**18. 统计学基础知识Ⅰ**

**（一）变量与随机变量**

**变量——**取值在不断变化的量；

例如，X, Y分别表示一组正常人中任何一人的身高和耳朵的个数，则X是变量，Y是常量。

**随机变量**——事先不知道变量将会取什么值，仅当试验或调查或测定之后，才知道变量的具体取值；

例如，X,Y分别表示一组动物注射某种药物的剂量，和药物在动物体内作用一定时间后的反应，则X是一般变量，Y是随机变量。

**（二）统计资料及其分类**

**统计资料——**具备三要素：（1）变量名；（2）专业含义；（3）变量的具体取值。

**定量资料**——测定每个观察单位某项指标量的大小，所得资料；

**计量资料**——取值带度量单位，还可以带小数（标志测量的精度）的定量资料；

**计数资料**——取值可以带度量单位，但不能带小数的定量资料；

例如，脉搏的次数；

**定性资料**——观测每个观察单位某项指标性质的不同状况，所得的资料；

**名义资料**——指标性质的不同状况之间，在本质上无数量大小或质量好坏或先后顺序之分，的定性资料；

例如，血型（A型、B型、AB型、O型）、职业等；

**有序资料**——若指标性质的不同状况（≥3）之间，在本质上有数量大小或质量好坏或先后顺序之分，的定性资料；

例如，药物疗效（治愈、显效、好转、无效、死亡）。

**（三）离散型随机变量概率分布**

**离散型随机变量**——随机变量在其取值区间内只取一些孤立的数值，通常是0或正整数；

例如，某药物对某病的治愈率是80%, 现在用该药物治疗该病患者100人，则能治愈的人数X是一个离散型随机变量，X的可能取值为：0、1、……100.

**离散型随机变量的概率分布**——将某个指定的离散型随机变量的所有可能的取值一一列举出来，再将该随机变量取每个特定值的可能性（即概率）也一一写出来，这两部分信息放一起。

例如，对于前例，治愈人数X的概率分布可表示为：



常见的离散型随机变量的概率分布有：二项分布（Binomial）、泊松分布（Poisson）、超几何分布（Hypergeometric）等。

**（四）连续型随机变量概率分布**

**一、连续型随机变量及其概率分布**

**连续型随机变量——**随机变量的取值充满一个区间，无法一一列出。刻画连续型随机变量是用概率分布函数、概率密度函数来描述的。

**概率密度函数**——设连续型随机变量X有概率分布函数F(x), 则F(x)的导数f(x)=F’(x), 称为X的概率密度函数。

解释：取定一个点x，按分布函数的定义，事件 {x < X < x+h}的概率，应为F(x+h)-F(x), 所以比值 [F(x+h)-F(x)]/h 可以解释为在单位长度的区间内随机变量取值的概率。令h趋于0, 则该比值的极限即F’(x)=f(x).

连续型随机变量X的密度函数f(x)满足如下性质：

（1）f(x) ≥ 0；

（2）

（3）对任意常数 a < b 都有



常见的连续型随机变量的概率分布有：正态分布（Normal）、分布（Chi Square）、t分布、F分布。

**二、连续型随机变量概率分布的应用**

**正态性检验**

正态分布是最重要的一类分布，生产与科学实验中很多随机变量的概率分布都可以近似地用正态分布来描述。一般若影响某一数量指标的随机因素很多，而每个因素所起的作用不太大，则这个指标的取值近似服从正态分布。

正态性检验，即检验一组数据是否服从正态分布。一般用W检验和D检验，SAS 规定：当样本含量n ≤2000 时，结果以W检验为准，当样本含量n >2000时，结果以D检验为准。

**卡方检验**，一般是检验名义型变量的实际观测次数和理论次数分布之间是否存在显著差异（要求理论次数≥5，否则需要进行校正；卡方值越大，代表统计量与理论值的差异越大）。主要应用于列联表的独立性检验和拟合性检验。

注：（1）**列联表**，是由两个以上的变量进行交叉分类的频数分布表，是用于提供基本调查结果的最常用形式，可以清楚地表示定性（定类）变量之间是否相互关联；

（2）**拟合性检验**，检验单个多项分类名义型变量各分类间的实际观测次数（样本数据的个数）与理论次数（根据理论或经验得到的期望次数）之间是否一致，其自由度通常为分类数减去1.

**t检验**，用于检验两样本平均数是否一致，可以对单样本、配对样本、两独立样本的均数进行t检验，适用条件是正态性、方差齐性。

**F检验，**又称方差齐性检验（方差齐性，即两方差没有明显差别）。主要用于：均数差异的明显性检验、分别各有关因素并估量其对总变异的作用、剖析因素间的交互作用、方差齐性（Equality of Variances）检验等。

**（五）总体和样本**

总体(population)：所研究对象的全体组成的集合。

样本(sample)：从总体中抽取的部分对象（个体）组成的集合。

样本中包含个体的个数称为样本容量。

容量为n的样本常用n个随机变量X1，X2，…，Xn表示，其观测值（样本数据）则表示为x1，...，xn，为简单起见，有时不加区别。

**（六）参数与统计量**

参数(parameter)：用来描述总体特征的概括性值。如总体平均值（μ）、总体方差（***σ*2**）、总体比例（π）等。

统计量(statistics)：用来描述样本特征的概括性值。如样本均值（）、样本方差（***s*2**）、样本比例（P）等。

1. **表示数据位置的统计量**

常见的位置统计量有：均值、中位数、分位数、众数等。

1. 均值（Mean）

均值是所有观测值的平均值，是描述数据取值中心位置的一个度量：



1. 中位数（Median）

中位数是描述观测值数据中心位置的统计量，大体上比中位数大和小的数据各占观测值的一半。

中位数的一个优点是它不受个别极端数据的影响，具有稳健性。其计算方法是：首先将数据从小到大排序为：x(1)，...，x(n)，然后计算

1. 众数（Mode）

观测值中出现最多的数称为众数。众数用得不如均值和中位数普遍。在属性变量分析中，常需考虑频数，因此众数用得多些。

1. 百分位数（Percentile）

分位数也是描述数据分布和位置的统计量。0.5分位数就是中位数，0.75分位数和0.25分位数又分别称为上、下四分位数，并分别记为Q3和Q1.

四、**表示数据分散程度的统计量**

1. 极差（Range）与半极差（Interquartile range）

极差就是数据中的最大值max{xi}和最小值min{xi}之差；

上、下四分位数之差Q3 – Q1称为四分位极差或半极差，它描述了中间半数观测值的散布情况。

2. 方差（Variance）

方差是由各观测值到均值距离的平方和除以观测量减1：



3. 标准差（Standard Deviation）

方差的开方称为标准差：. 标准差的量纲与原变量一致。

4. 变异系数（Coefficient of Variation）

变异系数是将标准差表示为均值的百分数，是观测数据分散性的一个度量，它在比较用不同单位测量的数据的分散性时是有用的：



**四、表示数据分布形状的统计量**

偏度和峰度是描述数据分布形状的指标。

1. 偏度（skewness）

偏度是刻画数据对称性的指标。其计算公式为：



在SAS中：（1）关于均值对称的数据其偏度为0；

（2）左侧更为分散的数据，其偏度为负，称为左偏；

（3）右侧更为分散的数据，其偏度为正，称为右偏。

1. 峰度（kurtosis）

峰度描述数据向分布尾端散布的趋势。其计算公式为：



利用峰度研究数据分布的形状是以正态分布为标准（假定正态分布的方差与所研究分布的方差相等）比较两端极端数据的分布情况，若

（1）近似于标准正态分布，则峰度接近于零；

（2）尾部较正态分布更分散，则峰度为正，称为轻尾；

（3）尾部较正态分布更集中，则峰度为负，称为厚尾。

**五、其它统计量**

1. 均值的标准误（Std Error Mean）



2. 校正平方和（Corrected sum of squares）



1. 未校正平方和（Uncorrected sum of squares）



1. k阶原点矩（origin moments）



注：A1即为均值.

1. k阶中心矩（central moments）

