**医学统计学--入门到精通**

**第一章：几个重要的统计学概念：**



**1、总体与样本**

**1）个体**（individual）：观察单位(observed unit)，统计研究的基本单位。

**2）总体**（population）：同质的所有观察单位的某种观察值的集合。

3）有限总体（finite population）：总体中明确规定了空间、时间、人群范围的有限个观察单位。

4）抽样研究（sampling research）：从总体中抽取样本，根据样本信息来推断总体特征的方法。

5）抽样（sampling）：从总体中抽取部分观察单位的过程。

为保证样本的代表性，抽样时必须遵循随机化（randomization）原则。

6）样本：从总体中随机抽取的部分观察单位的实测值的集合。

7）样本含量（sample size）：该样本所观察单位的数。

----研究的目的：获取样本，并且通过样本信息来推断总体特征。

**2、变量与资料**

1）变量（variable）：能表现观察单位变异性的某项/某些特征。

变量值：（value of variable）：变量的观察值

观察值：(observed value)

2）资料（data）：变量值构成资料。

按照定量/定性性质，分类为：

①计量资料：（measurement data） 观测每个观察单位某指标的大小。

又称定量资料（quantitative data）或**数值变量**（numerical variable）； 其变量值为定量的，表现为**数值大小，一般有度量单位**。

根据观察值是否连续，分为：

1）连续性continuous：在实数范围内任意取值，如：身高，体重，血压等。

2）离散型discrete：只取整数值，如：人数等

②计数资料：（enumeration data）每个观察单位按某种属性/类别进行分组计数，分组观测。

又称定性资料（qualitative data）或**无序分类变量**（unordered categorical variable），名义变量资料（nominal variable）；其变量值为定性的，表现为**互不相容的属性或类别**，如：试验结果的阴阳性、家族史有无。

分为：

1）二分类：观察两类互相对立，互不相容。如：疾病阴阳、有无；

2）多分类：互不相容的多个类别。如：血型分布：A,B,AB,O等。

③等级资料：（ranked data）每个观察单位按某种属性的不同程度进行分组计数，分组观测。

又称半定量资料（semi-quantitative data）或**有序分类变量**（ordinal categorical variable），其变量值是半定量的，表现为**等级大小/属性程度**，如：反应强度+，++，+++，++++，+++++五级；疗效：治愈、显效、好转、无效四级。

统计分类方法的选用，与资料类型密切联系。

各类资料之间可以相互转化。

一般上，能测量的指标——设计为定量资料；

**3、误差**

误差（error）：测量值与真实值之差。

按照产生原因和性质，分为：

随机误差（random error）

非随机误差（nonrandom error）：系统误差/非系统误差

1）随机误差：不恒定、随机变化的误差，由多种无法控制的因素引起。

随机误差是不可避免的，在大量重复测量/抽样中，可能呈现一定规律性分布。

正态分布的随机误差可以用医学统计学方法分析，统计分析主要针对抽样误差。

随机测量误差：（random error of measurement）:多次测量，每次测量结果依然出现一些随机变化。

抽样误差：(sampling error)：抽样过程中由于抽样的偶然性而出现的臭抽样误差。

2）系统误差：实验过程中产生的误差，值恒定不变或遵循一定变化规律，产生原因往往可知或可能掌握。

如：样本不均匀、分配不随机、不同操作者差异、不标准仪器、非实验因素不平衡。

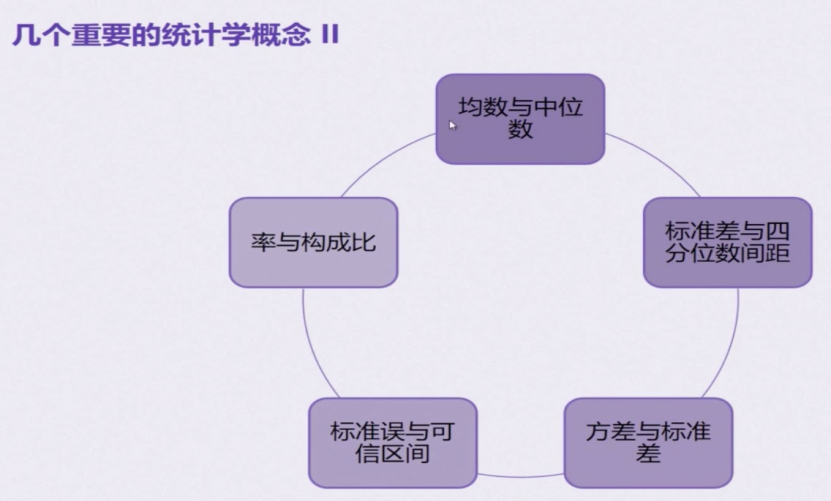
3）非系统误差：实验过程中由研究者偶然失误而造成的误差。

如：过失误差；

**4、频率与概率**

1）频率（relative frequency）：在实验测量下，出现某结果的比例。

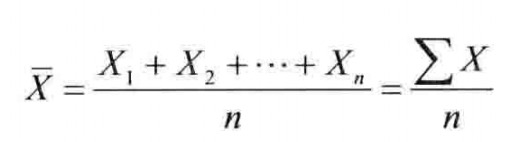
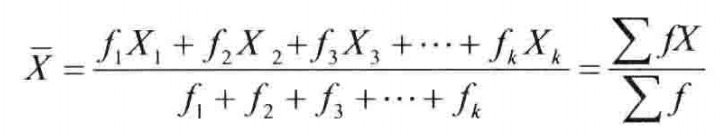
2）概率（probability）：描述随机事件发生的可能性大小的一个度量；



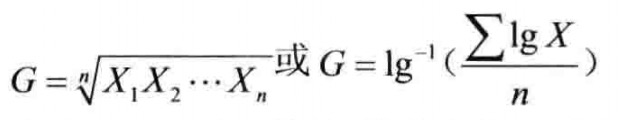
**5、均数及中位数**

常见平均数：算数均数、几何均数、中位数

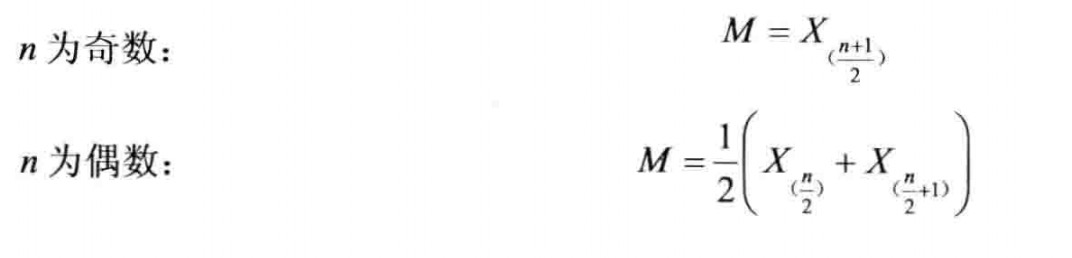
1）算数均数（mean）：反映一组呈现对称分布的变量在数量上的平均水平。

或

2）几何均数（mean）：反映一组经对数变换后呈现对称分布的变量在数量上的平均水平。（比如免疫学指标）



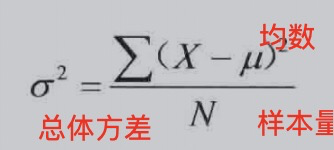
1. 中位数（median）：将n个变量值按照从小到大排列，位置位于中间的那个数。n为奇数时，取居中的变量值；n为偶数时去次居中两个变量值的均数。



4）四分位数间距：四分位数（quartile）把全部变量值分为四个部分的分位数；四分位数间距（quartile range）：第3四分位数与第1四分位数相减而得，记为QR；

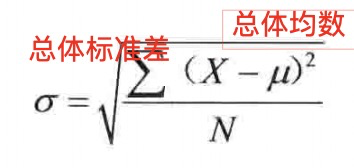
**6、方差与标准差**

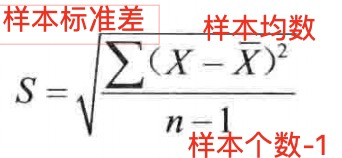
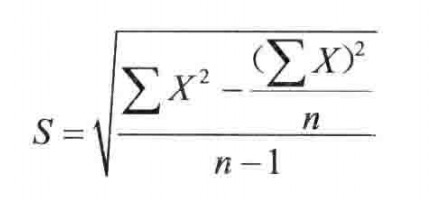
1. 方差（variance）：又称均方差，反映一组数据的平均离散水平。

总体方差：

样本方差：使用样本均数估计总体均数；使用样本个数n-1代替N；

1. 标准差（standard deviation）：方差的正平方根

总体标准差：

样本标准差： 

1. **参数估计与置信区间**
2. 参数估计：点估计&区间估计

用样本指标值（统计量）推断总体指标值（参数）

1. 点估计（point estimation）：用相应的样本统计量直接作为总体参数的估计量；方法简单但是**未考虑抽样误差的大小**。



1. 区间估计（interval estimation）：按预先给定的概率（1-a）所确定的包含未知参数的一个范围（置信区间：confidence bound/Confidence interval,CI）。常取95%或99%。

表达为：（L，U）；（可信上限，可信下限）

**医学统计工作的基本步骤**

1、设计：最关键环节

1）文献调研

2）研究目的和假说

3）研究总体、研究对象、观察单位

4）抽取样本：多少？怎样抽取？

5）施加干预措施

6）如何设置对照？

7）如何获得原始资料？

8）需要设置哪些指标研究观察结果？

9）资料如何整理和汇总？统计指标如何计算？

10）如何控制误差和偏倚？

11）预计得到什么结果？

12）经费和时间周期预算

2、收集资料

3、整理资料

1）原始数据净化：原始数据的清理、检查、核对、纠正错误。

2）系统化和条理化：根据研究目的，将原始数据合理分组并归纳汇总。

4、分析资料

又称统计分析，包括：有关统计指标的选择和计算、统计图表的绘制、统计方法的选取、spss/ststs/sas等软件的使用。

目的：在表达数据特征的基础上，阐明事物的内在联系和规律性。

包含：

1）统计描述（statistical description）：

选用恰当合适的统计指标（统计量，statistic）

选用合适的统计表和统计图

描述和测定：对资料的数量特征及分布规律

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。（estimation of parameter）

假设检验：由样本的差异来推断总体之间是否存在差异。（hypothesis test）

**第二章：计量资料的统计方法选择：**