多元统计分析陈钰芬课后答案

1、在数据处理时，为什么通常要进行标准化处理？

数据的标准化是将数据按比例缩放，使之落入一个小的特定区间。在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。其中最典型的就是0-1 标准化和 Z 标准化。

2、欧氏距离与马氏距离的优缺点是什么？

欧氏距离也称欧几里得度量、欧几里得度量，是一个通常采用的距离定义，它是在 m 维空间中两个点之间的真实距离。在二维和三维空间中的欧氏距离的就是两点之间的距离。

缺点：就大部分统计问题而言，欧氏距离是不能令人满意的。每个坐标对欧氏距离的贡献是同等的。当坐标表示测量值时，它们往往带有大小不等的随机波动，在这种情况下，合理的方法是对坐标加权，使变化较大的坐标比变化较小的坐标有较小的权系数，这就产生了各种距离。当各个分量为不同性质的量时，“距离”的大小与指标的单位有关。它将样品的不同属性之间的差别等同看待，这一点有时不能满足实际要求。没有考虑到总体变异对距离远近的影响。

马氏距离表示数据的协方差距离。为两个服从同一分布并且其协方差矩阵为 Σ 的随机变量与的差异程度:如果协方差矩阵为单位矩阵,那么马氏距离就简化为欧氏距离,如果协方差矩阵为对角阵,则其也可称为正规化的欧氏距离。

优点：它不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关。由标准化数据和中心化数据计算出的二点之间的马氏距离相同。马氏距离还可以排除变量之间的相关性的干扰。

缺点：夸大了变化微小的变量的作用。受协方差矩阵不稳定的影响，马氏距离并不总是能顺利计算出。

3、当变量 X1 和 X2 方向上的变差相等，且与互相独立时，采用欧氏距离与统计距离是否一致？

统计距离区别于欧式距离，此距离要依赖样本的方差和协方差，能够体现各变量在变差大小上的不同，以及优势存在的相关性，还要求距离与各变量所用的单位无关。如果各变量之间相互独立,即观测变量的协方差矩阵是对角矩阵，则马氏距离就退化为用各个观测指标的标准差的倒数作为权数的加权欧氏距离。

1.、聚类分析的基本思想和功能是什么？

聚类分析的基本思想是研究的样品或指标之间存着程度不同的相似性，于是根据一批样品的多个观测指标，具体找出一些能够度量样品或指标之间的相似程度的统计量，以这些统计量作为划分类型的依据，把一些相似程度较大的样品聚合为一类，把另外一些彼此之间相似程度较大的样品又聚合为另外一类，直到把所有的样品聚合完毕，形成一个有小到大的分类系统，最后再把整个分类系统画成一张分群图，用它把所有样品间的亲疏关系表示出来。功能是把相似的研究对象归类。

2、试述系统聚类法的原理和具体步骤。

系统聚类是将每个样品分成若干类的方法，其基本思想是先将各个样品各看成一类，然后规定类与类之间的距离，选择距离最小的一对合并成新的一类，计算新类与其他类之间的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的样品合为一类为止。

具体步骤：

1、对数据进行变换处理；（不是必须的，当数量级相差很大或指标变量具有不同单位时是必要的）

2、构造 n 个类，每个类只包含一个样本；

3、计算 n 个样本两两间的距离 ijd;

4、合并距离最近的两类为一新类；

5、计算新类与当前各类的距离，若类的个数等于 1，转到 6；否则回 4;

6、画聚类图；

7、决定类的个数，从而得出分类结果。

3、试述 K-均值聚类的方法原理。

K-均值法是一种非谱系聚类法，把每个样品聚集到其最近形心（均值）类中，它是把样品聚集成 K 个类的集合，类的个数 k 可以预先给定或者在聚类过程中确定，该方法应用于比系统聚类法大得多的数据组。步骤是把样品分为 K 个初始类，进行修改，逐个分派样品到期最近均值的类中（通常采用标准化数据或非标准化数据计算欧氏距离）重新计算接受新样品的类和失去样品的类的形心。重复这一步直到各类无元素进出。

4、试述模糊聚类的思想方法。

模糊聚类分析是根据客观事物间的特征、亲疏程度、相似性，通过建立模糊相似关系对客观事物进行聚类的分析方法，实质是根据研究对象本身的属性构造模糊矩阵，在此基础上根据一定的隶属度来确定其分类关系。基本思想是要把需要识别的事物与模板进行模糊比较，从而得到所属的类别。简单地说，模糊聚类事先不知道具体的分类类别，而模糊识别是在已知分类的情况下进行的。模糊聚类分析广泛应用在气象预报、地质、农业、林业等方面。它有两种基本方法:系统聚类法和逐步聚类法。该方法多用于定性变量的分类。

5、略

1、应用判别分析应该具备什么样的条件？

答：判别分析最基本的要求是，分组类型在两组以上，每组案例的规模必须至少在一个以上，解释变量必须是可测量的，才能够计算其平均值和方差。

对于判别分析有三个假设：

（1）每一个判别变量不能是其他判别变量的线性组合。有时一个判别变量与另外的判别变量高度相关，或与其的线性组合高度相关，也就是多重共线性。

（2）各组变量的协方差矩阵相等。判别分析最简单和最常用的的形式是采用现行判别函数，他们是判别变量的简单线性组合，在各组协方差矩阵相等的假设条件下，可以使用很简单的公式来计算判别函数和进行显着性检验。

（3）各判别变量之间具有多元正态分布，即每个变量对于所有其他变量的固定值有正态分布，在这种条件下可以精确计算显着性检验值和分组归属的概率。

2、试述贝叶斯判别法的思路。

答：贝叶斯判别法的思路是先假定对研究的对象已有一定的认识，常用先验概率分布来描述这种认识，然后我们取得一个样本，用样本来修正已有的认识（先验概率分布），得到后验概率分布，各种统计推断都通过后验概率分布来进行。将贝叶斯判别方法用于判别分析，就得到贝叶斯判别。

3、试述费歇判别法的基本思想。

答：费歇判别法的基本思想是将高维数据点投影到低维空间上来，然而利用方差分析的思想选出一个最优的投影方向。因此，严格的说费歇判别分析本身不是一种判别方法，只是利用费歇统计量进行数据预处理的方法，以使更有利于用判别分析方法解决问题。为了有利于判别，我们选择投影方向 a 应使投影后的 k 个一元总体能尽量分开（同一总体中的样品的投影值尽量靠近）。k 要做到

这一点，只要投影后的 k 个一元总体均值有显着差异，即可利用方差分析的方法使组间平方和尽可能的大。则选取投影方向 a 使Δ(a)达极大即可。

4、什么是逐步判别分析？

答：具有筛选变量能力的判别方法称为逐步判别分析法。逐步判别分析法就是先从所有因子中挑选一个具有最显着判别能力的因子，然后再挑选第二个因子，这因子是在第一因子的基础上具有最显着判别能力的因子，即第一个和第二个因子联合起来有显着判别能力的因子；接着挑选第三个因子，这因子是在第一、第二因子的基础上具有最显着判别能力的因子。由于因子之间的相互关系，当引进了新的因子之后，会使原来已引入的因子失去显着判别能力。因此，在引入第三个因子之后就要先检验已经引入的因子是否还具有显着判别能力，如果有就要剔除这个不显着的因子；接着再继续引入，直到再没有显着能力的因子可剔除为止，最后利用已选中的变量建立判别函数。

5、简要叙述判别分析的步骤及流程

答：（1）研究问题：选择对象，评估一个多元问题各组的差异，将观测个体归类，确定组与组之间的判别函数。

（2）设计要点：选择解释变量，样本量的考虑，建立分析样本的保留样本。

（3）假定：解释变量的正态性，线性关系，解释变量间不存在多重共线性，协方差阵相等。

（4）估计判别函数：联立估计或逐步估计，判别函数的显着性。

（5）使用分类矩阵评估预测的精度：确定最优临界得分，确定准则来评估判对比率，预测精确的统计显着性。

（6）判别函数的解释：需要多少个函数。评价单个函数主要从判别权重、判别载荷、偏 F 值几个方面；评价两个以上的判别函数，分为评价判别的函数和评价合并的函数。

（7）判别结果的验证：分开样本或交叉验证，刻画组间的差异。

6、略

1、主成分的基本思想是什么？

在对某一事物进行实证研究时，为更全面、准确地反映事物的特征及其发展规律，往往考虑与其有关的多个指标，在多元统计中也称为变量。一方避免遗漏

重要信息而考虑尽可能多的指标看，另一方面考虑指标的增多，又难以避免信息重叠。希望涉及的变量少，而得到的信息量有较多。

主成分的基本思想是研究如何通过原来的少数几个线性组合来解释原来变量绝大多数信息的一种多元统计方法。研究某一问题涉及的众多变量之间有一定的相关性，必然存在着支配作用的公共因素。通过对原始变量相关矩阵或协方差矩阵内部结构关系的研究，利用原始变量的线性组合形成几个无关的综合指标（主成分）来代替原来的指标。通常数学上的处理就是将原来 P 个指标作线性组合，作为新的综合指标。最经典的做法就是用 F1（选取的第一个线性组合，即第一个综合指标）的方差来表达，即 Var（F1）越大，表示 F1 包含的信息越多。因此在所有的线性组合中选取的 F1 应该是方差最大的，故称 F1 为第一主成分，如果第一主成分不足以代表原来 P 个指标的信息，再考虑选取 F2 即选第二个线性组合，为了有效地反映原来信息，F1 已有的信息就不需要再出现在 F2 中，用数学语言表达就是要求 Cov（F1，F2）=0 则称 F2 为第二主成分，依此类推可以构造出第三、第四·····，第 P 个主成分。

2、主成分在应用中的主要作用是什么？

作用：利用原始变量的线性组合形成几个综合指标（主成分），在保留原始变量主要信息的前提下起到降维与简化问题的作用，使得在研究复杂问题时更容易抓住主要矛盾。通过主成分分析，可以从事物之间错综复杂的关系中找出一些主要成分，从而能有效利用大量数据进行定量分析，解释变量之间的内在关系，得到对事物特征及其发展规律的一些深层次的启发，把研究工作引向深入。主成分分析能降低所研究的数据空间的维数，有时可通过因子载荷 aij 的结论，弄清 X 变量间的某些关系，多维数据的一种图形表示方法，用主成分分析筛选变量，可以用较少的计算量来选择，获得选择最佳变量子集合的效果。

3.由协方差阵出发和由相关阵出发求主成分有什么不同？

（1）由协方差阵出发

设随即向量 X=（X1，X2，X3,……Xp）’ 的协方差矩阵为Σ，?1≥?2≥……≥?p 为Σ的特征值，γ1，γ2，……γp 为矩阵 A 各特征值对应的标准正交特征向量，则第 i 个主成分为 Yi=γ1i*X1+γ2i*X2+……+γpi*Xp,i=1,2,……,p 此时 VAR（Yi）=?i，ＣＯＶ（Ｙｉ，Ｙｊ）＝Ｏ，ｉ≠ｊ

我们把 X1，X2，X3,……Xp 的协方差矩阵Σ的非零特征根?1≥?2≥……≥?p＞0 向量对应的标准化特征向量γ1，γ2，……γp 分别作为系数向量，Y1=γ1’*X，Y2=γ2’*X,……，Yp=γp’*X 分别称为随即向量 X 的第一主成分，第二主成分……第 p 主成分。Y 的分量 Y1，Y2，……，Yp 依次是 X 的第一主成分、第二主成分……第 p 主成分的充分必要条件是：（1）Y=P’*X，即 P 为 p 阶正交阵，（2）Y 的分量之间互不相关，即 D（Y）=diag(?1,?2,……,?p),(3)Y 的 p 个分量是按方差由大到小排列，即?1≥?2≥……≥?p。

（2）由相关阵出发

对原始变量 X 进行标准化，

Z=(Σ^1/2)^-1*(X-μ） cov（Z）=R

原始变量的相关矩阵实际上就是对原始变量标准化后的协方差矩阵，因此，有相关矩阵求主成分的过程与主成分个数的确定准则实际上是与由协方差矩阵出发求主成分的过程与主成分个数的确定准则相一致的。λi，γi 分别表示相关阵 R 的特征根值与对应的标准正交特征向量，此时，求得的主成分与原始变量的关系式为：

Yi=γi’*Z=γi’*(Σ^1/2)^-1*(X-μ）

在实际研究中,有时单个指标的方差对研究目的起关键作用,为了达到研究目的,此时用协方差矩阵进行主成分分析恰到好处。有些数据涉及到指标的不同度量尺度使指标方差之间不具有可比性,对于这类数据用协方差矩阵进行主成分分析也有不妥。相关系数矩阵计算主成分其优势效应仅体现在相关性大、相关指标数多的一类指标上。避免单个指标方差对主成分分析产生的负面影响，自然会想到把单个指标的方差从协方差矩阵中剥离，而相关系数矩阵恰好能达到此目的。

4、略

1、因子分析与主成分分析有什么本质不同？

答：（1）因子分析把诸多变量看成由对每一个变量都有作用的一些公共因子和一些仅对某一个变量有作用的特殊因子线性组合而成，因此，我们的目的就是要从数据中探查能对变量起解释作用的公共因子和特殊因子，以及公共因子和特殊因子的线性组合。主成分分析则简单一些，它只是从空间生成的角度寻找能解释诸多变量绝大部分变异的几组彼此不相关的新变量

（2）因子分析中，把变量表示成各因子的线性组合，而主成分分析中，把主成分表示成各变量的线性组合

（3）主成分分析中不需要有一些专门假设，因子分析则需要一些假设，因子分析的假设包括：各个因子之间不相关，特殊因子之间不相关，公共因子和特殊因子之间不相关。

（4）在因子分析中，提取主因子的方法不仅有主成分法，还有极大似然法等，基于这些不同算法得到的结果一般也不同。而主成分分析只能用主成分法提取。

（5）主成分分析中，当给定的协方差矩阵或者相关矩阵的特征根唯一时，主成分一般是固定；而因子分析中，因子不是固定的，可以旋转得到不同的因子。

（6）在因子分析中，因子个数需要分析者指定，结果随指定的因子数不同而不同。在主成分分析中，主成分的数量是一定的，一般有几个变量就有几个主成分。 （7）与主成分分析相比，由于因子分析可以使用旋转技术帮助解释因子，在解释方面更加有优势。而如果想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进行后续的分析，则可以使用主成分分析。

2、因子载荷a_{ij}的统计定义是什么？它在实际问题的分析中的作用是什么？

答：（1）因子载荷a_{ij}的统计定义：是原始变量X_i与公共因子F_j的协方差，X_i与$F_j(i=1,2,...,p;j=1,2,...,m)$都是均值为 0，方差为 1 的变量，因此$a_{ij}$同时也是$X_i$与$F_j$的相关系数。

（2）记$g_j^2=a_{1j}^2+a_{2j}^2+...+a_{pj}^2(j=1,2,...,m),$则$g_j^2$表示的是公共因子$F_j$对于$X$的每一分量$X_i(i=1,2,...,p)$所提供的方差的总和，称为公共因子$F_j$对原始变量$X$的方贡献，它是衡量公共因子相对重要性的指标。$g_j^2$越大，表明公共因子$F_j$对$X_i$的贡献越大，或者说对$X$的影响作用就越大。如果因子载荷矩阵对$A$的所有的$g_j^2(j=1,2,...,m)$都计算出来，并按大小排序，就可以依此提炼出最有影响的公共因子。

3、略

1、试述对应分析的思想方法及特点？

思想：对应分析又称为相应分析，也称 R—Q 分析。是因子分子基础发展起来的一种多元统计分析方法。它主要通过分析定性变量构成的列联表来揭示变量之间的关系。当我们对同一观测数据施加 R 和 Q 型因子分析，并分别保留两个公共因子，则是对应分析的初步。?对应分析的基本思想是将一个联列表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。它最大特点是能把众多的样品和众多的变量同时作到同一张图解上，将样品的大类及其属性在图上直观而又明了地表示出来，具有直观性。另外，它还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程，可以从因子载荷图上对样品进行直观的分类，而且能够指示分类的主要参数（主因子）以及分类的依据，是一种直观、简单、方便的多元统计方法。?

特点：对应分析的基本思想是将一个联列表的行和列中各元素的比例结构以点的形式在较低维的空间中表示出来。它最大特点是能把众多的样品和众多的变量同时作到同一张图解上，将样品的大类及其属性在图上直观而又明了地表示出来，具有直观性。另外，它还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程，可以从因子载荷图上对样品进行直观的分类，而且能够指示分类的主要参数（主因子）以及分类的依据，是一种直观、简单、方便的多元统计方法。

2、试述对应分析中总惯量的意义。?

总惯量不仅反映了行剖面集定义的各点与其重心加权距离的总和,同时与x^2统计量仅相差一个常数,而x^2统计量反映了列联表横联与纵联的相关关系,因此总惯量也反映了两个属性变量各状态之间的相关关系。对应分析就是在对总惯量信息损失最小的前提下，简化数据结构以反映两属性变量之间的相关关系。

1、试述典型相关分析的统计思想及该方法在研究实际问题中的作用。

答： 典型相关分析是研究两组变量之间相关关系的一种多元统计方法。用于揭示两组变 量之间的内在联系。典型相关分析的目的是识别并量化两组变量之间的联系。将两组变量相 关关系的分析转化为一组变量的线性组合与另一组变量线性组合之间的相关关系。

基本思想：

（1）在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数。

（2）选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对。

（3）如此继续下去，直到两组变量之间的相关性被提取完毕为此。

其作用为：

进行两组变量之间的相关性分析，用典型相关系数衡量两组变量之间的相关性。

2、简述典型相关分析中冗余分析的内容及作用。

答：典型型冗余分析的作用即分析每组变量提取出的典型变量所能解释的该组样本总方差的比 例，从而定量测度典型变量所包含的原始信息量。

第一组变量样本的总方差为 tr(R11 ） ?p，第二组变量样本的总方差为 tr(R22 ）?q。

$\hat{A}_z^*$ 和 $\hat{B}_z^*$ 是样本典型相关系数矩阵，典型系数向量是矩阵的行向量，

$\hat{U}_z^* = \hat{A}_z^* Z$，$\hat{V}_z^* = \hat{B}_z^* Z$

前 r 对典型变量对样本总方差的贡献为

则第一组样本方差由前 r 个典型变量解释的比例为：$$Rd_{z^{(1)}|\hat{U}} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{p} r^2_{z_k^{(1)},\hat{U}_i}}{p}$$

$$Rd_{z^{(2)}|\hat{V}} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{q} r^2_{z_k^{(2)},\hat{V}_i}}{q}$$

第二组样本方差由前 r 个典型变量解释的比例为：

3、典型变量的解释有什么具体方法？实际意义是什么？

答：主要使用三种方法：（1）典型权重（标准相关系数）：传统的解释典型函数的方法包括观察每个原始变量在它的典型变量中的典型权重，即标准化相关系数（Standardized Canonical Coefficients）的符号和大小。有较大的典型权重，则说明原始变量对它的典型变量的贡献较大，反之则相反。原始变量的典型权重有相反的符号说明变量之间存在一种反面关系，反之则有正面关系。但是这种解释遭到了很多批评。这些问题说明在解释典型相关的时候应慎用典型权重。

（2）典型载荷（结构系数）：由于典型载荷逐步成为解释典型相关分析结果的基础。典型载荷分析，即典型结构分析（Canonical Structure Analyse），是原始变量（自变量或者因变量）与它的典型变量间的简单线性相关系数。典型载荷反映原始变量与典型变量的共同方差，它的解释类似于因子载荷，就是每个原始变量对典型函数的相对贡献。

（3）典型交叉载荷（交叉结构系数）：它的提出时作为典型载荷的替代，也属于典型结构分析。计算典型交叉载荷包括每个原始因变量与自变量典型变量直接相关，反之亦然。交叉载荷提供了一个更直接地测量因变量组与自变量组之间的关系的指标。

实际意义：即使典型相关系数在统计上是显着的，典型根和冗余系数大小也是可接受的，研究者仍需对结果做大量的解释。这些解释包括研究典型函数中原始变量的相对重要性。

4.、略