

东北大学考研432统计学 简答题汇总

备考适用



2019-11-1

**标注记号说明：**

1. 高亮显示、红色字体:

标识语句中心词、关键词、重点词、重点知识点等。

1. **下划线加粗:**

**东北大学简答题往年考题。**

1. **波浪线加粗:**

**未考察但应该重点背诵的知识点。**

1. ★:

可与其他知识综合考察的知识点。

1. 只写编号而没有填充内容:例如:“**4、t分布:**”

(1) 已在前面或将在后面的总结中出现，不再赘述。

(2) 以公式为主体，请自行查阅对应书籍的相关章节内容。

1. （14-15）（16）、（17）、（18）、（19）、（20）：

考点及其出现年份，例如（20）表示考点在2020年考研真题中出现。

**注：有些考点并非以显式的形式考察，而是隐藏在题目信息之中，因此在真题中并不能直接看到对该考点的考察，请读者深入思考出题点。**

**补注：2021修改初试大纲，使用卯诗松《概率论与数理统计》，但简答题基本不变，本文档可延用，不必担心；**

# 《统计学》 贾俊平 第七版

**第一部分 数据的图表展示**

**1、频数**

**2、比例**

**3、比率**

**\*4、条形图**

**5、饼图**

**6、环形图与饼图的区别**

**\*7、数据分组**

**\*8、组距分组的步骤**

**\*9、组距分组的原则**

**\*10、组距分组的假定条件**

**\*11、直方图（20）**

**\*12、直方图与条形图的区别**

**13、茎叶图**

**14、茎叶图与直方图的比较**

**\*15、箱线图（20）**

**\*16、箱线图的绘制方法**

**\*17、散点图**

**18、气泡图**

**第二部分 数据的概括性度量**

**\*1、众数**

**\*2、众数的特点**

**\*3、中位数**

**\*4、四分位数**

**\*5、平均数**

**\*6、众数、中位数、平均数的比较**

**7、异众比率**

**\*8、四分位差（20）**

**\*9、极差（20）**

**10、平均差**

**\*11、方差和标准差**

**12、标准分数**

**\*13、3σ法则(经验法则)**

**\*14、切比雪夫不等式（20）**

**\*15、离散系数**

**16、为什么使用离散系数**

**\*17、离散程度测度值的选取**

**第三部分 概率与概率分布**

**1、古典概型的特点**

**2、概率的古典定义**

**3、概率的统计定义**

**\*4、概率的公理化定义**

**\*5、随机变量**

**\*6、离散型随机变量**

**\*7、连续性随机变量（20）**

**\*8、二项分布(n重伯努利实验)的性质（20）**

**\*9、泊松分布（20）**

**\*10、泊松定理**

**\*11、概率密度函数的性质**

**\*12、正态分布**

**第四部分 统计量及其抽样分布**

**\*1、统计量**

**\*2、抽样分布**

**\*3、χ2分布**

**\*4、t分布**

**\*5、F分布**

**\*6、中心极限定理**

**第五部分 参数估计**

**\*1、参数估计**

**2、估计量与估计值**

**\*3、点估计**

**\*4、区间估计**

**5、区间估计的三要素**

**\*6、置信区间**

**\*7、置信水平**

**8、置信区间与样本量**

**\*9、关于置信区间**

**10、区间估计的精度与可靠性**

**\*11、无偏性**

**\*12、有效性**

**\*13、一致性**

**\*14、两总体、小样本、均值之差的区间估计的假定**

**\*15、样本量与置信水平**

**第六部分 假设检验**

**\*1、参数估计和假设检验**

**\*2、两类错误（20）**

**\*3、两类错误的概率**

**\*4、两类错误的控制**

**\*5、假设检验的流程:（20）**

**\*6、P值**

**7、P的决定因素**

**\*8、显著性水平α与P值**

**\*9、显著性水平α的含义**

**10、两总体均值之差检验统计量的确定**

**\*两总体均值差的检验（20）**

**11、不接受与拒绝**

**第七部分 拟合优度检验和独立性检验**

**\*1、χ2统计量的特征**

**\*2、独立性检验**

**3、χ2分布的期望值准则**

**第八部分 方差分析**

**1、方差分析相比于假设检验的优点（20）**

**\*2、方差分析(ANOVA)（20）**

**3、误差分解**

**\*4、方差分析中的基本假定**

**\*5、单因素方差分析的步骤（20）**

**\*6、SSA、SSE、SST**

**\*7、三个平方和的自由度**

**8、关系强度的测量**

**9、多重比较(LSD)检验步骤**

**10、双因素方差分析步骤**

**11、双因素方差分析与单因素方差分析**

**第九部分 一元线性回归**

**1、函数关系**

**\*2、相关分析要解决的问题**

**\*3、进行相关分析时对总体的假定**

**\*4、相关系数**

**\*5、相关系数的性质**

**6、相关程度**

**7、r的显著性检验**

**\*8、回归分析主要解决的问题**

**9、回归模型**

**\*10、基本假定**

**11、回归方程**

**\*12、回归分析的一般过程**

**\*13、估计的回归方程中系数的解释**

**\*14、最小二乘法的思想**

**\*15、最小二乘估计的性质**

**\*16、平方和分解**

**\*17、判定系数R2**

**18、判定系数的实际意义**

**\*19、估计的标准误差（20）**

**20、线性关系的检验步骤**

**21、回归系数的检验步骤**

**22、回归结果的评价**

**23、残差**

**第十部分 多元线性回归**

**\*1、多元回归模型的基本假定**

**2、多重判定系数**

**\*3、调整的多重判定系数**

**4、线性关系的检验步骤**

**5、回归系数检验的步骤**

**\*6、多重共线性**

**\*7、多重共线性可能导致的问题**

**8、多重共线性的判别**

**9、多重共线性问题的处理**

**\*10、选择自变量的原则**

**\*11、变量选择的方法**

**第十一部分 时间序列分析和预测**

**\*1、时间序列的组成要素**

**2、时间序列的预测步骤**

**\*3、时间序列预测方法**

**4、复合型序列的分解预测步骤**

**5、季节指数计算方法**

# 《概率论与数理统计》 盛骤 浙大四版

**第一章 概率论基本概念**

**1、随机实验**

**2、频率**

**3、频率的性质**

**\*4、概率的公理化定义**

**\*5、实际推断原理（20）**

**\*6、两事件独立**

**第二章 随机变量及其分布**

**1、随机变量**

**\*2、二项分布的性质（20）**

**\*3、泊松定理**

**\*4、分布函数及其性质（20）**

**\*5、均匀分布的意义**

**\*6、指数分布的无记忆性**

**第三章 多维随机变量及其分布**

**\*1、二维随机变量的分布函数及其性质**

**\*2、随机变量的相互独立**

**\*3、随机变量的函数的分布（20）**

**\*4、联合分布与边缘分布的关系**

**第四章 随机变量的数字特征**

**1、数学期望的性质**

**2、方差及其性质**

**\*3、切比雪夫不等式（20）**

**\*4、不相关和独立**

**\*5、相关系数的概念及意义（20）**

**第五章 大数定律及中心极限定理**

**\*1、大数定律（20）**

**2、中心极限定理**

**3、弱大数定理(辛钦大数定理)**

**4、伯努利大数定理**

**5、独立同分布的中心极限定理**

**6、棣莫弗-拉普拉斯定理**

**第六章 样本及抽样分布**

**1、样本**

**\*2、样本分位数**

**\*3、箱线图（20）**

**4、箱线图的性质**

**5、经验分布函数**

**\*6、χ2分布及其性质**

**\*7、t分布及其性质**

**\*8、F分布及其性质**

**9、样本均值的抽样分布**

**10、抽样分布定理**

**第七章 参数估计**

**\*1、矩估计法**

**\*2、极大似然估计法及其性质**

**\*3、置信区间**

**第八章 假设检验**

**\*1、显著性检验**

**\*2、假设检验问题的步骤（20）**

**\*3、置信区间与假设检验之间的关系**

**\*4、p值**

**\*5、两类错误的地位**

**第九章 方差分析与回归分析**

**\*1、方差分析**

**\*2、方差分析的基本假定**

**3、平方和分解**

**\*4、平方和的自由度**

**5、平方和的统计特性**

**\*6、回归分析中最小二乘估计与极大似然估计的等价性**

**\*7、回归效果不显著的原因**

# 《概率论与数理统计》 梁之舜 中山三版

**第六章 抽样分布**

**\*1、顺序统计量**

**2、抽样分布定理**

**3、柯赫伦分解定理**

**第七章 估计理论**

**\*1、矩估计法的优缺点**

**\*2、最小方差无偏估计量**

**\*3、最小方差线性无偏估计**

**\*4、优效估计**

**5、有效率**

**\*6、相合估计**

**\*7、充分统计量**

**第八章 假设检验**

**1、皮尔逊χ2检验步骤**

**\*2、两类错误（20）**

**\*3、两类错误的关系及其控制**

**第九章 回归分析与方差分析**

**\*1、回归模型的基本假定**

**\*2、最小二乘法估计量的6条性质**

# 《多元统计分析》 何晓群 第四版

**第一部分 聚类分析**

**1、指标类型的尺度**

**2、聚类分析的两个方向**

**3、聚类分析的方法**

**4、类与类之间的距离**

**5、分类的准则**

**6、空间张度**

**\*7、系统聚类的步骤**

**8、模糊聚类分析计算步骤**

**\*9、聚类分析的计算步骤**

**第二部分 判别分析**

**1、判别分析的基本要求**

**\*2、判别分析的假设**

**第三部分 主成分分析**

**\*1、主成分分析及其思想**

**\*2、主成分与原始变量的关系**

**\*3、主成分分析步骤**

**第四部分 因子分析**

**\*1、因子分析及其思想**

**2、因子分析的种类**

**\*3、因子分析模型的假定**

**\*4、主成分分析与因子分析的区别**

**\*5、因子分析的步骤**

**第五部分 对应分析**

**1、对应分析**

**\*2、对应分析的步骤**

**第六部分 典型相关分析**

**1、典型相关分析**

**\*2、典型相关分析的步骤**

**简答题及概念补充**

**《概率论与数理统计》茆诗松 第二版**

**《数据分析》范金城 梅长林 第二版**

**\*统计量概念补充**

# 《统计学》 贾俊平 第七版

**第一部分 数据的图表展示**

**1、频数：**

落在某一特定类别或组别中的数据个数。

**2、比例：**

也称构成比，它是构成一个样本(或总体)中各个部分的数据与全部数据之比，通常用于反映样本(或总体)的构成或结构。

**3、比率:**

样本(或总体)中不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因而比值可能大于1。

**4、条形图:**

**用宽度相同的条形的高度或长短来表示数据的多少的图形。条形图可以横置或纵置，纵置时也称为柱形图。**

**5、饼图:**

用圆形及圆内扇形的角度来表示数值大小的图形，他主要用于表示一个样本(或总体)中各个组成部分的数据占全部数据的比例，对于研究结构性问题十分有用。

**6、环形图与饼图的区别和联系:**

①简单饼图只能显示一个样本各部分所占的比例。

②环形图中间有一个“空洞”，每个样本用一个环来表示，样本中的每一部分数据用环中的一段来表示。因此环形图可以显示多个样本各部分所占的相应比例，有利于对构成做比较研究。

**7、数据分组:**

概念:根据统计研究的需要，将原始数据按照某种标准分成不同的组别。

目的:观察数据的分布特征。

方法:单变量值分组（离散型）、组距分组（连续型）。

**8、组距分组的步骤:**

①确定组数:一般情况下，一组数据所分的组数不应少于5组且不多于15组，

即；

②确定各组组距:组距是一个组的上限与下限的差；组距=(最大值-最小值)÷组数；

③根据分组编制频数分布表。

**9、组距分组的原则:**

“不重不漏”(也称“穷尽原则”和“互斥原则”)；

①不重是指一项数据只能分在其中的某一组，不能在其他组中重复出现；为解决不重的问题，统计分组时习惯上规定“上组限不在内”。

②不漏是指组别能够穷尽，即在所分的组别中每项数据都能分在其中的某一组，不能遗漏；

**10、组距分组的假定条件:**

各组数据在本组内呈均匀分布，或在组中值两侧呈对称分布。

11、**直方图:（16）、（20）**

**用于展示分组数据分布的一种图形，它是用矩形的宽度和高度（即面积）来表示频数分布的。绘制该图时，在平面直角坐标系中，用横轴表示数据分组，纵轴表示频数或频率。**

**12、直方图与条形图的区别:**

**①条形图是用条形的长度(横置时)表示各类别频数的多少，其宽度(表示类别)则是固定的；直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或频率，宽度则表示各组的组距，因此其高度和宽度均有意义。**

**②由于分组数据具有连续性，直方图的各矩形通常是连续排列，而条形图则是分开排列。**

**③条形图主要是用于展示分类数据，而直方图则主要用于展示数值型数据。**

**13、茎叶图:**

反映原始数据分布的图形。它由茎和叶两部分构成，其图形是由数字组成的。

**14、茎叶图与直方图的比较:**

茎叶图类似于横置的直方图，与直方图相比，茎叶图既能给出数据的分布状况，又能给出每一个原始数值，即保留了原始数据的信息。而直方图虽然能很好的显示数据的分布，但不能保留原始的数值。在应用方面，直方图通常适用于大批量数据，茎叶图通常适用于小批量数据。

**15、箱线图:（16）、（17）、（20）**

**箱线图是根据一组数据的最大值、最小值、中位数、下四分位数、上四分位数这五个特征值绘制而成的，它主要用于反映原始数据分布的特征，还可以进行多组数据分布特征的比较。**

**16、箱线图的绘制方法:**

**①先找出一组数据的最大值、最小值、中位数和两个四分位数；**

**②连接两个四分位数画出箱子；**

**③再将最大值和最小值与箱子相连接，中位数在箱子中间。**

**17、散点图（16）:**

**用二维坐标展示两个变量之间关系的一种图形。**

**18、气泡图:**

用于展示三个变量之间的关系。与散点图类似，绘制时将一个变量放在横轴，另一个变量放在纵轴，第三个变量则用气泡的大小来表示。

**第二部分 数据的概括性度量**

1、**众数（16）:**

**一组数据中出现次数最多的变量值。众数主要用于测度分类数据的集中趋势，也可以作为顺序数据以及数值型数据集中趋势的测度值。一般情况下，只有在数据量较大的情况下众数才有意义**。

**2、众数的特点（16）：★注意与其他集中趋势度量的比较★**

**众数是一个位置代表值，它不受数据中极端值的影响。从分布的角度看，众数是具有明显集中趋势点的数值，一组数据分布的最高峰点所对应的数值即为众数。当然，如果数据的分布没有明显的集中趋势或最高峰点，众数可能不存在；如果有两个或多个最高峰点，则可以有两个或多个众数。**

**3、中位数（16）、（17）:**

**中位数是一组数据排序后处于中间位置上的变量值，用表示。显然，中位数将全部数据等分成两部分，每部分包含50％的数据，一部分数据比中位数大，另一部分则比中位数小。中位数主要用于测度顺序数据的集中趋势，当然也适用于测度数值型数据的集中趋势，但不适用于分类数据。中位数是一个位置代表值，其特点是不受极端值影响**。

**4、四分位数:**

也称四分位点，它是一组数据排序后处于25％和75％位置上的值。上四分位数和下四分位数之间包含了50％的数据。

**5、平均数:（16）、（17）**

**也称均值，它是一组数据相加后除以数据个数得到的结果。平均数是集中趋势的最主要测度值，它主要适用于数值型数据，而不适用于分类数据和顺序数据。平均数利用了全部数据信息**。

**6、众数、中位数、平均数的比较（16）:**

**①如果数据的分布是对称的，众数、中位数、平均数必定相等；**

**②如果数据是左偏分布，说明数据存在极小值，必然拉动数据向极小值一方靠，而众数和中位数由于是位置代表值，不受极值的影响，因此三者的关系表现为:平均数中位数众数；**

**③如果数据是右偏分布，说明数据存在极大值，必然拉动数据向极大值一方靠，因此众数中位数平均数。**

**7、异众比率:**

指非众数组的频数占总频数的比例。异众比率主要用于衡量众数对一组数据的代表程度。异众比率越大，说明非众数组频数占总频数的比重越大，众数的代表性越差；异众比率越小，说明非众数组的频数占总频数的比重越小，众数的代表性越好。异众比率适合测度分类数据的离散程度，当然，对于顺序数据以及数值型数据也可以计算异众比率。

**8、四分位差:（20）**

**也称内距或四分间距，它是上四分位数与下四分位数之差。四分位差反映了中间50％的数据的离散程度，数值越小，说明中间的数据越集中；数值越大，说明中间的数据越分散。四分位差不受极值的影响。此外，由于中位数处于数据的中间位置，因此，四分位差的大小在一定程度上说明了中位数对一组数据的代表程度。四分位差主要用于测度顺序数据的离散程度。对于数值型数据也可以计算四分位差，但它不适合分类数据。**

**9、极差:（20）**

**一组数据的最大值与最小值之差称为极差，也称全距，用表示。极差是最简单的描述数据离散程度的测度值，计算简单，易于理解，但它容易受极端值的影响。由于极差只是利用了一组数据两端的信息，不能反映出中间数据的分散状况，因而不能准确描述出数据的分散程度。**

**10、平均差:**

也称平均绝对离差，它是各变量值与其平均数离差绝对值的平均数。平均差以平均数为中心，反映了每个数据与平均数的平均差异程度，他能全面准确地反映一组数据的离散状况。平均差越大，说明数据的离散程度越大；反之，则说明数据的离散程度越小。为了避免离差之和等于零而无法计算平均差这一问题，平均差在计算时对离差取了绝对值，以离差的绝对值来表示总离差，这就给计算带来了不便、因而在实际中应用较少。但平均差的实际意义比较清楚，容易理解。

**11、方差和标准差:**

各变量值与其平均数离差平方和的平均数。方差的平方根称为标准差。方差(或标准差)能较好的反映出数据的离散程度，是应用最广的离散程度的测度值。标准差是有量纲的，它与变量值的计算单位相同，其实际意义要比方差清楚。

**12、标准分数:**

也称标准化值或z分数，它是变量值与其平均数的离差除以标准差后的值。在对多个具有不同量纲的变量进行处理时，常常需要对各变量进行标准化处理。标准分数具有平均数为0，标准差为1的特性。实际上分数只是将原始数据进行了线性变换，并没有改变一个数据在该组数据中的位置，也没有改变该组数据分布的形状，而只是将该组数据变为平均数为0，标准差为1。标准分数没有量纲。

**13、“法则”（经验法则）（17）:**

**约有68％的数据在平均数个标准差的范围内；**

**约有95％的数据在平均数个标准差的范围内；**

**约有99％的数据在平均数个标准差的范围内。**

**在平均数个标准差之外的数据，统计上称为离群点**。

**14、切比雪夫不等式的实际应用：★注意与经验法则对比★（20）**

**经验法则适合对称分布的数据。如果一组数据不是对称分布，经验法则不再适用，这时可以用切比雪夫不等式，它对任何分布形态的数据都适用。**

**至少有75％的数据在平均数个标准差的范围内；**

**至少有89％的数据在平均数个标准差的范围内；**

**至少有94％的数据在平均数个标准差的范围内。**

**15、离散系数:**

也称变异系数，它是一组数据的标准差与其相对应的平均数之比。离散系数是测度数据离散程度的统计量，主要用于比较不同样本数据的离散程度。离散系数越大，说明数据的离散程度也大；离散系数越小，说明数据的离散程度也小。

**16、为什么使用离散系数:**

方差和标准差是反映数据离散程度的绝对值，其数值的大小一方面受原变量值自身水平高低的影响，也就是与变量的平均数大小有关，变量值绝对水平高的，离散程度的测度值自然也就大，绝对水平低的，离散程度的测度值自然也就小；另一方面，他们与原变量值的计量单位相同，采用不同计量单位计量的变量值，其离散程度的测度值也就不同。因此对于平均水平不同或计量单位不同的不同组别的变量值，不能用标准差直接比较其离散程度。为消除变量值水平高低和计量单位对不同离散程度测度值的影响，需要计算离散系数

**17、离散程度测度值的选取:**

对于分类数据，主要用异众比率来测度其离散程度；

对于顺序数据，虽然也可以计算异众比率，但主要使用四分位差来测度其离散程度；

对于数值型数据，虽然可以计算异众比率、四分位差、极差和平均差等，但主要使用方差或标准差来测度其离散程度。

当需要对不同样本数据的离散程度进行比较时，则使用离散系数。

**第三部分 概率与概率分布**

**1、古典概型的特点:**

①结果有限:即基本空间中只含有限个元素。

②等可能性:各个结果出现的可能性被认为是相同的。

**2、概率的古典定义:**

如果某一随机实验的结果有限，而且各个结果出现的可能性相等，则某一事件发生的概率为该事件所包含的基本事件个数与样本空间中所包含的基本事件个数的比值。

古典概率局限在随机实验结果只有有限个可能结果的范围内，这使其应用受到了很大限制。

**3、概率的统计定义:**

在相同条件下随机实验次，某事件出现次，则比值称为事件发生的频率。随着的增大，该频率围绕某一常数上下波动，且波动的幅度逐渐减小，趋于稳定，这个频率的稳定值即为该事件的概率。

**4、概率的公理化定义:**

**5、随机变量:**

**在同一组条件下，如果每次实验可能出现这样或那样的结果，并且所有的结果都能列举出来，即的所有可能值都能列举出来，而且的可能值具有确定的概率，其中称为概率函数，则称为的随机变量，称为随机变量的概率函数。**

**6、离散型随机变量:**

如果随机变量的所有取值都可以逐个列举出来，则称为离散型随机变量。

**7、连续性随机变量:（20）**

**如果随机变量的所有取值无法逐个列举出来，而是取数轴上某一区间内的任一点，则称为连续性随机变量。**

**8、二项分布重伯努利实验或伯努利实验的性质:（20）**

**①包含个相同的实验；**

**②每次实验只有两个可能的结果:“成功”或“失败”；**

**③出现“成功”的概率对每一次实验是相同的，“失败”的概率也是如此，**

**且；**

**④实验室相互独立的；**

**⑤实验“成功”或“失败”可以计数，即实验结果对应于一个离散型随机变量。**

**9、泊松分布:（20）**

**用来描述在一指定时间范围内或在指定的面积或体积之内某一事件出现的次数的分布**

**10、泊松定理:**

**11、概率密度函数的性质:**

**①非负性:**

**②规范性:**

**12、正态分布:**

正态分布概率密度函数的特点:

1. 非负性:即整个概率密度曲线都在轴上方。

②曲线相对于对称，并在处达到最大值，，也称为位置参数。

1. 曲线的陡缓程度由决定，越大，曲线越平缓；越小，曲线越陡峭。也称为形状参数。
2. 当趋于无穷时，曲线以轴为其渐近线。

**第四部分 统计量及其抽样分布**

**1、统计量:**

**设是从总体中抽取的容量为的一个样本，如果由此样本构造一个函数，不依赖于任何未知参数，则称是一个统计量。**

**2、抽样分布:**

在总体的分布类型已知时，若对任一自然数都能导出统计量的分布的数学表达式，这种分布称为精确的抽样分布。

**3、分布:★(注意三大抽样分布的关系)**

**4、分布:★（18）、（19）**

**5、分布:★（18）、（19）**

**6、中心极限定理:**

**设从均值为方差(有限)为的任意一个总体中抽取样本量为的样本，当充分大时，样本均值的抽样分布近似服从均值为,方差为的正态分布。**

**第五部分 参数估计**

**1、参数估计:**

用样本统计量去估计总体的参数。

**2、估计量与估计值:**

在参数估计中，用来估计总体参数的统计量称为估计量。

根据一个具体样本计算出来的估计量的数值称为估计值。

**3、点估计:**

用样本统计量的某个取值直接作为总体参数的估计值。

**4、区间估计（18）:**

**在点估计的基础上，给出总体参数估计的一个区间范围，该区间通常由样本统计量加减估计误差得到。**

**5、区间估计三要素:**

**①点估计值；**

**②抽样平均误差；**

**③估计的可靠度**。

**6、置信区间:**

在区间估计中，由样本统计量所构造的总体参数的估计区间称为置信区间，其中区间的最小值称为置信下限，最大值称为置信上限。

7**、置信水平（14-15）：★注意与两类错误的概率比较★**

**一般地，如果将构造置信区间的步骤重复多次，置信区间中包含总体参数真值的次数所占的比例称为置信水平，也称为置信度或置信系数。**

**8、置信区间与样本量、置信水平的关系:**

当样本量给定时，置信区间的宽度随着置信系数的增大而增大；当置信水平固定时，置信区间的宽度随样本量的增大而减小，换言之，较大的样本量所提供的有关总体的信息要比较小的样本多。

**9、★关于置信区间★：(注意置信区间与假设检验的接受域之间的关系)**

**①如果用某种方法构造的所有区间中有95％的区间包含总体参数的真值，5％的区间不包含总体参数的真值，那么，用该方法构造的区间称为置信水平为95％的置信区间。**

**②总体参数的真值是固定的、未知的，而用样本构造的区间则是不固定的(置信区间是随机区间)。**

**③由一个(特定)样本所构造的区间是一个特定的区间，不再是随机区间，即一个特定的区间“总是包含”或“绝对不包含”参数的真值，不存在“以多大概率包含总体参数”的问题。**

**10、区间估计的精度与可靠性:**

对于一定的样本量，要想增大估计的可靠性，就会使区间变宽而降低精度；要想提高精度，就要求估计区间变得很窄，而这样，估计的可靠性就会大打折扣。增大样本量可以同时提高精度和可靠性。

**11、无偏性（14-15）、（17）:**

**估计量抽样分布的数学期望等于被估计的总体参数。总体参数为，所选择的估计量为，如果，则称为的无偏估计量。**

**12、有效性:**

**对同一总体参数的两个无偏估计量，有更小标准差的估计量更有效。假定分别是总体参数*θ*的两个无偏估计量，他们的抽样分布的方差用表示，如果**

**，则称比更有效。**

**13、一致性:**

随着样本量的增大，估计量的值越来越接近被估计总体的参数。

**14、两总体、小样本、均值之差的区间估计的假定:**

①两个总体都服从正态分布。

②两个随机样本独立地的分别抽自两个总体。

**15、样本量与置信水平:**

样本量与置信水平成正比，在其他条件不变的情况下，置信水平越大，所需样本量也越大；样本量与总体方差成反比，即可以接受的估计误差的平方越大，所需的样本量就越小。

**第六部分 假设检验**

**1、参数估计和假设检验（18）:**

**参数估计和假设检验是统计推断的两个组成部分，他们都是利用样本对总体进行某种推断，但推断角度不同。参数估计讨论的是用样本统计量估计总体参数的方法，总体参数在估计前是未知的。而在假设检验中，则是先对的值提出一个假设，然后利用样本信息去检验这个假设是否成立。**

2、**两类错误（14-15）、（17）、（19）、（20）:**

**第Ⅰ类错误是原假设为真却被我们拒绝了，犯这种错误的概率用表示，所以也称错误或弃真错误；第Ⅱ类错误是原假设为伪我们却没有拒绝，犯这种错误的概率用表示，所以也称错误或取伪错误。**

3、**两类错误的概率（19）:**

**对于一定的样本量，如果减少错误，就会增大犯错误的机会；若减少错误就会增大犯错误的机会。增大样本量可以同时减少和。**

4、**两类错误的控制（14-15）、（19）:**

**一般来说，哪一类错误所带来的后果越严重，危害越大，在假设检验中就应当把哪一类错误作为首要的控制目标。在假设检验中，一般首先控制犯错误。主要的原因在于，原假设是什么通常是明确的，而备择假设是什么通常是模糊的。**

**5、假设检验的流程:（18）、（20）**

**①提出原假设和备择假设；**

**②选择适当的检验统计量；**

**③计算检验统计量的值，并与给定显著性水平下的临界值比较；**

**④根据③的比较结果做出决策。**

**6、值:**

值是当原假设为真时，样本观察结果或更极端结果出现的概率。

值越小，拒绝原假设的理由越充分。

**7、值决定因素:**

①样本数据与原假设之间的差异；

②样本量；

③被假设参数的总体分布。

**8、显著性水平与值:**

**确定以后，拒绝域的位置也就相应确定了，其好处是进行决策的界限清晰，但缺陷是进行决策面临的风险是笼统的。**

**值的长处是它反映了观察到的实际数据与原假设之间不一致的概率值，与传统的拒绝域范围相比，是一个具体的值，这样就提供了更多的信息。**

**8、单总体检验统计量的确定:**

①样本量:

在样本量大的条件下，如果总体为正态分布，则样本统计量服从正态分布；

如果总体为非正态分布，则样本统计量渐进服从正态分布。

在这些情况下，可以把样本统计量视为正态分布，这时可以使用统计量。

②总体标准差是否已知:

在样本量较小的情况下:

如果总体标准差已知，则样本统计量服从正态分布，这时可以采用统计量；

如果总体标准差未知，进行检验所依赖的信息有所减少，这时只能使用样本标准差，样本统计量服从分布，应该采用统计量。

当样本量时，分布与分布非常接近，具备了用分布取代分布的理由。所以可以说:

当时，如果未知，则必须使用统计量；

在的条件下，选择统计量还是统计量可以根据使用者的偏好。

**9、显著性水平的含义：★注意与置信水平比较★**

**在假设检验中，它的含义是当原假设正确时却被拒绝的概率或风险，其实这就是假设检验中犯弃真错误的概率。**

**10、两总体均值之差检验统计量的确定:**

①总体方差，已知或未知:

在，已知的情况下，由抽样分布的理论可知，样本统计量服从分布；

而在，未知的条件下，样本统计量服从分布。

②，较大或较小:

当样本量，都较大时，如果总体方差，未知，可以用样本方差代替，这时，样本统计量近似服从分布，采用作为检验统计量是可行的。

但是当，不大时，如果，未知，就应该采用作为检验统计量。

**11、“不拒绝”与“接受”:**

从假设检验的原理看，不拒绝原假设意味着我们所构造的与原假设相矛盾的事件没有发生，但可能会有许多其他的与原假设矛盾的小概率事件，我们没有也无法证实所有的这些小概率事件不会发生，因此，我们把假设检验中出现接受的结果解释为“没有发现充足的证据反对”，或更严格地的解释为，“在显著性水平*α*下没有发现充足的证据反对”，而不是“接受原假设”，因为我们无法证明原假设是正确的。即接受备择假设一定意味着原假设错误；没有拒绝原假设并不能表明备择假设一定是错的。

**第七部分 拟合优度检验和独立性检验**

**1、统计量的特征:**

①；

②统计量的分布与自由度有关；

③统计量描述了观察值与期望值的接近程度。两者越接近，计算出的值就越小，反之计算出的值也越大。

**2、独立性检验:**

**目的:分析列联表中行变量与列变量是否相互独立。**

**原理:事件之间的独立性。**

**检验的自由度:**

**3、分布的期望值准则:**

①如果只有两个单元，则每个单元的期望频数必须是5或5以上；

②倘若有两个以上的单元，20％的单元的期望频数小于5，则不能使用检验。

**第八部分 方差分析**

**1、方差分析相比于假设检验的优点:**

①提高检验效率；

②将所有的样本信息结合到一起，增加了分析的可靠性。

**2、方差分析（14-15）、（16）、（17）、（20）**

**通过检验各总体的均值是否相等来判断分类型自变量对数值型因变量是否有显著影响。它是通过对数据误差来源的分析来判断不同总体的均值是否相等，进而分析自变量对因变量是否有显著性影响。**

**3、误差分解:**

来自水平内部的数据误差称为组内误差，它反映了一个样本内部数据的离散程度，组内误差只含有随机误差。

不同水平之间的数据误差称为组间误差。这种差异可能是由抽样本身形成的随机误差，也可能是由水平本身的系统性因素造成的系统误差。因此，组间误差是随机误差和系统误差的总和。

**4、方差分析中的基本假定（17）:**

**①（正态性）：每个总体(水平)都应服从正态分布；**

**②（方差齐性）：各个总体的方差必须相同；**

**③（独立性）：各观测值是相互独立的；**

**④ 各总体误差的期望值为零。**

**5、单因素方差分析的步骤:（17）、（20）**

**① 计算各水平的样本均值；**

1. **计算全部观测值的总均值；**
2. **计算各误差平方和；**
3. **计算统计量的值；**

**⑤ 做出决策。**

**6、**

是对随机误差和系统误差大小的度量，它反映了自变量对因变量的影响，也称自变量效应或因子效应；

是对随机误差大小的度量，它反映了除自变量对因变量的影响之外的其他因素对因变量的总影响，因此，也称为残差变量，它所引起的误差称为残差效应；

是对全部数据总误差程度的度量，它反映了自变量和残差变量的共同影响，因此它等于自变量效应加残差效应。

**7、三个平方和的自由度（14-15）**

**①为全部观测值的个数；**

**②为因素水平(总体)的个数；**

**③。**

**8、关系强度的测量:**

实际上只要组间平方和不等于零，就表明两个变量之间有关系(只是是否显著的问题)。当组间平方和比组内平方和大，而且大到一定程度时，就意味着两个变量之间的关系显著，大的越多，表明他们之前的关系就越强；反之，当组间平方和比组内平方和小时，就意味着两个变量之间的关系不显著，小的越多，表明他们之间的关系就越弱。

**9、多重比较检验步骤:**

①提出假设；

②计算检验统计量；

③计算；

④根据显著性水平做出决策。

**10、双因素方差分析步骤:**

①提出假设；

②构造检验统计量；

③做出统计决策。

**11、双因素方差分析与单因素方差分析:**

双因素方差分析中，误差平方和不包括两个自变量中的任何一个，从而减少了残差效应。而在分别做单因素方差分析时，将行因素作自变量是，列因素被包括在残差中，同样，将列因素作自变量时，行因素被包括在残差中。因此对于两个自变量而言，进行双因素方差分析要优于分别对两个因素进行单因素方差分析。

**第九部分 一元线性回归**

**1、函数关系:**

设有两个变量和，变量随变量一起变化，并完全依赖于，当变量取某个数值时，依据确定的关系取相应的值，则称是的函数，记为，其中称为自变量，称为因变量。

**2、★相关分析要解决的问题★：(注意与回归分析要解决的问题做对比)**

**① 变量之间是否存在关系；**

**② 如果存在关系，它们之间是什么样的关系？**

1. **变量之间的关系强度如何？**

**④ 样本所反映的变量之间的关系能否代表总体变量之间的关系？**

**3、★进行相关分析时对总体的假定★：(注意与回归分析的基本假定对比)**

**① 两个变量之间是线性关系；**

**② 两个变量都是随机变量。**

**4、相关系数:**

根据样本数据计算的度量两个变量之间线性关系强度的统计量。若相关系数是根据总体全部数据计算的，称为总体相关系数，记为；若是根据样本数据计算的，则称为样本相关系数，记为。

**5、相关系数的性质:**

①的取值范围是。当时，的取值完全依赖于，二者之间为函数关系；当时，的取值与无关，二者之间不存在线性相关关系。

②具有对称性；

③的数值大小与和的原点及尺度无关；

④仅仅是与之间线性关系的一个度量，他不能用于描述非线性关系。这意味着，只表示两个变量之间不存在线性相关关系，并不说明变量之间没有任何关系，它们之间可能存在非线性相关关系。

⑤虽然是两个变量之间线性关系的一个度量却不意味着与一定有因果关系。

**6、相关程度:**

①当时，可视为高度相关；

②当时，可视为中度相关；

③当时，可视为低度相关；

④当时，说明两个变量之间相关程度极弱，可视为不相关。

**7、相关系数的显著性检验:**

对的正态性假设具有很大的风险，因此通常情况下不采用正态检验，而采用费希尔提出的检验，该检验可以用于小样本，也可以用于大样本。

①提出假设:

②计算检验统计量；

③进行决策。

**8、★回归分析主要解决的问题★（17）：(注意与相关分析、方差分析对比)**

**①从一组样本数据出发，确定变量之间的数学关系式；**

**②对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响是显著的，哪些是不显著的；**

**③利用所求的关系式，根据一个或几个变量的取值来估计或预测另一个特定变量的取值，并给出这种估计或预测的可靠程度。**

**9、回归模型:**

描述因变量如何依赖于自变量和误差项的方程称为回归模型。

**10、★模型假定★（17）：(注意与相关分析、方差分析对比)**

**①因变量与自变量之间具有线性关系；**

**②在重复抽样中，自变量的取值是固定的，即假定是非随机的；**

**③误差项是一个期望值为的随机变量，即；**

**④对于所有的值，的方差都相同，这意味着对于一个特定的值，的方差都等于。即。**

**⑤误差项是一个服从正态分布的随机变量，且相互独立。即**

**11、回归方程:**

描述因变量的期望值如何依赖自变量的方程称为回归方程。

**12、回归分析的一般过程（17）:**

**①明确预测的具体目标，确定因变量和自变量；**

**②进行相关分析；**

**③建立预测模型；**

**④计算预测误差；**

**⑤进行显著性检验；**

**⑥确定预测值。**

**13、估计的回归方程中系数的解释（14-15）:**

**是估计的回归方程在轴上的截距；**

**是直线的斜率，表示每变动一个单位时，的平均变动值。**

**14、最小二乘法的思想（19）:**

**使因变量的观测值与预测值之间的离差平方和达到最小来估计和。**

**15、最小二乘估计的性质:**

①根据最小二乘法得到的回归直线能使离差平方和最小；

②由最小二乘法求得的回归直线可知和的估计量的抽样分布；

③在某些条件下，和的最小二乘估计量同其他估计量相比，其抽样分布具有较小的标准差。

**16、平方和的分解:**

因变量的取值是不同的，的取值的这种波动称为变差。变差的产生来自两个方面:一是由自变量的取值不同造成的；二是除以外的其他因素的影响。

称为回归平方和，反映了的总变差中由于与之间的线性关系引起的的变化部分，他是可以由回归直线来解释的的变差部分；

称为残差平方和，它是除了对的线性影响之外的其他因素引起的的变化部分，是不能由回归直线来解释的的变差部分。

**17、判定系数：**

**判定系数测度了回归直线对观测数据的拟合程度。**

**的取值范围是，越接近于1，表明回归平方和占总平方和的比例越大，回归直线与各观测点越接近，用的变化来解释值变差的部分就越多，回归直线的拟合程度就越好；反之，越接近于0，回归直线的拟合程度就越差。**

**在一元回归中，相关系数实际上是判定系数的平方根。**

**相关系数与回归系数的正负号是相同的。**

**18、判定系数的实际意义:**

在因变量的变差中有可以由自变量与因变量之间的线性关系解释。

**19、估计的标准误差:（20）**

**度量各实际观测点在直线周围的散布状况的统计量，他是均方残差的平方根。**

**它反映了估计的回归方程预测因变量时预测误差的大小。**

**20、线性关系的检验步骤:**

①提出假设；

②计算检验统计量；

③做出决策。

**21、回归系数的检验步骤:**

①提出检验；

②计算检验统计量；

③做出决策。

**22、回归分析结果的评价:**

①所估计的回归系数的符号是否与理论或事先预期的相一致；

②如果理论上认为与之间的关系不仅是正的，而且是统计上显著的，那么所建立的的回归方程也应该如此；

③回归模型在多大程度上解释了因变量取值的差异？

④考察关于误差项的正态性假定是否成立。

**23、残差:**

残差是因变量的观察值与根据估计的回归方程求出的预测值之差，用来表示。它反映了用估计的回归方程去预测而引起的误差。

**第十部分 多元线性回归**

**1、多元回归模型的基本假定（14-15）:**

**①误差项是一个期望值为的随机变量，即；**

**②对于自变量的所有值，的方差都相同，即；**

**③误差项是一个服从正态分布的随机变量，且相互独立，即，。正态性假定意味着因变量也是一个服从正态分布的随机变量。**

**2、多重判定系数:**

多元回归中的回归平方和占总平方和的比例，它是度量多元回归方程拟合程度的一个统计量，反映了因变量的变差中被估计的回归方程所解释的比例。

**3、调整的多重判定系数（19）:**

**由于增加自变量的个数总是使得因变量的变差中被回归方程所解释的比例增大，为避免增加自变量而高估，统计学家提出用样本量和自变量个数去调整。**

**4、线性关系的检验步骤:**

①提出假设；

②计算检验统计量；

③做出统计决策。

**5、回归系数检验的步骤:**

①提出假设；

②计算检验统计量；

③做出统计决策。

**6、多重共线性（19）:**

**当回归模型中两个或两个以上自变量彼此相关时，则称回归模型中存在多重共线性。**

**7、多重共线性可能导致的问题（19）:**

**①变量之间高度相关时，可能会使回归的结果混乱，甚至会把分析引入歧途。**

**②多重共线性可能对参数估计的正负号产生影响，特别是正负号有可能与预期的正负号相反。**

**8、多重共线性的判别:**

①模型中各对自变量之间显著相关；

②当模型的线性关系检验(检验)显著时，几乎所有的回归系数的t检验却不显著；

③回归系数的正负号与预期的相反；

④容忍度与方差扩大因子。容忍度越小，多重共线性越严重；方差扩大因子等于容忍度的倒数，方差扩大因子越大，多重共线性越严重，一般认为，方差扩大因子大于10时，存在严重的多重共线性。

**9、多重共线性问题的处理:**

①将相关的自变量从模型中剔除，使保留的自变量尽可能不相关；

②如果要在模型中保留所有自变量，那就应该:

ⅰ避免根据统计量对单个参数进行检验；

ⅱ对因变量的推断(估计或预测)限定在自变量样本值的范围内。

**10、选择自变量的原则:**

对统计量进行显著性检验，检验的根据是:将一个或一个以上的自变量引入回归模型中时，是否使残差平方和)显著减少。如果增加一个自变量使残差平方和显著减少，则说明有必要将这个自变量引入回归模型，否则就没有必要将这个自变量引入回归模型。

**11、变量选择的方法:（18）**

**①向前选择:**

**向前选择变量的方法是不停地向模型中增加自变量，直至增加自变量不能使得显著减少为止。**

**②向后剔除:**

**向后选择变量的方法是一直将自变量从模型中剔除，直至剔除一个自变量使得显著增加为止。**

**③逐步回归:**

**不停地增加变量并考虑剔除以前增加的变量的可能性，直至增加变量不会致使显著减小。逐步回归中，在前面步骤中增加的自变量在后面的步骤中有可能被剔除，而在前面步骤被剔除的自变量在后面的步骤中也可能重新进入模型。**

**第十一部分 时间序列分析和预测**

**1、时间序列的组成要素:**

**①趋势:**

**时间序列在长期内呈现出来的某种持续上升或持续下降的变动，也称长期趋势；**

**②季节性(季节变动):**

**时间序列在一年内重复出现的周期性波动；**

**③周期性(循环波动):**

**时间序列中呈现出来的围绕长期趋势的一种波浪形或震荡式变动。循环波动无固定规律，变动周期多在一年以上，且周期长短不一。**

**④随机性(不规则波动):**

**时间序列中除去趋势、周期性和季节性之后的偶然性波动。**

**2、时间序列的预测步骤:**

①确定时间序列所包含的成分，也就是确定时间序列的类型；

②找出适合此类时间序列的预测方法；

③对可能的预测方法进行评估，以确定最佳预测方案；

④利用最佳预测方案进行预测。

**3、时间序列预测方法:**

**①若时间序列不存在趋势且不存在季节性，则可以用平滑预测法，平滑预测法包括:简单平均法、移动平均法、指数平滑法；**

**②若时间序列不存在趋势但存在季节性，或时间序列存在趋势且存在季节性，则可以使用季节性预测法，季节性预测法包括:季节多元回归模型、季节自回归模型、时间序列分解；**

**③若时间序列存在趋势但不存在季节性，则可以使用趋势预测法，趋势预测法包括:线性趋势推测、非线性趋势推测、自回归预测模型。**

**4、复合型序列的分解预测步骤:**

①确定并分离季节成分:计算季节指数，以确定时间序列中的季节成分。然后将季节成分从时间序列中分离出去，即用每一个时间序列观察值除以相应的季节指数，以消除季节性；

②建立预测模型并进行预测:对消除了季节成分的季节指数建立适当的预测模型，并根据这一模型进行预测；

③计算最后的预测值:用预测值乘以相应的季节指数，得到最终的预测值。

**5、季节指数的计算方法:**

①计算移动平均值(如果是季度数据，则采用4项移动平均，月份数据则采用12项移动平均)，并对结果进行中心化处理，也就是将移动平均的结果再进行一次二项移动平均，即得出中心化移动平均值；

②计算移动平均的比值，也称为季节比率，即将序列的各观察值除以相应的中心化移动平均值，然后计算出各比值的季度(或月份)平均值；

③季节指数调整:将第二步计算的每个季节比率的平均值除以总平均值。

# 《概率论与数理统计》 盛骤 浙大四版

**第一章 概率论的基本概念**

**1、随机实验:**

①可以在相同条件下重复地进行；

②每次实验的可能结果不止一个，并且能事先明确实验的所有可能结果；

③进行一次实验之前不能确定哪一个结果会出现。

**2、频率:**

在相同条件下，进行了次实验，在这次实验中，事件发生的次数称为事件发生的频数。比值称为事件发生的频率，并记成。

**3、频率的性质:**

①非负有界性；

②规范性:样本空间，则事件的频率为；

③有限可加性。

**4、概率的公理化定义:**

**设 是随机实验，是它的样本空间。对于的每一事件赋予一个实数，记为，称为事件的概率，如果集合函数满足下列条件:**

**①非负性:对于每一事件，有；**

**②规范性:对于必然事件，有；**

**③可列可加性:设是两两互不相容的事件，即对于，有**

**5、★实际推断原理★：(与假设检验原理的联系，与置信区间原理的区别)（20）**

**概率很小的事件在一次实验中实际上几乎是不发生的。**

**6、★两事件独立★（16）：(与两变量相互独立的关系)**

**设是两事件，如果满足等式**

**则称事件相互独立，简称独立。**

**注：三事件相互独立、三事件两两独立、个事件相互独立、两个随机变量的独立。**

**第二章 随机变量及其分布**

**1、随机变量:**

设随机实验的样本空间为是定义在样本空间上的实值单值函数。称为随机变量。

**2、二项分布的性质（20）:**

**一般对于固定的及，二项分布具有这样的性质:当增加时，概率先是随之增加，直至达到最大值，随后单调减少。**

**3、泊松定理:**

**4、分布函数及其性质:（17）、（20）**

**设是一个随机变量，是任意实数，函数**

**称为的分布函数。**

**基本性质:**

**①是一个不减的函数；**

**② ，且；**

**③ ，即是右连续的。**

**5、均匀分布的意义:**

**在区间上服从均匀分布的随机变量，落在区间的任意等长度的子区间内的可能性是相同的。或者说它落在的子区间内的概率只依赖于子区间的长度而与子区间的位置无关。**

**6、指数分布的无记忆性（17）:**

**第三章 多维随机变量及其分布**

**1、二维随机变量的分布函数及其性质:（20）**

**设是二维随机变量，对于任意实数，二元函数:**

**称为二维随机变量的分布函数，或称为随机变量和的联合分布函数。**

**基本性质:**

**①是变量和的不减函数；**

**②，且**

**对于任意固定的，**

**对于任意固定的，**

**；**

**③，即关于右连续，关于右连续；**

**④ 对于任意，下述不等式成立:**

**。**

**2、★随机变量的相互独立★（16）、（19）：(与随机事件独立的关系)**

**设及，分别是二维随机变量的分布函数及边缘分布函数，若对于所有的和有**

**，**

**即**

**则称随机变量是相互独立的。**

**注：对于二维正态随机变量相互独立的充要条件是相关系数**

**3、随机变量的函数的分布:**

① 有限个相互独立的正态随机变量的线性组合仍然服从正态分布；

② 分布具有独立可加性；

③分布具有独立可加性。

**4、联合分布与边缘分布的关系:**

**对于而言，由的分布可以确定关于、关于的边缘分布。反之，由关于和关于的边缘分布一般是不能确定的联合分布的。只有当相互独立时，由两边缘分布能确定的分布。**

**5、随机变量的独立性是随机事件独立性的扩充。**

**第四章 随机变量的数字特征**

**1、数学期望的性质:**

① 设是常数，则有；

② 设是一个随机变量，C是常数，则有

；

③ 设是两个随机变量，则有

；

④ 设是相互独立的随机变量，则有

。

**2、方差及其性质:**

设是一个随机变量，若存在，则称

为的方差，记为或，即

。

性质:

① 设是常数，则；

② 设是随机变量，是常数，则有

；

③ 设是两个随机变量，则有

。

特别地，若相互独立，则有

④ 的充要条件是以概率1取常数，即

。

**3、切比雪夫不等式（20）:**

**4、不相关和独立（14-15）、（19）:**

**相关系数有时候也称为线性相关系数，它是一个可以用来描述随机变量的两个分量之间的线性关系紧密程度的数字特征。当较小时，的线性相关程度较差；当时，称不相关。不相关是指之间不存在线性关系，不相关，他们还可能存在除线性关系以外的关系。而相互独立是针对的一般关系而言的，即若相互独立，则之间不存在任何关系。**

**相互独立，则一定不相关；**

**相关，则一定不独立；**

**不独立，则有可能线性相关，也可能线性无关；**

**不相关，则有可能相互独立，也可能不相互独立。**

**特别地，对于二维正态随机变量不相关与相互独立是等价的**。

**第五章 大数定律及中心极限定理**

**1、大数定律:（20）**

**大数定律是叙述随机变量序列的前一些项的算数平均值在某种条件下收敛到这些项的均值的算数平均值。**

**2、中心极限定理:**

**中心极限定理是确定在什么条件下，大量随机变量之和的分布逼近于正态分布。**

**中心极限定理表明，在相当一般的条件下，当独立随机变量的个数不断增加时，其和的分布趋于正态分布。这一事实阐明了正态分布的重要性。也揭示了为什么在实际应用中会经常遇到正态分布，也就是揭示了产生正态分布变量的源泉。另一方面，它提供了独立同分布的随机变量之和的近似分布，只要和式中加项的个数充分大，就可以不必考虑和式中的随机变量服从什么分布，都可以用正态分布来近似。**

**3、弱大数定理(辛钦大数定理):**

**4、伯努利大数定理:**

**5、独立同分布的中心极限定理:**

**6、棣莫弗-拉普拉斯定理。**

**第六章 样本及抽样分布**

**1、样本:**

设是具有分布函数的随机变量，若是具有同一分布函数的，相互独立的随机变量，则称为从服从分布函数(或总体，或总体)得到的容量为的简单随机样本，简称样本。

**2、样本分位数:**

设有容量为的样本观察值，，样本分位数记为，它具有以下性质:

① 至少有个观察值小于或等于；

② 至少有个观察值大于或等于。

**3、箱线图（16）、（17）、（20）:**

**数据集的箱线图是由箱子和直线组成的图形，它是基于以下五个数的图形概括:最小值，第一四分位数，中位数，第三四分位数和最大值。它的作法如下:**

**① 画一水平数轴，在轴上标上，，，，。在数轴上方画一个上、下侧平行于数轴的矩形箱子，箱子的左右两侧分别位于，的上方。在点的上方画一条垂直线段。线段位于箱子内部。**

**② 自箱子左侧引一条水平线直至最小值；在同一水平高度自箱子右侧引一条水平线直至最大值。**

**4、★箱线图的性质★：(注意与其他数据图表展示的比较)**

**① 中心位置：中位数所在位置就是数据集的中心；**

**② 散布程度：全部数据都落在之内，在区间，，，的数据个数各约占。区间较短时，表示落在该区间的点比较集中，反之较为分散。**

**③ 关于对称性：若中位数位于箱子的中间位置，则数据分布较为对称。又若离的距离较离得距离大，则表示数据分布向左倾斜，反之表示数据向右倾斜，且能看出分布尾部的长短。**

**5、经验分布函数:**

**6、分布及其性质:**

**定义:**

**性质:**

**① 独立可加性；**

**② ；**

**7、分布及其性质:（18）、（19）**

**定义:**

**性质:**

**① 对称性；**

**② 当(有时也取)时，分布可用正态分布近似。**

**8、分布及其性质:（18）、（19）**

**定义:**

**性质:**

**若，则。**

**9、样本均值的抽样分布:**

**10、抽样分布定理:**

**第七章 参数估计**

**1、矩估计法（14-15）、（20）:**

**一般来说，样本矩依概率收敛于相应的总体矩，样本矩的连续函数依概率收敛于总体矩的连续函数，我们用样本矩作为相应的总体矩的估计量，而以样本矩的连续函数作为相应的总体矩的连续函数的估计量，这种估计方法称为矩估计法，其理论基础是大数定理。**

**2、极大似然估计法及其性质（14-15）:**

**固定样本观察值，在取值的可能范围内挑选使似然函数达到最大的参数值，作为参数的估计值，这种估计方法称为极大似然估计法。**

**性质：设的函数，具有单值反函数，。又假设是的概率分布中参数的最大似然估计，则是的极大似然估计，这一性质称为极大似然估计的不变性。**

**3、置信区间:**

置信水平为的置信区间并不是唯一的。(双侧等尾、双侧不等尾)

**第八章 假设检验**

**1、显著性检验:**

只对犯第Ⅰ类错误的概率加以控制，而不考虑犯第二类错误的概率的检验称为显著性检验。

**2、假设检验问题的步骤（18）、（20）:**

**① 根据实际问题的要求，提出原假设和备择假设；**

**② 给定显著性水平以及样本容量；**

**③ 确定检验统计量及拒绝域的形式；**

**④ 按{当为真拒绝取样；**

**⑤ 根据样本观察值做出决策，是接受原假设还是拒绝原假设。**

**3、置信区间与假设检验之间的关系（18）、（19）: 详见课本**

**4、值:**

**假设检验问题的值是由检验统计量的样本观测值得出的原假设可被拒绝的最小显著性水平。**

**对于任意给定的显著性水平:**

**① 若值，则在显著性水平下拒绝原假设；**

**② 若值，则在显著性水平下接受原假设。**

**值表示反对原假设的依据的强度，值越小，反对原假设的依据越强、越充分。**

**一般若值，称推断拒绝原假设的依据很强或称检验是高度显著的；若值，称推断拒绝原假设的依据是强的或称检验是显著的；若值，称推断拒绝原假设的理由是弱的或称检验是不显著的；若值，一般来说没有理由拒绝原假设。**

**5、两类错误的地位（14-15）、（19）:**

**在进行显著性检验时，犯第Ⅰ类错误的概率是由我们控制的，取得小，则概率{当为真拒绝就越小，这保证了当原假设为真时错误地拒绝原假设的可能性很小。这意味着原假设是受到保护的，也表明原假设和备择假设的地位不是对等的。**

**第九章 方差分析及回归分析**

**1、方差分析:（16）、（20）**

**方差分析就是通过对实验数据进行分析，检验方差相同的多个(多于两个)正态总体均值是否相等，用以判断各因素对实验指标的影响是否显著。**

**观察到的数据总是参差不齐的，用总偏差平方和来度量数据间总的变异(即离散程度)，将他分解为可追溯来源的部分变异(它是由随机误差引起的)与(它是由各水平效应的差异及随机误差引起的)之和。若比大的多，则有理由认为因素的各个水平对应的实验结果有显著差异，从而拒绝因素各水平对应的正态总体的均值相等这一原假设，这就是方差分析的思想。**

**2、方差分析的基本假定:**

**3、平方和分解:**

**4、平方和自由度:**

**5、平方和的统计特性:**

**6、回归分析中参数的极大似然估计与最小二乘估计的等价性:**

**7、回归效果不显著的原因**

**① 影响取值的，除及随机误差外还有其他不可忽略的因素；**

**② 与的关系不是线性的，而存在着其他的关系；**

**③ 与不存在关系。**

# 《概率论与数理统计》 梁之舜 中山三版

**第六章 抽样分布**

**1、顺序统计量:**

**设为总体的样本，现由样本建立个函数:**

**其中为这样的统计量，它的观察值为，为样本的观察值中由小到大排列之后的第位数值，则称为顺序统计量。**

**2、抽样分布定理:**

**3、柯赫伦分解定理:**

**第七章 估计理论**

**1、矩估计法的优缺点:**

**矩法估计是最古老的点估计方法，它直观而又简便，特别是在对总体的数学期望及方差数字特征作估计时，并不一定要知道的分布函数。但是矩法估计要求随机变量的原点矩存在，如柯西分布的原点矩不存在，那就不能用矩法了。再者，样本矩的表达式同总体分布函数的表达式无关，因而矩法还没有充分利用对参数所提供的信息。**

**2、最小方差无偏估计量:（19）**

**设为可估计函数的无偏估计量，若对于的任一无偏估计量，有**

**)**

**则称为的最小方差无偏估计量。最小方差无偏估计量简称为最优无偏估计量。**

**3、最小方差线性无偏估计**

**可估计函数的估计量称为他的最小方差线性无偏估计量，是指：如果**

**① (线性性)：为样本的线性函数；**

**② (无偏性)：；**

**③ (最优性)：对于满足①和②的任一估计量，对一切有**

**4、优效估计:**

**如果参数的函数的某个正规无偏估计量的方差达到不等式的下界，则称这个无偏估计量为的优效估计量(一致最小方差无偏估计,)。**

**5、有效率:**

设为参数的任一正规无偏估计量，为的优效估计量，则称

为正规无偏估计量的有效率。

**6、相合估计:**

**7、充分统计量:（18）**

**设总体的分布函数为为未知参数，，为样本，为不带有未知参数的统计量，若在给定统计量时，的条件分布与无关，则称为的充分统计量。**

**8、完备性统计量:**

**设总体的分布函数为为一随机变量，如果对一切，**

**成立时(即是无偏估计)，对于一切必有:**

**，**

**则称是完备的，若为样本，统计量的分布函数是完备的，则称为完备统计量。**

**9、区间估计:**

设为一给定的常数，满足，如果关系式:

成立，并用这个随机区间作为参数的估计，则称是参数的置信水平为的区间估计，称为显著性水平，分别为上、下置信限。

**第八章 假设检验**

**1、皮尔逊检验步骤:**

**① 抽取样本；**

**② 作直方图；**

**③ 计算理论频数；**

**④ 建立检验统计量；**

**⑤ 原假设的显著性检验。**

**2、两类错误:**

**3、两类错误的关系及其控制（14-15）:**

**第九章 回归分析与方差分析**

**1、回归模型的基本假定（14-15）:**

**① ；**

**②；**

**其中是未知参数，是单位矩阵，即:**

**① ；**

**② *D()＝；***

**③**

**上面的条件等价于:**

**①；**

**② ；**

**2、最小二乘法估计量的6条性质:**

# 《多元统计分析》 何晓群 第四版

**第一部分 聚类分析**

**1、指标类型的尺度:**

① 间隔尺度:变量用连续的量来表示；

② 有序尺度:指标用有序的等级来表示，有次序关系，但没有数量表示；

③ 名义尺度:指标用一些类来表示，这些类之间既没有等级关系，也没有数量关系。

**2、聚类分析的两个方向:**

① 对样品聚类，通常称为型聚类分析；

② 对指标聚类，通常称为型聚类分析；

**3、聚类分析的方法:**

① 系统聚类法:首先，将个样品看成类（一类包含一个样品），然后将性质最接近的两类合并成一个新类，得到类，再从中找出最接近的两类加以合并，变成类，如此下去，最后所有的样品均在一类，将上述并类过程画成一张图（称为聚类图）便可决定分多少类，每类各有哪些样品。

② 模糊聚类法:该方法多用于定性变量的分类。

③均值法：均值法是一种非谱系聚类法，它是把样品聚集成个类的集合。类的个数可以预先给定或者在聚类过程中确定。该方法可应用于比系统聚类法适用的大得多的数据组。

**4、类与类之间的距离:**

① 最短距离法；

② 最长距离法:选择最大的距离作为新类与其他类之间的距离，然后将类间距离最小的两类进行合并，一直合并到只有一类为止。

③ 类平均法:包括组内联结法和组间联结法。

④ 重心法；

⑤ 离差平方和法。

**5、分类的准则:**

**准则一：任何类都必须在临近各类中是突出的，即各类重心之间的距离必须大。**

**准则二：各类所包含的元素都不应过多。**

**准则三：分类的数目应该符合使用的目的。**

**准则四：若采用几种不同的聚类方法处理，则在各自的聚类图上应发现相同的类。**

**6、空间张度:**

｛最短距离法，重心法｝＜｛类平均法｝＜｛最长距离法，离差平方和法｝

**7、系统聚类法的步骤:**

**① 计算个样品两两间的距离；**

**② 构造个类，每个类只包含一个样品；**

**③ 合并距离最近的两类为一新类；**

**④ 计算新类与当前各类的距离，直到聚为一个类为止；**

**⑤ 画聚类图；**

**⑥ 决定分类个数和类。**

**8、模糊聚类分析计算步骤:**

**① 对原始数据进行变换。**

**② 计算模糊相似矩阵。**

**③ 建立模糊等价矩阵。**

**④ 进行聚类。**

**9、聚类分析的计算步骤:**

**① 分析所需要研究的问题，确定聚类分析所需要的多元变量；**

**② 选择对样品聚类还是对指标聚类；**

**③ 选择合适的聚类方法；**

**④ 选择所需的输出结果。**

**第二部分 判别分析**

**1、判别分析的基本要求:**

① 分组类型在两组以上，每组案例的规模必须至少一个以上；

② 解释变量必须是可测量的，这样才能够计算其平均值和方差，使其能合理的应用于统计函数；

**2、判别分析的假设:**

**① 每一个判别变量（解释变量）不能是其他判别变量的线性组合；**

**② 各组变量的协方差矩阵相等；**

**③ 各判别变量遵从多元正态分布，即每个变量对于所有其他变量的固定值有正态分布。**

**第三部分 主成分分析**

**1、主成分分析及其思想:（18）**

**主成分分析是利用降维的思想，在损失很少信息的前提下，把多个指标转化为几个综合指标的多元统计方法。**

**其思想是在保留原始变量尽可能多的信息的前提下达到降维的目的，从而简化问题的复杂性并抓住问题的主要矛盾。**

**2、主成分与原始变量的关系:**

**① 每一个主成分都是各原始变量的线性组合；**

**② 主成分的数目大大少于原始变量的数目；**

**③ 主成分保留了原始变量的绝大多数信息；**

**④ 各主成分之间互不相关。**

**3、主成分分析步骤（18）:**

**① 根据研究问题选取初始分析变量；**

**② 根据初始变量特性判断由协方差阵求主成分还是由相关阵求主成分；**

**③ 求协方差阵或相关阵的特征根与相应标准特征向量；**

**④ 判断是否存在明显的多重共线性，若存在，则回到第①步；**

**⑤ 得到主成分的表达式，并确定主成分个数，选取主成分；**

**⑥ 结合主成分对研究问题进行分析并深入研究。**

**第四部分 因子分析**

**1、因子分析及其思想:**

**因子分析是利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。相比主成分分析，因子分析更倾向于描述原始变量之间的相关关系，因此，因子分析的出发点是原始变量的相关矩阵。**

**因子分析的基本思想是根据相关性大小把原始变量分组，使得同组内的变量之间相关性较高，而不同组的变量间的相关性则较低。**

**2、因子分析的种类:**

① 研究变量之间的相关关系，则称为型因子分析；

② 研究样品之间的相关关系，则称为型因子分析。

**3、因子分析模型的假定:**

**① 是可观测随机向量，且均值向量**

**协方差阵且协方差阵和相关阵相等；**

**② 是不可观测的变量，其均值向量，协方差阵即向量的各个分量是相互独立的；**

**③ 与相互独立，且的协方差阵是对角方阵，即的各个分量之间也是相互独立的。**

**4、主成分分析与因子分析的区别:**

**① 因子分析的目的是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组合系数。主成分分析是从空间生成的角度寻找能解释诸多变量的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:各个公共因子之间不相关，特殊因子之间不相关，公共因子和特殊因子之间不相关。**

**④ 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。而主成分只能用主成分法提取。**

**⑤ 主成分分析中当给定协方差阵或者相关矩阵的特征根唯一时，主成分一般是固定的；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，随指定因子数量不同而结果不同。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 和主成分分析相比，由于因子分析可以使用因子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，则可以用主成分分析。**

**5、因子分析的步骤:**

**① 根据研究问题选取原始变量；**

**② 对原始变量进行标准化并求其相关阵，分析变量之间的相关性；**

**③ 求解初始公共因子及因子载荷矩阵，分析变量之间的相关性；**

**④ 因子旋转；**

**⑤ 计算因子得分；**

**⑥ 根据因子得分值进行进一步分析。**

**第五部分 对应分析**

**1、对应分析:**

对应分析是型因子分析与型因子分析的结合，它也是利用降维的思想来达到简化数据结构的目的，不过与因子分析不同的是，它同时对数据表中的行与列进行处理，寻求以低维图形表示表中行与列之间的关系。

**2、对应分析的步骤:**

**① 由原始列联表数据计算规格化的概率意义上的列联表；**

**② 计算矩阵；**

**③ 由或出发进行型因子分析或型因子分析，并由（或）型因子分析的结果推导出（或）型因子分析的结果；**

**④ 在二维图上画出原始变量的各个状态，并对原始变量相关性进行分析。**

**第六部分 典型相关分析**

**1、典型相关分析**

典型相关分析是研究两组变量之间相关关系的多元分析方法。他借助主成分分析降维的思想，分别对两组变量提取主成分，且使从两组变量提取的主成分之间的相关程度达到最大，而从一组内部提取的各主成分之间互不相关，用从两组分别提取的主成分的相关性来描述两组变量整体的线性相关关系。

**2、典型相关分析的步骤:**

**① 确定典型相关分析的目标；**

**② 设计典型相关分析；**

**③ 检验典型相关分析的基本假设；**

**④ 推导典型函数；**

**⑤ 解释典型变量；**

**⑥ 验证模型。**

**注：本部分内容选自《概率论与数理统计》茆诗松 第二版，主要为补充大数定律及中心极限定理方面的知识，这部分概念主要是针对复试计算题及证明题。**

**1、依概率收敛**

**设为一随机变量序列，为一随机变量，如果对任意的，有**

**则称序列依概率收敛于，记作。**

依概率收敛的含义是：对的偏差不小于任一给定量的可能性将随着的增大而愈来愈小。或者说，绝对偏差小于任意给定量的可能性将随的增大而愈来愈接近1。特别地，当为退化分布时，即，则称依概率收敛于，即。

设，是两个随机变量序列，是两个常数。如果

则有

(1)

(2)

(3)

**2、按分布收敛、弱收敛**

**设随机变量的分布函数分别为。若对于的任一连续点，都有**

**则称弱收敛于，记作**

**也称按分布收敛于，记作**

3、依概率收敛与按分布收敛的关系

(1)依概率收敛是一种比按分布收敛更强的收敛性，即有

(2)一般按分布收敛与依概率收敛是不等价的

(3)若为常数，则的充要条件是：。

**4、特征函数**

**设是一个随机变量，称**

**为的特征函数。**

因为，所以总是存在的，即任一随机变量的特征函数总是存在的。

当离散型随机变量的分布列为则的特征函数为

当连续型随机变量的密度函数为，则的特征函数为

与随机变量的数学期望、方差及各阶矩一样，特征函数只依赖于随机变量的分布，分布相同则特征函数也相同，所以也常称为某分布的特征函数。

注：常用分布的特征函数

(1)单点分布

，其特征函数为

(2)0-1分布 ，其特征函数为

(3)泊松分布 其特征函数为

(4)均匀分布 因为密度函数为

所以其特征函数为

(5)标准正态分布 因为密度函数为

所以其特征函数为

(6)指数分布 因为密度函数为

所以其特征函数为

**5、特征函数的性质**

(1)

(2)，其中表示的共轭。

(3)若，其中是常数，则

(4)独立随机变量和的特征函数为每个随机变量的特征函数的积，即设与相互独立，则

(5)若存在，则的特征函数可求次导数，且对，有

上式提供了一条求随机变量的各阶矩的途径，特别可用下式去求数学期望和方差

**6、伯努利大数定律**

**设为重伯努利实验中事件发生的次数，为每次实验中出现的概率，则对任意的，有**

伯努利大数定律说明：随着的增大，事件发生的频率大于预先给定的精度的可能性愈来愈小，要多小有多小。这就是频率稳定于概率的含义。

**7、切比雪夫大数定律**

**设为一列两两不相关的随机变量序列，若每个的方差存在，且有共同的上界，即则服从大数定律，即对任意的，成立**

切比雪夫大数定律只要求互不相关，并不要求他们是同分布的。因此，可以很容易的推出：如果是独立同分布的随机变量序列，且方差有限，则必定服从大数定律。伯努利大数定律是切比雪夫大数定律的特例。

**8、马尔可夫大数定律**

**马尔可夫条件：**

**对随机变量序列，若马尔可夫条件成立，则服从大数定律，即对任意的，成立**

**9、辛钦大数定律**

一个随机变量的方差存在，则其数学期望必定存在；但反之不成立，即若一个随机变量存在，则其方差不一定存在。前面的大数定律均假设随机变量序列的方差存在，下面的辛钦大数定律去掉了这一假设，仅设数学期望存在，但同时要求为独立同分布的随机变量序列。伯努利大数定律也是辛钦大数定律的特例。

**设为一独立同分布的随机变量序列，若数学期望存在，则服从大数定律，即对任意的，成立**

辛钦大数定律提供了求随机变量数学期望的近似方法。这样做的优点是可以不被去管的分布是怎样的。

由辛钦大数定律可以很容易的得到：如果为一独立同分布的随机变量序列，且存在，其中k为正整数，则服从大数定律。也就是可以将作为的近似值。

**10、泊松大数定律**

**设为次独立重复实验中，事件出现的次数，而事件在第次实验中出现的概率为则对任意的，有**

**注：大数定律讨论的是在什么条件下，随机变量序列的算术平均依概率收敛到其均值的算术平均。**

**11、林德伯格—莱维中心极限定理**

**设是独立同分布的随机变量序列，且存在，若记**

**则对任意实数，有**

此定理只假设独立同分布、方差存在，不管原来的分布是什么，只要n充分大，就可以用正态分布去逼近随机变量和的分布，所以这个定理有着广泛的应用，它同时揭示了这样的事实：测量误差近似服从正态分布。

**12、棣莫弗-拉普拉斯中心极限定理**

**设重伯努利实验中，事件在每次实验中出现的概率为，记为次实验中事件出现的次数，且记**

**则对任意实数，有**

它是专门针对二项分布的，因此称为“二项分布的正态近似”。与“二项分布的泊松近似”相比，一般在较小时，用泊松分布近似较好；而在和时，用正态分布近似较好。

**13、独立不同分布下的中心极限定理**

**林德伯格条件：对任意的，有**

**林德伯格中心极限定理：设独立同分布的随机变量序列妈祖林德伯格条件，则对任意的，有**

**14、李雅普诺夫中心极限定理：**

**设为独立随机变量序列，若存在*，*满足**

**则对任意的，有**

**15、样本**

为了了解总体的分布，我们从总体中随机的抽取个个体，及其指标值为，则称为总体的一个样本，称为样本容量，或简称为样本量，样本中的个体称为样品。

样本具有所谓的二重性：一方面，由于样本是从总体中随机抽取的，抽取前无法预知他们的数值，因此，样本是随机变量，用大写字母表示；另一方面，样本在抽取以后经观测就有确定的观测值，因此，样本有事一组数值，此时用小写字母表示是恰当的。

**简单随机样本的两个要求：**

**①样本具有随机性，即要求总体中每一个个体都有同等机会被选入样本，这就意味着每一样品与总体X有相同的分布。**

**②样本要有独立性，即要求样本中每一样品的取值不影响其他样品的取值，这就意味着相互独立。**

**16、格列文科定理**

设是取自总体分布函数为的样本，是其经验分布函数，当时，有

上述定理表明，当相当大时，经验分布函数是总体分布函数的一个良好的近似。

**17、样本均值的几个性质**

①若把样本中的数据与样本均值之差称为偏差，则样本所有偏差之和为0，即

②数据观测值与均值的偏差平方和最小，即在形如的函数中，最小，其中为任意给定常数。

③设是来自某个总体的样本，为样本均值

(1)若总体分布，则的精确分布为；

(2)若总体分布未知或不是正态分布，但，则较大时的渐进分布为，常记为，这里的渐进分布是指较大时的近似分布。

**18、次序统计量**

设是取自总体的样本，称为该样本的第i个次序统计量，他的取值是将样本观测值由小到大排列后得到的第i个观测值。其中称为最小次序统计量，称为最大次序统计量。

在简单随机样本中，是独立同分布的，而次序统计量则既不独立，分布也不相同。

**19、次序统计量的分布**

**设总体的密度函数为，分布函数为，为样本，则第个次序统计量的密度函数为**

**次序统计量的联合分布密度函数为**

**20、渐进正态性**

参数的相合估计称为渐进正态的，若存在趋于0的非负常数序列，使得依分布收敛于标准正态分布。这时也称服从渐进正态分布，记为，称为的渐进方差。

**21、均方误差**

一般而言，在样本量一定时，评价一个点估计的好坏使用的度量指标总是点估计值与参数真值的距离函数，最常用的函数是距离的平方。由于具有随机性，可以对该函数求期望，这就是下式给出的均方误差

均方误差是评价点估计的最一般的标准。自然，我们希望估计均方误差越小越好。均方误差有点估计的方差与偏差的平方两部分组成。如果是的无偏估计，则，此时用均方误差评价点估计与用方差是完全一致的。

**22、充分性原则**

如果无偏估计不是充分统计量的函数，则将之对充分统计量求条件期望可以得到一个新的无偏估计，该估计的方差比原来的估计的方差要小，从而降低了无偏估计的方差。换言之，考虑的估计问题只需要在基于充分统计量的函数中进行即可，该说法对所有的统计推断问题都是正确的，这就是所谓的充分性原则。

**23、有效估计一定是UMVUE。**

**《数据分析》梅长林 范金城 第二版**

**注：这部分只为复试作简单补充，本书绝大多数概念已经在前面几本书中提及。**

**1、偏度与峰度**

**偏度与峰度是刻画数据的偏态、尾重程度的度量。他们与数据的矩有关。**

**①偏度**

**偏度是刻画数据对称性的指标。关于均值对称的数据其偏度为0，右侧更分散的数据偏度为正，左侧更分散的数据偏度为负。**

**②峰度**

**当数据的总体分布为正态分布时，峰度近似为0；当分布较正态分布的尾部更分散时，峰度为正，否则峰度为负。当峰度为正时，两侧极端数据较多；当峰度为负时，两侧极端数据较少。**

**总体偏度是度量总体分布是否偏向某一侧的指标。对于对称的分布，偏度为0.**

**总体峰度是以同方差的正态分布为标准，比较总体分布尾部分散性的指标。**

2、若总体的分布未知，或者数据严重偏态，有若干异常值，此时计算数值特征需要用到次序统计量。

3、样本的分位数是总体分位数的相合估计

四分位差是度量样本分散性的重要数字特征，特别对于具有异常值的数据，它作为分散性的度量具有稳健性，因此它在稳健性数据分析中具有重要作用。

当样本是来自正态分布总体时，其总体上下四分位数为

四分位极差为

于是

它称为四分位标准差。当样本存在异常值时，标准差缺乏稳健性，可由上式得到标准差的一个稳健估计。

4、三均值

样本均值与中位数都是描述数据集中位置的数字特征。样本均值利用了样本的全部信息，而中位数仅仅利用了数据分布中的部分信息。正常情况下，使用样本均值比使用中位数描述数据的集中位置更优。然而，当数据中心存在异常值时，样本均值缺乏稳健性，而中位数具有很强的稳健性。考虑到要充分利用样本信息，又要具有较强的稳健性，可以使用三均值作为数据集中位置的数字特征，三均值的计算公式为：

5、上、下截断点

在探索性数据分析中，有一种判断数据为异常值的简便方法。称

为数据的上、下截断点。大于上截断点的数据为特大值，小于下截断点的为特小值。两者皆为异常值。

6、经验分布函数

①经验分布函数是单调不减的阶梯函数。

②在处的跃度为，若重复次，则跃度为。

③经验分布函数是总体分布函数的相合估计。

7、QQ图

现假定总体分布为正态分布。对于样本，其次序统计量为。设是标准正态分布的分布函数，是其反函数。对应正态分布的QQ图是由以下的点构成的散点图

若样本数据近似于正态分布，在QQ图上这些点近似地在直线

附近，此直线的斜率为标准差，截距为均值。所以利用QQ图可以做直观的正态性检验。

8、茎叶图与直方图

①茎叶图与直方图一样，可以直观地看出数据的分布状况。茎叶图用了所有的数据，没有信息损失。从茎叶图分析，可大致直观地看出这批数据是否接近对称，分散性如何，是否有异常值，数据中是否有间隙等。

②利用茎叶图，很自然地可以对所有数据排序。从茎叶图可以看出有原始数据得到的次序统计量。

**有关统计量的一些补充**

**注：这部分主要是针对复试计算题、证明题，题目不会直接考察，但有些题目会用到相关的概念和性质等，需要熟练掌握。**

1. 参数函数，若它存在无偏估计，则称参数函数为可估计的.

注：① 若记为参数函数的无偏估计全体，则为可估计的就等价于是非空的.

② 分布的均值就是可估计的，但并非每个参数函数都是可估计的.

例：容量为1的样本，为已知而为未知参数，就是不可估计的.

1. 若为参数函数的无偏估计，且对的任一无偏估计都成立

则称为参数函数的一致最小方差无偏估计，简记为.

注：① 若记为参数函数的无偏估计全体，则一致最小方差无偏估计就满足：

即一致最小方差无偏估计是无偏估计类中方差最小的那个.

② 参数函数的一致最小方差无偏估计若存在必唯一，即若同时为的一致最小方差无偏估计，则

1. 信息量通常用来表示样本中包含关于总体参数的信息量.

注：① 随着样本量的增加，样本所包含的信息量也成比例的增加.

② C-R不等式表明，样本包含参数的信息越多，估计方差的下界越小.

③ 若统计量为样本分布族的充分统计量，则与包含参数的信息量相同，即.

④ 若存在，且关于可在积分号下微分两次，则

⑤若记为参数函数的无偏估计全体，则C-R不等式表明在满足定理条件时，

若达到了下界，则它必是.

1. 若分布族与参数空间满足光滑性条件，参数函数不恒为常数，则存在的无偏估计量，满足光滑性条件且其方差达到C-R下界的充要条件可表为下列形式，

且，关于为可微的，这时必有

注：对于所规定的指数族分布，仅形式为的估计量作为其均值的无偏估计方差达到C-R下界.同样，由其规定的指数族分布，仅形式为

的参数函数存在方差达到C-R下界的无偏估计.若不计线性变换的差别，一个指数族分布至多只有一个参数函数存在这样的估计量.

1. 对参数函数的机遇样本量为的样本无偏估计量，又表示总体分布关于的信息量，

则称为估计量的效率.的估计称为有效的.

注：有效估计量一定是一致最小方差无偏估计.

1. 设是参数函数的无偏估计，统计量是分布族的充分统计量，则

i）是的无偏估计.

ii）且等式成立的充要条件是

注：① 定理表明若存在分布族的充分统计量，则对已有的无偏估计可利用其关于充分统计量的条件期望使估计量更有效.

② 若分布族存在充分统计量，为了寻找有效的无偏估计，可局限于在基于的无偏估计量中去寻找.

1. 分布族称为完备的，若对X的任一函数，由

可推出

上式表明，是恒为0的参数函数的无偏估计。所以分布族是完备的就意味着0的无偏估计是唯一的，这个就是恒为0的统计量。

对于统计量，它是样本的函数，若样本的可能分布族为，统计量有相应的分布族。若为完备的分布族，即

则称统计量为完备的。

上式表明，若参数函数，则依赖于的的无偏估计必是恒为0的统计量。由此可进一步推出，如果为分布族的完备统计量，那么，对一般的参数函数依赖于的无偏估计若存在，则必是唯一的。

1. 设的分布族为下列形式指数族

又若集合包含中的一个维小球，则为完备统计量.

注：① 由上述方法可以写出的完备统计量.

② 若统计量是完备的，则其函数也是完备的.

③ 对分布族，设分布族中支集与参数无关，，若分布族

为完备的，则分布族也是完备的.

9、 i) 若是分布族参数的充分完备统计量；

ii) 为参数函数的方差有限的无偏估计，

则为的一致最小方差无偏估计.

注：①若是分布族的充分完备统计量，估计量为的函数，，则就是的一致最小方差无偏估计.与有效估计相比，对于一个单参数指数型分布族，至多存在一个估计量及其线性函数为有效估计，但可以有许多不同的估计量分别是他们均值的.

10、 基于样本量的样本的参数函数的估计量若满足

则称为相合的*.*

注：样本阶原点矩是总体阶原点矩的相合估计量，样本阶中心矩是总体阶中心矩的相合估计量.

11、 若参数函数)的估计量满足

则是的相合估计.

12、 若参数函数的基于样本的估计量满足：

则称是的强相合估计.

注：强相合性包含了相合性.样本矩是总体相应矩的强相合估计，经验分布函数是总体分布函数的强相合估计.

13、 设样本取自分布族，为基于样本的统计量，若存在，使

则称统计量为渐进正态分布的，记为.

以上也记为

14、 若统计量为渐进正态分布的，则是的相合估计量.

注：渐进正态性是比相合性更强的一个要求.渐进正态性的意义在于他给出了统计量的近似分布，所以当样本量足够大时，就可以用它的近似分布代替原有的分布进行涉及分布的推断，而且这一近似分布由于样本原有的分布类型无关.

15、设总体分布族为离散概率或连续密度族，为参数直线上的开区间，且满足：

1)存在，且

2)当时，为不同的分布.

若为取自满足上述条件的总体，则对任何，当样本容量足够大时，似然方程

存在解，它满足

注：①上式表明是的强相合估计.

②在许多使用的情况中，最大似然估计是似然方程的唯一解，这是上述定理可用于推出的强相合性.

③一般地，似然方程的解不一定是唯一的，似然方程的解也不一定是最大似然估计，所以上述定理虽然给出了似然方程存在强相合解，但没有证明最大似然估计的存在性与强相合性.

④似然方程的相合解是渐进正态的，并且是渐进最优的.