（3）结构方程建模（针对潜变量之间的关系进行建模）。

（4）因子分析（调查设计和验证的探索型分析）。

（5）功效分析/试验设计（特别是基于仿真的试验设计，以避免分析过度）。

（6）非参数检验（MCMC 等）。

（7）k-means 聚类。

（8）贝叶斯方法（朴素贝叶斯、贝叶斯模型平均、贝叶斯适应性试验等）。

（9）惩罚性回归模型（弹性网络、Lasso、LARS 等）以及对通用模型（SVM、XGBoost 等）加罚分，这对于预测变量多于观测值的数据集很有用，在基因组学和社会科学研究中较为常用。

（10）样条模型（MARS 等），主要用于流程建模。

（11）马尔可夫链和随机过程（时间序列建模和预测建模的替代方法）。

（12）缺失数据插补方法及其假设（missForest、MICE 等）。

（13）生存分析（主要特点是考虑了每项观测出现某一结果的时间长短）。

（14）混合建模。

（15）统计推断和组群测试（A/B 测试以及用于营销活动的更复杂的方法）。

此外，建议读者根据自己所属领域重点学习面向该领域的专用模型。

习　题

一、选择题

1．从学科定位看，数据科学处于（　　）的重叠之处，具有显著的跨学科性。

A．数学与统计知识　　B．计算机科学

C．3C 精神与技能　　D．领域实务知识

2．以下提法中正确的是（　　）。

A．数据科学中的“数据”并不仅仅是“数值”，也不等同于“数值”

B．数据科学中的“计算”并不仅仅是加、减、乘、除等“数学计算”，还包括数据的查询、挖掘、洞见、分析、可视化等更多类型

C．数据科学关注的是“单一学科”的问题

D．数据科学强调的是“理论研究”，一般不涉及“领域实务知识”

3．数据科学领域常用的工具之一——（　　）是统计学家发明的语言。

A．Python　　B．R

C．Java　　D．C 语言

4．（　　）一般采用图表或数学方法描述数据的统计特征，如分布状态、数值特征等。

A．推断统计　　B．预测分析

C．描述统计　　D．诊断分析

5．2014年3月，Lazer D、Kennedy R和King G等在*Science*上发表了一篇标题为《谷歌流感的寓言：大数据分析的陷阱》（*The Parableof Google Flu: Trapsin Big Data Analysis*）的论文，提出GFT出现预测不准确性的主要原因是（　　）。

A．大数据浮夸　　B．算法动态性和用户行为的变化

C．原始算法的设计错误　　D．一直在虚假报道

6．迈尔-舍恩伯格与库克耶在其著名论著《大数据：一场改变我们生活、工作和思维方式的革命》中提出了大数据时代统计的思维变革包括（　　）。

A．不是知识驱动，而是数据驱动　　B．不是随机样本，而是总体数据

C．不是精确性，而是混杂性　　D．不是因果关系，而是相关关系

7．关于统计学与数据科学的内在联系，以下描述中正确的是（　　）。

A．统计学是数据科学的主要理论基础之一

B．统计学家在数据科学的发展中做出过突出贡献

C．数据科学是统计学的一个子学科

D．数据科学领域常用的工具之一——R是统计学家发明的语言

8．以下选项中属于描述统计的是（　　）。

A．集中趋势分析　　B．离中趋势分析

C．相关分析　　D．假设检验

9．“先对总体的参数μ的值提出一个假设，然后利用样本统计量来检验这个假设是否成立”的方法，属于统计学中的（　　）。

A．非参数检验　　B．点估计　　C．区间估计　　D．参数假设检验

10．在数据科学中，常用的元分析法有（　　）。

A．逻辑回归　　B．多项式回归　　C．加权平均法　　D．优化方法

11．在有异常值的情况下，中位数和均值哪个评价结果更合理和贴近实际（　　）。

A．均值　　B．中位数

C．中位数和均值均可以　　D．中位数和均值均不可以

12．对于具有单峰分布的大多数数据而言，如果数据是左偏分布，则众数、中位数和均值之间的关系是（　　）。

A．众数=中位数=算术平均数　　B．算术平均数＜中位数＜众数

C．众数＜中位数＜算术平均数　　D．中位数＜众数＜算术平均数

二、调研与分析题

1．调查并通过实验分析SPSS Statistics、SPSS Modeler、SPSS Analytic Server和SPSS Analytic Catalyst的区别与联系。

2．结合自己的专业领域，调研自己所属领域的统计分析方法、技术与工具。

3．调研常用统计分析工具软件（包括开源系统），并进行对比分析。