

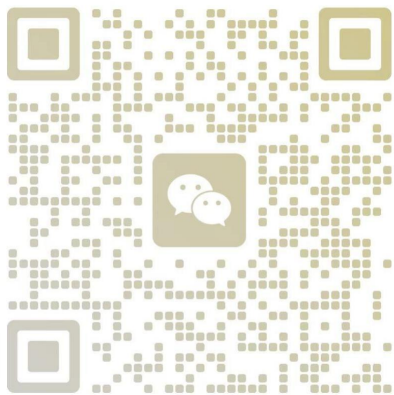
国家开放大学

期末考试考前复习资料

1318

《社会学》

伯仲教育出品



伯仲教育 国开期末 复习资料

《社会统计学》期末考试题 适用：【国开电大】【笔试】

题型： 单选(81) 简答(21) 名词解释(24) 计算题(17)

ps：资料考前整理，只供大家复习使用

单选(81)一

- 1、111、114、115，下列哪种图形最适合描述这些数据？（ ） -->D. 线图
- 2、90、92、72. 那么该班考试成绩的中位数是（ ）。 -->A. 89
- 3、残差平方和（SSE）反映了 y 的总变差中（ ） -->B. 除了 x 对 y 的现有影响之外的其他因素对 y 变的影响
- 4、从含有 N 个元素的总体中，抽取 n 个元素作为样本，同时保证总体中每个元素都有相同的机会入选样本，这样的抽样方式称为（ ）。 -->A. 简单随机抽样
- 5、从两个总体中各选取了 6 个观察值，得到组间平方和为 234. 组内平方和为 484. 则组间方差和组内方差分别为（ ）。 -->C. 234, 48. 4
- 6、从两个总体中共选取了 8 个观察值，得到组间平方和为 432. 组内平方和为 426. 则组间均方和组内均方分别为（ ） -->A. 432. 71
- 7、当我们用图形描述甲乙两地区的人口年龄结构时，适合选用哪种图形（ -->A. 环形图
- 8、对消费的回归分析中，学历、年龄、户口、性别、收入都是自变量，其中收入的回归系数为 1.6，这表明（ ）。 -->A. 收入每增加 1 元，消费增加 1.6 元
- 9、对于线性回归，在因变量的总离差平方和中，如果残差平方和所占比例越大，那么两个变量之间（ ） -->B. 相关程度越小

- 10、对于右偏分布，平均数、中位数和众数之间的关系是（ ）。 -->A. 平均数>中位数>众数
- 11、对于左偏分布，平均数、中位数和众数之间的关系是（ ）。 -->C. 众数>中位数>平均数
- 12、方差分析的目的是（ ）。 -->D. 研究各分类自变量对数值型因变量的影响是否显著
- 13、根据一个具体的样本求出的总体均值 90%的置信区间（ ）。 -->A. 以 90%的概率包含总体均值
- 14、根据一个具体的样本求出的总体均值 95%的置信区间（ -->A. 以 95%的概率包含总体均值
- 15、根据一个样本均值求出的 90%的置信区间表明（ ）。 -->C. 总体均值有 90%的概率会落入该区间内
- 16、根据一个样本均值求出置信水平为 90%的置信区间表明（ ）。 -->C. 总体均值有 90%的概率会落入该区间内
- 17、根据一个样本均值求出的 90%的置信区间表明（ ）。 -->C. 总体均值有 90%的概率会落入该区间内
- 18、回归平方和（SSR）反映了 y 的总变差中（ ）。 -->A. 由于 x 与 y 之间的线性关系引起的 y 的变差
- 19、经验法则表明，当一组数据正态分布时，在平均数加减 1 个标准差的范围之内大约有（ ）。 -->C. 95%的数据
- 20、离散系数的主要目的是（ ） -->D. 比较多组数据的离散程度
- 21、两个定类变量之间的相关分析可以使用（ ）。

答： A. λ 系数

- 22、两组数据的平均数不相等，但是标准差相等。那么（ ）。 -->C. 平均数大的，离散程度小
- 23、某班级 10 名同学期末统计课考试分数分别为 76、93、95、80、92、83、88、90、92、72. 那么该班考试成绩的中位数是（ ）。 -->A. 89
- 24、某班级学生平均每天上网时间可以分为以下六组：1) 1 小时及以下；2) 1-2 小时；3) 2-3 小时；4) 3-4 小时；5) 4-5 小时；6) 5 小时及以上，则 5 小时及以上这一组的组中值近似为（ ）。 -->C. 5.5 小时
- 25、某班级学生平均每天上网时间可以分为以下六组：(1) 1 小时及以下；(2) 1 — 2 小时；(3) 2 — 3 小时；(4) 3 — 4 小时；(5) 4 — 5 小时；(6) 5 小时及以上，则“5 小时及以上”这一组的组中值近似为（ ）。 -->C. 5.5 小时
- 26、某班级学生期末统计学考试平均成绩为 82 分，标准差为 5 分。如果已知这个班学生的考试分数服从正态分布，可以判断成绩在 77 — 87 之间的学生大约占全班学生的（ ） -->A. 68%
- 27、某班级有 100 名学生，为了了解学生消费水平，将所有学生按照学习成绩排序后，在前十名学生中随机抽出成绩为第 3 名的学生，后面依次选出第 13、23、33、43、53、63、73、83、93 名共九名同学进行调查。这种调查方法属于（ ）。 -->D. 系统抽样
- 28、某班级有 60 名男生，40 名女生，为了了解学生购书支出，从男生中抽取 12 名学生，从女生中抽取 8 名学生进行调查。这种调查方法属于（ ）。 -->C. 分层抽样
- 29、某单位对该厂第一加工车间废品率的估计高达 10%，而该车间主任认为该比例 偏高。如果要检验该说法是否正确，则假设形式应该为（ ）。

答： $H_0: p \leq 0.1$ $H_1: p > 0.1$

- 30、某单位对该厂第一加工车间废品率估计高于 13%，而该车间主任认为该比例偏高。如果要检验该说法是否正确，则假设形式应该为（ ）。

组别	按年龄分组(岁)	工作人员数(人)
1	20~24	6
2	25~29	14
3	30~34	24
4	35~39	18
5	40~44	12
6	45~49	18
7	50~54	14
8	55~59	6
	合计	112

答案：C,第3组

51、下表为某专业一年级学生平均每周上网时间的频率分布表，按照向上累积的方法计算第5组的累积频率是（）。

组别	按年龄分组(岁)	工作人员数(人)
1	20~24	6
2	25~29	14
3	30~34	24
4	35~39	18
5	40~44	12
6	45~49	18
7	50~54	14
8	55~59	6
	合计	112

答：D.66.1%

52、下列变量属于数值型变量的是（）。-->A. 工资收入

53、下列哪种情况不适合用方差分析（）。-->C. 年龄对收入的影响

54、下列四种变量属于数值型变量的是（）。-->A. 工资收入

55、下面哪一项不是方差分析中的假定（）。-->D. 各总体的方差等于

56、先将总体按某标志分为不同的类别或层次，然后在各个类别中采用简单随机抽样或系统抽样的方式抽取子样本，最后将所有子样本合起来作为总样本，这样的抽样方式称为（）-->D. 分层抽样

57、学校后勤集团想了解学校 42000 学生的每月生活费用，从中抽取 3200 名学生进行调查，以推断所有学生的每月生活费用水平。这项研究的总体是（）。-->A. 42000 名学生

58、一班学生的平均体重为 55 千克，二班学生的平均体重为 52 千克，两个班级学生体重的标准差均为 5 千克。那么（）-->B. 二班学生体重的离散程度大

59、已知某单位平均月收入标准差为 700 元，离散系数为 0.2. 那么他们月收入的均值为-->C. 3500

60、已知某单位平均月收入为 3500 元，离散系数为 0.2. 那么他们月收入的标准差为（）。-->A. 700

61、已知某单位职工平均每月工资为 3000 元，标准差为 500 元。如果职工的月收入是正态分布。可以判断月收入在 2500 元-3500 元之间的职工人数大约占总体的。-->B. 68%

62、以下关于条形图的表述，不正确的是（）。-->B. 条形图的矩形通常是紧密排列的

63、以下关于因变量与自变量的表述不正确的是（）。-->C. 自变量的变化是以因变量的变化为前提

64、用样本统计量的值构造一个置信区间，作为总体参数的估计，这种方法称为（）-->B. 区间估计

65、用样本统计量的值直接作为总体参数的估计值，这种方法称为（）。-->A. 点估计

66、用于说明回归方程中拟合优度的统计量主要是（）-->D. 判定系数

67、有甲、乙两人同时打靶，各打 10 靶，甲平均每靶为 8 环，标准差为 2；乙平均每靶 9 环，标准差为 3. 以下甲、乙两人打靶的稳定性水平表述正确的是（）。-->B. 甲的离散程度小，稳定性水平高

68、在方差分析中，某一水平下样本数据之间的误差称为（-->A. 组内误差

69、在回归方程中，若回归系数等于 0，这表明（）。-->B. 自变量 x 对因变量 y 的影响是不显著的

70、在假设检验中，不拒绝虚无假设意味着（）。-->D. 没有证据证明虚无假设是错误的

71、在假设检验中，如果所计算出的 P 值越小，那么检验的结果（）。-->A. 越显著

72、在假设检验中，虚无假设和备择假设（）-->C. 有且只有一个成立

73、在某市随机抽取 10 家企业，7 月份利润额(单位:万元)分别为 72.0, 63.1, 20.0, 23.0, 54.7, 5、54.3, 23.9, 25.0, 26.9, 29.0, 那么这 10 家企业 7 月份利润额均值为（-->A. 39.19

74、在频数分布表中，细率是指-->C 各组频数与总频数之比

75、在频数分布表中，将各个有序类别或组的百分比逐级累加起来称为（）。-->C. 累积频率

76、在频数分布表中，某一小组中数据个数占总数据个数的比例称为（）-->A. 频率

77、在因变量的总离差平方和中，如果回归平方和所占的比例越大，则两变量之间（）。-->A. 相关性越高

78、在因变量的总离差平方和中，如果回归平方和所占的比例越小，则自变量和因变量之间（）-->B. 相关程度越低

79、正态分布中， σ 值越小，则（）。-->A. 离散趋势越小

80、正态分布中，均值相等时 σ 值越小，则（）-->A. 离散趋势越小

81、中心极限定理认为不论总体分布是否服从正态分布

中心极限定理认为不论总体分布是否服从正态分布，从均值为 μ 、方差为 σ^2 的总体中，抽取容量为 n 的随机样本，当 n 充分大时（通常要求 $n \geq 30$ ），样本均值的抽样分布近似服从均值为 μ 、方差为 σ^2/n 的正态分布。（）

答：C. $\mu, \sigma^2/n$

简答(21)一

1、按测度水平，变量可分为哪几类？举例说明。...

2、等距分组和不等距分组有什么区别？请举例说明...

3、调查方法主要包括哪几种？简要说明各种方法的...

4、概率抽样方法主要包括哪几种？...

5、简述按照测量水平区分的四类变量，并举例说明...

6、简述什么是回归分析？其作用是什么？...

7、简述什么是简单回归分析？其作用是什么？...

8、简述什么是简单一元线性回归分析？其作用是什...

9、简述相关系数的取值与意义。

- 10、简要举例说明在分析双变量的关系时, t 检验和...
- 11、简要说明卡方的拟合优度检验和独立性检验的...
- 12、简要说明如何对配对样本进行 t 检验。...
- 13、简要说明社会研究过程。
- 14、举例说明什么是自变量和因变量, 二者之间是什...
- 15、劣简述众数、对位数和平均数作为测度中心趋...
- 16、判断以下随机变量是定性变量还是定量变量。...
- 17、判断以下随机变量是定性变量还是定量变量, 如...
- 18、如何对配对样本进行 t 检验。
- 19、什么是简单(一元线性)回归分析? 其作用是什么? ...
- 20、条形图和直方图有什么区别?
- 21、一元线性回归模型中有哪些基本假定? ...

1、按测度水平, 变量可分为哪几类? 举例说明。

答案: 按测度水平, 变量可分为哪几类? 并举例说明。按照测度水平, 变量可以分为定类变量、定序变量、定距变量和定比变量四类。一是定类变量: 当变量值的含义仅表示个体的不同类别, 而不能说明个体的大小、程度等其它特征时, 这种变量称为定类变量。定类变量是最高层次的变量。例如性别可以分为男、女两类。二是定序变量: 当变量值的含义不仅表示个体的不同类别, 还可以区分个体之间大小、程度等序次差异时, 这种变量称为定序变量。受教育程度是定序变量, 可以分为文盲、小学、初中、高中、大专及以上等。三是定距变量: 当变量值不仅可以将个体区分为不同类型并进行排序, 而且可以确定不同类别之间的数量差别和间隔差距时, 这样的变量称为定距变量。定距变量具有测量单位, 这些测量单位具有不变的相等区间的标准, 使个体之间的比较更具客观性。智商、温度等都是定距变量。四是定比变量: 除对上述三种变量的全部特征外, 还可以计算两个变量值之间的比值时, 这样的变量称为定比变量。定比变量是最高层次的变量。在社会调查中, 很多情况下我们使用的都是定比变量, 例如身高、年龄、收入、一个地区的人口数、某产品的生产量等。

2、等距分组和不等距分组有什么区别? 请举例说明。

答案: 等距分组和不等距分组有什么区别? 请举例说明。一是在对数据进行分组时, 如果各组组距相等, 则称为等距分组。例如, 分析某班同学期末统计课成绩时, 假如最低分为 73 分, 最高分为 98 分, 以 5 分为组距进行分组, 分为 70-75 分, 75-80 分, 80-85 分, 85-90 分, 90-95 分, 95-100 分。二是如果各组组距不相等, 则称为不等距分组。例如, 在分析人口时, 往往将人口分为婴幼儿组(0-6 岁), 少年儿童组(7-17 岁), 中青年组(18-59 岁), 老年人组(60 岁及以上), 该分类中各组组距不相等, 这就是不等距分组。

3、调查方法主要包括哪几种? 简要说明各种方法的优缺点。

答案: 调查方法主要包括哪几种? 简要说明各种方法的优缺点。一是问卷法问卷法是调查者根据一定的目的和要求, 采用事先设计好的问卷, 向被调查者了解情况, 征询意见的一种方法。问卷法广泛应用于社会学、人口学、教育学、管理学、心理学等领域中。问卷调查的主要优点是: 获得的资料便于进行定量分析; 节约时间、人力和经费; 应用范围广; 结果具有较高的信度。不足之处: 一是变量间的因果关系方面相对较弱; 对事物的理解和解释的深人性及效度相对较差; 所得资料比较表面化、简单化。二是访谈法访谈法是指由调查员直接对被调查员进行访问和交谈, 并记录调查结果的方法。访谈法是以口头语言为中介、调查者与被调查者面对面的交谈和互动的过程。访谈法的主要优点是: 调查的应答率较高; 调查资料的质量较高; 调查对象的适用范围较高。缺点主要有: 访谈员的语言、表情等可能会影响被调查者的回答; 对访谈员的要求比较高, 匿名性较差; 调查的时间比较长、费用比较高。三是观察法观察法是指研究者深入到所要研究对象的生活背景中, 在实

际参与研究对象的日常生活的过程中直接记录研究对象的有关资料的方法。观察法的主要优点是: 能获得更真实的结果; 调查结果不受研究对象的意愿和回答能力的影响; 可以更快更准确的收集到想要的的数据。观察法的缺点主要有: 观察到的只是行为, 而对于心理、态度、观点等没法公开的只能是通过观察到的行为推测, 而不能直接得到; 观察所得到的资料往往无法重复; 对观察者的要求较高, 观察者必须具备较强的敏感性、领悟能力。四是典型组讨论典型组讨论是由 8-12 个具有某些共同特征或经历的人, 在一个经过专门培训的主持人的引导下、围绕着某一专门话题进行互动小组讨论的一种定性调查方式。(1 分) 典型组讨论的主要优点是: 能够从一组人口中比较快得获得丰富的信息; 所得的信息更加丰富和深入, 分析和报告更加生动; 鼓励参与者互相回应和交流, 为对解差异和不同视角、确认答案提供对渠道; 典型组讨论的方式弹性比较大、灵活性比较强, 适用面广。缺点主要有: 参与者具有一定的偏差性; 实施比较复杂, 参与者的募集比较麻烦; 参与者之间的互动可能会产生不利的影响; 讨论内容和答案偏重于发现, 无暇确认信息; 得到的信息在汇总、分析、解释方面面临困难和复杂性。

4、概率抽样方法主要包括哪几种?

答案: 概率抽样方法主要包括哪几种? 概率抽样主要包括简单随机抽样、系统抽样、分层抽样、整群抽样和多阶段抽样五种。简单随机抽样是从包括总体 N 个单位的抽样框中随机的、一个一个的抽取 n ($n < N$) 个单位作为样本, 每个单位入选样本的概率是相等的。系统抽样也称为等距抽样, 是将总体中所有单位按照某一标志排序后, 在规定的范围内随机抽取一个单位作为初始单位, 然后按照一定的相等距离抽取调查单位。分层抽样也称类型抽样, 它是先将总体中的所有单位按某种特征或标志(如年龄、性别、职业等)划分成若干类型或层次, 然后再在各个类型或层次中采用简单随机抽样或系统抽样的方法抽样一个子样本, 最后将这些子样本合起来构成总体的样本的方法。整群抽样是先将总体按照某种标志或特征划分为一些子群体, 然后从总体中随机抽取一些子群体, 再将这些抽出的若干小群体内的所有元素构成总体样本的方法。多阶段抽样也称多级抽样或分段抽样, 是根据抽样元素的隶属关系或层次关系, 将抽样过程分为几个阶段来进行。

5、简述按照测量水平区分的四类变量, 并举例说明。

答: (1) 定类变量: 当变量值的含义仅表示个体的不同类别, 而不能说明个体的大小、程度等其它特征时, 这种变量称为定类变量。(2 分) 例如: 性别。(0.5 分)
(2) 定序变量: 当变量值的含义不仅表示个体的不同类别, 还可以区分个体之间大小、程度等序次差异时, 这种变量称为定序变量。(2 分) 例如: 学历。(0.5 分)
(3) 定距变量: 当变量值不仅可以将个体区分为不同类型并进行排序, 而且可以确定不同类别之间的数量差别和间隔差距时, 这样的变量称为定距变量。(2 分) 例如: 智商。(0.5 分)
(4) 定比变量: 除了上述三种变量的全部特征外, 还可以计算两个变量值之间的比值时, 这样的变量称为定比变量。(2 分) 例如: 收入。(0.5 分)

6、简述什么是回归分析? 其作用是什么?

答: (1) 回归分析是通过一定的数学表达式将变量间的关系进行描述, 确定一个变量或几个变量的变化对另一个特定变量的影响, 是进行估计或预测的→种方法, 侧重于考察变量之间的数量伴随关系。(4 分)
(2) 回归分析的作用包括:

- ①从已知数据出发, 确定变量之间的数学关系式;(2 分)
 - ②对变量间的关系式进行统计检验, 并从影响某一变量的多个变量中找出影响显著的变量;(2 分)
 - ③利用所求出的关系式, 根据→个变量或多个变量的取值估计或预测另一个特定变量的取值。(2 分)
- #### 7、简述什么是简单回归分析? 其作用是什么?

(1) 回归分析是通过一定的数学表达式将变量间的关系进行描述，确定一个变量或几个变量的变化对另一个特定变量的影响，是进行估计或预测的一种方法，侧重于考察变量之间的数量伴随关系。(4分)

(2) 回归分析的作用包括：

①从已知数据出发，确定变量之间的数学关系式；(2分)

②对变量间的关系式进行统计检验，并从影响某一变量的多个变量中找出影响显著的变量；(2分)

③利用所求出的关系式，根据一个变量或多个变量的取值估计或预测另一个特定变量的取值。(2分)

简述什么是一元线性回归分析？其作用是什么？

答案：简单回归分析是通过一定的数学表达式将两个变量间的线性关系进行描述，确定自变量的变化对因变量的影响，是进行估计或预测的一种方法，侧重于考察变量之间的数量伴随关系。(或者简单回归分析是对具有线性相关关系的两个变量之间(其中一个为自变量，另一个为因变量)数量变化的一般关系进行分析，确定相应的数学关系式，以便进行估计或预测。)其作用包括：一是从数据出发，确定变量之间的数学关系式；二是对变量间的关系式进行统计检验，考察自变量是否对因变量有显著影响；三是利用所求关系式，根据自变量的取值估计或预测因变量的取值。1-1之间；相关系数的正负号表示两个变量相关关系的方向，“+”表示正相关“-”表示负相关；相关系数的绝对值表示相关关系的程度，绝对值越大，相关程度越大，即 r 越接近 1；反之，绝对值越小，及 r 越接近 0，相关程度越弱；相关系数 $r=0$ 时，只能说变量之间不存在线性相关，而不能说它们之间不相关。

9、简述相关系数的取值与意义。

答：相关系数的取值在-1~1之间(2分) 相关系数的正负号表示两个变量相关关系的方向，“+”表示正相关，“-”表示负相关(2分)；相关系数的绝对值表示相关关系的程度，绝对值越大，相关程度越大，即 r 越接近 1(2分)；反之，绝对值越小，及 r 越接近 0，相关程度越弱(2分) 相关系数 $r=0$ 时，只能说变量之间不存在线性相关，而不能说它们之间不相关(2分)。

10、简要举例说明在分析双变量的关系时，t 检验和卡方检验的主要区别。

答案：简要举例说明在分析双变量的关系时，t 检验和卡方检验的主要区别。分析双变量关系时，t 检验和卡方检验都是主要用于检验这两个变量之间是否存在显著关系。t 检验主要用于对一个为数值型变量、另一个为分类变量且只有两个类别的变量的双变量关系的统计显著性检验。卡方检验主要用于对两个分类变量之间的相关性进行统计检验，判断变量之间是否存在显著关系。例如，我们想考察收入与性别是否存在关系，或者两性的收入是否存在显著差异，可以用两独立样本 t 检验。如果我们想考察职业与性别是否存在关系，而职业和性别都是分类变量，那么可以用卡方检验考察不同性别之间职业是否存在显著差异。

11、简要说明卡方的拟合优度检验和独立性检验的含义。

答案：简要说明卡方的拟合优度检验和独立性检验的含义。 χ^2 检验用于分类变量之间关系的检验。当用于检验不同类别的目标量之间是否存在显著差异时，称为拟合优度检验。例如，不同职业的人群中对某项改革措施的支持率是否一致。 χ^2 检验还可用于判断两个分类变量之间是否存在联系。如果两个分类变量之间没有关系，则称为独立，我们用 χ^2 判断它们之间是否关联，这时称为独立性检验。例如，对性行为的态度是否与受教育程度有关。

12、简要说明如何对配对样本进行 t 检验。

答案：配对样本检验主要是判断不同的处理或试验结果是否有差异。配对样本的 t 检验用于检验两个相关的样本是否来自具有相同均值的总体。在对配对样本进行 t 检验时，首先计算两个样本中每个对应变量之间的差值；然后再检验其差值的均值是否为零。如果差值的均值接近零(在给定的

置信区间内)，说明两个总体均值在给定的置信水平上没有差异。如果差值的均值在置信区间外，则说明两个总体均值在给定的置信水平上有差异。

13、简要说明社会研究过程。

答案：简要说明社会研究过程。一项系统、完整的社会研究通常遵循着比较固定的过程，一般包括五个主要过程：提出问题；形成假设；收集数据；分析数据；检验假设。一是提出问题是社会研究的第一步，研究者应首先明确要做什么，或者说进行研究的目的是什么。二是研究问题往往是无法直接检验的，而是要把研究问题形成假设才可以进行检验。假设是对两个变量间关系的陈述，是对它们关系的尝试性回答。假设中的变量是相对具体的，是可以通过经验可测的。三是收集数据是社会研究的关键步骤。这一阶段要确定收集的变量、收集的方法、收集的总量等。这一阶段所需投入的人力、物力最多，遇到的问题也最多，因而需要进行很好的组织和管理。四是分析数据主要是对收集的原始数据进行系统的审核、整理、归类、统计和分析。我们既要进行描述性统计，说明数据的总体情况，又要进行推断性统计，从收集的样本状况来推断研究总体的状况。五是完成对数据的统计分析后，我们要根据对数据的分析来检验我们之前的假设和理论。当结果支持我们的假设时，我们就对研究问题得到对解释；当结果不支持我们的假设时，要回过头来修改我们的问题和假设，开始新一轮检验。

14、举例说明什么是自变量和因变量，二者之间是什么关系？

答案：举例说明什么是自变量和因变量，二者之间是什么关系？一是我们将引起其他变量变化的变量称为自变量，而将由于其他变量的变化而导致自身发生变化的变量称为因变量。二是自变量与因变量之间的关系不仅仅是先后关系，还必须是因变量的变化是以自变量的变化为前提。三是例如，随着受教育程度增加，收入也增加，那么这两者就是因果关系，其中受教育程度是自变量，收入则是因变量。

15、简述众数、对位数和平均数作为测量中心趋势的指标所适用的效据类型

答案：简述众数、中位数和平均数作为测量中心趋势的指标所适用的数据类型。众数是对中心趋势低层次的侧及。是适用范围最大的一种测量方式，适用于任何类型的数据，即定类、定序、定距和定比数据都可以用众数来测其中心趋势。中位数适用于除对定类数据外的所有其他类型数据因为中位数是处于中间位置的变云值，而定类数据不存在类别的先后次序或大小。因而中位数不适用于定类数据。平均数是集中趋势主要的测量指标，但是平均数仅适用于定距效据和定比数据，定类数据和定序数据计算平均数没有意义。一般说来，定类数据的中心趋势测量指标是众数，定序数据的中心趋势测量指标是中位数，而数值型数据(定距和定比数据)的中心趋势测最指标是平均数。

16、判断以下随机变量是定性变量还是定量变量.如果是定量变量.确定是离散变量还是连续变量。

一是网络供应商的姓名：定性二是每月的网络服务费：定量、连续变量三是每月上网时间：定量、连续变量四是上网的主要目的：定性五是上月网上购物的次数：定量、离散变量。

17、判断以下随机变量是定性变量还是定量变量,如果是定量变量,确定是离散变量还是连续变量。

一是网络供应商的姓名：定性变量。二是每月的网络服务费：定量变量，连续变量。三是每月上网时间：定量变量，连续变量。四是上网的主要目的：定性变量。五是上周收到的电子邮件数量：定量变量，离散变量。六是每月用于网上购物的金额：定量变量，连续变量。七是上月网上购物的次数：定量变量.离散变量。八是使用的电脑的品牌：定性变量。九是上网是否玩游戏：定性变量。十是电脑是否带有光盘刻录机：定性变量。

18、如何对配对样本进行 t 检验。

答案：如何对配对样本进行 t 检验。配对样本检验主要是判断不同的处理或试验结果是否有差异。配对样本的 t 检验用于检验两个相关的样本是否来自具有相同均值的总体。在对配对样本进行 t 检验时，首先计算两个样本中每个对应变量之间的差值；然后再检验其差值的均值是否为零，如果差值

的均值接近零(在给定的置信区间内), 说明两个总体均值在给定的置信水平上没有差异; 如果差值的均值在置信区间外, 则说明两个总体均值在给定的置信水平上有差异。

19、什么是简单(一元线性)回归分析? 其作用是什么?

答案: 什么是简单(一元线性)回归分析? 其作用是什么? 简单回归分析是通过一定的数学表达式将两个变量间的线性关系进行描述, 确定自变量的变化对因变量的影响, 是进行估计或预测的一种方法, 侧重于考察变量之间的数量伴随关系。(或者简单回归分析是对具有线性相关关系的两个变量之间(其中一个为自变量, 另一个为因变量)数量变化的一般关系进行分析, 确定相应的数学关系式, 以便进行估计或预测。)其作用包括: 一是. 知数据出发, 确定变量之间的数学关系式; 二是对变量间的关系式进行统计检验, 考察自变量是否对因变量具有显著影响; 三是利用所求出的关系式, 根据自变量的取值估计或预测因变量的取值。

20、条形图和直方图有什么区别?

答案: 条形图和直方图有什么区别? 一是适用范围不同, 条形图适用于所有类型数据, 而直方图只适用于数值型数据。二是条形图中条形的宽度是固定的, 是用条形的长度(或高度)表示各类别频数的多少(1 分)而直方图则是用矩形的面积表示各类别频数的多少, 矩形的宽度和高度均有意义, 宽度表示组距, 高度表示每一组的频数密度, 而且宽度不一定是固定的, 只有等距分组时宽度是固定的。三是条形图中各条形是分开排列的, 而直方图中由于分组数据是连续的, 因而直方图的矩形通常是紧密排列的。

21、一元线性回归模型中有哪些基本假定?

答案: 一元线性回归模型中有哪些基本假定? 一元线性回归模型的三个基本假定为: 一是误差项是一个期望值为 0 的随机变量; 二是对于所有的 x 值, 误差项的方差都是相等的; 三是误差项是一个服从正态分布的随机变量, 相互独立, 其余自变量 x 不相关。

名词解释(24)一

1、Z 值一 Z 值又称为标准分数, 它是以平均数为参照点, 以标准差为单位的描述原始数据在总体中相对位置的量数, 通过计算 Z 值将一般正态分布转换为标准正态分布, Z 值的计算公式为: $Z = (X - \mu) / \sigma$ 。

2、参数与统计量一 \rightarrow 参数与统计量参数是研究者想要对解的总体的某种特征值。通常情况下, 我们的总体的参数主要有总体平均数、标准差、比例等。统计量是根据样本数据计算出来的一个量。与参数类似, 我们通常关心的样本统计量主要有样本平均数、样本标准差、样本比例等。

3、抽样单位与抽样框一 \rightarrow 抽样单位与抽样框抽样单位就是一次直接的抽样所使用的基本单位。抽样框是指一次直接抽样时所有抽样单位的名单。

4、抽样分布一 \rightarrow 抽样分布是指样本统计量的概率分布, 它是在重复选取容量为 n 的样本时, 由每个样本计算出来的统计量值的相对频数分布。

5、独立样本与配对样本一 \rightarrow 独立样本与配对样本独立样本是指我们得到的样本是相互独立的。配对样本就是一个样本中的数据与另一个样本中的数据相对应的两个样本。配对样本可以消除由于样本指定的不公平造成的差异。

6、二维表一 \rightarrow 二维表: 二维表就是行列交叉的表格, 将两个变量一个分行排放, 一个分列排放, 行列交叉处就是向属于两个变量的不同类的数据, 也称为列联表。

7、方差与标准差一 \rightarrow 方差与标准差方差 (variance) 是各数值与均值离差平方的平均数, 它是数值数据离散趋势最主要的测量值。标准差 (standard variance) 是方差的平方根, 用于测量数值型数据离散趋势。

8、非概率抽样

非概率抽样: 根据主观意愿、实际情况等进行抽样, 而不依据随机原则进行抽样, 这些不符合概率抽样要求的抽样都称为非概率抽样。(5 分)

9、概率抽样一 \rightarrow 概率抽样就是按照随机原则进行的抽样, 总体中每个个体都有一定的非零的概率入选样本, 并且入选样本的概率都是 C . 知的或可以计算的。

10、卡方检验一 \rightarrow 卡方检验: 卡方检验是对样本的频数分布所来自的总体分布是否服从某种理论分布或某种假设分布所作的假设检验. 即根据样本的频数分布来推断总体的分布。

11、离散变量与连续变量一 \rightarrow 离散变量与连续变量: 如果一个变量的变量值是间断的, 可以一一列举的, 这种变量称为离散变量; 如果一个变量的变量值是连续不断的, 即可以取无数多个数值, 这种变量称为连续变量。

12、离散系数一 \rightarrow 离散系数: 离散系数是一组数据的标准差与该组数据的均值之比, 也称为变异系数。

13、普查一 \rightarrow 普查是一种专门的调查, 它是为对某种特定的目的而对总体中所有的个体进行的一次全面调查. 普查一般是一次性或周期性的, 有统一的标准调查时点。普查具有信息全面、完整的特点, 可以为其它抽样调查提供依据。但是, 普查比较耗时、费力, 成本比较高, 而且适用范围比较难, 调查内容不深入, 只适合调查一些基本的、一般的社会现象。

14、普查与抽样调查一 \rightarrow 普查与抽样调查普查是一种专门的调查, 它是为对某种特定的目的而对总体中所有的个体进行的一次全面调查。抽样调查是从总体中选取部分个体组成样本进行调查的一种方式, 其目的在于根据样本的调查结果推断总体特征。

15、区间估计一 \rightarrow 区间估计: 区间估计是对总体参数估计的一种方法, 根据统计量的抽样分布的特点进行估计。相对于点估计而言区间估计是在估计总体参数时给出一个区间, 同时给出总体参数落入这一区间的可能性大小, 区间估计是在点估计的基础上进行的, 区间估计中的区间是根据样本统计量的抽样分布计算得到的, 这样使得估计 tt 更加准确。如果虚无假设不成立, 我们就拒绝虚无假设, 需要在另一个假设中进行选择, 这就是替换假设。

16、散点图一 \rightarrow 散点图散点图是在坐标系中, 用错轴表示自变量 x , 用 Y 轴表示因变量 y , 而变量组 (x, y) 则用坐标系中的点表示, 不同的变量组在坐标系中形成不同的散点, 用坐标系及其坐标系中的散点形成的二维图就是散点图。散点图是描述变量关系的一种直观方法, 我们可以从散点图中直观的看出两个变量之间是否存在相关关系、是正线性相关还是负线性相关, 也可以大致看出变量之间关系强度如何。

17、误差减少比例一 \rightarrow 误差减少比例: 在预测变量 Y 的值时, 知道变量 X 的值时所减少的误差 $(E1-E2)$ 与总误差 $E1$ 的比值称为误差减少比例, 简称 PRE。PRE 的取值范围为 $0 \rightarrow 1$, PRE 值越大, 说明用变量 X 去预测变量 Y 时能够减少的误差所占的比例越大, 即变量 X 与变量 Y 之间的相关性越大; 反之, PRE 越小, 说明变量 X 与变量 Y 之间的关系越弱。

18、相关系数一 \rightarrow 相关系数是对变量之间线性相关关系的程度和方向的度量。当研究两个变量之间的相关关系时我们称之为简单相关系数。相关系数 r 的取值在 $-1 \sim 1$ 之间。当 $r=1$ 时, 说明两变量之间存在完全正相关; 而当 $r=-1$ 时, 两变量之间完全负相关。当 $0 < r < 1$ 时, 两变量之间存在正线性相关; 而当 $-1 < r < 0$ 时, 两变量之间存在负线性相关。相关系数 $r=0$ 时, 只能说变量之间不存在线性相关, 而不能说它们之间一定不相关。相关系数 r 的取值在 $-1 \sim 1$ 之间。当 $r=1$ 时, 说明两变量之间存在完全正相关; 而当 $r=-1$ 时, 两变量之间完全负相关。当 $0 < r < 1$ 时, 两变量之间存在正线性相关; 而当 $-1 < r < 0$ 时, 两变量之间存在负线性相关。相关系数 $r=0$ 时, 只能说变量之间不存在线性相关, 而不能说它们之间一定不相关。

19、虚无假设与替换假设一 \rightarrow 虚无假设与替换假设 t 进行假设检验时, 我们将需要通过样本信息来推断正确与否的命题称为虚元假设, 也成为原假设或者零假设。样本就是从总体中按照一定方式抽取的一部分个体的集合。

20、整群抽样→整群抽样：先将总体按照某种标志或特征划分为一些子群体，然后从总体中随机抽取些子群体，再将这些抽出的若干小群体内的所有元素构成总体样本的方法。

21、正态分布→正态分布连续性随机变量中重要的分布是钟型概率分布，就是正态分布，也称为常态分布，是一种连续型随机变量的概率分布。正态分布是对称的，且正态分布的中央点最高。

22、置信水平

置信水平：置信水平就是将构造置信区间的步骤重复很多次，置信区间包含总体参数真值的次数所占的比例。(5分)

23、组间均方→是用各误差除以它们的自由度。组间均方记作 MSA，就是用组间平方和除以它的自由度，其自由度为因素水平个数减去 1，即 k-1。其计算公式为：MSA=组间平方和/自由度=SSA/(k-1)。

24、最小二乘法

答案：最小二乘法对于存在线性关系的变量 x 和 y 的观察值，我们可以用很多直线去描述，但我们需要选用距离各观测值最近的一条直线，用它来描述 x 与 y 之间的关系时实际的误差最小，根据这一思想来确定回归方程中参数的方法就是最小二乘法。(2分)最小二乘法是使因变量的观察值与

估计值之间的离差平方和达到最小来求参数 β_0 和 β_1 的方法。

计算题(17)一

- 1、甲单位人均月收入 4500 元，标准差 1200 元。...
- 2、某单位为研究其商品的广告费用（到对其销售量...）
- 3、某调查公司研究出租司机每天收入（元）与行驶里...）
- 4、某公司欲了解广告费用 x 对销售量 y 的影响，收集...）
- 5、某公司招聘职员时，要求对职员进行两项基本能...）
- 6、某农科院使用 4 种方法培育稻米，为确定哪种方...）
- 7、某企业使用 3 种方法组装一种新的产品，为确定...）
- 8、某汽车生产商欲了解广告费用（万元）对销售量（辆...）
- 9、某校社会学专业共有两个班级。...
- 10、某校社会学专业同学统计课成绩如下表所示。...
- 11、某行业管理局所属 40 个企业 2011 年产品销售额...）
- 12、为估计每个网络用户每天上网的平均时间是多...）
- 13、为研究某种商品的价格(x)对其销售量(y)的...）
- 14、为研究某种商品的价格(x)对其销售量(y)的...）
- 15、为研究某种商品的价格(x)对其销售量(y)的影响，收...）
- 16、一项关于大学生体重状况的研究发现，男生的平...）
- 17、一项关于大学生体重状况的研究发现，男生的平...）

1、甲单位人均月收入 4500 元，标准差 1200 元。乙单位月收入分布如下所示。

乙单位月收入分布表

按收入分组(元)	人数(个)
3000 分以下	120
3000~4000	420
4000~5000	540
5000~6000	420
6000 以上	300
合计	1800

(1) 计算乙单位员工月收入的均值和标准差。(2) 比较甲单位和乙单位哪个单位员工月收入的离散程度更大?

答案:

$$\begin{aligned} (1) \text{均值: } \bar{X} &= \frac{f_1 X_1 + f_2 X_2 + \dots + f_k X_k}{f_1 + f_2 + \dots + f_k} \\ &= (2500 \times 120 + 3500 \times 420 + 4500 \times 540 + 5500 \times 420 + 6500 \times 300) \div 1800 \\ &= 8460000 \div 1800 \\ &= 4700 \quad (4 \text{分}) \end{aligned}$$

方差:

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^k (X_i - \bar{X})^2 f_i}{N} \\ &= \frac{[(2500 - 4700)^2 \times 120 + (3500 - 4700)^2 \times 420 + (4500 - 4700)^2 \times 540 + (5500 - 4700)^2 \times 420 + (6500 - 4700)^2 \times 300]}{1800} \\ &= 2448000000 \div 1800 = 1360000 \end{aligned}$$

标准差: $\sigma = \sqrt{\sigma^2} = \sqrt{1360000} = 1166.19$ (4分)

(2) 甲单位人均月收入的离散系数为:

$$v_{\#} = \frac{S_{\#}}{X_{\#}} = 1200 \div 4500 = 0.2667 \quad (3 \text{分})$$

乙单位人均月收入的离散系数为:

$$v_z = \frac{S_z}{X_z} = 1166.19 \div 4700 = 0.2481 \quad (3 \text{分})$$

$V_{\#} > V_z$, 所以说甲单位人均月收入的离散程度大于乙单位人均月收入。(1分)

2、某单位为研究其商品的广告费用（到对其销售量（y）的影响，收集了过去 12 年的有关数据。

某单位为研究其商品的广告费用（到对其销售量（y）的影响，收集了过去 12 年的有关数据。通过分析得到以下结果:

方差分析表

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	B	D	F	0.000
残差	40158.08	C	E	—	—
总计	A	11	—	—	—

要求: (1) 计算上面方差分析表中 A、B、C、D、E、F 处的值。(2) 商品销售量的变差中有多少是由广告费用的差异引起的? (3) 销售量与广告费用之间的相关系数是多少?

答案: (1) 方差分析表: (12 分)

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	1	1602708.6	399.1	0.000
残差	40158.08	10	4015.808	—	—
总计	1642866.68	11	—	—	—

$$(2) R^2 = \frac{SSR}{SST} = \frac{1602708.6}{1642866.68} = 0.9756$$

即商品销售量的变差中有 97.56% 是由广告费用引起的。

(2 分)

$$(3) \text{销售量与广告费用之间的相关系数 } r = \sqrt{R^2} = \sqrt{0.9756} = 0.9877$$

(1 分)

3、某调查公司研究出租车司机每天收入(元)与行驶里程(公里)之向的关系。

某调查公司研究出租车司机每天收入(元)与行驶里程(公里)之向的关系。对 30 位出租车司机进行调查, 并根据每天的收入 y 行驶里程 x 进行回归, 得到。方程的截距为 162 回归系数为 0.6。回归平方和 SSR=2600。残差平方和 SSE=513。要求: (1) 写出每天的收入 y 与行驶里程 x 之间的线性回归方程。(2) 假如某司机某天行驶了 300 公里, 很据回归方程估计他该天的收入。(3) 计算判定系数 R², 并解释它的意义。

答案: (1) 回归方程为:

$$y = 162 + 0.6x \quad (4 \text{ 分})$$

(2) 当 x=300 时,

$$y = 162 + 0.6 \times 300 = 342 \text{ (元)} \quad (4 \text{ 分})$$

(3) 判定系数

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} \\ = \frac{2600}{2600 + 513} \\ = 0.8352 \quad (4 \text{ 分})$$

表明在每天收入的总变差中, 有 83.52% 可以由回归方程解释, 说明回归方程的拟合程度很高 (3 分)

4、某公司欲了解广告费用 x 对销售量 y 的影响, 收集了 16 个地区的数据, 并对 x、y 进行线性回归分析,

某公司欲了解广告费用 x 对销售量 y 的影响, 收集了 16 个地区的数据, 并对 x、y 进行线性回归分析, 得到: 方程的截距为 280。回归系数为 1.6。回归平方和 SSR=1503000。残差平方和 SSE=38000。

要求: (1) 写出广告费用 y 与销售量 x 之间的线性回归方程。(2) 假如广告费用投入 80000 元, 根据回归方程估计商品的销售量。(3) 计算判定系数 R², 并解释它的意义。

答案: (1) $y = 280 + 1.6x$ (5 分)

(2) 当 x=80000 时, $y = 280 + 1.6 \times 80000 = 128280$ (5 分)

$$(3) R^2 = \frac{SSR}{SSR + SSE} \\ = \frac{1503000}{1503000 + 38000} \\ = \frac{1503000}{1541000} \\ = 0.9753$$

表明在商品销售量的总变差中, 有 97.53% 可以由回归方程解释, 说明回归方程的拟合程度很高。 (5 分)

5、某公司招聘职员时, 要求对职员进行两项基本能力测试。

某公司招聘职员时, 要求对职员进行两项基本能力测试。已知, A 项测试中平均分数为 90 分, 标准差是 12 分; B 项测试中平均分数为 200 分, 标准差为 25 分。一位应试者在 A 项测试中得了 102 分, 在 B 项测试中得了 215 分。若两项测试的成绩均服从正态分布, 该位应试者哪一项测试更理想?

$$19. Z_A = \frac{x_A - \bar{x}_A}{S_A} = \frac{102 - 90}{12} = 1 \quad (5 \text{分})$$

$$Z_B = \frac{x_B - \bar{x}_B}{S_B} = \frac{215 - 200}{25} = 0.6 \quad (5 \text{分})$$

答案:

该测试者在八项考试中比平均分高出1个标准差，而在B项考试中比平均分高出0.6个标准差因而，可以说该测试者A项测试比较理想。

6、某农科院使用4种方法培育稻米，为确定哪种方法生产效率最高，随机划出40块试验田，并指定每块试验田使用其中的一种方法。

通过对每块试验田的产量进行分析得到下面的方差分析表。请完成方差分析表并写出你的计算步骤。

变差来源	SS	df	MS	F	Sig.
组间			320		0.000
组内	6048			—	—
总计		39	—	—	—

答案:960(3分)、7008(2分)、3(3分)、36(2分)、168(3分)、1.90(2分)

变差来源	SS	df	MS	F	Sig.
组间	960	3	320	1.90	0.000
组内	6048	36	168	—	—
总计	7008	39	—	—	—

变差来源	SS	df	MS	F	Sig.
组间			320		0.000
组内	6048			—	—
总计		39	—	—	—

答案: 960(3分)、7008(2分)、3(3分)、36(2分)、168(3分)、1.90(2分)

变差来源	SS	df	MS	F	Sig.
组间	960	3	320	1.90	0.000
组内	6048	36	168	—	—
总计	7008	39	—	—	—

7、某企业使用3种方法组装一种新的产品，为确定哪种方法生产效率最高，随机抽取30名工人，并指定每人使用其中的一种方法。

某企业使用3种方法组装一种新的产品，为确定哪种方法生产效率最高，随机抽取30名工人，并指定每人使用其中的一种方法。通过对每个工人生产的产品数进行分析得到下面的方差分析表。请完成方差分析表。

变差来源	SS	df	MS	F	Sig.
组间			210		0.000
组内	3836			—	—
总计		29	—	—	—

答案: 15分)

变差来源	SS	df	MS	F	Sig.
组间	420	2	210	1.478	0.000
组内	3836	27	142.07	—	—
总计	4256	29	—	—	—

8、某汽车生产商欲了解广告费用(万元)对销售量(辆)的影响。

某汽车生产商欲了解广告费用(万元)对销售量(辆)的影响。收集了过去12年的有关数据，通过分析得到2方程的截距为363，回归系数为1.42，回归平方和SSR=1600，残差平方和SSE=450。要求：

- 写出销售量y与广告费用x之间的线性回归方程。
- 假如明年计划投入广告费用为50万，根据回归方程估计明年汽车销售量。
- 计算判定系数R²，并解释它的意义。

答案：(1) 回归方程为：y=363+1.42x (4分)

(2) 当x=50时，y=363+1.42x=434(辆) (4分)

(3) 判定系数

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

=1600 / (1600+450)

=0.7805 (4分)

在汽车销售量的总变差中，有78.05%可以由回归方程解释，说明回归方程的拟合程度很高。(3分)

9、某校社会学专业共有两个班级。

某校社会学专业共有两个班级。期末考试时，一班同学社会学理论平均成绩为77分，标准差为12分，二班同学成绩如下所示。

二班同学社会学理论成绩分组数据表

按成绩分组(分)	人数(个)
60分以下	2
60~70	7
70~80	9
80~90	7
90~100	5
合计	30

要求：(1) 计算二班同学考试成绩的均值和标准差，(2) 比较一班和二班哪个班成绩离散程度更大，(提示：使用离散系数)

答案：(1) 均值：

$$\begin{aligned} \bar{X} &= \frac{f_1 X_1 + f_2 X_2 + \dots + f_k X_k}{f_1 + f_2 + \dots + f_k} \\ &= (55 \times 2 + 65 \times 7 + 75 \times 9 + 85 \times 7 + 95 \times 5) \div 30 \\ &= 2310 \div 30 \\ &= 77 \quad (4分) \end{aligned}$$

方差：

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^k (X_i - \bar{X})^2 f_i}{N} \\ &= \frac{[(55-77)^2 \times 2 + (65-77)^2 \times 7 + (75-77)^2 \times 9 + (85-77)^2 \times 7 + (95-77)^2 \times 5]}{30} \\ &= 4080 \div 30 \\ &= 136 \end{aligned}$$

标准差：

$$\sigma = \sqrt{\sigma^2} = \sqrt{136} \approx 11.6619 \quad (4分)$$

(2) 一班考试成绩的离散系数为：

$$v_{-班} = \frac{S_{-班}}{X_{-班}} = 12 \div 77 = 0.1559 \quad (3分)$$

二班考试成绩的离散系数为：

$$v_{二班} = \frac{S_{二班}}{X_{二班}} = 11.66 \div 77 = 0.1515 \quad (3分)$$

$v_{-班} < v_{二班}$ ，所以说一班成绩的离散程度小于二班。(1分)

10、某校社会学专业同学统计课成绩如下表所示。

某校社会学专业同学统计课成绩如下表所示。

社会学专业同学统计课成绩表

学号	成绩	学号	成绩	学号	成绩
101023	76	101037	75	101052	70
101024	91	101038	70	101053	88
101025	87	101039	76	101054	93
101026	78	101040	90	101055	62
101027	85	101041	76	101056	95
101028	96	101042	86	101057	95
101029	87	101043	97	101058	66
101030	86	101044	93	101059	82
101031	90	101045	92	101060	79
101032	91	101046	82	101061	76
101033	80	101047	80	101062	76
101034	81	101048	90	101063	68
101035	80	101049	88	101064	94
101036	83	101050	77	101065	83

要求：(1) 对考试成绩按由低到高进行排序，求出众数、中位数和平均数。
 (2) 对考试成绩进行适当分组，编制频数分布表，并计算累计频数和累计频率。

答案：(1) 考试成绩由低到高排序：

62,66,68,70,70,75,76,76,76,76,76,77,78,79,
 80,80,80,81,82,82,83,83,85,86,86,87,87,88,
 88,90,90,90,91,91,92,93,93,94,95,95,96,97 (1分)

众数:76 (2分)

中位数:83 (2分)

$$\begin{aligned} \text{平均数: } \bar{X} &= \frac{X_1 + X_2 + \dots + X_N}{N} \\ &= (62 + 66 + \dots + 96 + 97) \div 42 \\ &= 3490 \div 42 \\ &= 83.095 \quad (2分) \end{aligned}$$

(2)

按成绩 分组	频数	频率(%)	向上累积		向下累积	
			频数	频率(%)	频数	频率(%)
60—69	3	7.143	3	7.143	42	100.000
70—79	11	26.190	14	33.333	39	92.857
80—89	15	35.714	29	69.048	28	66.667
90—99	13	30.952	42	100.000	13	30.952

11、某行业管理局所属 40 个企业 2011 年产品销售额数据如下所示

某行业管理局所属 40 个企业 2011 年产品销售额数据如下所示。

40 个企业 2011 年产品销售额

企业编号	销售额	企业编号	销售额	企业编号	销售额	企业编号	销售额
1	152	11	105	21	103	31	136
2	105	12	123	22	103	32	146
3	117	13	116	23	137	33	127
4	97	14	115	24	138	34	135
5	124	15	110	25	91	35	117
6	119	16	115	26	118	36	113
7	108	17	100	27	120	37	104
8	88	18	87	28	112	38	125
9	129	19	107	29	95	39	108
10	115	20	119	30	142	40	126

要求：

(1) 对 2011 年销售额按由低到高进行排序，求出众数、中位数和平均数。

(2) 如果按照规定，销售额在 125 万元以上的为先进企业，115 万—125 万之间的为良好企业，105 万—115 万之间的为一般企业，105 万以下的为落后企业，请按先进企业、良好企业、一般企业、落后企业进行分组，编制频数分布表，并计算累积频数和累积频率。

答案：(1) 销售额由低到高排序：(5 分)

87、88、91、95、97、100、103、103、104、105、105、107、108、108、110、112、113、115、115、115、116、117、117、118、119、119、120、123、124、125、126、127、129、135、136、137、138、142、146、152

众数：115 (1 分)

中位数：115.5 (1 分)

$$\text{平均数: } \bar{X} = \frac{X_1 + X_2 + \dots + X_n}{N}$$

$$= (152 + 146 + \dots + 88 + 87) \div 40$$

$$= 4647 \div 40$$

$$= 116.175 \quad (3 \text{ 分})$$

(2) 40 个企业分组表 (5 分)

按销售额分组(万元)	企业个数	累计频数	频率(%)	累计频率
先进企业	11	11	27.5	27.5
良好企业	12	23	30	57.5
一般企业	8	31	20	77.5
落后企业	9	40	22.5	100
合计	40		100	

12、为估计每个网络用户每天上网的平均时间是多少，抽取了 225 个网络用户的简单随机样本，得到样本均值为 6.5 个小时，样本标准差为 2.5 个小时。

为估计每个网络用户每天上网的平均时间是多少，抽取了 225 个网络用户的简单随机样本，得到样本均值为 6.5 个小时，样本标准差为 2.5 个小时。(1) 试用 95% 的置信水平，计算网络用户每天平均上网时间的置信区间。(2) 在所调查的 225 个网络用户中，年龄在 20 岁以下的用户为 90 个。以 95% 的置信水平，计算年龄在 20 岁以下的网络用户比例的置信区间。

注： $Z_{0.025} = 1.96$

答案：

18. (1) 已知： $n=225, \bar{x}=6.5, s=2.5, Z_{0.025}=1.96$ (2 分)

用户平均上网时间的 95% 的置信区间为：

$$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}} = 6.5 \pm 1.96 \times \frac{2.5}{\sqrt{225}} = 6.5 \pm 0.33 \quad (4 \text{ 分})$$

即 (6.17, 6.83) (2 分)

(2) 样本比例： $P = \frac{90}{225} = 0.4$ (2 分)

年龄在 20 岁以下的网络用户比例的置信区间为：

$$P \pm Z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}} = 0.4 \pm 1.96 \times \sqrt{\frac{0.4 \times (1-0.4)}{225}} = 0.4 \pm 0.064 \quad (3 \text{ 分})$$

即 (33.6%, 46.4%) (2 分)

13、为研究某种商品的价格(x)对其销售量(y)的影响，收集了 12 个地区的有关数据。通过分析得到以下方差分析表：

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	B	D	F	0.000
残差	40158.08	C	E	—	—
总计	A	11	—	—	—

要求：

(1) 计算上方方差分析表中 A、B、C、D、E、F 处的值。

(2) 商品销售量的变差中有多少是由价格的差异引起的？

答：(1) 方差分析表：(12 分)

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	1	1602708.6	399.1	0.000
残差	40158.08	10	4015.808	—	—
总计	1642866.68	11	—	—	—

$$(2) R^2 = \frac{SSR}{SST} = \frac{1602708.6}{1642866.68} = 0.9756$$

即商品销售量的变差中有 97.56% 是由广告费用引起的。(3 分)

14、为研究某种商品的价格(x)对其销售量(y)的影响，收集了12个地区的有关数据。通过分析得到以下方差分析表：

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	B	D	F	0.000
残差	40158.08	C	E	—	—
总计	A	11	—	—	—

要求：

- 计算上面方差分析表中 A、B、C、D、E、F 处的值。
- 商品销售量的变差中有多少是由价格的差异引起的？

解：(1) 方差分析表：(12分)

变差来源	SS	df	MS	F	Sig.
回归	1602708.6	1	1602708.6	399.1	0.000
残差	40158.08	10	4015.808	—	—
总计	1642866.68	11	—	—	—

$$(2) R^2 = \frac{SSR}{SST} = \frac{1602708.6}{1642866.68} = 0.9756$$

即商品销售量的变差中有 97.56% 是由广告费用引起的。(3分)

15、为研究某种商品的价格(x)对其销售量(y)的影响，收集了12个地区的有关数据。

为研究某种商品的价格(x)对其销售量(y)的影响，收集了12个地区的有关数据。通过分析得到以下结果：

方差分析表

变差来源	SS	df	MS	F	Sig.
回归	A	B	C	D	0.000
残差	205158.07	E	F	—	—
总计	1642866.67	11	—	—	—

要求：(1) 计算上面方差分析表中 A、B、C、D、E、F 处的值。(2) 商品销售量的变差中有多少是由价格的差异引起的？(3) 销售量与价格之间的相关系数是多少？

答案：(1) 方差分析表：

变差来源	SS	df	MS	F	Sig.
回归	1437708.60	1	1437708.60	70.078	0.000
残差	205158.07	10	20515.81	—	—
总计	1642866.67	11	—	—	—

(7分)

$$(2) R^2 = \frac{SSR}{SST} = \frac{1437708.60}{1642866.67} = 0.875122 \quad (3分)$$

即商品销售量的变差中有 86.6% 是由价格引起的。(2分)

$$(3) r = \sqrt{R^2} = \sqrt{0.875122} = 0.9355 \quad (3分)$$

16、一项关于大学生体重状况的研究发现，男生的平均体重为 60 千克，标准差为 5 千克

一项关于大学生体重状况的研究发现，男生的平均体重为 60 千克，标准差为 5 千克；女生的平均体重为 50 千克，标准差为 5 千克。请问：(1) 是男生体重差异大还是女生体重差异大？为什么？

(2) 男生中有多少比重的人体重在 55 千克—65 千克之间？(3) 女生中有多少比重的人体重在 40 千克—60 千克之间？

答案：

$$(1) v_{男} = \frac{S_{男}}{x_{男}} = \frac{5}{60} = 0.083 \quad (4分)$$

$$v_{女} = \frac{S_{女}}{x_{女}} = \frac{5}{50} = 0.1 \quad (4分)$$

$v_{男} < v_{女}$ ，所以女生体重差异比男生大。(1分)

(2) 68% 的男生体重在 55 千克—65 千克之间。(3分) (3) 95% 的女生体重在 40 千克—60 千克之间。(3分)

17、一项关于大学生体重状况的研究发现，男生的平均体重为 60 千克，标准差为 5 千克

一项关于大学生体重状况的研究发现，男生的平均体重为 60 千克，标准差为 5 千克；女生的平均体重为 50 千克，标准差为 5 千克。请问：(1) 是男生体重差异大还是女生体重差异大？为什么？

(2) 男生中有多少比重的人体重在 55 千克—65 千克之间？(3) 女生中有多少比重的人体重在 40 千克—60 千克之间？

备注： $\Phi(1)=0.8413$ ， $\Phi(2)=0.9772$

答案：(1)

$$V_{\text{男}} = \frac{S_{\text{男}}}{x_{\text{男}}} = \frac{5}{60} = 0.083 \quad (3 \text{分}) \quad V_{\text{女}} = \frac{S_{\text{女}}}{x_{\text{女}}} = \frac{5}{50} = 0.1 \quad (3 \text{分})$$

$V_{\text{男}} < V_{\text{女}}$ ，所以女生体重差异比男生大。(1分)

$$\begin{aligned} (2) P(55 < X < 65) &= \Phi[(65-60)/5] - \Phi[(55-60)/5] \\ &= \Phi(1) - \Phi(-1) = \Phi(1) - [1 - \Phi(+1)] \\ &= 2\Phi(1) - 1 \quad (3 \text{分}) \end{aligned}$$

由题得， $\Phi(1)=0.8413$ ，所以 $2\Phi(1)-1=0.6826$ ，即 68% 的男生体重在 55 千克—65 千克之间。(1分)

$$\begin{aligned} (3) P(55 < X < 65) &= \Phi[(60-50)/5] - \Phi[(40-50)/5] \\ &= \Phi(2) - \Phi(-2) = \Phi(2) - [1 - \Phi(+2)] \\ &= 2\Phi(2) - 1 \quad (3 \text{分}) \end{aligned}$$

由题得， $\Phi(2)=0.9772$ ，所以 $2\Phi(2)-1=0.9544$ ，95% 的女生体重在 40 千克—60 千克之间。(1分)