CDA LEVEL Ⅰ考试大纲

CERTIFIED DATA ANALYST LEVEL I EXAMINATION OUTLINE

一、总体目标

「CDA 数据分析师人才行业标准」是面向全行业数据分析及大数据相关岗位的一套科学化、专业化、正规化、系统化的人才技能准则。经管之家 CDA 数据分析师认证考试是评判「标准化人才」的唯一考核路径。CDA 考试大纲规定并明确了数据分析师认证考试的具体范围、内容和知识点，考生可按照大纲要求进行相关知识的学习，获取技能，成为专业人才。

二、考试形式与试卷结构

考试方式：线下统考，上机答题

考试题型：客观题（单选+多选）

考试时间：120 分钟

考试成绩：分为 A、B、C、D 四个层次，A、B、C 为通过考试，D 为不通过。

三、知识要求

针对不同知识，掌握程度的要求分为【领会】、【熟知】、【应用】三个级别，考生应按照不同知识要求进行学习。

1．领会：考生能够领会了解规定的知识点，并能够了解规定知识点的内涵与外延，了解其内容要点和它们之间的区别与联系，并能做出正确的阐述、解释和说明。

2．熟知：考生须掌握知识的要点，并能够正确理解和记忆相关理论方法，能够根据不同要求，做出逻辑严密的解释、说明和阐述。此部分为考试的重点部分。

3．应用：考生须学会将知识点落地实践，并能够结合相关工具进行商业应用，能够根据具体要求，给出问题的具体实施流程和策略。

四、考试范围

◆ **PART 1 数据分析概念与统计学基础 （占比 30%）**

a. 数据分析概念、方法论、流程（占比 5%）

b. 描述性统计分析（占比 10%）

c. 推断性统计分析（占比 7%）

d. 方差分析（占比 2%）

e. 一元线性回归分析（占比 3%）

f. 机器学习的基本概念（占比 3%）

◆ **PART 2 SQL 数据库基础 （占比 10%）**

a. SQL 及关系型数据库基本概念（占比 1%）

b. SQL 数据类型、运算符、函数（占比 2%）

c. SQL 查询语句（占比 3%）

d. SQL 连接语句（占比 3%）

e. SQL 其它语句（占比 1%）

◆ **PART 3 数据采集与处理 （占比 20%）**

a. 数据采集方法（占比 5%）

b. 市场调研（占比 2%）

c. 数据探索与可视化（占比 5%）

d. 数据预处理方法（占比 8%）

◆ **PART 4 数据建模分析 （占比 40%）**

a. 主成分分析法（占比 4%）、因子分析法（占比 2%）

b. 多元回归分析法

多元线性回归（占比 10%）

逻辑回归（占比 10%）

c. 聚类分析法

系统聚类法（占比 4%）

K-Means 聚类法（占比 5%）

d. 时间序列（占比 5%）

五、考试内容

PART 1 数据分析概念与统计学基础

◆ **1、数据分析概述**

【领会】

数据分析和数据挖掘的概念

强调商业数据分析中对业务的理解

商业数据分析和预测的本质

大数据对传统小数据分析的拓展

【熟知】

明确数据分析目标及意义

数据分析的过程

数据分析与数据挖掘的常用方法

CRISP-DM、SEMMA 方法论

数据分析中不同人员的角色与职责

◆ 2、描述性统计分析

【领会】

数据的类型

数据的集中趋势、离中趋势和数据分布的概念

统计图的概念

各种统计图的含义和画法

【熟知】

分类、顺序、数值型数据；截面、时序、面板数据；以及它们之间的区别与联系。

衡量数据集中趋势、离中趋势和数据分布的常用指标及计算方法

统计图形的绘制、图形元素的调整、可视化效果，主要涉及条形图、线图、直方图、盒须图、散点图、气泡图、马赛克图、玫瑰图及其多种图形整合。

明确统计图形对统计指标表达上的对应关系

【应用】

根据不同数据类型选用不同的统计指标来进行数据的集中趋势、离中趋势和数据分布的衡量，不同统计图的使用场景。会写数据分析报告和结合业务需求对报告进行合理解释，对业务提出建设性意见建议。

◆ 3、抽样估计

【领会】

随机试验、随机事件、随机变量的概念

总体与样本的概念

抽样估计的理论基础

正态分布及三大分布的函数形式和图像形式

抽样的多种组织形式

确定必要样本容量的原因

大数定律与中心极限定理的意义与应用

【熟知】

随机事件的概率

抽样平均误差的概念与数学性质

点估计与区间估计方法的特点与优缺点

全体总体与样本总体

参数和统计量

重复抽样与不重复抽样

抽样误差的概念对总体平均数、总体成数和总体方差的区间估计方法

必要样本容量的影响因素

【应用】

随机变量及其概率分布

全部可能的样本单位数目的概念及其在不同抽样方法下的确定

抽样平均误差在实际数据分析中的计算方法

◆ **4、假设检验**

【领会】

假设检验的基本概念

其基本思想在数据分析中的作用

假设检验的基本步骤

假设检验与区间估计的联系

假设检验中的两类错误

【熟知】

检验统计量、显著性水平及对应临界值（Critical Value）的基本定义

P 值的含义及计算

如何利用 P 值进行检验

z 检验统计量

t 检验统计量

F 检验统计量

χ^2 检验统计量的函数形式和检验步骤

【应用】

实现单样本 t 检验

两独立样本 t 检验的步骤和检验中使用的统计量与原假设

两种检验应用的数据分析场景。

◆ **5、方差分析**

【领会】

方差分析的相关概念

单因素方差分析的原理

统计量构造过程

【熟知】

单因素方差分析的基本步骤

总离差平方和（SST）的含义及计算

组间离差平方和（SSA）的含义及计算

组内离差平方和（SSE）的含义及计算

单因素方差分析的原假设

【应用】

实现单因素方差分析的步骤

对方差分析表的分析以及多重比较表的分析

◆ **6、一元线性回归分析**

【领会】

相关图的绘制与作用

相关表的编制与作用

相关系数定义公式的字母含义

估计标准误差与相关系数的关系

【熟知】

相关关系的概念与特点

相关关系与函数关系的区别与联系

相关关系的种类

相关系数的意义以及利用相关系数的具体数值对现象相关等级的划分

回归分析的概念

回归分析的主要内容和特点

建立一元线性回归方程的条件

一元线性回归系数的最小二乘估计

应用回归分析应注意的问题

估计标准误差的意义及计算

【应用】

运用简捷法公式计算相关系数与回归系数

相关分析分析中应注意的问题

回归分析与相关分析的区别与联系

◆ **7、机器学习的基本概念**

【领会】

机器学习的基本概念

机器模型构建的一般流程

交叉验证方法与参数选择

模型评估方法：混淆矩阵、准确率、召回率、F 值、ROC 曲线

【熟知】

机器学习中有监督学习和无监督学习的概念和特点

常见有监督学习算法：knn 算法，决策树，朴素贝叶斯

常见无监督学习算法：k-means 算法，PCA 算法

【应用】

有能力根据训练样本集情况来确定采用有监督还是无监督算法，在特定场景中（比如反欺诈领域）知道采用哪种方法更优。

PART 2 SQL 数据库基础

◆ **1、SQL 基础概念**

【领会】

关系型数据库基本概念、属性

主键

外键

E-R 图

ANSI-SQL 以及不同的数据库实现的关系

【熟知】

逻辑运算符

比较运算符

算术运算符

通配符

◆ **2、SQL 查询语句**

【应用】

select 语句

包括查询单列

多列，去重，前 N 列

from 语句、where 语句、group by 语句、having 语句、order by 语句、子查询

SQL 聚合函数，包括 count、sum、avg、max、min 等

◆ **3、SQL 连接语句**

【领会】

表的连接类型，包括内连接（等值、不等值）、外连接（左、右、全）、交叉连接（笛卡尔连接）

查询的集合操作，只包括并集操作

【应用】

inner join 的用法

left/right 的用法

cross join 的用法

union 的用法

◆ **4、其它 SQL 语句**

【领会】

表的创建

视图及索引的概念及创建

数据插入、更新、删除

常用函数及正则表达式

PART 3 数据采集与处理

◆ 1、数据采集方法

【领会】

一手数据与二手数据来源渠道

优劣势分析

使用注意事项

【熟知】

一手数据采集中的概率抽样与非概率抽样的区别与优缺点

【运用】

概率抽样方法，包括简单随机抽样、分层抽样、系统抽样、分段抽样

明确每种抽样的优缺点

根据给定条件选择最可行的抽样方式

计算简单随机抽样所需的样本量

◆ 2、市场调研

【熟知】

市场调研的基本步骤（提出问题、理论推演、收集材料、构建模型、归因分析）

单选题及多项选择题的设置

数据编码及录入

◆ 3、数据探索与可视化

【领会】

数据探索的目的与意义

常用数据可视化工具软件（EXCEL BI, SPSS, PYTHON, Tableau 等其中之一）

【熟知】

数据探索与数据预处理之间的关系

数据探索常用数据描述方法：集中趋势分析，离中趋势分析，数据分布关系，图分析

数据探索常用数理统计方法：假设检验，方差检验，相关分析，回归分析，因子分析

【应用】

能够通过使用数据可视化工具（EXCEL BI, SPSS, PYTHON, Tableau 等其中之一）来完成相关数据分析项目的数据探索任务。（说明：考试中不会考核该部分工具和软件的使用方法）

◆ **4、数据预处理方法**

【熟知】

数据预处理的基本步骤，包括数据集成（不同数据源的整合）、数据探索、数据变换（标准化）、数据归约（维度归约技术、数值归约技术），这部分内容不需要涉及计算，只需要根据需求明确可选的处理技术即可。

【应用】

数据清洗，包括填补遗漏的数据值（根据业务场景使用常数、中位数、众数等方法，不涉及多重查补的方法）、平滑有噪声数据（移动平均）、识别或除去异常值（单变量根据中心标准化值，多变量使用快速聚类），以及解决不一致问题（熟知概念即可），查重（只考核 SQL 的语句，不涉及其它语言）。

PART 4 数据建模分析

总体要求

领会模型基本原理，数值模型操作流程，懂得模型应用场景，能够完成数据建模分析报告。

◆ **1、主成分分析**

【领会】

主成分分析的计算步骤

主成分分析中对变量自身分布和多变量之间关系的假设以及模型设置

【熟知】

适用于主成分分析的变量度量类型。通过分析结果，选取合适的保留主成分的个数，注意区分两种不同的分析目的（尽量压缩变量、避免共线性情况下保留更多信息）保留主成分个数的评判标准的差异。

【应用】

在深入理解主成分的意义的基础之上，在遇到业务问题时，有能力决定是否使用主成分分析

方法；有能力决定何时采用相关系数计算方法和协方差矩阵计算方法；有能力解释主成分得分的结果；根据变量分布情况进行函数转换。

◆ **2、因子分析**

【领会】

了解因子分析模型设置，只需要关注主成分法的计算步骤

【熟知】

适用于因子分析的变量度量类型。通过分析结果，选取合适的因子个数；

知道最常用的因子旋转的方法。

【应用】

在遇到业务问题时，有能力决定是否使用因子分析，还是使用主成分分析方法就可以了；有能力根据原始变量在各因子上的权重明确每个因子的意义；有能力对大量变量进行维度分析，分维度打分，并比较与专家打分（德尔菲法）的区别；在聚类前对数据进行描述，发现理想的聚类方式和数量。

◆ **3、回归分析**

【领会】

线性回归的综合应用

【熟知】

明确线性回归的 6 个经典假设（线性模型、不存在共线性、残差期望为 0（无内生性）、同方差、正态性、随机抽样）

独立同分布的概念

明确违反上述假设后出现的问题

模型是否违反经典假设的检验方法与模型纠正的方法

变量筛选方法

离群值、指标计算方法

明晰横截面和时间序列数据在回归建模上的差异

【应用】

结合业务构建回归模型并且解释回归系数

根据业务场景与变量分布情况进行函数转换

解释变量为分类变量时的处理方法

区分预测性建模与解释性建模的关系

使用结果进行新样本预测

进行客户价值分析的基本步骤与注意事项

◆ **4、分类分析**

【领会】

卡方检验计算公式

二分类逻辑回归的计算公式

【熟知】

分类变量是否存在相关关系的描述方法和检验方法，涉及列联表分析、卡方检验

似然比与 Logit 转换

二分类逻辑回归模型构建与变量筛选

模型评估的方法，涉及混淆矩阵、ROC 曲线

【应用】

结合业务构建回归模型并且解释回归系数

根据业务场景与变量分布情况进行函数转换

使用结果进行新样本预测

逻辑回归与多元线性回归模型的结合应用

进行客户流失预测、信用评级、精准营销等模型的基本步骤与注意事项

◆ **5、聚类分析**

【领会】

多种聚类算法的特点

迭代的概念与实现

【熟知】

聚类方法的基本逻辑

距离的计算

系统聚类和 K-Means 聚类的基本算法和优缺点

系统聚类的计算步骤，包括两点距离、两类合并的计算方法

系统聚类法中选择最优聚类数量的方法

K-Means 聚类的基本算法

聚类分析变量标准化的原因和计算方法

变量需要进行主成分分析的原因

变量进行函数转化的原因和计算方法

【应用】

结合客户画像、客户细分、商品聚类、离群值检验（欺诈、反洗钱）等业务运用场景，选取合适的聚类方法与步骤

聚类事后分析，根据聚类后变量分布情况获取每类的特征

◆ **6、时间序列**

【领会】

明确趋势分解法、ARIMA 方法、时间序列回归方法的差异和适用场景

明确 ARIMA 方法的计算方法

【熟知】

趋势分解法，涉及乘法模型、加法模型

ARIMA 方法的具体步骤；时间序列回归的方法

【应用】

结合业务（业绩预测、预警），选取合适的分析方法

进行业务时间序列预测等模型的基本步骤与注意事项

六、推荐学习书目

说明：推荐学习书目中，部分书籍结合软件，但考试不会考软件，考生可根据自身需求选择性学习。参考书目不需全部学完，根据考纲知识点进行针对性学习即可。

[1] 贾俊平，何晓群，金勇进. 统计学（第 7 版）[M]. 中国人民大学出版社，2018.（必读）

[2] 斯蒂芬森，晋劳，琼斯. SQL 入门经典（第 5 版）[M]. 人民邮电出版社，2011.（必读）

[3] 黄缙华. MySQL 入门很简单[M]. 清华大学出版社，2011.（选读）

[4] 经管之家，丁亚军. 统计分析：从小数据到大数据[M]. 电子工业出版社，2020. （必读）

[5] 何晓群. 多元统计分析（第 4 版）[M]. 中国人民大学出版社, 2015.（选读）

[6] 李子奈，潘文卿. 计量经济学（第四版）[M]. 高等教育出版社，2015.（选读）

[7] 盛骤，试式千，潘承毅等. 概率论与数理统计（第四版）[M]. 高等教育出版社，2018.（选读）

[8] 张颢. 概率论[M]. 高等教育出版社，2018.（选读）

[9] 张文彤. SPSS 统计分析基础教程[M]. 高等教育出版社，2017.（选读）

[10] 王燕. 应用时间序列分析（第四版）[M]. 中国人民大学出版社, 2015.（选读）

[11] 詹姆斯·D·汉密尔顿. 时间序列分析[M]. 中国人民大学出版社，2015.（选读）

[12] Daniel T. Larose，Chantal D. Larose. 数据挖掘与预测分析（第 2 版）[M]. 清华大学出版社，2017. （选读）

CDA Institute

经管之家 CDA 数据分析研究院