**第四讲：R语言的基础运算**

R的赋值

赋值可理解为命名。Ex：将一列数据录入进R，对这一列的数据命名就是赋值

运算符号：<- (R也支持=作为运算符，但偶尔报错，不推荐)

命名规则：名字可以包括英文字母、下划线、数字、英文环境下的点。

下划线和数字不能做开头。点可以开头但后面不能跟数字

用字母名字最稳妥

**第五讲：R语言的数据类型介绍**

R的数据类型

R最常处理的数据结构是：向量、数据框、矩阵、列表

* 向量是一系列相同元素的组合（age，bmi）。分为：数值型向量、字符型向量、逻辑型向量。可以简单理解为数据分析里面的变量，数值型向量为连续型变量（年龄、身高、体重），字符型变量就是名义变量（性别、婚姻状态等）文本, 信件

  描述已自动生成
* 数据框是若干个向量组合在一起（形成一个平面/表格）。常见的调研数据都属于数据框。常见的excel、SPSS格式的数据都是数据框。

表格

描述已自动生成表格

中度可信度描述已自动生成

* 矩阵：Matrix，形势比较特殊的数据框，改列名为colnames，行名为rownames
* 列表：列表的包容性很强，是一个大杂烩，把各种向量和数据框等等收纳整理

**第六讲：R软件工作空间的读取与保存**

简单来说，就是对R软件中，数据运算结果的保存，第二次打开即是读取

需保存三份文件，读取时需一起读取

R的三大内容：

1. 代码编写板块（R Script），保存的文件格式为：.R
2. 工作空间（Environment），也被称为工作环境；它存储着用户录入、导入或者定义的对象（向量、数据框、列表等），即R语言的数据环境，保存的文件格式为：.RData
3. 历史板块（History）即R语言的运算历史，保存的文件格式为：.Rhistory

R的工作目录

R的工作目录，是R用来读取数据、读取文件并保存结果的文件夹

即，我们录入进R的数据、编写的R的代码，需要保存在文件夹中，以便下次直接读取调用，这个保存数据文件的文件夹就是R的工作目录。

当我们安装好R软件之后，会有一个默认的文件夹，我们可以使用函数getwd()来查看默认的工作文件夹。

一般而言，R默认的工作目录路径会很深，会增加我们的工作量，因此，我们一般会自己设定方便自己工作的工作目录。

**第七讲：R语言的数据录入方法**

R的数据录入

无论是学术界还是企业界，我们通常会面对多种格式的数据。例如：Excel格式数据、spss格式数据、txt格式数据等等。90%以上数据是Excel电子表格数据。

将需要分析的数据正确导入进R软件是数据分析的第一步。Excel、spss等多种格式数据可以直接导入进R，此外，也可以通过手工录入的方式将数据逐个填写R软件中。

1. 键盘手工录入

第十一讲：数据管理—变量计算、重新编码

数据的预处理

第十二讲：r语言ifelse判断语言

上节课把患者按照bmi分成四组

偏瘦，小于18.5

正常，大于等于18.5-24

偏胖，大于等于24-28

肥胖，大于等于28

第十四讲：r语言for循环运算结果的导出

主要讲解for循环如何导出成Excel或csv

**第十八课：描述性统计分析**

什么是描述性统计分析

R是一个统计分析软件，是为数据的统计分析而生的软件

无论什么数据，对其进行统计分析的第一步就是描述性统计，描述性统计分析的本质就是描述一列数据样子（外形、轮廓等等）。类似于，作家在深入刻画一个人物性格之前，都会清晰描绘出这个人的外貌特征

作家会用不同的词汇描述男性和女性的外貌，同样，描述性统计分析对于不同类型的变量，也会采用不同的描述方法

数据变量类型：

变量类型

连续数值型变量：一服从正态分布—均值填补缺失值

二不服从正态分布---中位数填补缺失值

分类型变量：无序分类—众数填补缺失值

有序分类—众数填补缺失值

1. 连续数值型变量

定义：可以取到数轴上某一区间的任意值

举例：年龄、身高、体重等变量就是连续数值型变量。因为这几个变量可以取到某一区间内任意值。

技巧：一般而言，由单位的变量就是连续数值型变量。年龄变量，xx（岁）；身高变量，xxx（厘米/米）；体重变量xx（kg）。

**统计量描述**：反映连续数值型变量的统计量有平均值、标准差、方差、中位数、四分位数、最大值、最小值等

**在R语言描述统计操作**

1. summary（）函数

r自带函数不用安装程序包

1. describe（函数）

install.packages(“psych”)

library(psych)

describe ()

**正态分布**

正态分布在统计学中是最基本、最重要的一种分布。正态分布是针对连续型变量的一种分布，分类型变量和正态分布没有关系。连续数值型变量可能服从正态分布，也可能不服从正态分布。如果服从，那么均值和标准差能比较好地反映该变量集中趋势和波动状况；如果不服从，则是四分位数能比较好地反应集中趋势和波动状况。

**如何判断正态分布**

**R语言正态性检验操作**

1. **样本量<2000:Shapiro.test（）#p大于0.05表示服从正态分布**
2. **样本量大于等于2000：ks.test（x，“pnorm”） #p大于0.05表示服从正态分布**

**注：ks.test（）检验如果出现如下警告：**

**Warning message：**

**In ks.test default(w$test,“pnorm”):**

**Kolmogorov – smirnov检验里不应该有连接**

**这是因为，k-s检验是基于连续累计分布函数的，而连续累计分布中出现相同值的概率为0.所以变量中如果出现了两个一样的数字，就会出现这样的警告，不用在意。**

**如何用R绘制直方图，并添加正态曲线：**

**Hist(w$bmi, col=“red”,freq=false)**

**Curve(dnorm(x,mean=mean(w$bmi),sd=sd(w$bmi)),add=t,col=‘blue’)**

**#freq=true表示纵坐标是频数。频数直方图**

**#freq=false表示纵坐标是密度。频率直方图**

**#只有频率（密度）直方图才能添加正态曲线。**

**2、分类型变量**

定义：分类变量的数据是属性和类别

举例：颜色、性别、婚姻状况、学历等变量就是分类型变量。颜色变量下的数据分为绿色、蓝色、红色等等。

性别变量下的数据分为男性女性、婚姻状况变量下有已婚、未婚、离异等。学历变量下数据分为大专、本科、硕士、博士等等。

技巧：一般而言，没有单位的变量都是分类型变量

统计量：中暑、人数、占比

分类型变量：无序分类变量：变量下的各个属性不分顺序，相互平等。比如：性别

有序分类变量：变量下各个属性有顺序，递增或递减，比如学历

**第十九讲：交叉表卡方**

卡方检验：比较组间率的差异

交叉表卡方检验：两个或多个构成比（率）的差异检验。例如：中西医结合治疗组的有效率与单纯西医治疗组的有效率之间有无差异

1. 适用于分类型变量的统计学方法
2. 被检验的分类变量一定是无序分类变量；分组变量可以是有序也可以是无序

案例：

比较男性的糖尿病患病率与女性的糖尿病患病率有无显著差异

解析：

以上案例中，分组变量为性别，分为两组：男性组和女性组；被检验变量为糖尿病，分为两类：有和无；被检验变量是无序二分类变量，因此选择交叉表卡方检验比较患病率的差异。

重点：

1. n≥40，所有T≥5,选择Pearson卡方检验（非连续校正）
2. n≥至少一个单元格1≤T<5，选择连续校正卡方检验
3. n<40或T<1，选择fisher精确检验（该检验不属于卡方检验，但可以作为卡方检验在应用上的补充）

**第二十一讲：相关性分析**

相关性分析

相关性分析：探寻两个变量之间关联关系的分析方法

注意：仅仅只有连续性变量和有序分类变量才有相关性这个概念，如果涉及到无序分类变量，就需要通过差异分析来间接反映相关。具体情况如下：

比如学习时间越长是否考试成绩越好

**23讲：两独立样本非参数检验**

独立样本非参数检验：

比较两组在连续变量上的差异时，两组均服从正态分布，选择独立样本T检验；如果其中任意一组或者两组非正态，则需要选择独立样本非参数检验文本

描述已自动生成

两独立样本非参数检验：

比较有代谢综合征和无代谢综合征组在HDL上有无显著差异。

解析：

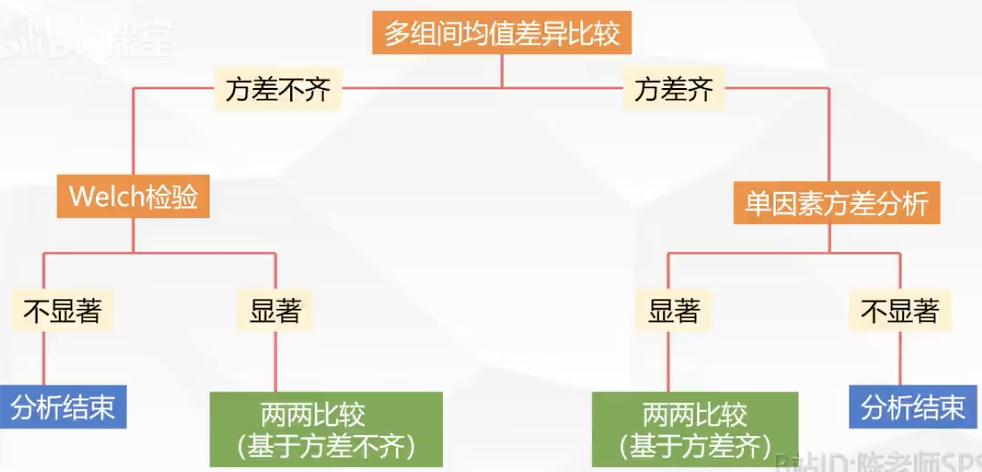
以上案例中，分组为两组，被检验变量是HDL，如果hdl服从正态分布，那么比较两组的hdl差异选择独立样本t检验；如果其中任意一组非正态分布，则选择两独立样本非参数检验。

**24讲：单因素方差分析（one-way ANOVA）**

**独立样本T检验：**比较两组均值差异

**单因素方差分析**:比较三组或三组以上均值的差异

**重点：只要是比较均值差异的统计方法，都要求数据服从的正态分布**

****

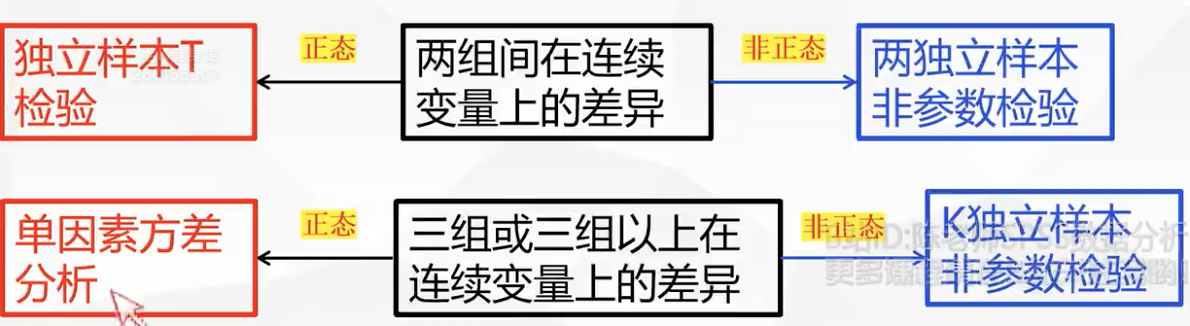
**单因素方差分析案例：**

比较A组（溃疡样型）、B组 （动力障碍样型）、C组（非特异性型）这三组的胃半排空时间有无显著差异。

解析：以上案例中，分为三组，被检验变量是胃半排空时间，如果胃半排空时间服从正态分布，那么比较三组的胃半排空时间差异选择单因素方差分析；如果不服从，则选择独立样本非参数检验

**25讲：k独立样本非参数检验**

独立样本非参数检验：比较两组或多组在连续变量上的差异时，如果数据不服从正态分布，选择非参数检验。如果数据服从正态分布，选择参数检验。



例子：k独立样本非参数检验

比较A组（溃疡样型）、B组（动力障碍样型）、C组（非特异性型）这三组的SS有无显著差异。

解析：

以上案例中，分组为三组，被检验变量是SS，如果SS服从正态分布，那么比较三组的SS差异选择单因素方差分析；如果不服从，则选择K独立样本非参数检验。

**26讲：秩和检验**

组间 有序变量差异比较时 使用秩和检验

日程表

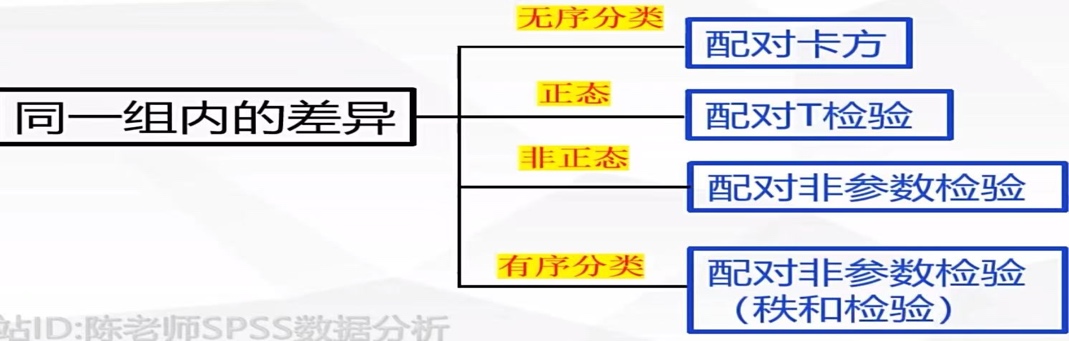
中度可信度描述已自动生成

案例：

1. 比较某乡镇，男性居民和女性居民的学历等级有无显著差异
2. 比较某乡镇，不同婚姻状况（未婚、已婚、离异）居民的学历等级有无显著差异

**27讲：配对样本差异比较**

配对样本：同一批样本前后两个状态；或者，同一批样本在两个环境下的状态。



有序变量的组间差异，也是秩和检验，独立样本非参数检验

所有参数、非参数都有独立样本和配对样本检验，在组间进行比较的用独立样本，在同一组内，前后做比较的用配对样本

配对样本t检验：如果样本的两个状态数据均为连续变量且服从正态分布，则选择配对样本t检验比较前后的差异是否显著，数据表达形式为均值±标准差。

举例：25名糖尿病患者的空腹血糖在治疗前和治疗后有无显著变化

解析：以上案例数据是25名患者前后两个状态的空腹血糖数据，空腹血糖为连续数值型变量，经正态性检验得出前后的空腹血糖均服从正态分布，因此选择配对样本t检验比较前后的差异是否显著。如果治疗前或治疗后出现非正态，则选择配对非参。

**28讲：配对样本非参数检验**

配对样本：同一批样本前后两个状态；或者，同一批样本在两个环境下的状态。

配对样本非参数检验：如果样本的两个状态数据均为连续变量但不服从正态分布，则选择配对样本非参数检验比较前后的差异是否显著，数据的表达形式为中位数（下四分位数～上四分位数）。

举例：43名患者疼痛评分，在治疗前和治疗后有没有发生显著变化

解析：以上案例的数据是43名患者前后两个状态的疼痛评分数据，疼痛评分为连续数值型变量，经正态性检验得出前后疼痛评分均不服从正态分布，因此选择配对样本非参数检验比较前后的差异是否显著。

**29讲：k个相关样本非参数检验**

K个相关样本：也是配对样本。

配对样本非参数检验：同一批样本前后两个状态

K个相关样本非参数检验：同一批样本多个（>2）状态。

K个相关样本非参数检验包含：1、friedman检验。2、kendall‘s W检验。3、Cochran‘s Q检验（）

在数据类型、统计目标上有一定区别

日程表

描述已自动生成

**第三十讲：**

线性回归的统计意义：

线性回归的统计目的：研究变量之间的因果关系。更进一步，一个或者多个原因变量对某一个结局变量的影响情况。

结局变量：也被称为因变量dv，必须是连续数值型变量。

原因变量：也被称作自变量iv，可以是连续型变量，也可以是分类变量。

举例：某公司通过问卷搜集了63名员工的工作满意度、环境满意度、薪资满意度、工作内容满意度、工作能力数据，现在需要用这些数据探寻环境满意度、薪资满意度、工作内容满意度、工作能力是否会显著影响员工的工作满意度。

解析：因变量：工作满意度（连续数值型变量）

自变量：环境满意度、薪资满意度、工作内容满意度、工作能力

**32讲**

案例背景：根据理论和实践经验，怀疑性别、年龄、维生素d、igg、igm、iga、糖尿病、冠心病、高血压这几个因素可能是术后感染的影响因素。因此，搜集295名患者的性别、年龄、维生素d、igg、igm、iga、糖尿病、冠心病、高血压以及是否出现术后感染的数据，通过对数据进行单因素分析和多因素分析，最终确定术后感染的独立影响因素。

统计方法说明：

探寻一个事件发生的影响因素，对应的统计学方法是单因素分析➕多因素分析。

1. 单因素分析

单因素分析，就是分别将每一个变量和术后感染结局做研究。具体方法是：比较术后感染患者和术后未感染患者在每一个变量上是否存在显著差异。如果存在显著差异，那么这个变量则是导致感染的可疑因素，保留，继续参与后续的多因素分析。如果不存在显著差异，则意味着感染和未感染两患者在这个因素上无差异，那么可以肯定这个因素一定不会是感染的影响因素，直接剔除。（剔除的好处排除混杂干扰，保证准确性）

比较两组的差异的统计学方法有三类，一类是针对被检验变量是连续型的方法；一类是针对被检验变量是有序分类型的方法；还有一类是针对被检验变量是无需类型的方法。

如果被检验变量是连续型：如果连续被检验变量服从正态分布，则选择独立样本t检验标两组的差异，数据的表达形式为均值±标准差；如果连续被检验变量不服从正态分布，则选择两独立样本非参数检验比较两组的差异，数据的表达形式为中位数（下四分位数～上四分位数）

对于被检验变量是有序分类变量：选择两独立样本非参数检验比较两组的差异，数据表达形式为人数（占比）对于被检验变量为无序分类变量：选择交叉表卡方检验比较两组的差异，数据表达形式为人数（占比）

2、多因素分析

基于以上单因素筛查出来显著的指标，需要继续保留，继续参与多因素研究，多因素分析得出显著的指标则是术后感染的独立影响因素。多因素分析选择的统计学方法是二元logistic回归