**统计学讲义**

**第一节：统计学基本概念**

**1. 统计学：**收集、处理、分析、解释数据，并从数据中得出结论的科学。方法分为描述性统计和推断统计。

**描述性统计：**研究数据搜集，处理，汇总，制图描述，概括分析等统计方法。

**推断统计：**研究如何利用样本数据来推断总体特征的统计方法。

**2. 统计数据的类型划分**

**（1）按数据形态分：**

**分类数据：**数据表现为类别，男女，行业类别等。此类数据仅有分类上的意义。

**定序数据：**有序类别的数据，例如产品分为一等品、二等品。此类数据不仅可以分类，还有高低之分，可以比较。

**定距数据：**既可以是离散的（0，1，2，3…），也可以是连续型的。例如温度、IQ，此类数据不仅可以比较高低，还可以做加减运算。但是其在零点处无意义，乘除运算无意义。

**定比数据：**在定距数据基础上，这类数据有绝对零点，即零点是有意义的，例如收入为0表示没有收入，做加减乘除运算均有意义。

**（2）按生成方式分：**

**观测数据：**经过调查或观测收集到的数据，特点是收集过程没有人为干预数据生成过程。

**实验数据：**实验室中通过人为的实验控制收集得到，是有干预的。

**（3）按数据结构分：**

**截面数据：**反应描述对象在某一时刻的状态，例如2010年我国各省经济情况

**时间序列：**在不同时间手机的数据，数据按时间顺序排列，描述现象随时间变化的情况，例如2010-2018年我国国内生产总值数据

注意：区分数据类型是数据分析的前提，不同类型的数据对应不同的建模方法，检验方法。

**3. 统计学基本术语**

**（1）总体：**包含所研究的全部个体的集合。理解总体概念注意以下两点：

**i.** 研究一个对象，首先要定义好它对应的总体。总体定义需要包括时间、地理范围、人群范围等等范围。

**ii.** 总体可分为有限总体和无限总体。区分有限和无限是为了研究每次抽取样本，是否是独立的。（影响每个个体被抽中的概率）

**（2）样本：**是从总体中抽取一部分元素的集合，样本元素的个数称为样本量。抽样的目的是根据样本提供的信息推断总体的特征。

**（3）参数：**用来描述总体特征的概括性数字度量，例如总体平均数、总体标准差等。总体数据是未知的，参数也是未知的，所以为了估计参数，才对总体进行抽样，用样本统计量来估计总体参数。

**（4）统计量：**用来描述样本特征的概括性数字度量，一个样本会生成一个样本统计量（例如样本均值），因此统计量是样本的函数。

**第二节：数据搜集**

数据搜集需要回答的问题：

（1）我从哪里获得数据？（数据来源）

（2）我怎样取得数据？（搜集方法）

（3）数据质量怎么样？（误差的估计和控制）

**1. 数据来源**

（1）我们一般通过调查或实验的方法获取数据，这是按生成方式划分的。也可以按一手数据和二手数据分，即我们是亲自收集的还是利用别人已经搜集好的数据。一手数据的优点是针对性强，各个环节把握程度高，数据质量较好，缺点是成本较高。二手数据优点是成本低，速度快。缺点是针对性相关性不够，口径不一致，误差较大，时效性不够等。但一般研究还是从搜索二手数据开始。

**（2）使用二手数据要注意的点：**

i. 资料是谁收集的？要选择具有权威性的机构公布的数据

ii. 搜集的目的是什么？是否会有潜在利益问题导致质量不可靠

iii. 数据搜集的方法是什么？ 范围口径、抽样方法等等

iv. 时效性如何？ 要与研究问题进行匹配

**2. 调查方法相关概念**

**（1）概率抽样：**概率抽样也称随机抽样，即按照一定的概率以随机原则抽取样本。每个单位被抽中的概率是已知的，样本估计量不仅与样本观测值有关，也与进入样本的概率有关（概率作为权重）。概率抽样的优点是可以在抽样之前就计算出预计的估计量误差，判断总体参数估计的可靠度。

i. 简单随机抽样：基本的抽样方法，每个个体进入样本概率相同，需要抽样框，执行难度大。

ii. 分层抽样：按某个特征或规则划分不同层（按性别、行业等），从不同层中独立随机抽取样本，代表性更强，估计精度更高。

iii. 整群抽样：将总体中若干单位合并为群，抽中群的所有个体都参与调查（先抽学校，抽中的学校所有学生都进入样本）。优点是方便实施，缺点是估计精度差，因为群内个体同质性较高。

iv. 系统抽样：将所有单位按一定顺序排列，随机抽取一个单位作为初始单位，然后按事先制定的规则，依次确定其他单位（例如每隔k个单位抽取下一个样本）。优点是操作简便，

缺点是标准误难估计。

v. 多阶段抽样：先抽取群，然后群中再抽取子群，最后抽取个体。该方法结合了整群抽样和简单抽样，多用于大型抽样调查中。

**（2）非概率抽样：**样本统计量不能用于推断总体

i. 方便抽样：调查员依据方便原则，自行确定作为样本的单位。例如街头拦访。该方法优点是成本低，方便操作。缺点是代表性不足，误差难以估计。此方法多用于研究初期的探索性研究，形成初步认识和建立假设。

ii. 判断抽烟：研究人员根据经验，有目的的选择一些单位作为样本，如重点抽样、典型抽样、代表抽样等。例如，调查国企中典型的大企业，或根据研究问题，将研究对象分几类，研究几个分类中的典型样本。多用于定性研究，探索性研究中。

iii. 滚雪球抽样：首先选择一个调查单位，然后根据该单位提供的线索，继续调查下一个单位。该方法多用于稀少群体的调查，比如研究吸毒者、卖淫者群体。

iv. 配额抽样：类似于概率样本中的分层抽样，广泛用于市场调查。根据几个分类变量，形成几类群体，然后配额抽样。

注意：选择概率抽样或非概率抽样没有绝对的优劣，必须按照研究问题的特点、调查对象的特征、费用和时间约束综合考虑。

**（3）调查数据取得方法：**自填式、面访式、电话式（略）

**3. 数据的误差**

指通过调查收集到的数据与研究对象真实情况之间的差异。分为两类：抽样误差和非抽样误差。

**（1）抽样误差：**由抽样的随机性引起的样本结果与总体真值之间的差异，是随机性的误差。这个差异来源于抽样的随机性，减少抽样误差的最主要方法是扩大样本量。其次，如果所调查的总体中，个体之间异质性特别大，抽样误差也会大（如果研究对象是铜原子，只需抽一个铜原子研究即可）。

**（2）非抽样误差：**除抽样误差以外的原因引起的误差，是系统性的误差。

i. **抽样框误差：**抽样框未能完整涵盖研究的总体，例如研究总体是某学校全体学生，但拿到的学生名单是去年的，因此今年新生就没在抽样框里，产生误差。

ii. **样本无应答误差：**一部分调查样本无法联系到，或联系到拒绝参与研究。如果这些无应答的样本，其某个特征与研究问题相关，则会产生系统性偏误。

iii. **应答误差：**被访者给出的答案与真实情况不符，又分为理解误差（主观题每个人标准不一样）、记忆误差（不容易回忆）、有意识误差（敏感问题）等。

iv. **调查员误差：**调查员粗心、记录错误、诱导答卷等

v. **测量误差：**指标、测量工具、测量方法选择不当或不精确，导致的误差。

**第三节：数据概括性度量**

随机变量都会有一个分布，概括分布的特征有三个角度：集中趋势、离中趋势、分布形状。低层次的数据集中趋势适用于高层次的测量数据，反之则不成立。

**1. 集中趋势的度量**

**（1）众数：**数据中出现次数最多的变量值，主要用于分类数据、定序数据的集中趋势。众数的优点是不受数据中极端值的影响，缺点是如果数据没有明显的集中趋势，则众数可能不存在或存在多个众数（无峰或多峰分布情况）。

**（2）分位数（包括中位数）：**找到一组数据中某个位置上的数据，中位数、四分位数、十分位数、百分位数等。中位数多用于测量数值型数据的集中趋势，优点是不受极端值影响，常用于研究收入分配。四分位数，排在25%和75%位置上的值

**注意：（1）（2）两种方法，不受数据分布形态的限制，比较稳健。**

**（3）平均数：**适用于数值型数据，不适合分类和顺序数据。平均数有较深的哲学基础，它反映的是一组数据中观测误差相互抵消后的结果，是一组数据的重心所在。当数据服从正态分布，或接近正态分布时，用平均数概括分布效果较好。（左偏分布，右偏分布的形态即三种方法的关系）

**2. 离散程度度量**

**（1）异众比率：**指非众数组的频数占总频数的比例。用于衡量众数对一组数据的代表程度，比率越大，说明众数的代表性越差，该指标适合测度分类数据的离散程度。

**（2）四分位差（内距，四分间距）：**上四分位数与下四分位数之差。四分位差反应了中间50%的数据的离散程度，值越小数据越集中，也间接测度中位数的代表程度。四分位差的优势在在于不受极值影响，主要应用于测量定序数据的离散程度。

**（3）方差与标准差：**用于数值型数据，是最常用的衡量离散程度的方法，标准差消除了方差的量纲。有了方差与标准差以及均值，就可以将数据标准化。

**（4）变异系数：**方差和标准差受到原变量自身水平高低的影响，也就是与平均数的绝对大小相关，变量水平高的方差自然大。为了消除水平高低对离散程度影响，产生了变异系数。

**变异系数是标准差与平均值之比。**

**（4）离群值：**根据经验法则，在正态分布下，68%的数据集中于+-1个标准差内，95%的数据集中于+-2个标准差内，99%的数据集中于+-3个标准差内，因此大于3个标准差的数据可被称为离群点。根据切比雪夫不等式，对于任何分布形态的数据，75%的数据落在2个标准差范围内，89%的数据落在3个标准差以内，94%的数据落在4个标准差以内。若一个数据中离群点非常多，提示数据质量有问题。

**3. 分布形态度量**

均值表示集中趋势，称为“一阶矩”；方差表示离中趋势，称为“二阶距”。但仅有这两个参数，也不能完全刻画一个分布的特点。因此，需要更高阶的测度，以刻画变量的分布。

**（1）偏度(skewness)：**皮尔逊提出的，用于刻画数据分布是否对称。偏度也称“三阶距”，如果数据分布是对称的，则偏度为0；如果偏度系数绝对值在0.5内，则中等偏态，大于1，则高度偏态。

**（2）峰度(kurtosis)：**峰度是相对于标准正态分布而言的，如果数据服从标准正态分布，则超额峰度值（峰度-3）为0。若超额峰度大于0，则为尖峰分布，即数据分布更集中。若超额峰度小于0，则为扁平分布，数据分布分散。

**第四节：统计量与抽样分布**

**1. 统计量的概念**

（1）原始数据是杂乱无章的，我们要从原始数据中整理出一些关键的代表性数字（数字特征），来“总结”出原始数据的特点，这个**总结出来的数字特征的函数，称为“统计量”（估计量），统计量是样本的函数。某样本下，具体计算出来的估计量的值称为“估计值”。**推断统计学的工作就是，通过从总体中抽取样本构造适当统计量，由样本性质推断总体性质。

（2）假设某市的家庭收入水平作为研究对象（也就是总体），我们通过抽样，得到了一个样本家庭的收入数据，可以用一个列向量X = (x1,x2,x3,…,xn) 表示。则T(x1,x2,x3,…,xn) 即为一个样本的统计量，任意给一组样本数据，都可以通过这个函数T转换为一个数字，这个数字可以是均值，可以是标准差（它们都有特定的公式），所以均值，标准差就是样本统计量。

（3）计算出来统计量后，我们就可以按这些统计量所各自遵循的理论分布，进行统计检验。例如，均值统计量，其理论分布服从t分布或正态分布（大样本下），因此要根据这些分布进行检验。

**2. 抽样分布**

**抽样分布、参数估计、假设检验是推断统计的三大核心内容**。

总体服从正态分布的条件下，卡方分布、t分布、F分布，通常称为统计三大分布

i. 卡方分布：n个相互独立的，服从标准正态分布的随机变量的平方和，服从自由度为n的卡方分布。自由度是独立变量的个数，或二次型的秩。卡方分布的极限分布，即随着样本数量n趋于无穷大时，卡方分布将趋于正态分布。

ii. t分布，也是基于正态分布的某个变体（公式略）。其极限分布也是正态分布，主要用于小样本统计推断中。

iii. F分布，两个相互独立的、服从两个不同自由度的卡方分布的、随机变量之比，需要两个自由度。主要用于方差分析、回归方程的显著性检验中，公式略。当回归方程中只有一个解释变量时，t检验与F检验等价。

**注意：以上理论的假设前提是总体分布已知，且为正态分布！**

**3. 样本均值的抽样分布和中心极限定理**

（1）抽样分布也称统计量分布。以均值这一统计量为例：按照相同的样本容量，相同的抽样方式，反复地抽取样本，每次可以计算一个平均数，所有可能样本的平均数所形成的分布，就是样本平均数的抽样分布。

（1）当总体服从正态分布，则其样本均值的抽样分布仍服从正态分布，抽样分布的期望即为总体均值，抽样方差为总体方差的1/n。因此，我们说这个均值的推断是“无偏的”，并且当n足够大，抽样分布的方差会越来越小，最终收敛至总体均值。

（2）若总体分布不服从正态分布，则根据中心极限定理，当样本数n足够大时（n>=30），不管总体服从什么分布，只要其方差是有限的，则其样本均值的抽样分布总是近似正态分布。

（3）大数定律和中心极限定理的区别：大数定律是针对某个样本而言的。当某一个样本中样本量n足够大，那么其中的分布规律会逐步浮现。例如，掷硬币，随着次数增多，正反面出现的频率应该逐步接近50%—理论上的真实概率。中心极限定理是针对抽样分布而言的，即单个样本的样本量增大，对总体推断的影响。某个随机变量可以服从大数定律，但不服从中心极限定理，或服从后者不服从前者，或两个都不服从。

**第五节：参数估计**

**注意：**

**参数估计是用样本统计量对总体参数的估计，如果拿到的数据是总体，就不需要“估计”了。**

**1. 参数估计的原理**

**（1）点估计：**用样本统计量的某个取值，直接作为总体参数的估计值。比如，估计一班学生的平均分，直接根据抽出的随机样本计算平均分，作为总体的平均分。

**（2）区间估计：**在点估计的基础上，给出总体参数估计的一个区间范围，区间由样本统计量加减估计误差得到。总体标准差是样本均值的抽样标准误的“根号n”倍。即，加大样本量n可以缩小估计误差。

**（3）置信区间：**由样本统计量所构造的总体参数的估计区间，样本量给定时，置信水平越高（alpha越小），置信区间也就越宽，也就越容易包含总体真值，也就越不容易拒绝原假设。置信水平固定，则样本量越大，置信区间越窄，估计精度越高。

例如，在95%置信水平（即容许5%的一类错误概率）的含义：抽取100个样本，由100个样本构造的总体参数的100个置信区间中，有95%的区间包含了总体参数的真实值。

例子：在95%的置信水平下，得到某班级学生成绩的置信区间为60-80分。以下说法对不对？

（1）60-80分这个区间以95%的概率包含全班学生的平均考试成绩真值

（2）全班平均考试成绩以95%的概率落在60-80分区间

**2. 评价估计量的标准**

**（1）无偏性：**指估计量抽样分布的数学期望等于被估计的总体参数

**（2）有效性：**对同一总体参数的两个无偏估计量，有更小标准误的估计量更有效。

**（3）一致性：**随着样本量增大，估计量的值越来越接近被估计总体的参数真实值。大样本给出的估计量比小样本给出的估计量更接近总体参数。**无偏性+有效性 = 一致性**

**3. 总体参数的区间估计**

总体参数比较重要的有均值、方差、比例等。

**（1）总体均值的区间估计**

对总体均值的估计要考虑总体是否为正态分布、总体方差是否已知，样本是大样本还是小样本。

i. 总体为正态分布，方差已知，或非正态总体、大样本：均值抽样分布服从标准正态分布，z检验。

ii. 总体为正态分布，方差未知，小样本：均值抽样分布服从t分布，样本方差代替总体方差，t检验。

**（2）总体比例的区间估计**

i. 大样本条件：比例P的抽样分布服从标准整体分布，用z检验

ii. 小样本条件：费希尔精确检验

**（3）总体方差的区间估计**

样本方差的抽样分布服从自由度为n-1的卡方分布（方差的公式），卡方检验

**4. 两个总体参数的区间估计**

（1）两个总体均值之差的区间估计

i. 独立样本：两个样本从两个总体中独立抽取，如果两个总体都服从正态分布（无论样本大小），或总体不服从正态分布但样本量足够大，则样本均值之差标准化后服从标准正态分布，两个总体方差未知可用样本方差代替。

ii. 小样本下，若方差未知，则方差经标准化处理后的抽样分布服从t分布

iii. 配对样本：两次抽取是不独立的，比如对同一组人测试两次产生的两个样本。T检验

（2）两个总体比例之差：z检验

（3）两个总体方差之比：现实中比较不同方法生产产品的稳定性，或测量工具的精度等。F检验

5. 样本量确定：根据估计量公式倒算得出，需要给定估计误差，置信水平，总体方差等参数来确定。

第六节：假设检验

1. 假设检验两部分：原假设、备择假设

2. 两类错误：

一类错误，弃真(alpha)：原假设为真，但我们错误拒绝了。

二类错误，纳伪(beta)：原假设为伪，但我们没拒绝

关系：二者此消彼长，要想同时减低，只能增大样本量。我们主要看重一类错误。

3. 假设检验流程：

（1）首先提出原假设和备择假设（单侧还是双侧检验）：

H0: mu = 3190 （克）

H1: mu != 3190 （克）

（2）计算统计量

（3）确定alpha值以及临界值

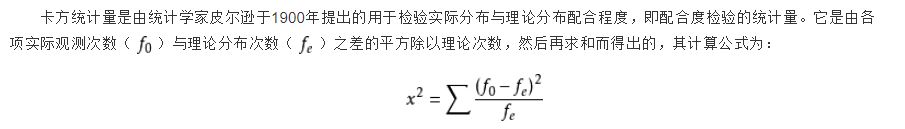
（4）将统计量的值与临界值比较，拒绝或不拒绝原假设（或比较P值）

4. 注意：统计检验不是非此即彼逻辑，有时可能会出现“单侧检验悖论”——有罪推定与无罪推定。当产品质量一贯很好，我们认为稍差的样本并不能作为整批产品非优的证据，如果一贯表现不好，我们认为测试合格的样本也不能成为整批产品为优的有力证据。如何设定原假设？一般把希望证明的命题放在备择假设上，把原有、传统的观点放在原假设上。

第七节：分类数据分析

1. 分类数据分析的数据类型是离散型的，结果是频数，因此主要利用卡方分布，进行拟合优度检验和独立性检验，主要工具是列联表。

2. 拟合优度检验 —— 一个分类变量



例子：泰坦尼克号海难中，我们知道船上有男性1738人，女性470人，共2208人。海难发生以后，幸存者有718人，其中男性374人，女性344人。因此，海难存活的总体比率 = 718/2208=0.325，如果死亡事件不受性别的影响，那么男性中会存活1738\*0.325 = 565人，女性应该存活470\*0.325 = 153人。565人和153人就是男性和女性的期望频数。因此，提出假设检验：

H0: 性别不影响生存概率

H1：性别影响生存概率

依据卡方统计量公式，可以算出卡方值 = (374-565)^2/565+(344-153)^2/153 =303。自由度为df = R-1,R为分类个数，本例中df = 2-1=1，经查卡方分布表，卡方（1）=2.706，因此拒绝原假设。

3. 独立性检验 —— 两个分类变量

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 地区 | 一级 | 二级 | 三级 | 合计 |
| 甲 | 52 | 64 | 24 | 140 |
| 乙 | 60 | 59 | 52 | 171 |
| 丙 | 50 | 65 | 74 | 189 |
| 合计 | 162 | 188 | 150 | 500 |

上表有3行，3列（不算合计），因此为3\*3列联表，独立性检验是看行变量与列变量是否相互独立，本例中就是产地与原材料质量之间是否有依赖关系。构建假设检验：

H0：地区与原料等级之间是独立的

H1：地区与原料等级之间不独立

分析：

关键是求各个单元格的期望频数。以第一个单元格（甲地区和一级品交叉）为例，从地区看，甲地区原料的频率 = 140/500，从一级品看，一级原料的频率=162/500，如果地区和原料等级之间相互独立，设A = 样本来自甲地区的概率，B=样本来自一级原料的概率，则：

第一个单元格的期望频率 = P(AB) = P(A)P(B) = (140/500)(162/500) = 0.09

第一个单元格的期望频数 = 0.09\*500 = 45.36

单元格频数一般公式 = RT\*CT/n，RT为单元所在行的合计，CT为单元所在列的合计，n为总样本量。

依次类推，可以算出整个表的卡方值 = 19.82。卡方自由度 =(R-1)(C-1)=2\*2=4,若alpha =0.05，则卡方临界值=9.488，拒绝原假设，地区和原料等级之间存在依赖关系。

4. 列联表中的相关测量

（1）φ相关系数 = sqrt(卡方/n)，φ=0，两个变量独立，φ=+-1，两个变量完全相关。φ的绝对值越接近1，则两个变量相关程度越高，φ系数无上限。

（2）列联系数：必小于1

（3）Cramer-V系数：值域在0-1之间，越接近1相关性越大。

5. 列联表的注意事项

（1）一般而言，列与行的变量没有明确规定，地位一样。但如果假设X是Y的原因，则一般把X放在列的位置，Y放在行的位置。例如;

|  |  |  |
| --- | --- | --- |
| 价值取向(Y) | 制造业（X） | 服务业（X） |
| 物质报酬 | 72% | 56% |
| 人情关系 | 28% | 44% |
| 合计 | 100% | 100% |

（2）在用卡方分布进行独立性检验时，要求样本量必须足够大，每个单元格中的期望频数不能太小，如果只有两个单元，则每个单元期望频数要在5以上。若有两个以上单元，则期望频数少于5的单元格不能多于20%。补救方法是合并部分单元格。

第七节：方差分析

1. 方差分析广泛用于分析心理学、工程、医药领域。基本原理是比较多个总体的均值是否相等，即自变量是一个分类变量，因变量是一个数值变量。主要分单因素方差、双因素方差等。

2. 方差分析比两两组间比较更有效率，同时因为将所有信息结合在一起，因此分析结果也更可靠。例如，4个样本，如果两两比较，就必须比较6次（C42），并且如果每一次的一类错误是0.05，那么6次比较会使错误率上升至1-（1-0.05）^6 = 0.265，即置信水平降至0.735，方差分析只需检验一次，因此避免了错误累积。

3. 方差分析中，要检验的对象成为因素或因子(factor)，因素的不同表现被称为水平或处理（treatment）。

例子：假设有一个数据，里面有4个行业（零售业、旅游业、航空业、家电制造业）以及一年内每个行业，各抽取5个公司，并记录收到的投诉次数。我们要分析行业是否对被投诉次数有显著影响。本例中，行业就是被检验的因素，行业下不同的类别，称为因素的具体表现（水平或处理），由于是单因素（行业）以及4个水平，因此称为单因素4水平试验。因素的每个水平，看作一个总体，例如四个行业就是四个总体，因变量就是投诉数量。如果四个总体下的投诉数均值不同，我们就认为行业这个因素对投诉数量有影响。

|  |  |  |  |
| --- | --- | --- | --- |
| 零售业 | 旅游业 | 航空业 | 家电制造业 |
| 57 | 68 | 31 | 44 |
| 66 | 39 | 49 | 51 |
| 49 | 29 | 21 | 65 |
| 40 | 45 | 34 | 77 |
| 34 | 56 | 40 | 58 |

4. 图形描述：画散点图，横轴是因素下不同水平，纵轴是因变量

5. 误差分解：来自水平内部的数据误差，称为组内误差，例如零售业内5家企业投诉次数之间的误差，我们假设是随机的。不同行业之间的观测值也是不同的，因此不同水平之间数据误差称为组间误差，这一差异可以来自抽样本身带来的随机误差，也可能来自组间系统性因素导致的系统误差。所以组间误差 = 系统误差+随机误差。

6. 全部数据误差的大小的平方和称为总平方和，即SST，反映了全部观测值的离散状况。反映组内误差大小的平方和称为组内平方和，也叫误差平方和或残差平方和，记为SSE。同理，组间误差大小称为组间平方和，SSA，它反映了样本均值之间的差异程度。

SST = SSA + SSE

7. 误差分析：如果行业对被投诉次数无影响，那么组间误差中应只包含随机误差而没有系统性误差。此时，组间误差与组内误差经过平均后的数值应该非常接近，即二者比值为1。如果，行业因素确有影响，那么不同水平的均值就会包含系统误差，那么组间误差平均后，将大于组内误差平均。

8. 方差分析三个基本假定：

（1）每个总体都服从正态分布，样本是从其中抽出的随机样本。

（2）各个总体的方差必须相同。

（3）观测值之间是独立的。

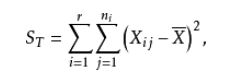
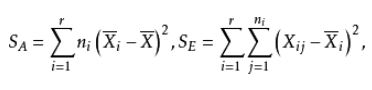
9. 单因素方差分析 —— 只有一个分类变量

（1）提出假设：

H0：mu1 = mu2 = … = muk —— 自变量对因变量没有显著影响

H1: mu1,mu2,…muk 不完全相等 —— 自变量对因变量有显著影响

（2）构造统计量：计算各样本均值，即求各个水平下的组内均值，计算全部样本的总均值，计算各误差平方和。SST为每个观测值与总均值的误差平方和（不乘1/(n-1)），SSA为各组均值与总均值的误差平方和（也称因素平方和）。



（3）计算统计量：由于各误差平方和的大小和观测值的个数有关，所以要消除观测值数量对误差平方和大小的影响，因此就要做自由度调整，其结果也被称为均方(mean square)。

SST自由度为n-1, SSA自由度为k-1，k为因素水平个数，SSE自由度为n-k。

MSA = 1/(k-1)\*SSA —— 组间均方（组间方差）

MSE = 1/(n-k)\*SSE —— 组内均方（组内方差）

F = MSA/MSE ~ F(k-1,n-k)

（4）检验决策：如果原假设成立，则MSA/MSE比值将接近1。

10. 关系强度测量

当组间平方和比组内平方和大，而且大到一定程度时，就意味着两个变量之间的关系显著。 因此，度量可以采用R-square = SSA/SST，其解释为行业（自变量）对被投诉次数（因变量）的影响效应占总效应的百分比，1-R2则为残差效应的占比，而R-square的平方根，则可以证明，等于自变量与因变量的相关系数—rho。

11. 双因素方差分析 —— 两个分类变量和一个数值变量

例子：5个地区，4个品牌的销量数据。分析究竟是一个因素在起作用，还是两个因素都起作用，还是两个因素都不起作用；有时还需要判断，是两个因素都起独立作用，还是存在交互作用（两个因素搭配在一起会对因变量产生新的效应）。

（1）SST = SSR（因素1误差平方和）+ SSC（因素2误差平方和）+ SSRC（因素1与因素2交互平方和）+ SSE

（2）FR = MSR/MSE ~ F（k-1,(k-1)(r-1)）, FC = MSC/MSE ~ F（r-1,(k-1)(r-1)）

检验：FR、FC分别与F分布的给定临界值比较。

第八节：一元线性回归

1.现实中，回归分析的例子很多，农业生产中药研究农作物产量与施肥量之间的关系，商业分析中，要研究广告费支出与销售量之间的关系等等。两个变量间的关系总体上分为：函数关系、相关关系。前者是确定关系，后者存在干扰因素。比如工资收入与教育水平，二者不是函数关系，但具有较强的相关关系。方差分析的自变量是分类变量，回归分析事实上是对它的扩展，将分类变量扩展到连续变量，它们的因变量都是数值型的变量。

**\*为了更好地理解回归分析，首先要理解相关分析**

2. 相关分析解决以下问题：

（1）变量之间是否存在关系？

（2）如果存在关系，是什么样的关系？

（3）关系强度如何？

（4）样本反映的关系是否能推断总体变量间的关系？

3. 分析步骤

（1）先把两个变量画一个散点图，观察一下形态，判断大致是线性关系、非线性关系、无关系，正相关、负相关等等信息。

（2）计算相关系数：r = cov(x,y)/var(x)var(y)，相关系数分为总体相关系数和样本相关系数，**相关系数只能度量两个连续变量的线性关系大小！**相关系数的值域在（-1,1）之间，相关系数与原数据的大小尺度、单位无关。r只能度量相关，但不能断因果。|r|越接近0，表示相关性越弱，|r|>=0.8，强相关，0.5<=|r|<=0.8，中相关，0.3<=|r|<0.5，弱相关，|r|<0.3基本无相关关系。

（3）只有当r服从正态分布时，才能应用正态分布检验，然而往往这个假设不满足，因此r 的检验一般用t检验。

4. 线性回归

（1）回归分析不仅能看两个变量间是否有关系，还能确定变量间的数学关系式，这一点优于方差分析和相关分析。

（2）通过对关系式的可信度进行统计检验，能够找出哪些变量的影响是显著的

（3）利用估计出的关系式，可以进行预测，并估计出预测值的可信度

（4）回归分析属于“有监督学习”，必须区分y和x

（5）非线性关系可以转化为线性关系，应用于线性回归模型中估计

5. 线性回归的假定

（1）x~y之间是线性关系

（2）各个解释变量不完全相关，即解释变量的样本观测值矩阵是满秩矩阵（不存在多重共线性）

**\*假设（1）和（2）保证了模型是可以估计的**

（3）x的取值是固定的或随机的，但不能与误差项相关

**\* 假设（3）保证了系数估计的无偏性**

（4）误差项是一个期望值为0 的随机变量，且服从整体分布。

**\* 假设（4）保证了回归方程是一条直线，并且可以用t检验和F检验**

（5）对于所有x，误差项的方差相同，即同方差假设。

（6）误差项之间是独立的，即不存在自相关。

**\* 假设（5）（6）保证了估计的标准误是无偏的，严重的多重共线性也会使标准误增大**

**6. 注意区分以下概念（请查询相关资料补全）：**

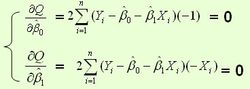
**（1）回归模型**

**（2）回归方程**

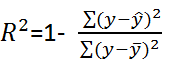
**（3）估计的回归方程**

**7. 最小二乘法的推导（一元回归，多元回归）**



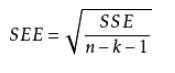


8. 回归直线的拟合优度，拟合优度越高越好吗？



9. 估计标准误

估计标准误是测度各实际观测点在直线周围的散布状况的统计量，它是均方残差（MSE）的平方根：



SSE是残差的平方和，即R2右侧分子部分，SSE/n-k-1 = MSE

思考：估计标准误对参数估计的意义是什么？

10. 显著性检验

（1）线性关系检验

构造F统计量，F=MSR/MSE ~ F(k,n-k-1) 。MSR是经过调正的回归平方和，MSE是经过调整的残差平方和，思想是若回归平方和显著大于残差平方和，即比值远大于1，那么几个解释变量就对y的解释是有贡献的。因此，F检验在一元回归中退化为t检验，多元回归中，检验几个变量是否联合起来对y有显著影响。

（2）回归系数检验

样本回归系数也是样本的函数，不同样本会有不同的回归系数，因此需要进行假设检验。回归系数的期望和标准误的公式要理解。

11. 回归方程的预测

（1）预测y均值与均值的置信区间

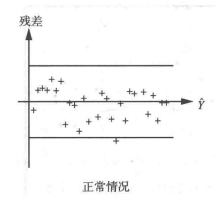
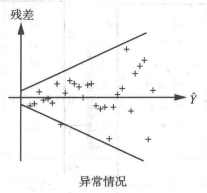
（公式略）

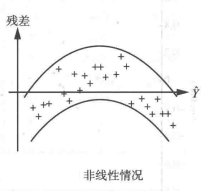
（2）预测的y的个别值与其置信区间

（公式略）

（3）后者置信区间要宽于前者

12. 残差分析 —— 模型假设的检测





**请思考：第二种和第三种残差可能预示模型存在什么问题？**