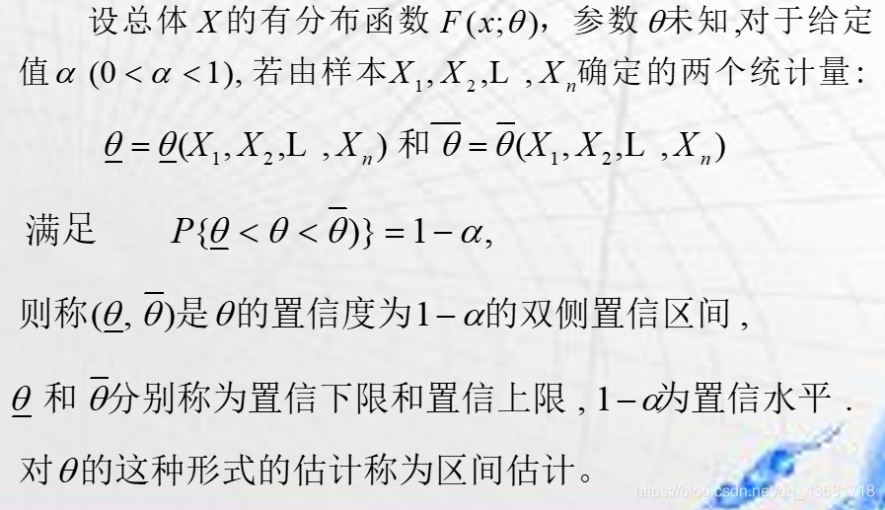
对于最小二乘估计来说，最合理的参数估计量应该使得模型能最好地拟合样本数据，也就是估计值与观测值之差的平方和最小。   
目标最小化估计值与观测值之差的平方和。Q表示误差平方和，Yi表示估计值，Ŷ i表示观测值，即

c、**极大似然估计**

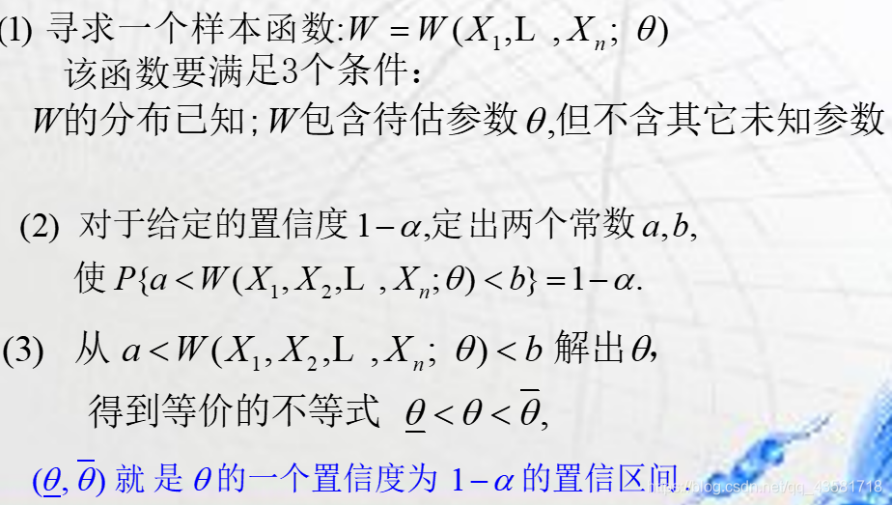
对于最大似然估计来说，最合理的参数估计量应该使得从模型中抽取该n组样本的观测值的概率最大，也就是概率分布函数或者似然函数最大。

**2）区间估计**

**定义：**区间估计 = 点估计 ± 边际误差   
根据样本求出未知参数的估计区间，并使这个区间包含未知参数的可靠程度达到预定要求（这个预定要求就是个置信度，用上α位分点来体现这个置信度）。



**步骤：**



**参数估计和假设检验的联系：**参数估计和假设检验都是样本去估计总体，都是建立在概率基础上的统计，可以相互转换

**参数估计和假设检验的区别是：**   
a.参数估计是用样本统计量估计总体参数的方法；假设检验是先对总体参数提出一个假设，然后利用样本信息去检验这个假设是否成立   
b.参数估计是以置信区间(大概率)估计总体参数；假设检验是利用小概率事件是否发生来判断假设是否成立

c.区间估计求得的是求以样本估计值为中心的双侧置信区间，假设检验既有双侧检验，也有单侧检验；

**6、 辛普森悖论**

**定义：**在对样本[集合](https://so.csdn.net/so/search?q=%E9%9B%86%E5%90%88&spm=1001.2101.3001.7020)进行分组研究时，在分组比较中都占优势的一方，在总评中有时反而是失势的一方，这种有悖常理的现象，就成为 “辛普森悖论”。

**实例：**

xx学校两个学院男生录取率均高于女生，但该学校女生录取率则高于男生，这个例子便是一典型的辛普森悖论现象



原因：为什么会出现辛普森悖论现象，从上面这个例子可以看到，男生虽然每个学院的录取率都高于女生，但男生中有100个人(5/6的比例)报了录取率较低的法学院，出现了严重的数据数量不平衡，因此导致男生整体的录取率偏低。

另外，录取率这个变量应该跟专业有较大关系，而不同性别的学生报考不同专业的比例不同，性别这个变量是因为跟报考专业相关，才呈现出跟录取率相关的现象，从本质上讲，这个变量设置并不合理。

**如何避免辛普森悖论：**

辛普森悖论会导致统计上的一些错误结论，要避免掉入辛普森悖论的陷阱，要从一下两方面着手：

1. 进行合理的变量设计，进行合理的变量关系验证

如上述实例，出现辛普森悖论的很大一个原因是由于变量设计不合理，在研究之初就漏掉了“专业”这个重要潜伏变量(，而直接将性别与录取率建立了联系。因此，在进行变量设计时，一是要尽量多查阅文献以了解自变量，因变量及其他重要变量之间的关系，二是积累相关的经验，能够敏锐地察觉到某个变量的重要性。

(这个重要潜伏变量指的是与实验分析的因果都有关系的变量，这一变量的缺失可能会掩盖或颠倒研究变量的原有关系。

比如在录取率这个例子中，专业既与录取率有关，不同专业难度不同，录取率自然有差异，又与性别有关，女生和男生趋向报考的专业不同。因此专业就是一个重要潜伏变量。)

2. 进行科学合理的抽样

在进行抽样调查时应选择合适的抽样方法或多种抽样方法并存，保证抽取的样本处于远离极端范围的中间区域，在进行非比例抽样时，应慎重考虑权重的问题。

**拓展知识点(低频、难度较大，选择性掌握)**

**1、幸存者偏差**

**定义**：幸存者偏差是一种选择偏倚，就是忽略了筛选条件，把经过筛选的结果当成随机结果。

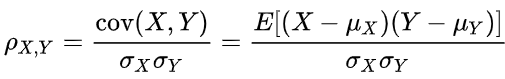
**实例**：统计结果显示，机翼中弹的飞机生还概率远高于驾驶舱和机尾中弹的飞机。但驾驶舱和机尾中弹的飞机很少回来，基本不在样本之中

**如何避免：**谨慎进行样本选择，当样本出现严重偏斜时，考虑出现原因及合理性**。**

**2、相关性分析（皮尔森person相关系数、spersman相关系数、肯德尔kendall秩相关系数）**

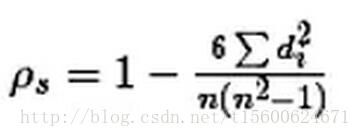
**定义：**相关性分析是指对两个或多个具备相关性的变量元素进行分析，从而衡量两个变量因素的相关密切程度。常用的相关系数有皮尔森person相关系数、spersman相关系数、肯德尔kendall秩相关系数

**皮尔森person相关系数：**两个变量之间的皮尔逊相关系数定义为两个变量之间的[协方差](https://baike.baidu.com/item/%E5%8D%8F%E6%96%B9%E5%B7%AE)和[标准差](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E5%B7%AE)的商。



person相关系数取值区间[-1,1]，若person相关系数接近0，无相关性；person相关系数接近-1或+1，有较强相关性

**spersman相关系数：**采用的是取值等级进行相关性计算，



**计算过程：**首先对两个变量（X, Y）的数据进行排序，然后记下排序以后的位置（X’, Y’），（X’, Y’）的值就称为秩次，秩次的差值就是上面公式中的，n就是变量中数据的个数，最后带入公式就可求解结果。举例说明，假设我们实验的数据如下：



利用计算出的秩次差的平方代入上述公式进行计算，便可得到spersman相关系数

**特点：**1. 即便在变量值没有变化的情况下，也不会出现像皮尔森系数那样分母为0而无法计算的情况。另外，即使出现异常值，由于异常值的秩次通常不会有明显的变化（比如过大或者过小，那要么排第一，要么排最后），相对于皮尔森相关系数，斯皮尔曼相关系数对于数据错误和极端值的反应不敏感。

2. 斯皮尔曼相关性系数没有数据条件要求，适用的范围较广，在生物实验数据分析中，尤其在分析多组交叉数据中说明不同组数据之间的相关性时，使用的频率很高。

**Kendall相关系数：**肯德尔(Kendall)系数又称一致性系数，是表示多列等级变量相关程度的一种方法，也是一种秩相关系数，计算对象是分类变量，适用于两个分类变量均为有序分类的情况。

**适用情况：**评委对选手的评分（优、中、差等），想看两个（或者多个）评委对几位选手的评价标准是否一致；或者医院的尿糖化验报告，想检验各个医院对尿糖的化验结果是否一致，这时候就可以使用肯德尔相关性系数进行衡量。

因计算公式过于复杂，不予展示

**面试问题**

1、简述中心极限定理、大数定律（字节、拼多多）

2、假设检验的定义、第一类错误、第二类错误（字节跳动、快手）

3、T检验、Z检验、F检验、卡方检验的定义、适用情况(快手、腾讯、小米)

4、辛普森悖论的定义及实例（腾讯、字节）

5、幸存者偏差现象（腾讯）

6、相关系数的取值范围及相关系数为0的意义(阿里)

7、参数估计有哪些方法

8、点估计和区间估计的区别和联系（字节）

9、极大似然估计是什么（字节）

10、参数估计和假设检验的联系和区别（阿里）