对于数据1，2，3，4，5，6，7，8，9，可以直观看出其均值、中位数、三均值分别为（ B ）。

A. 6、5、5 B. 5、5、5 C. 1、3、7 D. 5、1、9

已知甲班学生“统计学”的平均成绩为86分，标准差是12.8分，乙班学生“统计学”的平均成绩是90分，标准差是10.3分，下列表述正确的是（ A ）

A. 乙班平均成绩的代表性高于甲班

B. 甲班平均成绩的代表性高于乙班

C. 甲、乙两班平均成绩的代表性相同

D. 甲、乙两班平均成绩的代表性无法比较

描述一组对称（或正态）分布资料的离散趋势时，下列中最适宜选择的指标是（B）

A.极差 B.标准差 C.均数 D.变异系数

在我们进行数据分析的过程中，时常引入显著性水平，它是估计总体参数落在某一区间内，可能犯错误的[概率](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87/828845?fromModule=lemma_inlink)，并常用α表示，那么α一般取值为（D）

A.20% B.50% C.80% D.5%

在线性回归分析中，线性回归模型（一维）常被简单表示为（ C ）

A、y=a+b

B、y=na

C、y=a+bx

D、y=na+bx

当你用跑步时间、年龄、跑步时脉搏以及最高脉搏作为预测变量来对耗氧量进行回归时，年龄的参数估计是-2.78， 这意味着什么？（其余预测变量不改变）（ B ）

A、年龄每增加一岁，耗氧量就增大2.78

B、年龄每增加一岁，耗氧量就降低2.78

C、年龄每增加2.78岁，耗氧量就翻倍

D、年龄每减少2.78岁，耗氧量就翻倍

R2=SSR/SST，取值范围在[0，1]，反映回归曲线的拟合优度，当R2趋近于1，则回归曲线拟合优度（ A ）

A．越好       B. 越差       C. 适中       D. 以上都不对

对于数据1，2，3，5，5，5，5，7，8，9，可以直观看出其均值、中位数、三均值分别为（ B ）。

A. 6、5、5 B. 5、5、5 C. 1、3、7 D. 5、1、9

在本学期的学习中我们所提及并用到来进行检验的分布方式有（D）  
a. T分布  
b. F分布  
c. 卡方分布 (χ2分布)  
d. 以上都有使用

已知A班学生“线性代数”的平均成绩为88分，标准差是11.8分，B班学生“线性代数”的平均成绩是88分，标准差是7.8分，下列表述正确的是（ A ）

A. B班平均成绩的代表性高于A班

B. A班平均成绩的代表性高于B班

C. A、B两班平均成绩的代表性相同

D. A、B两班平均成绩的代表性无法比较

三均值的计算公式（A）。

A.

B.

C.

D.

由于在数据分析中时常出现异常数据，下列中具有对异常数据的较强抗扰性的指标是（B）

A.方差 B.四分位极差 C.均数 D.标准差

有一组数据的众数<中位数<均值，请问这组数据的分布（B）  
A.左偏  
B.右偏  
C.对称  
D.以上都不对

复相关系数在方差分析中用于衡量线性回归模型的拟合优度，复相关系数越小，则说明线性拟合优度（ B ）

A．越好       B. 越差       C. 适中       D. 以上都不对

假设有一个坐标系*xOy*，然后将原坐标系*xOy*逆时针旋转一个角度θ得到新的坐标系*mOn*，那么现在有一点A，它在坐标系*xOy*下的坐标为，那么请计算它在新的坐标系*mOn*下的坐标为（B）

A.  
B.  
C.  
D.

方差分析中的回归平方和，误差平方和，总离差平方和，三者的自由度分别为（ B ）。

A.a-1 , n-1 , n-a-1

B.a-1 , n-a , n-1

C.a-1 , n-a , n-a

D.a-1 , n-1 , n-a

方差的算术平方根称为（ D ）。

A.均值

B.极差

C.峰度

D.标准差

在聚类分析中常用的距离有（ D ）6.

A.绝对距离 B.马氏距离 C.欧氏距离 D.以上都是

单因素方差分析中，任何时候都成立的是（ C ）

A.SSE<SSA

B.SSA<SSE

C.SST=SSE+SSA

D.MST=MSE+MSA

F统计量的计算方法是( A )

A. 组间均方差和组内均方差的比值

B. 组间均方差和组内均方差的积

C. 组间均方差和组内均方差的和

D. 组间均方差和组内均方差的差

线性回归模型y=bx+a+e ( a和b为模型的未知参数)中，e称为（ B ）。

A.回归系数

B.随机误差

C.响应变量

D.回归变量

在单因素方差分析中，使用F统计量作进行检验，其分子的自由度和分母的自由度分别是（ C ）。

A.n-a，a-1

B.n-1，n-a

C.a-1，n-a

D.n-a，n-1

（ B ）的主要目的就是对原变量加以“改造”，在不致损失原变量太多信息的条件下尽可能地降低原变量的维数。

A.方差分析

B.主成分分析

C.回归分析

D.聚类分析

主成分分析即构造原变量的一系列线性组合，使其（ D ）达到最大。

A.均值

B.维数

C.最小值

D.方差

样本X到一个总体G的马氏平方距离，是关于X的二次函数，所以我们通常称之为（ B ）

A.距离判别函数 B.二次判别函数 C.一次函数 D.线性判别函数

在Bayes判别中，只考虑两个P维的总体Q1，Q2，在无先验信息可用的情况下，两者的先验概率满足以下哪个选项（ C ）

A.p1 > p2 或者 p1 < p2

B.p1+p2=2

C.p1=p2=

D.以上均满足

当前有两个总体G1，G2,与一个样本X，想得知X属于哪个总体，则下列说法中错误的是（ B ）

A.在两个总体的协方差矩阵相等时，通过计算出线性判别函数,两者相较得出结论

B.在两个总体的协方差矩阵不相等时，通过计算出线性判别函数,通过值的正负得出结论

C.在两个总体的协方差矩阵相等时，通过计算出X到总体G1，G2的马氏距离，两距离相较得出结论

D.在两个总体的协方差矩阵不相等时，通过计算出X到总体G1，G2的马氏距离，两距离相较得出结论

关于马氏距离d满足距离的性质中，假设x,y,z是来自总体G的三个样品，则（D）  
A. d(x , y) ≥ 0，当且仅当x = y 时d(x , y) = 0  
B. d(x , y) = d(y , x)   
C. d(x , y) ≤ d(x , y) + d(y , x)  
D. 以上性质都正确

距离判别法所用的距离是( A )

A.马氏距离 B.欧氏距离 C.绝对值距离 D.欧氏平方距离

下列哪个公式是贝叶斯公式（ C ）

A.P(B|A)=P(B)P(A|B)/P(A)

B. P(B|A)=P(B)P(B|A)/P(A)

C. P(B|A)=P(A)P(A|B)

D. P(B|A)=P(A|B)

贝叶斯判别中一般将总体假设为什么分布（ A ）

A.正态分布 B.均匀分布 C.二项分布 D.泊松分布

表示分布形状的数字特征有（ C ）和 偏度。

A.均值

B.方差

C.峰度

D.中位数

在日常生活中，大部分事情的发生都是存在0，1的关系，要么发生要么不发生，此时我们便学习到Logistic回归模型，其中可将进行Logit函数变换,以下式子中符合logit函数变换的函数为（ A ）。

A.L(P) =

B.L(P) =

C.L(P) =

D.L(P) =

主成分分析可以利用（ A ）求解主成分。

A.相关系数矩阵

B.Fisher信息矩阵

C.Hessian矩阵

D.拉普拉斯矩阵

快速聚类是指（ D ）

A.两步聚类 B.系统聚类 C.层次聚类 D.k-means聚类

有一组数据的众数>中位数>均值，请问这组数据的分布（A）  
A. 左偏  
B. 右偏  
C. 对称  
D. 以上都不对

k-means算法中的K指的是（ B ）

A.K个样本 B.K个质心 C.K次迭代 D.K次方

表示分布形状的数字特征有峰度和（ C ）。

A.均值

B.方差

C.偏度

D.中位数

方差的算术平方根称为（ D ）。

A.均值

B.极差

C.峰度

D.标准差

（ A ）是秩相关系数。

A. Spearman相关系数

B.Pearson相关系数

C.以上两种都是

D.以上两种都不是

在比较两个模型的拟合效果时，甲、乙两个模型的相关指数R2的值分别约为0.96和0.85，则拟合较好的模型是（ A ）。

A.甲

B.乙

C.无法比较

D.两者一样

单因素方差分析中，任何时候都成立的是（ D ）

A.SSE<SSA

B.SSA<SSE

C.MST=MSE+MSA

D.SST = SSA+SSE

线性回归模型y=bx+a+e ( a和b为模型的未知参数)中，e称为（ C ）。

A.回归系数

B.响应变量

C.随机误差

D.回归变量

下列关于主成分分析中标准化变量的说明正确的是（B）

A.由于不同的量纲会引起各变量取值的分散程度差异过大，此时总方差主要受方差较小的变量控制  
B.为了消除原变量彼此方差差异过大的影响，通常将原变量进行标准化再做主成分分析  
C.标准化变量纯属多此一举，无需进行  
D.以上都不对

当你用跑步时间、年龄、跑步时脉搏以及最高脉搏作为预测变量来对耗氧量进行回归时，年龄的参数估计是-2.78， 这意味着什么？（其余预测变量不改变）（ B ）

A、年龄每增加一岁，耗氧量就增大2.78

B、年龄每增加一岁，耗氧量就降低2.78

C、年龄每增加2.78岁，耗氧量就翻倍

D、年龄每减少2.78岁，耗氧量就翻倍

在比较两个模型地拟合效果是否优良时，目前有两个模型甲、乙的相关指数分别为0.88与0.97，两者中拟合效果最好的模型是（C）

A.两者相同

B.甲

C.乙

D.两者无法比较

关于标准差与标准误，以下说法正确的是（D）

A.样本数增大时，样本差减小，标准差不变

B.可信区间大小与标准差有关，而参考值范围与标准误有关

C.样本数增大时，标准差与标准误均减小

D.总体标准差一定时，增大样本数会减小标准误

希望描述一群用户在某页面停留时长的集中趋势，最好采用（C）

A.均值

B.众数

C.中位数

D.均值和中位数

主成分分析的目的是（ C ）

A.对原变量加以“改造”，在不知损失原变量太多信息的条件下尽可能地降低原变量的维数

B.用为数较少的“新变量”代替原来的各变量

C.以上两点均正确

D.以上两点均错误

贝叶斯判别是在判别分析中充分吸收（ B ）信息的一种判别方法。

A.距离

B.先验概率

C.均值

D.方差

下列哪个公式是贝叶斯公式（ C ）

A.P(B|A)=P(B)P(A|B)/P(A)

B. P(B|A)=P(B)P(B|A)/P(A)

C. P(B|A)=P(A)P(A|B)

D. P(B|A)=P(A|B)

贝叶斯判别中一般将总体假设为什么分布（ A ）

A.正态分布 B.均匀分布 C.二项分布 D.泊松分布

有一组数据的众数>中位数>均值，请问这组数据的分布（A）  
A. 左偏  
B. 右偏  
C. 对称  
D. 以上都不对

k-means算法中的K指的是（ B ）

A.K个样本 B.K个质心 C.K次迭代 D.K次方

表示分布形状的数字特征有峰度和（ C ）。

A.均值

B.方差

C.偏度

D.中位数

方差的算术平方根称为（ D ）。

A.均值

B.极差

C.峰度

D.标准差

（ A ）是秩相关系数。

A. Spearman相关系数

B.Pearson相关系数

C.以上两种都是

D.以上两种都不是

在比较两个模型的拟合效果时，甲、乙两个模型的相关指数R2的值分别约为0.96和0.85，则拟合较好的模型是（ A ）。

A.甲

B.乙

C.无法比较

D.两者一样

单因素方差分析中，任何时候都成立的是（ D ）

A.SSE<SSA

B.SSA<SSE

C.MST=MSE+MSA

D.SST = SSA+SSE

线性回归模型y=bx+a+e ( a和b为模型的未知参数)中，e称为（ C ）。

A.回归系数

B.响应变量

C.随机误差

D.回归变量

主成分分析的目的是（ C ）

A.对原变量加以“改造”，在不知损失原变量太多信息的条件下尽可能地降低原变量的维数

B.用为数较少的“新变量”代替原来的各变量

C.以上两点均正确

D.以上两点均错误

贝叶斯判别是在判别分析中充分吸收（ B ）信息的一种判别方法。

A.距离

B.先验概率

C.均值

D.方差

下列哪个公式是贝叶斯公式（ C ）

A.P(B|A)=P(B)P(A|B)/P(A)

B. P(B|A)=P(B)P(B|A)/P(A)

C. P(B|A)=P(A)P(A|B)

D. P(B|A)=P(A|B)

贝叶斯判别中一般将总体假设为什么分布（ A ）

A.正态分布 B.均匀分布 C.二项分布 D.泊松分布

有一组数据的众数>中位数>均值，请问这组数据的分布（A）  
A. 左偏  
B. 右偏  
C. 对称  
D. 以上都不对

下列关于主成分分析中标准化变量的说明正确的是（B）

A.由于不同的量纲会引起各变量取值的分散程度差异过大，此时总方差主要受方差较小的变量控制  
B.为了消除原变量彼此方差差异过大的影响，通常将原变量进行标准化再做主成分分析  
C.标准化变量纯属多此一举，无需进行  
D.以上都不对

当你用跑步时间、年龄、跑步时脉搏以及最高脉搏作为预测变量来对耗氧量进行回归时，年龄的参数估计是-2.78， 这意味着什么？（其余预测变量不改变）（ B ）

A、年龄每增加一岁，耗氧量就增大2.78

B、年龄每增加一岁，耗氧量就降低2.78

C、年龄每增加2.78岁，耗氧量就翻倍

D、年龄每减少2.78岁，耗氧量就翻倍

在比较两个模型地拟合效果是否优良时，目前有两个模型甲、乙的相关指数分别为0.88与0.97，两者中拟合效果最好的模型是（C）

A.两者相同

B.甲

C.乙

D.两者无法比较

关于标准差与标准误，以下说法正确的是（D）

A.样本数增大时，样本差减小，标准差不变

B.可信区间大小与标准差有关，而参考值范围与标准误有关

C.样本数增大时，标准差与标准误均减小

D.总体标准差一定时，增大样本数会减小标准误

快速聚类法又称为动态聚类法，该方法首先将样品粗略地分类，然后依据（ A ）按一定规则逐步调整。

A.样品间的距离

B.聚点

C.以上两点都对

D.以上两点都不对

（ D ）的关键是依据样品间的距离定义类与类间的距离，从而按照类间距离从小到大进行聚类。

A.快速聚类

B.k-means聚类

C.层次聚类

D.谱系聚类

主成分分析可以利用（ A ）求解主成分。

A.相关系数矩阵

B.Fisher信息矩阵

C.Hessian矩阵

D.拉普拉斯矩阵

在自然科学与社会科学的众多领域中，研究对象往往用某种方式已划分为若干类型，当得到一个新的样品时，要确定该样品属于已知类型中的哪一类，这类问题属于（ D ）。

A.回归分析

B.方差分析

C.主成分分析

D.判别分析

贝叶斯判别是在判别分析中充分吸收（ B ）信息的一种判别方法。

A.距离

B.先验概率

C.均值

D.方差

根据判别分析的（ B ）可以把判别分析分为两组判别分析和多组判别分析。

A.数据量

B.数据组数

C.数据分布情况

D.以上都对

快速聚类法又称为动态聚类法，该方法首先将样品粗略地分类，然后依据（ A ）按一定规则逐步调整。

A.样品间的距离

B.聚点

C.以上两点都对

D.以上两点都不对

在误判率回代估计的计算中，当回判结果如下时（假设训练样本足够）：

|  |  |  |
| --- | --- | --- |
| 实际归类 | 回判情况 | |
| G1 | G2 |
| G1 | a | b |
| G2 | c | d |

请问下列选项中与误判率最为接近的为（D）  
A.  
B.

C.  
D.

快速聚类是指（ C ）。

A.两步聚类

B.系统聚类

C.k-means聚类

D.层次聚类

问答题

1. 请写出快速聚类法的步骤。
2. 请简单介绍一下χ2检验法。

3、简述求线性回归方程的步骤

4、阐述主成分分析的主要目的

计算题：

1. 调查某市出租车使用年限x和该年支出维修费用y（万元），得到数据如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 使用年限x | 2 | 3 | 4 | 5 | 6 |
| 维修费用y | 2.2 | 3.8 | 5.5 | 6.5 | 7.0 |

1. 求线性回归方程；
2. 由（1）中结论预测第10年所支出的维修费用。（）

解析：（1）列表如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 |
| xi | 2 | 3 | 4 | 5 | 6 |
| yi | 2.2 | 3.8 | 5.5 | 6.5 | 7.0 |
| xiyi | 4.4 | 11.4 | 22.0 | 32.5 | 42.0 |
| xi2 | 4 | 9 | 16 | 25 | 36 |
| =4, =5, =90, =112.3 | | | | | |

于是

a=5-1.23x4=0.08

所以线性回归方程为：=1.23x+0.08

（2）当x=10时，=1.23x10+0.08=12.38（万元）

即估计使用10年时维修费用是12.38万元。

2、某班有30名学生，某门科目的考试成绩如下：

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **81** | **55** | **87** | **78** | **65** | **76** | **79** | **99** | **59** | **56** | **68** | **90** | **71** | **29** | **61** |
| **83** | **79** | **68** | **74** | **77** | **77** | **49** | **85** | **84** | **88** | **75** | **73** | **74** | **54** | **76** |

作出茎叶图。（可增加平均数，中位数，三均数，一三分位数等进行考察）

（数据不难，考察茎叶图格式）

3、设总体的相关系数矩阵为  
  
(1) 写出主成分分析中标准化变量的计算公式（描述各符号的意义）

(2)求出特征值与特征向量，并且利用其求出各主成分的贡献率

(3)写出各主成分的表达式   
（施密特正交化公式）

答：

(1)  
(这是简写的，像书上一样复杂的写，我认为没必要，只要解释好均值和方差就行)

(2)

(特征向量比例相同即可，2 3位置可换)

第一主成分贡献率：

第二主成分贡献率：

第三主成分贡献率：

(3)(以下的数据均可转为小数，但由于是无穷小数，建议使用分数代替)

或者

（Y2 Y3可交换）

4、某个医院早上收了六个门诊病人，如下表：

|  |  |  |
| --- | --- | --- |
| 症状 | 职业 | 疾病 |
| 打喷嚏 | 护士 | 感冒 |
| 打喷嚏 | 农夫 | 过敏 |
| 头痛 | 建筑工人 | 脑震荡 |
| 头痛 | 建筑工人 | 感冒 |
| 打喷嚏 | 教师 | 感冒 |
| 头痛 | 教师 | 脑震荡 |

现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

解析：根据贝叶斯公式，有：

P（感冒|打喷嚏x建筑工人）

=

=

=

=0.66

因此，这个打喷嚏的建筑工人有66%的概率是得了感冒。