**统计学方法的选取**

**前言**

**统计：**使用数学原理和方法，研究数据的搜集、整理与分析的科学，对不确定的数据作出科学的推断**。**

**医学统计学：**统计方法应用于医学实践

本课程的目的：不需要了解公式来源，只要求了解其意义、用途和应用条件

**1、几个重要概念：**

1）**研究单位**：也叫观察单位，unit，即：研究中的个体，依据研究目的而确定。

2）**变量**：研究单位的研究特征；变量值：即取值

3）**同质**：根据研究目的给研究单位确定相同的性质；

**变异**：同质研究单位中变量值的差异；

研究青岛市2004年7岁男孩身高范围；

同质：同一个市，同7岁，同男孩，同无影响身高的疾病。

变异：这些男孩身高有高有矮。

4）**总体**：同质研究单位的某种变量值的集合。

A有限总体：研究单位数是有限的。（某地2002年正常成年男性的红细胞正常数范围 ）

B无限总体：研究单位数是无限的。（高血压患者）

**样本**：从总体中随机抽样抽取的具有代表性的一部分。

**样本含量**：样本中包含的研究单位数。

5）**参数**：根据总体个体值统计计算出来的描述总体的特征量。（希腊字母）

**统计量**：根据样本个体值统计计算出来的描述样本的特征量。（拉丁字母）

总体参数通常未知，因此，用样本统计量估计总体参数。

6）**误差**：观察测量值与实际真值之差。

A 系统误差：由于仪器未矫正/观察者障碍/医生测量水准，使得观察值并非分布在真值两侧，而是有方向、系统性、周期性的偏离真值。

（可以通过实验设计和技术措施消除/减少）

B 非系统误差：由于观察者偶然操作失误造成的误差。（认真检查以消除）

C 抽样误差：由于抽样导致样本统计量与总体参数的差别（不可避免、有统计规律的）

1. **频率**：一次可重复的随机试验中产生的某结果次数占所有结果的比例；  
   **概率**：一次可重复的随机试验中产生的某结果的可能性；p(A)

p(A)=0，不可能事件；p(A)=1，必然事件；p(A)：（0，1）随机事件

小概率事件：统计分析中基于某置信程度下的概率推断，习惯称p(A)≤0.05或p(A)≤0.01称为小概率事件，认为在单次随机实验中不可能发生。

**2、推荐的统计分析软件：**

1）**excel**

**2) SPSS**

**3)python**

**4)R语言：绘图可视化**

**5）科研平台：描述统计：t检验，方差分析，卡方检验**

3、开展医学统计工作的步骤：

第一，实验设计

此环节是最为关键的一块，实验设计决定了我们的目标和分析维度有多少。

**这部分，我们希望大家能形成记录和整理成文档的习惯。**

**正是这些小工作能提高我们研究工作的效率。**

第二，收集资料

第三，整理资料

**这部分，我们同样希望大家能形成记录和整理表格的习惯。**

第四，分析资料

1）统计描述（statistical description）

选用恰当合适的统计指标/统计量/统计表/统计图，对资料的数量特征及分布规律进行描述和测定

2）统计推断（statistical inference）：在一定的可信程度下由样本信息推断总体特征。

参数估计：由样本的统计指标来推断总体的相应指标。

假设检验：由样本的差异来推断总体之间是否存在差异。

**------------------------------------------------------------------**

**本次关于医学统计学的培训主要分为三讲：**

**第一讲，医学统计方法的选择**

**第二讲，单因素分析和参数检验**

**第三讲，多因素分析：相关及回归**

**------------------------------------------------------------------**

**本次培训为第一讲，医学统计学方法的选择；**

**本次讲课的主要目的是：让大家对医学统计学的基本概念和流程进行了解。**

**目录：**

**第一部分，解读一篇文章**

**第二部分，统计方法**

**第三部分，统计方法的选择**

**第四部分，数据转换**

**最后是，一个附录，供大家进行查表。**

**--------------**

1. **，一篇SCI文章中用到的统计学方法：**

此文章为使用机器学习和组学方法提取冠状动脉周围脂肪的组学特征研究其在对不良心脏事件的预测能力。

我们把这篇文章进行拆解，首先是摘要部分，这里显示的是C统计量，95%置信区间；后面是C统计量，组间差异有统计学意义。

然后是方法部分；统计分析：分类变量：用到了皮尔逊，卡方检验；连续变量用了U检验；使用R语言和spss软件

最后是结果部分：第一个，统计表+参数检验：两个组间人口信息，做了参数检验；同样的，统计图+参数检验；置信区间+参数检验

可见在论文的各个部分需要使用统计学对我们的结果，在一定的检验水平下进行评价。

**下面我们就来学习统计方法怎么选择，也就是我们的第二章部分：统计方法及基本概念。**

首先，统计方法的选择，跟我们的数据类型是息息相关的，一般上的研究流程是：根据课题的研究目的，筛选我们需要分析的变量，然后根据这些变量的数据类型、分布情况和设计方法，选取恰当的统计学方法。

1. **：资料的类型。 对于一个变量，一般上分为：数值变量和分类变量；**

数值变量的值为数，对加减平均运算是有意义的。按照取值分为离散型和连续型。

分类变量的值为字符或数，按照等级情况分为，有序型和无序型。

有序型分类变量，例如尿蛋白水平，其取值之间是有等级的；

无序型分类变量，例如二分类和多分类变量。

同样，对这些资料也称为：计量资料、等级资料和计数资料。

**下面我们来举一些例子，下面是一些常见的临床项目。**

**第二节：研究目的**

通常，分为描述和推断；所谓的统计描述，是指对资料的固有基本特征进行统计分析。一般通过统计表、统计图和统计指标进行。

再者，是统计推断，指在一定的可信程度下由样本信息推断总体特征。

分为参数估计和假设检验

最后是，既有描述也有推断。分为相关分析和回归分析。

**第一部分：统计描述**。指：运用统计指标对数据的特征和分布规律进行客观的表达。根据资料的类型，下表给出了常用的指标。

一般上，对于正态分布的数值变量，我们通常使用均数和标准差进行描述；如果是非正态分布的，则使用中位数和四分位数间距进行描述；

对于分类资料，一般上使用率，或者构成比进行描述。

以下是一个临床特征统计表，红色部分是描述，p值为检验。

里面有数值变量也有分类变量，分别使用均数±标准差，率/构成比进行描述。

同样的，在英文文献中，临床人口特征也是这样进行描述的。

同样，也可以使用统计图进行描述，比如柱状图、饼图，可以直观的体现构成率，占比等。在我们的科研平台对于训练数据也有统计，在别的文献中对于组学特征也进行描述。

**第二部分：统计推断**。指在一定的可信程度或概率保证下，根据样本信息去推断总体特征。常用的很多参数检验和回归分析属于统计推断的研究。

**第三部分：统计描述+推断**。有些材料中把这部分归为推断了。一般上，是研究变量与变量之间的相关关系和因果关系。

在我们这里把相关和回归分析拿出来进行研究。

相关的变量研究，二者的地位平等，随机正态分布，使用相关系数来描述二者的线性相关程度。

回归的变量研究，二者有因果关系，使用回归系数揭示x对y的影响的大小，同时会通过回归方程进行预测和分析。

**第三节：研究的设计方法**

通常，临床研究在设计分组的时候，有几种方法：

第一：单组随机设计：实验组前，实验组2，对照组，均为随机设计，不使用人为实验干预。

第二：配对设计：例如，同一个患者的脑部左右进行对照；同一批胃癌患者，使用不同药物治疗，观察效果。

第三：重复测量设计：同一研究对象的同一测量指标在不同时间点的多次测量结果。

第四：多因素设计：是在一个实验中包含有两个或两个以上的自变量，且每个自变量的水平在两个或两个以上的实验设计

**第四节：数据的分布特征**

通常分为正态分布或非正态分布。大部分自然界的随机事件趋于正态分布，其概率密度函数为正态分布特性，中心发生概率高，两端概率低。

在正态分布的数据中，我们认为最终的估计的参数真值在头顶的检验水平下，一定的区间内取值，称为置信区间； 检验水平一般为0.05或0.01；当p值<0.05/0.01，则认为小概率事件发生，则两者的差异具有统计学意义；反之则二者差异无统计学意义。

一般上，我们建议先对数据做正态检验，根据正态分布正态分布的数据选择统计描述和统计推断方法。正态分布的数据一般使用均数、标准差描述，t检验/方差分析进行推断。非正态的数据使用中位数/四分位数间距进行描述，使用秩和检验进行推断。

-----------------------

因此我们重新研读一篇文献，可以发现，在这篇文献中用到的统计学方法，其目的无非是，第一，描述资料的特征；第二，对样本推测总体的结果，在统计学上看看是有意义。

1. **：方法选择**

**细化我们的统计分析流程，进一步来说：**

**根据不同的研究设计和不同的变量类型选择统计方法如下：**

1. **：数据分析：**

**第一部分：对于数值变量**

一般上先做正态分布检验，如果是正态分布的情况，我们就做参数检验；非正态分布使用非参数检验，当然非正态数据也可以转换成正态分布；

**第二部分，对于分类变量，分为单因素分析和多因素分析。**

单因素分析，指单一因素在两组之间的差异分析。一般通过参数检验方法。

多因素分析，通常指影响因变量的因素有多个，这种多个自变量影响一个因变量的问题可以通过多元回归分析来解决。

第三部分，对于一些变量我们可以进行转换，为了方便纳入研究，常用的有3种。一是，非正态数据转换成正态数据，通过数学处理方法。

二是，分类变量变为哑变量。当分类变量是3，4，5分类时，可以增加哑变量个数，通过重新编码，进行二值处理。

三是，连续变量转换为分类变量。例如拿到一组血压数据，但是想研究高、中、低血压的情况，可以找截断点进行分类。

----------

最后，给出了一个目录表，供大家进行查阅

最后，是本次培训的总结：

根据此次培训，我们学习了医学统计方法的基本概念和选取统计方法的步骤。下一讲我们讲围绕单因素分析和参数检验，对如何选取方法进行讲解。